Unterschrift (Betreuer)

MASTERARBEIT

Local Colour Features for Image Retrieval

ausgeführt am Institut für
Rechnergestützte Automation,
Arbeitsgruppe für Mustererkennung und Bildverarbeitung
der Technischen Universität Wien

unter der Anleitung von **Dr. Allan Hanbury**

durch

Julian Stöttinger
Außerpühret 34
4661 Roitham

27. April 2007

Unterschrift (Student)

Acknowledgements

All my thanks go out to my calm and considerate supervisor Dr Allan Hanbury and my supporters, the lively and communicative Prof Dr Nicu Sebe and the stylish and profound Prof Dr Theo Gevers for bringing me so far and always bearing with me. Now I like science. After all, you guys changed my life.
I am genuinely grateful to my parents Claudia and Toni for making my studies possible and always believing in me.
Special thanks to Andrea, for all the good times.

Julian Stöttinger

Contents

1 Introduction

The need for automatic image content analysis is increasing with the amount of digital images in private or business collections. This thesis deals with the automatic classification of image content. As the human visual perception builds the basis for an automatic approach, Section 1.1 gives an introduction to this field of science. Section 1.2 introduces a machine vision counterpart and the main research field of this work, content-based image retrieval (CBIR). This research focusses on the local features of images for content-based retrieval, which are outlined in Section 1.3.

1.1 Human Visual Perception

Primates obtain important information about their surroundings through visual information. Seeing, understanding and interpreting one's subjective environment is a fascinating process covering several fields of science. It is one of the rare connectors of humanities and science, where many different fields have their own theories. Probably beginning with ancient philosophers (see Socrates' definition of shapes [73]), visual understanding has always been a challenging field of investigation, including religious questions in former centuries as well as neurobiological research in the last one. With the invention of photography, the comparison between the human eye and optical devices arose, as shown in Figure 1.

Distinguishing and recognizing different objects is of great importance in the mental abstraction process. In many biological vision systems this process starts at the retina. The retina converts an electromagnetical stimulus of a light beam using the rods and cones into nerve activity. Besides registering the intensity and the wavelengths of the light, the retina also performs a multi-scale pre-processing of the elementary features present in the scene. These elementary features are for example blobs, bars and edges [38, 29].

In computer science, the retinal processes are simulated using digital information visualized as digital images. Pixel information is interpreted into light intensity or colour information, and image processing tries to extract early visual features to simulate the abilities of the retina. In the computer an image consisting of pixels can be seen as a snapshot of the electromagnetic stimulus of the eye at a certain moment. The first step is to derive the basic features from the image. At this stage, where no semantic information is present, there is no actual understanding of the scene taking place in the human visual perception. Nevertheless, some spots of a scene are regarded to be more important than other ones, even without prior knowledge [33].

This is one connection to a machine's task: There is no understanding of a visual scene happening when focussing towards the important parts of a scene. Therefore, it could be possible, that computers simulate this early feature extraction. From the information provided by the retina, the elementary features are grouped together into units that make up the objects. This is not fully understood, but we tend to do this according to a set of principles called "the gestalt principles of organization" (e.g. [94, 62, 28]), which are probably innately specified [26]. These steps can be seen as the syntactic steps.

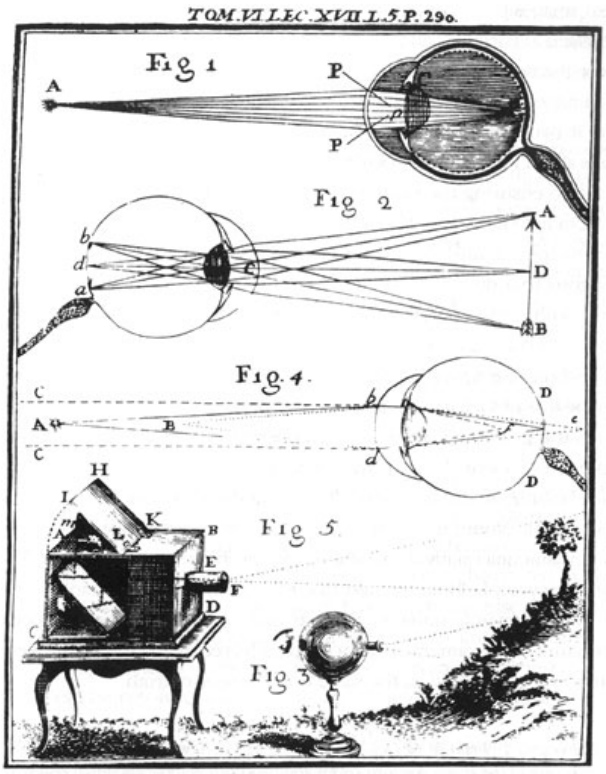


Figure 1: The comparison of the human eye and camera has been very obvious at a certain scale and point of view. From [4].

The next step is to identify the objects and associate a meaning with them; this can be considered the semantic step. The principles behind this process are not clearly understood although theories exist. The two dominant theories in this field are *template matching* (e.g. [68]) and *feature analysis* (e.g. [42]). In template matching, the object as perceived is compared to a model of the object. Based on the differences and similarities the approach judges whether the perceived object falls into the same category as the model or not. In this context, templates are often two dimensional.

Feature analysis theory covers the approaches in which the percieved data is segmented into elementary components. In this work, we will focus on this theory. The basic idea behind it is that an object can be interpreted as a configuration of known, simple objects. This means a hierarchical approach towards recognition, where the hierarchy is often related to the scale of inspection with respect to a primate's high abstraction and understanding possibilities it can also be a hierarchy of interest.

In [5] a theory is proposed in which objects are segmented into basic geometrical shapes like cones, cylinders, cubes, etc. which are called *geons*. The mutual positioning of the geons defines a pattern that can be matched to known patterns. However, at any time there must be a matching between the perceived data and prior knowledge. In template matching, we typically encounter a two dimensional template matched with the perceived data. In a feature analysis based approach, multiple high dimensional features are compared to prior features. The more

features of one object match, the more certain is the recognition.

All of the steps described above are performed at multiple scales. Based on their global structures, classes of objects can be recognized. At a finer scale, specific instances of an object class can be distinguished. This kind of multiscale analysis can be used to derive a hierarchical structure which abstracts the scene.

Another hierarchy that can be obtained by multi-scale analysis is collecting multiple objects into one. For example, if a large number of trees is perceived from a large distance, the individual trees cannot be distinguished. We recognize the trees as a forest. Moving closer allows us to observe the forest at a finer scale at which individual trees can be distinguished. When observing the trees at an even finer scale, leaves can be recognized. The basic idea behind these observations is that different things are observed, depending on the scale at which it is perceived. Note that this idea is not just restricted to object recognition. It applies also to other areas of science; for example when Mandelbrot wondered how long the coastline of Britain is [51].

With this biological model we are well equipped to do some computerization of the process, which can be used in many settings. In this research, the focus is on the classification and retrieval of image data in a large database. In the following, the task of content-based image retrieval is described.

1.2 Content - Based Image Retrieval (CBIR)

The term of CBIR first appeared in a paper by T. Kato [34] to describe automatic retrieval of images from a database based on visual features, such as colour and shape. The first commercial CBIR system, QBIC[1] (query by image content) was designed by IBM in the early 1990s.

In many areas of our daily life we encounter large collections of digital images. In various fields of commerce, government, academia, hospitals, and the media content creation business, a huge amount of images is being created. *Content-based image retrieval* is the task of retrieving relevant visual data from a large collection on the basis of a user query. Keyword archiving and keyword searching has been the usual way of indexing and retrieval, with their known advantages and drawbacks: Manual keyword indexing takes a lot of time for the actual indexing, but leads to a fast and reliable retrieval. Automatic keyword indexing on the other hand would be very fast [44]. Unfortunately, these automatic approaches tend to be inaccurate.

However, the possibilities for these annotations are always in predefined boundaries. If a needed property of the data has not been described, a retrieval based on this property is not possible. A good example is the use of image retrieval in medical databases. Using image data of other patients for diagnosis, the ways of querying will differ from case to case. Moreover, the important properties of the image data is not a priori known. The staging of a patient's therapy or the retrieval of images with similar diagnostic findings in large electronic archives is

[1] http://wwwqbic.almaden.ibm.com/

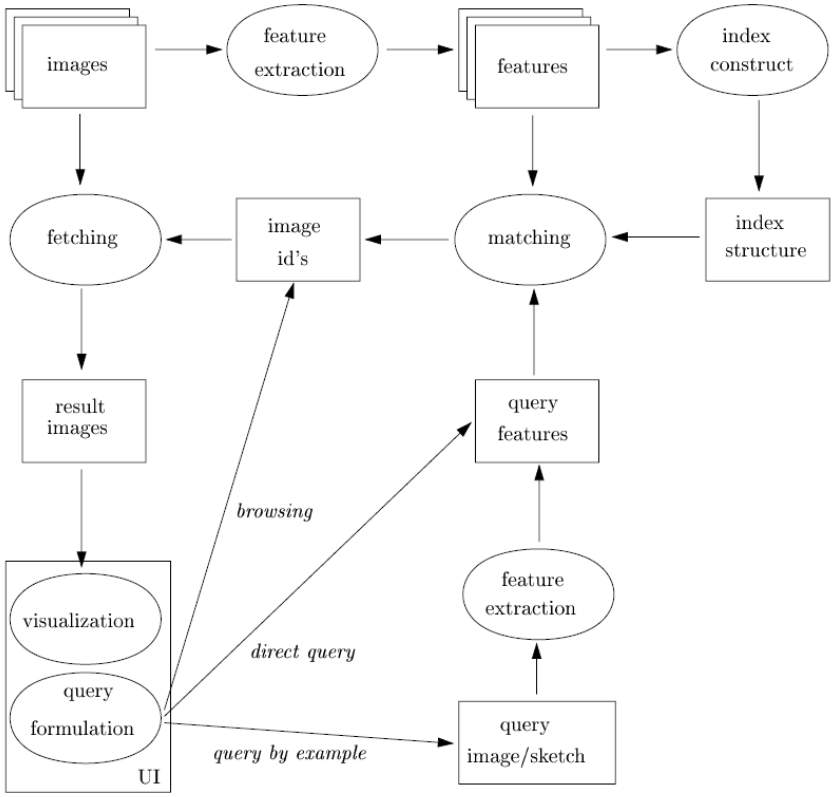


Figure 2: Content based image retrieval framework. From: [92]

a very diverse task. Additionally, the indexing process needs medical expert knowledge and is therefore very rare and expensive [40].

Computers do not understand the actual content of the image, but still there are some tasks that can be performed by machines. One way to achieve a retrieval of related data in huge databases is to use a *query image* as input. The user may also generate this image on the fly when drawing a sketch into the application itself. The goal is the computation of image features within images, and the use of those features as query terms. The basic components of a content based image retrieval system are shown in Figure 2. A survey of existing systems can be found in [13, 92].

Concerning personal digital image collections, there has been a rising interest in content-based query systems. The typical amount of images a contemprorary digital camera user owns has increased dramatically in the last years. Most of the time, the users neglect a proper organization or annotation of their personal image collections. When there is no proper description of the images besides the one provided by the camera itself (e.g. date, exposure time, focal length, aperture, etc.), queries based on the content would be the most appropriate way to search for images in private collections [65]. In the following, the three main ways of querying are described.

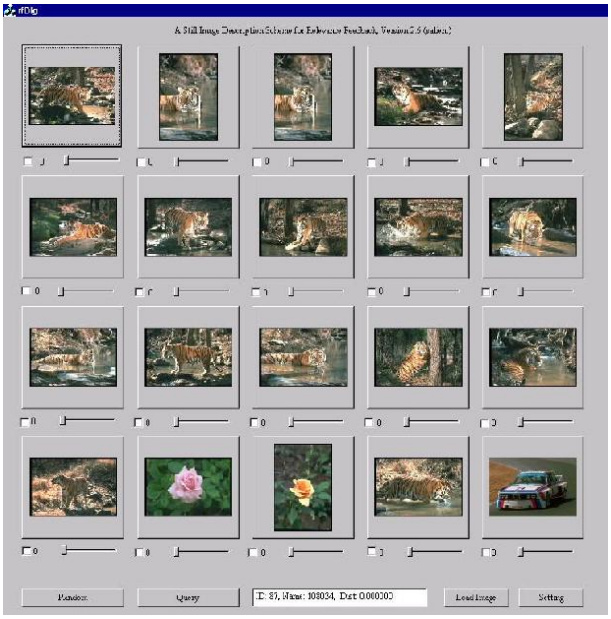


Figure 3: Query by example image, one way of image retrieval. The query image (top left) is here compared to the database based on global colour features. From: [23]

Query by example uses the users' chosen image as the basis for the retrieval process. A screen shot of a query and its results is given in Figure 3. Perhaps the most successful and famous applications are presented in [10, 75, 76]. Similarity is measured between the query image and images in the database. Relevant to this work, the approaches are often based on local low level features, and therefore these features are extracted from the query image and then compared to the local features of all other images. A ranking of these similarities provides then a list of the most probable matches. In this approach a ranked list of images is returned by the system. The user gets a small, convenient collection of images from which to choose.

Query by sketch The need of an example image can be a drawback. The approach from [14] overcomes this constraint. The user draws a sketch of the image or a colour distribution into the application, and the query is carried out with that sketch (see Figure 4). A sketch often consists of a multitude of overlapping lines which give a rough structure of the desired image. Colour distributions contain colourful areas to roughly match the resulting images. Providing a good sketch is nevertheless not always easy. Additionally, it makes no sense to compute local features from these sketches, as they contain only rough colour and structure information.

Semantic retrieval From the users' point of view, probably the most convenient way for retrieval would involve a *direct query*, also referred to as *semantic retrieval*. The user makes a written request about the content of the image, and the system retrieves relevant images. For current implementations, this is very hard to achieve.

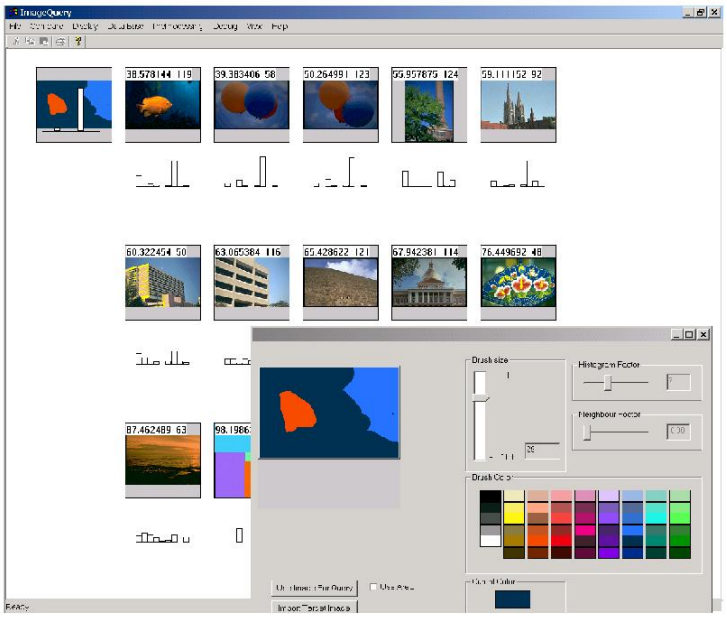


Figure 4: Query by sketch. The user draws a sketch of an image, or gives a colour distribution. The CBIR system compares it to the image database and gives results. From: [53]

1.3 Overview of the Thesis

Regarding the local features for image classification, this work focuses on the region extraction. This is the first stage in gathering local features of images, and it is broadly accepted that it is crucial for a usable description of the images. Focusing on features for colourful objects, the use of colour can have a distinct advantage in localizing the salient points of the image.

Various saliency based approaches that try to simulate the primate's visual cortex are reviewed. The review leads to a combination of the broadly established corner detection and a salient based approach: Corners provide very stable interest points, and corners of colour images can improve their distinctiveness. A statistical analysis of the colour distribution and the use of different colour spaces adapted to primate perception connect the fields of visual saliency and illumination based interest point detection to one stable saliency measurement.

All the recognition tasks of primates work at different scales. A new method for the automatic determination of the characteristic scale of a region is proposed. The scale determination is separated from the actual corner extraction, where the colour is transformed to a one dimensional distribution to provide another "perspective" on the colour distribution of the image.

Using a determination with colour information in different colour spaces, the interest points gain distinctiveness and stability. The shifting of interest points towards colour differences leads to a different focus of interest points, which can be very useful in colourful images and natural, cluttered scenes.

The strucutre of the thesis is presented in detail below:

Section 2 presents work that is related to this research. We begin with *visual saliency* in Section 2.1, explaining its need and use in computer vision. Giving an overview of its very different approaches, different definitions of regions of interest are shown. Focusing then on corner detectors, the state of the art of these interest points is given chronologically. Within Section 2.2, the origin of the contribution of this research is defined. Section 2.3 gives an overview of the ways of describing local structures for image retrieval, which is then finalized with the main ideas of descriptor matching in Section 2.4.

In **Section 3** the methods used in the software prototype developed to demonstrate the contribution of this thesis are presented. Based on the three steps of this image retrieval approach, the software is divided in three phases: The interest point extraction in Section 3.1 to gather useful and stable localisations for the subsequent description phase (Section 3.3). In this phase, the extracted interest points are used as the basis for the local description of the neighbourhood. Section 3.4 describes the matching algorithm in detail, which provides the final classification result of the retrieval task. The interest point extraction uses colour Harris corner detection in various colour spaces, and a newly proposed automatic scale selection based on statistical analysis of the given colour distribution.

Section 4 presents the experiments carried out with the new interest point detector and the software prototype. Repeatability experiments are shown in Section 4.2. To show that the new interest points gain performance in stability and information content, further experiments are done. In Section 4.3, the image retrieval experiments are described and the results are discussed. Both geometric and illumination changes are evaluated, and the new approach is compared to existing implementations in this field.

Section 5 summarizes the contribution and the results presented in the thesis. An outlook to future work and possible applications for the local colour features is given.

2 A Review of Local Image Description for CBIR

Approaches for matching images are very diverse in their principles and much research has been done in the last decades. Common in most approaches is the strategy to select some points or regions in the images and match these patches only. In this respect, certain image data is disregarded at an early stage. As will be shown in the following, this is not only a matter of processing time, but a matter of *focus*. The extraction of meaningful, stable and invariant locations is a crucial step for the whole process of recognizing visual input. The more discriminative the locations are, the more general the descriptors can become. The less ambiguous the locations that are described are, the more meaningful the descriptors that can be achieved will become. Therefore, the choice of right locations may result in a gain of performance.

In the literature, these locations are referred to as *salient points*, *regions of interest*, *focus points*, or *interest points*. The concept of an interest point was first defined by Moravec in 1977, who proposed a corner detector which is the basis for this research. We will follow this nomenclature and refer to extracted corners as interest points.

Other salient points rely on different assumptions: As the task of matching images and object recognition is fulfilled by primates very well, one strategy is to try to simulate the primate's visual system. Approaches to do so are presented in Section 2.1. All these approaches depend on observations in fields of neurobiology and psychology which try to explain the procedures taking place in a primate's image understanding process. Based on these insights, these algorithms try to simulate the procedures by machines. As it is broadly accepted that the first focusing takes place *before* the actual understanding of the scene, some of these mechanisms can be simulated by low level algorithms.

Section 2.2 describes the development of the Harris corner detector. This corner detector represents the non-saliency based approaches that are very powerful in many applications. Using corners is not very easy to explain from a neurobiological point of view, although some theories have arisen in the last years (Section 2.1). From a computer science point of view, corners provide very stable locations, which are highly needed for successful matching. Using different colour spaces as a basis of the interest point extraction (Sections 2.2.4 and 2.2.5), the primate colour perception is simulated. To follow one of the principles of visual saliency, the occurence probability of a colour is taken into account for the interest point extraction in Section 2.2.6.

Having extracted these locations, a smaller part of the image is selected for the description phase. For the matching, these regions have to be described. Section 2.3 gives an overview of common local description techniques. Beginning with the very straightforward patch sampling approaches, the most reliable SIFT descriptor (Section 2.3.3) and its enhancements are then shown. Section 2.3.4 points out newer colour description techniques, which can be used in parallel with the intensity based SIFT descriptor.

Section 2.4 summarizes the possibilities for matching the proposed local features. The similarities and the discriminative possibilities of these methods are shown.

