Dipl.-Ing. Marian Kogler, Bakk.techn.

VISUAL INFORMATION RETRIEVAL AND ITS APPLICATION TO USER INTENTIONS

DISSERTATION

zur Erlangung des akademischen Grades
Doktor
der Technischen Wissenschaften

Alpen-Adria Universität Klagenfurt
Fakultät für Technische Wissenschaften

1. Begutachter:	O.Univ.-Prof. Dr. Laszlo Böszörmenyi
Institut:	Institut für Informationstechnologie
2. Begutachter:	Prof. Dr. Michael Granitzer
Institut:	Universität Passau
Vorbegutachter:	Ass.-Prof. Dr. Mathias Lux
Institut:	Institut für Informationstechnologie

Juli 2012

Ehrenwörtliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende wissenschaftliche Arbeit selbständig angefertigt und die mit ihr unmittelbar verbundenen Tätigkeiten selbst erbracht habe. Ich erkläre weiters, dass ich keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle aus gedruckten, ungedruckten oder dem Internet im Wortlaut oder im wesentlichen Inhalt übernommenen Formulierungen und Konzepte sind gemäß den Regeln für wissenschaftliche Arbeiten zitiert und durch Fußnoten bzw. durch andere genaue Quellenangaben gekennzeichnet.

Die während des Arbeitsvorganges gewährte Unterstützung einschließlich signifikanter Betreuungshinweise ist vollständig angegeben.

Die wissenschaftliche Arbeit ist noch keiner anderen Prüfungsbehörde vorgelegt worden. Diese Arbeit wurde in gedruckter und elektronischer Form abgegeben. Ich bestätige, dass der Inhalt der digitalen Version vollständig mit dem der gedruckten Version übereinstimmt.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Unterschrift: _____

Klagenfurt, 31. Juli 2012

Contents

List of Tables

List of Figures

Abstract

The increasing number of videos and images in various multimedia repositories demands for clever content description and indexing techniques to retrieve the visual data in subsequent searches. The main search paradigm in large multimedia databases, whether they are distributed or local, is text-based and relies on additional meta data. Unfortunately meta data is often sparsely distributed over the image and video collection, which entails the need for additional and complementary search paradigms. Visual information retrieval leans onto content-based low level information in terms of color, texture and other features.

This work deals with various content-based techniques, applied for video and image retrieval, aiming at new methods to improve ongoing research in the respective fields. Low level approaches for video summarization and retrieval in combination with text as a complementary information source are presented. Well known visual descriptors comprising color and texture information are used in combination with a new motion based approach. Furthermore the Bag of Visual Words (BoVW) approach for visual information retrieval, which is a technique to quantize visual information in the feature space, is extended. The applied fuzzy rules for BoVW are thoroughly discussed in this thesis.

Beyond that visual information retrieval is regarded in context of user intentions. User intentions or the goal a user tries to achieve rely on user needs, which are expressed by the users' search behavior. A new approach for retrieval adaptation during content-based image search is presented, which tries to respond appropriately to different goals. Chapter 4 describes this idea, leveraging again BoVW.

Chapter 1

Introduction

The increasing amount of digital sensors like cameras and video recording tools has paved the way for an incredible quantity of new multimedia content, which needs to be properly administered and organized. The organization of multimedia content is not that easy and demands clever tools to extract valuable information and put this data into well defined structures for efficient search. Multimedia data such as tags or visual features like color are used during later retrieval and similarity searches. In the retrieval stage additional hurdles such as user interaction with the retrieval system come into play.

Typically users try to retrieve digital media by means of browsing, text-based and content-based retrieval. The prevalent techniques nowadays, to acquire images and videos, mostly rely on text information. Whereas popular search engines such as Google and Yahoo incorporate visual information, Flickr and YouTube leverage text, filenames of images and captions, in order to search large image archives for relevant digital content. Unfortunately text is sometimes misleading, because it doesn't express the visual content of an image appropriately. Filenames are sometimes less representative, what begs for additional information.

Tags, manually added when storing and indexing digital content, are used in subsequent search sessions as an additional source of information, in order to provide the user with the desired videos and images. The problems with manual tagging

are twofold. First and foremost the manual annotation of an exploding amount of digital data is a tedious process. Users are not willing to assign meta data to their multimedia content in order to retrieve it later on. Although automatic or semi-automatic tag recommendation systems try to attenuate this problem, they often do not recommend tags properly. The other issue of assigning meta data is that people use subjective tags. An image, tagged by k users, can cover k different tags (see [37]). This divergence affects the retrieval of images and videos negatively, hence claiming for additional retrieval paradigms, which do not solely rely on text.

Visual information retrieval[1] should be incorporated and should complement text based retrieval. Images and videos are organized and indexed based on their visual content. Features such as color, texture, shape or local information around interesting points within images and motion between subsequent frames are used to describe the visual content and are processed for later retrieval. A visual information retrieval (VIR) system extracts visual information of an image (or frame in case of a still image taken from a video stream), stores the information in a histogram interpreted as a vector and index the computed feature vector. During the retrieval stage a Query by Example (QBE) is submitted to the system and similar images or video frames according to a similarity function are retrieved by comparing the computed feature vector of the query image with the feature vectors in the database index.

Visual information retrieval systems can be usually divided into content-based video retrieval and content-based image retrieval as depicted Fig. 1.1.

[1] "Its purpose is to retrieve, from a database, images or image sequences that are relevant to a query. It is an extension of traditional information retrieval designed to include visual media." (cf. [7])

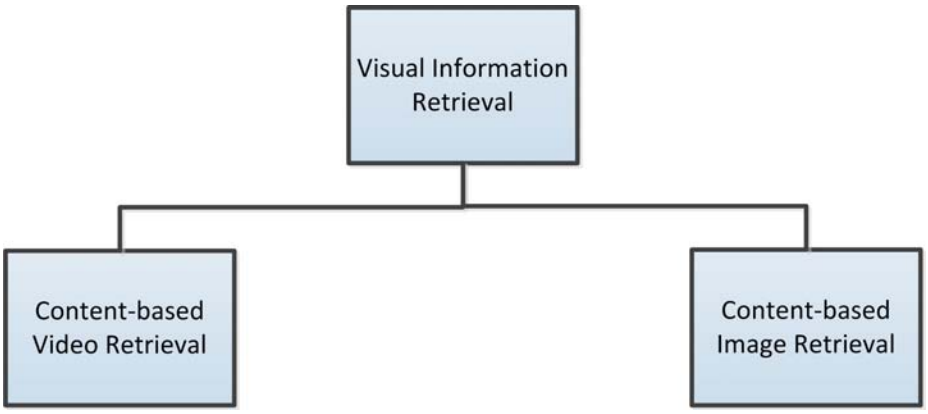

Figure 1.1: Visual information retrieval.

The following two subsections discuss the generally declared research questions and give an overview of new approaches, which were developed while writing this thesis. More detailed research questions are specified in the respective subsections.

General Research Question 1 *How can we describe and organize visual content effectively by means of visual information, in order to retrieve relevant images and videos from a plethora of multimedia data stored in repositories? (see section 1.1)*

General Research Question 2 *How can context information be used meaningfully while searching images in multimedia repositories? (see section 1.2)*

1.1 Content-based techniques for VIR

Both video and content-based image retrieval systems rely on low level features, whereas content-based video retrieval systems can also leverage temporal information. In VIR a video is understood as a sequence of frames and can be seen as a structured document, making the analysis and indexing of videos a little bit trickier than for images only. For analysis purposes still image features such as color and texture can be applied in combination with motion features, which represent the camera motion or the motion of an object.

In order to extract visual features, relying on still images, the video has to be parsed and pre-processed before. A video is decomposed into shots, representing subsequences of frames in it. They are determined by finding the transitions and boundaries between them. A shot is composed of a group of frames and single frames are analyzed to gather information about color, motion and so forth. Information about extracted features are represented by n-dimensional feature vectors. This information is typically indexed and used for later retrieval.

Additional techniques such as video summarization can be leveraged to build a subset of videos, which can be used during retrieval and browsing. Therefore not only frames or keyframes, representing a shot, are extracted, but also video summaries are created. Such summaries give a compressed representation of a video, allowing for fast assessment of its relevance during a search session (see Fig.1.2).

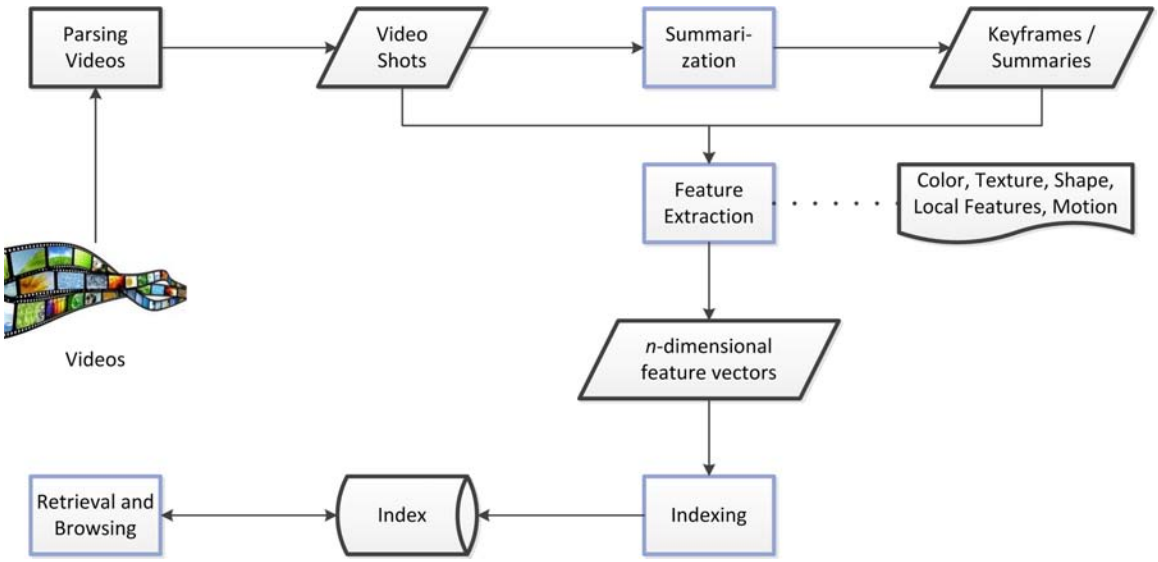

Figure 1.2: Video Retrieval.

Video summarization and retrieval are two topics in this thesis. Chapter 3 examines various techniques to process and analyze videos, in order to use the extracted information for later retrieval. A new approach for video retrieval, by combining

motion related information with Bag of Visual Words for visual information retrieval (see [91]), is presented (cf. Lux et al. [64]). The following research question 1 is elaborated in chapter 3.1.

Specific Research Question 1 *Can additional visual search methods increase the recall of badly annotated video clips?*

Moreover clustering methods together with the Bag of Visual Words approach are applied for video summarization. The new video summarization approach was published in Kogler et al. [47]. The arising research question 2 can be formulated as follows and is elaborate in chapter 3.2.

Specific Research Question 2 *Does the Bag of Visual Words approach yield more representative summaries of video clips than approaches with color and texture features?*

The corresponding process stages, which are addressed in this thesis, are highlighted in blue in Fig. 1.2.

Furthermore content-based image retrieval (CBIR) techniques, which also make use of low level features, are investigated. Content-based image retrieval has gained a lot of attention in recent years. It is comprised by a variety of research fields such as computer vision, information retrieval, human computer interaction and artificial intelligence. It can be applied to various use cases:

- In the Web on top of text retrieval systems, where text based search method acts as an entry point for a content-based search.
- In narrow domains such as medical image retrieval, where pictures are more closely related concerning their visual appearance.
- Searches conducted within categories such as cars or shoes, what is analogous to a search in a narrow domain.
- Image databases with sparse meta data, which need to be searched or browsed.

Usually a content-based image retrieval system expects an image or sketch (QBE) as input, from which visual information such as color, texture, etc. is extracted. Feature vectors are created, representing the visual information and similarity searches are conducted to retrieve similar images from an image database. Additionally relevance feedback can be employed by denoting relevant images during search, which are used to refine the query as depicted in Fig. 1.3. Chapter 3.3 deals with CBIR techniques, introducing a new visual content description approach. For this purpose the Bag of Visual Words approach is extended by applying fuzzy techniques (cf. Kogler and Lux [48] and [50]). The arising research question 3 is further investigated in chapter 3.3.

Specific Research Question 3 *Does fuzziness in visual vocabulary generation and visual word assignment lead to accurate and more robust results, when the number of visual words is varied? Does this allow for smaller and more efficient visual vocabularies?*

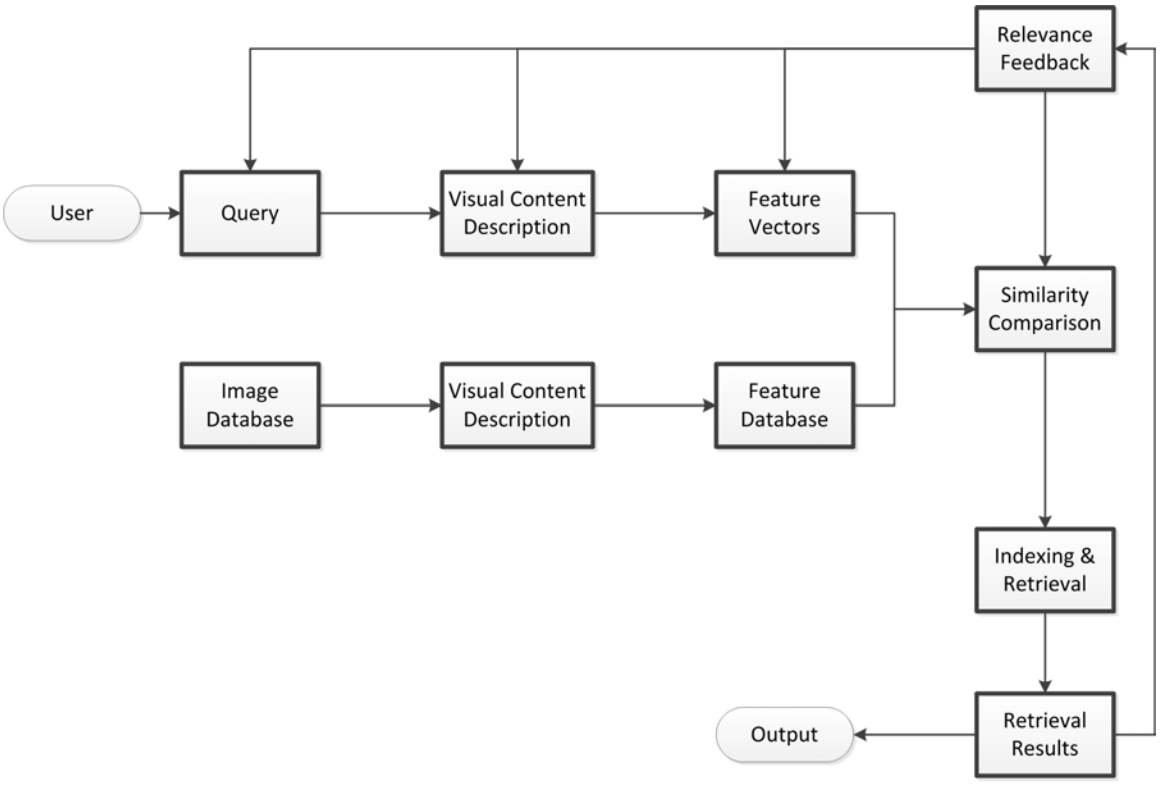

Figure 1.3: Content-based image retrieval (adopted from [26]).

1.2 VIR in Context of User Intentions

Retrieval systems like QBIC [77], PicToSeek [28], VisualSEEk [93] and many more have entered the multimedia community. Since they rely on visual information only, the semantic interpretation of the images is hardly covered. The semantic gap problem is a prevailing issue in many areas of computer science. Researchers try to bridge the gap by means of techniques such as:

- Ontologies, which assume already acquired domain knowledge.
- Machine learning, which can be either supervised by using support vector machines (SVM) and other learning models, or unsupervised by applying clustering

techniques.

- Relevance feedback, a technique inherited from information retrieval and used to incorporate user feedback into the retrieval process.

All mentioned approaches have their advantages and disadvantages, when returning result lists of images to search for, but all of them can benefit from context information. By using information concerning the context such as location, time, users search history and so forth the original search space is narrowed. Thus, visual content not fitting to the particular context is neglected, while preserving those images and videos, which are relevant to the users.
By considering visual information retrieval in context of user intentions result lists and relevance functions can be tailored to the users' needs.

An intention is a plan someone has in mind, in order to achieve a goal. It is difficult to measure, because its fuzzy and vague. On the other hand a goal is a state of affairs someone tries to achieve. Its (partial) success can be measured by observing a human's interaction with a computer. An example should illustrate the idea.

Example 1 *Imagine a student having the intention to add an appropriate image to her presentation slides. As soon as she expresses the objective to fetch a proper image (right color and quality for the presentation) a clear goal can be defined. If the goal is known, adequate constraints can be applied to guide the user's actions during her search. By restricting the color schemes for instance the user's needs can be properly served. On the other hand such constraints are not necessarily needed if a user wants an image just for entertainment purpose.*

User intentions during search or search goals, relying on an information need, are typically expressed by the users' search behavior. The search behavior can be tracked by observing user clicks, typed queries or in general the user's interaction with a visual information retrieval system. Industry pursues a strong interest in inferencing user intentions and search goals, in order to:

- adapt search result lists, which are adequate for a particular information need,
- adapt the visualization of the retrieved results,
- provide users with appropriate advertisement
- and/or increase the click-through rates during the search process, which inevitably leads to a higher user satisfaction and better surf experience.

Users pursuing a similar search behavior are likely to have similar user needs. Hence such users should be treated similarly from the retrieval system. Taxonomies are used to classify the search behavior automatically by applying intelligent classification algorithms. After the classification appropriate actions must be initiated, such as the presentation of results in a specific view or the application of relevance functions and retrieval mechanisms, which are appropriate for the particular class. Fig. 1.4 depicts this process, where the part, which is investigated in this thesis is highlighted with a dashed red line.

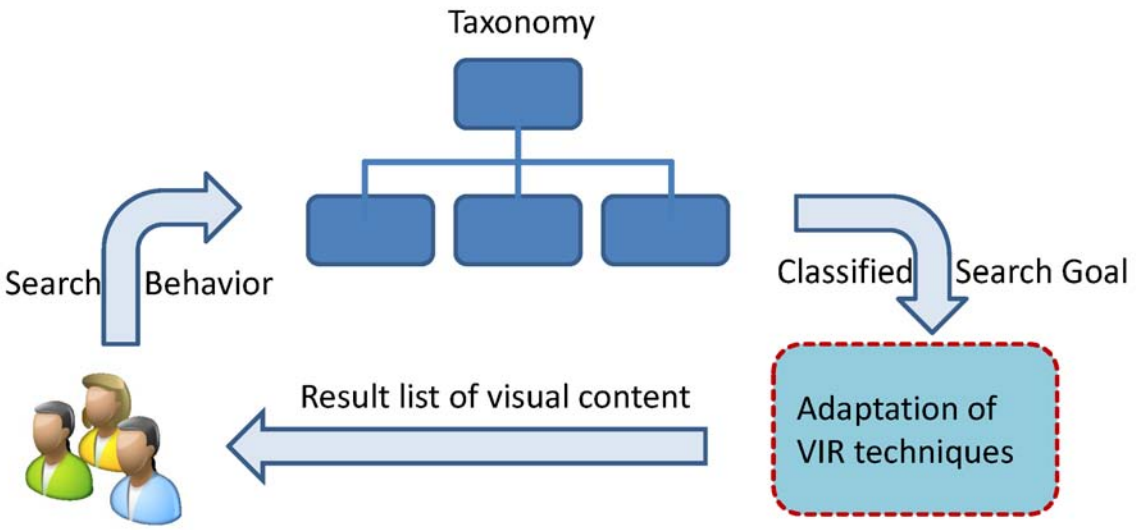

Figure 1.4: Adaptation of VIR techniques with respect to user intentions.

Chapter 4 of this work deals with adaptation of visual information retrieval techniques in context of users intentions during search. A new approach for the adaptation of retrieval mechanisms is presented, by leveraging the Bag of Visual Words technique (cf. Kogler and Lux [49] and [51]). The research question 4 is elaborated in chapter 4.

Specific Research Question 4 *Do users with different user needs benefit from distinctive content-based indexes?*

The remainder of this thesis is organized as follows:

- Chapter 2 gives a comprehensive overview of related work, comprising descriptions about a variety of visual features, Bag of Visual Words and user intentions in information and visual information retrieval.
- Chapter 3 addresses my own work concerning content-based techniques for visual information retrieval. Bag of Visual Words approaches are exploited for video retrieval, video summarization and content-based image retrieval. A new fuzzy approach, which extends the state of the art Bag of Visual Words approach, is presented.
- Chapter 4 addresses my own work concerning visual information retrieval in context of users intentions, by leveraging again different Bag of Visual Words approaches.
- Chapter 5 concludes this thesis, by summarizing and reflecting about the findings.

Chapter 2

Related Work

The retrieval of images and videos mainly relies on text based queries, which are submitted to a retrieval system. This may not work well, especially if multimedia data are badly annotated. Hence information concerning the visual content must be additionally used to improve image and video retrieval. Moreover user intentions should be considered in terms of visual information retrieval, in order to adapt retrieval results appropriately to various user needs. This chapter addresses related work in the field of visual information retrieval and user intentions respectively. It provides the reader with an overview of state of the art techniques.

2.1 Visual Information Retrieval

Visual information retrieval (VIR) has rapidly evolved over the years. Whereas early retrieval systems mostly rely on text queries and meta data, second generation VIR systems leverage the content of videos and images as an additional and complementary information source. Visual information reveals new possibilities to retrieval systems, by using features such as color, texture and shape to ascertain the similarity of indexed multimedia data.

2.1.1 Global Features

Global features characterize images or frames of videos as a whole. They are automatically extracted by applying various image processing and analysis techniques, in order to describe the visual content. The MPEG-7 visual standard for content description should be mentioned at this point. It comprises various descriptors, which describe color, texture and shape features (see [90] for further details). Different global features are presented in the following sub sections to provide a comprehensive overview of the topic.

Color Features

Color as one of the most perceptual features serves as a baseline for content-based image retrieval systems. Color features can be easily obtained by counting the pixel values of the respective color channel using a certain color model. A color model for digital image retrieval is the RGB model, which represents the color space by red, green, blue, which are the primary colors of light. The model is based on a cartesian coordinate system, which is depicted in the following Figure 2.1.

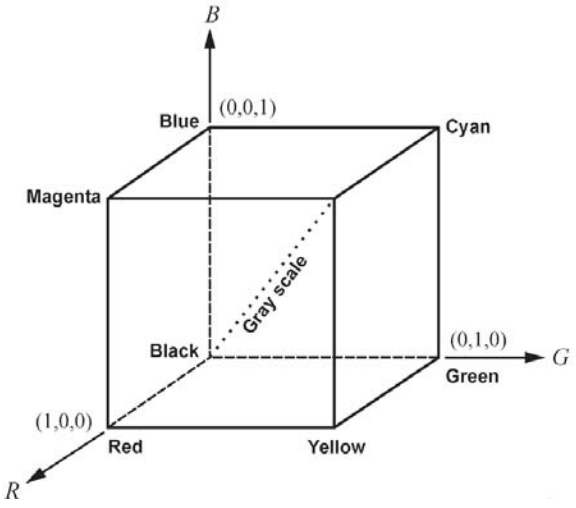

Figure 2.1: RGB color model. [66]

Alternatively a user oriented model, supporting the human perception of colors, such as HSV (2.2) can be used. The HSV color space is divided into:

- Hue, denoting the type of the color.
- Saturation, constituting how much the color is mixed with white and hence expressing the color's purity.
- Value, brightness or intensity of light.

The advantage of the HSV model compared to RGB is the segregation between color and intensity values, what is more related to the human perception of color. Humans ascribe more importance to intensity changes than to changes in color.

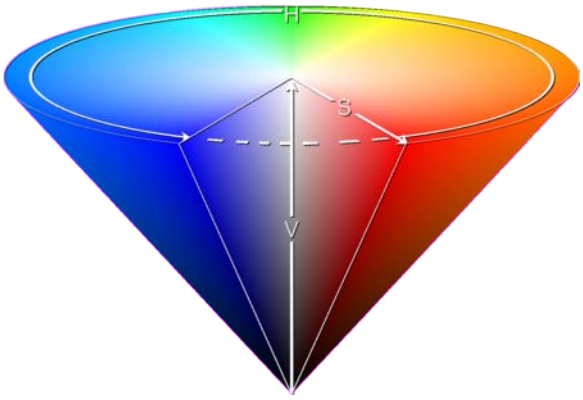

Figure 2.2: HSV color model.

The HMMD color space (see [44]) derives from the HSV and the RGB model by using the hue, the minimum value and maximum value in the RGB channels and the difference between the maximum and the minimum. More color models than the three mentioned exists such as YCbCr used in the JPEG compression and digital video, which divides the color space into luminance and chrominance color channels. A good overview of various color spaces is given in [67] and [7].

The visual representation of the images, according to the chosen color space, can be computed using various approaches. A straight forward approach as described in Deselaers et al. [21] is to down-sample the image to a predefined thumbnail size (eg. 32 x 32) and leverage the pixel values of the image as a feature. This simple approach has shown to be effective for the retrieval of medical images as stated by Deselaers.

Color Histograms are widely used in content-based image retrieval because they are easy to implement and serve as good baseline in order to describe the image content. The color space is usually quantized to reduce the amount of dimensions of the visual descriptor and pixel values within the range of a quantized color are summed up. The color distributions over each channel is measured or one distribution over all channels is computed to determine the visual representation of the image known as histogram. The computed histograms are invariant to affine transformations, namely translation and rotation.

Color moments (see [95]) represent information about the first, second and third statistical moment of the probability distribution. Statistical values such as the mean, the variance (the square of the standard deviation) and the skewness are computed over all color channels and are usually compared by computing the absolute difference between the values of two pictures. The listed formulae (taken from [26]) represent the mathematical definitions of the color moments, where N constitutes the number of pixels in the image and f_{ij} denotes the i-th color channel of pixel j in the image.

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} f_{ij} \tag{2.1}$$

$$\sigma_i = (\frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^2)^{\frac{1}{2}} \tag{2.2}$$

$$s_i = (\frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^3)^{\frac{1}{3}} \tag{2.3}$$

The pitfall of color histograms and color moments is that they do not include

spatial information. This will be a problem if two images share the same distribution of pixels, but differ in their layouts.

Color Correlograms ([35]) consider the spatial relationships between the image pixels by using co-occurrence matrixes of pixel values. The entries of the matrix constitute the amount of co-occurring colors within a spatial distance (depicted in Fig. 2.3. Given a color Cj an entry of the matrix denotes the amount of pixels of color Ci within a distance d. In the left image of Fig. 2.3 the color Ci with the pixel value 6 is located in the neighborhood of the color Cj with pixel value 2. The displacement vector d(0,1) means that colors being located in the same row and to the left of the color Cj must be considered. By counting the occurrences of the color Ci within the distance d(0,1) of the color Cj an amount of 3 must be stored in the co-occurrence matrix, as depicted on the right of Fig. 2.3. Different distances yield to different co-occurrence matrixes, thus requiring a huge amount of storage and being less efficient in retrieval in terms of speed. Hence only the main diagonal of the co-occurrence matrix is computed for the auto color correlogram (ACC) , covering only relationships between similar color values.

1	1	7	5	3	2
5	1	6	1	2	5
8	8	6	1	2	5
4	3	4	5	5	1
8	7	8	7	6	2
7	8	6	2	6	2

	1	2	3	4	5	6	7	8
1	1	2	0	0	0	1	1	0
2	0	0	0	0	1	1	0	0
3	0	1	0	1	0	0	0	0
4	0	0	1	0	1	0	0	0
5	2	0	1	0	1	0	0	0
6	1	3	0	0	0	0	0	1
7	0	0	0	0	1	1	0	2
8	1	0	0	0	0	2	2	1

Figure 2.3: Left image with pixel values. Right image denotes the co-occurrence matrix for a distance in pixels of (0,1) adopted from [66].

The depicted Co-occurrence matrix shows eight different color values and their correlations, hence preserving the spatial relation between them.

Texture Features

Texture features are widely used in Visual Information Retrieval, because they provide powerful and discriminating descriptors. The notion of texture, although no common definition prevails, can be described by certain image properties such as smoothness, regularity and coarseness. These properties, describing the correlation of brightness patterns within images or image regions, can be computed by means of statistical approaches. For this purpose statistical moments like the mean to get the overall brightness, the standard deviation, which is used to get an impression of the contrast, the skew, showing the asymmetry around the mean, the energy and the entropy are used. The first three descriptors have already been defined mathematically in the previous section (see Eq. 2.1, 2.2 and 2.3) based on color channels and are now extended with the two mathematical formulations of the energy and the entropy. r_j denotes the j-th gray-level and $p(r_j)$ constitutes it's probability of occurrence.

$$Energy = \sum_{j=0}^{L-1} [p(r_j)]^2 \tag{2.4}$$

$$Entropy = -\sum_{j=0}^{L-1} p(r_j) log_2[p(r_j)] \tag{2.5}$$

Images exhibiting high energy will have fewer grey-levels than images with a lower energy.

Example 2 *Lets consider two simple images img_1 and img_2 with 15 pixels and 10 possible grey-levels (L_1, L_2 ... L_{10}). The first image img_1 contains 2 pixels of grey-level L_1 up to grey-level L_7 and 1 pixel with grey-level L_8. Grey-levels L_9 and L_{10} are not present in image img_1. The corresponding energy of the image accounts for* $0.12\dot{8}$ *($(\frac{2}{15})^2 + (\frac{2}{15})^2 + (\frac{2}{15})^2 + (\frac{2}{15})^2 + (\frac{2}{15})^2 + (\frac{2}{15})^2 + (\frac{2}{15})^2 + (\frac{1}{15})^2 + (\frac{0}{15})^2 + (\frac{0}{15})^2$). The second image img_2 contains 10 pixels with the grey-level L_1 and 5 pixels with the grey-level*

L_2. Hence the energy accounts for $0.55\dot{5}$ *(*$(\frac{10}{15})^2+(\frac{5}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2+(\frac{0}{15})^2$*).*

Additionally the more complex an image is, the higher its entropy will be.

The statistical moments are usually computed on image histograms, what limits the descriptiveness of texture descriptors. The spatial information is neglected in this case, but will be considered by representing the co-occurring pixel values and their positions with respect to each other. A grey-level co-occurrence matrix stores the information about the spatial relationship of the pixel intensity values. It contains elements denoting the pair-wise occurrence of pixel intensities z_i and z_j around a position in an image within a certain range, constituted by a displacement vector d. As for grey-level histograms statistical moments are computed on normalized co-occurrence matrixes.

Other texture features, which were applied in image retrieval systems such as QBIC [77] and Photobook [81], are the Tamura features presented in Tamura et al. [98]. They comprise properties like coarseness, contrast and directionality and are very useful to get discriminating descriptors. To measure the texture granularity, known as the coarseness, moving averages around pixels within a certain window size are computed. The difference between the non-overlapping moving averages is calculated in both horizontal and vertical directions. For the window size 2^k x 2^k, a value k is selected, which maximizes the difference between moving averages. This k is then used to compute the sum for every pixel $g(i,j)$ intensity yielding a value, which represents the coarseness of an image and can be expressed by the following formula. The resolution of the image is represented by m and n. The variables i and j constitute the position of a pixel with its intensity value $g(i,j)$ in an image.

$$Coarseness = \frac{1}{mn}\sum_{i,j} 2^k g(i,j) \tag{2.6}$$

The contrast is defined with the following formula:

$$Contrast = \frac{\sigma}{(\frac{\mu_4}{\sigma^4})^{\frac{1}{4}}} \tag{2.7}$$

σ is the standard deviation and μ_4 denotes the 4-th statistical moment. For directionality the magnitude of the gradient vector, which points in the direction of the highest intensity change and the angle are computed.

Shape Features

Beside color and texture features, shape features constitute another possibility to represent images. Objects within images are described by several properties such as their area or their boundary and so forth. Shape features are usually applied on segmented images regions, which are retrieved through various approaches like thresholding, where image intensities are compared to a predefined or adaptive threshold to determine the membership to a region. Region based segmentation is another approach, where pixels are considered to be part of an object if a connectivity between them is existing (e.g. region growing). Further advanced segmentation strategies like watershed segmentation [102] or the GrabCut algorithm [86] can be used to prepare an image properly for the extraction of appropriate shape features.
State of the art methods for shape feature extraction and description can be roughly categorized in **boundary-based** and **region-based** approaches. Region-based approaches are simple but effective methods to describe objects by their area, the perimeter, the eccentricity, which is defined as the ratio between the major and minor axis of an object, the aspect ratio of the bounding box around the object and the central moments proposed by Ming-Kuei Hu in [33]. Describing an object by simple features like area and perimeter for instance yield to unsatisfactory projections to the feature space. The mathematical definition of the object's area represented by connected pixels is:

$$A_i = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} O_i(x, y) \tag{2.8}$$

Lets consider the following example taken from [66].

