Content-based image retrieval

An introduction to literature and applications

JÖRGEN AHLBERG, FREDRIK JOHANSSON,
RONNIE JOHANSSON, MAGNUS JÄNDEL,
ANNA LINDERHED, PONTUS SVENSON,
GUSTAV TOLT

FOI, Swedish Defence Research Agency, is a mainly assignment-funded agency under the Ministry of Defence. The core activities are research, method and technology development, as well as studies conducted in the interests of Swedish defence and the safety and security of society. The organisation employs approximately 1000 personnel of whom about 800 are scientists. This makes FOI Sweden's largest research institute. FOI gives its customers access to leading-edge expertise in a large number of fields such as security policy studies, defence and security related analyses, the assessment of various types of threat, systems for control and management of crises, protection against and management of hazardous substances, IT security and the potential offered by new sensors.

FOI
Swedish Defence Research Agency
Information and Aeronautical Systems
SE-164 90 Stockholm

Phone: +46 8 555 030 00
Fax: +46 8 555 031 00

www.foi.se

FOI-R—3395--SE
ISSN 1650-1942

User report
December 2011

Information and Aeronautical Systems

Jörgen Ahlberg, Fredrik Johansson, Ronnie Johansson, Magnus Jändel, Anna Linderhed, Pontus Svenson, Gustav Tolt

Content-based image retrieval

An introduction to literature and applications

Titel	Innehållsbaserad åtkomst av bilddata
Title	Content-based image retrieval

Rapportnr/Report no	FOI-R—3395--SE
Rapporttyp/ Report Type	Användarrapport
Sidor/Pages	44 p
Månad/Month	December
Utgivningsår/Year	2011
ISSN	ISSN 1650-1942
Kund/Customer	FM
Projektnr/Project no	E53353
Godkänd av/Approved by	Lars Höstbeck

FOI, Totalförsvarets Forskningsinstitut

Avdelningen för Informations- och aerosystem

164 90 Stockholm

FOI, Swedish Defence Research Agency

Information and Aeronautical Systems

164 90 Stockholm

Sammanfattning

Att söka i bildsamlingar baserat på visuellt innehåll är potentiellt en mycket kraftfull teknik. Problemområdet benämns Content-Based Image Retrieval, CBIR, och har lockat forskare från olika forskningsområden, bland annat datorseende, artificiell intelligens och maskininlärning. Det relativt unga forskningsområdet CBIR har resulterat i en enorm tillväxt av tillgängliga forskningsartiklar i ämnet de senaste åren. Huvudsyftet med denna rapport är att ge en kort introduktion till CBIR, litteratur och applikationer. Rapporten innehåller en översikt över användbara metoder och presenterar några av de största utmaningarna inom CBIR. De flesta av de föreslagna CBIR-metoderna förlitar sig på ett förbehandlingssteg med feature extraction, som syftar till att utvinna lämpliga bildegenskaper för att framgångsrikt kunna hämta relevanta bilder ur en databas som innehåller tusentals eller miljontals bilder. Även om lågnivå-funktioner som färg, textur och form är direkt relaterade till perceptuella aspekter av bildinnehåll, finns det också högnivå-funktioner i bilder som inte extraheras lika enkelt från det visuella innehållet. Att automatiskt dra semantiskt meningsfulla slutsatser från bilder är en svår utmaning som inte har några perfekta lösningar än. Det finns dock flera försök och förslag om hur man extraherar högnivåbegrepp från lågnivåfunktioner och beskriver dessa med hjälp av ontologier. Ontologier är formella beskrivningar av begrepp och relationer i en domän som används för att överbrygga det semantiska gapet. Hur sökfrågan skapas är viktigt för resultatet av sökningen. Mycket arbete sker på frågespråk för multimediasökningar baserade på metadata. Mindre görs på det svårare problemet att bara använda bilddata i sökfrågan. Ett antal kommersiella system finns på marknaden som har CBIR-kapacitet. Vi ger en begränsad översikt över system med CBIR-kapacitet omfattande prototyper, forskningssystem och kommersiella system. CBIR är också intressant för videodata, där man förutom att tillämpa CBIR-tekniker för enskilda bildrutor kan använda den temporala ordningen på bilderna för att upptäcka vissa handlingar, rörelser eller förändringar. Försvarsmakten har idag tillgång till stora mängder av bild-, video- och filmmaterial från internationella uppdrag, men saknar förmåga att effektivt söka i dessa arkiv. Det finns flera olika tillämpningsområden av intresse för det svenska försvaret, särskilt för counter-IED och analys för flygspaning. En slutsats från denna studie är att medan det faktiskt finns stora potentiella fördelar med att använda CBIR kräver förverkligandet av CBIR-system för militära applikationer att potentiella slutanvändare tar aktiv del i utvecklingen för att ta reda på var CBIR-funktionalitet har störst påverkan. Rapporten avslutas med en kort diskussion om de viktigaste slutsatserna från denna studie och presenterar våra tankar för nästa steg för att utreda behovet av att ge det svenska försvaret CBIR-förmågor. I en bilaga finns sammanfattningar av artiklar, tillsammans med en lista över publikationer som är resultatet av den litteratursökning som genomförts i projektet.

Nyckelord: CBIR, CBIR-system, content-based, image retrieval, information retrieval.

Summary

To search in image and video collections based on visual content is potentially a very powerful technique. Content-Based Image Retrieval, CBIR, has attracted researchers from various research fields: computer vision, artificial intelligence, human factors, and machine learning to name a few. The relatively young age of CBIR as a phenomenon and research area results in an enormous growth of research articles on the topic. The main purpose of this report is to give a brief introduction to the research field of CBIR, the literature and applications. The survey contains an overview of CBIR methods and presents some of the main challenges in CBIR. Most of the suggested CBIR approaches rely on a pre-processing step of feature extraction, aiming at identifying suitable image features to allow for successful retrieval of relevant images from a database containing thousands or millions of images. While low-level features based on colour, texture, and shape are directly related to simple perceptual aspects of image content, there are also higher-level features which are not extracted as easily from pixel data. To automatically derive semantically meaningful features or concepts from images is a hard challenge to which no perfect solutions exist. However, we review attempts and suggestions from the literature. Ontologies are formal descriptions of the concepts and relations in a domain. They are used for bridging the semantic gap between how humans and computers represent the world. Query creation is another vital issue which is critical for the result of the search. Research on multimedia query creation including metadata is thriving but less is done on the harder problem of only using image data in queries. This report includes a limited overview of systems including CBIR capabilities encompassing research prototypes as well as open-source and commercial systems. CBIR also extends into the realms of video data, where, in addition to applying CBIR techniques to individual video frames, the additional time dimension can be explored to detect certain actions, movements or changes. The Swedish Armed Forces today have access to vast amounts of image, video and film material from international missions, but lacks the ability to efficiently search such archives for information. We believe there are several different application areas of interest for the Swedish Armed Forces, particularly for counter-IED analysis, and analysis for air reconnaissance. A conclusion from this study is that while there are indeed large potential benefits of using CBIR, the realization of CBIR systems for military applications requires that potential end-users take active part in the development process in order to work out where CBIR functionality has the greatest impact. The report ends with a brief discussion on the main findings from this study and presents our thoughts on the next steps to be taken in investigating how the Swedish Armed Forces can be provided with relevant CBIR capabilities. An appendix with summaries of studied articles is provided, together with the list of publications resulting from our literature search.

Keywords: CBIR, CBIR-system, content-based, image retrieval, information retrieval

Table of contents

1 Introduction

1.1 Content-based image retrieval

Modern technology has lead to an accelerated growth of digital media collections, often containing both still images and videos. Storage devices are filled with terabytes of digital images[1], making it increasingly harder to retrieve images of interest from such collections. It is clear that search capabilities are needed for finding what we are looking for in such large collections, but how can we make such searches useful? Manual annotation of images with keywords describing the image content can make it easier to find images of interest, but this takes a lot of time, making this approach very costly. It is also only helpful to a certain extent, since we do not always know in advance what kind of searches that will be made in the future. Furthermore, different persons are likely to annotate the same image using different keywords, making it difficult to create a suitable taxonomy and annotate images with the "correct" keywords. For all of the above reasons, the use of content-based image retrieval (CBIR) has been suggested. CBIR is according to Datta el al. (2008) any technology that in principle helps to organize digital picture archives by their *visual content*.

To search in image collections based on visual content is potentially a very powerful technique. Imagine a system in which the user can query the system to retrieve all still images and video frames containing some type of IED, whereupon the system responds by presenting exactly those images. Likewise, imagine the same system, but where the user instead choose to query the system by providing a sample image of a vessel taken by a high-resolution camera, whereupon the system provides all images in the database in which the vessel is present, together with information regarding its previously documented locations. It is our belief that these types of techniques can be of great interest to the Swedish Armed Forces. Some of the aims of this project are therefore to get an overview of what the state-of-the-art of CBIR is, what kind of systems are available on the commercial market or as open source tools already today, what the main current limitations of CBIR are, and what kind of functionality that can be expected in the near future (defined as a couple of years from the present).

1.2 Purpose of this report

The main purpose of this report is to give a brief introduction to the research field of CBIR and how it potentially could be used by the Swedish Armed Forces. There is a vast amount of literature available in this research area; hence this report is not aimed at giving a complete description of the research frontier in CBIR. Instead, a selection of important papers is described, together with a presentation and limited evaluation of some available CBIR systems. The report also includes a chapter that discusses how CBIR could be used to enhance the capabilities of the Swedish Armed Forces.

The intended readers of the report are FOI researchers interested in information retrieval and CBIR, as well as technologically oriented personnel in the Swedish Armed Forces and at the Swedish Defence Materiel Administration (FMV).

The Swedish Armed Forces today have access to vast amounts of video and film material from international missions, but lack the ability to search in such information. Today, the information collected in international missions is therefore often not used, but instead disappears as the missions are rotated. A large amount of information is also available through different international information exchange systems (e.g., BICES). Here, too, the

[1] The term image will throughout the report be used to refer to still images as well as video, unless anything else is explicitly stated. The focus in the report is on still images, but the same techniques can generally be applied to video data that has been decoded to a sequence of still images.

information is not used to its full extent because of search limitations. If these search limitations are overcome, we think that the available information could be better used than is the case today.

The rest of this report is outlined as follows. In Chapter 2, we present a focused survey of the research field of CBIR. The survey contains an overview of low-level and high-level image features and presents some of the main challenges in CBIR. Next, we describe a set of available CBIR systems in Chapter 3, together with overall impressions of these systems. In Chapter 4 we present our general conclusions on the prospects of CBIR. The incorporation of CBIR techniques into the Swedish Armed Forces is discussed in Chapter 5. Finally, we end the report with Chapter 6, in which we briefly discuss the main findings from this study and present our thoughts for the next steps to be taken in investigating the need for providing the Swedish Armed Forces with CBIR capabilities. An appendix with summaries of studied articles is provided, together with the list of publications resulting from the literature search undertaken in this project.

1.3 Terminology

Here we present a list of terms that may be useful when reading the report.

- *Content-based*: indicates that the search will analyze the meaning of the image itself rather than metadata such as keywords, tags, and descriptions associated with the image. Content might refer to low-level features including colour patterns, shapes and textures or high-level features such as meaningful objects and relations between objects.
- *Content-based image retrieval* (CBIR): the application of computer vision techniques to the problem of searching for digital images or video in large databases.
- *Feature:* a summarizing quantity that describes the data or an aspect or part of the data.
- *Information retrieval*: finding documents of an unstructured nature (e.g., text or images) from large collections of data, satisfying some information needs.
- *Relevance feedback*: to involve the user in the information retrieval process for the purpose of improving the final result set.
- *Supervised learning:* the use of machine learning techniques to learn how to divide data into different classes (e.g., determine whether an image represents a landscape or a portrait) based on labelled training data.

2 CBIR research survey

CBIR has attracted researchers from many research fields including computer vision, artificial intelligence, human factors, and machine learning. In spite of, or perhaps because of, the relatively young age of CBIR as a phenomenon and research area, there has been an enormous growth of research articles on the topic. A rather early attempt to structure and review much of the initial work on CBIR (from the beginning of the 90's until year 2000) is found in Smeulders et al. (2000). Much of their findings are already outdated by now, but two very important aspects of CBIR that are highlighted in their article are still very relevant today: the *sensory gap* and the *semantic gap*.

The sensory gap is the unavoidable difference between the real object and the information observers can compile about the object. This gap can never be closed completely but it can to an ever increasing extent be narrowed by improved sensors, signal processing and sensor fusion.

The semantic gap is defined as "[...] the lack of coincidence between the information that one can extract from the visual data (by computational means) and the interpretation that the same data have for a user in a given situation" Smeulders et al. (2000). The semantic gap is hence about the difference between the significance of data for humans and the representation of the same data in information processing systems. As highlighted in the more recent review article by Datta et al. (2008) there has been little progress in bridging the semantic gap at least for generic CBIR-applications.

According to Datta et al., CBIR technology boils down to two fundamental problems:

1) To mathematically describe the image content (its *image signature*).
2) To assess the similarity between a pair of images based on their signatures.

In spite of the apparent simplicity of this, there are significant obstacles that need to be overcome in order to create a useful CBIR system. To find a good image signature using suitable image features is far from easy. We can of course represent an image as an array of pixel values, but this corresponds very poorly to how humans interpret and understand images. Subsections 2.1, 2.2 and 2.3. deal with the problem of building relevant image signatures while section 2.4 addresses the issue of image similarity assessment. The remaining subsections discuss other important aspects of CBIR technology including, ontologies, query creation, performance and user interaction.