2.1 Visual Saliency

Saliency is a broad term describing the quality of an entity of "standing out" relative to neighbouring entities. It can apply to any perception, item or feeling. This work investigates visual saliency only, and this is therefore referred to as saliency. Salient regions "pop out" of the visual input and create some form of visual arousal in the early, pre-attentive stages of the visual system [33]. For primates, a typical visual scene contains too many objects to be processed by the visual system at any given time. Therefore, additional mechanisms are needed to limit the data to visual objects that are currently relevant to the behaviour [15]. In cluttered visual environments, conspicuous objects are detected in real time by the use of saliency based attention processes [30]. Regarding image classification systems, the same issue arises: Diminishing the input data and discarding the irrelevant information is not only a matter of computational power and performance, but a mandatory constraint for interpreting a scene. The salient points are interesting for capturing the local image information, because they are located in visual focus points [82].

Visual attention is a subconscious process of the human visual system. The focus of attention is drawn to certain objects in that scene. The physiological and psychological processes of the primates' visual attention are not fully understood [7].

The predominant model for visual attention is that attention serves to enhance the response of neurons representing stimuli at a single behaviourally relevant location in the visual field [15]. Similar to numerous psychological models, a *spotlight of attention* [31] is shifted through the scene until the search task is fulfilled. There are a number of phenomena that are known to draw the attention in a visual scene including movement, shape, colour, or contrast in relation to the background [78]. Similarly, the model of *visual attention via selective tuning* [85] deals with the interaction between the input data itself and the relation between diverse perceptions using neural networks . The approach was proposed in [84] and is also referred to as the *inhibitory beam model*.

The *biased competition theory of attention* [15] is an alternative to these models. The effect of attention is best understood in the context of a competition between all other stimuli in the visual scene. The competition results in a focus on a few points of attention in the visual scene selected by anomalies in a scene. Regardless of whether the visual scene has been seen before, the attention is generally drawn to the anomalous object of the scene [80]. Another new aspect of this model is the fact that the subconscious process of attention focusing is a strict neurobiological process and is thus not affected by prior experience.

The important fact is that saliency-based approaches are not cognitive processes as there is no symbolic representation. The approaches are emergent as they try to adapt to scenes interactively and tend not to rely on prior specifications [93]. It is widely accepted that low level vision tasks are independent of context [33].

Saliency implies rarity [33]. The problem of this definition is the way in which the rarity is measured. If the parameters and descriptors are very distinctive, everything tends to be rare. On

the other hand, if the description is very general, salient regions will disappear. One of the main problems of saliency based algorithms is the setting of an appropriate level of discrimination. This is also addressed in the theory of *probabilistic saliency* [3]

In the literature, saliency is often divided into two parts: *Bottom-up models* and *top-down models*. Bottom-up models define neurobiological low level processes, most of the time without using prior knowledge or memory of the scene. It involves fast, and often compulsory, stimulus-driven mechanisms. Particular parts of the visual input are selected based on the properties of that input. It is a pre-attentive behaviour of the primate's visual system, often also referred to as *bottom-up saliency*, *bottom-up selection*, or *bottom-up effects*.

On the other hand, *top-down models*, *top-down saliency*, *top-down selection*, or *top-down effects* are part of the primates' conscious behavior. It is a slower, goal-directed mechanism where the observer's expectations or intentions influence the allocation of attention. For primates, they are caused by context understanding, memory, and different search tasks and motivations. Generally, attention selection and consequently, eye movements, are determined by the combination of stimulus-driven and goal-directed mechanisms [63].

In the following, several approaches for detecting salient regions are presented.

2.1.1 Instantaneous 2D Attention-Priority Maps

These maps give a straightforward measurement for saliency in image data. In this section, several instantaneous 2D attention-priority maps for simulating eye movement are described [9]. These maps are often part of the more complex approaches in the following sections. The algorithm is straightforward and high performance. It can be carried out in parallel, while for real time systems, the simple calculations can be programmed in special processing units.

The basic principle is that the variance of a certain property of the data gives a good idea of the rarity of a location. This is referred as a variance map of patch V. Having an image attribute $A(i,j)$ at the location i and j, and a $m \times n$ sized patch p, the contrast is given by

$$V_p = \sum_{i=1}^{m} \sum_{j=1}^{n} (A(i,j) - \overline{A}_p)^2 \tag{1}$$

where $\overline{A}_p$ is the mean of the attribute of the patch. This can be the intensity of the pixels, as is normally used. [9] suggest that eye movements tend to be guided toward regions where pixel intensities are decorrelated. This is refered to as the *intensity variance map*. Additionally, other attributes of the pixels can be used, which provide maps of orientation, colour moments, or motion, just to mention a few.

A very similar measurement is the contrast C_p proposed in [63]

$$C_p = \sqrt{(2r)^{-2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(A(i,j) - \overline{A}_p\right)^2} \tag{2}$$

where the nomenclature is as in Equation (1) and r is one half of the image patch width in pixels. It performs equally well, but is capable of dealing with different patch sizes. This allows a more general approach, which can be carried out in different scales and therefore provides a comparable attention measurement for different area sizes. This can be an important ability for a reasonable approach to simulating the primates' visual cortex.

2.1.2 Wavelet Based Salient Points

A wavelet representation provides information about variation in visual data at different scales. The idea of the wavelet based salient point extraction algorithm [82] is that wherever a high wavelet coefficient $W_{2^j}f(n)$ at a scale 2^j is encountered, it corresponds to a certain region with high global variation. If such a region is detected, coefficients at a higher scale 2^{j+1} are examined. Such a region can be seen as a *detail image*, which gives a finer location of the variations. In other words, it provides the convolution of the image with the wavelet function dilated at different scales.

In this approach, Haar wavelets are used, as they are symmetric and the simplest with orthogonal and compact support. Haar wavelets are fast to compute but not overlapping, which can lead to localization disadvantages. The compact support is important for knowing from which salient points each wavelet coefficient at a special scale was computed. This is done recursively with the ability to determine *children* $C(W_{2^j}f(n))$ of the sets of coefficients at higher scales (detail images of detail images).

To find the salient points, the maximum of a scale and then recursively the child with the highest coefficient is considered. This is called a *track* of a salient point. The coefficients of a track are summed up to extract salient points. A Saliency Value S for salient points is proposed.

$$S = \sum_{k=1}^{-j} |C^k(W_{2^j}f(n))|, 0 \leq n \leq 2^j N, -J_{max} \leq j \leq -1 \tag{3}$$

It is the sum of all absolute values of the wavelet coefficients considered in the recursive process of following the track. Therefore, coarse wavelet coefficients contribute to final salient points as well as finer coefficients [48]. In this respect, it is a multi-scale approach, which takes all scales into account for the final result of the saliency locations.

This recursive procedure is done for every wavelet coefficient. In thresholding S in relation to the desired number of extracted salient points, the most salient locations are obtained.

2.1.3 Saliency-Based Visual Attention for Rapid Scene Analysis

The model of saliency-based visual attention for rapid scene analysis was presented in [32]. The main inspiration of the model is to adapt a computer model to the early primate visual system. The authors compare each stage of the model with its biological equivalent. This is not the focus of this work and therefore will not be mentioned further. It is constructed to be processed

in parallel and is hence able to run in real time on dedicated hardware.

Basically, a global saliency map is generated to represent the conspicuity (or the saliency) of each position in the input image by a scalar value. As input, six intensity maps, 12 colour maps and 24 orientation maps are computed, making 42 feature maps for the basic feature extraction.

In the following, the approach is described in more detail. First, all the features are extracted. Then, a global saliency map is generated by comparing these to each other. This is carried out in a neural network, which builds the main idea of the approach.

Extraction of Early Visual Features Having an *RGB* picture as an input the intensity I is calculated by the average of these three values. These values are used to build a Gaussian pyramid with a scale $\sigma = [0..8]$.

As a threshold, each location where I is less than ten percent of the global maximum intensity value is discarded and becomes black. For all remaining pixels, each original colour channel is normalized by the location's intensity in order to decouple hue from intensity. Then the colour channels are converted to four broadly-tuned colour channels representing red, green, blue and yellow. Similarly, four Gaussian pyramids are created from these colour channels.

As a first set of feature maps, 6 *intensity maps* are created between the scales of the intensity pyramids. The second set is similarly constructed with the four colour pyramids using the "colour double-opponent system" obtaining 12 *colour maps*. Finally, oriented Gabor pyramids of the intensity image lead to 24 *orientation maps*.

The Saliency Map The combination of this collection of feature maps is not straightforward, as they represent different attributes in different modalities. A certain normalization method must therefore be found. Also, salient objects represented only in a few maps should be considered as well.

A *map normalization operator* is proposed, which promotes maps having few strong peaks of activity globally, while it suppresses maps with numerous comparable peaks.

Each set of feature maps is then combined into one of three *conspicuity maps*. An across-scale addition is done by reducing each map to the medium scale of four and then summing them point-by-point. The three resulting maps are averaged per pixel to generate the final saliency map. At any given time, the *focus of attention* (FOA) is directed to the global maxima of the saliency map.

The saliency map is used as the input for a *winner take all* neural network. All the neurons are independent. Their potential increases with the value of the saliency level. Corresponding winner take all neurons fire and therefore a new focus of attention is given. This leads to a time dependent FOA, which simulates the attention shift of primates.

2.1.4 Attention Based Similarity

Following the biased competition theory of attention model mentioned previously, a system for estimating attention based similarity is given in [79]. The system does not need any prior knowledge of the scene, and the response of the algorithm tries to *stimulate by surprise*. A pixel is surprising, and hence worthy of attention, if its neighborhood is not similar to other pixels' neighbourhoods. Such a pixel is unique in its features and thus salient.

The main tool of the system is the *fork* of pixels. For an analysed pixel, another pixel in a defined distance radius is taken into account and both are selected. This fork is than matched with other random pixel forks in the image. For each mismatch, the attention score is incremented. Forks can be rotated and scaled. Therefore, scale and rotation invariant features can be obtained.

The approach chooses forks randomly. No other strategy allows the complete lack of assumptions in analyzing visual data. As the different forks are matched to find a salient value, no distinct feature is focussed on. The forks can vary in scale, rotation, and location, and therefore no real asumptions about the result of the algorithm can be made. This defines the approach as an *emergent* system.

The matching of the forks is an independent operation, so the computation can be carried out in parallel. The major drawback of the approach is the large number of computations for the similarity estimation which increases dramatically with the size of the input image.

Concerning the large number of randomized, *exploring* matches, this is interpreted as an advantage by the author, as the approach is emergent rather than cognitivist. It does not rely upon prior specification, but explores the scene adaptively and thus interacts with the scene. The author claims that all other approaches to finding and describing features in images use predefined point selection routines and metrics that can limit performance on unseen data.

2.2 Interest Points

Because the definition of what makes a point salient is not very clear, assumptions about the salient point properties are made. Two main approaches can be found:

Region based approaches have gained lot of performance and therefore popularity in the last years (eg. [59]). They can be both used for finding salient points and the regions around them, or finding regions directly.

Scale space blobs [46] can be computed very efficiently in a number of ways. A blob is a thresholded, uniform region where the background is darker or lighter. There are many ways to detect blobs, including the Laplacian of Gaussian (Section 2.2.3) or simplified as Difference of Gaussian (Section 2.3.3), the determinant of the Hessian or simplified as Haar wavelets [2]. Affine shape adaption can be included [47].

Maximally stable extremum regions (MSERs) [52] are obtained by a watershed like algorithm. Connected regions of a certain thresholded range are selected if they remain stable over a set of thresholds. The algorithm is very efficient both in run-time performance and detection rate. The region priority or importance is measured in the number of thresholds where the region remains stable.

Corner detectors are one of the most popular methods to detect salient points. The idea is to assume that the salient points are corners. These methods usually refer to the points as *interest* points. In general this assumption is not valid since not all salient points are corners, and not all corners are salient points. However, corners are more stable than edges and are invariant to several transformations as described in Section 2.2.1. They are useful in a wide variety of applications, including stereo matching and object recognition.

In the following, several corner detectors are presented. These are efficient to calculate and hence a practical salient point extraction approach. The focus of our attention in this work lies on these approaches, as corners showed to give very useful locations in retrieval scenarios.

Corner detection can be traced back to Moravec [61] who measured the average change of intensity by shifting a local window by a small amount in different directions. The approach is discussed in Section 2.2.2.

Harris and Stephens [27] improved the repeatability of the Moravec detector under small image variations and near edges by an analytic expansion of the Moravec detector. The local autocorrelation matrix of the approach is derived using first order derivatives. The principle theory and realisation is shown in Section 2.2.3. The Harris detector, in combination with a rotational invariant descriptor, was also used by Schmid and Mohr [71] when they extended local feature matching to general object recognition.

A low-level approach to corner finding is proposed by Smith and Brady: the SUSAN detector [77]. Their corner detector compares the intensity of a pixel with the intensities of neighbouring pixels. If few of the neighboring pixels have approximately the same value, the center pixel is considered a corner point.

Lindeberg [46] proposed an "interesting scale level" detector which is based on determining maxima over scale of a normalized blob measure. The Laplacian-of-Gaussian (LoG) function is used for building the scale space. Mikolajczyk [57] showed that this function is very suitable for automatic scale selection of structures. An efficient algorithm for use in object recognition was proposed by Lowe [50]. This algorithm constructs a scale space pyramid using Difference-of-Gaussian (DoG) filters. There are several way of calculating the function, but the main advantage is the high performance in building a pyramid hierarchy. The DoG are used to obtain an efficient approximation of the LoG. From the local 3D maxima a robust descriptor is built for matching purposes. The disadvantage of using DoG or LoG is that the repeatability of the extracted features is not optimal since both DoG and LoG not only respond to blobs, but also to high gradients in one direction. Because of this, the location of the features may not be very accurate.

An approach that intuitively arises from this observation is the separation of the feature detector and the scale selection. The original Harris detector [27] is robust to noise and lighting variations, but only to a small extent to scale changes [72]. To deal with this Dufournoud et al. [17] proposed the scale adapted Harris operator. Given the scale adapted Harris operator, a scale space can be created. Local 3D maxima in this scale space can be taken as salient points. Mikolajczyk points out that the scale adapted Harris operator rarely attains a maximum over scales [57]. This results in very few points, which are not representative enough for capturing the image content. To address this problem, Mikolajczyk [57] proposed the Harris-Laplace detector that merges the scale-adapted Harris corner detector and the Laplacian based scale selection, which is described in Section 2.2.3.

All the approaches presented above are intensity based. Since the luminance axis is the major axis of colour variation in the RGB colour cube, most interest points are found using only intensity. The additional colour based interest points might not dramatically increase the number of interest points. The distinctiveness of these colour based interest points is however much larger, and therefore colour can be of great importance when matching images. A colour extension of the Harris detector is proposed in [60] and discussed in Section 2.2.4. This is an important step towards saliency based approaches, as colour plays an important role in the preattentive stage in which features are detected. This means that the saliency value of a point also depends on the colour information that is present.

Very relevant to this work is the research of van de Weijer et al. [88]. Using colour information and further colour invariance, an important fact of the human perception of interpreting colour information (e.g. [6, 32, 83]) is then modeled in Section 2.2.5. They aim at incorporating colour distinctiveness into the design of interest point detectors. In their work, the colour derivatives form the basis of a colour saliency boosting function since they are used in both the detection of the interest points, and the determination of the information content of the points.

Furthermore, the histograms of colour image derivatives show distinctive statistical properties which are used in a colour saliency boosting function. As described previously, the distinctiveness of a perception is an argument for its saliency. These color statistics in Section 2.2.6 can therefore be seen as the missing link between visual saliency and interest points. As colour is processed in a way inspired by human perception, the main idea of these approaches is integrated into a mathematically sound approach.

In the following, the desired properties for interest points are described. The idea of locations being invariant to changes from one image to another is explained. Focussing on the Harris corner detector, the basic principles to achieve these properties are given.

2.2.1 Goals and Properties of Local Interest Points

Extracting interest points from image data should lead in to comparable results under varying conditions. For example specific, salient locations of an object in different images should be extracted from local interest point algorithms in an invariant way. Having different images of

an object under different view points, comparable locations are desired, not with respect to the image, but to the object in the image.

The degree of invariance of a particular local interest point algorithm determines the success and the range of applications in which it can be used [65]. In this section, the main types of invariances regarding local interest point detectors are presented.

Spatial Invariance is the property of providing the same results after the translation of an object in an image, or the cropping or translation of the image. Local interest points should remain after such transformations. Obviously information about cropped image content will be lost, and possibly new content will arise, but the same locations should persist. It is considered to be the basis of all invariance properties as it requires the same locations to be present without regarding the actual position in an image. This property contradicts some of the saliency properties of the image, as data can become less or more salient when the surroundings change. However, it is an important fact in the idea of interest points, as it provides comparable location for image retrieval.

Scale Invariance describes the ability to provide the same result after camera zooming or image resizing. Zooming of a camera results not only in change of scale, but also in other non-affine transformations, but this is generally not considered in the issue of scale invariance. In [95], it is proposed that scale should regarded as a continuous parameter for image representation, and [45] showed that under some rather general assumptions, the Gaussian kernel

$$G(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{4}$$

and its derivatives are the only possible smoothing kernels for scale space analysis. This *scale space theory* making scale invariance possible, is the basis for one of the most important properties in an image retrieval context, as it allows the analysis of image data under varying resolution and size.

Separability is a very attractive property of the Gaussian kernel. It allows to have a multi dimensional kernel as a product of one dimensional kernels.

$$G(\sigma) = G_x(\sigma) G_y(\sigma) \tag{5}$$

where G denotes the Gaussian kernel, x, y the direction in the image and σ the size of the kernel. The one dimensional Gaussian kernels

$$\begin{aligned} G_x(\sigma) &= \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2)}{2\sigma^2}} \\ G_y(\sigma) &= \frac{1}{2\pi\sigma^2} e^{-\frac{(y^2)}{2\sigma^2}} \end{aligned} \tag{6}$$

can have great advantages for the run time of a convolution.

Another important property is the *commutative semi-group property*, which allows one to calculate a successive smoothing with a Gaussian kernel (i.e. multiple convolutions $\otimes$ with image I) to be carried out in one calculation step.

$$G(\sigma_1) \otimes ... \otimes G(\sigma_n) \otimes I = G(\sqrt{\sigma_1^2 + ... + \sigma_n^2}) \otimes I \tag{7}$$

Using different scales σ and normalizing the detector results with σ^2 [46], the interest points become comparable to each other over different scales. Therefore, scale-invariant detection is possible. The *characteristic scale* is then chosen by the highest result of the detector.

Affine Invariance is a generalization of the scale invariance when scale changes are not isotropic. It is handled by using a more general Gaussian kernel defined by

$$G(\Sigma) = \frac{1}{2\pi\sqrt{\det\Sigma}} e^{-\frac{\mathbf{x}^T\Sigma^{-1}\mathbf{x}}{2}} \tag{8}$$

where Σ is the covariance matrix defining the affine transformation of the image. It is often refered to as the *shape adapted Gaussian kernel*. This approach has four parameters to deal with, instead of the one in non affine cases. This leads to a more complicated and therefore more time consuming calculation for detection of the local interest points. Therefore, detectors typically use just the detected scale invariant region for further affine invariant transformations.

In the following, interest point detectors are described. Beginning with the Moravec corner detector, we develop the state of the art in this field of the interest point extraction.

2.2.2 Moravec Corner Detector

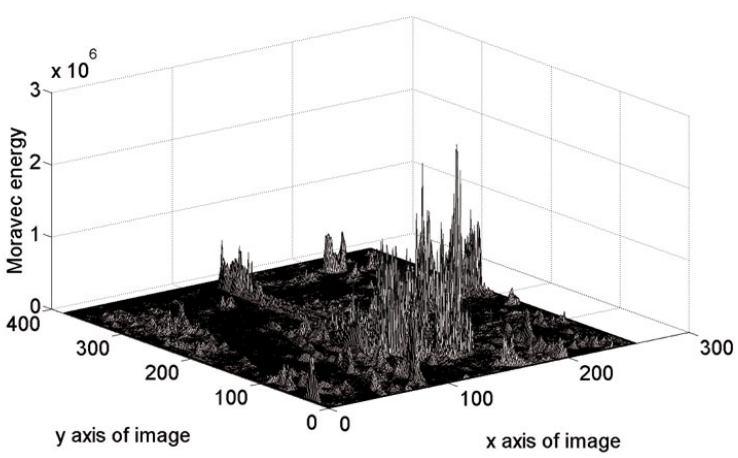
The basis for the Harris corner detector algorithm used in this research is the earlier Moravec low-level corner detector [61]. The Moravec corner detector estimates the similarity of each point to its neighbour pixels by the sum of the squared differences (referred to as SSD).

