Content-based retrieval of visual information

Oerlemans, A.A.J.

Chapter 1

Introduction

We live in an Age of Information, a period in time where almost limitless amounts of information are available from a multitude of sources containing text, images, video, audio and other types of information. Take for example Facebook, which has over 10 billion photos, or Google, which has indexed tens of billions of web-pages, or YouTube, which hosts over 140 million videos. Beyond these publicly available sources, there are for example the digitized contents of the libraries and museums worldwide.

Storing this information in a database is not enough to take advantage of the knowledge stored in this data. We also need to be able to search through it. In many situations, text annotation is incomplete or missing in which case it is necessary to turn to content analysis techniques, that is, methods which analyze the pictorial content of the media. It is also noteworthy that even when text annotation is available, it may be possible to improve the quality of search results by also using the pictorial content information.

Searching through digital data is a very active field of research. For each of the types of digital information, specific search methods exist and this thesis aims to add to that research by exploring a specific part of searching in digital information: content-based image retrieval (CBIR). In this type of searching, the pictorial contents of images are automatically analyzed and indexed, to allow search methods to use these contents, instead of relying on descriptions.

In this thesis, we extend existing methods for performing content-analysis of images, but we also try to extend the search process itself by adding an interactive component, which is called relevance feedback. We also look at relevance feedback procedures in video-analysis, or more specifically, object tracking.

When searching for information, a user in general starts with supplying a description of what the user wants to find, also known as the query. The search engine processes the query and presents the results to the user. These results are possibly ranked by relevance, which is usually the similarity of the search results

to the query. This is a very common way of searching, used by all the well-known text search engines on the Internet.

The most widely used method of searching on the Internet is text-based searching. The user supplies a set of descriptive words and the search engine retrieves documents that contain these words.

The common technique for text based searching, is the inverted index [48]. An inverted index contains all known words from the documents in the database and for each word it contains a list of documents that contain that specific word. Search speeds are greatly improved, because not every document has to be compared to the query.

Relevance feedback was originally designed to extend the search process by asking the user to give feedback on the search results to the search system. The search system can combine this feedback with the original query to run a new search with hopefully more relevant results.

An example of this would be a person asking for a book about Africa in a book store. Initially, the employee of the book store will come up with a few books that have Africa as their topic. However, after looking at these books, the customer decides that some of these books describe the African culture and some others describe the species of animals that live there. At this point the customer decides that it was actually these species of animals that he or she was interested in and not the culture.

The customer points out a book that is more like what he or she was looking for and also points out another book that does not contain the desired type of information. Now the person working at the book store knows more about the type of books the customer is looking for and can search for another set of books to show to the customer. Essentially, by giving feedback, the query has been changed in a direction that will result in better search results.

The original relevance feedback algorithm was designed in 1971 by J.J. Rocchio [65] and it was applied to text-based searching. Later, Salton and Buckley [69] improved the original formula, to get to the following result:

$$Q_{i+1} = \alpha Q_i + \beta \sum_{rel} \frac{D_i}{|D_i|} - \gamma \sum_{nonrel} \frac{D_i}{|D_i|} \tag{1.1}$$

In words, this means that the query is adjusted by including knowledge of relevant and non-relevant documents. The new query Q is based on a weighted sum of the previous query and the relevant and non-relevant documents that were selected by the user. Eventually, this process will result in a query point that is at the optimal location for separating the two classes of relevant and non-relevant documents.

For text based searches, this translates to using a weight vector for the words that are used for the document retrieval. Relevant documents increase weights on certain words (or even add new words) and non-relevant documents decrease the weights on other words.

1.1 Content-based image retrieval

Content-based image retrieval (CBIR) uses the actual pictorial contents of images in the search process. In this case, the query is an image, instead of a set of words. The search system uses contents of the image to search for matching images.

There are many reasons to use image contents in the search process, instead of user-supplied tags. Some of them are:

- *Tags could be missing*
  A set of pictures taken at a vacation, is usually not tagged. The entire set probably has a description, but the contents of each individual image are not described.
- *Tags could be incorrect or not descriptive of the contents*
  Users may supply tags that are incorrect, for example, these could be a representation of the situation in which the picture was taken, but not what can be seen on the picture.

  Note that users still might want to search for these higher-level descriptions. This is probably one step further than CBIR, because in this case the contents of images are linked to a notion of a situation or a location. (Photos taken of the crowd at the inauguration of Obama will probably not show the president, but when using this image as input for a search, people expect the search engine to return images of a crowd at this specific event only.)
- *Tags are not always able to capture the true contents*
  For example, for more complex textures such as a view of the Rocky Mountains, there are no words that truly describe the image contents.