2.1 Low-level image features

Most of the suggested CBIR approaches rely on a pre-processing step of *feature extraction*, aiming at extracting suitable image features (*descriptors, properties*) that carry enough information to allow for successful retrieval of relevant images from a database containing thousands or millions of images. A good overview of CBIR-related feature extraction approaches are found in Datta et al. (2008).

The perhaps most intuitive image features are based on *colour*, and such features have been used very frequently from the beginning of CBIR's short history. One of the most common (and simple) colour features is *colour histograms* (see e.g., Heczko et al., 2004). Each bin in a colour histogram represents a range of colours and the value of the bin counts the number of image pixels that fall in the colour range. Colour histograms hence describe the overall colour content of the image. Colour histograms are easily computed and are invariant to image transformations such as rotation and translation, but in order to combat problems related to sensitivity to cropping and the lack of spatial information, various region-based histogram approaches (Wang et al., 2010) have been proposed.

Colour alone is generally not enough to recognize objects; therefore it is important to consider complementing aspects such as *texture* and *shape*. A texture-based CBIR approach, along with an overview of some commonly used texture features (e.g., wavelets and gray-level co-occurrence matrices, GLCMs) is presented in Hazra (2011).

The use of *edge histograms* has also been proposed (Balasubramani and Kannan, 2009). For the purpose of computing an edge histogram the image is first divided into a grid of sub-images. Each bin in an edge histogram represent the occurrence of a significant brightness change in a specific direction in a given sub-image e.g. a vertical edge in the upper left block. Edge histograms describe hence approximately the location and direction of significant intensity changes in the image. The popular SIFT (Scale-Invariant Feature Transform) features (Lowe, 2004), used for describing local image properties around certain *interest* points, is based on similar edge information. Among tools for finding interest points in images, the general purpose corner detector (Harris and Stephens, 1988) and the Difference-of-Gaussians (DoG) operator (Lowe, 2004) are particularly fashionable.

Shape features can be roughly divided into two categories, region-based and contour-based. In the first category, objects of interest are segmented and described with features such as moments, elongation, convexity, etc. Frequently encountered features in the latter category, where the boundary of the object rather than the interior is extracted, may be based on Fourier analysis (Granlund, 1972) or Shape Contexts (Belongie and Malik, 2000). In Sinjur and Zazula (2008), another shape-based retrieval approach is presented, in which objects are associated with a sequence of convex hulls, each described by their respective vertices. A convex hull is a geometrical concept that defines an envelope enclosing a set of vertices. We can imagine the convex hull as a taut rubber sheet supported by the vertices. A "fuzzification" of shape information in order to handle vagueness and uncertain information is presented in (Banerjee et al., 2004).

In general, better results are achieved with local features than with global features (Bartolini et al., 2010). Basically, global features describe the whole image while the local features describe smaller parts of the image (often corresponding to the objects that can be found in the image). According to Liu et al. (2007), most current CBIR systems are region-based, since this is more relevant to the workings of human perception.

2.2 Higher-level image features

While low-level features based on colour, texture, and shape are comparatively easy to compute and at best may be related to low-level human perception, there are also higher-level features in images which are not extracted as easily from pixels. Higher-level features are intended to represent semantically meaningful concepts in the image (e.g., activities taking place in the image or objects seen in the image), which are of more direct interest to a humans.

To automatically derive semantically meaningful features or concepts from images is a hard challenge to which no high-performance generic solutions exist. However, there are various attempts and suggestions in the literature as described in a survey by Liu et al. (2007) where a number of approaches are identified:

1) to use *object ontologies* to define high-level concepts,
2) to use *machine learning methods* to associate low-level features with query concepts,
3) to use *semantic templates* for mapping high-level concepts to low-level visual features.

In Mezaris et al. (2004), an object ontology is used to allow for querying with semantically meaningful concepts without requiring manual annotation of images. The user describes objects using intermediate-level descriptors such as intensity, position, and size. As an

example, the concept “tiger” can be represented as having size=(small. medium) and an intensity consisting of luminance=(high, medium), green-red=(red low, red medium), and blue-yellow=(yellow medium, yellow high), assuming the use of L*a*b* colour space (which is designed to approximate human vision). Once the “tiger” concept has been defined, it can be used for querying, in which the concept's intermediate-level descriptors are compared with all image regions in the database. Searches can be improved using relevance feedback, the process in which the user selects results that are relevant or not relevant to the query, in order to improve the accuracy of the results (see Section 2.9).

Tsai and Hung (2008) review research where machine learning has been used to automatically annotate images. Often used supervised machine learning techniques for learning high-level concepts from lower-level features are support vector machines and Bayesian classifiers (Liu et al., 2007).

Semantic templates are according to Liu et al. (2007) a promising but seldom used technique for semantic-based image retrieval. An example of how it can be used is that a user makes a search using query-by-example, but where keywords/concepts are added to the search. After some iteration with relevance feedback, the system provides images that are judged as relevant to the query. A feature centroid is calculated for the set of relevant images, which becomes a representation of the provided keyword. A potential problem with this type of approach is that much manual work is needed for the system to learn new concepts/keywords.

In order to capture high-level concepts, such as relations between objects in the image, spatial relationships have also been proposed. Wang et al. (2010) address this problem, introduce a number of spatial relationship concepts and propose the use of spatial relationship semantic similarity. Hollink et al. (2004) present a tool for supporting users to add spatial information semi-automatically to images. An interesting observation by Hollink et al. is that people in their test groups often used “three-dimensional” concepts to describe the image, such as “far” and “near”. Since ordinary images do not contain any explicit information about distance to objects, successful extraction of such concepts is particularly challenging. In Ahmed et al. (2008), an approach is proposed for representing the image by a Fuzzy Attributed Relational Graph (FARG) that describes each object in the image, including attributes and spatial relation between the objects. Texture and colour attributes are computed in a way that models the human visual system.

2.3 Image segmentation

As mentioned above, a recent trend in CBIR is to extract local (region-based) features rather than just relying on global features. Earlier approaches such as the ones presented in Niblack et al. (1993), Pentland et al. (1994), and Stricker and Orengo (1995) used only global features such as colour and texture for the whole image, but local features soon started to gain more attention. Examples of this are Carson et al. (2002), Chen and Wang (2002), and Li et al. (2000). This development is natural since the semantic meaning in images is made up of the objects/regions and their relationships, rather than the image as a whole entity. Hence, many contemporary CBIR techniques rely on image segmentation, aiming at dividing the image into its main components (objects). An example of such a technique is Blobworld, presented in Carson et al. (2002). The first step in the Blobworld algorithm is feature extraction encompassing colour, texture, and position features. Next, pixels are grouped into regions by using the well-known Expectation-Maximization (EM) algorithm to estimate a mixture of Gaussians, representing a joint distribution of the features (pixel colour, texture, position). In the reported results, it appears that Blobworld performs better than a global histogram approach in terms of precision and recall in image retrieval tasks on images consisting of distinctive objects such as tigers and zebras. However, for distinctive scenes such as airplane images, the results indicate that the simple global histogram-based approach actually performs better.

There are many other alternatives for how to segment images. A simple image segmentation method is to partition an image into a set of fixed-sized blocks, as explained in, e.g., Tsai and Hung (2008). A downside with this method is that parts of the same object are likely to end up in different partitions. Recently, graph-cut segmentation, based on concepts from graph theory, has been become quite popular in the image analysis field. The segmentation problem is formulated as a graph with nodes representing image pixels. These are connected to each other through edges, which are assigned weights that reflect the similarity between the nodes (pixels). The task is then to segment the image by "cutting" the graph so as to minimize a certain criterion, e.g., the sum of the weights associated with the edges between the different segments (or in practice, to come as close as possible to the optimal solution, as the original problem may typically be computationally intractable). In Malcolm et al. (2007) it is demonstrated how the inclusion of more information than just pixel intensity/colour in the (graph cut) segmentation process can be used to give better segmentation results. The problems of over-segmentation or under-segmentation hamper, however, the use of advanced shape features.

Despite research progress, segmentation accuracy similar to human perception is far from reality. A disadvantage of region-based approaches in general is the increased complexity. Current region-based image retrieval systems are not scalable enough to cope with large image collections (Bartolini et al., 2010).

2.4 Identify similar images

Given a set of image features, be it lower-level or higher-level features, local or global, we need a way to measure the *similarity* between a query image and images in the database. As for feature extraction, a large number of similarity estimation frameworks have been suggested (Datta et al., 2008).

Important aspects to consider when comparing image similarity measures are:

- the degree to which they agree with semantics (i.e., whether "they make sense" for the application of interest)
- how robust they are with respect to perturbations and noise.
- how computationally efficient they are (for example, whether they lend themselves towards indexing techniques such as hashing)

There are three main approaches to image similarity measures.

Distance measures include Euclidean distance, weighted sums, Hausdorff distance, and Kullback-Leibler divergence, to name a few. Obtaining good performance by means of distance measures can be quite difficult since they often relate poorly to what human finds meaningful in the image.

Supervised machine learning algorithms can learn to classify images in meaningful categories. Users just need to provide a sufficiently large and diverse set of training examples. This is particularly attractive for large or complex data sets where the system can learn the spectrum of variations associated with each class or semantic term. Tools such as Support vector machines (SVM) or Bayes classifiers are often used to learn high-level classifications from low-level image features.

Unsupervised machine learning techniques discover categories without human assistance. Such categories are not always well matched to human opinion but they can sometimes reveal similarities that would be difficult for humans to detect. Unsupervised methods are also important for making CBIR scalable to large collections of images. Important unsupervised methods include principal component analysis (PCA), self-organized maps, multi-dimensional indexing techniques (e.g., kd-trees), and data mining techniques (Flores-Pulido et al., 2010).

2.5 Ontologies

Ontologies are formal descriptions of the concepts and relations in a domain. They are used for bridging the *semantic gap* between how humans and computers represent the world. For CBIR it is necessary to have well-defined ontologies for the knowledge domains that are relevant for the application. If the task is to retrieve images of military vehicles from road traffic surveillance data the ontology should e.g. cover vehicles and other objects and relations that are likely to occur in this context. CBIR systems should relate the concepts and relations of the ontology to image features possibly via lower-level ontologies encompassing lower-level image features. Separate ontologies can also describe the queries users can make to the CBIR system. There are some attempts at making ontologies suitable for image and video descriptions, such as ONVIF 1.0, MPEG-7, and the ontologies of PETS, TRECVID (http://trecvid.nist.gov/trecvid.data.html) and I-Lids.

2.6 Query creation

The form of the query is vital for the result of the search. While much present research on multimedia query creation involve metadata search we shall focus on the harder problem of using image data with or without auxiliary text or annotation. The main query types are:

- Query by description where the queries include textual descriptions, names and tags, as well as MPEG-7 descriptions and/or description schemes (MPEG-7 is a standard for describing multimedia content with metadata)

- Query by example in which the query language should support retrieval based on representative examples of the desired content.

- Spatial-temporal queries where the query language supports retrieval based on spatial and/or temporal relationships of image features, e.g., search for images where a red car is in front of a white house.

SQL is the standard querying language for text-based databases, and hence most multimedia query languages developed prior to the standardization of MPEG-7 are derived from SQL. It is very difficult to implement query-by-example using SQL since SQL cannot embed multimedia data as part of the query. The MPEG Query Format (MPQF) is a part of MPEG-7 and was established in 2009. MPQF is an XML-based query language that defines the format of the queries and replies exchanged between clients and servers in a distributed multimedia search and retrieval system. It provides extended functionalities for service discovery, service selection and service capability description (Döller, 2009).

Targeted browsing is a technique for multimedia query-by-example designed especially for mobile devices. Query streaming is used for continually updating queries by adding additional terms to an existing query (Adistambha, 2010). This is similar to searching within a result set with an important difference: a new query is not created; the original query is just updated to include new filtering terms.

2.7 Handling large image databases

Although the main idea of CBIR is to allow for searches in massive image databases, most of the early research has ignored scalability issues of the suggested algorithms and instead been devoted to find good image features, accurate image segmentation techniques, and so on. To test the suggested algorithms, quite small image collections often have been used. However, more lately the problem of handling massive amounts of data has attracted increased research attention. In Batko et al. (2010), a system for web-scale image similarity search is described, that finds similar images in a test collection of 50 million Flickr images. The proposed approach is based on pre-computing five MPEG-7 based descriptors for all images and storing them as an XML structure in a database along with an URL to the original image and an image thumbnail. To provide an easily enlargeable

capacity, the authors propose a distributed similarity search structure based on a peer-to-peer network. The search for similar images (neighbours) uses a technique called M-Chord for partitioning the data space between peers.

A common method for speeding up search is to divide the search space using kd-trees. The idea is to partition the space using hyperplanes so that all points to the left of the hyperplane represent one sub-tree and the points to the right of the hyperplane represent another sub-tree. When the dimensionality of the data is fairly low, kd-trees are quite efficient, but for higher dimensionality the efficiency is not better than straight-forward exhaustive search. Instead, approximate nearest neighbour techniques such as Locality-Sensitive Hashing (LSH) (Wong et al., 2007) or Randomly Projected kd-trees (Wu et al., 2011) can be used. The former approach was demonstrated on a collection of 500,000 images, each represented as a 238-dimensional feature vector. The query time for the proposed LSH method on a standard PC platform was reported to be about 70 ms for the full 500,000 image database and scaled almost linearly with the number of images. Wu et al. reported speed-ups compared to LSH when testing their Randomly Projected kd-trees technique on a database of similar size with 297-dimensional feature vectors. Other (or complementary) strategies to speed up the search for similar images, and also to reduce the computational resources spent on each image, includes using a multi-resolution approach (see e.g. Wichert, 2008).