The SSD can be seen as a local autocorrelation function of the image. It is is computed in four directions and takes the lowest result as the measure of interest. Hence it detects points where there are large intensity variations in every direction. Moravec did not only develop the first corner detector but was the first to introduce the idea of *point of interest*.

Given the image I, the cornernes measurement *SSD*, also refered to as Moravec energy, on the pixel (x,y) is estimated for the location (u,v) with a patch of the width w as

$$SSD_{x,y}(u,v) = \sum_{i=-w}^{w} \sum_{i=-w}^{w} \left[I(x+i,y+i) - I(x+i+u,y+j+v)\right]^2 \tag{9}$$

In Figure 5, the *SSD* and the maxima of this function are shown. The are several drawbacks of this fast and straightforward way of cornerness measurement. The main one is its anisotropy. Hence it is not orientation invariant, as only discrete orientational shifts of 45° are considered. Therewith connected, edges, which are considered as unstable, are rated very highly compared



(a) $SSD; u, v = 5$

(b) 30 highest maxima extracted.

Figure 5: Visualization of the algorithm to extract a fixed number of Moravec corners.

to corners. The measurement is not very stable against noise, either.

These are three highly needed properties for the extraction of stable and salient locations and the main criticism of Harris and Stephens [27]. Their approach is described in the following section.

2.2.3 Harris Corner Detector

The Harris corner detector introduced in [27] provides a cornerness measure for image data. Overcoming the main drawbacks of the Moravec corner detector, the Harris corner detector extends this approach in analyzing the gradients in a patch.

A major objective is to suppress the noise without suppressing the anisotropic shape of the structure. For this purpose the gradients around the point are computed using a Gaussian derivative:

$$X = I \otimes (-1,0,1) = \frac{\partial I}{\partial x} \qquad (10)$$
$$Y = I \otimes (-1,0,1)^T = \frac{\partial I}{\partial y}$$

The gradients are averaged by a Gaussian kernel of the differentation scale σ_D. This process smooths down the responses to noise. The result is three directional components

$$
\begin{aligned}
L_x^2(x,y,\sigma_D) &= X^2 \otimes G(\sigma_D) \\
L_y^2(x,y,\sigma_D) &= Y^2 \otimes G(\sigma_D) \\
L_xL_y(x,y,\sigma_D) &= (XY) \otimes G(\sigma_D)
\end{aligned} \tag{11}
$$

For any shift (x,y) the cornerness measurement can then be formulated as

$$
E(x,y) = (x,y)M \begin{pmatrix} x \\ y \end{pmatrix} \tag{12}
$$

The symmetric second moment matrix M describes the gradient distribution in the local neighbourhood of a point.

$$
M = \mu(x,y) = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} = \begin{bmatrix} L_x^2(x,y) & L_xL_y(x,y) \\ L_xL_y(x,y) & L_y^2(x,y) \end{bmatrix} \tag{13}
$$

Therefore it is possible to consider multiple different directions and the dicrimination between edges and corners becomes more effective. The key properties of this matrix are extracted from its eigenvalues. They represent two principal curvatures of the location. The eigenvalues of the second moment matrix ($\lambda_1(\mu)$ and $\lambda_2(\mu)$) are proportional to the principal curvatures of the considered area, and are invariant to rotation.

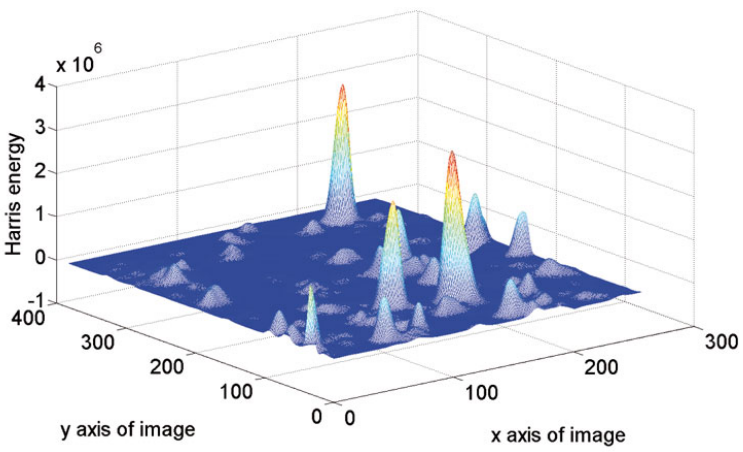
Based on these two eigenvalues, a classification can be made: If they are near zero, there is obviously no curvature in any direction. One notable larger value than the other one indicates a predominant curvature direction: an edge. If both eigenvalues are large, curvature in different directions is encountered. The main idea of the Harris corner detector is the fact that exactly these spots are the most interesting and stable ones in an image.

This two dimensional cornerness measurement is not very intuitive, and the decision between interesting corner and uninteresting uniformness depends on several assumptions. Therefore, the pixels are not classified into the three categories, but analysed to result in a one dimensional Harris energy, the measure proposed is based on the trace and determinant of the second moment matrix:

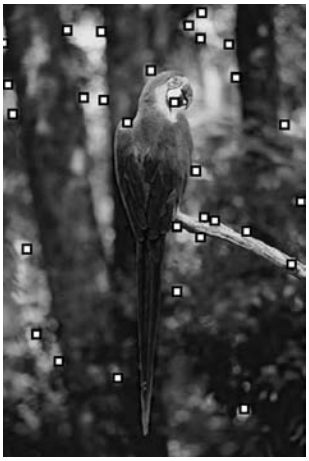
$$
\begin{aligned}
C_H(\mu) &= \det(\mu) - \alpha \mathrm{trace}^2(\mu) \\
\det(\mu) &= \lambda_1(\mu)\lambda_2(\mu) = \mu_{11}\mu_{22} - \mu_{12}\mu_{21} \\
\mathrm{trace}(\mu) &= \lambda_1(\mu) + \lambda_2(\mu) = \mu_{11} + \mu_{22}
\end{aligned} \tag{14}
$$

As the only constant, α indicates the slope of the *zero line*. In the Harris energy, 0 is the border between corner and edge. The Harris energy and the extraction of the maxima shown in Figure 6.

Scale Invariant Harris corner detection Using a fixed scale has one drawback: too small or too big structures are not taken into account. The goal is to develop a scale invariant description



(a) Harris energy; σ_I=1.4, σ_D=5



(b) 30 highest maxima extracted.

Figure 6: Visualization of the algorithm to extract a fixed number of Harris corners.

of corners in an image. In our retrieval context, this idea provides us with the same locations, regardless of the size of the object in the image.

An extension of the Harris corner detector is done to achieve scale invariance. The derivative scale σ_D is included as the scale of the corner detector. The second moment matrix is then

$$M(x,y,\sigma_I,\sigma_D) = \sigma_D^2 G(\sigma_I) \otimes \begin{bmatrix} L_x^2(x,y,\sigma_D) & L_x L_y(\mathbf{x},\sigma_D) \\ L_x L_y(x,y,\sigma_D) & L_y^2(\mathbf{x},\sigma_D) \end{bmatrix} \tag{15}$$

where σ_I is the integration scale. As shown in several experiments [56, 57], the following relation performs best:

$$3\sigma_D = \sigma_I \tag{16}$$

To make the different scales more comparable, the matrix components are averaged by the Gaussian Kernel $G(\sigma_I)$. The factor σ_D^2 is for scale normalization.

Using different normalized cornerness measurements, characteristic scales can be found by setting a threshold on the results. Mikolajczyk and Schmid [56] applied the Laplacian of Gaussian to the detector for the automatic scale selection. This is further called the *Harris-Laplacian* detector.

The scale space of the Harris function is built by iteratively calculate the cornerness measurement E under varying σ_D and σ_I (Equation (12)). Using scale steps s with a factor t of σ_D from 1.2 to $\sqrt{2}$, the cornerness measurement E is extended to

$$E(x,y,s) = (x,y) M(x,y,t^s\sigma_I,t^s\sigma_D) \begin{pmatrix} x \\ y \end{pmatrix} \tag{17}$$

The second moment matrix M, developed in Equation (13), is not changed at all. The amount of scale change is chosen by the need for preciseness of the corner location. In reasonably big objects in a retrieval context, $t = \sqrt{2}$ showed to be precise enough, and is therefore used in the experiments in Section 4. For more precise results, a value of $t \approx 1.2$ is recommended.

The next step is to choose the characteristic scale. The Laplacian of Gaussian function Λ has been used to detect the characteristic scale automatically [47, 55]. Λ is defined by

$$\begin{aligned}\Lambda(\sigma_D) &= -\frac{1}{\pi(t^s\sigma_D)^4}\left(1-\frac{x^2+y^2}{(t^s\sigma_D)^2}\right)e^{-\frac{x^2+y^2}{2(t^s\sigma_D)^2}} \otimes c_{u,v} \\ &= \left((\frac{\partial I}{\partial x})^2+(\frac{\partial I}{\partial y})^2\right)\otimes G(\sigma_D)\otimes c_{u,v} \\ &= \left(L_x^2(x,y,\sigma_D)+L_y^2(x,y,\sigma_D)\right)\otimes c_{u,v}\end{aligned} \tag{18}$$

where $c_{u,v}$ is the raised cosine kernel, as described below. Using the sum of the precomputed values L_x and L_y the computation can be done very efficiently.

To make the maxima more stable, a raised cosine kernel

$$c_{u,v} = \frac{1+\left((\frac{1}{2}-cos(\pi u))+(\frac{1}{2}-cos(\pi v))\right)}{3} \tag{19}$$

is used to smooth the resulting data. As suggested by [11], this kernel gives more smoothed borders than the Gaussian kernel G for scale decision.

A characteristic scale of a possible region is found if both the Harris Energy and the Laplacian of Gaussian are extrema

$$\nabla\Lambda(x,y,s(\sigma_D)) = \nabla M(x,y,\sigma_I,\sigma_D) = \overline{0} \tag{20}$$

where $\overline{0}$ is the Null vector. With this non-maxima suppression, the locations with their according scales are found.

Affine Invariant Harris corner detection The second moment matrix can also be used to determine the affine deformation of a structure. It is shown [56] that an affine normalized image patch can be obtained by transforming the image with the square root of the shape adapted second moment matrix, which is obtained by using the shape adapted Gaussian kernel

$$G(\Sigma) = \frac{1}{2\pi\sqrt{\det\Sigma}}e^{-\frac{\mathbf{x}^T\Sigma^{-1}\mathbf{x}}{2}} \tag{21}$$

Two normalized regions from an image and an affine transformed copy are shown to be related by a rotation. Under certain affine transformations, this property may be crucial.

The uniform second moment matrix can be generalized to non uniform scale space [57]

$$\mu(x,y,\Sigma_I,\Sigma_D) = \det(\Sigma_D)G(\Sigma_I)\cdot((\nabla L)(x,y,\Sigma_D)(\nabla L)(x,y,\Sigma_D)^T) \tag{22}$$

where Σ_D and Σ_I represent the differentiation and integration kernels, and $L(\mathbf{x},\Sigma)$ represents the affine Gaussian scale space.

In an iterative procedure proposed proposed by Lindeberg and Garding [47] the shape adapted second moment matrix M that has the following properties is derived:

$$\begin{aligned} \mu(x,y,\Sigma_I,\Sigma_D) &= M \\ \Sigma_I &= \sigma_I M^{-1} \\ \Sigma_D &= \sigma_D M^{-1} \end{aligned} \tag{23}$$

where scalars σ_I and σ_D are the integration and differentiation scales. Local moment-based image descriptors computed under these conditions have shown to be stable under affine transformations.

2.2.4 Colour Harris Corner Detector

A basic extension of the intensity based Harris detector is proposed by Montesinos et al. [60]. The second moment matrix they used is defined as

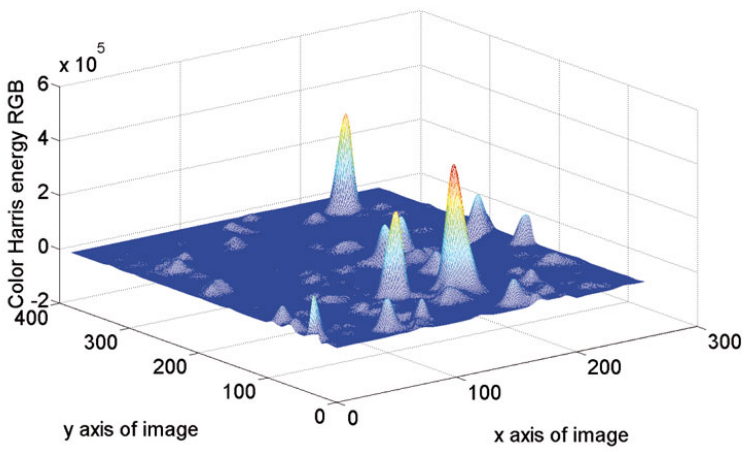
$$M = \mu(x,y,\sigma_I,\sigma_D) = \sigma_D^2 G(\sigma_I) \otimes \begin{bmatrix} R_x^2+G_x^2+B_x^2 & R_xR_y+G_xG_y+B_xB_y \\ R_xR_y+G_xG_y+B_xB_y & R_y^2+G_y^2+B_y^2 \end{bmatrix} \tag{24}$$

Instead of using just the intensity gradient, the gradient for each colour channel is determined. These values are summed and averaged using a Gaussian kernel with size σ_D.

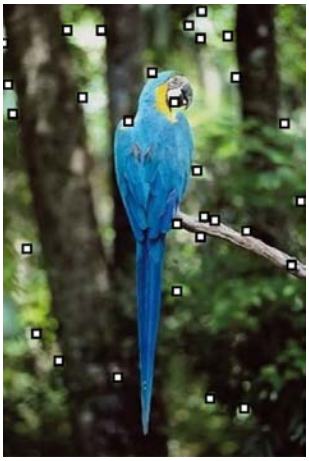
As suggested in [88], the second moment matrix can be computed using different colour models. The first step is to determine the gradients of each component of the *RGB* colour system. This is done using a convolution with the differentiation kernels of size σ_D. The gradients are then transformed into the desired colour system. By multiplication and summation of the transformed gradients, all components of the second moment matrix are computed. The values are averaged by a Gaussian integration kernel with size σ_I. Scale normalisation is done again using the factor σ_D^2. The Harris energy is shown in Figure 7(a) and extracted corners can be seen in Figure 7(b).

Because of common photometric variations in imaging conditions such as shading, shadows, specularities and object reflectance, the components of the *RGB* colour system are correlated. By transforming the *RGB* colour coordinates to other systems, photometric causes for features in images can be distinguished.

Normalizing the *RGB* colour space to *rgb*, illumination is rectified and equals 1 for each colour:



(a) Color Harris Energy; σ_I=1.4, σ_D=5



(b) 30 highest maxima extracted.

Figure 7: Visualization of the algorithm to extract a fixed number of Color Harris corners based on *RGB*.

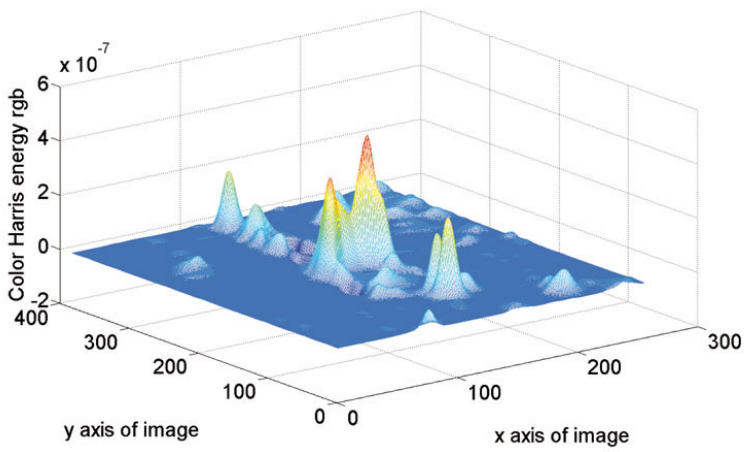
$$rgb = \begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix} \tag{25}$$

In this transformation, all the illumination information is discarded. The cornerness measurement changes to prioritize the colour changes between foreground and background and therefore, the silhouette of the parrot can be recognized in the graph in Figure 8(a). The main drawback of this colour space is the fact, that *rgb* is very unstable near zero illumination and the approach will lead to highly prioritized corners within smallest colour changes in dark regions of the image. This is encountered in the lower right part of the image in Figure 8(b).

2.2.5 Quasi Invariant Colour Space

To overcome this weakness and to enhance the ability to be stable against lighting changes, different colour spaces can be used. The goal would be to have an *invariant* colour space to all manner of light, shadow and reflectance effects that may be encountered in an image. The following colour spaces have some invariant properties, leading into a *quasi-invariant* colour space which is invariant to shadow and specular effects.

The spherical colour space *S* (Equation 26), the opponent colour space *OCS* (Equation 27), and the quasi invariant *HSI* colour space (Equation 28) [90] are developed:



(a) Color Harris Energy; σ_I=1.4, σ_D=5

(b) 30 highest maxima extracted.

Figure 8: Visualization of the algorithm to extract a fixed number of Color Harris corners based on normalized *rgb*.

$$S = \begin{pmatrix} \theta \\ \varphi \\ r \end{pmatrix} = \begin{pmatrix} \tan^{-1}\left(\frac{G}{R}\right) \\ \sin^{-1}\left(\frac{\sqrt{R^2+G^2}}{\sqrt{R^2+G^2+B^2}}\right) \\ \sqrt{R^2+G^2+B^2} \end{pmatrix} \tag{26}$$

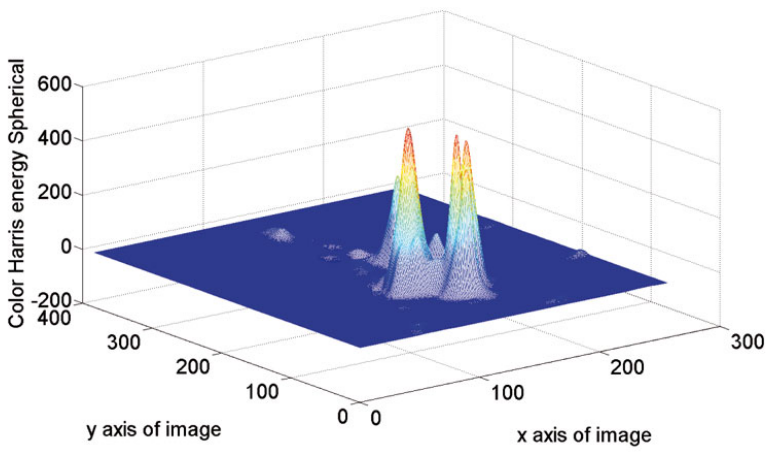
$$OCS = \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{27}$$

$$HSI = \begin{pmatrix} h \\ s \\ i \end{pmatrix} = \begin{pmatrix} \tan^{-1}\left(\frac{o_1}{o_2}\right) \\ \sqrt{o_1^2+o_2^2} \\ o_3 \end{pmatrix} \tag{28}$$

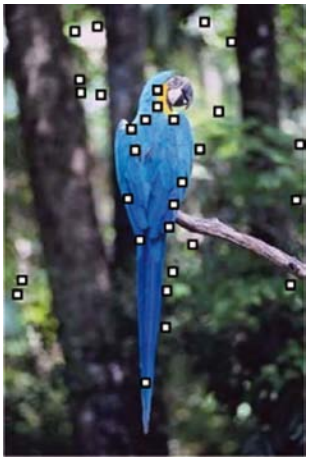
In these decorrelated color spaces only the photometric axes are influenced by the common photometric variations. In [88] the spatial derivatives are separated into photometric variant and invariant parts.

The spherical colour transformation (Equation (26)) has the shadow-shading direction as the r coordinate. This overcomes the instability of *rgb* and is very stable in colours with low illumination. As shown in Figure 9, dark regions are more stable and there are fewer corners at shadowing effects involved. The orthonormal transformation into *OCS* (Equation (27)) provides specular variance. In Figure 10, the prioritization of the yellow - blue edge is shown.