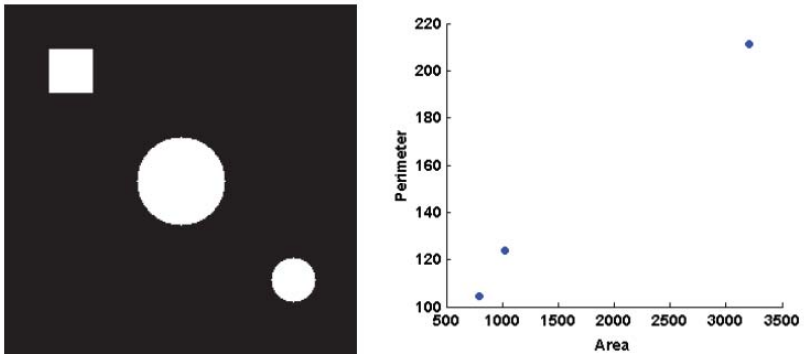

Figure 2.4: Shape features adopted from [66]

The feature point denoting the big circle is farther away from the feature point denoting the small circle than the feature point of the small rectangle, although the big and the small circle tend to be more similar. The description of the image content by means of such simple features do not yield a desired representation, which is discriminating enough to position their description adequately in the features space. Hence additional information must be considered to meet the requirements of proper representation.

Boundary-based features encapsulate a description of the object's contour. Such descriptions turn out to be more or less accurate, depending on chosen method. One simple approach would be to compute the bounding box of the object, which is defined as the minimum rectangle, needed to encapsulate the particular shape completely. Such a descriptor can be applied, if only the size and the location of the object are required. A more compact representation of the object is retrieved by applying a polynomial approximation of it. The convex hull is a more meaningful descriptor by spanning a minimal convex polygon around the shape, which is comparable with an elastic band, wrapped around the object.

Chain codes constitute another technique to describe an object's contour by coding lines of a predefined length and direction of the boundary, while tracing the contour of the shape. The Freeman code, which is an implementation of a chain code uses 8

different directions, in order to describe the object's contour. As a consequence the shape to describe is approximated by the computed chain code.
Fourier descriptors store the information of an object's boundary through coordinate pairs computed along the boundary, which are represented as complex numbers. The complex coordinates are transformed via the Discrete Fourier Transform to build the Fourier descriptor. The advantage of the Fourier descriptors is their descriptiveness of the contour by using only a small number of points to restore the appearance of the shape. A more detailed description is given in [26].

Compact Composite Descriptors

Compact Composite Descriptors combine visual features to one joint histogram, hence drawing on the richness of visual information. The Fuzzy Color And Texture Histogram (FCTH) [13] puts both color information and texture information in one single histogram. Three fuzzy systems for smooth color and texture computation are used, what results in 192-valued feature vector. An image is divided into a multiple of blocks, which passes a fuzzy membership computation to predefined colors belonging to the HSV color model. Additionally texture information is extracted, by computing features that represent energy in high frequency bands from a Haar wavelet transform on the luminosity. The computed features are classified by a fuzzy system to different texture areas. Each texture area is made up of 24 subregions denoting the predefined colors, what accounts for a 192 bin feature vector. Each block is assigned to the corresponding bin based on color and texture information.

The Color and Edge Directivity Descriptor (CEDD) [12] is another compact composite descriptor used for natural color images. It uses the same fuzzy system to get the membership of a block to colors of a 24-color palette. The difference between CEDD and FCTH is that CEDD uses digital filters from the MPEG-7 Edge Histogram Descriptor [106] in order to get information about the directionality of an edge. The assignment to different bins, constituting the directionality of an edge, is fuzzy, which means that an analyzed image block can contain more than one type of

edges.

The two mentioned descriptors are combined to a Joint Composite Descriptor (JCD) in [11], which uses 7 texture areas with 24 sub regions for each area. Since color information is extracted identically for the CEDD and the FCTH descriptor, only texture related features must be combined.

One more compact composite descriptor based on fuzzy rules has to be mentioned. The Brightness and Texture Directionality Histogram (BTDH) [14] combines brightness and texture information and captures their spatial distribution. Fuzzy classification is used again, in order to get the brightness values of image pixels and to compute the directionality, which conveys texture characteristics. This descriptor can be applied for images, where texture information plays and important role such as medical radiology images or in general grey-level images.

An overview of the global features, mentioned in this section, is given in Table 2.1.

feature	**descriptors**	**characteristics and merits**	**demerits**
color features	Color Histogram	effective representation of color content; robust to translation and rotation; fairly insensitive to changes in image resolution and partial occlusion	sensitive to changes in lighting; do not include spatial information
continued on next page			

feature	**descriptors**	**characteristics and merits**	**demerits**
	Color Moments	compact representation, since only 3 numbers are used for each of the 3 color components	may reduce the discriminative power due to their compactness; do not include spatial information, what's bad for images whose color distribution is identical
	Color Correlogram	considers spatial information, by using co-occurrence matrices	big storage requirements, hence only the main diagonal of a matrix is stored
statistical texture features	statistical moments, energy and entropy	compact descriptors; standard deviation is a concise representation of overall contrast; one floating point value for each descriptor is required for storage; rotation, scale and translation invariant	limited discriminative power; spatial relationships among pixels is only regarded, if co-occurrence matrixes are used
	tamura texture descriptor	discriminative descriptor, due to coarseness, contrast and directionality;	sensitive to lighting
shape features	region based approaches	valuable descriptors to find similar objects in a constrained context; compact representation and small storage requirements	rely on segmentation, in order to get meaningful objects; difficult to measure accurate and meaningful shape-based similarity; limited discriminative power

continued on next page

feature	descriptors	characteristics and merits	demerits
	boundary based approaches	meaningful description of objects, by considering their contour; bounding boxes are less descriptive, but easier to compute than advanced descriptors such as Fourier descriptors	preceding segmentation is needed
composite features	CEDD and FCTH	combination of color and texture features; suitable for retrieving natural color images; small storage requirements (CEDD needs 54 bytes per image, FCTH needs 72 bytes per image)	sensitive to lighting
	BDTH	combines brightness and texture; suitable for retrieving grayscale and radiology medical images	not suitable for color-based retrieval

Table 2.1: Overview of global features for visual information retrieval.

2.1.2 Local Features

Instead of regarding images in a global way, local features describe the visual content around special points, also called interest points, features points and key points in an image. Key points have been thoroughly used in object recognition and visual tracking and are described in this chapter (cf. [46]) to get an overview of various algorithms used for feature point detection and description. The common strategy to compute local features is to determine edges, corners and regions with high contrast

in an image and extract visual information from local image patches. Information such as gradient magnitudes in various directions are considered to compute local descriptors.

A short introduction to edge detection

Edge detection is a prerequisite of interest point detection and hence an important step towards it. During edge detection high frequencies of the image content are retained while low frequencies are disregarded. In other words high pass filters perform edge detection by considering dominant intensity changes in an arbitrary direction. A commonly used high pass and directional filter is the Sobel edge detector, which considers horizontal and vertical intensity changes. Filtering is conducted by using convolution either looking at the intensity changes in x or y direction. Convolution is a term used to describe linear filtering with a kernel (a mask) moving over each pixel of an image. Hence a new value for each pixel is computed, in order to get a blurred version of an image or to find high intensity changes in case of edge detection. Mathematically a Sobel filter (operator) is looking for gradients with high magnitudes, which are defined as 2D vectors and constitute an edge.

$$grad(I) = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})^T \tag{2.9}$$

The gradients are obtained by computing the difference of intensity values of image pixels. Meaningful edges with high intensity changes are retrieved by applying a threshold. Unfortunately its not that easy to define a good threshold. A low threshold would lead to many edges probably containing less representative ones, whereas a high threshold would keep only few edges. An illustration of an edge in a grey-level image is depicted in Fig. 2.5. It shows the first and second order derivatives of such an edge. The first order derivative constitutes the difference in intensity change and the second order derivative gives information about the maximum of an edge by looking at zero crossings.

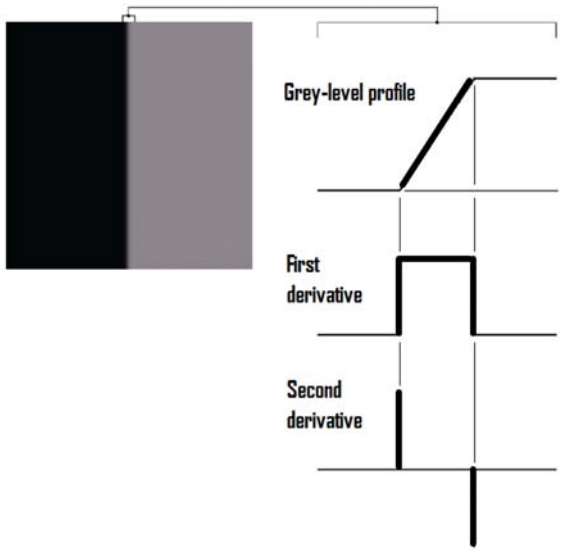

Figure 2.5: Edge detection using first and second order derivative. Adopted from [66]

Noise effects the validity of the first and second order derivatives as depicted in Fig. 2.6. The derivatives are more distorted the higher the amount of noise is, due to the additional appearance of high frequency components. The noise in the grey-level picture represented by its grey-level intensity profile on the left increases from top to bottom. The distortion is rendering the derivatives more and more useless. Therefore image smoothing is often applied before edges are computed, in order to reduce noise while keeping significant information about edges. Several low pass filters, rejecting high intensity values and preserving low ones, are applied in practise such as: the Gaussian filter, the mean filter or the median filter.

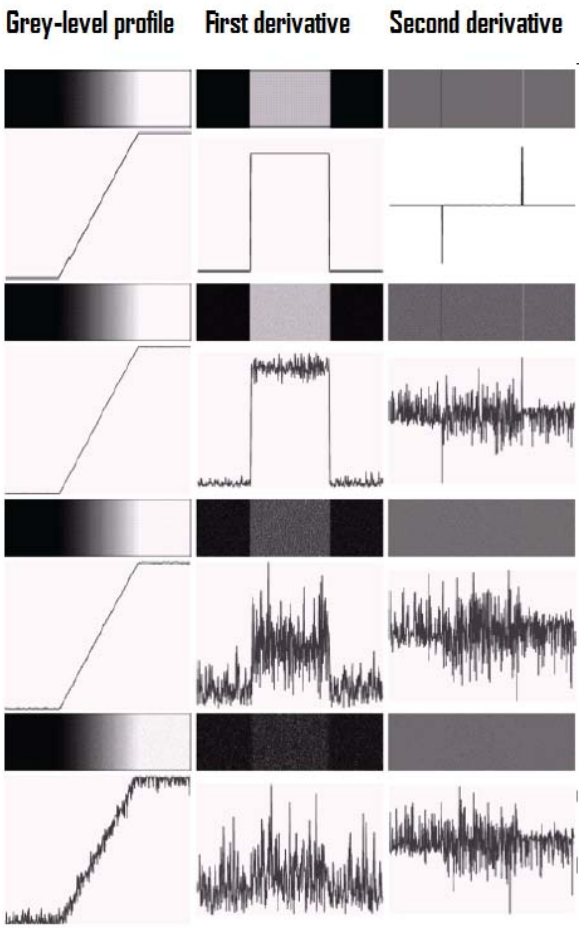

Figure 2.6: Impact of noise on edge detection.Adopted from [66]

Harris Corners - Interest Point Detection

Corners or more general interest points are valuable features, which can be extracted from an image and used for further object detection and tracking. In general corners appear where two dominant edges point into orthogonal directions. To be more precise the average intensity change in a small window around an interest point is computed and can be mathematically described as:

$$\sum_{u,v}(I(x+u,y+v)-I(x,y))^2 \tag{2.10}$$

The variables u and v denote the dimensions of the window, which is considered for calculation. The average intensity change is computed in all possible directions. After obtaining the direction, where a maximal intensity change occurred, the intensity change in the orthogonal direction is examined too. If this intensity change is also high a corner is obtained. A Sobel filter can be used for edge detecion.
The result of the corner detection process can be further improved by utilizing a non-maxima suppression. Corners on thick edges are suppressed and only local maxima of corner values are retained. Hence only meaningful corners (see Harris corners [30]) with a large value over a certain threshold not adjacent to each other (local maxima), are chosen. This further boosts the corner detection process. The computation of uniformly distributed Harris corners across the image, where corners appearing near to each other are neglected, is describe by Shi and Tomasi in [89].

Features From Accelerated Segment Test (FAST) - Interest Point Detection

In place of applying the Harris corner detector, which examines the rate of intensity changes in orthogonal directions by computing the image derivatives, the FAST detector (see [85]) looks at intensity changes in a certain area. This is performed without the expensive calculation of image derivatives and lead to a faster but probably more unstable computation of possible interest points. Nevertheless the quick detection of interest points allows for real time analysis of video sequences.
Features From Accelerated Segment Test measures pixel intensities on an arc around a center point p (the possible interest point). Point p is considered an interest point if the pixel intensity values around p differ significantly from p's value.
In order to find reliable interest points, the radius of the investigated arc should be provided as a parameter. Empirical knowledge states that a radius of 3 pixels yields good result while keeping high efficiency. Fig. 2.7 depicts the examined pixels and the arc with the predefined radius. 16 pixels in the neighborhood need to be investigated in total. A trick can be applied to lower computation time, by looking only at those

pixels separated by 90 degrees (e.g. top, bottom, right and left). Hence center points can be rejected as interest points if they do not pass the preceding test, where 3 of the surrounding pixels must be all brighter or darker.

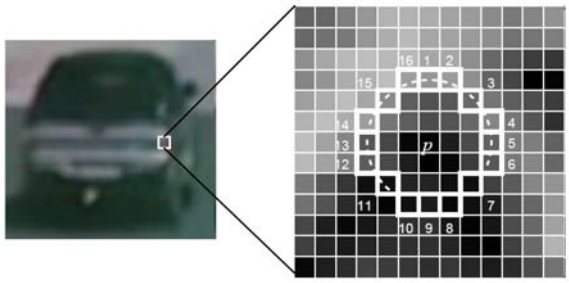

Figure 2.7: Interest point detection using the FAST algorithm.

Speeded-Up Robust Features - Fast Hessian Detector

Computer vision application often face the problem of scale changes (level of detail changes). The previously defined FAST algorithm struggles with matching objects of interests at different scales. Interest points representing an object in an image are hard to preserve if the object is moving across the image plane and its size is reduced. If we use a fixed size neighborhood around a center pixel p, p will probably not be regarded as an interest point anymore, whereas it is preserved by using a scale invariant feature detector.

This section describes the Speeded-Up Robust Feature (SURF) algorithm ([6]) for interest point detection. The algorithm is more reliable in computing stable interest points than FAST but computationally more expensive due to its various processing steps.

In order to consider scale changes, Gaussian blur filters are applied. Hence differently scaled image versions are computed of an original image by retaining low frequency components. In other words a Gaussian filter rejects those pixels with high intensity values while preserving lower intensity values.

Scale invariant features consider interest points at different scales (various blurred

versions of an image), which differ in detail. Only those points in an image, which survive a large set of scales become interest points. The SURF algorithm determines such interest points by computing the Hessian matrix at each pixel, which contains the second order partial derivatives of a function.

$$H(x,y) = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix} \tag{2.11}$$

It measures the local curvature of a function and its determinant constitutes the strength of the curvature. Locations with a high local curvature are considered as corners or interest points, which describe a high variation of intensity in more than one direction.

The Hessian matrix can be computed by applying Laplacian of Gaussian kernels at different scales, which is a costly process. Instead of smoothing the image repeatedly with Gaussian kernels, SURF uses box filters for approximation (see Fig.2.8). For implementing box filters in constant time, integral images (see [103]), which constitute subregions within images, are used. The applied filters in x,y and xy directions are depicted in Fig. 2.8.

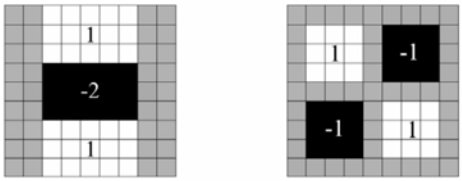

Figure 2.8: Approximated Gaussian Kernels [6]

The left kernel computes an approximation of the second-order derivative in the vertical direction, whereas the rotated version computes the approximation of the second-order derivative in the horizontal direction. The right kernel computes the second-order partial derivatives in xy-direction.

Scale Invariant Feature Transform - Difference of Gaussian

The Scale Invariant Feature Transform (SIFT) algorithm (see [59] and [60]) comprises four main steps to compute scale invariant features, which are invariant to rotation, scale and other image transformations. The first two steps are used to localize interest points:

1. Scale-space extrema detection: All possible locations at different scales are searched, in order to determine potential interest points.
2. Stable interest points are localized.
3. Orientation assignment is conducted.
4. Key point descriptors are computed.

In order to detect potential interest points, different scales (usually 3 scales are considered) of an image must be investigated. This is accomplished by applying a Gaussian scale-space kernel to all possible locations of the respective image. Fig. 2.9 depicts Gaussian smoothed images over different octaves, where the scale factor to smooth the images is enlarged. An octave constitutes the resolution of the image, which is steadily sub-sampled by a factor of two. Hence only every second row and second column is maintained to get a sub-sampled image. After that the difference between two consecutive Gaussian blurred images is computed for every octave.

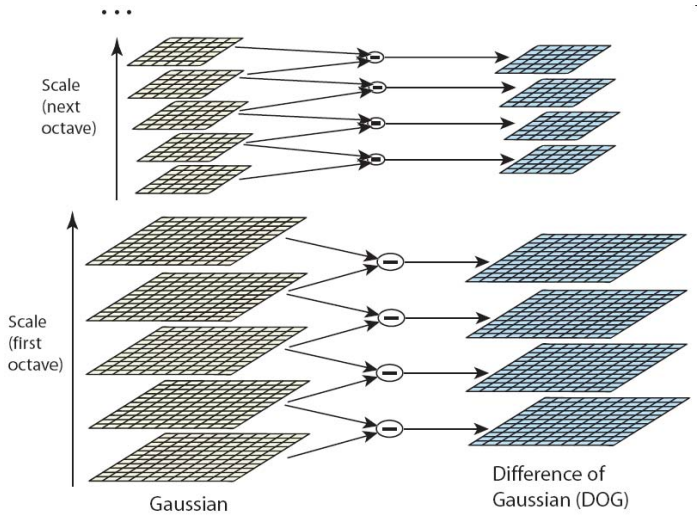

Figure 2.9: Difference of Gaussian images [60].

Local maxima and minima in the smoothed images are detected, by comparing a pixel to its 26 neighbors, which are composed by its eight surrounding pixels in the current scale and the 9 neighbors in the two adjacent scales. This step is depicted in Fig. 2.10.

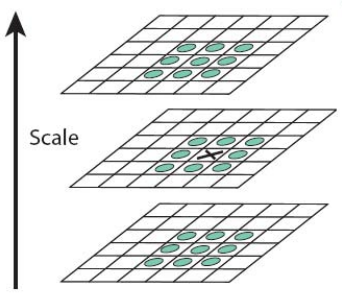

Figure 2.10: Detection of local maxima and minima in the scale space [60].

After the detection of potential interest points, those interest points having a low contrast and which are poorly localized along an edge are rejected, hence retaining stable and dominant points. The next two figures exemplarily depict located interest

points in two different images with a resolution of 352x288. (Fig. 2.11 and Fig. 2.12). A thorough comparison of region and interest point detectors is given in [70].

Figure 2.11: Interest point detection example 1

Figure 2.12: Interest point detection example 2

Interest point detector comparison

Because of the expensive computation of Difference of Gaussian images over various scales and octaves the SIFT detector performs worst. The computation time accounts for more than 300 milliseconds, followed by the Harris Corner detector and the SURF detector with a computation time of nearly 150 milliseconds (see Fig. 2.13). The FAST detector takes the lead by only looking at pixels on an arc around a putative

interest point. Interest points are computed in less the 50 milliseconds. Features From Accelerated Segment Test performs at least 3 times faster than the aforementioned detectors.

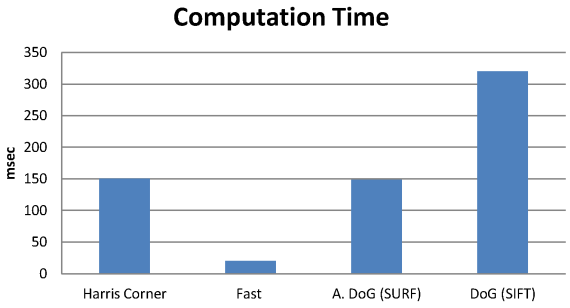

Figure 2.13: Computation time, to detect interest points. Test Environment: Intel Core 2 Duo CPU 2.3 GHz with 4GB Ram. (cf. [46])

For a qualitative evaluation of interest point detection snapshots of a car moving along the road are taken at time t and time $t + 1$. Referring to Fig. 2.14 all the detectors perform quite well in detecting feature points between the consecutive frames. However, the application of the FAST detector yield more false positives, than the use of SIFT, SURF and Harris Corner detectors. FAST only considers pixel intensities around an interest point and computes absolute differences of pixel intensities. Image derivatives or various scales are not taken into account.

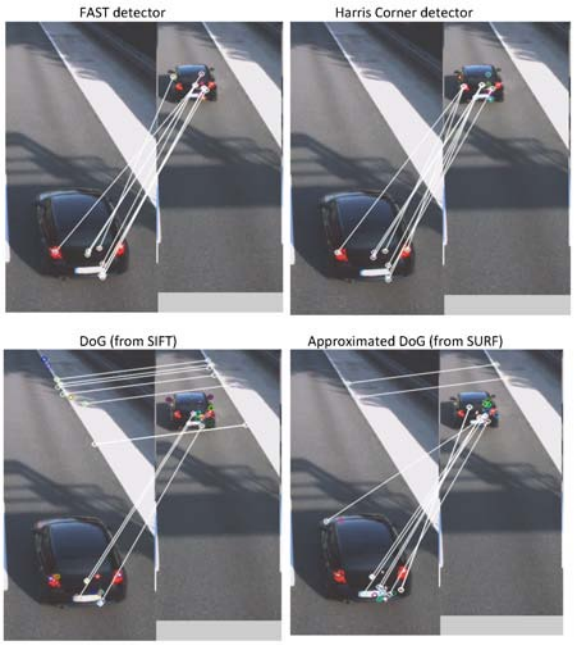

Figure 2.14: Qualitative evaluation of interest point detectors.

The performance evaluations, presented in Bay et al. [6], can be considered to gain further insights.

Descriptor computation of local interest points

After the detection of interesting points in an image, feature descriptors are computed, which store information of local image patches around the interest points. An image patch constitutes a rectangular region around an interest point. The size of the region depends on the selected descriptor. Various approaches of describing image patches exist in literature such as Histogram of Oriented Gradients, RGB-SIFT and so forth ([72] and [71]). The most prevailing approaches for object recognition are SIFT ([59], [60]) and SURF ([6]) related descriptors.
The Scale-Invariant Feature Transform creates the descriptors by computing local gradient magnitudes and their orientation in a 16x16 region around the interest point.

4x4 8 bin descriptors are computed, where each bin corresponds to a dominant gradient vector pointing in the particular direction. This yields to a 128 dimensional feature vector, which describes the local image patch, being invariant to affine transformations, lighting and contrast change. The following Fig. 2.15 exemplarily depicts the creation of the descriptor. Gradient orientations and gradient magnitudes are computed around the interest point (see right image of Fig. 2.15) . They are accumulated and put into orientation histograms, which describe the subregions around the interest point (see left image of Fig. 2.15).

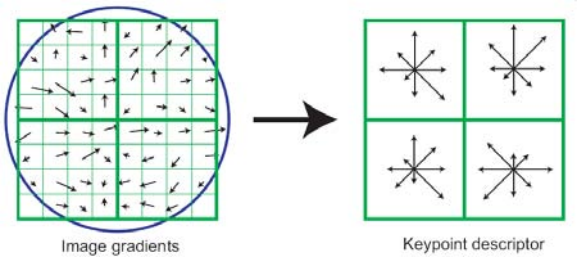

Figure 2.15: Local interest point descriptor based on a 2x2 subregions with 8x8 samples leading to 4 8 bin orientation histograms. Picture was taken from[60].

The SURF descriptor is usually computed on 4x4 sub regions around an image interest point. Wavelet Responses are computed and summed up to measure the intensity change in both horizontal and vertical direction. Each subregion is expressed by a four dimensional descriptor, what leads to a 64 dimensional descriptor for all subregions or in other words for the region around the interest point, representing the image patch. SURF descriptors can be computed faster than SIFT descriptors due to the aforementioned techniques to detect interest points and describe the patches around them, what makes the descriptor a good choice for real-time applications. Referencing to [6], it is in no way inferior to SIFT concerning its reliability in computing good features to match. The following Table 2.1.2 summarizes SIFT and SURF related approaches and techniques.

	SURF	SIFT
used interest point detector	Fast Hessian	Difference of Gaussian
descriptor length	usually 64 element vector; variants of SURF can either increase (128) or decrease (36) the length; the lower the descriptor length the worse the matching performance of objects get; but small descriptors yield faster matching	128 element vector; the vector length can be reduced by using dimensionality reduction techniques
computation time	depends on the descriptor size; usually SURF is faster than SIFT	SIFT is 3 times slower than SURF
reliability of detected features	similar local features in different images (e.g. distorted version of a reference image) are detected fairly well; both approaches provide good performance measurements, concerning object recognition (see [6] for further details)	

A thorough comparison between different local interest point descriptors is provided by Bay et al. in [6] and by Krystian Mikolajczyk in [72].

The extraction of local information of image patches around salient points leads to many local features. The amount of extracted local features depends on the resolution and the visual content of the image and accounts for more than 100 features per image. To search for similar images all local feature vectors of all images must be compared, which is a time consuming operation. Hence an additional quantization step is conducted to reduce the amount of time during a similarity search. The Bag

of Visual Words technique, which is explained in the next section, is employed to perform this additional step.

2.1.3 Bag of Visual Words

The Bag of Visual Words (BoVW) approach has its roots in information retrieval, where text documents are indexed as term vectors, composed of all possible terms in the document collection. The so called vocabulary, representing an unordered collection of all terms in the document collection, is used together with the term frequencies of each document to computed the document specific term vectors. Relating to the vector space model ([87]) from information retrieval, these term vectors are compared during document retrieval, in order to find similar documents. Similarity models such as the cosine similarity, measuring the angle between to vectors, the Euclidean or Manhattan distance and many more can be used to return a ranked list of documents, fitting a search query in the best possible way.

Cosine Similarity:

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \tag{2.12}$$

- $\vec{d_j}$ is a term vector denoting the term frequencies in a document
- $\vec{q}$ is a term vector denoting the term frequencies in a query

Similarity using the Euclidean Distance:

$$sim(d_j, q) = \sqrt{\sum_{i=1}^{i=n} (d_j[i] - q[i])^2} \tag{2.13}$$

Similarity using the Manhatten Distance:

$$sim(d_j, q) = \sum_{i=1}^{i=n} |(d_j[i] - q[i])| \tag{2.14}$$

- d_j denotes a document and $d_j[i]$ denotes the frequency of term i in the document

- q denotes a query and $q[i]$ denotes the frequency of term i in the query

Example 3 *Assume that our collection contains two documents:*
Document 1: "Information Retrieval and Image Retrieval are challenging research topics."
Document 2: "Content-based Image Retrieval is challenging and relies on low level features."
The resulting vocabulary, comprising all the terms in the two documents, reads as follows:
{*Information, and, Image, Retrieval, are, Content, based, challenging, research, low, level, features, is, and, relies, on, topics*}.

Based on the vocabulary the occurrences of the terms for the respective documents are counted and summed up, to build 17-dimensional feature vectors.
Feature vector of document 1 d_1*:* $[1, 1, 1, 2, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1]$
Feature vector of document 2 d_2*:* $[0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0]$

A query q could be: "Image Retrieval".
This leads to a feature vector $[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$*, which is compared to the feature vectors representing the document collection by using a similarity metric, such as the Manhatten Distance.*
$sim(d_1, q) = |1-0| + |1-0| + |1-1| + |2-1| + |1-0| + |0-0| + |0-0| + |1-0| + |1-0| + |0-0| + |0-0| + |0-0| + |0-0| + |0-0| + |0-0| + |0-0| + |1-0| = 7$

$sim(d_2, q) = |0-0| + |1-0| + |1-1| + |1-1| + |0-0| + |1-0| + |1-0| + |1-0| + |0-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |0-0| = 11$
The query is more similar to document 1 than to document 2.

This approach can be adopted and used for Visual Information Retrieval. The Bag of Visual Words approach, to produce visual vocabularies, can be divided into two main steps:

1. Extraction of local features, appearing at salient points (corners, edges) in images

2. Clustering of previously extracted local features of an image collection by using a clustering algorithm such as k-means ([36]). The cluster centers denote the visual words (related to terms within vocabularies in information retrieval) and constitute the visual vocabulary (codebook).

The process of visual vocabulary generation is depicted at the left hand side of the following Fig. 2.16.

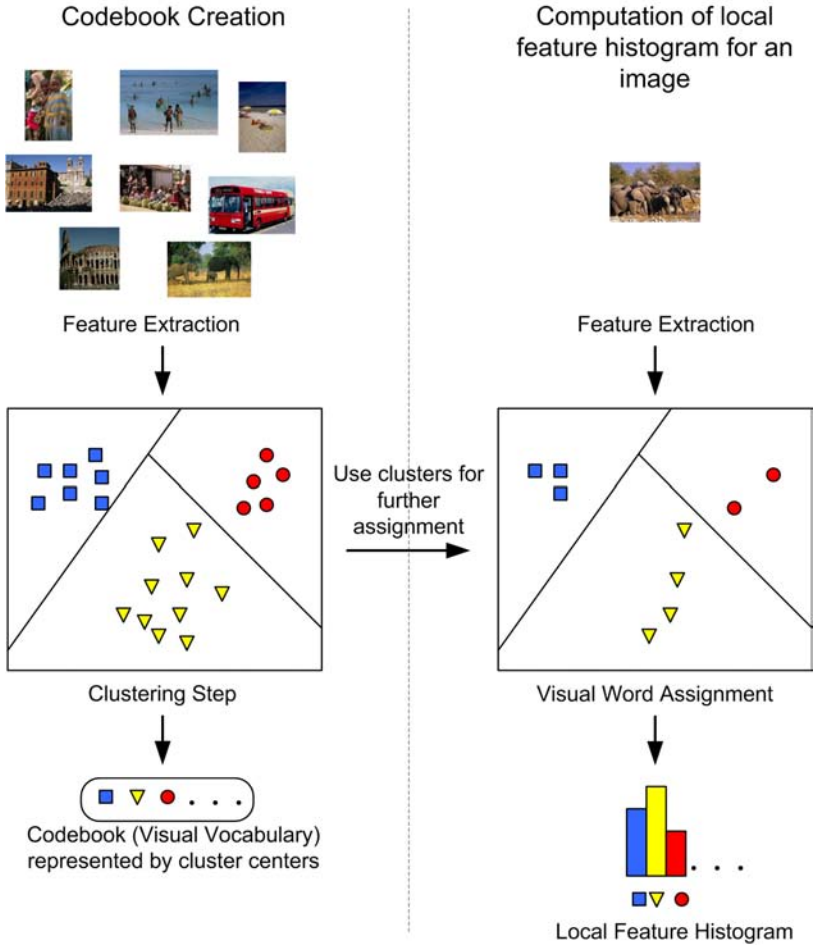

Figure 2.16: Visual vocabulary generation and local feature histogram creation.

After the computation of the visual vocabulary, which has been conducted on a training set of images, local feature histograms are created to represent the images. Local feature vectors are extracted from the image, which must be indexed, and assigned to the nearest cluster center (visual word) from the previously computed visual vocabulary. The assignments are counted and one local feature histogram for the respective image is computed (depicted at the right hand side of the picture 2.16). The count of the assignments is similar to the term-frequencies of vectors representing documents in a document collection. Referring to Fig. 2.16 the local feature histogram of the image on the right side contains 3 local feature vectors, which are assigned to the cluster center illustrated as the blue rectangle, 2 vectors correspond to the red circled cluster center and 4 vectors belong to the yellow triangle.

Bag of words for visual information retrieval, introduced in Sivic and Zisserman [91], has been addressed in various publications. A general overview of global and local features together with an evaluation comparing their performance can be looked up in [21].

Jiang et al. [41] experiment with different key point detectors, visual vocabulary sizes and suggested a new soft assignment approach of feature vectors to visual words. They regarded the distance of feature vectors, describing local image patches, to cluster centers (the visual words), by assigning these vectors to the top n nearest visual words. Hence the significance of a local feature vector to a visual word will get higher, if the vector is more closely located to it. According to the authors their approach performs better than traditional approaches from text retrieval such as term frequency (tf) or term frequency * inverse document frequency (tf*idf)[1] (see [87]). This sustains their assumption that a local feature vector should be assigned to more than one visual word, in order to retain important visual information.