These examples explain the need for using different techniques than text-based searching. Content-based techniques do not depend on external descriptions to perform a search task. However, it is also possible to combine the two types of searching.

The contents of images can be analyzed in various ways. Low-level features such as color, texture and shape are commonly used, but higher level features, or concepts are also available for describing images.

A good overview of the history of CBIR systems is given by Veltkamp et al. [91] and Smeulders et al. [83]. However, we would like to emphasize a few notable systems from the past.

QBIC (Query-By-Image-Content) [16] was developed by IBM and presented in 1995 as one of the first systems that enabled searching through image and video databases based on image contents. Even today, the QBIC technology is still commercially used in DB2, a data-management product by IBM. The system can use image properties such as color percentages, color layout and textures in the search process.

In the same year, Chabot [59] was presented. By integrating image information

stored in a database, which can be text and other data types, in combination with properties of image contents, the user can search for 'concepts'.

One of the first systems that used relevance feedback in an image retrieval system was MARS, Multimedia Analysis and Retrieval System. It was first demonstrated in 1996 as a basic image retrieval system [29] by Thomas Huang and was later extended with the relevance feedback component [68] by Rui.

The ImageScape image retrieval system [37] by Michael Lew, used several methods for searching through images, one of them being query by icons, a method that used predefined visual concepts, which made it one of the first systems to use visual concepts for image retrieval. The concepts could be placed on a canvas by the user in the form of icons and the system would then retrieve images for which the concept was detected at the user-specified locations.

In content-based retrieval, several promising research directions have emerged. Some try to reduce content-based searching to text-based searching, others focus on the problems of interest point detection or sub-image searching and yet another direction is the use of relevance feedback techniques.

Usually, image database lack user-supplied tags, so automatically tagging these would be a desirable option. As described in [41], real-time automated tagging of images is already a promising research direction. This research combines low-level features into a concept that can be described with words. Searching for images is then reduced to text-based searching. The query image is translated into tags (in real-time) and the database is queried for the best matching concepts.

Interest points are locations within an image that can be automatically calculated and that define the best input for other algorithms, such as object matching, tracking and image retrieval.

One of the earliest interest point detectors was Moravecs corner detector [52]. Other well-known more recent algorithms are SIFT [44] and SURF [1].

Searching for images, or image contents, is not bound by the area of the entire image. The query contents can be part of a larger image. The research area for sub-image searching tries to solve this problem by subdividing database images into smaller subimages that can be matched to the query.

The same method can also be applied by subdividing the query image into sub-images and to use these as separate queries. The ImageScape system [37] did this by handling each user-placed icon as a query for a visual concept.

Some of the challenges in this research area are:

- How can we subdivide an image into regions that are meaningful to be used for sub-image searching
- What features can be used to describe the sub-images so that they can be matched to other sub-images, that possibly have different shapes or sizes

In a CBIR task, the text-based relevance feedback process can be translated to changing the image contents that the user is searching for. The user-supplied

image is combined with feedback on the search results, resulting in a virtual query image that contains elements of both the user input and the feedback images.

As an example: if the original query contained the color green and a round shape, but the user has given positive feedback for an image that contains the color blue, the new query would probably result in images that contain the color blue and round shapes.

1.2 Research areas in CBIR

This paragraph describes some of the topics in content based image retrieval that have drawn the attention of researchers in previous years and it introduces a few challenges of CBIR that will probably be the subject of many research projects in the future.

1.2.1 Image segmentation

In partial image searches, the question is how to define the image parts. A straightforward way would be to linearly divide the image into several rectangular regions, but this will have problems in that real object boundaries will rarely coincide with the rectangular regions. A better way would be to use image properties as a segmentation guide, so that the segmented regions have the same properties. There are several ways of selecting segmentation properties, but image intensity, color and texture are common choices.

A recent example of such a segmentation method is fuzzy regions [63], used in the FReBIR system.

1.2.2 Curse of dimensionality

One of the first, logical, steps in setting up an image retrieval system is to select a large number of different features, to increase the chance of finding perfectly matching images. For example, one could choose multiple color features to improve color-based matching.

However, there is a downside to increasing the number of features that are used for similarity matching and this is expressed by the 'curse of dimensionality'. This term, which was first mentioned by the mathematician Richard Bellman [2], is used to express the difficulties that arise with using distances between high-dimensional vectors. In high-dimensional spaces, every vector seems to be at a very large distance from any other vector and then the question is, what the usefulness of these distances is in finding the best match based on the selected features.