As this colour space is often motivated by simulating primate retinal processes, these opponent colours are the end points of one axis of the colour space, and therefore have the largest

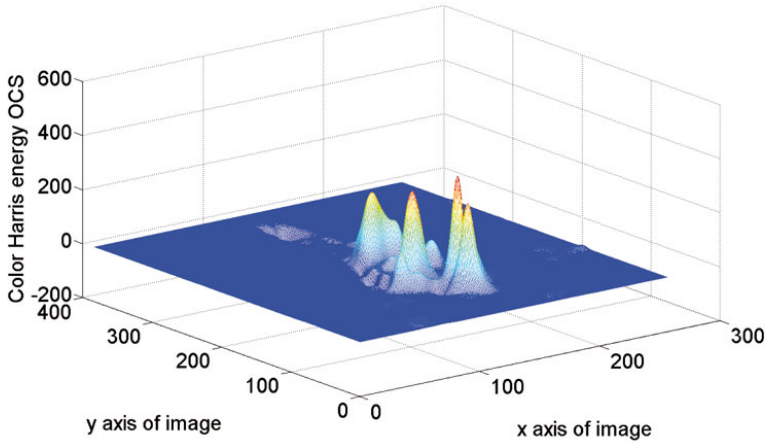


(a) Color Harris Energy; σ_I=1.4, σ_D=5

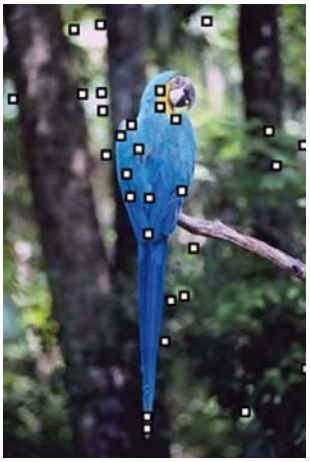


(b) 30 highest maxima extracted.

Figure 9: Visualization of the algorithm to extract a fixed number of Colour Harris corners based on the spherical color space.



(a) Color Harris Energy; σ_I=1.4, σ_D=5

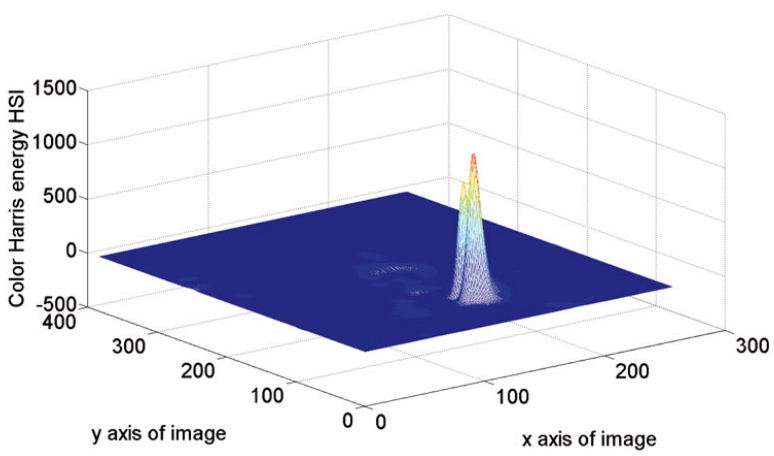


(b) 30 highest maxima extracted.

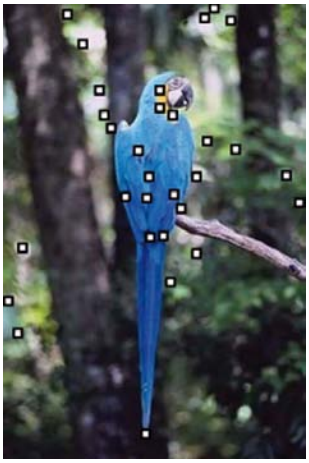
Figure 10: Visualization of the algorithm to extract a fixed number of Colour Harris corners based on *OCS*.

distance. The specular variance can also be seen at the right shoulder of the parrot in Figure 10(b), where the blue feathers become lighter and lighter, and these corners are prioritized in the *OCS*.

A polar transformation on the first two axes of the *OCS* leads to the *HSI* colour space (Equation (28)). The derivative of the hue component is both the shading and the specular quasi-invariant [88], which means that those light effects should not change the coefficient. Ob-



(a) Color Harris Energy; σ_I=1.4, σ_D=5



(b) 30 highest maxima extracted.

Figure 11: Visualization of the algorithm to extract a fixed number of Colour Harris corners based on the quasi invariant *HSI*.

viously, this applies not to every specular and shadowing effect in a natural scene.

The main drawback is the undefinedness between black and white. Further, small changes around the gray axis result in large changes in the color direction as can be seen in the upper left corner of Figure 11(b). As shown in Figure 11, colour changes between opponent colours are highly prioritized, as it is motivated by the primates' visual cortex.

2.2.6 Colour Statistics and Boosting

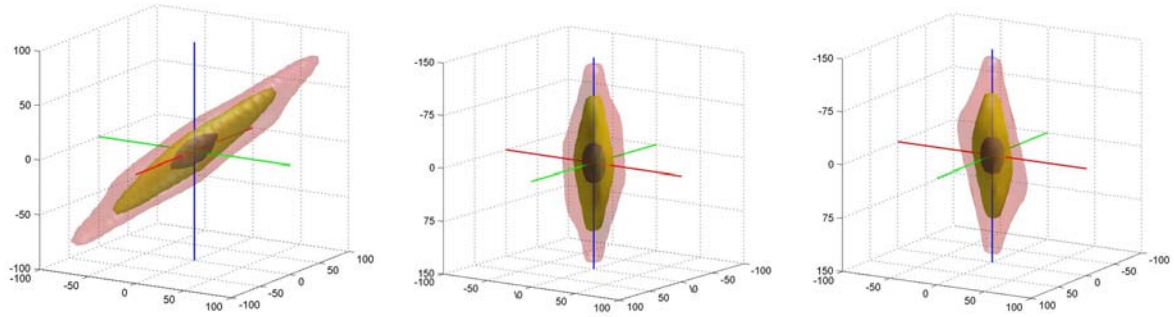
As proposed in [89], colours have different occurrence probabilities $p(v)$ and therefore, different information content $I(v)$ of the descriptor v:

$$I(v) = -log(p(v)) \tag{29}$$

The assumption is now to boost rare colours for having a higher saliency in the cornerness measurement. Looking for rare colours, statistics for the Corel Database[2] containing 40 000 colour images showed that the three dimensional colour distribution was remarkably significant. For all considered colour spaces, one coordinate coincides with the axis of the maximum variation (see Figure 12).

Following these properties, a boosting function can be found so that pixel information in the image data has equal impact on the saliency function as its information content. The strength of gradients is considered as the decorrelated information content. This is a transformation $g : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ such that

[2] www.corel.com



(a) RGB (b) Opponent Colour Space OCS (c) Spherical Colour Space SCS

Figure 12: Three dimensional colour distribution of the transformed image derivatives. The darkest shapes contain 90% of all pixels, the lighter ones 99% and 99.9%, respectively. From [89].

$$p(f) = p(f') \leftrightarrow |g(f)| = \left|g(f')\right| \tag{30}$$

where $p(f)$ is the probability for pixel information in an image, or of its gradient $p(f')$, respectively.

The transformation is obtained by deriving a function describing the surface of the three-dimensional colour distribution, which can be approximated by an ellipsoid in the origin (see Figure 12) after alignment of the data as described in the following:

The coordinates have to be aligned to the axis of the coordinate system to achieve one simple boosting factor per channel. A rotation matrix R^{ϕ} can be found, that so the data is aligned in *RGB* (Figure 12(a)) following Equation (31), *OCS* (Figure 12(b)) by Equation (32), and finally the spherical colour space (Figure 12(c)) by Equation (33).

$$\begin{pmatrix} \widetilde{G}_x \\ \widetilde{R}_x \end{pmatrix} = R^{\phi} \begin{pmatrix} G_x \\ R_x \end{pmatrix} \tag{31}$$

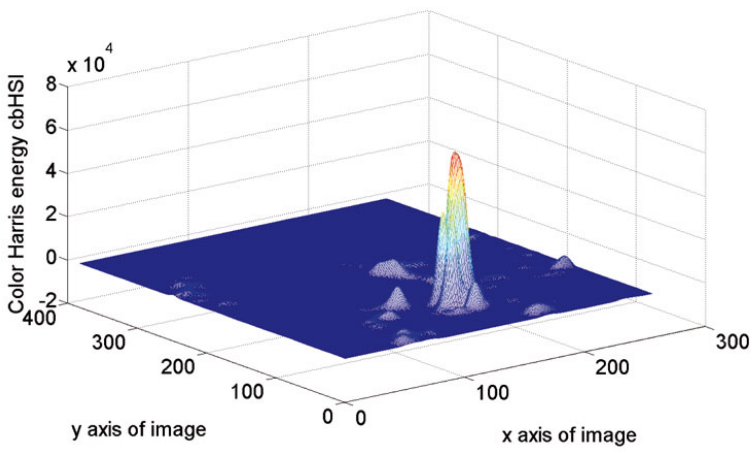
$$\begin{pmatrix} \widetilde{o1}_x \\ \widetilde{o2}_x \end{pmatrix} = R^{\phi} \begin{pmatrix} o1_x \\ o2_x \end{pmatrix} \tag{32}$$

$$\begin{pmatrix} r\sin\widetilde{\varphi}\widetilde{\theta}_x \\ r\widetilde{\varphi}_x \end{pmatrix} = R^{\phi} \begin{pmatrix} r\sin\varphi\theta_x \\ r\varphi_x \end{pmatrix} \tag{33}$$

The colour distributions can then be approximated by ellipsoids having the definition

$$(\alpha h^1)^2 + (\beta h^2)^2 + (\gamma h^3)^2 = R^2 \tag{34}$$

where the $h^{[1..3]}$ factors are the axes in Figure 12 or transformed axis in Equation (31) to Equation (33), respectively.



(a) Color boosted Harris Energy;σ_D=1.4σ_I=5

(b) 30 highest maxima extracted.

Figure 13: Visualization of the algorithm to extract a fixed number of colour boosted Harris corners in opponent colour space.

To find the transformation in Equation (30), the ellipsoid has to be transformed from the elliptic shape to a sphere, so that vectors of equal saliency lead to vectors of equal length. The definition

$$g(f_x) = \Xi h(f_x) \tag{35}$$

leads to a saliency boosting factor for each component of the colour space. For the opponent colour space, the diagonal matrix Ξ is given by

$$\Xi = \begin{bmatrix} 0.850 & 0 & 0 \\ 0 & 0.524 & 0 \\ 0 & 0 & 0.065 \end{bmatrix} \tag{36}$$

The idea is that these factors not only hold for the analysed images, but for other natural images too. The large scale of the experiment in the Corel database promises good results in using the extracted function even in new images. In Figure 13, the statistically extracted function is used for the extraction of colour Harris interest points in an image which is not part of experiment data.

2.3 Region Description

A localized region should be described in a compact and complete way in order to make a similarity measurement possible. This is done with *local descriptors*. The features which describe the region must be highly distinctive, and also invariant to changes in the images. They are often referred to as *local interest point descriptors* although it is obvious that they have to describe a

local interest area.

Generally, the description should be invariant to the same changes as the localization stage of the interest points described previously including invariance to noise, scale, rotation, transformation, etc. Unfortunately, invariance and distinctiveness are inversely dependent. Distinctiveness is lost when obtaining a more invariant descriptor, and vice versa. The more the description gets invariant, the less information is expressed. For example, somebody telling a story about a artistic performance would probably include if somebody was standing on his hands. Choosing the rotation invariant way of describing the show and saying that there was just a person present would probably take away the whole point of the story.

As already pointed out in Section 2.1, setting the appropriate level of discrimination is again one of the main problems, especially because this discrimination has to be done automatically for unseen general image data. However, there is one guideline for modern applications: as the interest points gain more stability and invariance, the regions get more stable and distinct. The increasing invariance of the location phase leads to less influence of the invariance of the description regarding the overall performance of a CBIR system. This allows one to have more specific descriptors of more invariant locations for a more discriminative and precise retrieval [65].

These two properties are desireable for a successful local description:

Distinction The description for different image content should be different, and description of similar content should be similar (see matching approaches in section 2.4). This property is hard to measure, and is related to the term of *rareness* and *discrimination* discussed in Section 2.1

Compact Size a compact description of features is both desireable in storing the data, matching features and in run-time issues for applications [96].

In the following, different ways of describing regions of interest are shown.

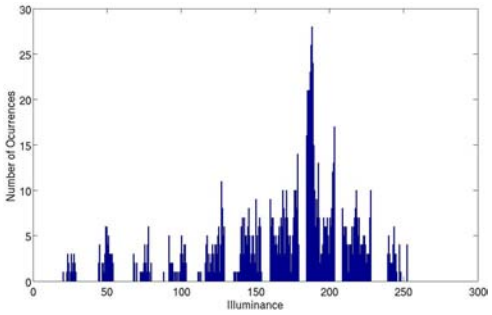
2.3.1 Patch Sampling

The simplest way to describe a region of interest is simply to describe the luminance values inside the region. With the already found region of interest, size n for the patch $P(i, j)$ is given. Then an n^2 dimensional feature vector is produced where the local descriptor D

$$D = \begin{pmatrix} l_1 \\ \vdots \\ l_{n^2} \end{pmatrix} \tag{37}$$

contains every luminance value l of $P(i, j)$.

(a) Illuminance patch of the region of interest



(b) Histogram of illuminance values

Figure 14: Visualization of a straightforward descriptor: The region of interest provides the localization. The patch in (a) is analyzed. In order to have a description of the luminance distribution, the histogram in (b) is generated.

This approach is highly unstable with respect to noise, illumination or affine transformation changes. On the other hand, it is therefore the most distinct descriptor. Its simplicity and high performance makes it useful in some cases [43].

This approach can be enhanced to get more stable descriptors. As shown in Figure 14(a), the region of interest can be sampled to a fixed size to get a uniform dimensional feature vector. The complete head of the parrot is sampled to a 32×32 region. This leads to a very weak ability for scale invariance. To overcome the weakness of describing the exact position of every pixel, histograms are used. In Figure 14(b), a histogram of the intensity values is shown. The description consists of ranges of luminance values and has no spatial information any more. The distinctiveness of this descriptor becomes much lower than the first approach, but it is more stable against noise, small illumination changes and transformations. To get more stable to bigger illumination changes, the luminance values are normalized.

2.3.2 Derivative Description

Using derivatives for image description provides a more invariant and stable description than using the original values, even if they are normalized. The different orders of the derivative at the point x, y are built from the convolution with the n^{th} order derivative of the Gaussian Kernel $G(\sigma_D)$ (see Figures 15 and 16).

$$L_n(x, y, \sigma_D) = I \otimes G_n(\sigma_D) \tag{38}$$

The resulting functions $L_n(x, y, \sigma_D)$ are visualized in Figure 16.

The 0th order of an image I (shown in Figure 16(a)) provides stability against noise, additionally it is the basis for building a scale space [45]. The 1st order results in the differences

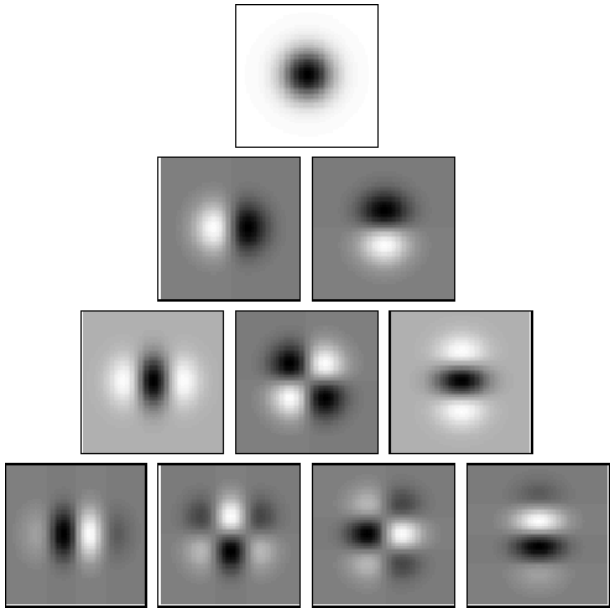
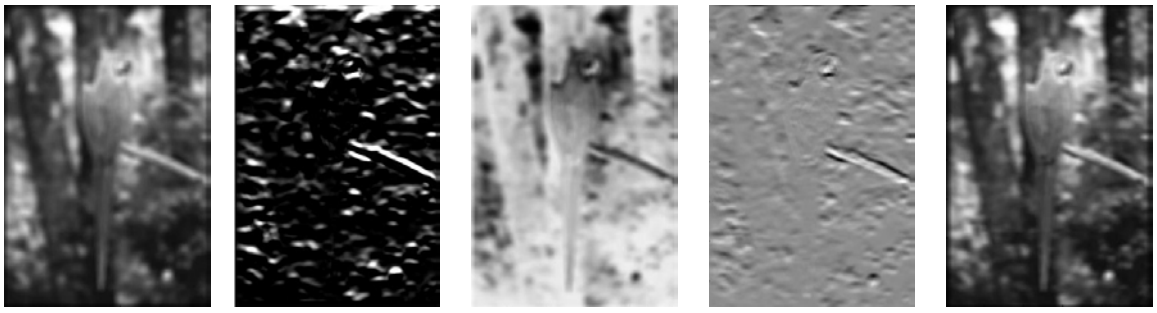


Figure 15: Uniform Gaussian derivatives up to the fourth order in a pyramid. Every line shows the filters for the possible directions (x, y, xy, respectively). From: [45].



(a) $L_0(x,y,10)$ (b) $L_1(x,y,10)$ (c) $L_2(x,y,10)$ (d) $L_3(x,y,10)$ (e) $L_4(x,y,10)$

Figure 16: Convolved images in ascending derivative order in x direction $L_{ix}(x,y,\sigma_D)$. Luminance range scaled for better visualization.

between adjacent neighbours. It overcomes some drawbacks of normalization. There are i.e. no numerical problems for small values.

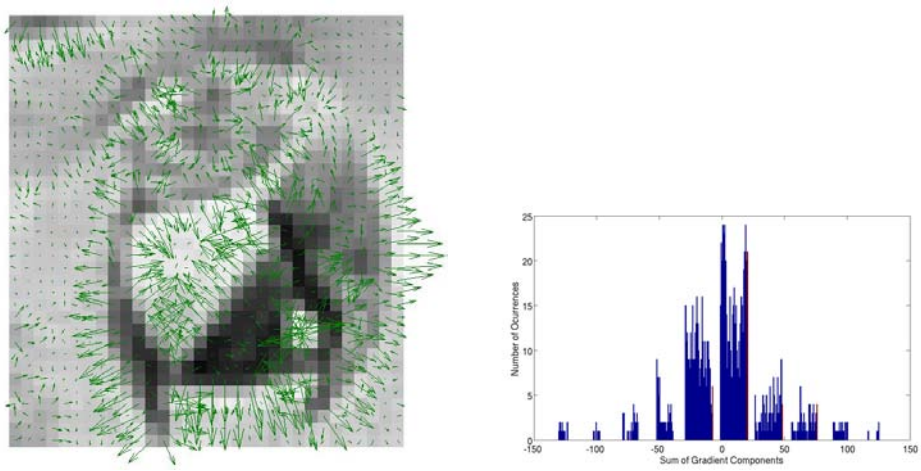
The idea of gradients in different directions is probably best visualized to take the derivatives as components of a vector. Figure 17(a) shows the first order gradients. Descriptors based on these gradients are able to achieve additional scale, rotation and affine transformation invariance. A naive approach can be seen in Figure 17(b), where the sum of the gradient components build a histogram.

One example is the *local jet* descriptor of the order p [37]. Into a predefined gradient basis, the local interest area description is a decomposition of the gradient information. The full range

of jets is defined as

$$J_p(I,x,y,\sigma_D) = \{L_{1\ldots n}(x,y,\sigma_D)\}, \forall n \in [1\ldots p] \tag{39}$$

To use a set of different Gaussian kernels in multiple direction is proposed by [1, 70] and referred to as *steerable filters*. The direction of the derivative can be chosen freely and leads so to a better invariance to rotation.



(a) Gradients $I_{[x|y]}$ visualized as vectors on the according pixels

(b) Histogram of sum of gradient vectors

Figure 17: (a) The gradients of the region of interest visualized as vectors. (b) One possible description is a histogram of the sum of the gradient components I_x and I_y.

2.3.3 SIFT

The *scale invariant feature transform* (SIFT) was introduced by Lowe in [49]. The method provides local image features developed for reliable object matching. One of the weaker points of the original approach is the extraction of the salient points [16, 58]. Improvements of the original approach to increase the stability are done in [8, 50]. Although several enhancements of this descriptor have been made (e.g. PCA-SIFT [35] based on [20], GLOH [58], SURF [2]), the original method is still state of the art in a general context of experiments (e.g. [59]), especially under heavy transformations.

In the following, details of the implementation are shown and the main methods described.