Van Gemert et al. [27] propose a kernel based approach, which allows to assign

[1]tf*idf measures the importance of a query term by its occurrence in a document and its overall occurrence in all documents. The more frequently a term appears in a document the more important the document will be for the related query and hence will be ranked higher in the result list. Additionally the importance of a document decreases if the distribution of the query term over all documents gets higher.

a local feature vector ambiguously to two different clusters. Their evaluation show that their soft assignment technique improves retrieval accuracy.

Beside different weighting schemes such as binary weighing, tf or tf*idf Yang et al. [111] evaluated various vocabulary sizes, ranging from 200 up to 80,000. Their study reveals that an ideal vocabulary size does not exists and that a large amount of clusters and hence high dimensional feature histograms pose an additional burden to classifiers in terms of computation speed.

In [52] the traditional bag of visual words approach, which results in an unordered collection of local features, is augmented by spatial information by subdividing an image in equally sized regions. Key points are extracted region-wise to build a local feature histogram for every sub region.

In [22] local feature histograms are created by applying *Difference of Gaussians* for key point detection. Image patches around the key points are extracted and PCA (Principal Component Analysis) transformed to reduce dimensionality. By using image patches color information of key points is maintained, which is not the case with SIFT or SURF. During clustering a Gaussian mixture model for density estimation is used.

Snoek et al. [94] provide a comprehensive report on various Bag of Visual Words techniques, looking at concept detection in videos. Different stages are considered. Their investigations range from a variety of feature detectors and descriptors up to assignment approaches to visual words and support vector machines for classification.

Wu et al. [107] extend the bag of words approach by applying distance metric learning. The codebooks are generated for each of their investigated object categories, employing a distance metric learned during the k-means clustering process, under the assumption that this allows for preservation of semantics of the clusters. The union of the generated codebooks then forms a global codebook.

In [40] the authors address the relatedness of visual words using a visual ontology, which is constructed from a generated codebook. This approach to model the relatedness should overcome the quantization effect resulting from clustering, which

could lead to the distribution of two similar visual words over different clusters.

In [101] Uijlings et al. extend their work from [100] on Bag of Visual Words. They apply various techniques to enhance the computation speed and the retrieval effectiveness in terms of Mean Average Precision. The applied algorithms for descriptor extraction comprise SIFT and SURF descriptors and lower dimensional variants of them, in order to gain computation speed while keeping a high classification accuracy. Dimensionality reduction is both achieved by investigating a smaller region of gradient responses around the interest point and the application of Principal Component Analysis, to get rid of redundant information in high dimensional feature vectors. For visual vocabulary generation and word assignment, k-means with nearest neighbor searches and Random Forests as a faster alternative for visual word assignment are used. Object classification is conducted with support vector machines.

In [54] the authors investigate the impact of conceptual relations of visual words and contextual information by means of neighboring relations. Instead of regarding the visual words solely on the visual appearance of local image patches, the statistical distribution of local image patches with respect to various categories like coast, office, forest, etc. serves as an indicator of finding relations among visual words. The idea of their approach is that visual words, representing conceptual related concepts like eye and nose, expose similar probability distributions over the various categories, because they appear in such images, where faces are prevalent. To measure this relatedness Li et al. apply the Kullback-Leibler Divergence using the statistical information of visual words occurrences over all investigated image categories. The KL Divergence constitutes a measure for the dissimilarity of probability distributions. Visual words showing a similar distribution over the categories are grouped together in a agglomerative way by using the measurement as a similarity criterion. Hence semantic relations of local image patches belonging to the respective visual words can be considered from multiple levels. Additionally local images patches appearing together are regarded to convey valuable information. By employing the N-gram model from text retrieval, where complete texts or words are split to fragments, visual patches with neighboring

relations incorporate contextual information. In order to classify the images in one or more of the investigated categories, the authors apply support vector machines, utilizing a one-versus-all approach. Thus one SVM is generated for each category and trained with images from the respective category as positive examples and the rest of the training set as negative examples.

Another technique to incorporate semantic meaning into the visual words process is to apply latent semantic analysis on visual words. Latent semantic analysis is usually used in text retrieval to map the original feature space into an semantic space having a reduced dimensionality of feature vectors. Synonyms are grouped together by applying statistical analysis over all the documents in the document corpus. Hence words like Ferrari and Porsche are grouped to the concept car, yielding a more common representation of the words. Li et al. investigated in [55] the classification performance of various burning states of a rotary kiln, by leveraging tf-idf weighting and LSA for visual words. The application of their approach shows a good performance concerning classification accuracy.

Kesorn and Poslad propose a three stage approach ([43]) for an improvement of the BoVW process. First redundant keypoints, extracted using the SIFT algorithm, are reduced by grouping nearby keypoints through a x-mean clustering algorithm (see [80]). Hence the number of similar keypoints is diminished, what increases the computation speed of visual words, while keeping a good visual representation of the image content. The authors state that noisy keypoints, which can reduce the quality of visual words, are neglected in this way. As a result of the clustering process groups of keypoints are obtained. These groups are represented by the center points of the cluster and used to generate the visual words. While their claim to reduce computation time during the visual words generation step sounds feasible, due to a reduced amount of feature vectors in the feature space, their assumption that more noisy keypoints are neglected is questionable, because of the previous clustering stage. Not all of the generated visual words are useful. Non-informative visual words, not

contributing to the accuracy during a classification or retrieval task, are therefor dismissed.
By applying statistical methods, referring to stop word removal from information retrieval for instance, the authors aim at retaining less but more informative visual words, what affects both the computation speed and the retrieval accuracy in a positive manner. Furthermore an ontology model for the sports domain is manually created by separating important foreground objects like pole, horizontal bar or human from the background. Each object is described by its extracted visual descriptors and bag of visual words models are built for each object respectively. The established ontology, showing the relations between the concepts, objects belonging to the concepts and the visual words at the lowest level, is then used for assigning visual words from test images to the corresponding concepts. The manual construction of such an ontology is a tedious process and cannot be applied in general to large image databases. However the construction of the ontology is only conducted during training phase. Test images are processed automatically by applying all mentioned methods to receive informative visual words. These visual words are mapped to the previously constructed ontology, by assigning the visual words of the test images to one or more concepts, which are represented by concept's centroids. Visual words lying in the range of the concept are regarded to be members of it. For classification of eight sport genres they used a Naive Bayes classifier and support vector machines with a linear and a Radial Basis Function (RBF) kernel. The SVM with the RBF kernel together with a term frequency weighting of normalized visual words achieved the best results.

While in most Bag of Visual Words approaches local features like SIFT and SURF are used to extract local information of the image content, Xu et al. [110] propose 54 dimensional descriptors based on color and texture information around local interest points. The mean and the standard deviation of each color channel of a RGB image and texture features from grey level co-occurrence matrixes are computed. The proposed features are tested on a database containing aerial images of ponds, crops,

woodland and so forth.

In [2] Alqasrawi et al. leverage the spatial pyramid approach introduced in [52] for both local features and color features. They consider multiple coarse image grids to compute various feature histograms, which are concatenated to a combined feature vector. Color Moments are additionally calculated based on HSV color channel and concatenated to the final feature vector, representing the image.

In [112] Zagoris et al. compare the retrieval effectiveness of the TOP-SURF (see [99]) descriptor and the Bag of Visual Words model with two global Compact Composite descriptors (see [10]). While the TOP-SURF descriptor and the BoVW words model are an effective approach for Near-duplicate searches and objects recognition, the inspected global features provide better retrieval results when the database contains visually diversified images. They evaluate their approach by using the ImageCLEF 2010 Wikipedia[2] test collection and performed the mentioned CBIR techniques to re-rank the top k images, initially retrieved from a text search. While the multimodal approach using text and global features improves the retrieval results to text-only search the application of the TOP-SURF descriptor in combination to BoVW deteriorates the initial result, retrieved from text search. However the authors generate two large vocabularies, one with 10,000 and one with 200,000 visual words, what probably influences the effectiveness of the visual representation of the visual vocabularies.

Besides dense sampling of local image patches, by looking at every pixel at each scale, can be applied to produce a large number of local patches described by texture features. It can improve the description of images, because more local information is covered. Nevertheless more information possibly mean more noise, evoked through featureless regions as described in [78], which gives a good overview of sampling strategies and their impact on descriptive visual vocabularies.

Fig. 2.17 gives an overview of BoVW methods for interest point detection, feature extraction, vocabulary generation and visual words assignment. The corresponding

[2]http://www.imageclef.org/2010

processing pipeline is depicted in Fig. 2.18, where the numbers within the picture describe the order of processing. Table 2.2 describes the characteristics of the BoVW methods in the respective processing stages.

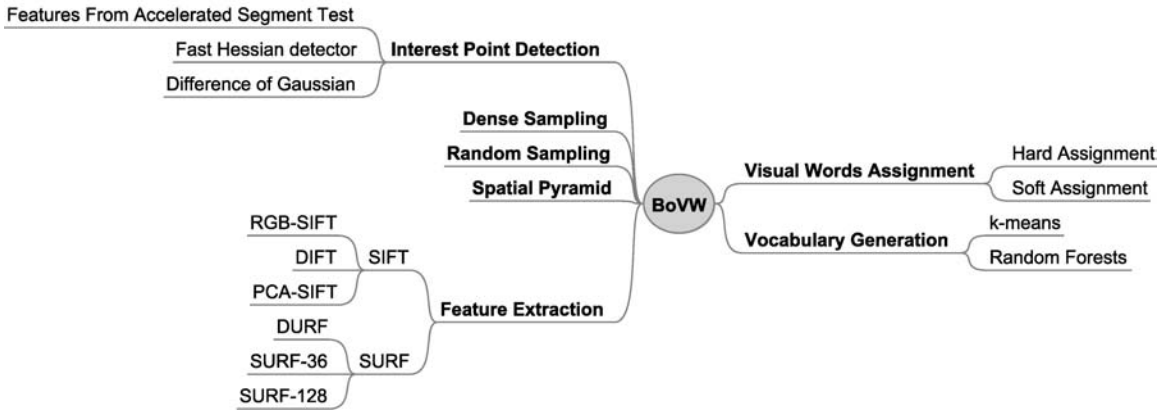

Figure 2.17: Overview of Bag of Visual Words techniques.

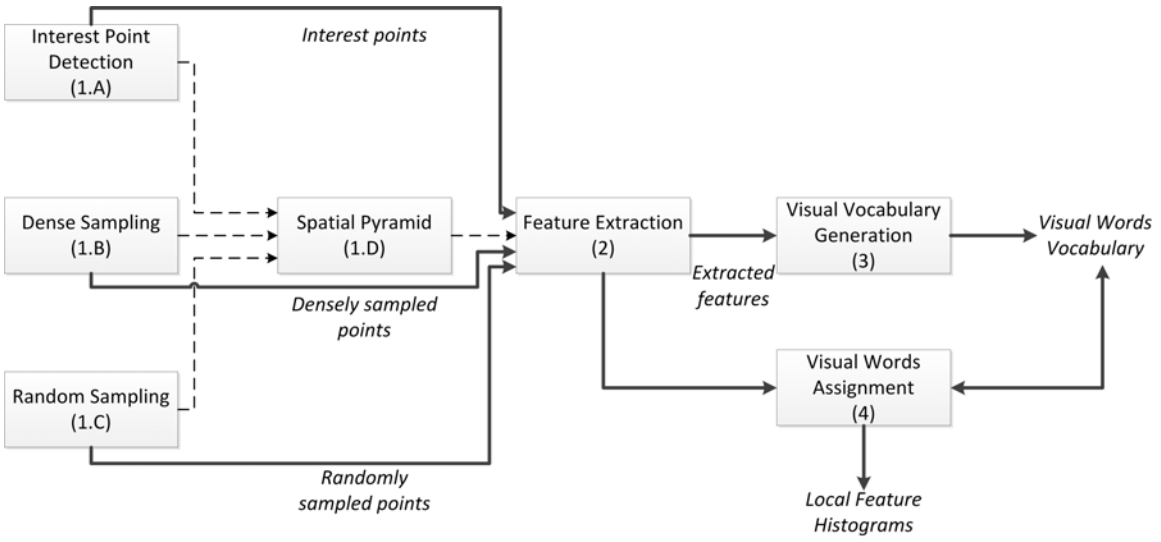

Figure 2.18: Flow chart, representing the processing order of the Bag of Visual Words pipeline.

stage	**techniques**	**characteristics**
Interest Point Detection (1.A)	FAST	examines pixel intensities around a putative interest point; very fast method
	Fast Hessian	uses box filters, in order to detect interest points over a predefined set of scales; differently sized box filters can be applied in parallel; faster than Difference of Gaussian used by SIFT, but not as fast as Features From Accelerated Segment Test
	Difference of Gaussian	Difference of Gaussian images are processed, in order to find interest points; the slowest of the three described methods
Dense Sampling (1.B)	an image is sampled on a regular grid; a fixed pixel interval is used; useful strategy to get a lot of points, especially for those images having many homogenous regions	
Random Sampling (1.C)	random samples are computed and used as interest points; useful for homogenous regions	
Spatial Pyramid (1.D)	an image is increasingly subdivided into regions; visual words frequency histograms are computed for every region; spatial information is retained; the authors of [52] report a 5% - 10% performance increase by using a spatial pyramid	
Feature Extraction (2)	SURF	uses Haar wavelet responses, in order to measure the intensity change around an interest point; variants of SURF exist such as DURF (dense sampling in combination with SURF), SURF-36 (36 dimensions for the descriptor, instead of 64), SURF-128 (128 dimenstions)

continued on next page

stage	**techniques**	**characteristics**
	SIFT	oriented gradient responses are computed, which is more expensive in terms of speed than the SURF approach; variants of SIFT exist such as DIFT (dense sampling in combination with SIFT), PCA-SIFT(dimensionality reduction, applied on the 128 SIFT descriptor), RGB-SIFT(computation of SIFT descriptors on RGB color channels)
Vocabulary Generation (3)	k-means	fast clustering algorithm, which can be executed in linear time; standard procedure for computing visual vocabularies; often terminates at a local optimum; the disadvantage is that cluster computation is affected by noisy outliers; k-means is sensitive to the initial points, used to start the clustering process; the number of clusters must be determined before the execution of the algorithm; linear assignment time for visual word assignment, by using a nearest neighbor search

continued on next page

stage	techniques	characteristics
	Random Forests (clustering)	is a collection of binary decision trees, which are build by using randomly taken local descriptors from the training set; the trees represent the spatial partitioning of the descriptors, where descriptors are clustered according to their spatial appearance; leave nodes (the visual words) define the partitioning (grouping); during visual word assignment local descriptors of a query image are assigned to the nearest visual word for each tree respectively; the final assignment to a visual word is conducted by applying majority voting of all trees (the visual word with the most votes is selected); see ([73] and [57]) for further information; logarithmic assignment time for visual words assignment, due to the tree structure; Mossmann [73] report improvements over k-means in terms of retrieval performance; number of clusters need not be specified a priori

continued on next page

stage	techniques	characteristics
Visual Words Assignment (4)	Hard Assignment	assigns a local descriptor to only one visual word; information is lost, because a descriptor can belong to more than one visual word
	Soft Assignment	local descriptors can be assigned to more than one visual word; increases retrieval accuracy; slower than hard assignment

Table 2.2: Description of Bag of Visual Words techniques.

Various Bag of Visual Words techniques have been suggested in literature, reporting among other things that soft assignment is superior to hard assignment. However, fuzziness in actual codebook creation process and fuzzy assignment to visual words have not been considered yet.

2.2 User Intentions

The actual section comprises a literature review of user search intentions in information and visual information retrieval. Frameworks for search intents as well as various approaches to infer a user's search intent and respond to it with clever retrieval techniques are addressed.

2.2.1 Information Retrieval

The correct interpretation of user intentions is a challenging topic in information retrieval. Users long for meaningful result sets when expressing an information need by submitting a query. Retrieval systems are supported by high level semantic models and low level mechanisms to infer a specific user intent (a goal), based on a certain usage behavior.

Conceptual frameworks for user intentions in information retrieval

In [109] Hong investigates the relation between information seeking strategies and user goals. He states that further insights in retrieval strategies can be achieved by understanding what users actually intend to achieve during their search sessions. The effectiveness of information retrieval systems can be enhanced by incorporating user goals into the retrieval process.

Broder [8] adopts a categorization of user intent, where he identifies three different classes representing search behavior. This trichotomy distinguishes between informational intent to retrieve specific information, a navigational intent to reach a webpage over a particular URL and a transactional intent in order to reach sites where a user pursues a transaction (action) like downloading a file. This manual classification of search intent is not restricted to text documents, but can be used and adapted to image retrieval.

	Informational	Navigational	Transactional
Queries	• wide: Vienna, cars • narrow: Latent Semantic Analysis, Gaussian Bell Curve	• Infineon technologies • National transportation system • Compaq	• Harry Potter 7 DVD • World of Warcraft game (to play the online game)

This taxonomy can be applied in generic search systems, where detailed information such as the price of certain products or the picture qualities are not available to retrieval systems. Although Broder's taxonomy provides a good starting point to

model user intentions, the presented classes are somewhat to restrictive to allow for a smooth classification. Let's consider an example: The query 'Infineon Technologies' can be a navigational query, what causes the retrieval system to direct the user to the enterprise's homepage. On the contrary the query can also possess an informational character, because the user want's to retrieve all available information about the company, which is spread over different sides. Hence a fuzzy system, considering the overlaps of certain classes, would probably lead to a better solution.

In [84] Levinson and Rose leverage Broder's taxonomy to create their own goal classification framework, which refines the already known intent classes. This refinement leads to sub categories where different queries from the Alta Vista query log are manually classified. The authors talk about 3 main steps, which need to be pursued when trying to carry user oriented search to the next level:

1. building a conceptual framework for user goals
2. relate search queries to user goals
3. adapt searching to the particular user goal

They focus on the first and the initial part of the second step and give insights how searching can be improved. The outcome of their study show that more than 80% of user goals can be classified to informational and resource seeking goals. The manual classification is a predecessor for automatic classification with intelligent supervised learning algorithms like support vector machines, Bayesian Networks, etc. and is based on the search behavior, which is comprised by a search query, results returned by the search engine, clicked results and further searches. In either way, wether the classification is conducted manually or automatically, it is a tedious work. Even if it's done automatically, training samples must be labeled by hand in the first place, before performing automatic guesses.

Inferring user intentions by means of query analysis in information retrieval

Inferring the user intent from queries is manually done in both ways and therefore a laborious work. An automatic algorithm for user intent classification is presented in ([38], [39]) and applied to 400 queries of a Dogpile transaction log [3]. The classification rate accounts for 74%. The classification model itself covers Broder's approach and is slightly different to Levinson's and Rose's refinement into sub categories. First of all the authors try to find certain characteristics of various manually observed queries for the different intention classes. Based on their findings they implemented an automatic classifier to quantitatively verify the accuracy of their approach, which features the partitioning of search queries to distinctive intention classes.

Another interesting approach to categorize user intent is described in [31], where the authors apply a two stage classification. They preprocess user queries, by grouping them either into navigational, informational or resource-seeking snippets by using hint verbs, acquired through a supervised learning method, URL information and title information. Based on this categorization, latent user goals are uncovered with the help of three different models, which are applied independently to the corresponding category and therefore only to relevant result snippets.

In [23] Downey et al. try to discover how search queries correlate with user intents and found out that user behavior for rare and common queries differ significantly from each other. In order to benefit from semantic similar queries, clustering seems to be an indispensable method.

Query recommendation is supported in [5] by a clustering process, utilizing a k-means algorithm, which tries to propose related queries to a submitted query during the search process. In addition related queries are ordered by relevance to provide a user with the best clustering results. The relevance criterion is inspired by the tf * idf weighting scheme from information retrieval. Yates et al. adapt this weighting scheme slightly by changing the inverse document frequency with a popularity measure of

[3] http://www.Dogpile.com

clicked URLs. The weighting of a query term increases if the quantity of the clicks for the particular URL gets higher. Query terms are not only ranked by the term frequency in the particular URL, but also by the URLs popularity. By doing so different query terms of various search sessions, which seems to be unrelated in the first place can be ranked similar, because the same URLs were clicked in the past. Yates approach could also be adapted to visual information retrieval by enriching visual words with the popularity of images.

In [4] the authors identify user intentions in an automatic way by analyzing query logs and applying supervised and unsupervised learning. They distinguish between three types of goals:

- informational, with the aim to obtain information on the Web
- not informational, focusing on target and resource seeking search (download, buy, etc.)
- ambiguous, not able to tell the goal solely from the queries

Support vector machines are used to associate the queries, expressing the goals, to one of the classes of the defined taxonomy. Additionally, probabilistic latent semantic analysis is conducted to find relationships of queries between different topics like business, society, computers, education, health, etc.

Speaking of intent, people have to consider that user queries, which represent an articulation of a goal, lead to good or bad results. This can be explained with the occurrence of queries, which are short and/or shallow. In order to decrease the cognitive gap between a query and the user's goal, Strohmaier and Prettenhofer explain in [97] that queries result in implicit or explicit user goal. Implicit user goals are vague and precursors of bad result sets and therefore have to be changed to explicit ones. This goal refinement process yields a more accurate understanding of user intent and can help search engines to perform more effectively. For reasons of simplicity the authors trained a Bayesian classifier to partition the queries in implicit and explicit ones, although it must be mentioned that a finer grained spectrum of

explicitness exists. The authors define a continuous spectrum of different degrees of intentional queries, denoting the user's traceable search behavior as an intentional artefact. Intentional artefacts are used to distinguish between different degrees of explicitness, which can be utilized in web search to support the user in finding the information s/he actually wants depending on the particular information need. The authors make this distinction by analyzing query logs cumbersome in the execution but apparently fruitful for the detection of different degrees of explicitness.

In [96] a prototypical, parametric algorithm is specified to support query suggestion and to clarify a user's intent by making it more explicit. Leaving the implicitness behind, document precision is increased because the searcher is guided by reformulated (explicit) queries, which express the user intent in a more detailed way. This decreases click through and as a result saves time. As mentioned before implicit queries are sometimes too short in order to infer a specific user intention. The authors of [96] use a bipartite graph to establish a relationship between implicit and explicit queries (in seven out of ten cases the suggested explicit queries are correct). While common query refinement or suggestion techniques seek to provide relevant documents based on semantic relations with Latent Semantic Analysis for instance, where words belong the the same concept if they statistically appear together in a bunch of documents, Strohmair et al. try to find the intention behind submitted queries via graph analysis.

Inferring user intentions by means of search behavior analysis in information retrieval

Beside proposing user queries and making them more explicit user intentions can be determined by tracking past user click behavior and anchor-link distribution as suggested in [53]. The authors use the aforementioned methods and several features like the mean value, median, kurtosis and skewness to predict a navigational or informational intent automatically and state that a combination of the features lead to better results. In order to facilitate their study they skipped the transactional class

and focused on a dichotomy, which is still challenging because you cannot guarantee a correct classification due to insufficient user-click data for instance. Nevertheless past user-click behavior seems to be a good feature in predicting the user's intention. By utilizing this feature to predict user intentions the authors reach a prediction accuracy of 80%. They present interesting plots showing the click distribution for informational and navigational queries of all submitted search queries submitted in the past. Clicks on different URLs in the result list for a submitted query denote an informational goal, because a user tries to get as much information as possible. Beside this even distribution of user clicks, navigational queries often lead to a "one click answer" meaning that a URL, which links to the desired web page, is often clicked, while neglecting the other URLs. Past user click behavior is also leveraged in [5].

In [56] Li et al. present an automatic query classification which is achieved by click graphs. The bipartite graph consists of nodes representing queries and nodes representing URLs which will be interconnected if they relate to each other. A relationship denotes a possible user goal when issuing a query. Due to the fact that click graphs are noisy and sparse a correct automatic association cannot always be procured. In order to diminish erroneous edges, the authors use a regularization approach on a click graph with the support of content-based classification and pack the two methods into an unified framework.

User intentions are addressed in [58] in order to provide a user during a search session with appropriate documents. For this purpose user profiles are learning by looking at the user's search history, which comprises the categories like Cooking, Soccer, etc. and the associated weighted query terms. The weighting reflects the importance of a query term within a particular category for a user. The search process itself comprises two steps:

1. The user submits a query, which in turn delivers a result list of appropriate categories

2. Relevant categories are chosen either manually by the user or automatically by the retrieval system. The selected categories are used as a context to enhance

retrieval effectiveness.

2.2.2 Visual Information Retrieval

All the work presented so far hardly covers the topic of visual information retrieval under the consideration of user intents. This work is partially addressed in [25] where Fan et al. developed a hyperbolic visualization to provide the users with a user friendly view of clustered images, which are tagged. Images with similar Flickr tags belong to the same cluster. A Cluster is semantically interlinked with other clusters and can be explored by zooming into it. Such a kind of visualization approach, which benefits from user interaction, helps to express user goals in an efficient manner. But the explicit user intentions are not addressed in the work of Fan et al.

Conceptual frameworks for user intentions in visual information retrieval

User intentions in image search are addressed in [45], where the authors correlate the search goals of a users with their search behavior in multimedia retrieval. They pose the question, wether and how users search and browsing behaviors fit different intention classes, adopted from information retrieval. They conducted an exploratory study, confronting participants with various image retrieval tasks, which were properly chosen for the different intention classes. The chosen taxonomies arise from already proposed classification systems introduced in [84] and [39] for information retrieval. The specified image retrieval tasks were conducted with Flickr[4], a web based photo sharing and retrieval system. The findings of their study show that the classification schemes do not apply easily to retrieval tasks, related to image search. They further mention that the user's search behavior, while browsing the result lists, might correlate to the particular intention classes. Features like average number of viewed images and average duration time, needed for solving a tasks, were leveraged to model and describe the user's behavior. Direct or informational tasks carrying an advice character tend to be rather short, with few clicks within result lists, whereas

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

undirected information intention leads to the longest duration time with the highest amount of viewed images. The results must be considered with caution, due to a small number of tasks investigated in their study. Besides, the number of the various tasks is not evenly distributed.

Text based taxonomies for the classification of user intentions must be adapted and extended to fit the requirements needed for representing user intent classification in image retrieval. Lux et al. [62] extend Broder's taxonomy by adding a mental image class, meaning that a user has the visual representation of the desired images in mind. Query by Example or sketches depicting the visual appearance can be applied as the query paradigm, in order to retrieve relevant image content. The transactional class, which is adopted from Broder's taxonomy, contains user goals focused on finding images for further use, e.g for illustration in a brochure. Image semantics as well as visual analysis can be used by an image retrieval system to provide the user with appropriate pictures. The last two classes (Knowledge Orientation and the Navigational) heavily depend on a textual and conceptual search paradigm and can be traced back to taxonomies and search solutions presented in information retrieval so far. The search paradigm always depends on the task, which need to be fulfilled, the knowledge of the user, concerning the search topic and the retrieval system's capabilities to offer various search strategies. The various classes overlap as depicted in Fig. 2.19, what speaks for a fuzzy model of user intentions.

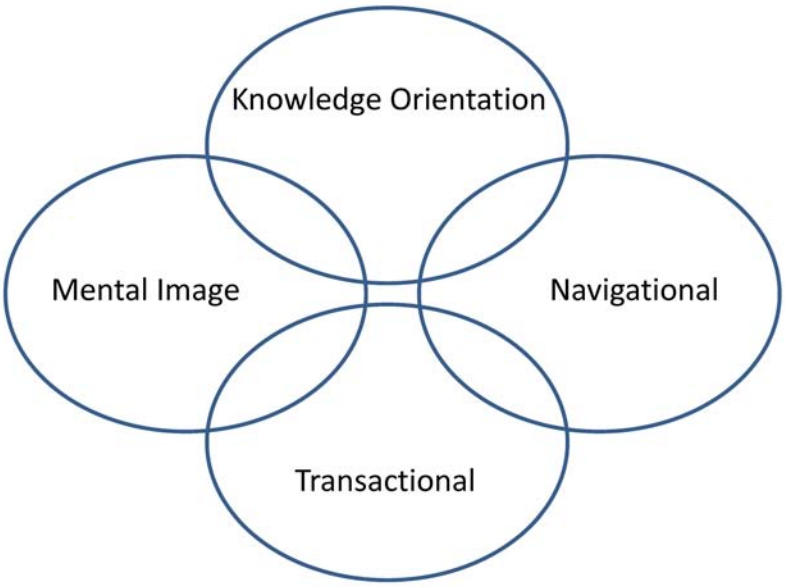

Figure 2.19: User Intention taxonomy in image retrieval adopted from [62].

Armitage and Enser [3] conducted an analysis of user queries for image archives of still and moving images, whereby they categorized user requests addressed to seven picture libraries. Their general model contains four categories:

	Search by artist	Known item search	unique subjects	non-unique subjects
Description	the name of the artist	a painting or image someone is aware of	particular objects, locations, named individuals	queries describing a broader spectrum of images and paintings
Examples	Van Gogh	Whristler's Flying Rocket	Adam and Eve, Alan Turing	Japanese people, women jockeys

Table 2.3: Categorization of user requests by Armitage and Enser.

The authors state that their findings can be used in order to design good user interfaces for visual information retrieval systems.

In [20] Datta et al. address that image retrieval systems have to consider a human-centered perspective and classify a user intent in three main classes with respect to clarity of search: the browser, the surfer and the searcher. Based on the classification they presume that different representations will be helpful for the user and remark that this has not been studied extensively. The proposed intents serve as a basis for the classes of user intents during search, presented in chapter 4.

	Browser	Surfer	Searcher
Description	A typical browser is a user pursuing no clear goal, who aims at retrieving images from diversified topics. As a result uncorrelated searches are performed, to entertain the user in a certain way, where a pattern to recognize search behavior can be hardly recognized.	Is a user with a moderate clarity of the end goal, who follows a search behavior with an exploratory character in the first place. After some search iterations the user's idea of what s/he actually needs gets clearer, what should be supported by the retrieval systems through valuable hints.	The searcher is a user with a clear end goal, with high clarity about what s/he is searching for. A session should be rather short by retrieving accurate results.
Nice to have	A browser values the suggestions of random search hints, providing her with interesting information like most popular pictures. Such suggestions will increase the interaction of the browser with the retrieval systems, while keeping a high value of entertainment.	A user looking for holiday destinations, would be happy about recommendations, guiding him to most popular destinations, which speak for a pleasant holiday.	On the contrary a searcher looking for a picture of the Eiffel Tower would need an accurate representation of the desired images.

Table 2.4: Taxonomy representing the clarity of intent during search.

Inferring user intentions by means of relevance feedback in visual information retrieval

An interesting approach, to display relevant images, is referenced in [92] and described comprehensively in [16], where a content-based image retrieval system named PicHunter is introduced. An image search is conducted with a certain amount of iterations, while analyzing a user's relevance feedback until the process ends with the presentation of the desired target image. During the search process images are associated with probabilities of being the possible target image, which are stored in a vector. Although the authors mention three classes to express user intentions: target search, category search, where both has an informational intent and open-ended search, where a user follows a navigational intent, they strongly focus on the target search in order to pursue visual information retrieval and neglect the other two classes.

The main search paradigm on the Web is text based, due to its intuitive way of expressing an information need, which is conveyed to the search and retrieval system. Visual information can be leveraged, in order to complement text-based searches.

The authors of [17] propose a visual search system, which operates on top of Microsoft Live Image Search. Their system re-ranks results from text retrieval with respect to user intentions and chosen query images. A user operating the system can choose among various result images, mark them as relevant and submit the relevant images to the retrieval system. The submitted query images are visually analyzed and automatically assigned to a category, describing portraits, people, scenes and so forth. The assignment is conducted by using an off-line trained decision tree. Once a category has been found a previously determined linear combination of all features is used to re-rank the former mentioned result list. As a matter of fact portrait images lead to a higher weighting of facial features, whereas pictures containing wonderful landscapes do not need facial features at all. The relevant images are presented by choice in either a grid view or a rank collage view, which organizes the relevant images around the query images, presenting them larger and nearer to the queries than less

relevant images.

In [19] Cui et al. present the aforementioned intention based search re-ranking approach more deeply. They look at intentions in terms of categories like portraits, scenery, objects, etc., arguing that user intention are interwoven with categories, because the chosen linear combination of visual features for similarity searches reflect user intentions, when submitting a query image. Hence they name the investigated categories intentions, what must be taken cautiously, because intentions are abstract and hard to measure, while categories describe semantically similar images. Nevertheless their presented approach to apply different linear combination of features is interesting. They mention the features and when they are used when issuing a query image, retrieved from the initial text based search. The authors use a test data collection of 451,352 images with 483 keywords retrieved from Google and Microsoft Image Search and compared similarity measures for single visual features with their proposed feature fusion approach. The linear combination of different features performs better than the single feature approach, when re-ranking the images, which is not that surprising, because an optimal feature weighting scheme in terms of a linear combination is predefined for every category.

User intention modeling is further driven in [18] where the authors present three interaction methods, which support the users in finding the image content they actually want. They propose Smart Intention Lists, already introduced in [19] and [17], with five different categories (scene, object, people, portrait, general) designed as intentions. The authors state that their search system uses the complete image index and smartly adapt to the conveyed intention, instead of searching within a sub-category only.