2.8 Video retrieval

So far, we have focused on CBIR for single images. CBIR is, however, also highly relevant for video data. Applying the image analysis methods that we have discussed frame by frame is possible but engenders significant performance problems for large video databases since compressed video must be decoded before it can be processed as a sequence of still images. The compressed format includes, however, readily available information about the motion flow that potentially could be utilized for CBIR purposes.

Several research groups explore the time dimension of video data for the purpose of detecting actions, movements or changes. Kläser et al. (2010) address the problem of distinguishing between different human actions, although only with moderate success. DARPA´s Mind´s Eye program focuses on automatic recognition of human behaviour in video surveillance data. The target is to capture 48 generic human actions such as e.g. “give”, “follow” and “digging” (Lawlor, 2011). Hu et al. (2007) demonstrate how traffic surveillance video can be processed automatically to cluster the movement of objects into certain paths. By doing so, the system can be queried on a semantic level (e.g., “retrieve all sequences where a car made a U-turn after the traffic light”) or through a sketch of the sought-after path.

A fundamental problem of video analysis is how to handle movement that may have occurred between two frames, so that correspondences between objects (or pixels) in successive images can be established. The problem arises both due to movement of the sensor (maybe on a moving platform) and movement of objects in the scene. In both papers mentioned above, the set-up was quite favourable as no complex movement of the sensor or the objects was involved. Estimating and compensating for sensor motion without the need for external positioning devices (GPS, IMU, etc.) is an active field of research where a great deal of progress has been made recently.

2.9 System-user interaction

Generally, system-user interaction should form an integral component in any modern image retrieval system, rather than just being a last resort when the automatic methods fail. Already from the beginning, interaction can play an important role to help the user navigate through the query space (Smeulders et al., 2000). A fundamental difference

between a computer vision pattern recognition system and a content-based image retrieval system is that a human is an indispensable part of the latter system (Rui et al., 1999).

Most CBIR algorithms and systems rely on the concept of query-by-example, i.e., that the user presents a sample image to the system, which replies by presenting a set of images from its database that in some sense are most similar to the sample image. Whether the user agrees on the selected set of images being similar or not depends to a large extent on which image signature that has been used, and how well this image signature fits the human point of view. But what if we don't have a query image? A few CBIR systems allow the user to construct a hand-drawn image (sketch) that can be used as query image, but this demands some level of artistic talent, and still does not cover all kind of queries a user can be interested in. Other ways in which the user can make searches is by using more high-level concepts as explained briefly in Section 2.2. Such technology is however quite immature, therefore the use of query-by-example is most commonly used, potentially improved by the user feedback

A straightforward way of getting user feedback would be to ask the user to tune the system parameters during the retrieval process, but it is often too complicated for an untrained user to learn and manipulate the detailed parameters of the system.

Relevance feedback (RF) is a query modification technique that attempts to capture the user's needs through iterative feedback and query refinement (Datta et al., 2008). The user grades the relevance of the current retrieval results. The interaction is a complex interplay between the user, the images, and their semantic interpretations (Smeulders, 2000). A practical problem with relevance feedback is, according to Liu et al. (2007), that most current CBIR systems relying on relevance feedback may need five or more iterations for converging to a stable performance level, whereas users often are more impatient than that. Nevertheless, relevance feedback algorithms have overall been shown to provide dramatic performance boosts in retrieval systems (Zhou and Huang, 2003).

2.10 Performance assessment

A fundamental problem in CBIR research is how to evaluate the suggested algorithms and systems. Many papers have been published on everything from low-level feature extraction and image matching criteria to learning mechanisms and search strategies but since much of the presented work includes only very little comparison with other approaches (if any) it is difficult to assess the usefulness of novel techniques and ideas.

Traditional metrics for measuring retrieval performance in CBIR systems are *precision* and *recall* (Liu et al., 2007), where precision is defined as the ratio of the number of relevant images retrieved (true positives) to the number of total retrieved images (true positives + false positives), and recall is the ratio of the number of relevant images retrieved to the number of relevant images in the database (true positives + false negatives). A good system should in theory have high precision and recall, but in practice the recall is often quite low for most CBIR systems (Liu et al, 2007), while the precision often is better. This means that top results often are relevant to the query, while many relevant images in the database are missed. Precision and recall can also be summarized into accuracy (number of images classified correctly divided with the total number of images classified) or be presented in the form of ROC curves or precision vs. recall curves (Müller et al., 2004). Yet other parameters to look at when evaluating CBIR algorithms and systems is the search time and the average normalized rank (Sivic and Zisserman, 2003).

Although it is easy to define performance measures like precision, recall and accuracy in theory, it is harder to measure them in practice. To judge the quality of image retrieval is highly subjective, therefore some common benchmark datasets such as Caltech101 and Caltech256 (see http://www.cvpapers.com/datasets.html) have been developed. It should

however be noticed that ground truth for images is hard to establish since people often associate a given image with a wide range of high-level semantics.

On the video side, TRECVID (TREC Video Retrieval Evaluation) is widely used (see http://trecvid.nist.gov/trecvid.data.html). The main goal of TRECVID is to promote progress in content-based analysis and retrieval from digital video via open, metrics-based evaluation. TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations.

3 Existing CBIR systems

Over the years, many CBIR systems have been developed. Most of these are prototype systems and hence not intended for public usage. However, some are released under open source licenses so that they can be used and developed further by others. There are also a number of commercial systems on the market, into which CBIR capabilities have been incorporated. We have made a limited overview of various systems, research prototypes as well as commercial, to get some insights to how useful these systems might be for our purposes at the moment.

Before we made any attempt to identify (and in some cases also install) existing CBIR systems, we have first checked whether there are any systematic reviews of CBIR systems in the literature. The most thorough review that we have been able to find is the one of Venters and Cooper (2000). In their investigation, 74 systems are identified. A majority of those are however research prototypes and only five of the systems are more thoroughly reviewed. Three of them (ImageFinder, IMatch, and QBIC) are tested using a small-scale retrieval experiment. Additionally, the user interfaces of ImageFinder and IMatch are further evaluated. The results indicate that the performance in terms of precision and recall was very questionable at the time of the experiments. Moreover, the user interfaces of the tested products could be significantly improved according to Venters and Cooper. The only newer review of CBIR systems that we have identified is the one by Müller et al. (2004), only covering CBIR systems for medical applications.

Based on the lists of CBIR systems in Venters and Coopers (2000) and the Wikipedia list of CBIR systems[2], we have identified a number of CBIR systems (research projects as well as commercial systems) for a closer look at the potential of current systems. Note that many of the systems listed in Venters and Coopers (2000) have ceased to exist (probably since the research projects in which the prototype systems have been developed have ended).

IBM's QBIC (Query By Image Content) system (Flickner et al., 1995) is one of the most well-known CBIR systems and has a long history. The CBIR engine used in QBIC has e.g. been used for searching in archives of world-famous art at the Hermitage Web site[3]. It seems that QBIC-functionality today has been integrated into IBM software for Enterprise Content Management, but it has been hard to find any useful information about this. It is therefore unknown what the functionality of QBIC looks like today.

Another CBIR system that has been developed during many years is Virage[4]. Today, Virage is part of Autonomy (very recently bought by the company Hewlett-Packard). There are many different Autonomy Virage products for image processing such as number plate recognition, face recognition and intelligent scene analysis. The CBIR system has probably been integrated to one or many of these products, but it is not evident from their webpage. This issue could however be worth to look into since Autonomy products already are well-known to the Swedish Armed Forces.

A third commercial CBIR system of interest is IMatch[5]. Rather than being a standalone CBIR system, IMatch is a commercial image management system containing CBIR functionality. We have downloaded an evaluation version of the system in order to test it. IMatch has a nice and intuitive GUI. We have made some simple evaluations with an image database consisting of two datasets from the Caltech image database: 1074 images of airplanes and 126 images of cars. A randomly picked car from the dataset was used as query image, whereupon IMatch returned the images in the database that according to its implemented algorithm were most similar to the query image. The results are

[2] http://en.wikipedia.org/wiki/List_of_CBIR_engines
[3] http://www.hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English
[4] http://www.virage.com/
[5] http://wp.photools.com/imatch-3-overview/

disappointing since it returned images that in our opinion are very dissimilar to the query image (it also gave a lot of airplanes as results when using a car as query image and vice versa when trying with an airplane as query image). The individual weights of the used features (colour and shape) can be adjusted, but this did not give any better results. There is also functionality in IMatch for searching by drawing a sketch instead of providing a query image.

We have also tested a number of free CBIR systems. The first such systemsto be tested was Octagon[6], which is a quite simple and free Java software for CBIR. There is no open API available, so the functionality of the software can't be extended. The program allows for searching for images by their visual content (colour and image structure). Not much time has been spent on developing the Octagon GUI, but due to its limited functionality it is straightforward how to use it. Basically, what can be done is that an image database is created, after which one can select to import jpg-images into the database (where the import of 10 images takes approximately 1 second). When a database has been established, the user can choose an image as a query image, whereupon the system quickly returns a set of images most similar to the query image. There is a kind of simple relevance feedback functionality where the user can indicate if returned images are relevant, neutral, or not relevant to the query, but it is not clear if the selections actually improve the search or not. There is also a very basic beta-functionality for keyword annotation and search. To make a quick test of the system, we used the same test data as for IMatch. When a randomly selected airplane was used as a query image, only aircrafts were returned among the 30 most similar images and vice versa when using a car as a query image. This is of course good, but we were not very impressed with the similarity of the top ranked images to the query image (i.e., many images in the database were in our opinion more similar to the query image than the images returned by the system).

Another free CBIR system that has been tested is Emir, described in Lux (2008). Emir is part of the software package Caliph and Emir, downloadable from Sourceforge[7]. In Emir, it is possible to search for similar images based on colour layout, scalable colour, and edge histograms. Since we for some reason could not index the images used for Octagon and IMatch, test data included with Emir were used to test the system. The test data contained very few photos, but the system seems to work well, at least on this limited test set. Except for the CBIR functionality in Emir, there is also functionality for image annotation in the Caliph software. Caliph and Emir are released under a GPL license. There is also a library called LIRE that is part of the project, which aims to provide the CBIR features of Caliph and Emir to other Java projects in an easily accessible form.

We have also tested the so called MUFIN Project[8]. For some kinds of images, the functionality to retrieve visually similar images work very well with MUFIN (e.g., images of flowers and castles), but when it comes to more heterogeneous images, the results are not better than for the other tested CBIR systems.

Two open CBIR frameworks that have been identified but not tested are GIFT[9] and Windsurf[10]. GIFT stands for GNU Image-Finding Tool and allows for query by image example and has functionality for relevance feedback. The reason for not evaluating GIFT is that is has not been updated since 2005. Finally, Windsurf stands for Wavelet-based INDexing of ImageS Using Region Fragmentation. Relying on a region-based approach to CBIR, Windsurf is written in Java and has a brief Javadoc describing how to use the API. Because of time limitations we have not tested Windsurf, but it appears to have potential and should be considered for can future work.

[6] http://octagon.viitala.eu/
[7] http://sourceforge.net/projects/caliph-emir
[8] http://mufin.fi.muni.cz/tiki-index.php
[9] The source code is downloadable from http://www.gnu.org/software/gift/
[10] Source code available from http://www-db.deis.unibo.it/Windsurf/

A very recent addition to consumer products with CBIR capabilities is Photoshop Elements. Its "Auto-Analyzer", which first showed up in the Windows version of Elements 8 evaluates and identifies the most interesting and important aspects of an image. The analyzer automatically searches the photo collection and assigns a tag based on qualities such as lighting, focus or contrast. In Photoshop Elements 9 (as well as in Apple's iPhoto) one of the key technologies is "People Recognition", which tags images that contain faces. The program basically learns the identity of faces frequently captured in the images and makes suggestions about who is in each photo based on user notes made during import. Over time, the program becomes more proficient and automatically associates specific faces with names. Photoshop Elements 10 brings new features such as the ability to search for photos containing specific tangible objects. A catalogue of photos can be searched for duplicated photos, similar photos or objects in the photo, based on visual similarities. The search is done by colour or shape or a mix of these. We have made some simple tests with the same image databases for the testing of the other CBIR methods. The results are, however, not satisfying. When trying to find the registration plate in the set of car images we get many false detections. When searching for registration plates in a mixed image set with both cars and aeroplanes we get many answers indicating that an aeroplane is a registration plate. Some of the results obtained in the experiments are shown below.