Scale Space The scale space $L(x,y,\sigma)$ is defined by

$$L(x,y,\sigma) = G(x,y,\sigma) \otimes I(x,y) \tag{40}$$

and provides the main information for the SIFT description, where $G(x,y,\sigma)$ is the multi-scale Gaussian Kernel (Equation (4)) with the convolution $\otimes$ with the Image $I(x,y)$.

Every smoothing step is calculated with the 1D Gaussian function in the horizontal and vertical direction and with the paramter $\sigma = \sqrt{2}$. To approximate this function, seven sample points are taken into account. Experiments showed that this number of points build the best trade-off between performance and preciseness.

In the first step, the image is smoothed and stored separately. One image with a smoothing of $\sigma = \sqrt{2}$, one with the effective smoothing of $\sigma = 2$ is produced, where the first image is subtracted from the second one to obtain the Difference of Gaussian function $D(x, y, \sigma)$

$$D(x, y, \sigma) = (G(x, y, \sqrt{2}\sigma) - G(x, y, \sigma)) \otimes I(x, y) = L(x, y, \sqrt{2}\sigma) - L(x, y, \sigma) \tag{41}$$

For the next pyramid level, the $\sigma = 2$ smoothed image is bilinearly interpolated by 1.5 pixels in each direction. So each new pixel is a result of 4 original pixels. To maintain the highest spatial frequencies, the first level of the pyramid is the expanded image by the factor 2.

Notably, the function $D(x, y, \sigma)$ is a close approximation to the Laplacian of Gaussian function Λ(x,y,σ) shown in Equation (18).

$$G(x, y, \sqrt{2}\sigma) - G(x, y, \sigma) \approx \Lambda(x, y, \sigma) \tag{42}$$

It has been used as the decision measurement for the characteristic scale of a region. Here, the function uses illumination information only, a fact which is one drawback for the overall retrieval result, but a necessary fact for the sound comparison between the meaning of the interest points in a retrieval context.

Key Point Localisation Key points are used in the SIFT implementation to localize the most important feature in a region of interest. To detect the extrema, a pixel is compared to its 8 neighbours in one pyramid level. If it is a maxima or minima, it is compared to the corresponding pixel in the next level. If this pixel is still a local extrema, further testing in the next level is done. This leads to an early diminished dataset, as the majority of the pixels are discarded right away in the first scale.

To make this location more accurate, the Hessian and derivative of $D(x, y, \sigma)$ are approximated by the differences to the prior analysed 8 neighbours. If there is an offset $\widehat{x}$ larger than 0.5 in any direction, the point is moved to the next pixel in this direction. Otherwise, the offset is added to the current location to get an interpolated extremum.

$$D(\widehat{x}) = D + \frac{\partial D^T}{2\partial x}\widehat{x} \tag{43}$$

Using $D(\widehat{x})$ as a threshold (0.03 in the implementation), this function mainly discards locations with low contrast, thereby stabilizing them.

To discard edges and prioritize corners, the already calculated Hessian matrix is used to process an adapted Harris corner detection algorithm (compare Section 2.2.3 to Equation 44):

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{44}$$

The eigenvalues of this matrix gives a measurement of the curvature at this location. There is no need to calculate them implicitly, since the ratio of them gives an idea of the local structure. This can be done efficiently substituting Equation (14) and $\lambda_1 = r\lambda_2$ into Equation (45) (compare again with properties in section 2.2.3):

$$\frac{\text{trace}(\mu)^2}{\det(\mu)} = \frac{(\lambda_1 + \lambda_2)^2}{\lambda_1 \lambda_2} = \frac{(r\lambda_2 \lambda_2)^2}{r\lambda_2^2} < \frac{(r+1)^2}{r} \tag{45}$$

r gives then the relation between the two eigenvalues. In case the determinant is negative, the point is discarded. In the implementation, a threshold of $r = 10$ is used for a valid location.

Orientation Assignment Using now the estimated position and the scale of a key point, the gradient magnitude $m(x,y)$ and $\theta(x,y)$ is precomputed for every $L(x,y)$ using

$$\begin{aligned} m(x,y) &= sqrt(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2 \\ \theta(x,y) &= \arctan \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)} \end{aligned} \tag{46}$$

The orientations $\theta(x,y)$ are stored in a 36 bin histogram covering 10 degrees per bin. They are weighted by its magnitude $m(x,y)$ and smoothed by a 1.5 times larger Gaussian window $G(1.5\sigma)$ than the previously computed scale σ. The highest peak in this histogram builds the dominant direction in the gradients. If there is any other peak within 80% of the highest one, another keypoint with this orientation is built.

The description of these key points is inspired by [18], a model of biological vision, in which it is argued that the orientation and frequency of a gradient is more important to the primal visual cortex than its precise localization.

The Feature Vector One goal of the SIFT descriptor is to be invariant against small shifts in the relative gradient position. Creating 16 histograms per description with 8 orientation bins for the sample regions leads to a very stable but not too distinct description of a region. Positions of the gradients may change in their quarter of the neighbourhood. Due to the fixed number of bins, the resulting feature vector for a key point consists then of 128 dimensions.

The feature vector is created by first computing the gradient magnitude and orientation at each image sample point in a region around the key point location. These are weighted by a Gaussian window. These samples are accumulated into orientation histograms summarizing the contents over 4x4 subregions, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region.

To reduce the effect of non-linear illumination change, the highest values of the normalized feature vector are diminished. Experiments showed that a threshold value of 0.2 performed best. Every value in the feature vector higher than this threshold is scaled back. Then, the whole vector is renormalized. This is done to reduce the priority of large gradients.

PCA-SIFT The PCA-SIFT [35] was introduced to achieve a more compact local descriptor than the original SIFT [50]. The development aimed towards wide-baseline matching, and the more compact description should perform equally well in these tasks. This is achieved by using the PCA projection of the gradient map to describe the region of interest. The resulting descriptors were shown to perform best with 20 dimension. The two main steps can be summarized as the following: the *training* step pre-computes an eigenspace to express the gradient maps of the region of interest using a set of training data. In the *testing* step, the gradients of the new region of interest are projected into the eigenspace to obtain a decomposition of the gradient map. The eigenspace coefficients are the actual elements of the PCA-SIFT. Under certain circumstances, it even outperforms the SIFT descriptor.

SURF - Speeded up Robust Features [2] was highly influenced by the success of the SIFT descriptor. The main idea is that the performance in both processing time and description stability can be increased when using less, but more essential features. It is rather different to the original SIFT, but it is often referred to as the high performance extension of the SIFT descriptor. Regarding the whole implementation - including the initial feature extraction - the detection is done with mean and average box filters to estimate $L_{x|y}$. With non-maximum suppression and interpolation of blob-like features, the scale of the locations is determined. With this information, the Haar wavelet responses are represented as vectors and summed within sections of 60°. The predominant vector is the longest one, and second longest is ignored. The description is then the sum of absolute values of the responses in the sections and leads into a feature vector of length 64. This diminished dimensionality also improves the processing time for matching.

2.3.4 Local Colour Description

Generally, there is no need to restrict the description approaches described previously to the illumination property, as long as it is one dimensional input data. Regarding more dimensional colour information, there are several methods providing the desired properties for local descriptors as described previously.

In the following, a method to describe local colour information in images is given [91]. It is a concatenated approach including three state of the art ideas for colour description:

Colour constancy For colour constancy in different images, colour normalization has to be done. The description should provide independence from diffuse lighting and specular reflectance. Invariance to the dominant illumination and to the diffuse illuminant is achieved by normalizing the colour channels with their average derivative. This correction for the illuminant is related to the *grey-edge hypothesis* proposed in [87].

Colour invariance A colour descriptor has to be robust to colour variations because of shadows, shading, specularities and different light sources. Therefore, [90] showed that the derivatives from the opponent colours computed after the normalization step are the only invariant insensitive to a diffuse illuminant. For a combined illumination and geometric invariant feature, the comprehensive colour image normalization (CCIN) [21] can be applied as a local feature.

Geometric robustness The local feature requires robustness to changes of viewpoint, zoom, and object orientation. An affine invariant feature detection partially solves this problem already. However, under transformations, colour edges tend to change and due to aliasing and perspective effects, undesired effects occur. *Colour angles* overcome variance to certain blurring which is encountered by zoom or focal changes.

Descriptors can be built from the previous computations. If the colour description is used in combination with a shape description, no spatial information is needed. The descriptors do not provide stable shape and structure information. Therefore, the SIFT descriptor is concatenated to the colour description to provide both. This results in a higher dimensionality of the descriptor, but it tends to be more discriminative. In certain CBIR scenarios, it outperforms the orginal SIFT approach [91].

In the following section, the last phase of the task of CBIR is described. Several approaches to measure the similarity of local descriptors are shown.

2.4 Descriptor Matching

The next stage of the typical retrieval method is the matching of the extracted features. The most important part of matching is the distance measurement. It describes the similarity of features and should therefore be the measurement for a match of features. The multi-dimensional distance is described in Section 2.4.1.

To speed up the process of matching, which is crucial for an image retrieval application, several methods for enhancing the search inside the feature database are presented in Section 2.4.2.

2.4.1 Feature Distance

For feature matching, the Euclidean distance gives the similarity measurement between two candidates. This can be carried out with multi-dimensional features, where the nearest neighbour is the vector with the minimal distance:

$$D_{pq} = \overline{pq} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{47}$$

where p and q are two feature vectors of the same dimensionality. Typically, the vectors are normalized for a more robust matching, so that

$$\overline{p} = \overline{q} = 1 \tag{48}$$

For matching of features, the distance function is called often. In the image retrieval experiment in Section 4.3.4, each query image with an estimated 250 features had to be compared with estimated 5 000 000 features in the database. Therefore, a short calculation time for the distance function is essential for the efficient performance of a query. In this respect, not only the function run-time is crucial for the performance, but also the requirement of fast data fetching in the feature database. This is both a matter of descriptor compactness, and information encoding in the database itself. A fast decoding often conflicts with the the property of small data size for features (compare properties of local features in Section 2.3).

Other examples of multidimensional distance measurements are the Canberra distance (Equation 49), squared chord distance (Equation 50), squared Chi-squared distance (Equation 51), which can also be used in clustering algorithms (Section 2.4.2).

$$D_{Cpq} = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i + q_i|} \tag{49}$$

$$D_{Spq} = \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2 \tag{50}$$

$$D_{CHpq} = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{(p_i + q_i)} \tag{51}$$

The cosine similarity function [69] treats the set of descriptors as components of an M dimensional vector, and the similarity is the cosine of the angle between these vectors (their dot product divided by their magnitudes). This similarity, which is also known as the Ochini coefficient, is given by the expression

$$OC_{pq} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2 \sum_{i=1}^{n} q_i^2}} \tag{52}$$

Similar descriptors have an Ochini coefficient approaching 1.0 and a distance approaching 0.0. The $arccos(OC_{pq})$ can therefore be used for a distance function.

Normally, a *similarity threshold* gives then the response whether two local descriptors match or not. For object recognition, an additional threshold of the number of matched features is needed to generate a positive recognition result.

The search method inside the feature database is a crucial performance issue in large databases, as already pointed out. The method widely used to speed up the process of finding the nearest neighbours is the clustering of the features into similar descriptor clusters and/or indexing the descriptors. In the next section, several approaches are described.

2.4.2 Descriptor Organization

Hierarchical or tree based approaches are adaptive in the partitioning of the data. The access time can be reduced when taking advantage of hierarchical structures.

K-d Tree [24] is a special case of the binary space partitioning tree and is a very popular and powerful hierarchical structure for computer vision. It stands for a complete binary tree having smaller bins in the higher-density regions of the space. Performance can be raised and look up time reduced, but ends up in more complex tree generation algorithms, as several constraints must be fulfilled (e.g [36]).

Approximate Searches One way to make matching faster is to be satisfied with a *good* result, and not looking for the best one possible. When loosening the constraint for the nearest neighbour, processing time can be highly reduced. Internal nodes of graphs can be the centers of the mass of the nodes of a lower level [54] and therefore, similar descriptors can be looked up quickly. Neighbourhood graphs show to gain good results in higher dimensionalities, too.

Clustering is a powerful tool for finding structures in large data. It aims to partition data into clusters. In this respect, it is a classification task within data sets which share some common attributes within a certain distance. Two of the most popular approaches are *K-means* and *agglomerative clustering*.

K-means The most straightforward algorithm for clustering high dimensional data is the k-means algorithm. Randomly, k seed points are initialized for the clusters. In the subsequent iterations, every data point is assigned to the nearest cluster center. The cluster centers are then computed as the mean center of the assigned data points [41]. It performs very well on large data sets [75]. The run-time depends on the relation between the number of seed points k, the dimensionality d, number N of data points, and the number of iterations l: $O(Nkld)$.

Improvements can be achieved by using hierarchical approaches like *kd-trees* [67] or the triangular inequality [19]. The issue of having random seed points leads to different results in different runs on the same database. Several methods to overcome this problem are presented in [64].

Agglomerative Clustering starts by assigning each data point to a new cluster. Pairs of clusters are then iteratively selected and merged. A hierarchical structure tree is thereby iteratively built.

The crucial part of the algorithm is the criterion for merging clusters. Several methods are useful for this criterion, including Equations 47 to 52. Other criteria can be found in [39].

The crucial performance issue is the fact that clusters should be merged in decreasing order of their distances to each other. After a merge, the distances have to be recalculated. This makes the standard approach not applicable on large data sets. The possibility to overcome the fixed

merging order is proposed in [12], where additionally, the recomputation of merged clusters becomes more efficient.

2.5 Summary

In this section, state of the art of methods used in CBIR systems are described. This thesis focusses on the impact of salient locations in an image retrieval context. Therefore, an overview of approaches regarding visual saliency in computer vision is given. As we chose the Harris corner detector as the basis for this research, this approach is discussed in more detail and described from its first form as an autocorrelation measurement to the state of the art colour Harris corner detector.

Several approaches for local descriptors are given. The local descriptor chosen for this research, the SIFT descriptor, is described in more detail. The different advantages and drawbacks of descriptors and their dependence on the provided locations in the previous phase in a CBIR system are shown.

Matching of descriptors is the last step of defining the similarity of image data. In most of the retrieval scenarios, this is a crucial calculation time issue. Therefore lot of research has been done in this field, and a brief overview is given.

3 An Image Retrieval System Based on Colour Interest Points

In this chapter, a software prototype for image retrieval is developed. Due to its organization, all the operations can be carried out in parallel. The system processes each task, query and data set independently so that multiple machines can allocate the processing time.

As shown in Figure 18, the image retrieval scenario is structured in independent tasks: The first stage of the retrieval scenario contains the extraction of the salient points in the image (Section 3.1). A form of a general *saliency implies rarity* [33] approach for weighting colours is used for focusing interest points. A new approach to extend automatic scale selection to multiple colour spaces is shown in Section 3.1.2. A visualization of the performance of the proposed regions is given in Section 3.2.

Having these locations, the basis for the forthcoming local description in Section 2.3 is found: Local description takes place in the defined locations only. The matching of these descriptors is described in Section 3.4. The methods for exchanging and storing data are described in Section 3.5.

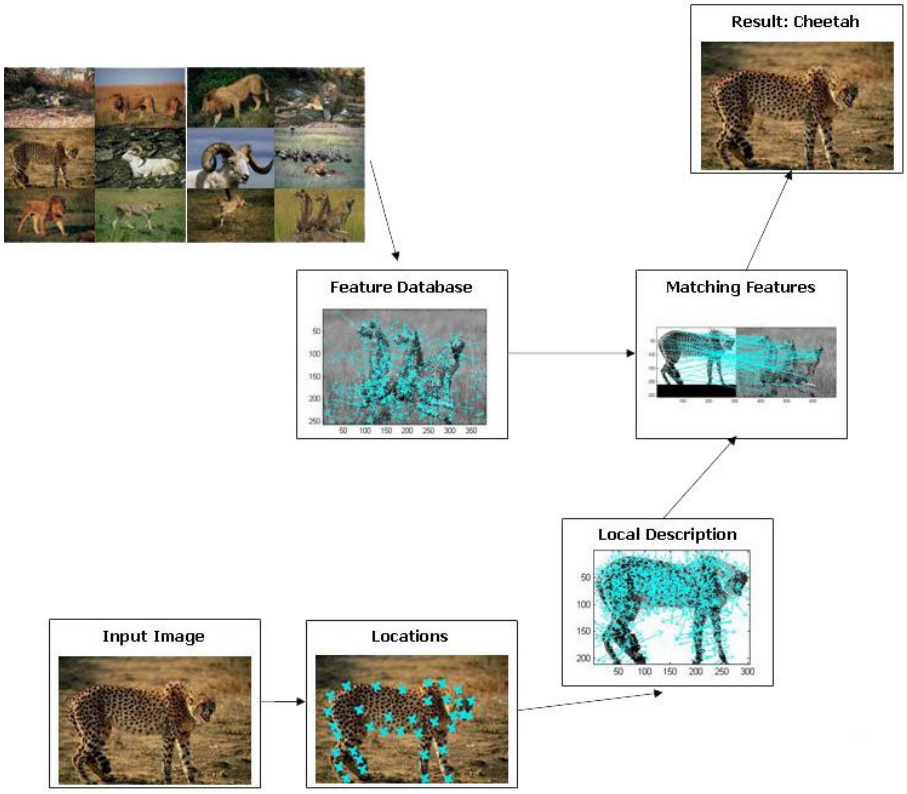


Figure 18: Overview of the software prototype. Locations refers to Section 3.1, local description is described in Section 3.3. The matching of the feature and the generation of the final result is explained in Section 3.4. The feature database is created with a set of data files, which are defined in Section 3.5.

3.1 Interest Point Extraction

The stage of interest point extraction provides the location for the local description. From the input image, colour Harris corners (see Section 3.1.1) are extracted at different scales. The decision, which scale to prioritize is described in Section 3.1.2, which builds the main contribution of this research.

3.1.1 Corner Extraction

As already pointed out, the *RGB* colour space is highly correlated and therefore not stable under varying illumination circumstances. In natural high contrast images, same colours appear completely different under varying illumination in this colour space. The *RGB* changes are very similar to the changes in illumination. Therefore, a colour Harris in *RGB* does not dramatically change the position and priority of the corners compared to an illumination based approach (see Section 2.2.4). This is not the case for photos with low contrast changes as usually encountered in studio photography, artificial images or highly post processed images.

The second moment matrix can be computed using different colour models. The first step is to determine the gradients of each component of the RGB colour system, as this is this matrix is only defined in this colour space. This is done using a convolution with the differentiation kernels of size σ_D. The gradients are then transformed into the desired colour system. By multiplication and summation of the transformed gradients, all components of the second moment matrix are computed. The values are averaged by a Gaussian integration kernel with size σ_I . Scale normalization is done again using a factor σ_D.

To write this procedure in symbolic form, we use a more general notation not restricted to one colour space. Colour space C is used with its components $[c1,...,cn]^T$, where n is the number of colour channels.

$$C = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \tag{53}$$

The second moment matrix is defined as

$$M(\mathbf{x},\sigma_I,\sigma_D) = \sigma_D^2 G(\sigma_I) \otimes \begin{bmatrix} L_x^2(\mathbf{x},\sigma_D) & L_xL_y(\mathbf{x},\sigma_D) \\ L_xL_y(\mathbf{x},\sigma_D) & L_y^2(\mathbf{x},\sigma_D) \end{bmatrix} \tag{54}$$

with the components L_x^2, L_xL_y and L_y^2 defined as:

$$L_x^2(\mathbf{x},\sigma_D) = \sum_{i=1}^{n} c_{i,x}^2(\mathbf{x},\sigma_D) \tag{55}$$
$$L_xL_y(\mathbf{x},\sigma_D) = \sum_{i=1}^{n} c_{i,x}(\mathbf{x},\sigma_D)c_{i,y}(\mathbf{x},\sigma_D)$$
$$L_y^2(\mathbf{x},\sigma_D) = \sum_{i=1}^{n} c_{i,y}^2(\mathbf{x},\sigma_D)$$

where $c_{i,x}$ and $c_{i,y}$ denote the components of the transformed colour channel gradients, with $i \in [1,...,n]$, and where the subscript x or y indicates the direction of the gradient. These colour spaces can be for example *OCS*, *HSI*, etc. as described in Section 2.2.5.

3.1.2 Colored Automatic Scale Decision

To achieve scale invariant locations, the Harris corner detector has to build a scale pyramid consisting of the Harris energy over different scales. We propose a new, adapted method for finding the characteristic scale in different colour spaces [81] to extend the scale decision with colour information (see Figure 19).