The second interaction method offers region selection within images. The user can draw a line around the interesting region of a picture in order to tell the system, which subpart should be emphasized for search. The third proposed method encompasses a relevance feedback mechanisms by dragging relevant images to an "I like" image basket, which seems to be a user friendly and intuitive way to inform the retrieval

system about positive examples. The authors define search tasks and compare different retrieval systems like Picasa and QBIC in terms of percentage of task completion, elapsed time to fulfill a task and qualitative measurements like user satisfaction or efficiency of the retrieval systems with their image retrieval system. Although it is interesting to see that their proposed retrieval system performs much better than common picture administration applications like Picasa, it must be mentioned that Picasa heavily focuses on image organization by time. This is adequate for personal image collections, when users want to memorize certain events in life, but not for images on the Web, not related to the user in any kind.

Tang et al. [108] further extend the approach from Cui described in [17] and [19], to improve the precision of retrieval results of images during an online search session. The authors try to capture the user's intention with just one click by taking both textual and visual information of user queries into account as depicted in Fig. 2.20. After retrieving the images by means of a text query (step (a) in Fig. 2.20), the user clicks on an appropriate image, what initiates a bunch of steps to expand the retrieval list and re-order the results automatically.

The image, clicked by the user, is automatically classified with a trained decision tree in one of the predefined categories such as scene, portrait, etc. Depending on the assigned category a similarity measure, leveraging various features and weights for each category, is applied to to re-order the images from the initial retrieved image pool (step (b) and (c) in Fig. 2.20). Images visually appearing more closely to the query image are ranked higher. The re-ordering depends on the used features and the adaptive weights, which are different from category to category. So for instance facial features are more relevant for a query image containing people, what emphasizes the weight for such features during the similarity search. Once the images are re-ranked, query expansion is conducted (step (d) in Fig. 2.20), by grabbing relevant keywords from the surrounding text, filenames, etc. of the k most similar images to the query image. The tf-idf method is then used to rank the words in descending order, only keeping a small set of candidate words to reduce the computational cost.

To get more reliable keywords for query expansion all the images containing the previously computed candidate words are clustered in terms of their visual similarity (step (e) in Fig. 2.20). Clusters therefore contain those images having the same candidate words and are similar concerning their visual content. The k-means clustering algorithm is used to compute different clusters, where the cluster size depends on the amount of images to cluster. After the computation of the clusters the difference between the query image and the images within a cluster is computed. The keyword of the cluster exhibiting the highest visual similarity to the query images is selected for keyword expansion. The initial image pool is enlarged with images containing the expanded keyword (step (f) in Fig. 2.20) and re-ranked (step (g) in Fig. 2.20) according to learned query-specific and textual similarity metrics.

Their approach seems to be promising, outperforming pure text based search approaches of visual content. Images containing the same keywords can be different in their visual appearance. On the contrary, images appearing closely together in the feature space sometimes exhibit a semantic difference. A combination of textual and visual approaches is to prefer, when searching for images in large images collections. Hence additional efforts must be brought up, which comprises not only effective text retrieval methods but also content-based image retrieval techniques to exploit the visual information within the images.

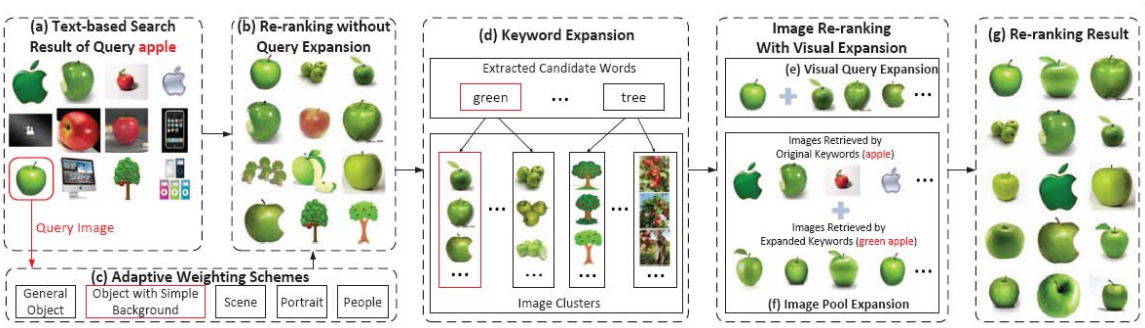

Figure 2.20: Search process in the image retrieval system proposed by Tang et al. (Figure adopted from [108])

In [34] Hua and Tian give insights on how text based image search systems can be improved by applying visual retrieval mechanisms on top of them. They address visual filters in order to restrict the result list when searching for images with specific keywords, such that only a fraction of the possible candidate pictures is returned. Users searching for Paris, but only interested in pictures of the celebrity Paris Hilton, can apply such filters, which in this case rely on face detection algorithms. They further mention visual based re-ranking of pictures in result lists, where images primarily retrieved through a text search paradigm, are presented nearby a chosen image from the result list, according to their visual similarity. Such a re-ranking could be beneficial to enhance user satisfaction during search and could better fulfill search goals of users to fit their intention.

When searching for images the recognition of an object's appearance, it's shape and color and the scene where it is embedded can be very helpful in retrieving the desired visual information. Many people possess the capability of keeping visual information in mind, which can be used during the retrieval process. Users draw sketches or provide a query image as a starting point for their search to retrieve visual similar ones. Especially art historians as described in [32] have the ability to remember objects with a certain shape and color, hence providing the retrieval systems with valuable illustrations.

Users, trying to receive art images, are more likely to convey their information need through a visual search paradigm, like: 'give me those images containing azure oval objects'. However Chen [32] states that users find it difficult to express their information need, either in text or visual mode. Given a search topic users struggle in expressing key words or visual illustrations, which are suitable for their search task. The specificity of their search query heavily depends on their knowledge and especially on their intent, the goal they try to achieve.

Inferring user intentions by means of search behavior analysis in visual information retrieval

User intentions or goals are expressed through the users' interaction with the retrieval systems. Submitted text queries, queries by example, the click history or in general the search behavior of a user are leveraged to infer the goal of a user during her/his search session. The analyzed search behavior is classified with a predefined model describing user intentions and retrieval mechanisms fitting the respective classes are applied, in order to retrieve appropriate images.

Chen et al. [15] talk about action intentions (basic user actions such as mouse clicks, keyboard typing) and semantic intentions (e.g. "I want to bay a DVD from Amazon"). Semantic intentions, the high level user goals, are very hard to grasp, but can be approached by observing the user's interaction with the computing system. The user's interactions like user clicks or keyboard typing for instance represent the user's search behavior and are used to infer the overall goal of a search session. A search session comprises subgoals, expressed by various user actions. Those user actions or action intentions are investigated in this paper. The authors trained a Naive Bayes classifier based on a user search log, which consisted of various actions and used the trained classifier to infer the next probabilistic user action. A user performing several interaction steps should be supported by a software agent capable of predicting the next action step by highlighting important hyperlinks for instance. This could relieve the user from the burden to navigate over undesired links.

Mueller et al. [[76], [75]] use log files for the analysis of user search behavior and try to weight features, used for finding similar pictures, depending on the user interaction activity with the retrieval system. User feedback is used to boost the weighting of color and texture features, from images marked as relevant during a search session. This kind of positive relevance feedback should help to enhance the relevance and precision of image result sets.

Medical image retrieval is becoming a popular subfield of image retrieval, where users operate and conduct searches in a narrow domain. Wang et al. [105] state that

few studies have been conducted so far on the behavior of users search tactics within this rising field. They investigated search behaviors of experts and novices on an image database containing images related to radiology. Their findings are very interesting and show that users' search moves distinguish between the two investigated groups and depend on the amount of domain knowledge. The authors defined 6 different search tasks and asked the users to fulfill them, while tracking the search moves applied by the users, in order to achieve their goals. 29 participants were asked to search and browse for images. One interesting point in their findings is that experts tend to browse more deeply within result lists than novice users and explore more screens of returned images, while users with less experience stop searching for images, when the retrieval of relevant images lasts too long. The users' willingness to browse should be supported by an intelligent retrieval system to return those images, which are adequate for users and their particular goals.
Besides experts used more query terms for their searches to express the specificity of their information need and refine them more often. Novice users instead rely more on the retrieval system's recommendation, what paves the way for reliable relevance functions and different indexes, tailored for particular user needs.

Visual information retrieval techniques are hardly used in health care systems. Either search queries rely on patient names propagated using the DICOM[5] standard or on text meta words describing a certain pathology. In [74] the authors investigate the usage of images and search behaviors of health care professionals. They point out that visual information of medical images has experienced little attention so far and need to be included into image search systems, hence taking a multi modal approach of image search into account. Additional visual search can complement text search, in order to achieve higher recall rates. In their paper they interviewed different groups of people, related to health care systems, who need medical images for teaching, research, presentation purpose, for finding similar cases, etc. The sources they tap, in order to get such information, are the internet (mainly using Google's image search), personal

[5]http://medical.nema.org/

image collections and medical systems, which store a rapidly increasing amount of pictures. The used search paradigms are:

- search by text
- hierarchical search within pathologies
- simple search by patient names using certain picture archives specific for medical image retrieval

Visual search is desired but hardly used, due to the lack of completeness of search systems, which do not consider visual information as an additional information source. Hence future medical information systems should incorporate visual search strategies, in order to answer queries such as: 'get me similar medical cases based on hand drawn sketches, query by example pictures, or regions of interest within an image', satisfyingly. Visual search can be used to refine text based search, by re-organizing and clustering visual similar images in the result list. This can further enhance user satisfaction, because relevant images, describing medical cases, are recommended to the user. Especially novices, who find it difficult to express their information need for medical images, rely on a good recommendation on behalf of the retrieval system.

Digital photo retrieval in the field of journalism can be a tedious process. Journalists searching for illustrations to be incorporated into their future articles, face the obstacle to search and browse within large newspaper photo archives, which are sometimes poorly indexed, in the sense of caption, in order to retrieve proper illustrations. Markkula et al. [65] investigated the user needs of journalists, when searching for illustration within newspaper archives. Their findings reveal that even journalists find it difficult to express their information need properly. The typed textual queries to retrieve relevant images often consist of only one word. As a consequence the result set of images is very large, what makes it cumbersome to browse. The authors state that the observed interaction behavior of the journalists with the system mainly relies on browsing.

This could be supported by clever indexing mechanisms either based on textual information or on visual information. Visual information to offer content-based retrieval can be used in addition to textual methods. Especially in the browsing case visual similar images can be searched and grouped together.
The definition of visual similarity in the sense of similar color, similar objects and so forth within images can differ, depending on the particular information need and goal a user tries to achieve. A journalist could search for adequate portrait pictures of a certain politician, what would cause the retrieval system to use face detection descriptors and the upper part of a person to be in the illustration. Another journalists could be interested in group photos covering a council of politicians, what would stress the need for people counting and the illustration of the whole body. A further example would be the need of retrieving all kind of pictures of Paris Hilton, whereas the need to retrieve pictures of Paris Hilton in black dresses would be more specific. Certainly it must be admitted that the search method also depends on the task a user is working on. If the journalist needs pictures of a certain event it is rather easy to browse the picture gallery by only considering those pictures with the particular time stamp and location. Visual retrieval methods can complement such a search, only providing those pictures, which are visually relevant for a certain event.

McDonald and Tait [68] conducted a user study with twenty participants, to find out which search strategies (search by sketch, browsing) are used for particular search tasks such as:

- find images visually similar to an already seen image
- find images containing particular objects, yet unknown to the user
- find images conveying an abstract meaning, like innocence.

The outcomes show that searching by sketch is sometimes tedious, because users are not quite familiar with this kind of search strategy. Apart from that browsing seem to be a desirable and intuitive search method to retrieve relevant images.

The investigation of the search behavior, when searching for images on the web or various image databases, is an indispensable aspect, in order to improve current search systems. Menard [69] investigate users' search behavior on museums objects by conducting qualitative user studies. The interviews, carried out with thirty participants, give insights concerning the search tools used, the preferences of users and interaction behavior of users during search and fruitful suggestions, how image search could be improved by taking more search modalities into account. The preferred search paradigm is text based, because it's intuitive. But additional visual searches and searches within result lists, what speaks for browsing, are suggested by the users for the various search engines. These alternative search modalities are not yet common and must be adopted to gain higher attention from the users.

In [82] Rodden et al. conducted a study on the importance of grouping visual similar images, in order to assist browsing. They figured out that grouping by visual similarity improves the quality of image result list, regarding the convenience for users to find images they actually want. Although visual similar images grouped together seem to be less eye-catching, users prefer the visually organized image result lists to randomly organized result lists, which cover more visual diversified results.

The usage of visual information to support browsing and retrieval is also addressed in [83], where Rodden and Wood investigate the search and browsing behavior of users in private photo collections. The primary approach to find digital photographs, taken at a certain event, location, etc. is based on time information, which is conveyed with any photograph recorded by a digital camera. People tend to look at pictures of recent events in chronological order presented as thumbnail images. Text queries are rarely used due to the lack of good annotated images. Usually people are too lazy to tag their images, because it's cumbersome process. Hence pure text queries will lead to a bad recall, neglecting those images, which were not annotated before. Additional visual searches can help by serving as a complementary retrieval technique, to get more images, visually similar to those already retrieved by a text search. Nevertheless browsing seem to be an appreciated method to find photographs in private image

collections. The consideration of visual information retrieval techniques can positively support browsing in finding desired images more quickly.

Chapter 3

Content-based techniques for Visual Information Retrieval

Recent years have witnessed an increasing importance of visual information retrieval, due to a large amount of produced images and videos. Text based search is not always sufficient, because meta information is sparsely spread over the digital content. Hence low level features, incorporating color, texture and edge information are leveraged to add additional capabilities for analyzing, organizing and searching visual content. In this chapter various techniques for video and image analysis, organization and retrieval are presented. Beside global features local features, employing a Bag of Visual Words approach, are investigated and applied to various tasks.

Starting from video retrieval, by using motion information in combination with BoVW, an effective approach for organizing videos is shown. Still image summaries, useful for browsing in video collections, are automatically generated by using low level features and clustering strategies. At the end of this chapter a thorough description of a new visual vocabulary (codebook) generation approach is presented and applied for image retrieval.

3.1 Known-Item Search in Video Retrieval by exploiting Motion Codebooks and Color Features

This section addresses a video retrieval approach, introduced in Lux et al. (cf. [64]). State of the art algorithms for text-based searches as well as visual features like CEDD (see [12] and 2.1.1 for further details) and (G)lobal (M)otion Histograms (see [64] for further details) are applied. Besides a motion based search approach, which is described in the respective subsection (see 3.1.1), is presented. This motion based approach leverages the Bag of Visual Words technique and constitutes a part of the own work, explained in this thesis.

Searching for videos in video archives can either be text-based, query by example searches (or sketches) or a combination of both. Usually text-based searches perform quite well in delivering the video content a searcher is looking for. YouTube as common streaming on demand portal provides this capability. Text search requires well annotated video content, which is sometimes not present, due to laborious manual annotation tasks. A part of videos stored in repositories can be retrieved by entering text queries with the disadvantage of not satisfying recall rates. The video we are searching for was probably not annotated before, hence leading to result sets, which do not contain the desired video. By taking visual information into account recall rates are improved, which opens the chances of delivering the wanted video in the result list. Nevertheless pure visual search can lead to unwanted result lists as well, because the visual content lacks in conveying the semantic meaning of video scenes. Therefore a combination of text and visual search should be preferred.

This multi modal approach is shown in Fig. 3.1 where a search for a video clip already seen before and contained in a collection of videos is initiated by a text query. A text-based search is conducted, in order to retrieve previously annotated video clips. Respective meta data and speech transcripts are indexed, to search for corresponding video material based on a tf-idf weighting scheme. Text searches conducted on an inverted index are performed very quickly and top results are used

to apply additional content-based searches, leveraging various visual features. Query-by-example searches are performed next for every feature and respective result lists from all searches are merged. The applied visual features and the process, how the results get merged, are explained in the following subsections (see 3.1.1 where a new approach for visual content description, comprising motion features and Bag of Visual Words, is described and 3.1.2). The idea to combine text and visual search is based on the assumption that those searches gain higher recall rates, retrieving also relevant videos not annotated before. The initial result list retrieved by a text-based search contains videos, which represent the desired search topic visually, hence constituting a good starting point for additional content-based searches.

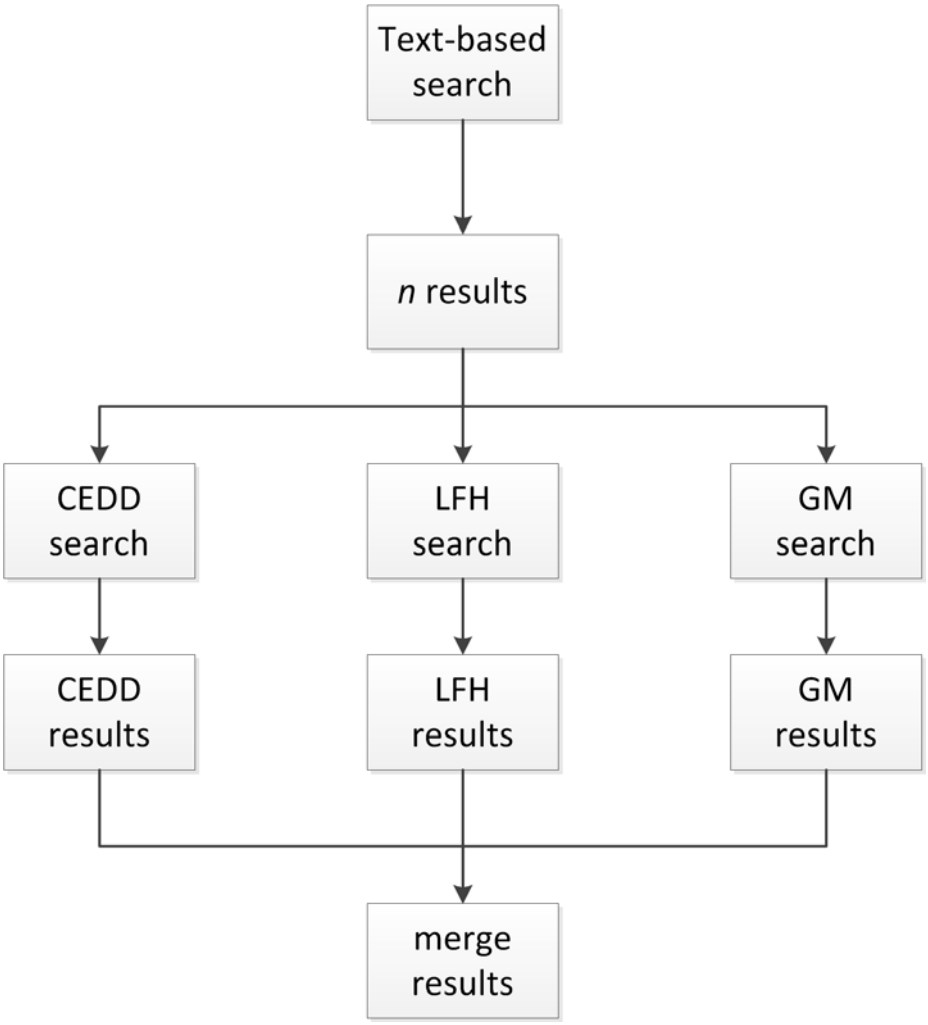

Figure 3.1: Architecture of the proposed approach adopted from [64]

3.1.1 Content-based searches

CEDD features (see [12] and 2.1.1), comprising a fuzzy color scheme and edge information, and motion features are used for content-based retrieval. Visual information concerning motion is either utilized by a **Global Motion Histogram (GM)** (see [64] for further details), which describes a whole video file by its motion, or by a so called **Local Feature Histograms (LFH)**. These local feature histograms are

inspired by the Bag of Visual Words approach. But instead of containing edge and intensity related information, motion vectors from macroblocks of H.264 encoded videos (compare [1]), representing camera motion and object motion, are leveraged.

Motion Codebooks and Local Feature Histograms (LFH)

Forward predicted motion vectors from frame $k-1$ to frame k are computed for each 16x16 pixel size macroblock. This computation is conducted over all frames within a shot in order to gather the distribution of motion over the whole shot (see A.1 and B.1 of Fig. 3.3). Regional information is considered by subdividing each frame into 16x16 blocks, where each macroblock is assigned to the corresponding block. Hence the evolution of motion related information is tracked on a block basis. A 13-bin motion histogram is calculated for every block by classifying macroblock related motion information (macroblock based motion estimation), consisting of direction and magnitude. The classification scheme introduced in Schoeffmann et al. [88] is depicted in Fig. 3.2, showing the possible directions a motion vector can be assigned to. This assignment is conducted for all macroblocks of a frame for the whole number of frames, comprising a shot.

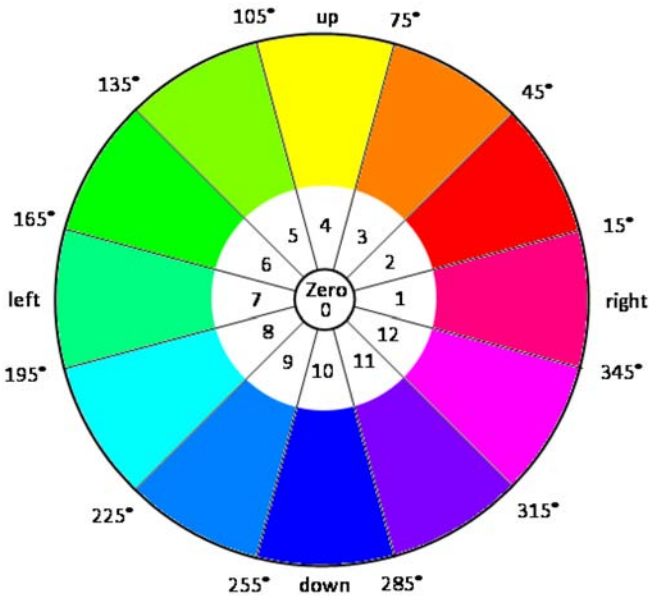

Figure 3.2: Particular directions used for the classification of motion vectors [88].

After the computation of several 13-bin motion histograms a motion codebook is created (see A.2 in the left image of Fig. 3.3). Therefore all histograms stored as vectors of all shots of the videos in the training database are k-means clustered. The cluster centers denote quantized visual information in terms of motion information and are used to compute the local feature histograms. The local feature histograms are computed for every shot (see right image of Fig. 3.3) of the test videos provided from TRECVID[1]. 13-bin motion vectors from a single shot are assigned to the most nearby cluster center. The number of the assignments for every respective cluster center is calculated and constitutes the local feature histograms for one shot, which is needed for later retrieval by performing content-based similarity searches (see B.2 in the right image of Fig. 3.3).

The computation of the motion codebook and the local features histograms is inspired

[1]http://trecvid.nist.gov/

by the Bag of Visual Words technique for visual information retrieval and constitutes a part of the own work explained in this thesis. Instead of using visual information such as patches around local interest points (see chapter 2.1.3), motion features from Schoeffmann et al. [88] are leveraged.

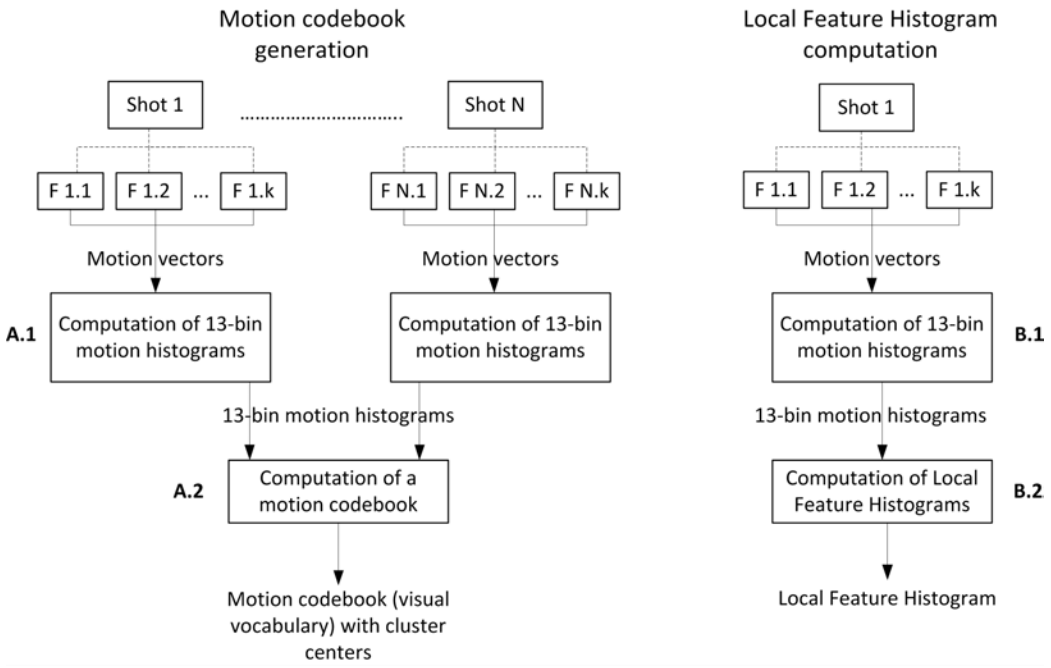

Figure 3.3: Left image: Motion Codebook creation. Right image: Local Feature Histogram computation, comprising motion information.

3.1.2 Combine results from text-based and visual searches

The first z ($z < n$)results from text search are used to perform content-based searches with the aforementioned features. In Fig. 3.4 the scheme for content-based search relying on the results of the text search is given. Starting from video 'Video-1' similarity searches with the features CEDD, LFH and GM are conducted separately and new result lists based on visual similarity are retrieved. One result list for each video and feature is returned, yielding $z * 3$ new result lists. For content-based similarity

searches histograms computed of the center shot (shot with the smallest euclidean distance to all other shots in a video) of each video are used for CEDD and LFH, because both features describe shots in the proposed approach. Searches based on global motion histograms operate on entire videos. Hence lists of shots representing the videos are returned for CEDD and LFH, whereas lists of videos are provided by GM.

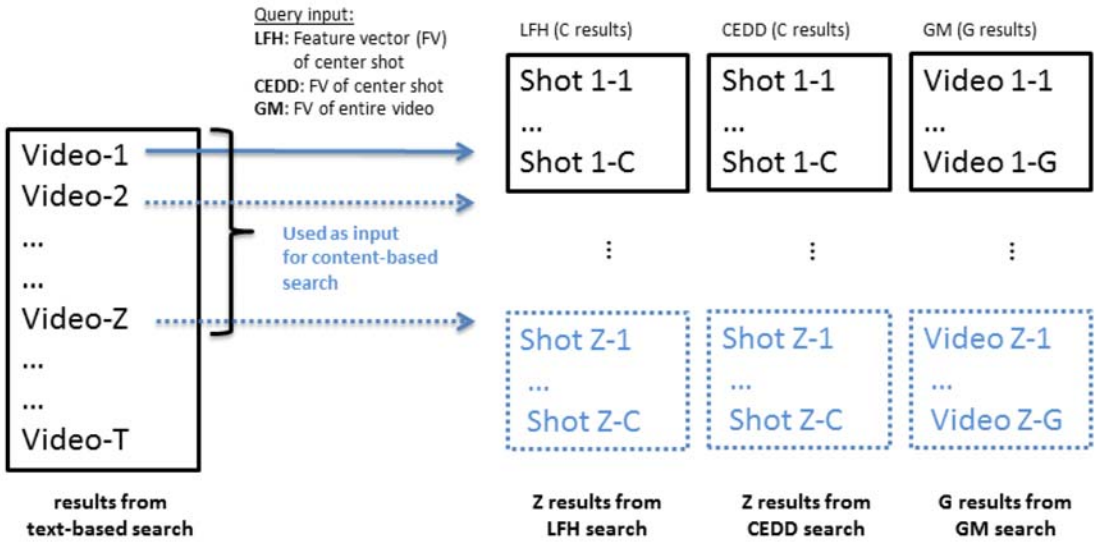

Figure 3.4: Overview of content-based search result lists starting from text-based searches (adopted from [64]).

The results of different result lists of content-based searches for one feature (CEDD, LFH or GM) are merged by applying an interlacing scheme, where the first result 'Result 1-1' constituting the first place of the first result list is placed before the first result of the second result list and so forth. The second result 'Result 1-2' of the first result list is placed after the first result of the last result list, what is depicted in figure 3.5 (stated as fusion search for simplicity from now on). Duplicates, retrieved by content-based searches, are eliminated. The reason to chose this merge strategy is that a top ranked target video, found by text search, is still found by fusion search.

Whereas a target video, which is badly annotated and which does not appear among the top ranked videos after text search, can be among the top ranked videos after fusion search.

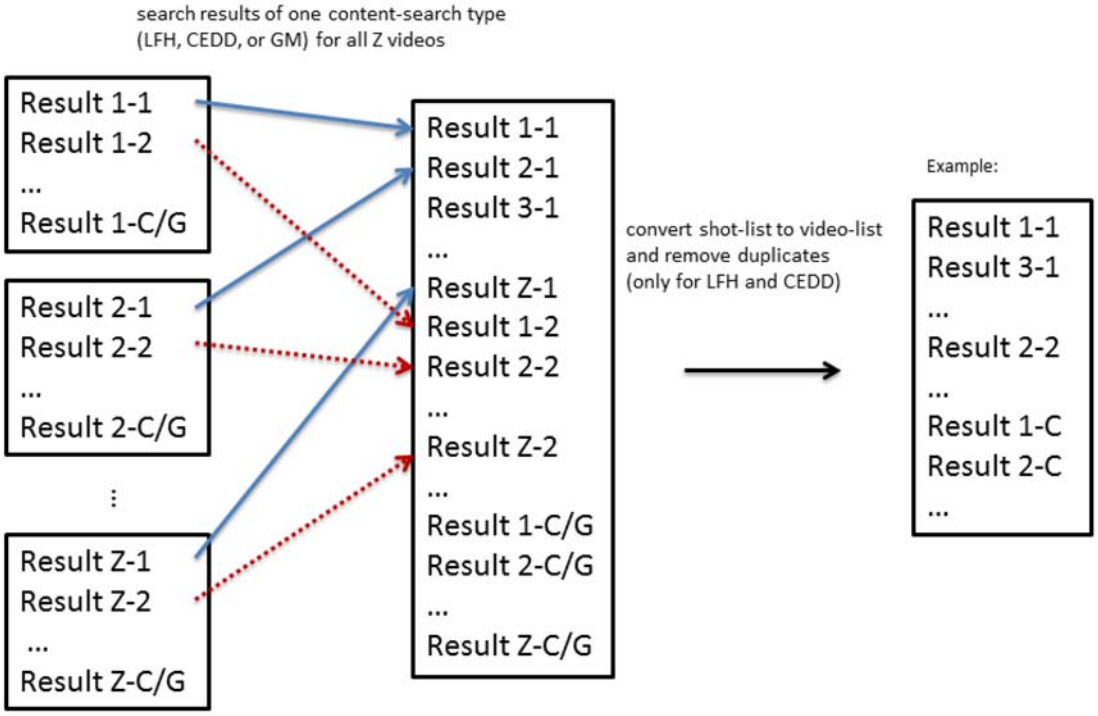

Figure 3.5: Overview of search results for one content-based search type (adopted from [64]).

After the application of content-based searches and interlacing of respective results lists, results from text search are merged with results from visual similarity searches. Therefore the averaged inverted rank[2] of videos contained in the respective result lists is computed. The result with the best rank is placed first and so forth, yielding a final result list of fused text and content-based searches (depicted in Fig. 3.6).

[2]proposed for evaluation of Known-item search 2010. see http://www-nlpir.nist.gov/projects/tv2010/tv2010.html

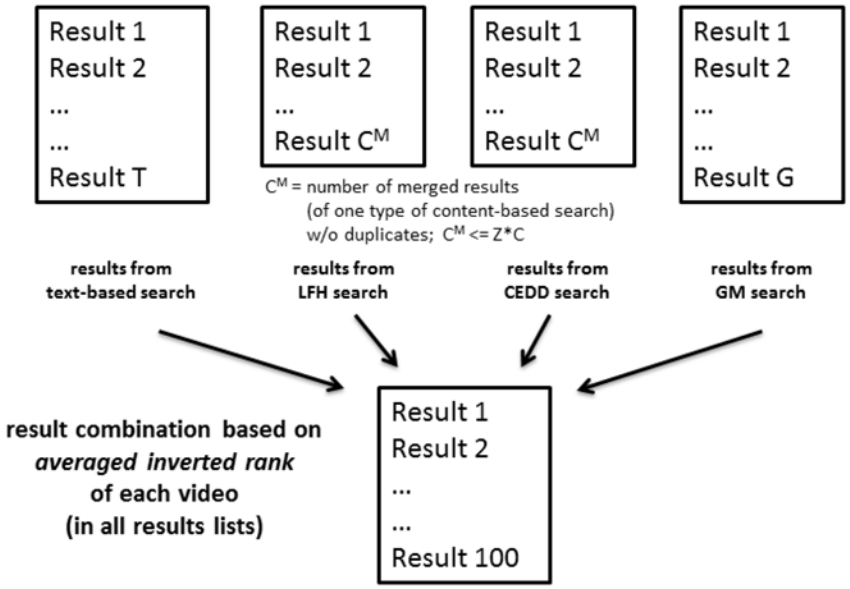

Figure 3.6: Overview of the interlacing scheme of a text-based search result list and content-based search result lists (adopted from [64]).