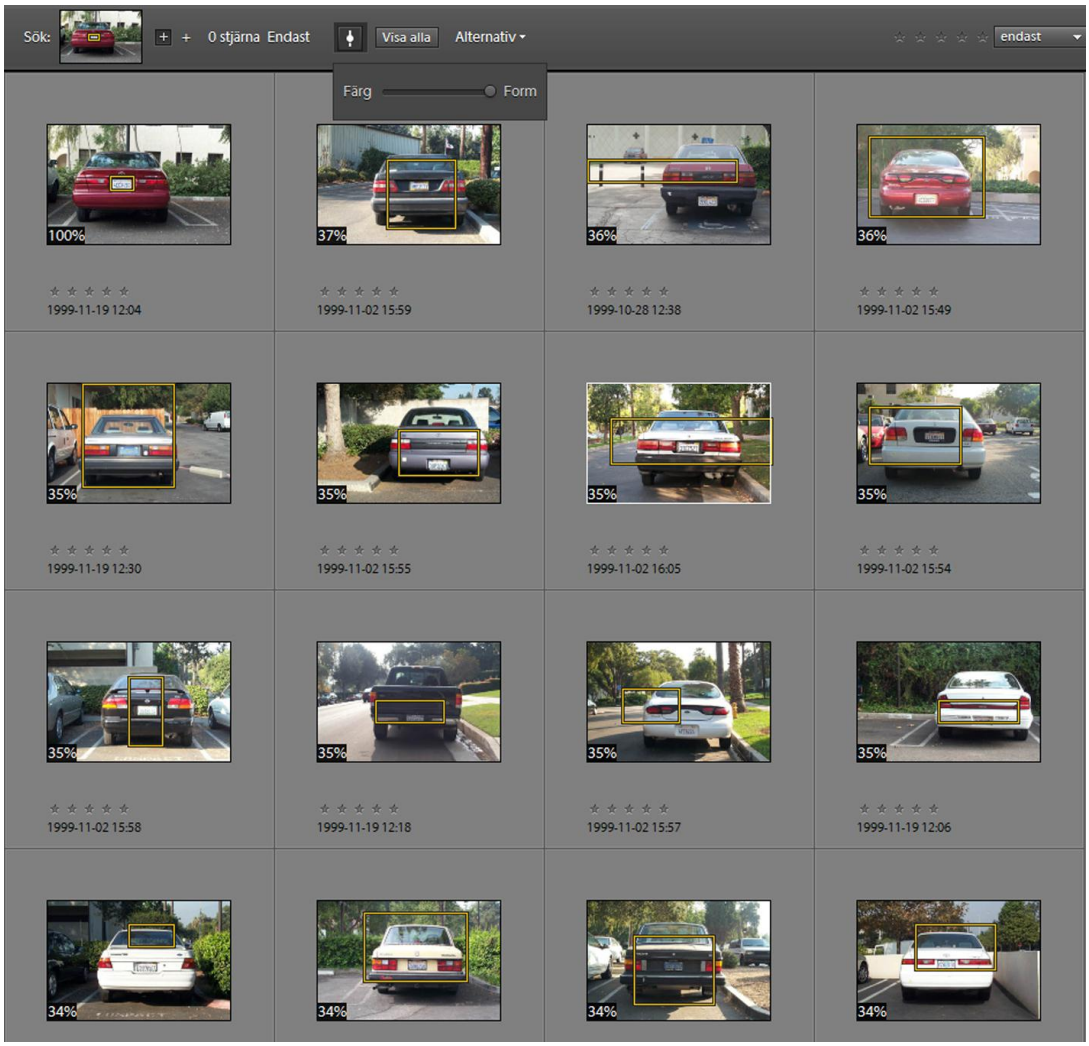

Figure 1: Search for registration plate in car database using shape.

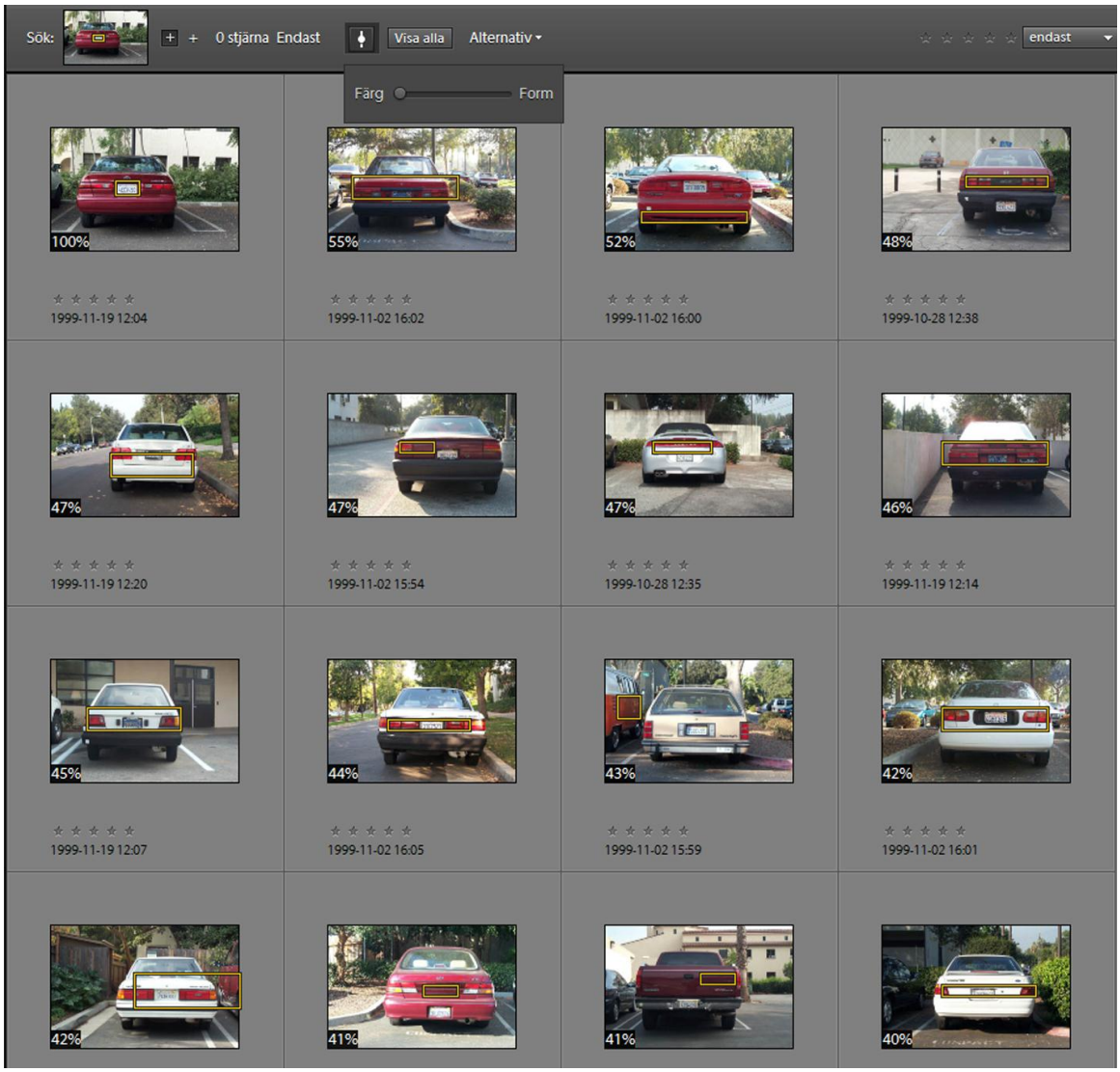

Figure 2: Search for registration plate in car data base using colour.

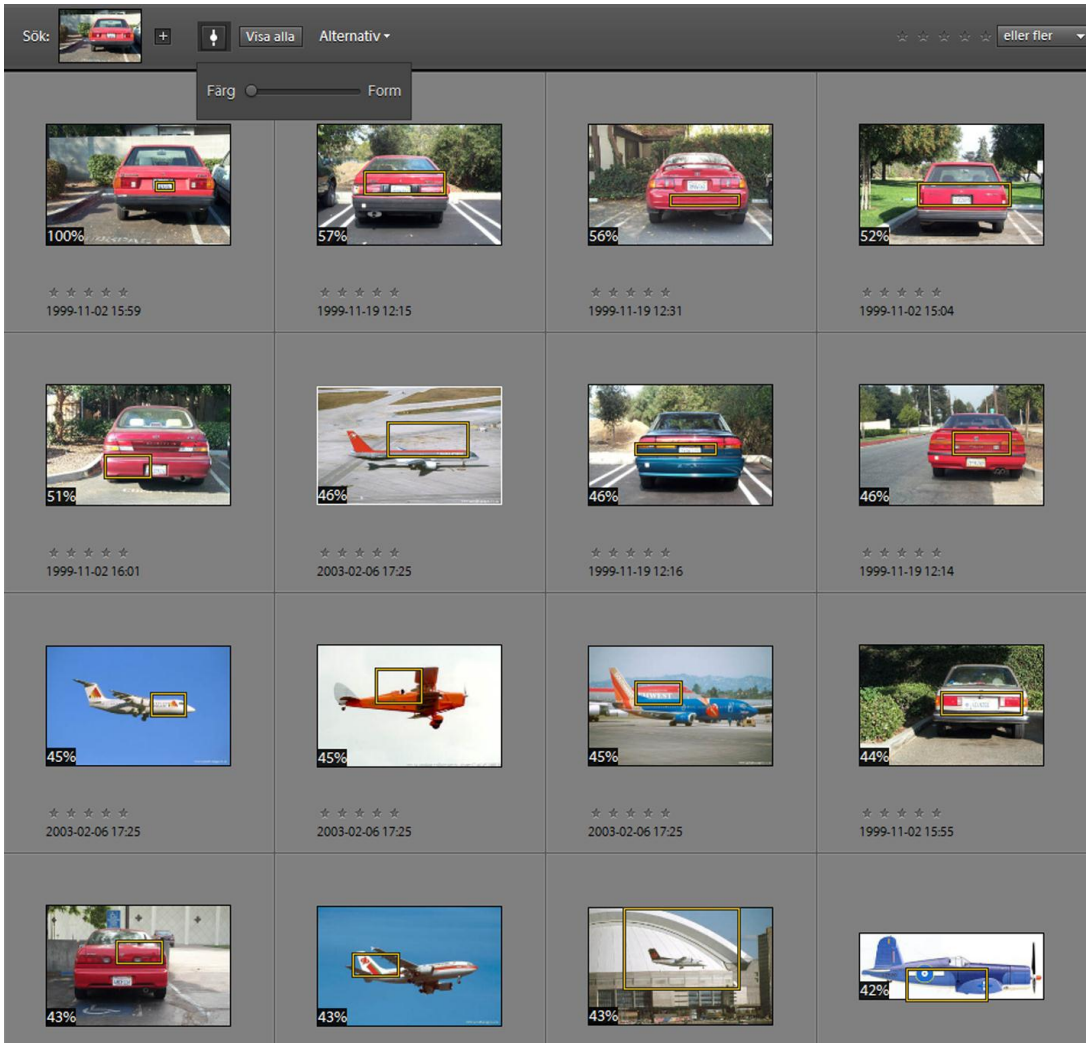

Figure 3: Search for registration plate in the mixed data base using colour.

Another quite recent product is Google Image search, which is a feature in the Google search engine “Search by Image”. It is optimized to work well for content that is reasonably well described on the web, and hence, will most likely give more relevant results for famous landmarks or paintings than for more personal images. Google uses computer vision techniques to match the query image to other images in the Google Images index and additional image collections. From those matches, the algorithm tries to generate an accurate "best guess" text description of the query image, as well as to find other images that have the same content as the query image. The search results page can show results for the generated text description as well as related images. When using an image of an airplane as input, the images shown in Figure 4 were returned as the result. After some trials the algorithm learns that the image is an airplane and only returns pictures of airplanes. However, the search is then no longer an image search but becomes a meta data search. After testing Google image search it is clear that colour information is vital for the matching, while the structure is not. Searching for a specific object in an image is not possible with Google image search.

Figure 4: Result from the first search for visually similar images with Google “search by image”.

4 Prospects of CBIR

In this report we have provided a quite pessimistic view of the current status of CBIR. Many published papers show promising results on controlled datasets, but when applying the algorithms or systems to real-world problems, the obtained precision and recall is not always as good as wished for. However, the recent development where CBIR-functionality has started to be included in mass market products (e.g., Photoshop Elements, iPhoto and Google Image Search) can be a sign of a more mature development phase where CBIR is becoming useful for practical purposes. Moreover, even though no general breakthrough has been achieved for large-scale heterogeneous image datasets, this does not mean that CBIR is not useful for more constrained domains.

Many failures of CBIR are caused by overreliance on the ability of readily computable statistical measures or generic pattern recognition algorithms to generate meaningful image features from a training set. The result is usually that essentially random features are found. After some tinkering with the algorithms, the training set is, however, classified with high accuracy. The reason for this is that any sufficiently large random set of features can be used for classifying a given limited training set. Performance drops, however, dramatically as soon as the method is tried on a larger virgin set of examples.

On the contrary, successful application of CBIR requires that the features are strongly related to the relevant semantics of the application. This will only happen as a result of purposeful engineering as demonstrated by the limited set of success stories that we have found.

Consider an archive of text documents where each document has been scanned and saved in digital format as an image. It is eminently possible to build a high-performance CBIR-application where users input an image containing scanned text and the application outputs all images in the database containing the same text sample. The key technology for this application would be Optical Character Recognition (OCR) which is based on a very mature set of algorithms that essentially identifies the same features that children are taught for the purpose of distinguishing the letters of the alphabet. A spell-checker is usually applied as a second layer of semantic filtering.

A second mature CBIR application is finger print identification where automatic tools identify the same set of features as human operators have used for several generations of forensic work. Automatic face recognition is the basis of third successful CBIR-technology as manifested e.g. in Photoshop Elements 9 “People recognition” which was discussed in Chapter 3 or Apple iPhoto. The enabling factor is again that algorithms can identify precisely the same features (eyes, nose, lips …) that humans find interesting.