The input image is transformed to the same colour space as used for the extraction of the Harris energy. After that, a principal component analysis (PCA) is applied to diminish the three colour dimensions of the input image to a one dimensional dataset $\widehat{I}(x,y)$

$$\widehat{I}(x,y) = \sqrt{3}v_\lambda I(x,y)^T \otimes c_{u,v} \tag{56}$$

by having the dot product of the colour information $I(x,y)$ and the corresponding eigenvector v_λ. To keep the relative pixel value differences in the same range, the result is scaled back by $\sqrt{3}$.

This analysis leads to a transformed one-dimensional function which includes many of the advantages of the corresponding colour space as described in Section 2.2.5. Based on $\widehat{I}(x,y)$,

$$\Lambda(x,y,\sigma_D) = -\frac{1}{\pi(t^s\sigma_D)^4}\left(1-\frac{x^2+y^2}{(t^s\sigma_D)^2}\right)e^{-\frac{x^2+y^2}{2(t^s\sigma_D)^2}} \otimes c_{u,v} \tag{57}$$

can be applied and the characteristic is chosen as described in Section 2.2.3.

As the discrimination vector is chosen due to the maxima of the sum of the distances between the values, the PCA as the basis for the scale decision criterion ensures that a trade-off between the prioritization of rare colours and not losing information on similar colours is achieved. Therefore, it can be seen as a relaxed colour boosting function within the dimension reduction.

(a) Illumination (b) HSI

Figure 19: 30 extracted regions based on illumination and HSI information $t = 1,2; s = 10; \sigma_I = 0.7$. The regions shift towards colour differences, specular and shading changes are not regarded any more. The parrot is therefore highly prioritized.

If values are salient and have therefore greater distances to the other values, the less salient colours are disregarded as corners. They get very small distances to each other, and are therefore not prioritized. If the distance to the rarest colours is not big enough, the transformation prioritizes the common colours, which get more distance to each other, then. This is a saliency examination on global scope for a local scale decision.

However, it tends to lose less distance information than other transformations $f : \mathbb{R}^3 \rightarrow \mathbb{R}^1$ e.g. the usually used illumination transform. The PCA analysis does not have a direct perception counterpart in real-world, but tend to adapt itself to the colour distribution of the image.

Combining different colour spaces (e.g. extracting colours from human perception motivated opponent colour space and choosing the scale on illumination invariant normalized *rgb*) between the corner extraction and the scale selection turned out to not work at all, as the most informative responses tend to extinguish each other.

Aiming for just one region per location and a balanced distribution of regions over the input

image, the following decision criteria showed to perform best:

$$\widehat{R}(x,y) = \begin{pmatrix} \max\left(E(x,y,*)\right) \\ 3t^{\arg\max\left(\widehat{\Lambda}(x,y,*)\right)}\sigma_D \end{pmatrix} \tag{58}$$

where $E(x,y,*)$ is the Harris energy over all regarded scales. With this decision, only one region per location is chosen, and the Harris energy and the scale is estimated independently. This leads to the function $\widehat{R}(x,y)$ defining all candidates for interesting points and the corresponding region size.

The number of different scales regarded is a crucial matter for the processing time. Each step must be calculated on its own (but independently and therefore possibly in parallel) and the processing time increases with the size of the kernels. Building this scale space of the Harris energy leads to a pyramid of the cornerness measurement E. In several other implementations of a scale invariant corner detector (e.g. [55, 74]) the Harris energy of scale levels is scale normalized by the factor σ_I^2 because the values tend to get lower the higher the scale gets. This is not done in this implementation, because of the different automatic scale decision approach.

Within this step, there is not one interest point chosen, but a region of interest with a center interest point. It examines the surrounding structure of the image and selects the predominant size of a region. This area information is then used for the description phase in Section 3.3.

3.2 Shifting Interest Points Towards Colour

Taking colour information into account leads to a different definition of interest points. Every colour transformation provides other properties. In the case of the quasi invariant colour space, only colour is regarded.

The colour only interest points consider only changes in colour, not in illumination, specular, or shadow changes. It is more likely to describe meaningful objects in real circumstances, as the results are not changed by different lighting conditions. Working with natural cluttered images, this helps to overcome one of the major problem: two natural image may show the same objects in the image, but taken under completely different circumstances. Then, e.g. under completely different lighting conditions, the images will be different. For illumination changes, illumination based methods will have completely different output for obvious reasons.

Figures 20(a)-20(f) show the shifting of interest points towards the colour edges as more colour specific interest point detectors are used. Note that the background of Figure 20(f) is not disregarded, but the colour edges have higher priority. As the colour only edges tend to have higher contrast, the corner measurement expands heavily under the quasi invariant colour space. Therfore, the Harris energy has a wider range of data, and pure colour edges are more repeatable. This is shown in Section 4.2.

Figure 20(a) shows the illumination based corners. The background with lot of self shadowing effects tends to provide many illumination corners. These shadow effects are highly

unstable under varying lighting circumstances. *RGB* based corners in Figure 20(b) show to provide nearly the same results. Normalized *rgb* overcomes the illumination based point of view. The example image contains lot of dark regions. where *rgb* is highly unstable and provides ambiguous results. Therefore no stable corners are extracted in this case (Figure 20(c)).

OCS (Figure 20(d)) shifts the focus of the colour differences towards colour, and less locations caused by pure shadow effects apppear. Using colour boosting in Figure 20(e), rare colours are more prioritized and therefore the colourful crab provides the most stable interest points. Figure 20(f) extracts corners based on the quasi-invariant *HSI* colour space, and regards pure colour edges only. In this case, the 30 corners with the highest Harris energy are found on the crab only.

Taking more maxima into account, a more constant distribution of the interest points can be achieved. This is important for CBIR systems. For visualization reasons, only the 30 highest maxima of the Harris energy are shown.

3.3 Region Description

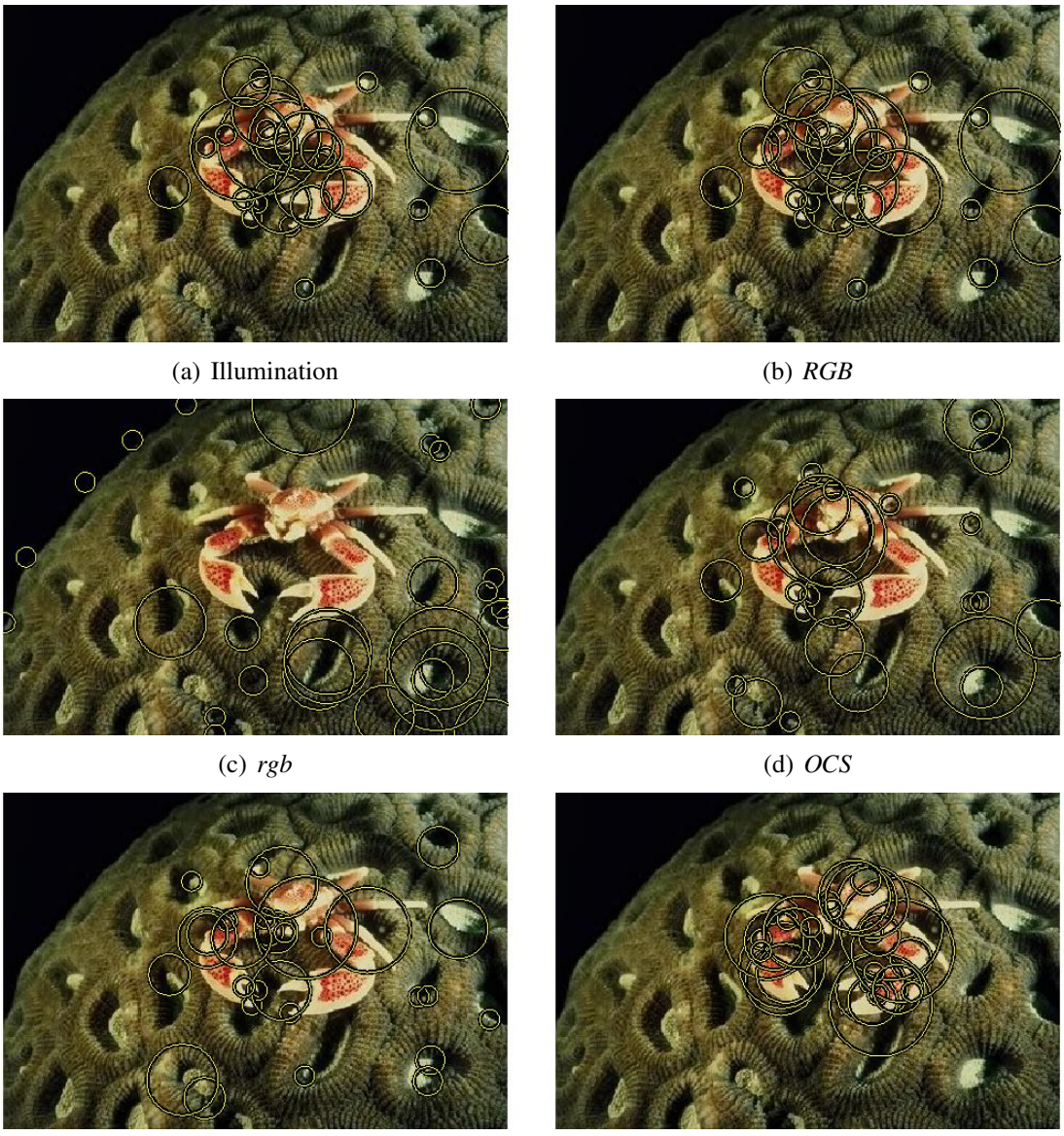
This research uses the previously extracted interest regions (see Section 3.1) and calculates the SIFT description of these regions. Throughout this work, the original SIFT is used. SIFT features have become widely used for both wide-baseline matching and quantized local descriptor approaches and have been found to perform best for many tasks, as evaluated by several authors, e.g. [22, 43, 50, 58, 59, 65, 66, 72]. The SIFT description is calculated by the binary provided by Mikolaczyk[3].

SIFT is chosen because this method is broadly accepted and its performance has been evaluated under many different circumstances. As this thesis investigates the impact of different interest points in a retrieval scenario, the local description must not change during the experiments and it should provide stable, well known results for all the tested interest point detectors under all circumstances.

The illumination based description method lacks in performance regarding colour based locations. When providing colour corners for the SIFT description, the method takes the illumination structure into account. Therefore the discrimination between colour and illumination location diminishes. In special cases, colour corners provide locations where the SIFT methods finds no description at all. This is a drawback of our method of evaluation, because the performance of the colour based approaches decreases under certain circumstances. On the other hand, it provides precise results of the impact of locations for local description.

SIFT analyses the gradient directions of the extracted regions and thresholds theses values. This leads possibly to more than one description for one region when there is another predominant direction within 80 % of the most powerful one. In cases where this illumination only based description finds not enough contrast and therefore no gradients over a certain threshold,

[3] http://www.robots.ox.ac.uk/∼vgg/research/affine/



(a) Illumination (b) *RGB*

(c) *rgb* (d) *OCS*

(e) colour boosted *OCS* (f) Quasi invariant *HSI*

Figure 20: Scale invariant interest point extraction for different colour spaces. The background is highly structured with high illumination changes. $t = 1.2; s = 10; \sigma_I = 0.7$. 30 highest values extracted. Example image from the MUSCLE database

it discards the description.

The key locations are determined at the maxima and minima of Difference of Gaussian function applied in scale space. With image pyramids, this can be done very efficiently. This is described in more detail in Section 2.3.3.

3.4 Descriptor Matching

The task of matching the descriptors in the feature database is done with the Euclidean distance as described in Section 2.4.1. The more sophisticated task of finding the nearest neighbours efficiently and fast is not considered in this software prototype, as it does not affect the overall retrieval result. However, the processing time of the software prototype in a query takes several minutes for a database of 5 000 000 descriptors (see Section 4.3). This would not be tolerable in a real application. On the other hand, the approaches proposed in Section 2.4.2 require heavy calculations in ordering the feature database before the actual queries.

In the following, the algorithm to extract the final retrieval result is presented.

3.4.1 Score of Query

The similarity between two images is determined by first calculating the Euclidean distances between each possible pair of (normalised) descriptors. The minimal distances for each descriptor in the query image to the corresponding database image is then extracted.

$$S(d_q) = min(\overline{d_q d_{d1..dn}}) \tag{59}$$

where $S(d_q)$ gives the score for the matching attempt for one descriptor. For every descriptor in the image, an array $S(d_{1..n})$ is built. d_q is a descriptors of the query image and $d_{d1..dn}$ all the descriptors of the database image which is processed in this step.

Then, the distances are sorted that $S(d_{qn}) \leq S(d_{qn+1})$. The sum of the $N = 100$ smallest distances is then taken to be the distance between the images giving the final similarity between the query image and the processed database image:

$$D(q_d) = \sum_{n=1}^{N} S(d_{qn}) \tag{60}$$

where $D(q_d)$ denotes the overall distance between two images and is the basis for the overall retrieval result. The resulting Chamfer distance $\delta(q)$ of a query image is then a list of the smallest overall distances where $D(q_d n) \leq D(q_d n+1)$.

$$\delta(q) = D(q_d 1), ..., D(q_d n) \tag{61}$$

3.5 File Formats

All data are stored in text files according to the definition suggested by Mikolajczyk. An elliptic region in an image is defined as

$$a(x-u)(x-u)+2b(x-u)(y-v)+c(y-v)(y-v)=1 \tag{62}$$

and is stored in the format

$$\begin{matrix} u_1 & v_1 & a_1 & b_1 & c_1 \\ : & & & & \\ : & & & & \\ u_m & v_m & a_m & b_m & c_m \end{matrix}$$

beginning with $(0,0)$ at the top left corner of the image. Descriptors are then simply added to the corresponding region:

$$\begin{matrix} u_1 & v_1 & a_1 & b_1 & c_1 & d_{1,1} & d_{1,2}...d_{1,n} \\ : & & & & & & \\ : & & & & & & \\ u_m & v_m & a_m & b_m & c_m & d_{m,1} & d_{m,2}...d_{m,n} \end{matrix}$$

Notably, the SIFT descriptor can lead to more than one description per regions (see Section 2.3.3), so there may be regions being defined by multiple descriptors in one file.

This file format is able to store all information needed in the prototype, except for the final retrieval result. After the matching of the descriptors, the ranks of the 100 best matches are stored in a text file. Matches are stored in a list ordered by similarity. The ranks are referred to by their file name and their Chamfer distance $\delta(q)$ as described in Section 3.4.

$$\begin{matrix} f_1 & \delta_1 \\ : & \\ : & \\ f_{100} & \delta_{100} \end{matrix}$$

where $f_{1..100}$ identifies both the object category and the specific image and $\delta_n \leq \delta_{n+1}$.

3.6 Summary

In this section, the methods used for the software prototype are presented and described in detail. Beginning with an overview of the structure of the piece of software, three main stages of the image retrieval software are discussed. Beginning with the interest point extraction, a new

method for automatic scale detection is proposed.

The next stage of the region description gives an overview of the main reasons for the decision of using the SIFT descriptor in the description phase of the software prototype. Drawbacks of the evaluation with this descriptor are pointed out.

Matching a query image with the feature database is the final phase of the CBIR system. More technically, the file formats used and developed for storing and exchanging data between the different stages are defined.

In the next section, the experiments carried out with the software prototype are shown and a large-scale evaluation of the proposed interest point detector is given.

4 Results

The software prototype is now used in different CBIR scenarios. Carrying out diverse experiments, the proposed methods can be evaluated. Beginning with a description of the given image data in Section 4.1, an overview of the experiments is given.

Using a very important property of interest points, the repeatability, performance of different interest region detectors is shown in Section 4.2. Performance in large scale image retrieval scenarios is then tested in Section 4.3. First different scale selection techniques are compared in Section 4.3.2, then retrieval performance under illumination change is evaluated in Section 4.3.3. The largest experiment measures performance under object rotation and is described in Section 4.3.4.

4.1 Image Data

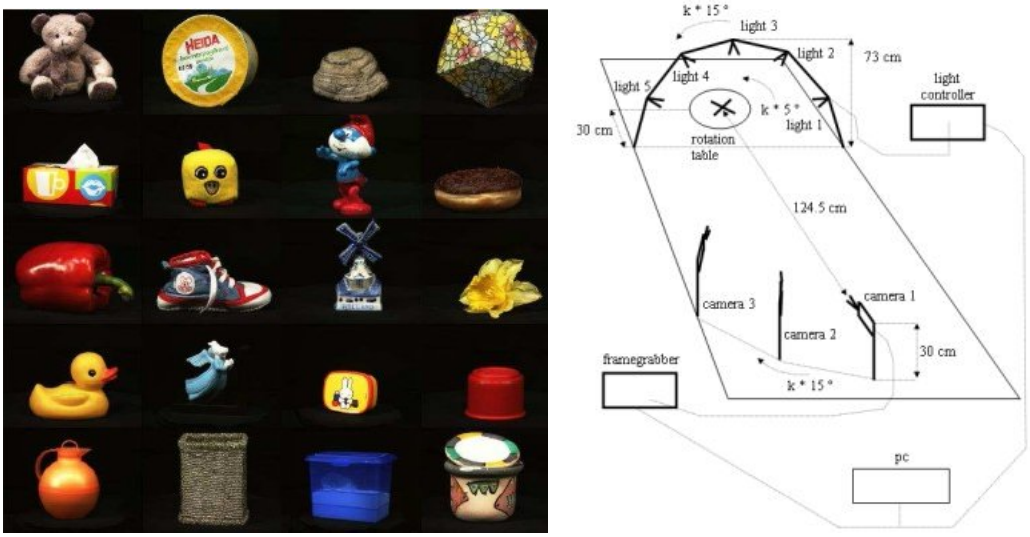
The Amsterdam Library of Object Images (ALOI) provides images of 1000 objects under supervised, predefined conditions on a dark background[4]. Having the possibility to regard single objects only, better conclusions can be drawn than in images under natural circumstances. All transformations are precisely applied, including viewpoint transformation and varying light conditions [25].

As shown in Figure 21(b), 5 OSRAM Tungsten Halogen 64637, 12V, 100W, 3100K lights are placed around a rotation plate. The light controllers (Dimmer Osram HT 1-10 DIM, Transformer Osram HT 150/230/12L) provide stable, deterministic lighting conditions. Pictures are taken with three Sony DXC390P 3CCD with 6 dB gain and Computar lenses of the type 12.5-75 mm, 1:1.2 (settings: f=5.6mm, zoom=48mm (objects 1-750), zoom=15mm (objects 751-1000)). The rotation table is set up for 800 steps per revolution (Parker Hannifin Corporation 20505RT). As the frame grabber, the Matrox Electronic Systems Ltd. CORONA II PCI framegrabber is used. All images are taken from a 124.5 cm distance, the camera positioned at a height of 30 cm.

The database consists of 110 250 PNG images, having a resolution of 768×576 pixels and a colour depth of 24 bpp in the highest resolution. The following data configurations are available:

- **Illumination Direction** provides 24 images in different configurations. Each image was recorded with just one of the five lights turned on, with the three cameras in different positions. Furthermore, combinations of lights were used to illuminate the object. With two lights turned on at the sides of the object, an oblique illumination from right and left is established. Turning on all lights yields a sort of hemispherical illumination, although restricted to a more narrow illumination sector than a true hemisphere.
- **Illumination Colour** provides 12 configurations, all taken with all 5 lights turned on. Colour temperature is successively increased from 2175K to 3075K.

[4] http://staff.science.uva.nl/∼aloi/



(a) Sample of ALOI objects

(b) Configuration of object images recordings

Figure 21: The Amsterdam Library of Object Images (ALOI) is a colour image collection of 1000 small objects, recorded for scientific purposes. From: [25]

- **Object Viewpoint** provides views of an object from 72 different directions. The images are taken by rotating the rotation table in steps of 5 degrees. This collection is similar to the COIL[5] collection.
- **Wide-baseline Stereo** is recorded for 750 images only. The three cameras provide a 15 or 30 degree baseline stereo pair.

With this data, the proposed interest points are evaluated in terms of the repeatability measurement and in a large scale image retrieval context.

4.2 Repeatability Experiment

The experimental setup is the following: a colourful, flat object such as the one shown in Figure 22(a) is turned on a rotation stage in steps of 5 degrees. Rotations up to 50 degrees in both directions are used.

Mikolajczyk provides software to estimate the homography matrix and to evaluate repeatability performance. We used these binaries in order to compare our colour based system with the intensity based approaches. The user selects corresponding pixels in two images, and the 3×3 matrix is calculated.