3.1.3 Experiments and Results

The proposed approach was applied for TRECVID's known-item search task in the year 2010[3]. This task models the behavior of a user, who is looking for a video, contained in a video collection, s/he has already seen before. The search process starts by typing a text query to retrieve the desired target video constituting a particular topic, which contains objects, a person a location and so forth.

The challenge was to automatically return a result list of 100 videos, which fit the need of a user in best manner. The test data collection (IACC.1.A) comprised more than 200 hours of video clips (10 seconds up to 3.5 minutes per clip, approximately 8000 Internet Archive H.264 videos, 50 GB). 99% of the videos featured text annotations such as title, keywords and description. Most of the annotations contained a

[3]http://www-nlpir.nist.gov/projects/tv2010/tv2010.html#kis

subject meta data field. About half of the annotations contained creator, source and color fields. Only few of the video annotations had notes or keyword fields. A total amount of 298 search topics were made available and results of search runs had to be submitted to TRECVID. The results of the automatic runs were scored by computing the mean inverted rank of found topics. A sample query to look for a video containing the desired topic can look like this: 'Find the video of a man in orange outfit throwing an apple for a black dog with red collar to retrieve and the dog retrieves but eats the apple.' Visual cues such as 'man-orange outfit, apple, black dog, red collar' should appear in the desired target video. However it is not guaranteed that query terms will overlap with meta data. Hence visual searches should be additionally conducted. More example topics with links to the respective target video are listed at http://www-nlpir.nist.gov/projects/tv2010/ki.examples.html.

Table 3.1 gives an overview of the evaluated results. The second column shows results of the text search only and the third column gives an overview of results related to combined searches (both text and content-based search with all features). The following parameter setting were used:

- **Text Search**: 100 results
- **Fusion Search**: 60 results from text search. The best 10 results were selected for content-based searches. The best 40 results of each content-based search were used for merging.

Metric	Text Search	Fusion Search
mean inverted rank	0.266	0.260
topics found	155	130
topics with rank 1	65	64
topics with rank ≤ 10	103	101
topics with rank ≤ 20	107	107
# of topics not found by text search but by fusion search	-	3
# of topics found by text search but not by fusion search	28	-
mean rank over 127 topics (found by text and fusion search)	5.73	11
rank standard deviation over 127 topics	10.27	24.93
# of topics having the same rank in text and in fusion search	91	91
# of topics having a worse rank in fusion search than in text search	36	36
# of topics having a better rank in fusion search than in text search	2	2

Table 3.1: Overall statistics of text-based searches and fusion searches, regarding all 298 search topics from TRECVID.

The mean inverted rank of text searches reached a score of 0,266, whereas the fusion of text searches and content-based searches received a score of 0,260. A total amount of 155 search topics were found with the first 100 results for text searches. The fusion of searches only retrieved 130 results. Some videos retrieved by fusion searches are not the desired target videos and are placed before them, hence decreasing their rank. The idea to apply a combination of text and content-based searches was to

increase the recall, by receiving additional candidate videos, which lacked for meta data or were badly tagged. Unfortunately only 3 videos could be found by fusion search, which were not found by text search. Desired target videos with a low amount of meta data are retrieved by considering additional visual information. They are ranked among the first 100 results for fusion search, but not for text search. Text based queries are not sufficient due to the lack of appropriate meta data. In contrast to that 28 videos were found by applying text search only, which were not found using a combination of searches. Around 99% of the videos in the test data collection featured usable meta data. Although the quality and the extent of annotations varied within video data set, the text search performed better than fusion search. However topics ranked top twenty account for 107 for both search types.

Distributions showing the ranks of text-based searches and fusion searches over all 298 topics are depicted in the following figures (3.7 and 3.8, topics not found have a value below zero).

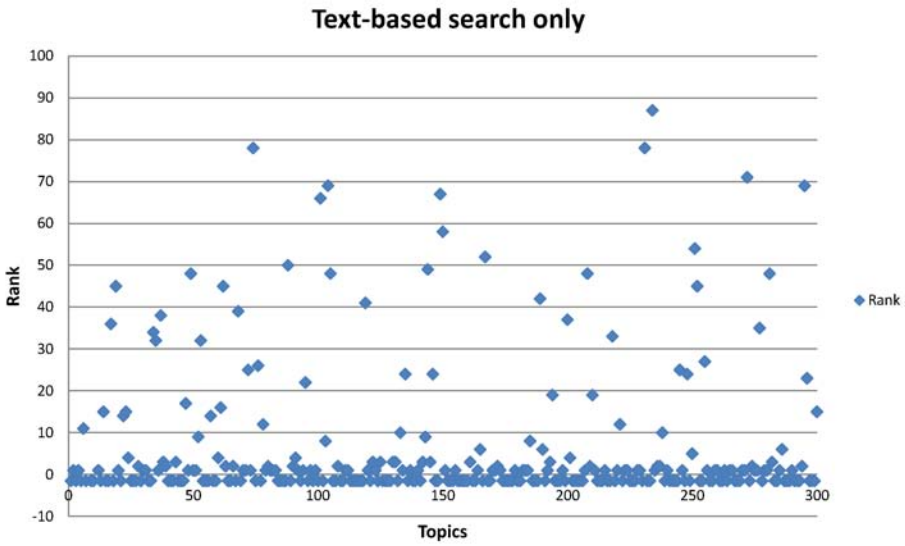

Figure 3.7: Distribution of ranks in results lists for text-based video search over all 298 topics.

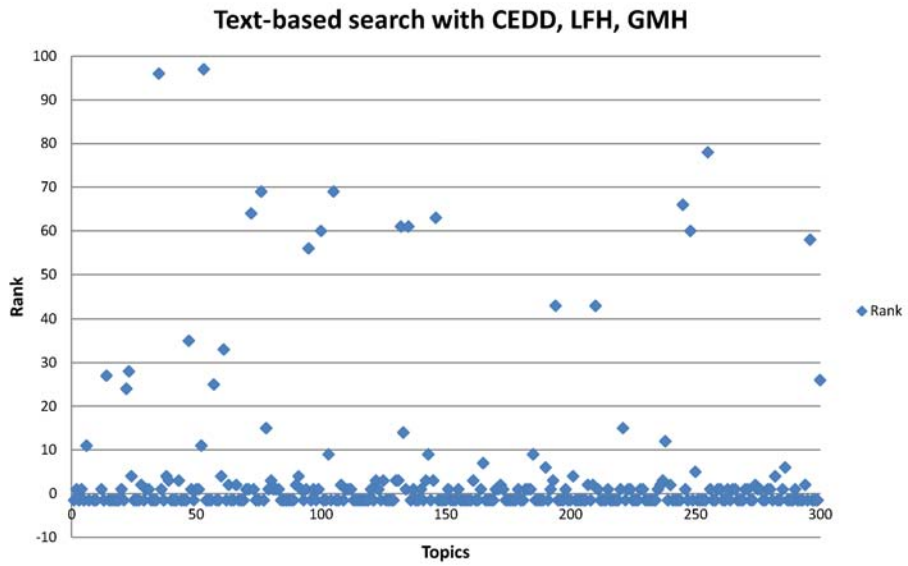

Figure 3.8: Distribution of ranks in results lists for fusion based video search over all 298 topics.

A total amount of 127 topics were both found by text-based searches and fusion searches. The mean rank over all 127 topics for text-based searches accounts for 5,73, whereas the mean rank for fusion searches accounts for 11. Moreover the ranks of the desired target videos for fusion searches spread more than the ranks for text-based searches.

Topics having a worse rank regarding fusion searches account for 36, whereas only 2 topics have a better rank than text-based search results. Fig. 3.9 depicts the detailed results of text-based and fusion searches over all 298 topics.

It is arguable that fusion search slightly improves the recall rate in case of badly annotated videos, which cannot be found by a text-based search paradigm. On the other hand pure text search performs quite well, due to a fairly amount of meta data distributed over the whole test collection. This speaks for the employment of advanced methods for text search. Nevertheless additional visual search can be applied, wherever a sparse amount of meta data prevails.

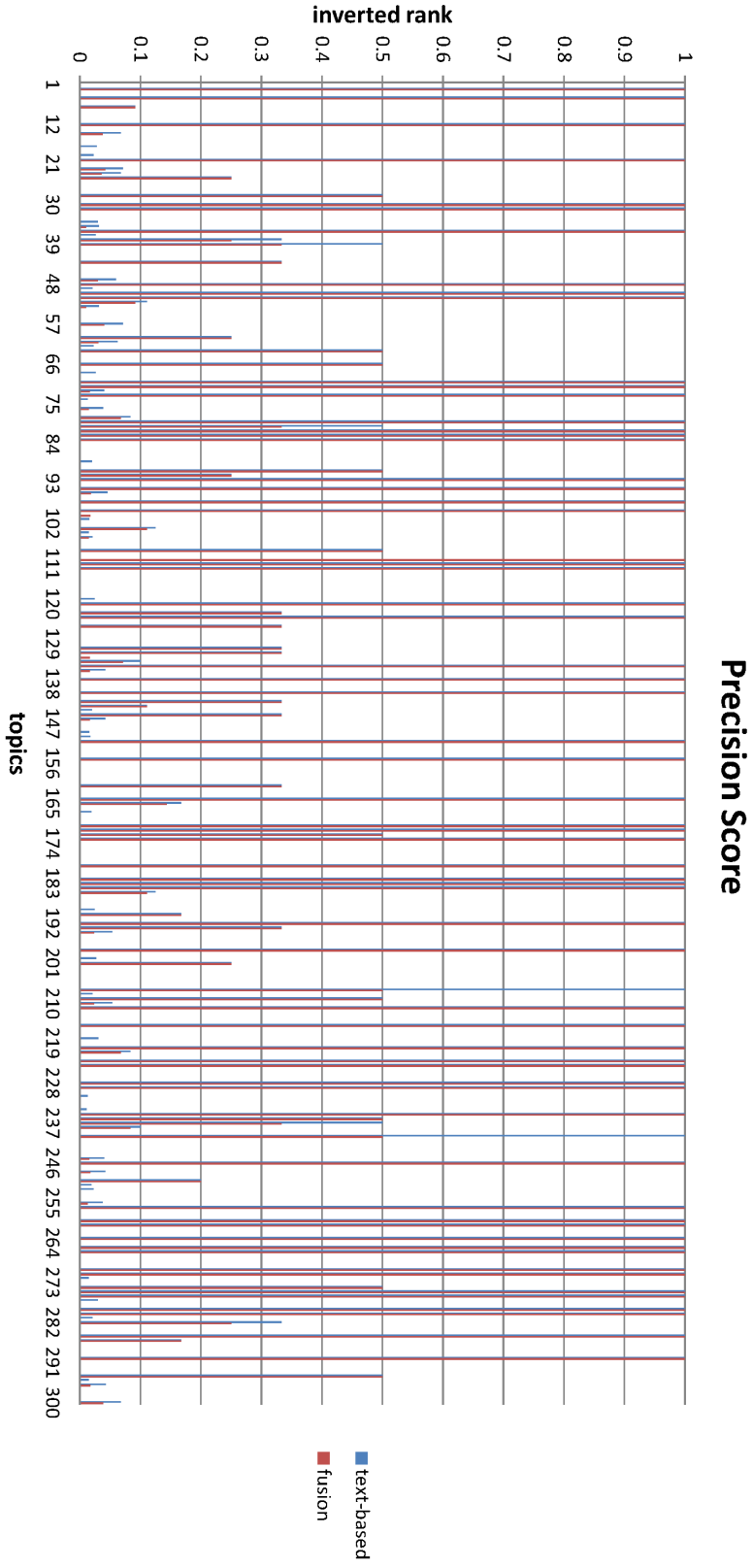

Figure 3.9: Inverted ranks of text-based search and fusion search for all search topics.

3.2 Bag of Visual Words for Automatic Video Summarization

Effective video retrieval and browsing is an important task to retrieve the desired video content in large multimedia repositories. Instead of typing search queries users often tend to browse video collections. As a consequence overviews of videos are desirable, which present the content of a video in a concise manner. A technique for video content representation, leveraging the Bag of Visual Words approach and further low level analysis techniques, is introduced in the next section (cf. [47]).

In the last decade the importance of videos conveying information has increased, which is accompanied by the need to store, organize and index the multimedia content appropriately, in order to support the user in retrieving videos. A lot of video clips are produced, broadcasted, shared and stored every day by professionals, amateurs and hobbyists. Finding videos matching the actual information need of a user proves to be a hard problem. *Video abstracts*, or video summaries, aim at presenting the semantics and content of a clip in minimized time and space to allow fast assessment of video clip relevance. In this section the focus lies on static methods: still image summaries showing keyframes of the video.

3.2.1 Approach

Generally speaking a video abstract should maximize the (semantic) information and minimize the length of a video. A method for keyframe selection and summarization is discussed in detail in [63]. It implements summarization of video clips by keyframe extraction based on global image features.

The method is extended in this thesis by supporting the extraction of representative images based on local features in order to find out, whether this new approach yield more representative summaries of video clips than the approach with global features. Figure 3.10 depicts the new approach, in order to get a better understanding. SIFT features proposed by Lowe [59] (see also chapter 2.1.2) to extract feature

vectors from salient keypoints of an image are applied. The salient points and their 128 dimensional feature vectors are interpreted as local features describing a video frame (a still image) of an uncompressed input video (see step 1 in Fig. 3.10). For pairwise comparison of frames the *Bag of Visual Words* approach is employed. All local features are clustered using k-means [36] (see step 2 in Fig. 3.10). The cluster centers are interpreted as reference feature for the whole cluster and are called *visual word*. A single frame is then represented by a histogram, called *local feature histogram*, denoting the occurrence of visual words within the frame (see step 3 and step 4 in Fig. 3.10). For keyframe selection the local feature histograms are k-medoid clustered [29] and cluster medoids are selected as representative keyframes of a frame cluster (see step 5 in Fig. 3.10). Cluster medoids are ranked based on the cluster size.

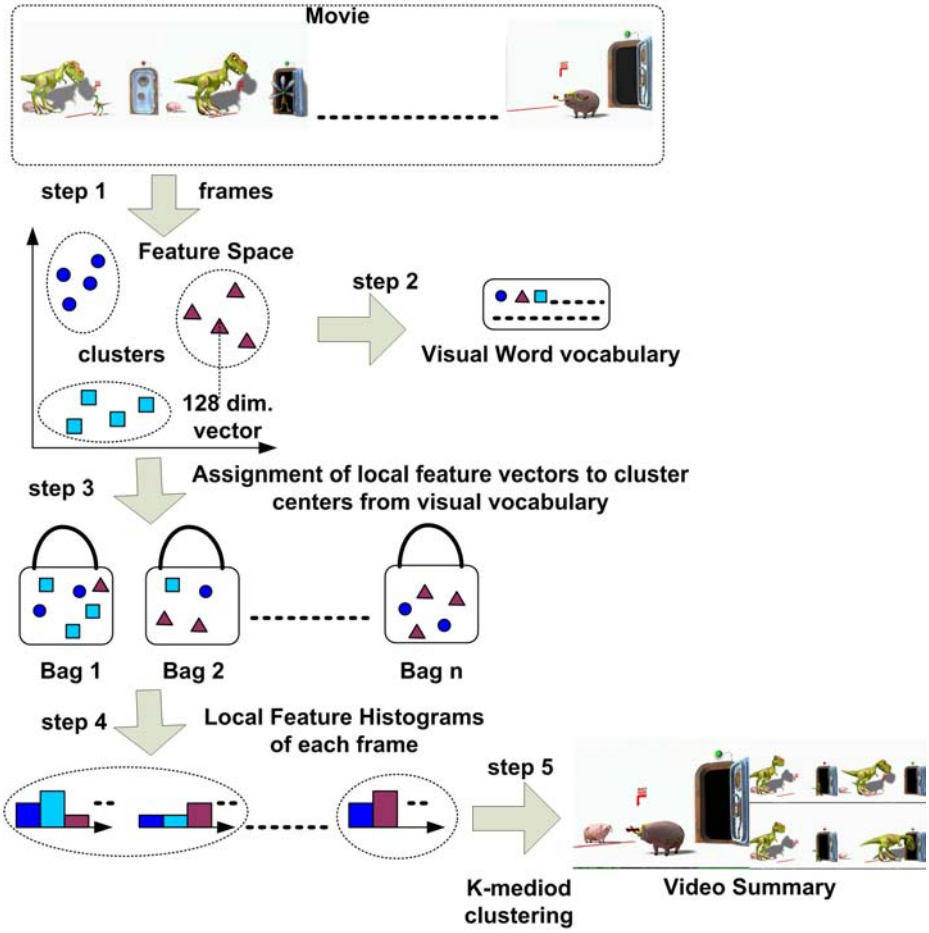

Figure 3.10: Bag of Visual Words for Video Summarization

The Bag of Visual Words technique is applied for summarizing a video, in order to get a video summary, which is more suitable for the user to assess the videos relevance. This should facilitate the search and browsing process of the user in a huge multimedia database, by depicting more meaningful images in a video summary.

3.2.2 Experiments and Results

The experiments were conducted on a off-the-shelf Intel Core 2 Duo CPU 2.3 GHz with 4GB Ram, by using short video clips ranging from 30 seconds to 76 seconds. Four

low level features (ACC, CEDD, RGB color histograms SIFT, all explained in section 2.1) were investigated, which led to four different summaries for each video clip. A number of global features were investigated in a published study [63]. Summaries generated on the basis of the ACC, CEDD and RGB features were favored by the users and therefore selected to compete with local SIFT features and Bag of Visual Words. For feature extraction the Lire library (see [61]) was utilized.

First and foremost quantitative statistics are presented showing the time needed to extract the visual features and time to build the feature vectors of every keyframe from the video. Therefore a 30 second H.264 video clip with a resolution of 480x360 and a frame rate of 29.92 frames per second was taken and analyzed. Figure 3.11 shows that feature extraction with global features is faster than the extraction of visual information based on local interest points. The Bag of Visual Words approach (k-means clustering with 1024 clusters and random initialization of cluster centers) utilizing SIFT features needs the most computation time with a mean value of 508 ms per frame in order to retrieve the visual information and to build the local feature vectors. Visual information extraction, by using 512 bins quantized RGB color histograms, needs 34 ms in average and performs best, followed by CEDD with 52 ms and ACC with 53 ms.

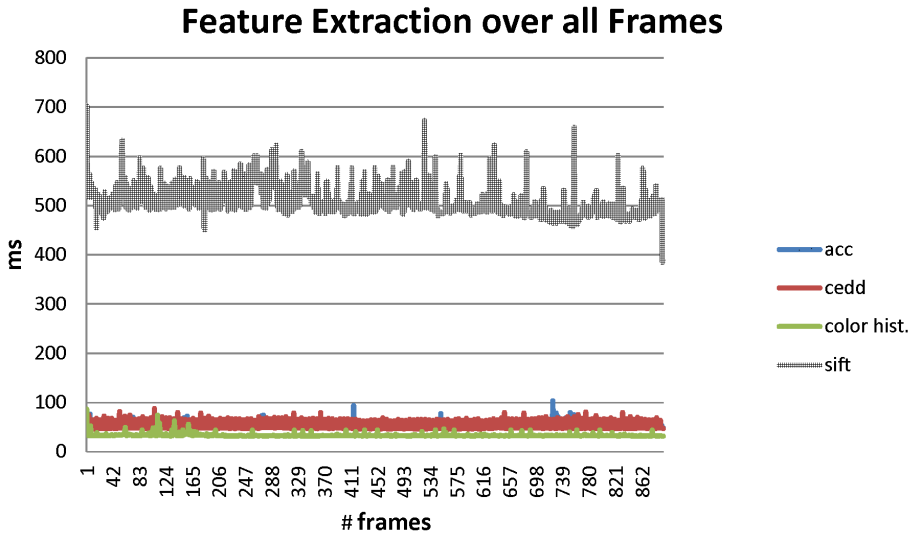

Figure 3.11: Performance evaluation - computation speed of feature extraction for frames

The overall computation speed of the whole summary process (see Fig. 3.12) is dominantly influenced by the feature extraction and feature vector building stage. The clustering of feature vectors and the summarization take less time. The best performance is again achieved by global features namely CEDD and color histogram features, due to the low amount of dimensions used to represent the visual information in the feature vectors. Hence the k-medoid clustering can be conducted in a more efficient way.

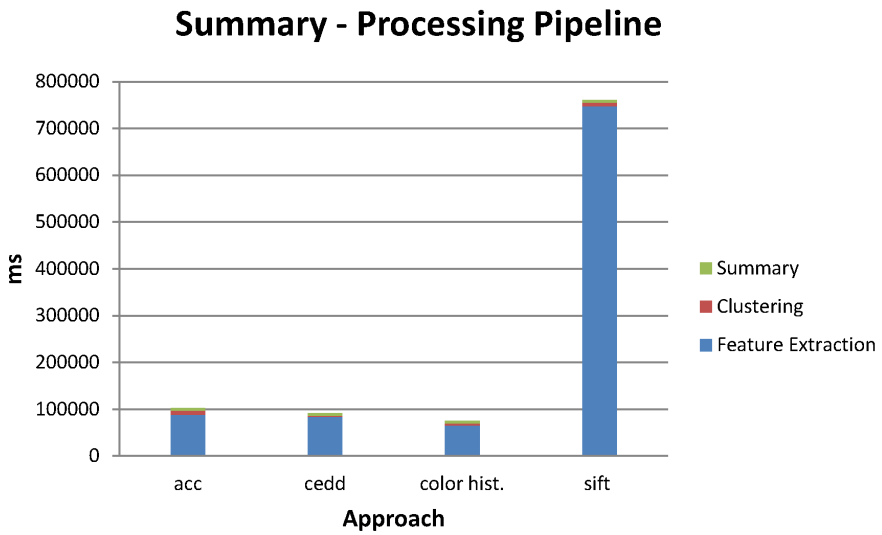

Figure 3.12: Performance evaluation - computation speed to generate video summaries, incorporating feature extraction, clustering and summary creation

Further more an exploratory evaluation was conducted, where users had to choose their favorite summaries depicting the corresponding videos in best manner. One summary consisted of five keyframes extracted by the particular approach. These keyframes were arranged in a single summary image which was presented to the user (see Fig. 3.13). As mentioned before four short video clips ranging from news to animations were analyzed. Because videos longer than five minutes probably cover too much information, which cannot be depicted properly in a video summary consisting of five still images, only short video clips were investigated. A further reason for selecting short clips is, that video clips recorded by users, in order to retain a moment of attraction, usually do not take longer than three minutes. This assumption is based on the observation that the average length of a video clip posted on YouTube is 2

Title	**Length**
iPhone commercial[2]	76 s
dinosaurs vault[3]	48 s
hurricane IKE - news reporter almost washed away[4]	30 s
shrek[5]	48 s

Table 3.2: videos used for exploratory study

minutes and 46.17 seconds[4].

Each video is summarized by a full sized frame of the biggest cluster (the cluster with most frames) on the left, followed by four frames half in width and height on the right representing smaller clusters. Figures (3.13, 3.14, 3.15 and 3.16) show the visualizations of the video summaries created by using the presented approach for the Shrek, the dinosaur, the iPhone and the news video.

Figure 3.13: 'shrek' video summary containing representative images of the five dominant clusters.

[4]Statistics from http://ksudigg.wetpaint.com/page/YouTube+Statistics
[5]http://www.youtube.com/watch?v=2k3zvI2tyPM
[6]http://www.youtube.com/watch?v=Dim0INyvJdw
[7]http://www.youtube.com/watch?v=SYl9mgFhe2o
[8]http://www.youtube.com/watch?v=uvyelwDA0Ws
(last checked: 2009-09-22)

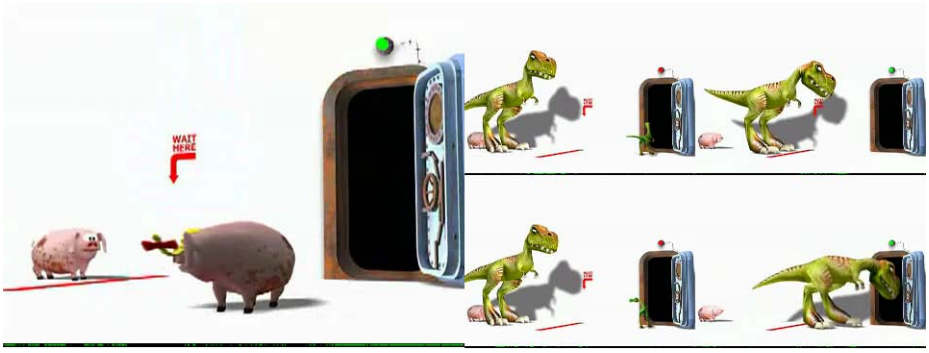

Figure 3.14: 'dinosaur vault' video summary.

Figure 3.15: 'iPhone commercial' video summary.

Figure 3.16: 'hurricane IKE - news reporter almost washed away' video summary

Each participating user had to assess four summaries (4 points for the best, down

	ACC	CEDD	color histogram	SIFT
iPhone	**29**	23	17	21
News	20	**31**	16	23
Shrek	11	**33**	25	21
Dinosaur	**27**	17	20	26

Table 3.3: Rating of the features for each video

to 1 point for the worst) for each video clip, which led to a total of 16 summaries. The user group consisted of 9 people (5 female and 4 male students); ages ranging from 20 to 30 years.

There was no clear winner in the experiment. All four selected image features got similar ratings from the test persons as Figure 3.17 shows. The summaries based on CEDD have reached the highest score (104 points), followed by the SIFT based visual bag of words approach (91 points), ACC (87 points) and the color histogram (78 points). The scores for each single video are shown in Table 3.3.

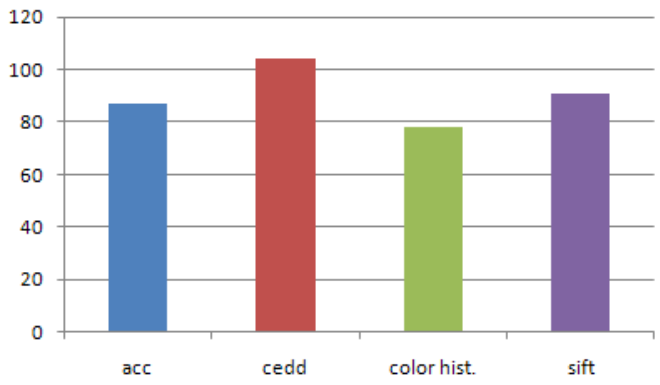

Figure 3.17: User ratings of each low level approach summed up for every video.

While CEDD produces very good results for the news video and the clip of the movie Shrek, it performs rather poor for the animation with the dinosaurs. On the other side, ACC reaches a high score for the iPhone commercial and the dinosaurs animation, but it is a bad choice for the Shrek clip. The SIFT-based Bag of Visual

Feature	Standard deviation
ACC	1.18
CEDD	1.14
color histogram	1.21
SIFT	0.84

Table 3.4: Standard deviation for the ratings of the selected visual features

Words approach never reached the best score, but also never performed worst, which can be seen easily in Table 3.3. In three cases (iPhone, news and dinosaurs) it reached the second highest score and in the fourth case (Shrek) it reached the third place. Therefore, it seems that this approach based on local image features produces more stable results than the ones based on global image features. This assumption is also supported by the deviation of the samples, given in Table 3.4. The local feature approach in the conducted experiments features lowest standard deviation (SIFT, 0.84) and can be considered the most stable approach for the selected test set.

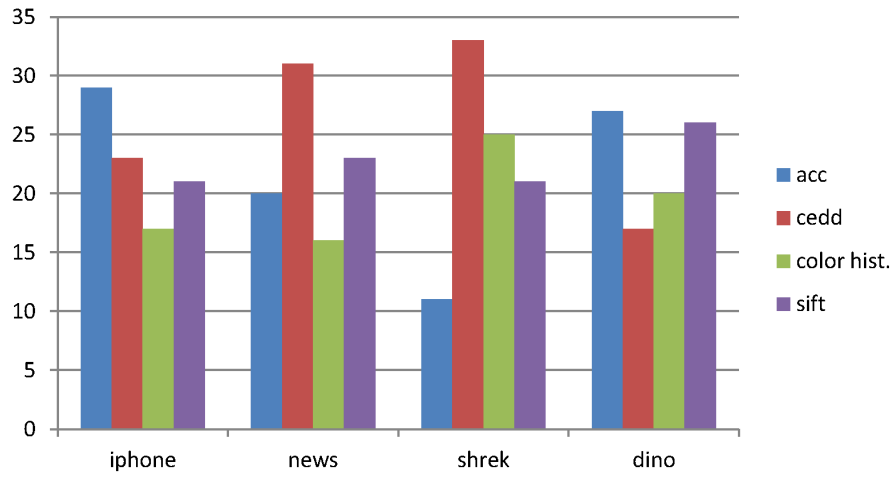

Figure 3.18: User ratings for every video.

The summarization of short video clips plays an important role for later retrieval.

Users browsing or searching for desired videos will get a first impression of the content, hence finding the video, they actually want to watch, much faster. The video summaries represented by still images constitute a good presentation of the video content, allowing for a fast assessment of the videos.
Video and image retrieval is becoming more important, due to a higher production of multimedia content. The next section, presented in this work, covers the topic of image retrieval by means of low level features.

3.3 Fuzzy Codebooks for Image Analysis and Content-based Image Retrieval

This section describes an approach for visual vocabulary generation (cf. [48] and [50]).

The Bag of Word approach for content-based image retrieval consists of several different sequential steps:

1. **Extraction of key points.** Multiple salient points are extracted from each image. SIFT and SURF are common choices for key point description. This work focuses on SURF utilizing the implementation provided in [61].

2. **Codebook creation.** All local features (one for each key point) are clustered. For each cluster a visual word is found. Usually a mean vector (in case of k-means) or a medoid is used as visual word. In this work a different, fuzzy clustering approach is employed and compared to the traditional approach utilizing k-means.

3. **Local feature histogram creation.** Each local feature of an image is assigned to the visual word most similar to the local feature. These assignments are quantized in a histogram, where each bin reflects a visual word. We speak about hard assignment, if a local feature is assigned to one single bin. We speak about

fuzzy (soft) assignment, if each local feature has a degree of membership to one or several bins.

The whole process results in one local feature histogram per image. These local feature histograms can then be used for content-based retrieval or classification. Pair wise distance can be determined by different metrics. Typically the Manhatten or Euclidean distance (see chapter 2.1.3) are used. In this work the focus lies on steps two and three of the pipeline. Fuzzy clustering and fuzzy (soft) assignment is employed to generate the visual vocabularies (codebooks) and the respective local feature histograms.

Fuzzy set theory uses a membership function $\mu_A : X \to [0, 1]$, which determines the degree of an element $x \in X$ belonging to a set A. Concerning the BoVW approach the set A denotes a cluster of local features represented by visual word. The sum of membership values of an element x to all visual words is one:

$$\sum_{A_i} \mu_{A_i}(x) = 1 \tag{3.1}$$

A fuzzy c-means clustering algorithm [9] is employed. In fuzzy c-means a membership function is iteratively used to assign data points $\vec{d} \in D$ to clusters $c_i \in C$ with $\bigcup_{c_i \in C} c_i = D$ and to compute cluster centers $\vec{m}_i \in M$:

$$\vec{m}_i = \frac{\sum_{\vec{d} \in D} \mu_{c_i}(\vec{d})^m \vec{d}}{\sum_{\vec{d} \in D} \mu_{c_i}(\vec{d})^m} \tag{3.2}$$

$$\mu_{c_i} = \frac{1}{\sum_{m_k \in M} \left(\frac{L_2(m_i,d)}{L_2(m_k,d)}\right)^{\frac{2}{m-1}}} \tag{3.3}$$

Parameter $m \in [1, \infty)$ is called *fuzzyfier* and controls the membership function. The larger m gets the fuzzier the membership will be. The algorithm terminates when a global optimization function

$$f = \sum_{\vec{d} \in D} \sum_{\vec{m}_i=1}^{c} L_2(\vec{d}, \vec{m}_i)^2 \mu_{c_i}(\vec{d})^m \tag{3.4}$$

reaches a minimum. The fuzzy c-means algorithm is defined as:

1. Randomly select n cluster centers.
2. Determine membership of each data point to each cluster (using the cluster center).
3. Compute f_{last}.
4. Recompute cluster centers based on the determined membership values.
5. Determine membership of each data point to each cluster (using the cluster center).
6. Compute f_{actual}.
7. (a) If $|f_{actual} - f_{last}| < \epsilon$ stop.
   (b) Else set f_{last} to f_{actual} and start over with step 4.

The same membership function μ_A (see Eq. 3.1) is used for soft assignments of local features to visual words and therefore, to create the local feature histograms. Figure 3.19 shows an overview of the proposed Bag of Visual Words pipeline for the applied experiments and the evaluation. Both traditional codebooks (codebook generation with k-means) and fuzzy codebooks (codebook generation with fuzzy c-means) are created based on SURF features and hard and fuzzy (soft) assignment, which is described in the next section, are applied, in order to compute the local feature histograms. Based on the created local feature histograms a k-nearest neighbor search in the Wang Simplicity data set (see [104]) set is conducted and results are evaluated as described in Section 3.3.2.