All CBIR-successes that we have mentioned take advantage of the tacitly assumed context of the images. Text is printed on a uniform background and applies a reasonably limited set of fonts. Finger print identification software gets only finger prints on a standard format as input. Face recognition software assume that people are photographed in fairly standard settings.

In conclusion we believe that while truly generic CBIR requires breakthroughs in semantic processing an increasing range of feasible applications will open up in the near-term, particularly in application domains such as industrial inspection, biometry, medical technology and security. For each such application it will be important to take advantage of image features that are based on hard-won human knowledge about the domain.

5 CBIR for the Swedish Armed Forces

CBIR potentially allows for completely new kinds of searches in image and video material collected by the Swedish Armed Forces. Instead of search based on filenames or tags manually inserted in the data, it is for suitable application domains possible to search for characteristics automatically extracted from the data. As discussed in chapter 4, the present level of CBIR technology requires that the image or video material is constrained to a well-defined context and human expertise may be necessary for defining meaningful base-line features. For instance, it is in an appropriately constrained image set possible to search for all images showing a vehicle of a given type, or images containing special kinds of terrains. Video analysis could, with the same caveats as before, prove useful for detecting people pointing a gun, throwing something, lying down, etc. Applications concerning static surveillance (of camps, parked vehicles, etc.) are considered especially feasible in the near future, since problems due to sensor motion can be kept at a minimum.

We hence believe that there are several different application areas of interest for the Swedish Armed Forces. FOI is presently researching face recognition for a biometric application of interest to the Swedish Armed Forces. Some further application areas are suggested in the following three subsections.

5.1 C-IED

One application where we consider CBIR particularly useful for the Swedish Armed Forces is for counter-IED (C-IED). The IED threat constantly and rapidly changes its nature over time, as both technology and tactics for IED-related attacks develop. After having been successfully used at some site, some new tactics or technological modifications often spread across the area of conflict, often resulting in similar attacks at other parts of the country. In order to reduce the risk of own personnel being subject to attacks and to be able to track and defeat the insurgents behind the attacks, the military needs tools to analyze and understand the current situation and to make predictions concerning developments in the future.

Here CBIR has an important – and currently unexplored – role to play. Today, every time an IED is found, be it through detonation or not, the incident is documented in terms of several images taken at the site, as well as written reports. The images may show the IED itself or the effects of a detonation, the site, the surrounding area, some critical technical components, e.g., triggering mechanism, circuit board or capsule, and so on. This documentation work results in huge amounts of IED-related images being collected every year, and the data is to be analyzed by human operators. Here, CBIR could provide the operator with tools to facilitate the search for relevant information, including the following aspects:

- CBIR can be used to help the operator to find similar images in a database through “query-by-example”, based on image similarity measures. For example, to find images that contain IEDs similar to the query object, of the same size, with the same colour and shape, etc (note that photographs of objects laid out on a uniform background often are available). It could also be used to search for IEDs embedded in similar terrain (e.g., rocks, gravel or vegetation). Possible with some human assistance such as marking the area of lettering it could be possible to identify IEDs having the same letters, codes or symbols.
- CBIR may also be used without the intervention of an operator, for example through *image clustering*, i.e., automatic grouping of a set of images in a particular context with respect to a certain similarity measure, or *recognition* of a particular kind of structure/object/detail in the images. This would help the

operator to target his/her efforts to where it matters the most, to ensure that the images are searched efficiently.

- CBIR also enables the detection of *anomalies* in and *changes* between images. For example, the CBIR system may recognize two IED devices to be similar enough to point the operator to look at the images, but there may be also subtle differences that can be automatically identified as anomalies by the system.

What makes the IED application potentially very fruitful in the near future is that much of the analysis can be based on low-level image features such as colour and texture or high-level features that have been defined by humans (e.g. related to shape and size) rather than automatically identified semantic features. A great deal of CBIR work has been on the low-level aspects, and we believe it is mature enough to really make a difference in the counter-IED work.

It should be pointed out that relevant information could also be added manually to the images as metadata (tags, keywords), which could be used alongside with CBIR functionality in the quest for relevant information. In fact, adding information manually is necessary to couple the images to attributes or circumstances that are difficult or impossible to extract automatically from the image content. However, it would be virtually impossible to anticipate all the metadata that will be useful in the future. Therefore, we believe that CBIR tools for finding relations between images through analysis of the image content would be of great relevance to C-IED work.

5.2 Video surveillance

Another example of an area of application where we believe that CBIR can be of great use for the Swedish Armed Forces is for video analysis. A starting point could be video acquired in a surveillance context, e.g., with one or several video cameras mounted at a camp for recording activity in the proximity of the base. Among other things, collected video material could be processed automatically in order to:

- track, characterize and cluster the movements of different objects (persons, vehicles) in the scene. Such trajectory data can be queried to find all sequences that an object has moved in a particular fashion, stood still a certain period of time or behaved anomalously (in a manner not previously seen by the system)
- retrieve sequences containing a particular type of object, e.g., a person or a vehicle. The query could be by *example* (“retrieve all objects that look like this one”) or through semantic search (“retrieve all blue cars”)

Similarly, video data collected with cameras mounted on vehicles can be used to recognize objects (persons, vehicles, traffic signs, etc.) and to detect patterns/clusters in the data (”are we regularly followed by a particular car?”). Data from several different types of sensors including radar can be combined with the video data to improve performance.

5.3 Image analysis for air reconnaissance

One currently challenging domain where there is a large amount of image data available is IMINT (imagery intelligence) collected from aerial reconnaissance missions. Today’s modern reconnaissance aircraft are capable of collecting huge amounts of imagery during a flight, which today must be analysed manually. This manual processing and analysis is a bottleneck in the current process, as evidenced for instance in the missions in Libya recently.

In this domain CBIR could help in several ways. The most advanced solution would be to automatically tag collected imagery data and automatically provide intelligence analysts and mission planners with relevant image data based on search queries that they have entered. This requires advances in both low-level image processing to extract suitable

features and convert them into semantic tags and in high-level query processing to automatically match the information needs of a user to the tags.

Another use, which is perhaps more realistic in the near future, is to use “query by example” on suitably constrained subsets of the IMINT data, in a similar way as described for the IED application above.

6 Discussion and further work

In this report, we have given a short summary of the research area of CBIR, and some of the possibilities and challenges related to applying this research to military problems.

A key conclusion from this study is that while there are indeed large potential benefits of using CBIR, the realization of CBIR systems for military applications requires that potential end-users work closely together with CBIR experts in order to work out where CBIR functionality both is technically feasible and has great impact. Conclusions on technical feasibility are found in Chapter 4 while concepts for military applications are discussed in Chapter 5. In the follow-on project during 2012, we will work more closely with representatives of the Swedish Armed Forces in some selected problem domains and produce a plan for how CBIR could enable improvements.

Based on the outcome of the workshops, and the survey presented here, one of the anticipated activities in such a follow-on project is to investigate in more detail the capabilities of the publically available CBIR systems presented in Chapter 3, and if possible, to obtain evaluation licenses for commercial tools.

A central requirement for development and performance assessment of CBIR systems is to have access to relevant data. While there are several relevant applications (some outlined in Chapter 5), practical data availability issues are likely to influence the choice of case study.

During a visit to SWEDEC, the project discussed opportunities for enhancing information search capabilities in the context of Explosive Ordnance Disposal (EOD). SWEDEC distributes a software application (EOD IS) to EOD professionals world-wide. Presently, it is possible to identify objects by entering text-based queries and then selecting matching objects from a list of candidates. Deminers can then use retrieved information for safely clearing mines and IEDs. The project will investigate if it is possible to apply CBIR query-by-example to EOD IS. Demining staff should ideally be able to take a snap-shot of an unknown object, send the picture to an EOD IS terminal and immediately retrieve one or a few database entries describing the object.

The follow-on project also needs to study what ontology is needed for implementing CBIR in the proposed application domain(s). In addition to specifying an initial ontology, a process for how it should be updated and changed as new information is made available needs to be described. Changing the ontology also requires changing already used metadata (tags), and a process for this needs to be developed.

There is also a need to determine what the limits of current signal processing are when it comes to automatically computing semantic tags that the image data can be marked with.

FOI has included CBIR related research in two applications to the 2011 call of the security research programme of the EU. One of the applications deals with decision support systems for operators of x-ray scanning equipment for air cargo. In this application, FOI's focus is on the high-level aspects of CBIR, i.e., ontologies and query construction as well as anomaly detection based on the extracted features. The other application also deals with air cargo security, but here the focus is on information systems for handling air cargo data and making threat analyses.

A more detailed plan of the project during 2012 will be completed in January of 2012, where choices between the options outlined above will be made based on available resources.

References and literature list

The following is the result of our extensive search for CBIR literature. Many of the most important articles and papers are reviewed in chapter 7 of this report and the list also includes the references of chapters 1-6. However, not all of the papers in the list have been read and summarized but references are provided for further study.

Adistambhaa, K., Davisa, S., Ritz, C., and Burnett, I., "Efficient multimedia query-by-content from mobile devices", Computers and Electrical Engineering, Vol. 36, Issue 4, 2010.

Agarwal, M., Maheshwari, R., "HOG feature and vocabulary tree for Content-based Image Retrieval", International Journal of Signal and Imaging Systems Engineering. Vol. 3, no. 4, pp. 246-254. 2011.

Ahmed, H.A., El Gayar, N., and Onsi, H., "A New Approach in Content-Based Image Retrieval Using Fuzzy Logic" Proceedings of INFOS'2008, 2008.

Akguel, C., Rubin, D., Napel, S., Beaulieu, C., Greenspan, H. and Acar, B., "Content-Based Image Retrieval in Radiology: Current Status and Future Directions", Journal of Digital Imaging Vol. 24, no. 2, pp. 208-222. Apr 2011.

Ali, S. and Shah, M., "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 32, Issue: 2, pp. 288–303, 2010

Arpah, A., Alfred, S., Lim, L.H.S., and Sarinder, K.K.S., "Monogenean image data mining using Taxonomy ontology", International Conference on Networking and Information Technology (ICNIT2010), June 2010.

Balasubramani, R. and Kannan, V., "Efficient use of MPEG-7 Color Layout and Edge Histogram Descriptors in CBIR Systems", Global Journal of Computer Science and Technology; Vol 9, No 4 (2009).

Banerjee, M., Kundu, M.K., and Das, P.K., "Image retrieval with visually prominent features using fuzzy set theoretic evaluation", Indian Conference on Computer Vision, Graphics and Image Processing (VGIP2004), Dec 2004.

Bartolini, I., Ciaccia, P., Patella, M., "Query processing issues in region-based image databases", Knowl. Inf. Syst., Vol. 25, No. 2, 2010.

Batko M, Falchi F, Lucchese C, et al., "Building a web-scale image similarity search system", Multimedia tools and applications, Vol. 47, Issue 3, pp. 599-629, 2010.

Beecks, C.; Uysal, M.S.; Seidl, T., "A comparative study of similarity measures for content-based multimedia retrieval", IEEE International Conference on Multimedia and Expo (ICME) 2010, pp.1552 – 1557, 2010.

Belongie, S. and Malik, J. "Matching with Shape Contexts". IEEE Workshop on Content-based Access of Image and Video Libraries, 2000.

Bursuc, A.; Zaharia, T.; Prêteux, F.,"Online interactive video content retrieval", 2011 IEEE International Conference on Consumer Electronics (ICCE), 2011.

Carson, C., Belongie, S., Greenspan, H., and Malik, J., “Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying”, IEEE Trans. on PAMI, vol. 24, No.8, pp. 1026-1038, 2002.

Chatterjee, K.; Shu-Ching C. "GeM-Tree: Towards a Generalized Multidimensional Index Structure Supporting Image and Video Retrieval", Tenth IEEE International Symposium on Multimedia, pp. 631 – 636, 2008.

Chen, Y. and Wang J. Z., “A Region-Based Fuzzy Feature Matching Approach to Content-Based Image Retrieval,” IEEE Trans. on PAMI, vol. 24, No.9, pp. 1252-1267, 2002.

Chuohao, Y; Ahammad, P.; Ramchandran, K.; Sastry, S.S, "High-Speed Action Recognition and Localization in Compressed Domain Videos", IEEE Transactions on Circuits and Systems for Video Technology, Volume: 18 , Issue: 8, pp. 1006–1015, 2008.

Colombino T, Martin D, Grasso A, et al, "A Reformulation of the Semantic Gap Problem in Content-Based Image Retrieval Scenarios", 9th International Conference on Designing Cooperative Systems, MAY 18-21,2010

Datta, R., Joshi, D., Li, J. and Wang, J., “Image Retrieval: Ideas, Influences, and Trends of the New Age”, ACM Computing Surveys, vol. 40, No. 2, 2008.

Döller, M., Mayer, T., Fong, K-L., Beck, S., Kosch, H., and Coquil, D. ”Standardized Mobile Multimedia Query Composer", 2nd International Symposium on Intelligent Interactive Multimedia Systems and Services, 2009

El Sayad, I., Martinet, J., Urruty, T., Amir, S., and Djeraba, C., "Effective object-based image retrieval using higher-level visual representation”, 2010 International Conference on Machine and Web Intelligence, ICMWI 2010.

Fan, Z-G., Li, J., Wu, B., and Wu, Y., “Local Patterns Constrained Image Histograms for Image Retrieval”, Proceedings of the Fifteenth International Conference on Image Processing, 2008.

Flickner, M., et al., “Query by Image and Video Content: The QBIC System”, IEEE Computer, Vol. 28, No. 9, 1995.