1.2.3 Semantic gap

In many image retrieval systems, low-level features such as color, texture and shape are commonly used to describe images or parts of images. On the other hand, users tend to think in higher level concepts, such as house, person or desert. (Or even higher level concepts such as 'inauguration of Obama'.) The relation between a set of low-level features and a high-level concept is still a challenge for researchers in the CBIR community and the term 'semantic gap' is often used to describe the lack of a solid theory or methods to overcome this.

In other words, the semantic gap is used to describe the unclear relation, if any, between low-level features and high-level concepts. One would like to say 'if the texture of the area is this and the color is that, there must be a car in this area'. However, there are still no systems that truly bridge the semantic gap by providing these kinds of rules.

1.2.4 Searching with relevance feedback

If only the contents of an image are used as a query for an image retrieval system, ambiguities will definitely arise. A well known saying is 'an image is worth a thousand words' and this also applies to the images that are used as input for image retrieval: one image can have many different meanings to many different users. In other words, two different users may have significantly different goals for their query when the same image is used as a query.

In text-based searches this effect can also be seen, when a word has several meanings, such as 'monitor'. The Wikipedia disambiguation page for monitor lists several different meanings, from the computer monitor to a town in Indiana, US. Without asking the user for feedback, there is no way of knowing what a user is searching for.

1.2.5 Future CBIR challenges

There are many challenges in the field of CBIR research that still need to be addressed. An overview of these challenges was recently given in [40]. The authors conclude that the following five challenges are noteworthy:

- Concept detection in the presence of complex backgrounds
- Multi-modal analysis and retrieval
- Experiential multimedia exploration
- Interactive search
- Performance evaluation

1.3 Thesis contents

This research has focused on two types of digital information: images and video. Chapters 2, 3 and 4 give a general overview of the image features, machine learning techniques and performance evaluation methods that were used. Chapters 5 to 8 contain techniques that are applied to image searching. Chapters 9 and 10 show the results of relevance feedback on object tracking in video. A more detailed description of each chapter is given below.

Chapter 2 gives an overview of existing image features and similarity methods that are used in this research. Chapter 3 gives an overview of the machine learning techniques used in this research. In chapter 4 various performance measures are described that were used to evaluate the experiments.

In Chapter 5 a new interest point detector is presented. The detector uses local dissimilarity to determine the most distinctive points in an area, based on a selected feature or combination of features. We presented this work at the 10th ACM International Conference on Multimedia Information Retrieval (MIR) in Vancouver, Canada in 2008.

Chapter 6 demonstrates the use of relevance feedback for visual concept detection. A visual concept is learned by asking the user for positive and negative examples of the concept. This concept is then used for pointing out parts of images that contain the concept. This contribution was published in the proceedings of the 21st Benelux Artificial Intelligence Conference (BNAIC) in Eindhoven, The Netherlands in 2009.

An improved version of the paper used our new interest point detector combined with an enhanced wavelet representation feature and shows results of experiments on the MIRFLICKR-25000 dataset. This paper was presented at the 11th ACM International conference on Multimedia Information Retrieval (MIR) in Philadelphia, Pennsylvania, USA in 2010.

Chapter 7 presents a novel similarity measure that uses the coincidence of feature values in a training set of similar images and maps this in a 3D space. The resulting surface is used as the similarity measure when searching for new images.

In Chapter 8, a new texture feature is described, which is a generalization of the well-known 3x3 texture unit paradigm, that has shown that the statistical distribution of 3x3 blocks is a very good classifier for textures [25]. The novel texture feature was published in the proceedings of the 6th IEEE International Symposium on Image and Signal Processing and Analysis (ISPA) in Salzburg, Austria in 2009.

Chapter 9 presents a robust, adaptive object tracking system that was presented at the 11th Annual Conference on Computing and Imaging (ASCI) conference in 2005. It was also used as the basis for further research for this thesis.

Chapter 10 builds on the new similarity measure based on multidimensional maximum likelihood. This work was presented at the IEEE International Workshop

on Human Computer Interaction (HCI)in Rio de Janeiro, Brasil in 2007.

Chapter 10 also demonstrates the use of relevance feedback to object tracking. Tracked objects can be selected as positive or negative examples and the tracking system can keep tracking these objects when they are standing still, or it can ignore them. A paper based on this techniques was published in the ACM International Conference on Image and Video Retrieval (CIVR) in Amsterdam, The Netherlands in 2007.

Appendix A describes RetrievalLab, an educational and research tool to illuminate the process of content-based retrieval. RetrievalLab was presented at the ACM International Conference on Multimedia Retrieval (ICMR) in Trento, Italy in 2011.