3.3.1 Soft Assignment using Fuzzy Set Theory

Usually a local feature vector does not only belong to one single cluster. The distribution in the feature space and the statistical mining with a clustering algorithm yield to an appearance of local feature vectors nearby several cluster centers (visual

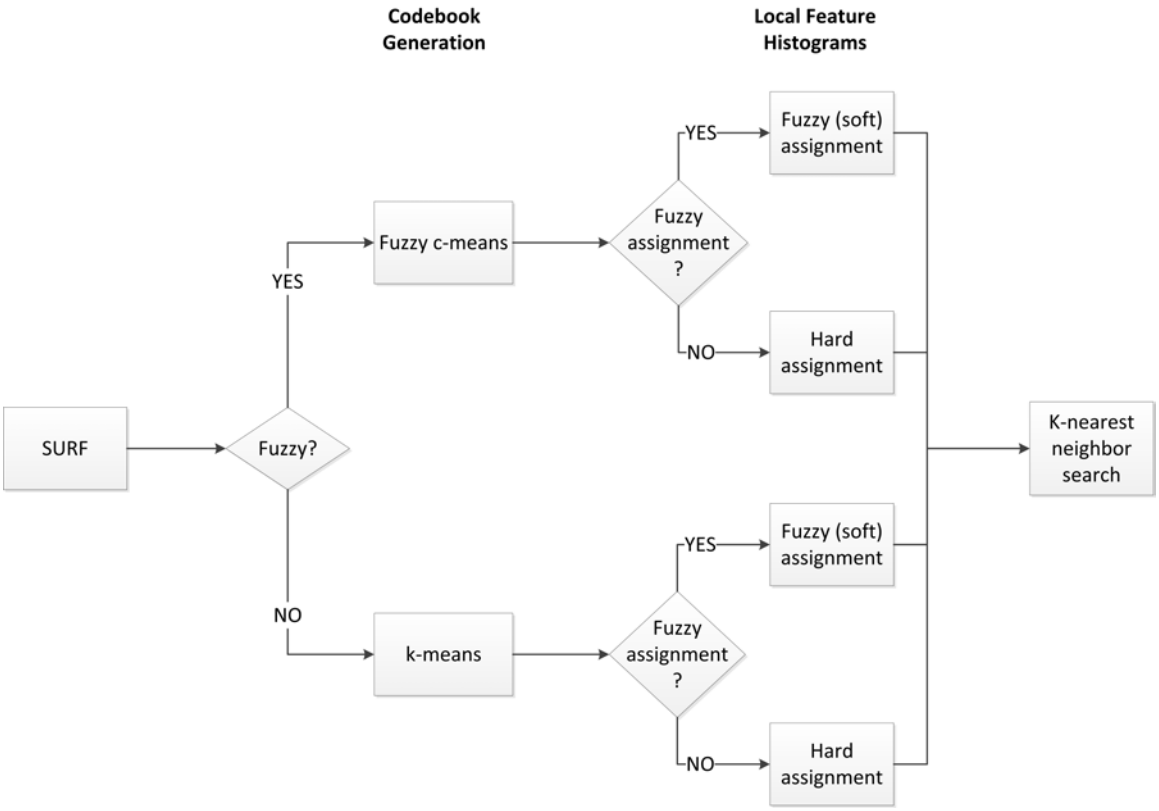

Figure 3.19: Visual Words pipeline

words). Hence an assignment strategy, which considers only one visual word, tends to be too restrictive. Soft assignment instead regards the membership of feature vectors to more than one cluster (see Fig. 3.20). The local feature vector, which needs to be assigned to the respective cluster centers is depicted as a black 'X'. The cluster centers or the visual words surround the vector and are depicted by the other shapes, which appear in the picture. The upper part of Fig. 3.20 represents the distance of the feature vector to the various visual words. The lower part of Fig. 3.20 depicts the assignment of the feature vector to the visual words either with the hard assignment or the fuzzy (soft) assignment approach.

For computing the assignment values to various visual words the fuzzy membership

function from the fuzzy c-means algorithm is employed and slightly adapted. Instead of computing the membership degree over all visual words, only a subset of them is used. The subset comprises the k-nearest visual words to a local feature vector, hence only looking at the most dominant clusters a vector belongs to. In order to find the k-nearest visual words, a k-nearest neighbor search is used. After the computation of the nearest visual words the fuzzy membership function is applied:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{C} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}} \tag{3.5}$$

- μ_{ik} denotes the membership value of the k-th local feature vector to the i-th visual word
- C denotes the amount of the previously computed nearest visual words
- d_{ik} denotes a distance metric (we used the Euclidean distance) of the k-th local feature vector to the i-th visual word
- d_{jk} denotes a distance metric (we used the Euclidean distance) of the k-th local feature vector to the j-th visual word
- m is called the fuzzyfier. A low value of m means a crisp assignment, whereas a high value signifies a more uniform distribution over the various visual words.

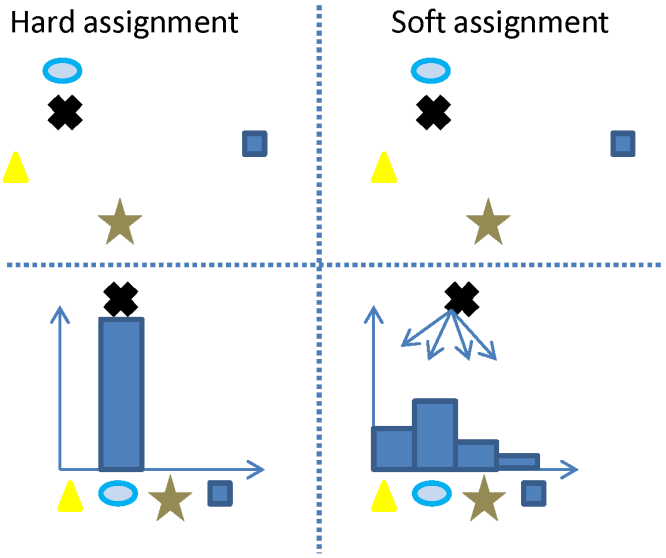

Figure 3.20: Hard vs. Soft Assignment.

3.3.2 Experiments and Results

The experiments were conducted on the Wang Simplicity data set, which contains 1,000 images. The images belong to 10 different concepts, each containing 100 images respectively. For visual vocabulary creation 250 images (25 of each concept) of the data set were processed both with the k-means and the fuzzy c-means algorithm. The vocabulary sizes range from 10 up to 512 cluster centers (10, 20, 30, 40, 50, 100, 200, 300, 400, 512), what accounts for 20 different vocabularies, 10 for both described clustering approach. Local feature vectors were computed with hard and soft assignment for each vocabulary, what leads to 40 different content-based indexes. These indexes were searched by conducting a nearest neighbor search. Overall 1,000 query images (100 per concept) were used for a query-by-example search to retrieve similar ones, belonging to the same concept as the query image. From the results of the k-nearest neighbor search the mean average precision (MAP) and the error rate (ER) were computed to allow objective comparison between both approaches. MAP

is based on the average precision (AP). For a single query $q \in Q$ the average precision denotes the mean of the precision scores after each retrieved relevant item.

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q)$$

The error rate reflects how often the first hit in the result list is not a correct one.

$$ER = \frac{1}{|Q|} \sum_{q \in Q} \begin{cases} 0 \ldots \text{if first element is correct hit} \\ 1 \ldots \text{otherwise} \end{cases}$$

In order to understand the following figures and tables the approaches and their abbreviations are explained:

approach	clustering method	assignment method
Fuzzy Codebooks_1_1 (HA)	fuzzy c-means with fuzziness parameter $m = 1.1$	hard assignment
Fuzzy Codebooks_1_1 (SA - 2)	fuzzy c-means with fuzziness parameter $m = 1.1$	fuzzy (soft) assignment - 2 clusters
Fuzzy Codebooks_1_1 (SA - 4)	fuzzy c-means with fuzziness parameter $m = 1.1$	fuzzy (soft) assignment - 4 clusters
Fuzzy Codebooks_1_4 (HA)	fuzzy c-means with fuzziness parameter $m = 1.4$	hard assignment
Fuzzy Codebooks_1_4 (SA - 2)	fuzzy c-means with fuzziness parameter $m = 1.4$	fuzzy (soft) assignment - 2 clusters
Fuzzy Codebooks_1_4 (SA - 4)	fuzzy c-means with fuzziness parameter $m = 1.4$	fuzzy (soft) assignment - 4 clusters
Fuzzy Codebooks_1_7 (HA)	fuzzy c-means with fuzziness parameter $m = 1.7$	hard assignment
Fuzzy Codebooks_1_7 (SA - 2)	fuzzy c-means with fuzziness parameter $m = 1.7$	fuzzy (soft) assignment - 2 clusters
Fuzzy Codebooks_1_7 (SA - 4)	fuzzy c-means with fuzziness parameter $m = 1.7$	fuzzy (soft) assignment - 4 clusters
k-means Codebooks (HA)	k-means	hard assignment
k-means Codebooks (SA - 2)	k-means	fuzzy (soft) assignment - 2 clusters
k-means Codebooks (SA - 4)	k-means	fuzzy (soft) assignment - 4 clusters

Table 3.5: Definition and abbreviations of all approaches used in the following graphs.

Fig. 3.21 depicts the mean average precision for content-based searches in a subset of the indexes. The visual vocabularies were computed with the k-means and the fuzzy c-means approach. For local feature histogram creation hard assignment was

applied in order to compare the two different codebook approaches in the first place. The graph shows the traditional approach with k-means and the new approach with fuzzy clustering, by using a fuzziness parameter m of 1.1, 1.4 and 1.7. The fuzzy approach with a fuzziness parameter $m = 1.1$ performs slightly better than the k-means approach for smaller vocabulary sizes from 10 up to 40, which takes over at a vocabulary size of 50. This can be reasoned in so far that dissimilar feature vectors are assigned to the same cluster, due to a smaller amount of clusters. This is attenuated by the fuzzy approach by taking the membership values to the various clusters into account.

This advantage disappears if the amount of computed clusters gets bigger. The leadership of k-means reaches a peak at a cluster amount of 300. From that on, the fuzzy approaches, especially the approach with a fuzziness paramter of 1.4, perform better. Since more clusters are involved, features representing visually similar content are assigned to different clusters. Fuzzy codebooks mitigate this effect. Interestingly the fuzzy approach with a fuzziness parameter $m = 1.7$ performs worst for nearly all vocabulary sizes, which reflects that a high fuzzyness parameter is counterproductive for good retrieval results.

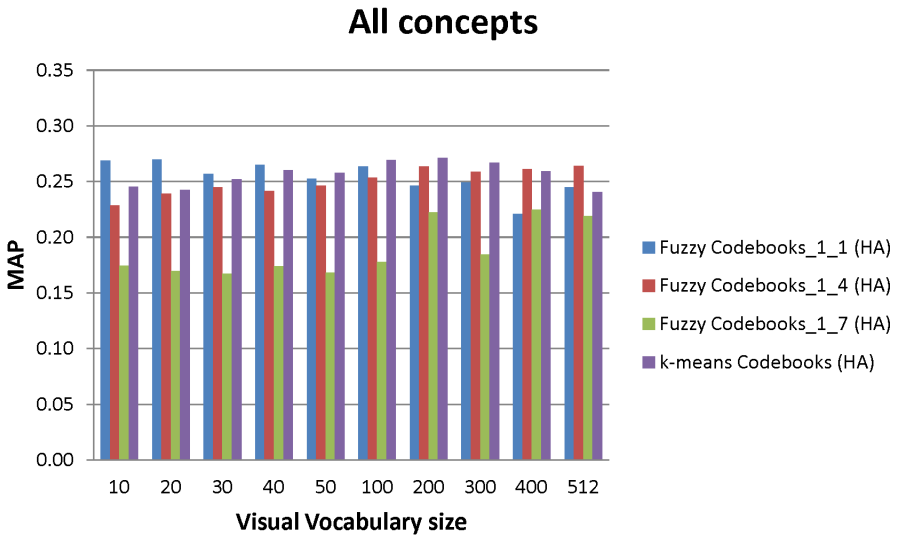

Figure 3.21: MAP for k-means codebooks and fuzzy codebooks with hard assignment for all concepts of the Wang Simplicity data set.

After looking at the two codebook generation approaches and their impact on retrieval performance utilizing hard assignment, hard and soft assignment are compared with respect to a particular clustering method. Figure 3.22 shows that the fuzzy assignment approach is better than hard assignment, starting at a cluster size of 40 clusters. The best performance is achieved by assignments to 4 different clusters, followed by assignments to 2 different clusters. Local feature vectors, appearing in the region of more than one cluster are better distributed over the clusters, hence preserving important visual information. By assigning a vector to only one cluster visual information is lost, what speaks for the application of the fuzzy assignment approach.

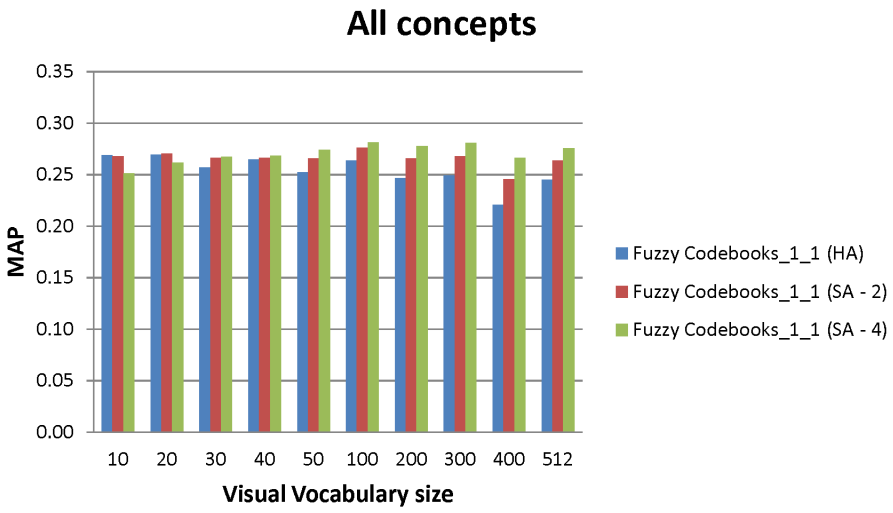

Figure 3.22: MAP for fuzzy c-means with $m = 1.1$ and hard / soft assignment.

The advantage of the fuzzy assignment approach is diminished, as can be seen in the results of Fig. 3.23 and 3.24. The mean average precision scores hardly get any better, due to a higher fuzziness parameter during the visual vocabulary generation process. This observation stands in contrast to the traditional codebook generation approach with k-means clustering (see Fig. 3.25). The result scores for soft assignment to a maximum of 4 clusters are always better than the other approaches, which are expelled to the rear places.

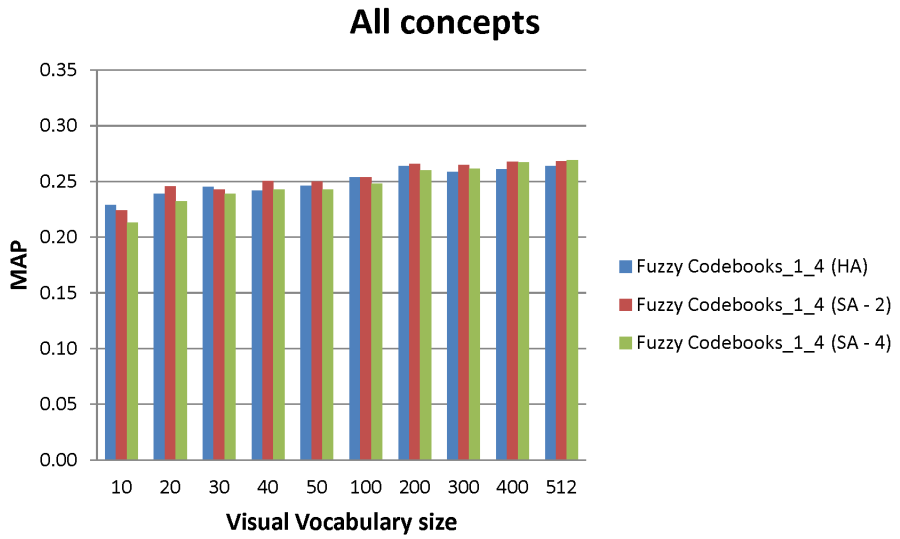

Figure 3.23: MAP for fuzzy c-means with $m = 1.4$ and hard / soft assignment.

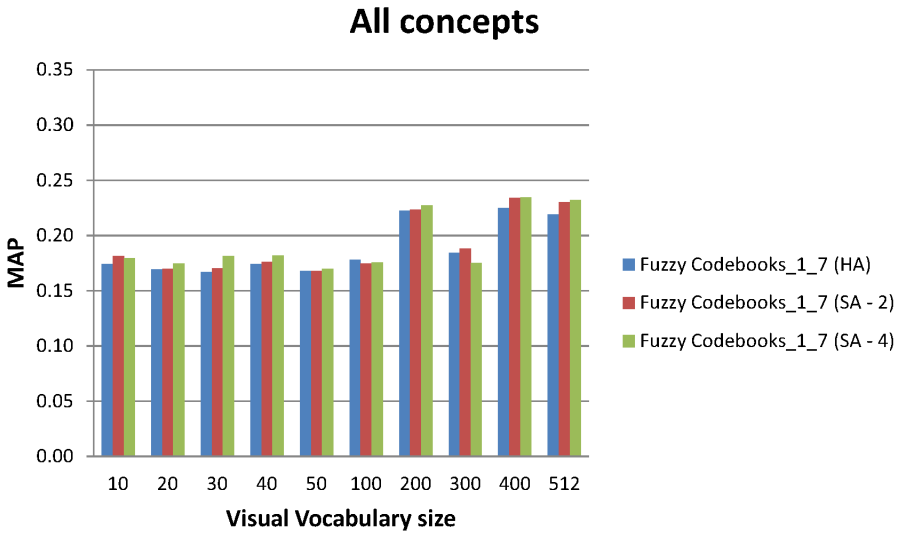

Figure 3.24: MAP for fuzzy c-means with $m = 1.7$ and hard / soft assignment.

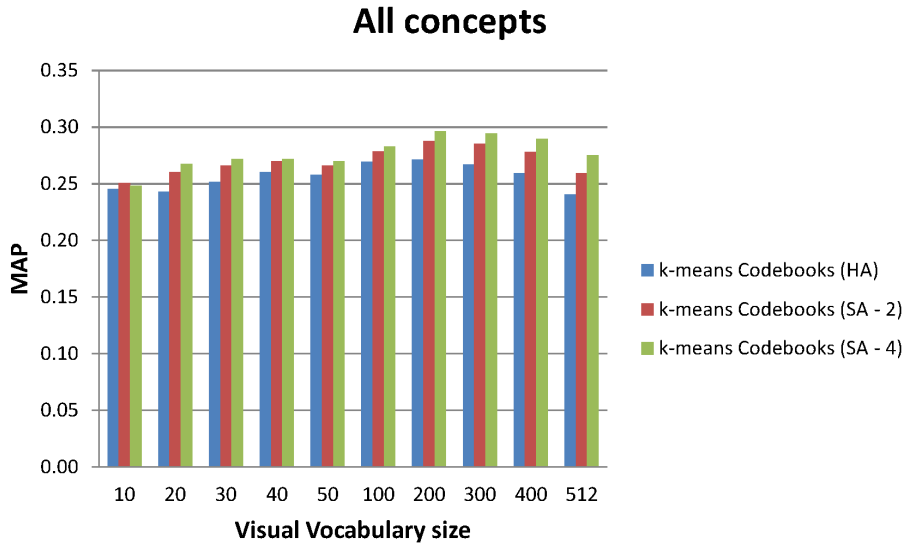

Figure 3.25: MAP for k-means codebooks and hard / soft assignment.

A thorough overview of the mean average precision scores for every approach over all concepts is presented in Table 3.6. The best results achieved are highlighted in bold. Fuzzy codebooks with a fuzziness parameter $m = 1.1$ and hard assignment provide the best result at cluster size of 10. The fuzzy assignment technique with assignments to a maximum of two clusters take the lead at an amount of 20 clusters, whereas the fuzzy assignment up to four clusters takes the first place with 50 and 512 clusters. The remaining cluster sizes (30, 40, 100, 200, 300, 400) benefit from the traditional codebook approach, utilizing k-means clustering, with a fuzzy assignment to four clusters.

	10	20	30	40	50	100	200	300	400	512
Fuzzy Codebooks _1_1 (HA)	**0.269**	0.270	0.257	0.265	0.252	0.264	0.246	0.249	0.221	0.245
Fuzzy Codebooks_1_1 (SA - 2)	0.268	**0.271**	0.266	0.266	0.266	0.276	0.266	0.268	0.246	0.264
Fuzzy Codebooks_1_1 (SA - 4)	0.251	0.262	0.267	0.268	**0.274**	0.281	0.278	0.281	0.266	**0.276**
Fuzzy Codebooks_1_4 (HA)	0.229	0.239	0.245	0.242	0.246	0.254	0.264	0.259	0.261	0.264
Fuzzy Codebooks_1_4 (SA - 2)	0.224	0.245	0.243	0.250	0.250	0.254	0.266	0.265	0.268	0.268
Fuzzy Codebooks_1_4 (SA - 4)	0.213	0.232	0.239	0.243	0.243	0.248	0.260	0.261	0.267	0.269
Fuzzy Codebooks_1_7 (HA)	0.174	0.169	0.167	0.174	0.168	0.178	0.222	0.184	0.225	0.219
Fuzzy Codebooks_1_7 (SA - 2)	0.181	0.170	0.170	0.176	0.168	0.175	0.223	0.188	0.234	0.230
Fuzzy Codebooks_1_7 (SA - 4)	0.179	0.175	0.182	0.182	0.170	0.176	0.227	0.175	0.234	0.232
k-means Codebooks (HA)	0.245	0.243	0.252	0.260	0.258	0.269	0.271	0.267	0.259	0.241
k-means Codebooks (SA - 2)	0.251	0.260	0.266	0.270	0.266	0.278	0.287	0.285	0.278	0.259
k-means Codebooks (SA - 4)	0.248	0.268	**0.272**	**0.272**	0.270	**0.283**	**0.296**	**0.295**	**0.290**	0.275

Table 3.6: MAP for every approach over all concepts.

Traditional codebooks with a fuzzy assignment to four clusters perform best, what can be derived from the mean of the MAP values over all cluster sizes (depicted in table 3.7). The fuzzy codebook generation technique with four assignments and a fuzziness parameter $m = 1.1$ follows the aforementioned approach. The fuzzy codebook generation approach with $m = 1.7$ performs worst.
However by looking at the standard deviation over all codebook sizes for the various techniques, the codebook generation with fuzzy c-means clustering and a parameter $m = 1.1$ provides the most stable results. Fuzzy c-means clustering with a fuzzy assignment to two clusters provides a standard deviation of 0.008, followed by an assignment to four clusters. The k-means clustering approach with hard assignment reaches the third place.

	mean	standard deviation
Fuzzy Codebooks_1_1 (HA)	0.254	0.015
Fuzzy Codebooks_1_1 (SA - 2)	0.266	0.008
Fuzzy Codebooks_1_1 (SA - 4)	0.270	0.009
Fuzzy Codebooks_1_4 (HA)	0.250	0.012
Fuzzy Codebooks_1_4 (SA - 2)	0.253	0.014
Fuzzy Codebooks_1_4 (SA - 4)	0.248	0.017
Fuzzy Codebooks_1_7 (HA)	0.188	0.024
Fuzzy Codebooks_1_7 (SA - 2)	0.192	0.027
Fuzzy Codebooks_1_7 (SA - 4)	0.193	0.027
k-means Codebooks (HA)	0.257	0.011
k-means Codebooks (SA - 2)	0.270	0.012
k-means Codebooks (SA - 4)	0.277	0.015

Table 3.7: Mean and standard deviation for every approach over all codebook sizes and concepts.

In order to gain insights into the MAP scores for some concepts Fig. 3.26 shows the results of the two best approaches (the fuzzy codebook generation approach with $m = 1.1$ and fuzzy assignment to four clusters and the k-means codebook generation approach with fuzzy assignment to four clusters) and the worst approach (fuzzy codebook generation with $m = 1.7$ and hard assignment). The images with concepts buildings and food are more diversified concerning the appearance of the visual content, hence leading to modest MAP values. The visual information contained in the descriptors is not sufficient enough to distinguish between images of other concepts. By looking at the three other graphs covering the concepts buses, flower and dinosaur, the overall MAP increases. The descriptors are discriminating enough to provide more similar results during search. Moreover the concept dinosaurs contains more near duplicates than the other concepts, what explains the good results for this specific concept.

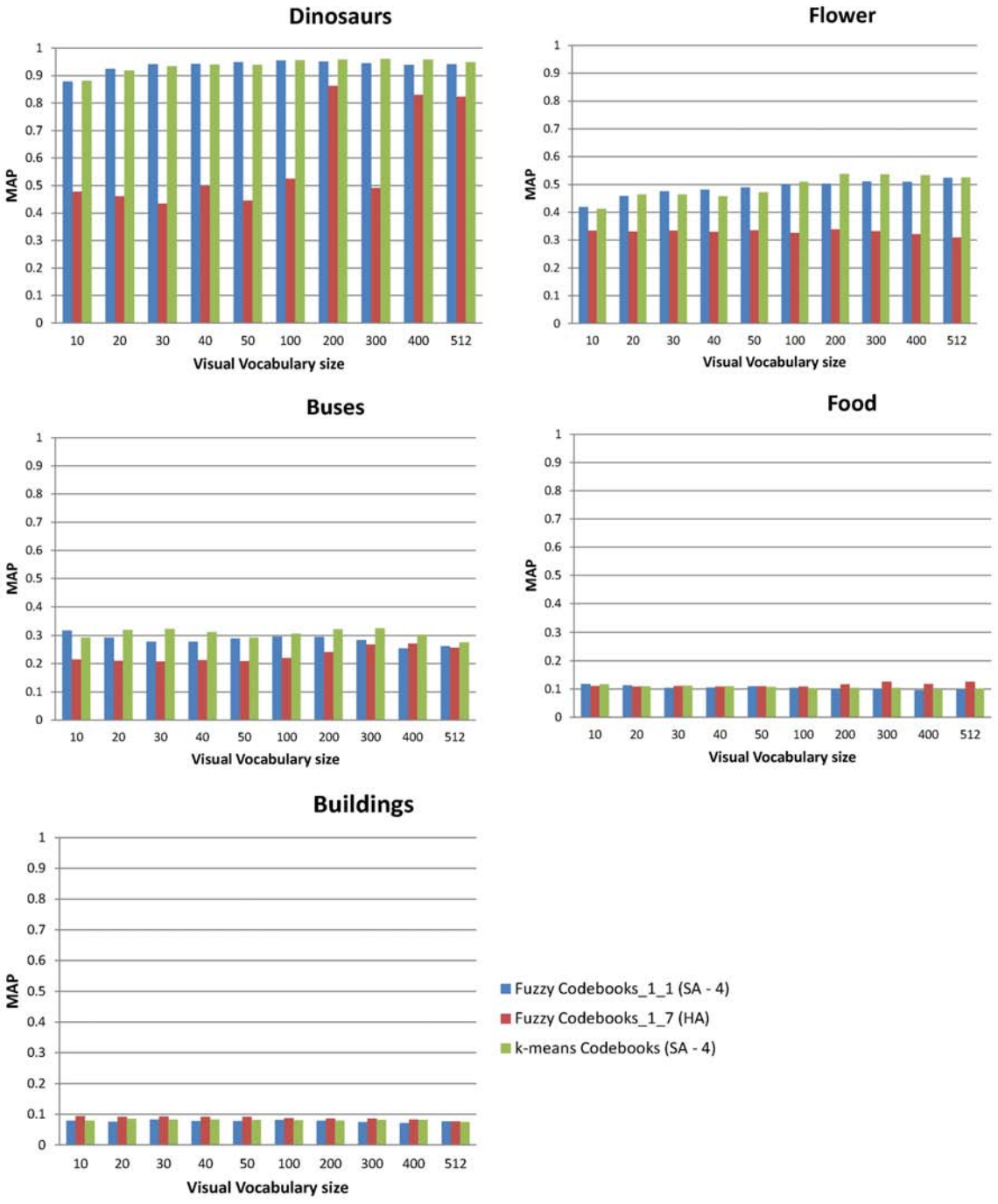

Figure 3.26: MAP of concepts: dinosaurs, flower, buses, food, buildings, leveraging the two best approaches (k-means codebooks with soft assignment 4 and fuzzy codebooks with $m = 1.1$ and soft assignment 4) and the worst approach (fuzzy codebooks with $m = 1.7$ and hard assignment).

In addition to the mean average precision scores the error rate is shown in the following figures. The lower the error rate the more often a correct hit appears at the first rank. First and foremost the codebook generation approaches are again compared against each other (see Fig. 3.27. The fuzzy codebook generation approach with a fuzziness parameter $m = 1.1$ and the k-means codebook generation approach perform best, exhibiting more first place hits than the fuzzy approach with $m = 1.4$ and $m = 1.7$. While the fuzzy approach ($m = 1.1$) achieves slightly better results than k-means codebook approach for visual vocabulary sizes less than 50, k-means clustering leads to better results for higher codebook sizes.

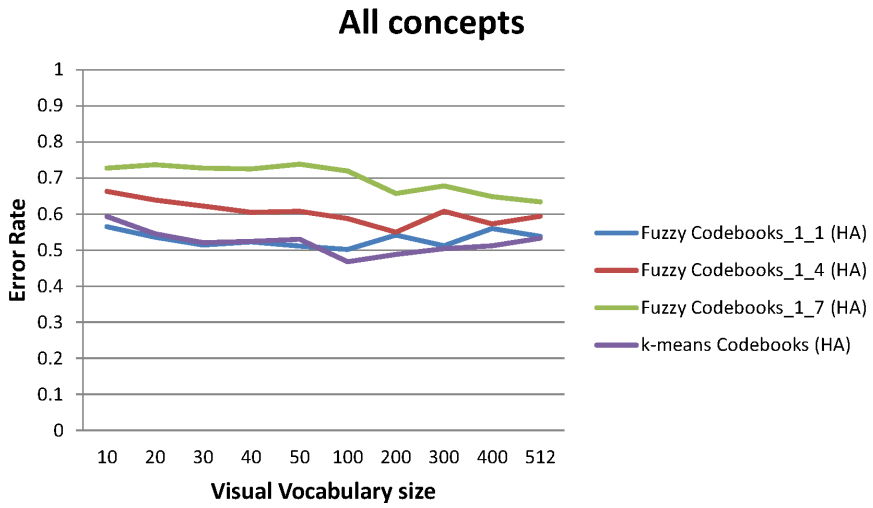

Figure 3.27: Error rates for k-means codebooks and fuzzy codebooks with hard assignment for all concepts.

Next the error rates for a particular codebook generation technique with hard and fuzzy assignment are represented. While one can observe a performance gain in terms of lower error rates for the fuzzy codebook generation approach with $m = 1.1$ and the k-means clustering method utilizing fuzzy assignment (see Fig. 3.28 and 3.31), the scores for fuzzy codebook generation with a fuzziness parameter $m = 1.4$, $m = 1.7$ and fuzzy assignment do not provide much better results than hard assignment (see

Fig. 3.29 and 3.30).

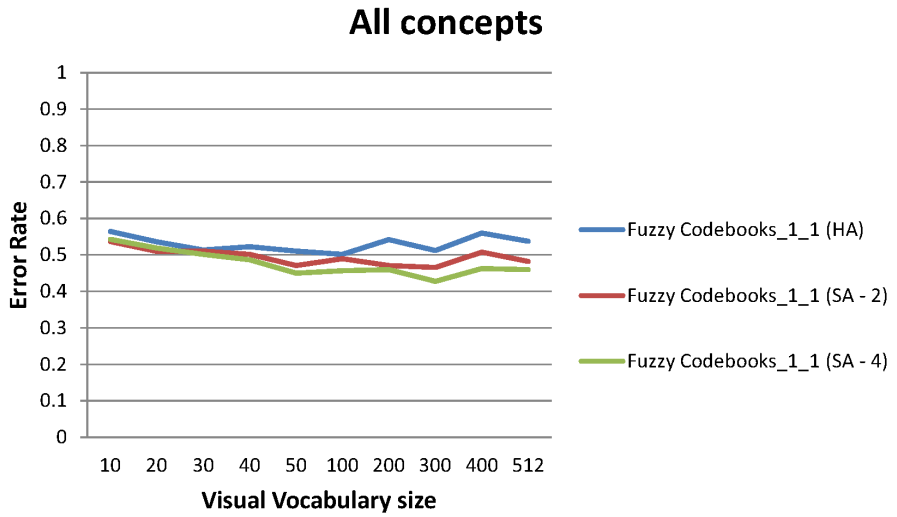

Figure 3.28: Error rates for fuzzy c-means with $m = 1.1$ and hard / soft assignment.