Flores-Pulido, L.; Rodríguez-Gómez, G.; Starostenko, O.; Santacruz-Olmos, C., "Radial Basis Function for Visual Image Retrieval", Electronics, Robotics and Automotive Mechanics Conference (CERMA), 2010.

Garcia-Arteaga JD, Kybic J, "Regional Image Similarity Criteria Based on the Kozachenko-Leonenko Entropy", IEEE Conference on Computer Vision and Pattern Recognition, 2008

Gevers, T. and Smeuiders. A.W.M., “Combining colour and shape invariant features for image retrieval”, Image and Vision computing, vol.17(7), pp. 475-488, 1999.

Ghoshal, A.; Khudanpur, S.; Klakow, D.,"Impact of novel sources on content-based image and video retrieval" , IEEE International Conference on Acoustics, Speech and Signal Processing, 2009

Gonde, A.B.; Maheshwari, R.P.; Balasubramanian, R., "Texton co-occurrence matrix: A new feature for image retrieval", India Conference (INDICON), 2010.

Gorisse, D.; Cord, M.; Precioso, F.,"Scalable active learning strategy for object category retrieval", 7th IEEE International Conference on Image Processing (ICIP), 2010.

Granlund, G., “Fourier Preprocessing for Hand Print Character Recognition”, IEEE Transactions on Computers, pp. 195-201, 1972.

Guldogan, E.; Lagerstam, E.; Olsson, T.; Gabbouj, M., "Multi-form hierarchical representation of image categories for browsing and retrieval", 5th International Workshop on Semantic Media Adaptation and Personalization (SMAP), pp. 64 – 69, 2010

Harris, C. and Stephens, M., “A combined corner and edge detectors”, 4th Alvey Vision Conference, pp. 147-151, 1988.

Haiying, G., Antani, S., Long, LR, and Thoma, G.R. ," Bridging the semantic gap using Ranking SVM for image retrieval, IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009.

Hazra, D. “Texture Recognition with combined GLCM, Wavelet and Rotated Wavelet Features”, International Journal of Computer and Electrical Engineering, Vol.3, No.1, February, 2011.

Heczko, Martin; Hinneburg, Alexander; Keim, Daniel; Wawryniuk, Markus,"Multiresolution similarity search in image databases", Multimedia Systems, Volume 10, Number 1, June 2004 , pp. 28-40.

Hoiem, D. Sukhtankar, R., Schneiderman, H. and Huston, L., “Object-Based Image retrieval Using Statistical structure of images”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR04), 2004.

Hollink, L., Schreiber, G., Wielinga, B., and Worring, M. “Classification of user image descriptions”, International Journal of Human Computer Studies, Vol. 61, pp. 601-626, 2004.

Hu W., et al., “Semantic-based surveillance video retrieval”, IEEE Trans. Image Processing, vol 16, no 4, pp. 1168-1181, 2007.

Hu, W. ; Xie, N. ; Li, L. ; Zeng, X. ; Maybank, S.,"A Survey on Visual Content-Based Video Indexing and Retrieval.", Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2011.

Huang W, Gao Y, Chan KL,"A Review of Region-Based Image Retrieval", Journal of Signal Processing Systems for Signal Image and Video Technology, Volume: 59, pp. 143-161, 2010.

Israel, M.; van der Schaar, J.; van den Broek, E.L.; den Uyl, M.; van der Putten, P., "Multi-level visual alphabets", 2nd International Conference on Image Processing Theory Tools and Applications (IPTA2010), pp. 349 – 354, 2010.

Jain A. and Vailalya, A., “Image retrieval using colour and shape”, Pattern recognition, vol. 29, pp. 1233-1244, 1996.

Jiang, X.H., Sun, T.F. and Wang S.L., "An automatic video content classification scheme based on combined visual features model with modified DAGSVM", 2nd International Congress on Image and Signal Processing, pp.17-19, 2009.

Kachouri, R.; Djemal, K.; Maaref, H.;, "Adaptive feature selection for heterogeneous image databases", 2nd International Conference on Image Processing Theory Tools and Applications (IPTA), 2010.

Kläser, A., "Learning human actions in video", PhD thesis, Université de Grenoble, 2010.

Koyuncu, M.; Yilmaz, T.; Yildirim, Y.; Yazici, A., "A framework for fuzzy video content extraction, storage and retrieval", Fuzzy Systems (FUZZ), 2010 IEEE International Conference on Digital Object Identifier: 2010.

Kwitt, R.; Meerwald, P.; Uhl, A.,"Efficient Texture Image Retrieval Using Copulas in a Bayesian Framework", IEEE Transactions on Image Processing, 2011

Kwitt R, Uhl A, "Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retreival", 15th IEEE International Conference on Image Processing (ICIP 2008), 2008.

Lawlor, M., "Seeing eye systems learn to discern", Signal, pp. 27-30, May 2011.

Li, Z., Shi, Z., Liu, X., and Shi, Z., "Modeling continuous visual features for semantic image annotation and retrieval", Pattern Recognition Letters, Volume 32, Issue 3, pp. 516-523, 2011

Li, J., Wang, J.Z. and Wiederhold, G., "IRM: Integrated Region Matching for Image Retrieval," in Proceedings of the 8th ACM Int. Conf. on Multimedia, pp. 147-156, Oct. 2000.

Li J., Yu H.N., Tian Y.H., "Salient object extraction for user-targeted video content association", Journal of Zhejiang University-Science, Computers & Electronics, Vol. 11, Issue: 11, pp.: 850-859, 2010.

Lin CH, Chen RT, Chan YK, "A smart content-based image retrieval system based on color and texture feature", Image and Vision Computing, Vol. 27, Issue: 6, pp. 658-665, 2009

Liu, D. and Chena, T., "Video retrieval based on object discovery", Computer Vision and Image Understanding, Vol. 113, Issue 3, 2009.

Liu, Y., Zhang, D. and Lu, G., "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, Vol. 40, pp. 262-282, 2007.

Liu, Y., Xin, C., Chengcui, Z. Sprague, A., , "Semantic clustering for region-based image retrieval", Journal of Visual Communication and Image Representation, Volume 20,Issue 2, pp. 157-166, February 2009

Loupias, E. and Sebe, N., "Wavelet-based salient points: Applications to image retrieval using colour and texture features", Proceedings of the 4th International Conference on Advances in visual Information systems, pp. 223-232, 2000.

Lowe, D. “Distinctive image features from scale invariant keypoints”, International Journal of Computer vision, vol. 2(6), pp.91-110, 2004.

Lux, M., and Savvas, C. “Lire: Lucene Image Retrieval – An Extensible Java CBIR Library”. Proceedings of the 16th ACM International Conference on Multimedia, pp. 1085-1088, 2008

Ma, W.Y. and Manjunath, B.S., “NETRA: A Toolbox for Navigating Large Image Databases,” in Proc. IEEE Int. Conf. on Image Processing, vol. I, Santa Barbara, CA, pp. 568--571, 1997.

Mejdoub, M., Fonteles, L., BenAmar, C., and Antonini, M., "Embedded lattices tree: An efficient indexing scheme for content based retrieval on image databases", Journal of Visual Communication and Image Representation, Volume 20, Issue 2, pp. 145-156, 2009

Malcolm, J., Rathi, Y., Tannenbaum, A. ”A graph cut approach to image segmentation in tensor space”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR07), 2007.

Mallik, J., Samal A., and Gardner, S., "A content based image retrieval system for a biological specimen collection", Computer Vision and Image Understanding, Vol. 114, no. 7, pp. 745-757. Jul 2010.

Mang T., Gong S.G. ,"Activity based surveillance video content modelling", Pattern Recognition, Volume: 41 Issue: 7, pp. 2309-2326, 2008.

Manjunath, B.S. and Ma, W.Y., “Texture features for browsing and retrieval of image data”, *IEEE Trans. PAMI*, vol.18, no 8, Aug 1996.

Meixner, A., and Uhl, A., "Robustness and security of a wavelet-based CBIR hashing algorithm", Proceeding of the 8th workshop on Multimedia and security; 26-27 Sept. 2006.

Melbourne, A; Ridgway, G; Hawkes, DJ," Image similarity metrics in image registration", Proceedings of SPIE - The International Society for Optical Engineering [Proc. SPIE Int. Soc Opt. Eng.]. Vol. 7623, Mar 2010.

Memar, S.; Ektefa, M.; Affendey, L.S., "Developing context model supporting spatial relations for semantic video "2010 International Conference on Digital retrieval, Information Retrieval & Knowledge Management, (CAMP), pp. 40 – 43, 2010.

Mezaris, V., Kompatsiaris, I., and Strintzis M. G., “Region-based Image Retrieval Using an Object Ontology and Relevance Feedback,” Eurasip Journal on Applied Signal Processing, vol. 2004, No. 6, pp. 886-901, 2004.

Mikolajezyk, K. and Schmid, C., “Scale and affine invariant interest point detectors”, International Journal of Computer Vision, vol. 1(60), pp. 63-86, 2004.

Min HS, Choi JY, and De Neve W., "Semantic annotation of personal video content using an image folksonomy”, 16th IEEE International Conference on Image Processing, pp.257-260, 2009.

Min HS, Lee S, De Neve W, et al., "Semantic Concept Detection for User-Generated Video Content Using a Refined Image Folksonomy", 16th International Conference Multimedia Modeling (MMM2010), 2010.

Mustaffa MR, Ahmad F, Rahmat RWOK, et al., "Content-based image retrieval based on color-spatial features", Malaysian Journal of Computer Science, Vol. 21, pp. 1-12, 2008

Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A., "A review of content-based image retrieval systems in medical applications - clinical benefits and future directions", International Journal of Medical Informatics Vol 73, pp. 1-23, 2004.

Natsev, A., Rastogi, R., and Shim, K., "WALRUS: A Similarity Retrieval Algorithm for Image Databases," in Proc. ACM SIGMOD Int. Conf. Management of Data, pp. 395--406, 1999.

Niblack, W., et al., "The QBIC Project: Querying Images by Content Using Colour, Texture, and Shape," in Proc. SPIE, vol. 1908, San Jose, CA, pp. 173--187, Feb. 1993.

Paradowski M, Sluzek A, "Keypoint-Based Detection of Near-Duplicate Image Fragments Using Image Geometry and Topology", International Conference on Computer Vision and Graphics, Lecture Notes in Computer Science, Volume: 6375, 2010.

Pentland, A., Picard, R., and Sclaroff, S., "Photobook: Content-based Manipulation of Image Databases," in Proc. SPIE Storage and Retrieval for Image and Video Databases II, San Jose, CA, pp. 34--47, Feb. 1994.

Rubner, Y., Guibas, L.J., and Tomasi, C., "The earth mover's distance, multi-dimensional scaling, and colour-based image retrieval", Proceedings of DARPA Image understanding Workshop, pp. 661-668, 1997.

Rena, W., Singhb, S., Singhb, M., and Zhua Y.S., "State-of-the-art on spatio-temporal information-based video retrieval", Pattern Recognition, Volume 42, Issue 2, pp. 267-282, 2009

Rui, Y., Huang, T., and Chang, Y.S., "Image retrieval: Current techniques, promising directions and open issues", *J. Vis. Comm. and Im. Repr.*, vol 10, pp. 39-62, 1999.

Schreer, O, "Multimedia Indexing and Retrieval of Unedited Audio-Visual Footage", 9th International Workshop on Image Analysis for Multimedia Interactive Services, 2008.

Sebastine SC, Thuraisingham B, Prabhakaran B, "Semantic Web for Content Based Video Retrieval", 3rd International Conference on Semantic Computing (ICSC 2009), SEP 14-16, 2009

Sinjur S, Zazula D,"Image Similarity Search in Large Databases Using a Fast Machine Learning Approach", Conference Information: 1st International Symposium on Intelligent Interactive Multimedia Systems and Services, JUL 09-11, 2008

Smeulders A.W.M. et al., "Content-Based Image Retrieval at the End of the Early Years" *IEEE Trans. PAMI*, vol 22, no 12, pp. 1349-1380, Dec. 2000.

Snoek, C. et al., "The Mediamill TRECVID 2009 semantic video search engine," *Proceeding of the 7th TRECVID Workshop*, 2009.

Stassinopoulos G.I., Papastefanos S.S.,"Efficient selection of candidates in video content search", International Journal of Electronics and Communications, Volume 64, Issue 7, pp. 650-662, 2010.

Stejić, Zoran; Takama, Yasufumi; Hirota, Kaoru, "Variants of evolutionary learning for interactive image retrieval", Soft Computing, Volume 11, Number 7, pp. 669-678, 2007.

Stricker, M. and Orengo, M., "Similarity of Colour Images," in Proc. SPIE Storage and Retrieval for Image and Video Databases, pp. 381-392, 1995.

Su, JH., Huanga, YT, Yeha, HH and Tseng, V., "Effective content-based video retrieval using pattern-indexing and matching techniques", Expert Systems with Applications, Vol. 37, Issue 7, pp. 5068-5085, 2010.

Sumana IJ, Islam M, Zhang DS, et al., "Content Based Image Retrieval Using Curvelet Transform", 10th IEEE Workshop on Multimedia Signal Processing, 2008

Talbar, S. and Varma, S. "iMATCH: Image Matching and Retrieval for Digital Image Libraries", Second International Conference on Emerging Trends In Engineering and Technology, 2009.