The performance is measured by the repeatability rate, which is the percentage of corresponding points detected in two images. The higher the repeatability rate between two images,

[5] http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html

(a) front view (b) 25 degree turned (c) 50 degree turned.

Figure 22: ALOI object number 46 under viewpoint transformations.

the more points can potentially be matched and the better are the matching and recognition results. A match is counted if the transformation of one image to the other one using the provided homography matrix leads to interest regions overlapping by more than 40% [59].

The experiments to compare affine region detectors were done with a test set of images of natural scenes [59]. Compared to the precise transformations of the ALOI database, the transformations are done in larger and looser steps. Therefore, the results of these experiments give higher repeatability rates than those with other datasets (e.g [58, 56]). Having simple, precise transformations on flat objects provide a more stable repeatability rate than previous experiments.

Our proposed interest point extraction method is evaluated in the following variations:

- The *Quasi Invariant ScIv Harris* refers to the scale invariant Harris corner detector with the proposed automatic scale decision under the quasi invariant *HSI* colour space.
- *Colour boosted ScIv Harris* uses the proposed approach under the opponent colour space *OCS* and the colour statistics to boost rare colours.
- *RGB ScIv Harris* uses *RGB* information for the proposed scale invariant Harris.

For evaluation, the following state of the art approaches are tested in the experiment:

- The *Harris Laplacian* corner detector is the scale invariant approach from the Mikolajczyk implementation. It is a Harris implementation using LoG for the scale determination.
- *Harris Affine* is an extension of this approach, using the results of the Harris Laplacian algorithm to detect the affine transformation of the region (Mikolajczyk's implementation is used for this algorithm, too).

As shown in Figure 23, the Harris Laplacian detector performs steadily about 5% better than the Harris Affine detector, a result which is explainable by the repeatability criteria. Both approaches use the same locations, as the majority of the algorithm is the same. Just the final stage of converting the scale invariant regions into affine invariant regions diminishes the area

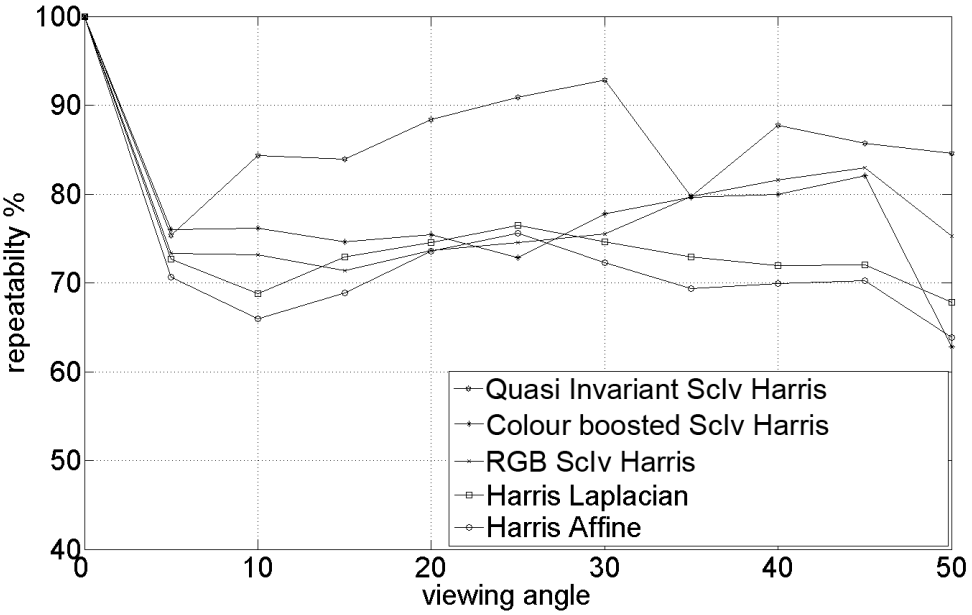


Figure 23: Repeatability experiment with ALOI database on full size resolution. Colour information increases stability under viewpoint transformation on colorful, flat objects.

of the region by the direction of the gradients. These smaller, elliptic regions have sometimes smaller overlapping areas after transforming them back according to the given homography matrix. However, the result remains relatively stable beginning at over 70% and ending below 70% after the 50 degree transformation.

Performing this experiment in the *RGB* colour space, the results are quite similar to the luminance only approaches, until the transformation reaches a level of 35 degrees. From this point, all colour based approaches perform better than those using only luminance information. Apparently, colour corners remain more stable under these transformations. Using colour statistics (Section 2.2.6) and the *OCS* colour space, the salient colour differences become more distinct, and therefore the results improve. A drawback of this method is the instability due to aliasing effects of the transformation, as seen in the 25 degree transformation. Aliasing effects occur when a new colour is introduced into the image because of a transformation of a colour corner. In colour spaces like the quasi invariant *HSI* or in colour boosted spaces, this new colour can cause a major shift in the result.

The quasi invariant colour space performs best, as this approach takes only colour differences into account. Many parts of the image, like the blue shading in the ALOI object nr. 46 seen in Figure 22, are disregarded completely, and only the colour changes between different ones are taken as interest points. This leads to a over 90% repeatability rate at a 30° transformation and an 85% rate after the full 50 degree transformation.

The fact that the measurement increases is also a matter of the diminished area which is regarded after the transformation of the object. When turning the object, only the former front side is regarded for the measurement and therefore gets smaller and less regions are considered.

4.3 Image Retrieval

In this section, the software prototype for image retrieval is tested and evaluated. In retrieval scenarios, the performance of different scale selection algorithms and colour spaces used for interest point extraction is presented and compared.

The image retrieval experiments are carried out with the data described in Section 4.1. The following variations of the proposed interest point extraction method are used:

- The *OCS ScIv Harris* is the proposed scale invariant approach in the opponent colour space.
- The *colour boosted OCS Harris* boosts the opponent colour space with the colour statistics and extracts scale invariant interest points based on this weighted colours.
- The *Quasi Invariant ScIv Harris* uses the quasi Invariant *HSI* colour space as the basis for the interest point extraction.

The next algorithms use the same colour Harris on the same colour spaces and the Laplacian of Gaussian scale selection as described in Section 2.2.3.

- The *Ref. Harris Laplacian* defines the performance of a reference implementation of the Colour Harris with Laplacian of Gaussian scale selection on Illumination information only. This implementation tries to use the same thresholds as the implementation by Mikolajczyk, so as to get comparable results. It uses the proposed Colour Harris and the Laplacian of Gaussian scale selection. Therefore, it is similar to the *Harris Laplacian* approach described below.
- The *colour boosted Ref. Harris Laplacian* uses the reference implementation in the colour boosted *OCS* colour space. It is similar to the *colour boosted OCS Harris* except for the scale decision method.

They are provided by [86] used for evaluation of the proposed scale selection. The rest of the algorithms is similar. For evaluation with other methods, the following state of the art implementations are evaluated in the same scenarios.

- The *Harris Laplacian* indicates the scale invariant Harris approach from the reference implementation by Mikolajczyk. As the scale determination, the Laplacian of Gaussian decision is used..
- *DoG* uses the original SIFT binaries from Lowe[6] for the interest point extraction. This is the original SIFT algorithm as described in Section 2.3.3, where the key point extraction is used for the interest point extraction.

[6] http://www.cs.ubc.ca/∼lowe/keypoints/

All these scale invariant interest points provide the locations for the calculation of the SIFT descriptors (also obtained using the implementation by Mikolajczyk). Therefore, the only difference between the four different retrieval tests is in the interest point extraction stage (see Section 3.1).

The difference between two images is determined by first calculating the Euclidean distances between each possible pair of (normalised) descriptors. The $N = 100$ smallest distances are then taken to calculate the distance between the images. As the retrieval performance measurement, the precision and recall values are calculated for the 30 best matches to the query image (see Section 3.4). The retrieval performance is described in more detail in the following section.

In Section 4.3.2, the scale selection approaches are compared in a retrieval scenario with illumination direction change. In Section 4.3.3 the experiment is taken further in comparing the performance of the colour boosted *OCS* and the *HSI* colour space. Section 4.3.4 describes then the result of the retrieval scenario in object rotation.

4.3.1 Retrieval Performance

As the retrieval performance measurement, the precision and recall values are calculated iteratively for the first results $\delta(q)$ to the query image q. The number of results considered depends on the number of possible true positives in the data set. The values are defined as

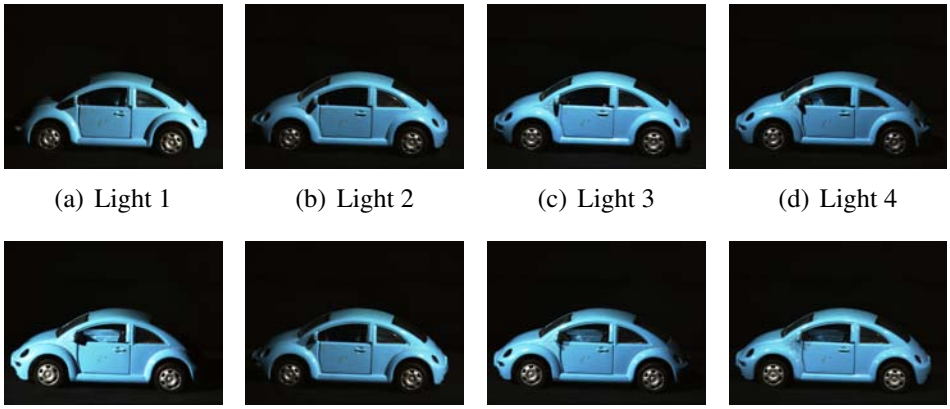
$$\begin{aligned} Recall &= \frac{true\ positive}{true\ positive + false\ negative} \\ Precision &= \frac{true\ positive}{true\ positive + false\ postive} \end{aligned} \tag{63}$$

with the retrieval results categorised as

	true *true result* wrong	
retrieved result true/wrong	true positive	false positive
	false negative	true negative

where $true\ positive$ defines a correct match according to the ground truth, a $true\ negative$ a match correctly marked as a wrong match (e.g. higher $\delta(q)$ than the true positives). False positives are defined as wrong images that are marked as corresponding ones, also known as a *type one error*. This means simply a wrong retrieval result. The most tricky error is the $false\ negative$ which means a correct result not retrieved from the system. This *type two error* is usually considered worse than the other one, as the matching then lacks in similarity and discriminative power.

Recall is defined by the probability that a correct retrieval result occurs, the *precision* is a measurement for the probability that a retrieved result is correct.



(a) Light 1 (b) Light 2 (c) Light 3 (d) Light 4

(e) Light 5 (f) Light 1 and 2 (g) Light 4 and 5 (h) All lights

Figure 24: Illumination direction change on ALOI object nr. 138, camera 1. As the query image, image (e) is used. The total data set consists of 8000 images.

The *precision-recall graphs* used for the visualization of the retrieval results are built iteratively. Each data point in the graph shows the averaged precision and recall over all queries on the database. The data points increase the number of the regarded retrieval results iteratively to show the distribution of the correct results.

4.3.2 Comparison of Scale Selection

The illumination direction change provides images from one object in completely equal position, except that the light sources are changed (see Figure 24). The query image uses the most left light (see Figure 24(e)), so that a lighting change up to 60° can be retrieved (see Figure 24(a)). Other circumstances seen in Figure 24(f) to Figure 24(h) are produced by switching on multiple lights. Therefore, illumination direction and illumination power is changed over the image.

Figure 25 shows the retrieval results using four of the algorithms listed above. As the most popular and widely used, the Harris Laplacian implementation by Mikolajczyk provides the state of the art performance in image retrieval. It is closed source, therefore no precise information can be obtained about the fine implementation details.

The two Ref. Harris Laplacian rely on the approach by Mikolaczyk, but use the proposed Colour Harris corner detector. Therefore, the illumination based Ref. Harris Laplacian should perform equally to the Harris Laplacian. The colour boosted Ref. Harris Laplacian uses the colour boosted *OCS* colour space for the region extraction.

Generally, all the algorithms show a corner in the precision/recall graph after the first seven retrieval results. This is caused by the fact that a perfect retrieval result consists of seven true results. These perfect retrieval results have obviously a perfect precision and recall score. For the other queries, the true results are not disriminative enough with respect to the wrong ones

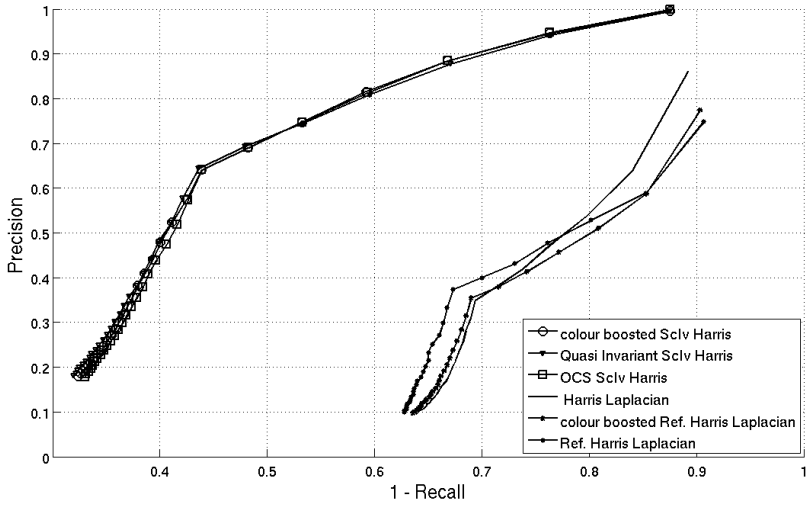


Figure 25: Retrieval performance under varying illumination direction. Both the colour boosting and the quasi invariant colour space based interest point extraction outperform other approaches. The difference between *colour boosted ScIv Harris* and *colour boosted Ref. Harris Laplacian* is in the scale selection only, differences between *Harris Laplacian* and *Ref. Harris Laplacian* are in implementational details only.

or not similar enough to the query image. For this dataset of very similar images, this means that the retrieval result of these cases is unlikely to become better in the majority of the cases. Therefore, the curve changes its shape at the 7th rank.

The original Harris Laplacian Algorithm outperforms the two Ref. Harris Laplacian in precision in the first three retrieval ranks. Therefore, for a very precise retrieval result, e.g. where just one result is taken into account, it is more suitable.

After the first three ranks of the retrieval results, the reference implementation improves the performance and both gain stability in their results, and the slope of the curve decreases. The colour boosted Ref. Harris Laplacian is more precise in its retrieval rate in the first rank compared to the Ref. Harris Laplacian, but is then outperformed. Apparently, the illumination transformation is more stable against smaller changes in the illumination direction than the colour boosting, which may lead to unwanted major changes when the colours are changed due to lighting changes. The proposed scale selection shows better results than the other implementations. The results show to be more stable and in the first ranks, nearly perfect retrieval results are encountered.

This leads to the conclusion that having more stable and distinct localisations lead to better retrieval performance. In many cases, it is not the crucial problem to extract the correct and important localisations, but to avoid the wrong and ambiguous ones.

(a) Uni coloured ALOI object nr. 161 (b) Many fine colour changes on ALOI object nr. 196

Figure 26: Two types of ALOI objects: Colourful ones compared to uni-coloured object hold different challenges for the colour based description.

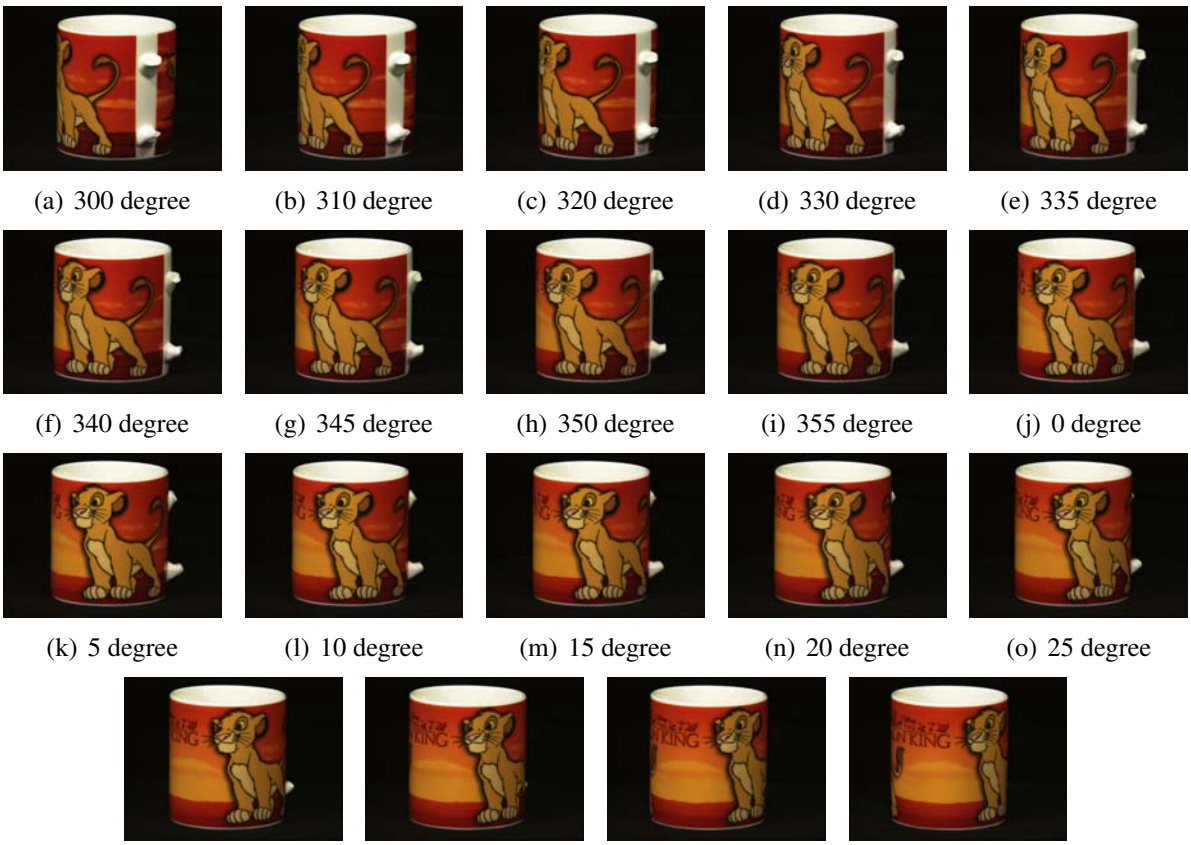
4.3.3 Illumination Direction

This retrieval experiment has been carried out with the same data as the previous one. The main idea of this experiment is to evaluate the impact of different colour spaces in retrieval scenarios with varying lighting conditions.

To overcome problems of illumination changes, the quasi invariant *HSI* colour space has been proposed and showed good results on colourful images. One example can be seen in Figure 26(b), where many different primary colours give stable results even under heavy lightning changes. But as a matter of fact, much information is lost in the transformation, as it discards all specular and shadowing changes and keeps pure colour only. It already showed good performance on colourful images, but for a uni-coloured object as seen in Figure 26(a), nearly all information is lost. The silhoutte is not properly found either, because the black background is not defined in this colour space.

This is the other way around for the colour boosted *OCS*: As it boosts rare colours, it is very likely that it boosts one of the shading effects on a uni coloured surface. Additionally it is not specular, shadowing nor illumination invariant, so many localisations will occur. They will not be stable under transformations, but will be more distinct than pure illumination based methods. On fine multi coloured objects, colour boosting normally outperforms all other methods, as will be shown in the next experiments, because the boosted colour edges remain more stable than other illumination or coloured ones. Unfortunately, in fine colour edges, aliasing effects may occur, which are than amplified by the boosting: Colour edges may provide new colours under transformation, depending on the image quality. New colours may have great impact on the resulting regions in the quasi invariant *HSI* and the colour boosted *HSI*, as the interest points can change unexpectedly.

In having these drawbacks in the new colour spaces, a colour transformation to *OCS* only is compared. As can be seen in Figure 25, the drawbacks of the different colour spaces nearly extinguish each other: As there are several uni coloured objects in the ALOI database, the quasi invariant colour space lacks slightly in distinction, but averaged over 1000 object classes, the performance gain on colourful objects compensates this drawback. The same for *OCS* compared



(a) 300 degree (b) 310 degree (c) 320 degree (d) 330 degree (e) 335 degree

(f) 340 degree (g) 345 degree (h) 350 degree (i) 355 degree (j) 0 degree

(k) 5 degree (l) 10 degree (m) 15 degree (n) 20 degree (o) 25 degree

(p) 30 degree (q) 40 degree (r) 50 degree (s) 60 degree

Figure 27: Data set for ALOI object nr. 290 for the object rotation image retrieval experiment. As the query image, (j) was used. The whole experiment dealt with 19000 images.

to colour boosted *OCS* based approaches: Although they lead to many different results, the averaged retrieval rates are comparable. The more distinct colour boosted edges compensate the aliasing effects, and therefore the two approaches perform nearly equally well, with a slightly better result for the *OCS* ScIv Harris.