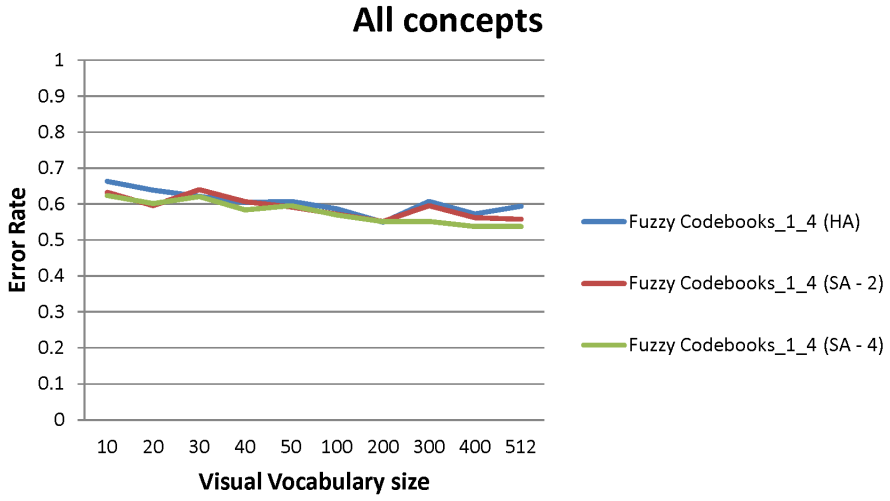

Figure 3.29: Error rates for fuzzy c-means with $m = 1.4$ and hard / soft assignment.

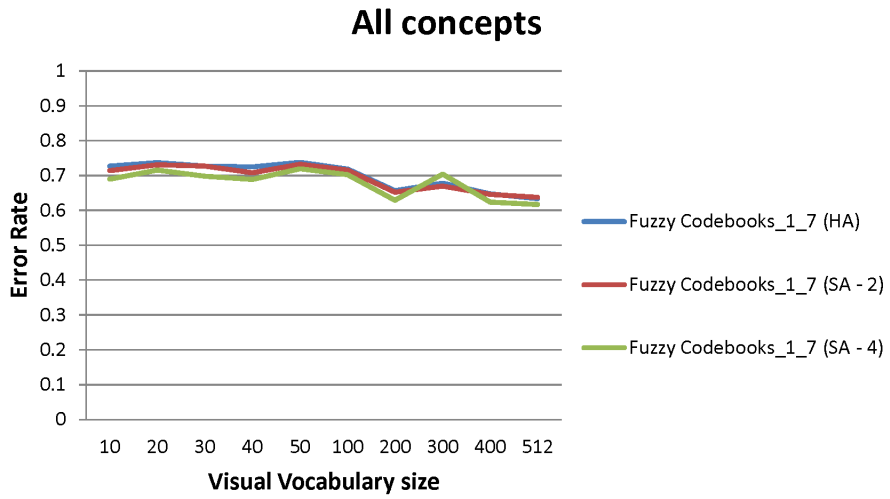

Figure 3.30: Error rates for fuzzy c-means with $m = 1.7$ and hard / soft assignment.

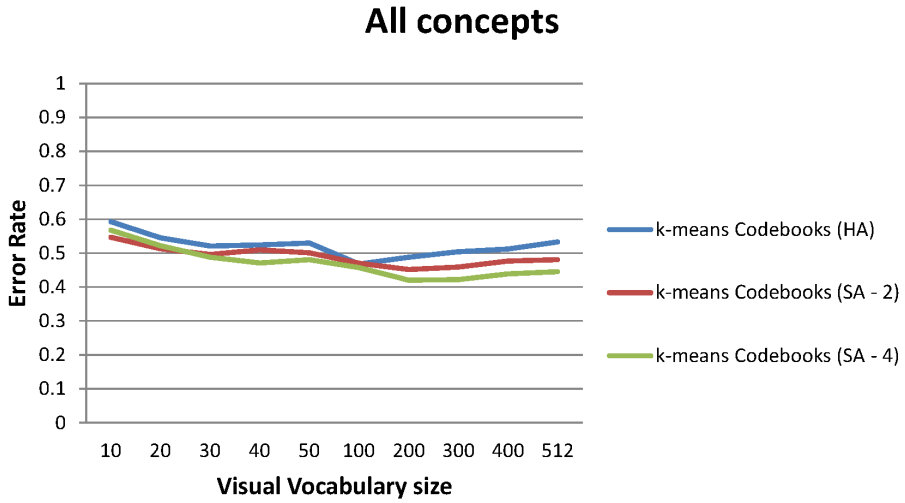

Figure 3.31: Error rates for k-means codebooks and hard / soft assignment.

An overview of all error rates is given in table 3.8. The fuzzy codebook generation approach with a fuzziness parameter $m = 1.1$ and a fuzzy assignment to two clusters provides the best results for small codebook sizes such as 10 and 20. For 30 and 40

clusters the k-means approach takes over, being superseded at an amount of 50 cluster from fuzzy codebooks with a fuzziness parameter $m = 1.1$ and a fuzzy assignment to four clusters. For 100 clusters the hard clustering and the fuzzy clustering both with fuzzy assignment to four clusters provide the best results. From 200 clusters up to 512 the hard clustering approach together with fuzzy assignment wins the competition.

	10	20	30	40	50	100	200	300	400	512
Fuzzy Codebooks _1_1 (HA)	0.565	0.536	0.514	0.523	0.511	0.502	0.542	0.512	0.560	0.538
Fuzzy Codebooks_1_1 (SA - 2)	**0.537**	**0.510**	0.511	0.502	0.471	0.490	0.471	0.466	0.508	0.482
Fuzzy Codebooks_1_1 (SA - 4)	0.543	0.519	0.502	0.487	**0.450**	**0.457**	0.460	0.428	0.463	0.460
Fuzzy Codebooks_1_4 (HA)	0.663	0.639	0.622	0.605	0.608	0.587	0.550	0.608	0.573	0.594
Fuzzy Codebooks_1_4 (SA - 2)	0.633	0.596	0.640	0.607	0.592	0.572	0.552	0.596	0.562	0.558
Fuzzy Codebooks_1_4 (SA - 4)	0.624	0.602	0.621	0.584	0.596	0.570	0.551	0.552	0.538	0.538
Fuzzy Codebooks_1_7 (HA)	0.727	0.737	0.727	0.725	0.738	0.719	0.657	0.678	0.648	0.634
Fuzzy Codebooks_1_7 (SA - 2)	0.714	0.731	0.727	0.708	0.733	0.716	0.652	0.670	0.646	0.638
Fuzzy Codebooks_1_7 (SA - 4)	0.690	0.716	0.698	0.689	0.720	0.703	0.629	0.704	0.624	0.617
k-means Codebooks (HA)	0.593	0.545	0.521	0.524	0.530	0.468	0.488	0.504	0.512	0.533
k-means Codebooks (SA - 2)	0.547	0.513	0.496	0.510	0.501	0.470	0.452	0.459	0.477	0.481
k-means Codebooks (SA - 4)	0.568	0.522	**0.488**	**0.471**	0.481	**0.457**	**0.420**	**0.422**	**0.439**	**0.445**

Table 3.8: Error Rate

3.4 Summary

Low level features such as SIFT and SURF in combination with Bag of Visual Words approaches were employed for video retrieval, representation of videos for fast assessment by conducting video summarization and image retrieval. More specifically Bag of Visual Words related motion codebooks were introduced, which leverage motion information of motion vectors computed for macro blocks. Local feature histograms computed with motion codebooks were applied together with CEDD features in order to re-rank videos in combination with text-based searches.

In order to present the videos in a concise way, so that a user can retrieve them later on during search and browsing sessions, video summaries leveraging global features and SIFT features with Bag of Visual Words were created. Still image video summaries were automatically generated by applying a k-medoid clustering algorithm. A user test was conducted to gain insights in the expressiveness of the video summaries computed with different low level features.

The last section of this chapter introduced a new codebook generation technique for the BoVW approach. Fuzzy codebooks were generated by means of a fuzzy c-means clustering algorithm and fuzzy (soft) assignment was leveraged to compute local feature histograms. Tests on the Wang Simplicity data set were performed and the various BoVW approach were compared.

Chapter 4

Leveraging Visual Information in Context of User Intentions during Search

The overwhelming availability of visual content on the Internet poses a serious problem: although there are huge resources to tap, users often cannot find the content they actually want. By considering context information such as user intentions the search space can be narrowed (see example 1), in order to retrieve the desired visual content.

An intention (see chapter 1.2 and 2.2) is a plan someone has in mind to achieve a goal. This plan is expressed by the user's search behavior (e.g. user clicks, search queries, etc.), which reflects the information need of a user during search. For instance, if someone wants to buy a cheap motorcycle, the information need – find photos of motorcycles for sale – is based on the intention to buy a motorcycle. Each user has a certain information need, while searching for visual content. This information need must be communicated to the search system, in order to produce the best possible retrieval results.

In this chapter visual information retrieval is regarded in terms of specific goals and information needs (e.g. "find images of of black & white Adidas sport shoes") and

less specific goals and broader information needs (e.g. "find images of sport shoes, different in size and shape of various brands") during search. A novel approach to retrieval adaptation based on the degree of intentionality is presented. Tests on two well known data sets demonstrate the potential of the approach (cf. [49] and [51]).

4.1 Visual Vocabularies in Context of User Intentions

User goals during search in context of user intentions and their benefit for visual information retrieval (VIR) have not yet been extensively investigated and pose a challenging research area. Once a user's goal is known – whether it is explicitly communicated or deduced by intelligent classification algorithms – various system aspects can be adapted:

- by adapting the relevance function to the user's needs,
- by adapting the result view,
- by offering options that are relevant to the user's goal or inferring parameters from the user's goal.

More specifically, this work is written under the assumption that users, pursuing different goals, will benefit from relevance functions tailored for a particular information need; therefore, retrieval mechanisms are adapted with regard to user goals and information needs in context of user intentions.

As stated by Datta et al. in [20] the importance of building human-centered multimedia systems is crucial to increase the user satisfaction while searching for visual content on the web. Hence an adequate representation of visual content and the recall are important factors, in order to satisfy the user's information need. The research activities presented in this chapter, explicitly address this point. Similarity searches on different content-based indexes are conducted, with respect to various information

needs. The indexes are built leveraging the BoVW approach for visual information retrieval. Clustering techniques such as k-means and hierarchical clustering are exploited, to get different visual vocabularies, which are related to distinctive degrees of intentionality (see section 4.1.1). The assumption is that smaller vocabularies model vague information needs, such as: "get me images of sport shoes, different in size and shape of various brands", in a better way than larger once, due to a rigourously more quantized feature space. As a result visual dissimilar information is preserved, which leads to a broader result set during the retrieval process. Larger vocabularies however are rather suitable for more specific information needs, by keeping a higher degree of granularity and thus a more accurate representation of visual features. A user with a specific information need like: "get me images of black & white Adidas sport shoes", will benefit from an accurate description of the visual content, which is supported by less quantized vocabularies. Figure 4.1 depicts the idea of combining user goals during search in context of user intentions with BoVW, by using smaller vocabularies for vague and larger vocabularies for more specific information needs.

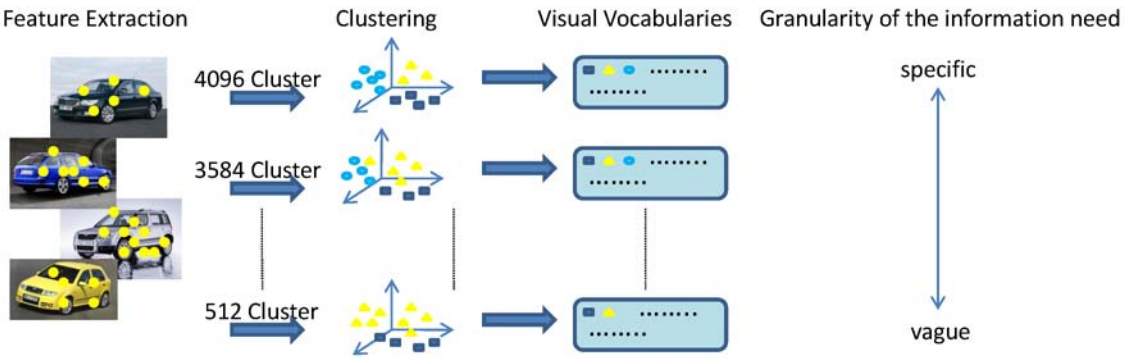

Figure 4.1: Intention modeling

In the following specific information needs and less specific more vague information needs, while searching for images, are distinguished. Table 4.1 shows examples of the assumed classes.

specific information need	vague information need
get me images of collies	get me images of different dog breeds
get me images of Tarmac road bikes of a specific shape	get me images of mountain bikes, comfort and leisure bikes not regarding any shape restrictions
get me images of blue Chevrolet Camaros	get me images of muscle cars in the 70s
get me images of black & white Adidas sport shoes	get me images of sport shoes, different in size and shape of various brands

Table 4.1: Examples modeling different information needs.

Users with different degrees of intentionality benefit from distinctive vocabulary sizes, different in their granularity of visual accuracy. K-means clustering is used to build the visual vocabularies and hard assignment to clusters is applied, in order to build the content-based indexes. The particular indexes, different in their granularity when describing the visual content, are employed for near-duplicate searches and a user simulation to gain insights into the connection between them and the respective degree of intentionality. The assumed hypothesis are stated as follows:

- H1: "Users with a more specific information need benefit from larger vocabularies of visual words."
- H2: "Users with vague information needs benefit from smaller vocabularies."

The two hypothesis are sustained by applying various Bag of Visual Words approaches additionally. Agglomerative hierarchical clustering (HCL) of visual words for visual vocabulary generation in combination with hard assignment is utilized, to build a subset of all used indexes for the experiments. By applying HCL on visual words directly visual vocabularies can be generated more efficiently in terms of speed,

instead of applying k-means clustering of all feature vectors over the whole feature space for every single vocabulary. Besides fuzzy (soft) assignment is used, in order to build additional indexes. The assignments are based on k-means generated visual vocabularies. As already mentioned in chapter 3.3.1 a local feature vector, which describes a local image patch, belongs to more than one visual word. Hard assignment assigns a local feature vector to only one visual word, hence losing important information. By using a soft assignment approach a local feature vector can be assigned to more than one visual word, hence preserving important information about the visual content. In addition to the assumption that users with different information needs benefit from distinctive vocabulary sizes, the impact of soft assignment on different information needs is investigated and compared with hard assignment.

4.1.1 HCL and Hard Assignment in Context of User Intentions

The proposed approach adopts a hierarchical clustering of visual words, which provides a principled way of adjusting the vocabulary size to reflect the granularity of users' needs. If we consider a clustering routine with k hierarchy levels, each image can be described by k different local feature histograms. The hierarchy levels constitute different visual vocabularies. The first histogram is built on the visual vocabulary at the lowest level, the second histogram is computed on the visual vocabulary at the second lowest level, and so forth. Histograms built with vocabularies of lower levels characterize the content of an image with a lot more different visual words and hence in more detail than histograms built with vocabularies on a higher level, which contain more visual diversified information. Users with a vague information need benefit from a less specific more diversified content-based representation of an image. On the contrary users with a more specific information need benefit from vocabularies with more visual words, since a more accurate content based description of the images exists. For vocabulary generation SURF descriptors of all images from the two training data collections are extracted. The computed descriptors are k-means clustered,

separately for each data set. An amount of 4,096 clusters is used, i.e. a vocabulary of 4,096 visual words, to perform agglomerative hierarchical clustering, where similar visual words are merged to clusters in every iteration step. The computation stops when all visual words are assigned to one cluster, which denotes the root element of the hierarchy tree (as depicted in Fig. 4.7). Local feature histograms based on different hierarchy levels, and hence on distinctive vocabulary sizes, are created for the images belonging to the test data collection. For the creation of local feature histograms hard assignment is applied, which means that a local descriptor of an image was only assigned to one cluster.

Eight different indexes for each data set with 512, 1024, 1536, 2048, 2560, 3072, 3584, and 4096-dimensional local feature histograms are computed in total.

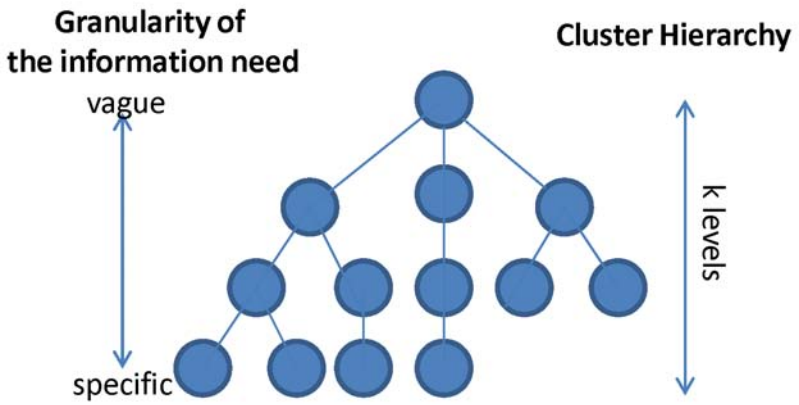

Figure 4.2: Hierarchy levels with respect to the degree of intentionality and the granularity of the information need.

4.1.2 Soft Assignment in Context of User Intentions

In addition to k-means clustering and hierarchical clustering with hard assignment content-based indexes are created by utilizing the soft assignment strategy mentioned in chapter 3.3.1. The indexes are built with 2, 4 and 8 nearest visual words, what accounts for 3 different indexes for one data set and one visual vocabulary. The overall

amount of indexes over the eight different visual vocabularies and the two data sets accounts for 48. Beside the assumption that smaller vocabulary sizes are better for users with a vague information need and vice versa, a further research question is posed:"Will the amount of used visual words for soft assignment reflect users' degree of intentionality?"

The assumption is that users with a somewhat vague information need will benefit from a higher amount of used visual words during the soft assignment process, because visual information is distributed over more visual words. On the other hand users with a more specific information need are better served by indexes, which are constructed with a smaller amount of visual words used during the membership computation.

4.2 Experiments and Results

4.2.1 Used data sets

The tests were conducted on the Pascal VOC 2007 ([24]) and the Wang Simplicity data set. The Pascal VOC 2007 data set contains 9,963 images, 5,011 from the training data collection and 4,952 images from the test data collection. It comprises 20 different semantic concepts with different cardinalities. The amount of images per concept varies from 200 up to 700, except for the concept *person*, which accounts for roughly 2000 pictures. Several pictures cover more than one concept, e.g. a picture where a person is standing in front of a car belongs to the two concepts, *person* and *car*. The Wang Simplicity data set contains 1,000 pictures, with 10 concepts, where the amount of pictures is evenly distributed and accounts for 100 pictures per concept. It is not a priori divided in a test and training data collection, what led to the procedure of picking 250 images (25 of each concept) from the collection as training instances. Each image belongs to one single concept. Eight visual vocabularies of different size (4096, 3584, 3072, 2560, 2048, 1536, 1024, 512) for each of the two data sets were computed, by utilizing the k-means clustering algorithm. Additionally eight visual vocabularies by using hierarchical clustering for each data set were generated.

Content-based indexes were computed by leveraging hard assignment for k-means and HCL generated visual vocabularies. The fuzzy (soft) assignment method was additionally applied to compute indexes based on k-means visual vocabularies.

4.2.2 Result Set Diversity

The first step in the evaluation was to show that the results retrieved with different vocabulary sizes are different. This is a prerequisite to the claim that such results may map to different degrees of intentionality. In order to demonstrate this a similarity search leveraging the Euclidean distance as a similarity metric was conducted and the Jaccard coefficient J (Equation 4.1) to show the diversity of different result sets R_1 and R_2 was computed.

$$J = \frac{\mid R_1 \cap R_2 \mid}{\mid R_1 \cup R_2 \mid} \tag{4.1}$$

For each pair of result sets among the eight sample sets (the eight indexes computed with k-means and hard assignment) for every single concept for each data set the mean value of all Jaccard coefficients across all 20 concepts for the Pascal VOC 2007 and across all 10 concepts for the Wang Simplicity data set is determined. As depicted in Fig. 4.2 the value decreases as the selected vocabulary sizes get more different. The underlined and bold values represent the results for the Wang Simplicity data set and the remaining values represent the results for the Pascal VOC 2007 data set. This indicates that the larger the difference in the vocabulary size, the more diverse are the result sets. The findings support a basic constraint to the assumption that different degrees of intentionality might be properly served with different amounts of clusters, as different sizes lead to distinctive results. The same holds for hierarchical clustering and hard assignment as depicted in figure 4.3 and k-means clustering with the fuzzy assignment approach during local features histogram creation 4.4.

	512	1024	1536	2048	2560	3072	3584	4096
512	1 / 1	***0.489*** / 0.358	***0.424*** / 0.327	***0.377*** / 0.302	***0.368*** / 0.280	***0.368*** / 0.264	***0.363*** / 0.254	***0.313*** / 0.247
1024		1 / 1	***0.460*** / 0.392	***0.420*** / 0.339	***0.399*** / 0.311	***0.387*** / 0.293	***0.380*** / 0.281	***0.334*** / 0.275
1536			1 / 1	***0.472*** / 0.455	***0.479*** / 0.383	***0.462*** / 0.352	***0.466*** / 0.333	***0.423*** / 0.324
2048				1 / 1	***0.503*** / 0.434	***0.494*** / 0.394	***0.470*** / 0.354	***0.449*** / 0.342
2560					1 / 1	***0.557*** / 0.476	***0.559*** / 0.402	***0.510*** / 0.379
3072						1 / 1	***0.567*** / 0.462	***0.514*** / 0.423
3584							1 / 1	***0.531*** / 0.546
4096								1 / 1

Table 4.2: Jaccard coefficients for different content-based indexes, which were produced with k-means clustering and hard assignment.

	512	1024	1536	2048	2560	3072	3584	4096
512	1 / 1	***0.761*** / 0.525	***0.690*** / 0.423	***0.633*** / 0.379	***0.602*** / 0.353	***0.559*** / 0.314	***0.543*** / 0.299	***0.457*** / 0.289
1024		1 / 1	***0.817*** / 0.646	***0.737*** / 0.571	***0.698*** / 0.528	***0.643*** / 0.456	***0.623*** / 0.433	***0.515*** / 0.414
1536			1 / 1	***0.822*** / 0.758	***0.776*** / 0.688	***0.709*** / 0.574	***0.687*** / 0.545	***0.565*** / 0.517
2048				1 / 1	***0.881*** / 0.809	***0.792*** / 0.651	***0.764*** / 0.615	***0.622*** / 0.581
2560					1 / 1	***0.842*** / 0.723	***0.809*** / 0.675	***0.651*** / 0.635
3072						1 / 1	***0.914*** / 0.844	***0.707*** / 0.786
3584							1 / 1	***0.730*** / 0.866
4096								1 / 1

Table 4.3: Jaccard coefficients for different content-based indexes, which were produced with HCL clustering and hard assignment.

	512	1024	1536	2048	2560	3072	3584	4096
512	1 / 1	***0.611*** / 0.485	***0.529*** / 0.442	***0.482*** / 0.404	***0.470*** / 0.378	***0.470*** / 0.360	***0.466*** / 0.344	***0.404*** / 0.336
1024		1 / 1	***0.579*** / 0.517	***0.537*** / 0.452	***0.521*** / 0.4202	***0.503*** / 0.400	***0.496*** / 0.382	***0.443*** / 0.375
1536			1 / 1	***0.576*** / 0.568	***0.570*** / 0.497	***0.552*** / 0.458	***0.551*** / 0.438	***0.504*** / 0.428
2048				1 / 1	***0.604*** / 0.542	***0.576*** / 0.491	***0.556*** / 0.453	***0.521*** / 0.439
2560					1 / 1	***0.651*** / 0.581	***0.648*** / 0.511	***0.588*** / 0.486
3072						1 / 1	***0.663*** / 0.564	***0.597*** / 0.527
3584							1 / 1	***0.620*** / 0.636
4096								1 / 1

Table 4.4: Jaccard coefficients for different content-based indexes, which were produced with k-means clustering and fuzzy (soft) assignment.

4.2.3 Specific information need - Near-duplicate Detection

The rationale behind the choice of a near-duplicate search task (NDD) as described in [42] is to relate specificity in a NDD task with different degrees of intentionality. Especially on the web, where a lot of images are distributed within different social networks, an NDD task is helpful in finding forged images, in order to prevent copyright violation. Finding specific altered images is correlated with a specific user need, what leads to the assumption that larger vocabularies, as stated in hypothesis H1, are likely to be more adequate than smaller vocabularies. By using a near-duplicate Detection paradigm, the interrelatedness of vocabulary size and degree of intentionality shall be shown. The NDD task consists of finding forged images, altered by applying various image transformations on an original image such as change of contrast and brightness, cropping, blurring, etc.

The experimental setup for the NDD task consisted of 8,752 test images from the PASCAL VOC 2007 image collection. 10 representative images from each of the 20 concepts of the test collection were selected[1]. Then, a series of image distortions were applied to each of the 200 "original" images to generate "distorted" versions that should ideally be returned as results when searching for near duplicates. The selected operations were:

- Brightness: Increased and decreased by 20%, 30% and 40% of the original intensity values.

[1] A list is available at http://www-itec.uni-klu.ac.at/∼mkogler/queryImages

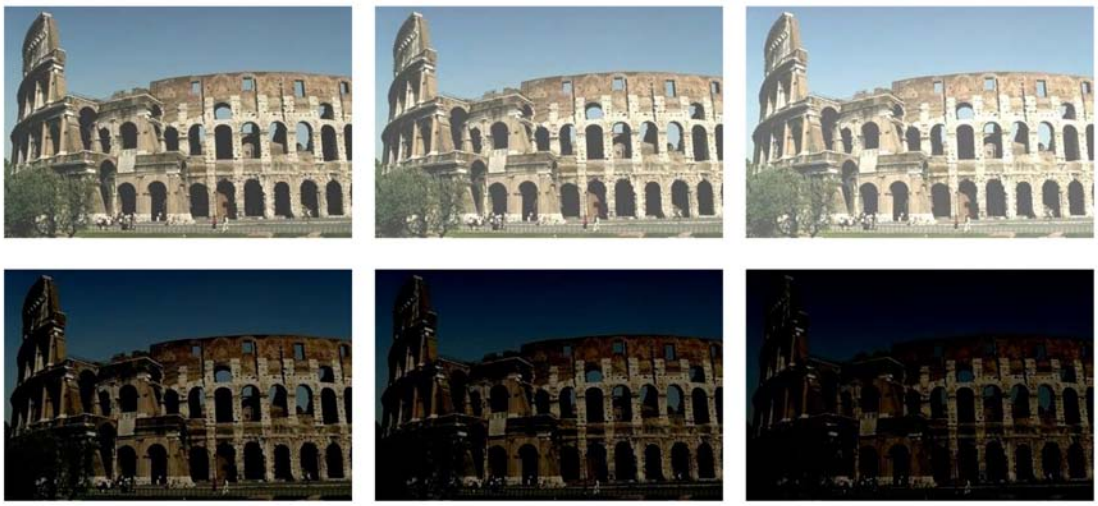

Figure 4.3: Brightness change.

- Contrast: Increased and decreased by 20%, 30% and 40% of the original intensity values.

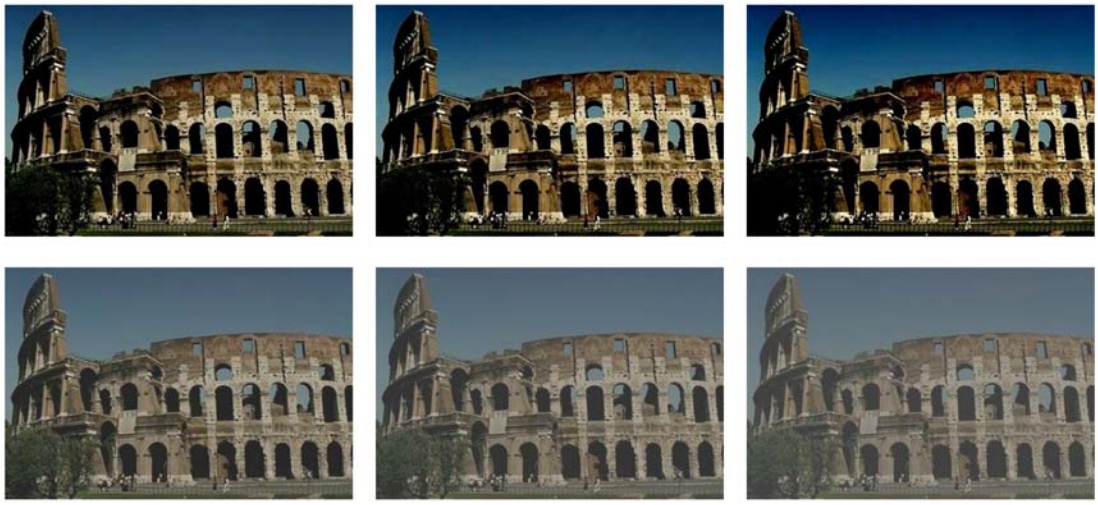

Figure 4.4: Contrast change.

- Motion blur: Changed by a standard deviation of Gaussian in pixels (10,15).

Figure 4.5: Motion blur.

- Gaussian blur: Changed by a applying a gaussian filter with a standard deviation of 2 and 4. The higher the standard deviation, the more blurred the image is.

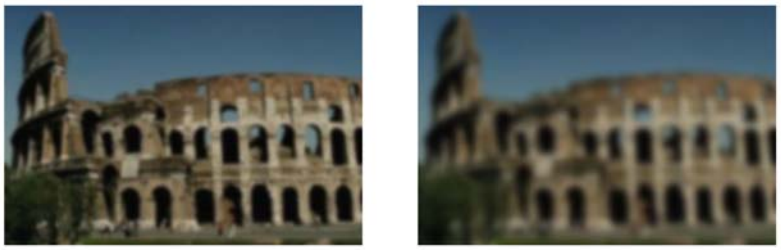

Figure 4.6: Gaussian blur.

- Cropping: Crop the image by 25%(12,5% top and bottom, 12,5% right and left), 50% and 75%

Figure 4.7: Cropped images.

The applied operations produced 3,800 near duplicates, which were added to the original test sample collection with it's 4,952 images and led to an amount of 8,752 pictures, reduced by the formerly selected 200 query images.
As for the second test bed the Wang Simplicity data set, containing 10 concepts with 100 images each, was selected. The training data collection comprised 25 pictures per concept, what accounts for an overall amount of 250 training instances for the vocabulary generation. 10 pictures of each concept were randomly taken as query images and the aforementioned 19 image distortions were conducted on each picture, which accounts for 1,900 images. Together with the remaining 900 images the 1,900 distorted versions constitute the test data collection. Various indexes based on the two test collections were created, by applying hard and soft assignment. Similarity searches were conducted, to find the distorted images to the corresponding query images. Table 4.5 shows an overview of the applied approaches and their abbreviations, which are used in the graphs.

abbreviations	clustering method	assignment method
k-means	k-means clustering	hard assignment
HCL	agglomerative hierarchical clustering	hard assignment
assignments2	k-means clustering	fuzzy (soft) assignment to 2 visual words
assignments4	k-means clustering	fuzzy (soft) assignment to 4 visual words
assignments8	k-means clustering	fuzzy (soft) assignment to 8 visual words

Table 4.5: Definitions and abbreviations for all approaches, used in the following graphs.

Figure 4.8 depicts the development of the MAP over the various vocabulary sizes for the Wang Simplicity data set. Starting from 512 visual words the MAP steadily increases for all applied approaches. This includes HCL with hard assignment, k-means with soft assignment by a fuzzy membership function labeled as assignments 2,4 and 8 in the respective figures and k-means with hard assignment.

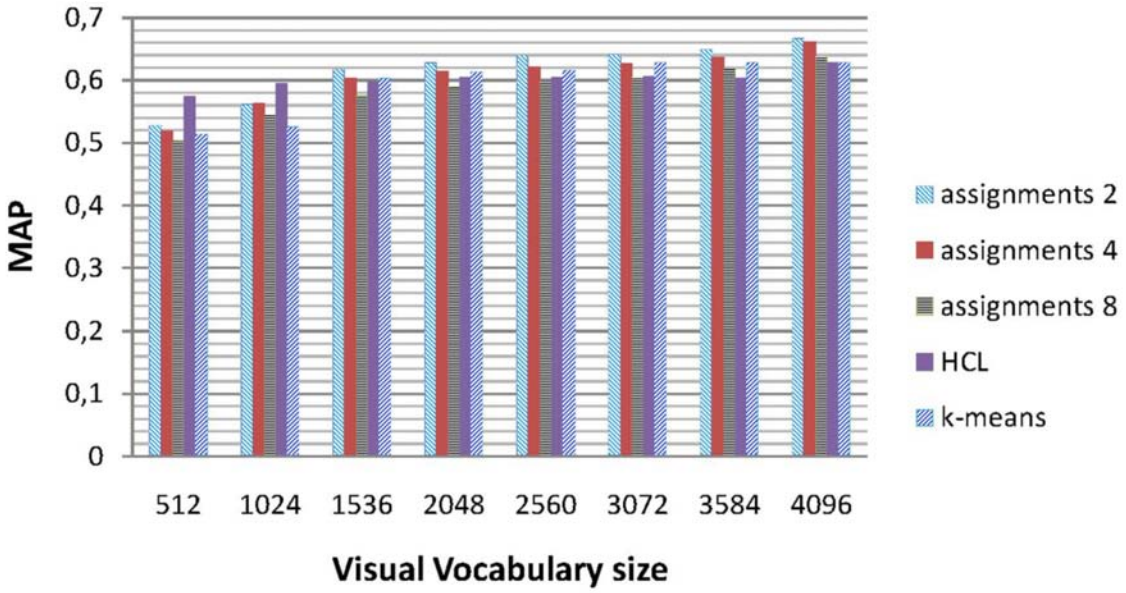

Figure 4.8: MAP of the NDD task for the Wang Simplicity data set.