Thanh-Toan Do, Ewa Kijak, Teddy Furon, Laurent Amsaleg, "Challenging the Security of Content-Based Image Retrieval Systems", 2010 IEEE International Workshop on Multimedia Signal Processing, MMSP2010 - December 01, 2010

Thurau, C.; Hlavac, V., "Pose primitive based human action recognition in videos or still images",Computer Vision and Pattern Recognition, 2008.

Tsai, C-F and Hung, C, "Automatically Annotating Images with Keywords: A Review of Image Annotation Systems", Recent Patents on Computer Science, vol 1, pp. 55-68, 2008.

Wang, B., Zhang, X., Wang, M., and Zhao, P., "Saliency distinguishing and applications to semantics extraction and retrieval of natural image", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 2, pp. 802 – 807, 2010.

Wang WY, Zhang DM, Zhang YD, et al.,"Fast and robust spatial matching for object retrieval, IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, 2010.

Wang, C., Zhao, J., He, X., Chen, C., and Bu, J., "Image retrieval using nonlinear manifold embedding", Neurocomputing, Vol. 72, Issues 16-18, pp. 3922-3929, 2009.

Venters, CC, and Cooper M., "Content-based image retrieval", Tech. Rep. JTAP-054, JISC Technology Application Program, 2000

Wichert, A., "Content-based image retrieval by hierarchical linear subspace method", Journal of Intelligent Information Systems, Vol. 31 Issue: 1 pp. 85-107, 2008

Wong, YM, Hoi, S., and Lyu, M., “An Empirical Study on Large-Scale Content-Based Image Retrieval”, IEEE International Conference on Multimedia & Expo (ICME2007), 2007.

Wu, K. Yap, KH., "Content-based image retrieval using fuzzy perceptual feedback", Multimedia Tools and Applications, Volume 32, Number 3, pp. 235-251(17), 2007.

Wu, P., Hoi, S., Nguyen, D., and He, Y., “Randomly projected KD-trees with distance metric learning for image retrieval”, Proceedings of the 17th international conference on Advances in multimedia modelling”, 2011.

Yang CL, Ga YL, Fan JP, "Benchmark of Multiple Approaches for Feature Extraction and Image Similarity Characterization", Conference on Imaging and Printing in a Web 2.0 World; and Multimedia content Access - Algorithms and Systems IV, 2010.

Yu, M., Lizeng, M, Li, L., "Shot Boundary Detection Using Directional Empirical Mode Decomposition", Second International Conference on Multimedia and Information Technology (MMIT), Vol. 1, pp. 107 – 110, 2010

Yuan, J., Wu, Y., and Yang, M., "From Frequent Itemsets to Semantically Meaningful Visual Patterns", Proceedings of the 13th International ACM Conference on Knowledge Discovery and Data Mining, pp. 864-873, 2007

Zargari, F., Mehrabi, M. ; Ghanbari, M.,"Compressed domain texture based visual information retrieval method for I-frame coded pictures.", IEEE Transactions on Consumer Electronics, 2010

Zhang, J and Ye, L, "Series feature aggregation for content-based image retrieval", Computers & Electrical Engineering Vol.36, no. 4, pp. 691-701. Jul 2010.

Zhaozhong, W.; Min, L., "Locally Affine Invariant Descriptors for Shape Matching and Retrieval", IEEE Signal Processing Letters, Vol. 17 , Issue: 9, 2010.

Zheng YT, Neo SY, Chua TS, et al.,"Object-based image retrieval beyond visual appearances", 14th International Multimedia Modeling Conference, 2007.

Zheng, QF., and Gao, W.,"Constructing visual phrases for effective and efficient object-based image retrieval", ACM Transactions on Multimedia Computing, Communications and Applications. Vol. 5, no. 1. Oct. 2008

Zhou, Z; Liu, JY; Ma, LH; Chen, D., "Video content analysis technique", Jisuanji Gongcheng yu Sheji (Computer Engineering and Design). Vol. 29, no. 7, pp.1766-1769. Apr. 2008.

Zhu, X. et al., "Video Data Mining: Semantic Indexing and Event Detection from the Association Perspective," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no 5, May 2005.

Zhou, X.S. and Huang T.S., "Relevance Feedback in Image Retrieval: A Comprehensive Review," *Multimedia Systems*, vol. 8, no. 6, pp. 536-544, 2003.

Zadeh, L.; Joshi, A.; Ramamoorthy, C.; Yu, H.; Sheu, P., "Visual Ontology Construction and Concept Detection for Multimedia Indexing and Retrieval", Semantic Computing, 2010

Appendix A: Short summaries of selected articles and papers

In this appendix, we list the papers we have reviewed in this study, together with short summaries of the content, as well as our initial comments given when reviewing the publications.

iMATCH: Image Matching and Retrieval for Digital Image Libraries

Authors: Sanjay N. Talbar, Satishkumar L. Varma

Source: Second International Conference on Emerging Trends in Engineering and Technology, ICETET, 2009

Summary: The problem addressed in the paper is to retrieve images in a database consisting of different classes of images. The features used for image description are derived from a DCT (Discrete Cosine Transform) applied on 8x8 pixel regions of the image after it has been transformed from RGB to YCbCr colour space. In all, twelve features are extracted from each such region. Then, the feature vectors are clustered using the k-means algorithm in order to find the five largest clusters in the image, the idea being that those five clusters carry most of the significant information about the image. The technique is evaluated with a test set of 360 images.

Comments: The choice of the number of clusters (k) is not discussed. It should, because that is always an issue with clustering.

The evaluation data set is very small (360 images) and still the results are less than impressive (precision and recall rates of around 20-40 %).

Image Similarity Search in Large Databases Using a Fast Machine Learning Approach

Authors: Smiljan Sinjur, Damjan Zazula

Source: 1st International Symposium on Intelligent Interactive Multimedia Systems and Services, JUL 09-11, 2008 Piraeus, GREECE, NEW DIRECTIONS IN INTELLIGENT INTERACTIVE MULTIMEDIA Book Series: STUDIES IN COMPUTATIONAL INTELLIGENCE Volume: 142 Pages: 85-93.

Summary: This paper presents a novel image similarity search method based on the extraction of multiple convex hulls from an image. First, the image is subject to a RGB-> HSV colour space transformation. The H (Hue) image is then thresholded, based on the rationale that foreground and background objects have different Hue values. Then, the convex hull of the foreground pixels is determined. All pixels belonging to this convex hull (the vertices) are removed, after which the convex hull is constructed anew with the remaining pixels. The corresponding vertices are removed, a new convex hull is constructed, and so on until no more pixels remain. From the resulting set of convex hulls, a set of features is extracted, namely the number of vertices in each convex hull. After some de-trending of the data, a standard correlation between feature vectors is used to find similar images. Then a second round of analysis is performed using a SVM. As training data, the feature vectors of similar and dissimilar images (as found through correlation) are used. The method was tested on data from a movie (154892 frames). The best result reported was a sensitivity of 83%.

Comments: The authors do not motivate or discuss whether foreground pixels are below or above the threshold. This detail, either overlooked or avoided, is important as the two pixel sets imply two entirely different convex hull sets and consequently completely different feature vectors. The convex hull is also very sensitive to outliers.

The method was not compared to other techniques so it is impossible to say anything about its qualities. The paper gives the impression of being an interesting idea worth trying

out because no one else has done it, rather than a robust method that is very useful in practice.

Multiresolution Similarity Search in Image Databases

Authors: Martin Heczko, Alexander Hinneburg, Daniel Klein

Source: Multimedia Systems, Volume 10, Number 1, June 2004 , pp. 28-40(13), Springer

Summary: In this paper, an approach for multi-scale matching of colour histograms is presented. Rather than computing similarities between the entire histograms of two images, they propose a generalization based on first subdividing the histograms, computing features for each subdivision and then summing the similarities. For (dis)similarity measure, the Euclidean distance between feature vectors was used. They consider colour images, but treat the colours independently by using the three 1D histograms. Based on tests involving a database of 10,000 images, the proposed approach yielded better result than some other techniques (that did not involve subdivision). By choosing different coefficients, the matching can be done on coarse or fine level, the former giving matches between images of similar overall colour and the latter resulting in finding images with similar fine details (e.g. texture).

Comments: The starting point for the method is to make three separate histograms for the entire image, one for each colour. By doing so, all information about the spatial distribution of the colour values across the image is lost and so is information about the cooccurence of colours. Due on such inherent shortcomings, this work is unlikely to be a milestone in CBIR research, although it might be worth a look if the task is to match 1D histograms.

Building a web-scale image similarity system

Authors: Michal Batko, Fabrizio Falchi, Claudio Lucchese, David Novak, Raffaele Perego, Fausto Rabitti, Jan Sedmidubsky, Pavel Zezula

Source: *Multimedia Tools and Applications*, Springer Netherlands, USA. ISSN 1380-7501, vol. 47, no. 3, pp. 599-629, 2010

Summary: A system for finding similar images in web-scale (i.e. enormous) databases is presented. For image features, five MPEG-7 features are used: Scalable colour, colour structure, colour layout, edge histogram and homogeneous texture. Those features are calculated for each image and stored as an XML structure along with the URL of the original image, and an image thumbnail. The search task is distributed among several nodes in a network, and the space that each node searches through is partitioned using the M-Chord data mapping. At each node, pivoting M-Trees are used for performing nearest neighbour queries. The system has been used with about 50 Million images.

Comments: The paper addresses the problem of how to handle all those enormous, ever-increasing amounts of image data on the web.

The particular choice of using those five MPEG-7 features is not motivated. However, in another paper (Bolettieri et al., 2009, "CoPhIR: a Test Collection for Content-Based Image Retrieval") it is claimed that "Many experiences suggest that retrieval based on these five MPEG-7 standard descriptors can be acceptable on non-specialized images, such as the ones in our collection." So if any other features would prove better for a particular military application, those could/should be used instead.

The main contribution of the paper is efficient and distributed search for similar images in a huge search space across several computer nodes, something that comes in a bit later in the process of designing a CBR system: first, one must solve the problem of actually finding image features and similarity functions that do the job.

Content-based image retrieval by hierarchical linear subspace method

Authors: Andeas Wichert
Source: Journal of Intelligent Information Systems 31:95-107, 2008

Summary: This paper presents a hierarchical linear subspace to query large on-line image databases. The search starts in the subspace corresponding to the lowest image resolution. All images considered similar to the query object in this subspace are brought to the nearest higher subspace in a sequence (e.g. higher resolution), where the set is reduced using additional metric information (more features).The approach is evaluated on three databases with 1000 to 10000 images.

Comments: The paper claims to address query of large databases, but 10000 images are not really that many... Still, for that amount of images, the search time is the order of 10-30 s per query.

The image descriptors are very simple, if not "naïve", as claimed by the author himself: the pixel colour values themselves. Probably too simple to be of much use in practice.

The main (only?) contribution of the paper is the idea of defining certain subspaces and to traverse those in sequence to narrow down the initial search results and to save time compared to searching the highest subspace directly, but the results are not really that impressive.

Semantic-Based Surveillance Video Retrieval

Authors: W. Hu, D. Xie, Z. Fu, W. Zeng, S. Maybank

Source: IEEE Trans. Proc., Vol. 16, No. 4, April 2007

Summary: A clustering-based tracking algorithm for obtaining trajectories is proposed. Moving objects are detected in traffic surveillance video data, and their paths are clustered in the spatio-temporal domain. The result is a set of activity patterns to which semantic meanings are given. The database of trajectories can be queried with a trajectory sketch ("retrieve all trajectories that looked like this") or via keywords ("retrieve all instances where an object went fast and turned left in this crossing"

Comments: The approach does not involve recognition of particular object types (e.g. cars, bicycles, pedestrians), but all movements are detected and tracked. In fact, it is simple from an image processing viewpoint, as the camera is fixed. Moving the sensor would create additional, significant image processing problems.

Detecting that/how something moves can probably be quite robust and useful in practice.

Automatically Annotating Images with Keywords: A Review of Image Annotation Systems

Authors: Tsai, Chih-Fong and Hung, Chihli,

Source: Recent Patents on Computer Science, vol 1, pp. 55-68, 2008.

Summary: This article provides a review of various machine learning approaches that have been suggested for automatic image annotation. The main reason for automatic image annotation is to overcome the semantic gap between low-level image feature content and the high-level concepts/semantics of the image. Automatic image annotation can be thought of as a pattern recognition/classification problem, in which supervised learning is used to teach a classifier to identify the correct class of an image based on its low-level features (attributes in the machine learning-terminology). Many references are given to literature in which various supervised classifiers have been used for automatic image annotation, e.g. naive Bayes, ANNs, SVMs, and decision trees. References are also given to work where ensembles of classifiers have been used, using techniques such as majority voting, bagging, and boosting. In most of the cited works, only one keyword is

assigned per image, but there are also some references to work in which several keywords are generated per image.

When evaluating (and training) a supervised image annotation classifier, Corel and similar datasets are most often used (where the category of the image has been decided by a human professional indexer). A problem with this is that only one class is considered to be the correct one, although many different classes can match a certain image.

Comments: Automatic image annotation can probably be used to come a little bit closer to high-level semantics, but this is not the final solution. Rather than to map a class to a whole image, it is much more useful if we can map (at least) one class to each object (i.e. each segment if the segmentation is made properly) in the image. However, even though we succeed in accomplishing this, this is only a description of the object contents of the image. This would of course be very useful, but in the long run we also would like to be able to label images with e.g. events that are present in the image (a man riding a horse, etc.). Hence, the current state-of-the-art of automatic image annotation seems to be far away from the manual annotation performed by humans.