4.3.4 Object Rotation

For this retrieval experiment, the impact of the extraction of interest points in a retrieval scenario is examined under object rotation up to 120° as seen in Figure 27.

The retrieval scenario consists of 1000 objects captured as described in Section 4.1. For every object, 9 images are taken rotating the object 60° in both directions. From 5° to 30° and 355° to 330° degree rotation, the steps are taken in 5° increments. Up to 60° and 300°, respectively, the steps are carried out in 10° increments. This results in a dataset of 18000 images with an additional 1000 query images.

Since the ALOI database delivers images of objects on a dark background and image masks

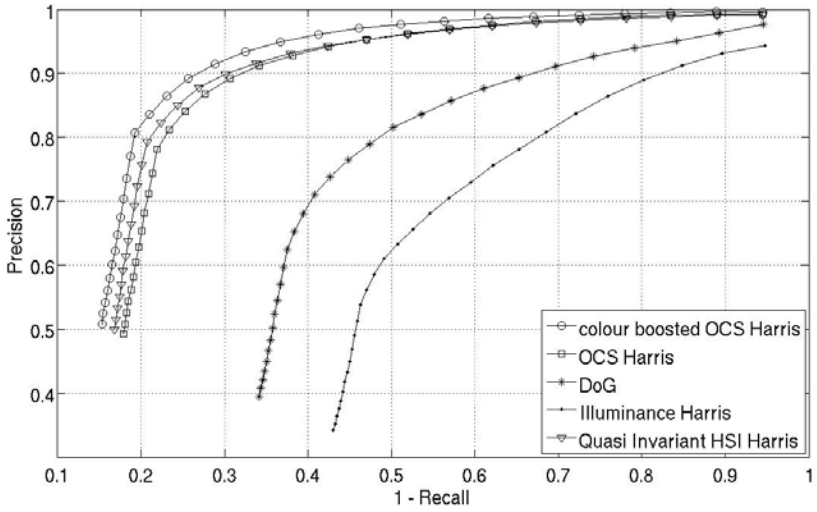


Figure 28: Precision recall graph of the retrieval experiment on the whole ALOI database. All SIFT descriptors are calculated by the Lowe binary, differences in performance are caused by different interest regions only.

to completely disregard the background, the retrieval is obtained by object characteristics only. Query images are captured from the front view, the position which was omitted in the database. Every query image is processed as described above, except that no mask is applied to it. Therefore, it is possible, but not very likely, that descriptors located in the background can occur in the query image.

As shown in Figure 28, the overall performance of the approach based on the colour boosted ScIv Harris is better than the two illumination based methods. The DoG in the original SIFT implementation outperforms the illuminance Harris, especially in low contrast, dark images. It focusses more on the silhouette of the object, and therefore improves the performance in this experiment. The applied colour boosting factors for the colour boosted OCS improve the ability to describe an object under heavy viewpoint changes, as the performance is steadily better than in OCS ScIv Harris.

For colourful objects, the Quasi Invariant ScIv Harris provides better results than the other colour based approaches, as the colour edges remain stable even under heavy transformations. Unfortunately, not all the objects are colourful (compare Section 4.3.3 and Figure 26(a)) and therefore, the better performance is diminished on these objects and the approach gains the second best performance in this experiment.

The black background provides predominantly dark images lightend with constant lighting under all the transformations. From this point of view, the illumination only based methods should not have problems in stability in changing contrast. In another context, the illumination based methods suffer in great instability: when the rotating object moves a surface away from the light source, the illumination on this surface gets less and less as there is almost no ambient

light. This is the case for objects which provide self-shadows while turning. In these cases, the colour boosted OCS based method performs better, as the colours get more distinctive, especially under dark circumstances. Illumination based approaches tend to lose meaningful locations under these illumination changes.

However, in this large scale experiment the need for a stable scale selection and meaningful colour spaces are shown. The performance of retrieval scenarios can be raised by focussing on distinct localisations only. The discriminative power of corners based on colour differences are shown on a large dataset with 1000 classes, which is a huge number for retrieval experiments.

4.4 Summary

In this section, the proposed approaches for interest point extraction are tested and compared to existing implementations. Beginning with a general overview of the impact of using different colour spaces for corner extraction, a discussion of the focus of corners is started.

Repeatability experiments on various colour spaces, which showed to gain performance on corners of colour differences, different scale determination approaches are evaluated under illumination changes.

When testing in image retrieval scenarios, the difficult question is to find the real reason for certain performance changes. In our experiments, the precise circumstances of the ALOI database provide fixed image attributes, which leads to an objective discussion of the different localisation properties. Always using the same local descriptors for different localisations, the impact of this stage in the retrieval scenario is evaluated.

For all the experiments, the proposed approaches outperformed existing approaches in both preciseness and overall performance. This evaluations under different, predefined circumstances lead us into a promising conclusion and outlook for future work in the next and final section.

5 Conclusion

In this section, a summary of this thesis is provided. Related work and the main contribution are discussed. Results of the proposed approches are discussed from a general point of view. Section 5.1 gives the overview of this thesis.

Having these new interest points, many new questions and possibilites arise: Some directions for future work are given in Section 5.2, as the new method can be used in many other contexts. Regarding the results, the proposed algorithms may give good results in other applications, too.

5.1 Summary and Discussion

In this thesis, the main topic is image retrieval. Focussing on local interest point, the question of *what* to describe is discussed and several approaches from computer science and neurobiology are described. A state of the art of visual saliency, which builds the basis for this work, starts the overview of automatic approaches to find salient regions in image data. Going more into history of computer vision, the very successful corner detection approach is developed chronologically.

Having an overview of local description and some approaches for descriptor matching, all stages of an typical image retrieval scenario are covered. The main issues of these stages are pointed out and certain drawbacks and advantages of different colour spaces are given.

Equipped with these approaches, a software prototype is developed. Explaining why to take certain colour properties, the piece of software is described in detail and the main contribution is given. A new scale selection for Harris corners gives less, but more distinct, better distributed and more stable localisations than previous methods. It uses arbitrary colour information for its scale decision, as it processes the distribution of the colour space and not the colour information itself. Therefore, a relaxed visual saliency function is included in the localisation decision.

Extensive experiments are carried out with different interest point extraction algorithms. The proposed scale selection is compared to the existing one, repeatability experiments are carried out and image retrieval experiments on large data sets are given. In all the experiments, the proposed method outperforms existing methods, both using illumination or colour information.

Using colour distances for corner measurement can shift the interest points to more equally distributed, stable and distinct locations than luminance based methods. A colour scale selection leads to a better stability under transformations. Both the corner measurement and the scale selection can be transformed into various colour spaces, and advantage can be taken of different properties of these transformations. Using correlated colour, boosted colour or colour invariant information, the method gains performance over illuminance based methods.

5.2 Outlook

In addition to the results presented in this thesis, some question arise: the most obvious one is the question for colour local description. There are lots of possibilities to add colour information to the description phase, which should provide more discriminative results in image retrieval scenarios. Then, the gathered information in the localisation phase can have full impact on the retrieval result.

The software prototype can be redeveloped to a high performance application, to carry out more experiments in less time. Here, the first issue is to add some faster matching algorithms to overcome the main bottleneck of the implementation.

The extension of other localisation extraction algorithms to variable colour spaces could be another promising field of research. This could lead to more distinct and stable localisation without corners. One example would be the very powerful MSER approach and using the quasi invariant colour space for image segmentation.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 774–781, 2000.

[2] H. Bay, T. Tuytelaars, and L. van Gool. SURF: Speeded up robust features. In *Proceedings of the ninth European Conference on Computer Vision*, pages 404–417, May 2006.

[3] A. Berengolts and M. Lindenbaum. On the distribution of saliency. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 543–549. IEEE Computer Society, 2004.

[4] D. Berg. What is an image and what is image power? In *Image and Narrative*, volume 8, May 2004.

[5] I. Biederman. Recognition-by-components: A theory of human image understanding. In *Psychological Review*, volume 94, pages 115–147. American Psychological Association, 1987.

[6] S. Borer and S. Süstrunk. Opponent color space motivated by retinal processing. In *Proc. IS&T Conference on Color in Graphics, Imaging and Vision (CGIV)*, volume 1, pages 187–189, 2003.

[7] A. Bradley and F. Stentiford. Visual attention for region of interest coding in JPEG 2000. In *Journal of Visual Communication and Image Representation*, volume 14, pages 232–250, September 2003.

[8] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference, BMVC 2002, Cardiff*, pages 656–665, 2002.

[9] R. Carmi and L. Itti. Causal saliency effects during natural vision. In *Proc. ACM Eye Tracking Research and Applications*, pages 1–9, 2006.

[10] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: image segmentation using expectation-maximization and its application to image querying. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, volume 24, pages 1026–1038, 2002.

[11] C.Kenney, M. Zuliani, and B. Manjunath. An axiomatic approach to corner detection. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2005.

[12] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proc. ECCV*, pages 16–29, 2006.

[13] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262. ACM Press, 2005.

[14] A. del Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 19, pages 121–132. IEEE Computer Society, 1997.

[15] R. Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. In *Philosophical Transactions of the Royal Society B: Biological Sciences*, volume 353, pages 1245 – 1255, August 1998.

[16] G. Dorkó and C. Schmid. Maximally stable local description for scale selection. In *European Conference on Computer Vision, Graz, Austria*, volume 4, pages 504–516, 2006.

[17] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *CVPR*, pages 612–618, 2000.

[18] S. Edelman, N. Intrator, and T. Poggio. Complex cells and object recognition. unpublished script, http://kybele.psych.cornell.edu/∼edelman/archive.html, 1997.

[19] C. Elkan. Using the triangle inequality to accelerate kmeans. In *In Proceedings of the Twentieth International Conference on Machine Learning*, pages 147–153, 2003.

[20] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 264–272, 2003.

[21] G. Finlayson, B. Schiele, and J. Crowley. Comprehensive colour image normalization. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I*, pages 475–490. Springer-Verlag, 1998.

[22] F. Fraundorfer and H. Bischof. A novel performance evaluation method of local detectors on non-planar scenes. In *Workshop Proceedings Empirical Evaluation Methods in Computer Vision, Conference on Computer Vision and Pattern Recognition (CVPR)*, page 33. IEEE Computer Society, 2005.

[23] J. French, J. Watson, X. Jin, and W. Martin. An exogenous approach for adding multiple image representations to content-based image retrieval systems. In *Seventh International Symposium on Signal Processing and its Applications*, volume 1, pages 201–204. ISSPA, July 2003.

[24] J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. In *ACM Trans. Math. Softw.*, volume 3, pages 209–226. ACM Press, 1977.

[25] J. Geusebroek, G. Burghouts, and A. Smeulders. The Amsterdam library of object images. In *International Journal of Computer Vision*, volume 61, pages 103–112, 2005.

[26] C. Granrud. Visual size constancy in newborn infants. In *Invest. Ophthalmology & Visual Science*, volume 28, pages 265–315, 1987.

[27] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings 4th Alvey Visual Conference, UK*, volume 15, pages 147–151, 1988.

[28] D. D. Hoffman and W. A. Richards. Parts of recognition. In *Readings in computer vision: issues, problems, principles, and paradigms*, pages 227–242. Morgan Kaufmann Publishers Inc., 1987.

[29] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. In *The Journal of Physiology*, volume 160, pages 106–154.2, January 1962.

[30] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA*, volume 3644, pages 473–82. SPIE Press, Jan 1999.

[31] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. In *Vision Research*, volume 40, pages 1489–1506, May 2000.

[32] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 20, pages 1254–1259. IEEE Computer Society, 1998.

[33] T. Kadir and M. Brady. Saliency, scale and image description. In *International Journal of Computer Vision*, volume 45, pages 83–105. Kluwer Academic Publishers, 2001.

[34] T. Kato. Data-base architecture for context-based image retrieval. In *Proc. SPIE: Image storage and retrieval systems*, pages 112–123, 1992.

[35] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, 2004.

[36] B. Kim and S. Park. A fast k nearest neighbor finding algorithm based on the ordered partition. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 8, pages 761–766. IEEE Computer Society, 1986.

[37] J. Koenderink and A. Doom. Representation of local geometry in the visual system. In *Biological Cybernetics*, volume 55, pages 367–375. Springer-Verlag New York, Inc., 1987.

[38] S. Kuffler. Discharge patterns and functional organization of mammalian retina. In *Journal of Neurophysiology*, volume 16, pages 37–68, January 1953.

[39] G. Lance and W. Williams. A general theory of classificatory sorting strategies II: Clustering systems. In *Computer Journal*, volume 10, pages 271–277. British Computer Society, 1967.

[40] T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen. Content-based image retrieval in medical applications: a novel multistep approach. In M. M. Yeung, B.-L. Yeo, and C. A. Bouman, editors, *Proceedings of SPIE: Storage and Retrieval for Media Databases 2000*, volume 3972, pages 312–320, 2000.

[41] B. Leibe, K. Mikolajczyk, and B. Schiele. Efficient clustering and matching for object class recognition. In *British Machine Vision Conference (BMVC)*, volume 2, pages 789–799, September 2006.

[42] J. Lettvin, H. Maturana, W. McCulloch, and W. Pitts. What the frog's eye tells the frog's brain. In *In Proceedings of IRE*, volume 47, pages 1940–1951, 1959.

[43] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531. IEEE Computer Society, 2005.

[44] J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 25, pages 1075–1088. IEEE Computer Society, 2003.

[45] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[46] T. Lindeberg. Feature detection with automatic scale selection. In *International Journal of Computer Vision*, volume 30, pages 77–116, 1998.

[47] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. In *Image and Vision Computing*, volume 15, pages 415–434, June 1997.

[48] E. Loupias and N. Sebe. Wavelet-based salient points for image retrieval. Technical Report RR 99.11, Laboratoire Reconnaissance de Formes et Vision., 1999.

[49] D. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.

[50] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 60, pages 91–110, 2004.

[51] B. Mandelbrot. How long is the coast of Britain? statistical self-similarity and fractional dimension. In *Science*, volume 156, pages 636–638, May 1967.

[52] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *Image and Vision Computing*, volume 22, pages 761–767, September 2004.

[53] K. Matkovic, L. Neumann, J. Siglaer, M. Kompast, and W. Purgathofer. Visual image query. In *SMARTGRAPH '02: Proceedings of the 2nd international symposium on Smart graphics*, pages 116–123. ACM Press, 2002.

[54] L. Miclet and M. Dabouz. Approximative fast nearest neighbor recognition. In *Pattern Recognition Letters*, volume 1, pages 277–285, 1983.

[55] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 525–531, 2001.

[56] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision*, volume 1, pages 128–142, 2002.

[57] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. In *International Journal of Computer Vision*, volume 60, pages 63–86. Kluwer Academic Publishers, 2004.

[58] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 27, pages 1615–1630. IEEE Computer Society, 2005.

[59] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Gool. A comparison of affine region detectors. In *International Journal of Computer Vision*, volume 65, pages 43–72, 2005.

[60] P. Montesinos, V. Gouet, and R. Deriche. Differential invariants for color images. In *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition*, volume 1, pages 838–841. IEEE Computer Society, 1998.

[61] H. Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. Tech. report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University, doctoral dissertation, Stanford University, September 1980.

[62] S. Palmer. Hierarchical structure in perceptual representation. In *Cognitive Psychology*, volume 9, pages 441–474, October 1977.

[63] D. Parkhurst and E. Niebur. Scene content selected by active vision. In *Spatial Vision*, volume 6, pages 125–154, 2003.

[64] K. Popat and R. Picard. Cluster based probability model and its application to image and texture processing. In *IEEE Trans. Image Processing*, volume 6, pages 268–284, 1997.

[65] P. Quelhas. *Scene image classification and segmentation with quantized local descriptors and latent aspect modeling*. PhD thesis, 2007.

[66] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 1, pages 883–890. IEEE Computer Society, 2005.

[67] V. Ramasubramanian and K. Paliwal. A generalized optimization of the k-d tree for fast nearest-neighbour search. In *TENCON '89. Fourth IEEE Region 10 International Conference*, pages 565–568, Nov 1989.

[68] I. Rock and J. DiVita. A case of viewer-centered object perception. In *Cognitive Psychology*, pages 280–293, 1987.

[69] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[70] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 414–431. Springer-Verlag, 2002.

[71] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, pages 530–535. IEEE Computer Society, 1997.

[72] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. In *International Journal of Computer Vision*, volume 37, pages 151–172, 2000.

[73] A. Schubert. Plato: Der Staat. Paderborn, 1995.

[74] N. Sebe, T. Gevers, S. Dijkstra, and J. Weijer. Evaluation of intensity and color corner detectors for affine invariant salient regions. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 18. IEEE Computer Society, 2006.

[75] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1470. IEEE Computer Society, 2003.

[76] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 22, pages 1349–1380. IEEE Computer Society, 2000.

[77] S. Smith and J. Brady. Susan - a new approach to low level image processing. In *Int. Journal of Computer Vision*, volume 15, pages 76–84, May 1997.

[78] F. Stentiford. An evolutionary programming approach to the simulation of visual attention. In *Proceedings of the 2001 Congress on Evolutionary Computation CEC2001*, pages 851–858. IEEE Press, 27-30 2001.

[79] F. Stentiford. Attention based similarity. In *Pattern Recognition*, volume 40, pages 771–783, 2006.

[80] F. Stentiford and M. D. Walker. Attention based color correction. In *SPIE Human Vision and Electronic Imaging*, volume 6057, pages 58–167, January 2006.

[81] J. Stöttinger, N. Sebe, T. Gevers, and A. Hanbury. Colour interest points for image retrieval. In *Proceedings of the 12th Computer Vision Winter Workshop*, pages 83–91, February 2007.

[82] Q. Tian, N. Sebe, M. Lew, E. Loupias, and T. Huang. Content based image retrieval using wavelet-based salient points. In *Journal of Electronic Imaging*, volume 10, pages 835–849, October 2001.

[83] S. Treue. Visual attention: the where, what, how and why of saliency. In *Curr Opin Neurobiol*, volume 13, pages 428–432, August 2003.

[84] J. Tsotsos. An inhibitory beam for attentional selection. In *Proceedings of the 1991 York conference on spatial vision in humans and robots*, pages 313–331. Cambridge University Press, 1993.

[85] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. In *Artif. Intell.*, volume 78, pages 507–545. Elsevier Science Publishers Ltd., 1995.

[86] K. E. A. van de Sande. Coloring concept detection in video using interest regions. Master's thesis, University of Amsterdam, March 2007.

[87] J. van der Weijer and T. Gevers. Color constancy based on the grey-edge hypothesis. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 972–983, Oct 2005.

[88] J. van der Weijer and T. Gevers. Edge and corner detection by photometric quasi-invariants. In *IEEE Transactions on Pattern Analysis & Machine Intelligence*, volume 27, pages 625–630, 2005.

[89] J. van der Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. In *IEEE Transactions on Pattern Analysis & Machine Intelligence*, volume 28, pages 150–156, Jan 2006.

[90] J. van der Weijer, T. Gevers, and A. Smeulders. Robust photometric invariant features from the color tensor. In *IEEE Transactions on Image Processing*, volume 15, pages 118–127,, Januar 2006.

[91] J. van der Weijer and C. Schmid. Coloring local feature extraction. In *European Conference on Computer Vision*, volume Part II, pages 334–348. Springer, 2006.

[92] R. Veltkamp and M. Tanase. A survey of content-based image retrieval system. In *Content-based image and video retrieval, by Oge Marques and Borko Furht*, pages 47–101. Kluwer Academic Publishers, 2002.

[93] D. Vernon. Cognitive vision - the development of a discipline. In *European Research Network for Cognitive AI-enabled Computer Vision Systems, EC Vision project report*, August 2004.

[94] M. Wertheimer. Experimentelle Studien über das Sehen von Bewegung. In *Zeitschift für Psychologie*, number 61, pages 161–265, 1912/1932.

[95] P. Witkin. Scale-space filtering. In *8th Int. Joint Conf. Artificial Intelligence*, volume 2, pages 1019–1022, August 1983.

[96] C. Wolf, J.-M. Jolion, W. Kropatsch, and H. Bischof. Content based image retrieval using interest points and texture features. In *Proceedings of the ICPR2000*, volume 4, pages 234–237. IEEE Computer Society, September 2000.