For the Pascal VOC 2007 data set the MAP values increase starting from 512 visual words and reach a peak at 3072 visual words as shown in Fig. 4.9. Whereas this observation further sustains hypothesis H1, similarity searches on indexes, which were generated with the soft assignment method yield higher MAP values than similarity searches over indexes based on hard assignment. This observation holds for indexes created with 2 and 4 assignments. Similarity searches over indexes with 8 assignments, show similar retrieval performance to HCL with hard assignment and slightly worse retrieval performances than searches over indexes built with k-means and hard assignment.

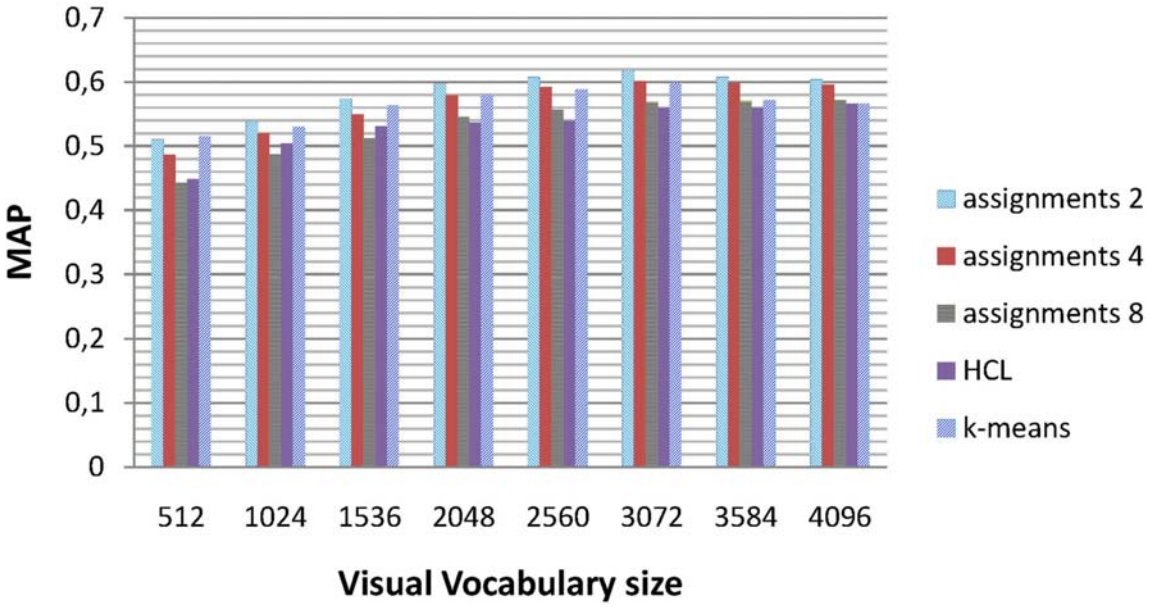

Figure 4.9: MAP of the NDD task for the Pascal VOC 2007 data set.

In both cases NDD searches on indexes produced with larger vocabularies perform better, which supports hypothesis H1.

4.2.4 Vague information need - User Simulation

The third step entailed to model the behavior of a user with a vague or less specific information need. Motivated by the random surfer model used in the motivation of the PageRank algorithm [79] a software agent was implemented, that simulated a behavioral pattern:

1. Choose a single query image I_q randomly from a concept C in the data set.
2. Using I_q find a result set R with 20 results.
3. Add all images from result set R to the set of found images R_{all}.
4. Add true positives from result set R, that feature the same concept C, to set E collecting all true positives.

5. If $|E| \geq 5$, assume that the user has found enough images and stop the search.

6. Else, randomly select a previously found image from R_{all}, use it as new I_q and go back to step 2. If either there is no previously found image that hasn't been searched for or more than 20 different query images I_q have been used, then abort.

The simulation is run 100 times per concept in the PASCAL VOC 2007 data set and the Wang Simplicity data set, resulting in 2,000 runs for the Pascal VOC 2007 data set with it's 20 concepts and 1,000 runs for the Wang Simplicity data set with it's 10 concepts. For each concept the mean number of search queries needed to find a total number of five true positives is computed. As a baseline measure a scenario was added, where 20 images from the data set were selected *randomly* as a result set (instead of searching for them based on a query).

Fig. 4.10 and Fig. 4.11 show the share of successful searches for the Wang Simplicity data set and the Pascal VOC 2007 data set regarding all respective concepts based on k-means and HCL generated vocabularies with hard and soft assignment. The figures depict that the number of conducted searches decreases, if the the vocabulary size gets smaller. Starting from an arbitrary search image, which can belong to any concept within the image collection, indexes created with smaller vocabularies, provide more adequate results sets, due to the visual diversity. Results retrieved, belong to different concepts, hence covering a greater spectrum of visual dissimilar images, more suitable for a user with a vague information need, what further support hypothesis H2. Beside this observation one can state that more assignments result in better results than less assignments in terms of average search iterations. This observation speaks for the distribution of visual information over more visual words during soft assignment, when users with a less specific information need are considered. The amount of average search iterations for indexes based on more assignments is lower than the amount of average search iterations on indexes based on less assignments. Soft assignment approaches with a distribution of visual information over 4 and 8 visual words provide better results than hierarchical clustering and k-means clustering

with hard assignment and seem to be a promising solution to retrieve more adequate pictures for users with a vague information need.

Not surprisingly, the random approach performs very well on both data sets, especially for the Pascal VOC 2007 data set, where it achieves slightly better results than the best results of the k-means algorithm. This can be reasoned with the distribution of instances over different concepts in the Pascal VOC 2007 data set, which is not uniform. Concepts like person and car contain 2,000 and nearly 800 pictures, whereas concepts like sheep comprise less than 100 images. Random searches for concepts with a larger amount of images lead to good results, since the probability of matching images belonging to such concepts is higher than for concepts with less images. Search results for the Wang Simplicity data set, where the assignment of images to concepts is more uniform, show that smaller vocabularies achieve better results than random searches. Moreover the number of average search iterations over smaller indexes and indexes generated with a soft assignment approach with a distribution over 4 and 8 visual words lead to better results.

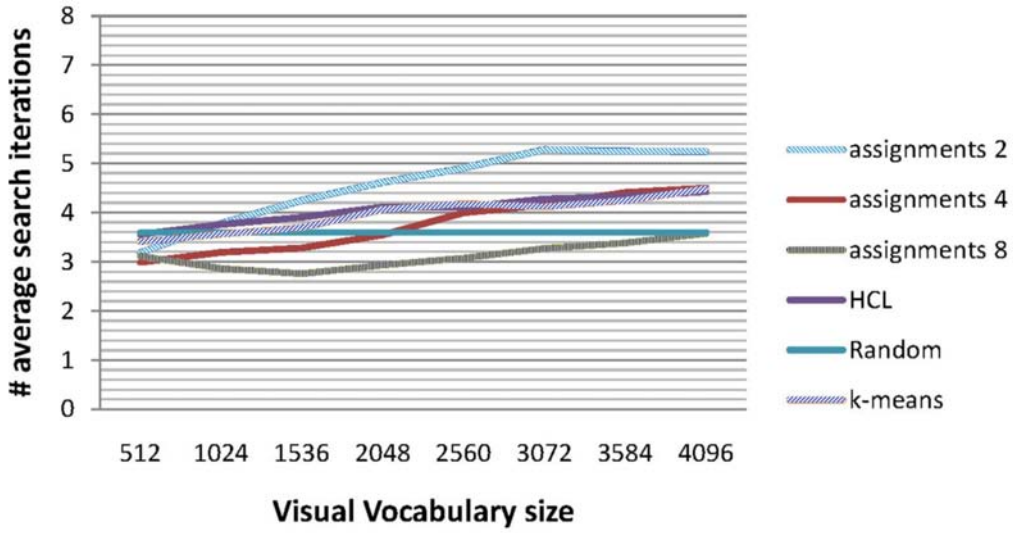

Figure 4.10: Average search iterations over all concepts for the Wang Simplicity data set.

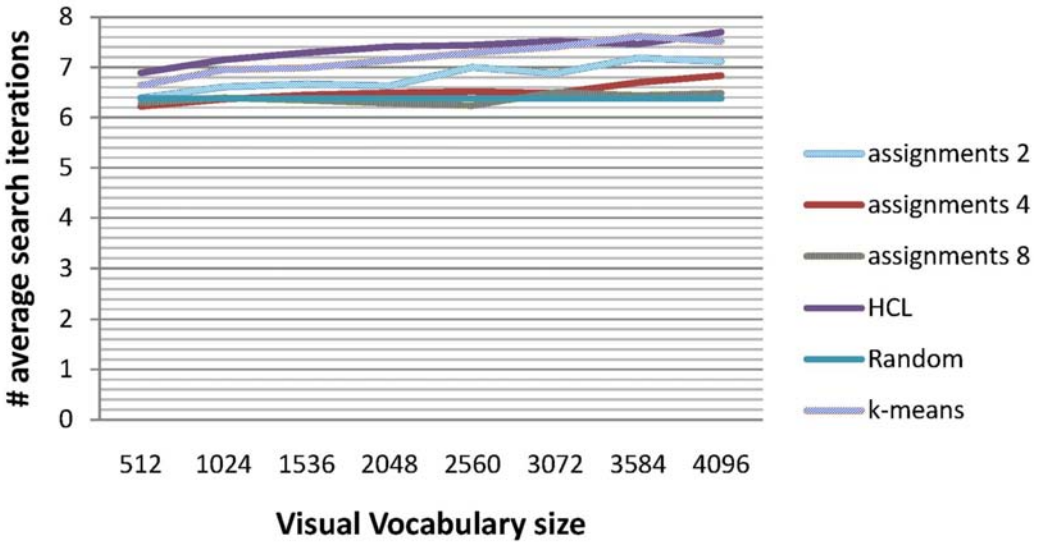

Figure 4.11: Average search iterations over all concepts for the Pascal VOC 2007 data set.

In summary, smaller vocabularies, less specific than larger ones, contain visual important information, which is used during retrieval, in order to provide a user with a vague information need with an adequate result set. Obviously search results based on larger vocabularies are not that appropriate for users with a less specific information need.

4.2.5 Discussion

The results from the NDD searches for both data sets sustain the hypothesis H1 that users with a more specific information need benefit from larger vocabularies. One can argue that larger vocabularies represent the visual content more accurately due to their fine grained visual representation. Visual information represented by local feature vectors is more accurately assigned to visual words, whereas the quantization of smaller vocabularies leads to the assignment of dissimilar feature vectors to similar visual words. Similarity searches conducted over such indexes retrieve images, which

cover a greater spectrum of visual diversified information. However a user with a more specific information need expects more accurate search results, which is expressed through the MAP values of the NDD searches and speaks for larger vocabularies.

Additionally indexes computed with the fuzzy (soft) assignment approach show higher retrieval performance in terms of MAP than hard assignment, which can be reasoned insofar, that local feature vectors not only belong to one visual word, but with a certain degree to two or more. By applying hard assignment information is lost, because the membership of local feature vectors to other visual words is neglected, whereas the soft assignment approach preserves the degree of membership. Although the results of our NDD search show, that soft assignment performs better than hard assignment, the use of more assignments seem to be counterproductive in terms of retrieval performance, when dealing with a user with a specific information need. By using more assignments, visual information is spread over more visual words, which affects the retrieval performance. More assignments yield a worse MAP.
On the contrary accurate retrieval results are not desirable for users with a somewhat vague information need. Users with vague information needs benefit from more visual diversified result lists, which speaks for the use of small vocabularies. Smaller visual vocabularies contain more visual diversified information, because visual dissimilar local feature vectors are assigned to similar visual words, when dealing with more coarse-grained visual vocabularies. Considering the results of the user simulation, where a search starts from an arbitrary search image, searches on indexes produced with smaller vocabularies lead to less iterations. The retrieved images are visually more dissimilar and fit users with a less specific information need better, because the users can chose among a broader spectrum of visual dissimilar images.

The soft assignment approach exposes that the distribution of one local feature vector over more visual words performs quite well. The results of the user simulation show that users with a vague information need benefit from indexes, which were generated with soft assignments over more visual words. Due to a higher distribution of visual information over various visual words, visually dissimilar representations of

images are preserved. During a search more visual dissimilar images are retrieved, which is advantageous for a user with a vague information need.

4.3 Summary

A new approach for retrieval adaptation in context of user intentions was introduced in this chapter. Bag of Visual Words techniques were applied in order to compute visual vocabularies and content-based indexes, which differed in their visual granularity, when describing the visual content. The granularity was related to users following a specific information need and users with a less specific more vague information need. Various tests were conducted, showing that users with a more specific information need benefit from larger vocabularies and more fine grained content-based indexes. On the other hand users with a more vague information need profit from smaller vocabularies and coarse grained content-based indexes.

Chapter 5

Conclusion

Visual information retrieval is a broad field, comprising various research areas in computer science such as human-computer interaction, computer vision, information retrieval and so forth. This work deals with content-based techniques, which are applied in visual information retrieval to analyze, organize and search multimedia related content. Low level features were utilized to re-organize videos, which had been retrieved by text-based searches. The selection of low level features comprised state of the art visual features, such as color, texture and motion features. The Bag of Visual Words approach was applied and extended by regarding motion related information. The findings show that a decent annotation of video clips paves the way for advanced text-based search methods. Content-based search approaches are needed in case of badly annotated videos, in order to improve the recall. Especially searches in video collections with a sparse amount of meta data can benefit from a combination of text and visual search methods.

Moreover automatic video summarization by color, texture and local features, using the Bag of Visual Words approach, was performed and evaluated. Video summaries were automatically generated for short video clips by means of different visual features and presented to participants of a user study. The users had to compare the generated summaries qualitatively and had to weight their appropriateness to describe short video clips. The outcomes of the user study indicate that local features

together with Bag of Visual Words cannot outperform approaches based on global features for the investigated type of video summaries. Nevertheless the BoVW approach provides the most stable results for the test data collection of different video clips. This speaks for its usage in different domains including user captured single shot videos and medical videos.
Furthermore a fuzzy approach for visual vocabulary generation and fuzzy assignment was introduced. The rational behind the application of fuzzy approaches for vocabulary generation as well as histogram creation was to achieve more accurate and more robust retrieval results over different vocabulary sizes. The findings of the experiments show that fuzzy approaches lead to better retrieval results than the state of the art BoVW approach. Smaller vocabularies benefit from fuzzy approaches, which paves the way for a reduction of dimensionality of descriptors. While more compact descriptors can be computed, the retrieval performance does not suffer that much.
There is still room for further research regarding Bag of Visual Words, despite it has gained much attention over the past years. Especially the vocabulary generation and the local feature assignment as well as the choice of local features themselves can still be improved to guarantee higher recall rates and precision scores.
Although visual information retrieval has gained importance, it still has to find its way into common search and retrieval systems, which mostly rely on text-based information. Nevertheless visual search can act as a complementary paradigm and step in for text-based searches, where meta information such as tags hardly exist.

By solely regarding visual information, while searching for images and videos, the user as an important information source is neglected. Users express an information need in order to reach a certain goal or intent during their search and retrieval session. The goal a user tries to achieve, can be measured and classified manually or automatically into a pre-defined model, representing the intent of a user during search.
Depending on the user search intent, retrieval mechanisms were adapted in this work. Therefore the Bag of Visual Words approach was leveraged and examined, by relating the granularity of the visual content to different information needs.

Tests show that users with a more specific information need (e.g. get me images of collies) benefit from larger vocabularies and more fine grained content-based indexes. On the contrary users with a more vague information need (e.g. get me images of different dog breeds) benefit from smaller vocabularies and coarse grained content-based indexes. The presented approach is only a first step towards the linkage of user intentions and visual information retrieval. There is still a long way to go, in order to promote user oriented search to the next level.

Bibliography

[1] L. Aimar, L. Merritt, E. Petit, Chen M., J. Clay, M. Rullgard, R. Czyz, C. Heine, A. Izvorski, and Wrigh. x264 - a free h264/avc encoder. Online, 2010.

[2] Yousef Alqasrawi, Daniel Neagu, and Peter Cowling. Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification. In *Signal, Image and Video Processing*, 2011.

[3] Linda H. Armitage and Peter G.B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.

[4] Ricardo Baeza-Yates, Liliana Calderon-Benavides, and Cristina Gonzalez-Caro. The intention behind web queries. In *String Processing and Information Retrieval*, pages 98–109. Springer Berlin / Heidelberg, 2006.

[5] Ricardo A. Baeza-Yates, Carlos A. Hurtado, and Marcelo Mendoza. Query recommendation using query logs in search engines. In Wolfgang Lindner, Marco Mesiti, Can Tuerker, Yannis Tzitzikas, and Athena Vakali, editors, *EDBT Workshops*, volume 3268 of *Lecture Notes in Computer Science*, pages 588–596. Springer, 2004.

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[7] Alberto Del Bimbo. *Visual Information Retrieval*. Number 1-55860-624-6. Kaufmann, 1999.

[8] Andrei Broder. A taxonomy of web search. volume 36, pages 3–10, New York, NY, USA, September 2002. ACM.

[9] R. L. Cannon, J. V. Dave, and J. C. Bezdek. Efficient implementation of the fuzzy c-means clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8:248–255, 1986.

[10] S. A. Chatzichristofis, A. Arampatzis, and Y. S. Boutalis. Investigating the behavior of compact composite descriptors in early fusion, late fusion and distributed image retrieval. *Radioengineering*, 19(4):725–733, 2010.

[11] S.A. Chatzichristofis, Y.S. Boutalis, and Mathias Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Signal Processing, Pattern Recognition and Applications (SPPRA 2009)*, 2009.

[12] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Cedd: Color and edge directivity descriptor. a compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer Vision Systems*, pages 312–322. Springer-Verlag, 2008.

[13] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, WIAMIS '08, pages 191–196, Washington, DC, USA, 2008. IEEE Computer Society.

[14] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Content based radiology image retrieval using a fuzzy rule based scalable composite descriptor. *Multimedia Tools Appl.*, 46(2-3):493–519, January 2010.

[15] Zheng Chen, Fan Lin, Huan Liu, Yin Liu, Wei-Ying Ma, and Liu Wenyin. User intention modeling in web applications using data mining. volume 5, pages 181–191, Hingham, MA, USA, November 2002. Kluwer Academic Publishers.

[16] Ingemar J. Cox, Matthew L. Miller, Thomas P. Minka, Thomas V. Papathomas, and Peter N. Yianilos. The bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, 2000.

[17] J. Cui, Fang Wen, and Xiaoou Tang. Intentsearch: Interactive on-line image search re-ranking. In *MULTIMEDIA '08: Proceedings of the 16th international conference on Multimedia*. ACM, 2008.

[18] J. Cui, Fang Wen, and Xiaoou Tang. User intention modeling for interactive image retrieval. In *IEEE International Conference on Multimedia & Expo, International Workshop on Visual Content Identification and Search (VCIDS)*, 2010.

[19] Jingyu Cui, Fang Wen, and Xiaoou Tang. Real time google and live image search re-ranking. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 729–732, New York, NY, USA, 2008. ACM.

[20] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40, 2008.

[21] Thomas Deselaers, Daniel Keysers, and Hermann Ney. Features for image retrieval: an experimental comparison. *Inf. Retr.*, 11:77–107, April 2008.

[22] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. Bag-of-visual-words models for adult image classification and filtering. In *ICPR*, pages 1–4. IEEE, 2008.

[23] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. Understanding the relationship between searchers' queries and information goals. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, pages 449–458, New York, NY, USA, 2008. ACM.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[25] Jianping Fan, Daniel A. Keim, Yuli Gao, Hangzai Luo, and Zongmin Li. Justclick: personalized image recommendation via exploratory search from large-scale flickr images. volume 19, pages 273–288. Institute of Electrical and Electronics Engineers Inc., The, February 2009.

[26] David Dagan Feng. *Multimedia information retrieval and management*. Number 978-3-540-00244-4. Springer, 2003.

[27] Jan C. Gemert, Jan-Mark Geusebroek, Cor J. Veenman, and Arnold W. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 696–709, Berlin, Heidelberg, 2008. Springer-Verlag.

[28] T. Gevers and A.W.M. Smeulders. Pictoseek: combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, 2000.

[29] Yousef Hadi, Fedwa Essannouni, and Rachid Oulad Haj Thami. Video summarization by k-medoid clustering. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 1400 – 1401. ACM, 2006.

[30] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proceedings of Alvey Vision Conference*, 1988.

[31] Kuan-Yu He, Yao-Sheng Chang, and Wen-Hsiang Lu. Improving identification of latent user goals through search-result snippet classification. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 683–686, Washington, DC, USA, 2007. IEEE Computer Society.

[32] Hsin-liang Chen. An analysis of image retrieval tasks in the field of art history. *Inf. Process. Manage.*, pages 701–720, 2001.

[33] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Transactions of Information Theory*, 8:179 – 187, 1962.

[34] Gang Hua and Qi Tian. What can visual content analysis do for text based image search? In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME'09, pages 1480–1483, Piscataway, NJ, USA, 2009. IEEE Press.

[35] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 762–, Washington, DC, USA, 1997. IEEE Computer Society.

[36] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31:264–323, September 1999.

[37] Ramesh Jain and Pinaki Sinha. Content without context is meaningless. In Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, editors, *ACM Multimedia*, pages 1259–1268. ACM, 2010.

[38] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1149–1150, New York, NY, USA, 2007. ACM.

[39] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. volume 44, pages 1251–1266, Tarrytown, NY, USA, May 2008. Pergamon Press, Inc.

[40] Yu-Gang Jiang and Chong-Wah Ngo. Bag-of-visual-words expansion using visual relatedness for video indexing. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 769–770, New York, NY, USA, 2008. ACM.

[41] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 494–501, New York, NY, USA, 2007. ACM.

[42] Y. Ke, R. Sukthankar, and L. Huston. An efficient parts-based near-duplicate and sub-image retrieval system. In *MULTIMEDIA '04*, pages 869–876, New York, NY, 2004. ACM.

[43] Kraisak Kesorn and Stefan Poslad. An enhanced bag-of-visual word vector space model to represent visual content in athletics images. *IEEE TRANSACTIONS ON MULTIMEDIA*, 14:211–222, 2012.

[44] Hyeon Jun Kim and Jin Soo Lee. Method for quantizing colors using hue , min, max, difference (hmmd) color space. European Patent Application, 1999.

[45] Christoph Kofler and Mathias Lux. An exploratory study on the explicitness of user intentions in digital photo retrieval. In *In Proceedings of th International Conference on Knowledge Management (I-KNOW 09)*, Graz, Austria, Sept. 2009.

[46] Marian Kogler. Video-based sight distance and speed computation for traffic surveillance. Technical report, Klagenfurt University (Alpen-Adria Universität Klagenfurt), Tranportation Informatics, 2012.

[47] Marian Kogler, Manfred del Fabro, Mathias Lux, Laszlo Boeszoermenyi, and Klaus Schoeffmann. Global vs. local feature in video summarization: Experimental result. In *Proceedings of the 10th International Workshop of the Multimedia Metadata Community on Semantic Multimedia Database Technologies (SeMuDaTe'09) in conjunction with the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009)*, 2009.

[48] Marian Kogler and Mathias Lux. Bag of visual words revisited: an exploratory study on robust image retrieval exploiting fuzzy codebooks. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 3:1–3:6, New York, NY, USA, 2010. ACM.

[49] Marian Kogler and Mathias Lux. Pursuing the holy grail by interrelating user intentions and bag of visual words to perform retrieval adaptation. In *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*, SBNMA '11, pages 3–8, New York, NY, USA, 2011. ACM.

[50] Marian Kogler and Mathias Lux. Robust image retrieval using bag of visual words with fuzzy codebooks and fuzzy assignment. submitted to the 12th International Conference on Knowledge Management and Knowledge Technologies (accepted), 2012.

[51] Marian Kogler, Mathias Lux, and Oge Marques. Adaptive visual information retrieval by changing visual vocabulary sizes in context of user intentions. In *In Proceedings of the Workshop on Multimedia on the Web (MMWeb) 2011*. IEEE, 2011.

[52] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[53] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 391–400, New York, NY, USA, 2005. ACM Press.

[54] Teng Li, Tao Mei, In-So Kweon, and Xian-Sheng Hua. Contextual bag-of-words for visual categorization. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, 21:381–392, 2011.

[55] Weitao Li, Xiaojie Zhou, and Tianyou Chai. Bag of visual words and latent semantic analysis-based burning state recognition for rotary kiln sintering process. In *Control and Decision Conference (CCDC), 2011 Chinese*, 2011.

[56] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 339–346, New York, NY, USA, 2008. ACM.

[57] Bing Liu, Yiyuan Xia, and Philip S. Yu. Clustering through decision tree construction. In *Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, pages 20–29, New York, NY, USA, 2000. ACM.

[58] Fang Liu, Clement Yu, and Weiyi Meng. Personalized web search for improving retrieval effectiveness. *IEEE Trans. on Knowl. and Data Eng.*, 16:28–40, January 2004.

[59] David G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999.

[60] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110., 2004.

[61] Mathias Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In *MM '08: Proceeding of the 16th ACM international conference*, 2008.

[62] Mathias Lux, Christoph Kofler, and Oge Marques. A classification scheme for user intentions in image search. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, pages 3913–3918, New York, NY, USA, 2010. ACM.

[63] Mathias Lux, Oge Marques, Klaus Schöffmann, Laszlo Böszörmenyi, and Georg Lajtai. A novel tool for summarization of arthroscopic videos. volume 46, pages 521–544, Hingham, MA, USA, January 2010. Kluwer Academic Publishers.

[64] Mathias Lux, Klaus Schoeffmann, Manfred del Fabro, Marian Kogler, and Mario Taschwer. Itec-uniklu known-item search submission. In *TRECVID*, 2010.

[65] Marjo Markkula and Eero Sormunen. End-user searching challenges indexing practices inthe digital newspaper photo archive. *Inf. Retr.*, 1:259–285, January 2000.

[66] Oge Marques. *Practical Image and Video Processing Using MATLAB*. Number 978-0470048153. John Wiley & Sons, 2011.

[67] Oge Marques and Borko Furht. *Content-based image and video retrieval*. Number 1-4020-7004-7. Kluwer Academic, 2002.

[68] Sharon McDonald and John Tait. Search strategies in content-based image retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 80–87, New York, NY, USA, 2003. ACM.

[69] Elaine Menard. Search behaviours of image users: A pilot study on museum objects. *Canadian Journal of Library and Information Practice and Research*, 6, 2011.

[70] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65:43–72, November 2005.

[71] Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. Local features for object class recognition. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, 2005.

[72] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 27:1615 – 1630, 2005.

[73] Frank Moosmann, Eric Nowak, and Frederic Jurie. Randomized clustering forests for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1632–1646, September 2008.

[74] Henning Mueller, Christelle Despont-Gros, William Hersh, Jeffery Jensen, Christian Lovis, and Antoine Geissbuhler. Health care professionals' image use and search behaviour. In *Proceedings of Medical Informatics Europe (MIE 2006)*, Maastricht, Netherlands, 2006.

[75] Henning Mueller, Wolfgang Mueller, David McG. Squire, Stephane Marchand-Maillet, and Thierry Pun. Long-term learning from user behavior in content-based image retrieval. Technical Report 00.04, University of Geneva, 2000.

[76] Henning Müller, Thierry Pun, and David Squire. Learning from user behavior in image retrieval: Application of market basket analysis. volume 56, pages 65–77, Hingham, MA, USA, January 2004. Kluwer Academic Publishers.

[77] Carlton W. Niblack, Ron Barber, Will Equitz, Myron D. Flickner, Eduardo H. Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, and Gabriel Taubin. Qbic project: querying images by content, using color, texture, and shape. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pages 173–187, San Jose, 1993.

[78] Eric Nowak, Frederic Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *In Proc. ECCV*, pages 490–503. Springer, 2006.

[79] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[80] Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[81] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *Int. J. Comput. Vision*, 18:233–254, June 1996.

[82] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '01, pages 190–197, New York, NY, USA, 2001. ACM.

[83] Kerry Rodden and Kenneth R. Wood. How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, pages 409–416, New York, NY, USA, 2003. ACM.

[84] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM.

[85] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *In European Conference on Computer Vision*, pages 430–443, 2006.

[86] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23:309–314, August 2004.

[87] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[88] Klaus Schoeffmann, Mathias Lux, Mario Taschwer, and Laszlo Boeszoermenyi. Visualization of video motion in context of video browsing. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME'09, pages 658–661, Piscataway, NJ, USA, 2009. IEEE Press.

[89] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, CVPR*, pages 593–600. IEEE Computer Society, 1994.

[90] Thomas Sikora. The mpeg-7 visual standard for content description-an overview. *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, 11:696 – 702, 2001.

[91] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477. IEEE Computer Society, 2003.

[92] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1349–1380, December 2000.

[93] J.R. Smith and S-F. Chang. *Intelligent Multimedia Information Retrieval*, chapter 2, pages 23–41. American Assiciation for Artificial Intelligence, 1997.

[94] de Rooij O. Huurnink B. van Gemert J C. Uijlings J R R. He J. Li X. Everts I. Nedovic V. van Liempt M. van Balen R. Yan F. Tahir M A. Mikolajczyk K. Kittler J. de Rijke M. Geusebroek J M. Gevers T.Th. Worring M. Smeulders A W M. Koelma D C. Snoek C G M., van de Sande K E A. The mediamill trecvid 2008 semantic video search engine. In *TRECvid 2008 Working Notes*. NIST, 2008.

[95] M. Stricker and M. Orengo. Similarity of color images. In *In Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1995.

[96] M. Strohmaier, M. Kröll, and C. Körner. Intentional query suggestion: Making user goals more explicit during search. In *Proceedings of the Workshop on Web Search Click Data WSCD'09, in conjunction with WSDM 2009*, Barcelona, Spain, 2009. ACM.

[97] Markus Strohmaier, Peter Prettenhofer, and Mathias Lux. Different degrees of explicitness in intentional artifacts - studying user goals in a large search query log. In *CSKGOI'08 International Workshop on Commonsense Knowledge and Goal Oriented Interfaces, in conjunction with IUI'08*, Canary Islands, Spain, 2008.

[98] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transaction on*, 8:460–473, 1978.

[99] Bart Thomee, Erwin M. Bakker, and Michael S. Lew. Top-surf: a visual words toolkit. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1473–1476, New York, NY, USA, 2010. ACM.

[100] Jasper R. R. Uijlings, Arnold W. M. Smeulders, and Remko J. H. Scha. Real-time bag of words, approximately. In Stephane Marchand-Maillet and Yiannis Kompatsiaris, editors, *CIVR*. ACM, 2009.

[101] Jasper R. R. Uijlings, Arnold W. M. Smeulders, and Remko J. H. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665–681, 2010.

[102] Corinne Vachier and Fernand Meyer. The viscous watershed transform. *J. Math. Imaging Vis.*, 22:251–267, May 2005.

[103] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[104] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (9):947 – 963, 2001.

[105] Xin Wang, Sanda Erdelez, Yunhui Lu, Carla Allen, Blake Anderson, Hongfei Cao, and Chi-Ren Shyu. Search tactics for medical image retrieval. In *Proceedings of the American Society for Information Science and Technology*, volume 48, pages 1–4, 2011.

[106] Chee Sun Won, Dong Kwon Park, and Soo-Jun Park. fficient use of mpeg-7 edge histogram descriptor. *ETRI Journal*, 24:23–30, 2002.

[107] Lei Wu, Steven C.H. Hoi, and Nenghai Yu. Semantics-preserving bag-of-words models for efficient image annotation. In *Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining*, LS-MMRM '09, pages 19–26, New York, NY, USA, 2009. ACM.

[108] Tang Xiaoou, Liu Ke, and Cui Jingyu. Intentsearch: Capturing user intention for one-click internet image search. *IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 6:13, 2011.

[109] Hong (Iris) Xie. Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management*, 38:55–77, 2002.

[110] Sheng Xu, Toa Fang, Deren Li, and Shiwei Wang. Object classification of aerial images with bag-of-visual words. *Geoscience and Remote Sensing Letters, IEEE*, 7:366 – 370, 2010.

[111] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, MIR '07, pages 197–206, New York, NY, USA, 2007. ACM.

[112] Konstaninos Zagoris, Savvas A. Chatzichristofis, and Avi Arampatzis. Bag-of-visual-words vs global image descriptors on two-stage multimodal retrieval. In *SIGIR '11 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011.