A review of content-based image retrieval systems in medical applications

Authors: Müller H, Michoux N, Bandon D, Geissbuhler A

Source: Int J Med Inform, vol. 73, pp. 1-23, 2004.

Summary: The article presents an overview of CBIR in the context of medical applications. Although quite different from the military domain, the article is interesting in that it gives an overview also of more general CBIR systems (the commercial systems QBIC and Virage and a number of systems from academia such as Candid, Photobook, Netra, GIFT, and Viper). It is also mentioned that the principal components of all CBIR systems are the retrieval engine, visual feature extraction, distance measures and similarity calculations, storage and access methods, and the GUI.

Comments: The perhaps most interesting part of the article is the idea of creating a database with images representing normal (non-pathologic) cases, to which images from new cases are compared using dissimilarity retrieval (rather than similarity retrieval). Combining this with techniques that highlight regions with strongest dissimilarity, it is argued that e.g. tumours and fractures can be discovered. In the same manner this type of technique perhaps also can be used to discover changes (anomaly detection) in areas monitored by e.g. satellite photos or SAR images. Most likely those kinds of techniques are already used, but if not, this could be a useful tool for e.g. intelligence analysts.

Region-based image retrieval using an object ontology and relevance feedback

Authors: Mezaris, Vasileios and Kompatsiaris, Ioannis and Strintzis, Michael G.

Source: EURASIP J. Appl. Signal Process, pp. 886-901, 2004.

Summary: Various CBIR approaches have been suggested, and this is yet another. The authors propose a methodology that extracts low-level features from regions that are identified by segmentation, where the segmentation algorithm used is based on a variant of KMCC (K-Means with connectivity constraints). The produced segments are compared to those generated by the Blobworld segmentation algorithm (where it is shown that the suggested algorithm produces better results in most cases). A potential problem is that the segmentation takes ~27 seconds on average, which sounds quite long if such an approach is to be used in a system containing a large number of images.

The segmentation is however perhaps not the most interesting part of the article. Most CBIR algorithms/systems require query by example (image or sketch), but Mezaris et al. propose the use of a simple object ontology, which allows for more high-level searches. In their ontology, intermediate-level descriptors such as intensity, position, and size are used. These can be used directly for search, but can also be used to define high-level concepts such as "tiger" or "rose" which then can be used for search (i.e. "give me all images

containing a tiger and grass"). The mapping of semantic keyword definitions to intermediate-level descriptors is very coarse (a tiger is defined as a concept that has luminance=[medium, high], green-red=[red low, red medium], blue-yellow=[yellow medium, yellow high] and size = [small, medium]). Hence, the search for a tiger will certainly result in many images without any tigers, but this should rather be seen as a first preprocessing filter in which the majority of regions that clearly are irrelevant to the query are excluded. The remaining images (or at least a subset of these) are presented to the user, who then is assumed to refine the search using relevance feedback (checking whether images are relevant to the query or not). This relevance feedback is used as training samples for a SVM that learns the boundaries between positive and negative samples, so that better results can be presented to the user.

Comments: This is interesting from the point of view that it does not require query by example. The approach will most likely demand a lot of relevance feedback which can be frustrating for the user, but it is still nice to see an approach where the aim is higher than just comparing low-level features directly. This kind of functionality could be useful when searching for high-level concepts such as IED, weapon, etc., but since the object ontology is so coarse, the false alarm rate will be very high. There is however also an advantage with this coarseness, since it allows for applying it to generic image collections.

Person-based video retrieval with automatic query face generation

Authors: X. Sun, K. Liu, T. Zhang

Source: Review paper for IEEE Int. Conf on Image Processing (ICIP), 2011

Summary: The paper presents a quite extensive system for retrieving videos containing specific persons. The contributions are a) a method for generating query faces, and b) a method for re-ranking search results using the query faces. The process is as follows:

1. A textual search (based on tags/annotations) return a large number of videos (eg, search on "Obama" on YouTube).
2. In the returned videos, faces are detected (unknown how), features extracted and clustered (across all videos). The dominant cluster(s) will likely represent the sought-for person, and representative faces are shown to the user. Query faces for each cluster are selected automatically.
3. Each video sequence is matched to the query faces and ranked in terms of similarity to the query, total time in the video in which the target person appears, and popularity of the video. These three factors are weighted and combined and a final ranking is created.

The results are of course dependent on the face detection and face matching methods used. The face detection is not mentioned at all. Face feature extraction is done using Gabor filters and LBP (probably on a dense grid), then dimensionality-reduced using PCA and LDA, and then clustered (ie, the BoG/BoW approach). Clustering is done by Hierarchical Agglomerative Clustering (HAC) and the CANNOT-link constraint.

The third step contains an intra-video clustering of all detected faces, representative faces are selected, and the query faces are matched with these faces.

The results are compared to a text-based search & ranking to create query faces (using the Lemur toolkit, www.lemurproject.org). It is unclear in the review paper if the Lemur approach is compared to the entire proposed solution or to a combination.

Experiments are performed on 3416 videos (with 103 celebrities) from YouTube.

Comment: The main interest of this paper is that it gives a good example of methodologies to search for persons in video archives.

Query processing issues in region-based image databases

Authors: Bartolini, Ilaria and Ciaccia, Paolo and Patella, Marco

Source: Knowledge and Information Systems, vol. 25, pp. 389-420, 2010.

Summary: In many CBIR systems, a region-based paradigm is used, i.e. images are segmented into homogeneous regions for which features are locally extracted. This is referred to as region-based image retrieval (RBIR) and can be contrasted to approaches where features are extracted globally for the whole image. Comparison of RBIR approaches to global paradigms has shown that RBIR often allows for better performance in terms of precision and recall. However, a drawback with RBIR is that the issue of efficient query processing becomes even more important than usual. According to Bartolini et al., most current RBIR systems are using a sequential evaluation strategy when searching for images (or regions of images) that are similar to a query, despite that this is an approach that does not scale to large collections of images. A number of more efficient index-based strategies are suggested and experimentally tested on top of the WINDSURF systems.

Comments: The article is quite interesting from a database perspective, since it highlights scalability issues that will arise when dealing with very large image databases. Similar problems are likely to occur for video data, since this (ignoring the cost of decoding) can be thought of as very large collections of still images.

Bridging the Semantic Gap in Image Retrieval

Authors: Rong Zhao and William I. Grosky

Source: Chapter in the book Distributed Multimedia Databases: Techniques and Applications, Idea Group Publishing, pp. 14-36, 2002.

Summary: A fundamental problem in CBIR is the gap between low-level features of visual data and more high-level semantic (conceptual) features, where the latter often are more useful for users of CBIR systems. Zhao and Grosky are trying to, at least partially, bridge this gap by finding latent correlation between visual features and high-level semantics. To accomplish this, they make use of latent semantic indexing (LSI), which earlier has been used for improving text retrieval performance when matching words of queries on a conceptual level with words of documents. Moreover, they propose a framework in which a feature vector space model which otherwise represents visual elements only is extended to also incorporate textual elements resulting from automatic image annotation.

Comments: This is a well-written article that highlights some of the challenges with CBIR. It also explains LSI on a quite detailed level.

Automatic Annotation of Images from the Practitioner Perspective

Authors: Enser, Peter and Sandom, Christine and Lewis, Paul

Source: Image and Video Retrieval, Springer, vol. 3568, 2005

Summary: As the title suggests, this paper deals with the topic of automatic image annotation. According to the authors, the recurrent mentioning of the semantic gap within literature on visual image retrieval is a reflection of that researchers are becoming more and more aware of the limited functionality of CBIR in realistic commercial systems. One way to try to overcome the semantic gap is to integrate CBIR techniques with traditional textual metadata. By having a training set of pre-annotated images, there are various supervised learning approaches that can be used to connect visual features in images with textual descriptions. Two such approaches are latent semantic analysis (LSA) and probabilistic latent semantic analysis (PLSA). However, according to the authors, such approaches to bridge the semantic gap suffer from what they refer to as the visibility limitation and the generic object limitation. The visibility limitation has to do with the significance of the image, which is something that cannot be derived from low-level features of an image, since it demands a high level of context knowledge. In the same way, the generic object limitation also calls for more information that can be derived from the visual content of the image alone, since it has to do with a common desire from users to recover images of features uniquely identified by name (e.g. Abraham Lincoln). For this

reason, the authors argue for the need of enhancing the functionality of current automatic annotation techniques with ontology-supported content annotation. They also present a simple taxonomy for still images.

Comments: This is a very shallow paper that does not add much on its own. There are however references to work on other work on automatic image annotation that may be well worth to look into. Moreover, the book "Image retrieval: theory and research" by Jörgensen seems to be a relevant reference in which potential problems with CBIR are identified ("the emphasis in the computer science literature has been largely on what is computationally possible, and not on discovering whether essential generic visual primitives can in fact facilitate image retrieval in 'real-world' applications.").

Learning To Count Objects in Images

Authors: Victor Lempitsky, Andrew Zisserman

Source: Advances in Neural Information Processing Systems (NIPS), 2010

Summary: A supervised learning framework for visual object counting tasks is presented. Based on training images annotated with dots (one dot per object) and some features, an optimal set of weights is computed by using the MESA distance as cost function. The scalar product between the feature vector and the weight vector gives a feature density in each pixel. The density field is then integrated across the area of interest to yield an estimate of the object count.

Comment: Interesting approach. The idea is to avoid segmentation of the images, something that could be very tricky to do reliably when objects overlap, etc. Instead they estimate a density and integrate that over an area. The people counting algorithm they present requires that the camera is fixed (because frames are compared on a pixel-by-pixel basis). Plus you have to come up with useful features for the task at hand. For the pedestrian detection several features are used: the image itself, frame-to-frame difference, frame-to-background difference. The weights used for the linear combination of features are found through training.

The system needs to be trained on labelled data (hence it's a supervised method), but the data needs only be labelled with dots. That is, an operator could just click on the positives without the need to carefully segment the images into objects.

Demo is available from: http://www.robots.ox.ac.uk/~vgg/research/counting/

Human Focused Action Localization in Video

Authors: Alexander Kläser, Marcin Marszalek, Cordelia Schmid, Andrew Zisserman

Source: International Workshop on Sign, Gesture, Activity, September 2010

Summary: Abstract. We propose a novel human-centric approach to detect and localize human actions in challenging video data, such as Hollywood movies. Our goal is to localize actions in time through the video and spatially in each frame. We achieve this by first obtaining generic spatio-temporal human tracks and then detecting specific actions within these using a sliding window classifier.

A human-centric approach to detect and localize human actions in challenging video data is proposed. The goal is to localize actions in time through the video and spatially in each frame. This is achieved through obtaining generic spatio-temporal human tracks and then detecting specific actions within these. A three-dimensional extension of Histograms of Oriented Gradients (3D-HOG) are used as features (the temporal dimension as no. 3). The spatio-temporal domain is divided into several "cuboids", in which 3D-HOGs are created (quantized to ten directions).

Only two actions were evaluated - drinking and smoking. AP (average precision?) was 54.1% and 24.5%, respectively.

Comment: This paper is about detecting human movement patterns in video, something that could be interesting for the SwAF. People standing still, raising their hands, pointing,

aiming with a gun, cheering, jumping, etc could (in principle) be detected this way. Addresses the semantic gap in that it couples low-level image features to high-level concepts. Even though quite a bit of work remains to be done before the system can discriminate between similar movements (like drinking vs. smoking), maybe distinctively different movement patterns are manageable.

The approach requires detection and tracking of people (upper bodies) between a number of frames in order to succeed. The authors claim that it may be difficult to detect movement of people in a crowd, for example.

Demo is at: http://www.robots.ox.ac.uk/~vgg/research/actions_interactions/

Efficient multimedia query-by-content from mobile devices

Authors: Kevin Adistambha , Stephen J. Davis, Christian H. Ritz, Ian S. Burnett

Source: Computers & Electrical Engineering, Volume 36, Issue 4, July 2010, Pages 626-642, Signal Processing and Communication Systems

Summary: The article deals with multimedia query formats that need to be applicable to mobile devices, which, compared to desktop PCs, have specific limitations such as small screen size, limited memory and processing power and high bandwidth cost. It presents ideas that can be a potential solution to multimedia querying in mobile environments, such as query streaming and its application as targeted browsing. Targeted browsing is a technique for multimedia query-by-content designed especially for mobile devices while query streaming is a method for continually updating a query by sending additional terms to an existing query. The paper also describes an implementation of query streaming that combines the Multimedia Query Format (MQF), a standard communication language for querying multimedia databases, with Fragment Request Units (FRU) and Fragment Update Units (FUU) which provide a standard way of randomly accessing fragments of XML documents.
Comments: Searching multimedia items is a non-trivial task that requires metadata to be attached to the multimedia items in question. Although the focus of this paper is the more specific problem of developing a standard communication language between clients and database solutions, the paper includes some knowledge about query construction**.**

Standardized Mobile Multimedia Query Composer

Authors: Doller M, Mayer T, Fong KL, Beck S, Kosch H, Coquil D

Source: 2nd International Symposium on Intelligent Interactive Multimedia Systems and Services, Mogliano Veneto, ITALY, JUL 16-17, 2009

Summary: The article presents an extension to a video search framework, introducing a new query engine which generates multimedia using the MPEG Query Format
Comments: This paper gives us references to the new MPEG Query Format (part 12 of MPEG-7) now established as the MPEG Query Format (MPQF) standard.