Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Detection and Evaluation Methods for Local Image and Video Features

DISSERTATION

ausgeführt zum Zwecke der Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Julian Stöttinger
Matrikelnummer 9926328

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuer: Priv-Doz. Dr. Allan Hanbury

Wien, 15. November 2010 _____ _____

(Unterschrift Verfasser) (Unterschrift Betreuer)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Erklärung zur Verfassung der Arbeit

Julian Stöttinger
Mariahilferstr 142/1/12
1150 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. November 2010

(Unterschrift Verfasser)

Kurzfassung

Lokale Features, also räumlich begrenzte Beschreibungen von visuellem Inhalt, sind in der Computer Vision das Werkzeug der Wahl zum Erkennen von Bildern und Videos. Diese Doktorarbeit beschäftigt sich mit dem Auffinden der besten Positionen und der richtigen Skalierung von lokalen Features. Der wissenschaftlicher Beitrag der Arbeit besteht zum einen im Entdecken von neuen Wegen, die Lage von Features zu bestimmen, zum anderen im Erkunden von neuen Evaluierungsmethoden derselbigen. Dies beinhaltet sowohl rein räumliche Features ("2D" oder Bildfeatures) als auch räumlich-zeitliche Features ("3D" oder Videofeatures).

Die Arbeit zeigt, dass das Auffinden von robusten und wiedererkennbaren Features auf die Erkennungsrate von aktuellen Klassifikationssystemen einen großen Einfluss hat. Deswegen ist es entscheidend, die richtigen Features für bestimmte Aufgaben zu nutzen. Im Bereich der Bildfeatures beschäftigt sich die Arbeit mit der Frage, ob es möglich ist, die Anzahl der Features zu reduzieren und gleichzeitig die Erkennungsrate zu erhalten. Da die Featureextrakion den ersten Schritt eines Klassifizierungssystems darstellt, reduziert ein Minimum der Features jeden nachfolgenden Berechnungsschritt und verkürzt so die Berechnungszeit entscheidend. In Bereichen, wo Rechenzeit sehr knapp bemessen ist, könnte dies neue Anwendungen ermöglichen, zum Beispiel in mobilen – und Echtzeit Systemen.

Forschungsschwerpunkt ist das Nutzen von Farbinvarianzen und –salienzen beim Auffinden von skalierungsunabhängigen Bildfeatures. Diese neuartigen Bildfeatures erweisen sich als äußerst stabil gegen Veränderung durch Beleuchtung und Schatten und erlauben so eine robustere Beschreibung des Bildinhaltes. Mit dieser Methode können Bilder in großen Datenbanken mit weniger Features leichter gefunden werden. In einem internationalen Wettbewerb von aktuellen Bildfeatures erreichten die vorgestellten farbbasierten Bildfeatures die beste Erkennungsrate in vier von 20 Klassen, während gleichwertige Methoden ein Vielfaches der Anzahl der Features benutzten.

Weiters untersucht die Arbeit den Gradient Vector Flow zum Finden von Bildfeatures. Da diese Methode Bildstrukturen im größeren Umfang als bestehende Methoden untersucht, erlaubt sie eine äußerst stabile, skalierungsunabhängige Featureextraktion von hoher Dichte.

In den letzten Jahren wurde das Auffinden von stabilen Videofeatures ein begehrtes Forschungsgebiet der Computer Vision. Die erfolgreichsten Bildfeatures wurden um die zeitliche Dimension erweitert. Klassifizierungssysteme erlauben an Hand dieser Features das Erkennen von Handlungen in Videos. Im Unterschied zu Bildfeatures wurden Videofeatures noch nicht in einer stringenten und systematischen Art und Weise auf ihre Robustheit untersucht. Diese Arbeit schließt diese Lücke und stellt eine neuartige Datenbank zur Verfügung. Diese Datenbank von 1710 Videos erlaubt Forschern neue Videofeatures unter acht verschiedenen, wohl definierten, iterativen Veränderungen der Videos zu testen.

Die Evaluierung erfolgt mit einem effizienten 3D Repeatablity Test für Videofeatures. Um die Robustheit von lokalen Beschreibungen in Videos zu messen, wurde ein neuartiges Verfahren entwickelt, das die Evaluierung der Videobeschreibung abhängig von den Veränderungen in den Videos ermöglicht.

Abstract

In computer vision, local image descriptors computed in areas around salient interest points are the state-of-the-art in visual matching. This doctoral thesis aims at finding more stable and more informative interest points in the domain of images and videos. The research interest is the development of relevant evaluation methods for visual matching approaches. The contribution of this work lies on one hand in the introduction of new features to the computer vision community. On the other hand, there is a strong demand for valid evaluation methods and approaches gaining new insights for general recognition tasks. This work presents research in the detection of local features both in the spatial ("2D" or image) domain as well for spatio-temporal ("3D" or video) features.

For state-of-the-art classification the extraction of discriminative interest points has an impact on the final classification performance. It is crucial to find which interest points are of use in a specific task. One question is for example whether it is possible to reduce the number of interest points extracted while still obtaining state-of-the-art image retrieval or object recognition results. This would gain a significant reduction in processing time and would possibly allow for new applications e.g. in the domain of mobile computing.

Therefore, the work investigates different corner detection approaches and evaluates their repeatability under varying alterations. The proposed sparse color interest point detector gives a stable number of features and thus a better comparable image representation. By taking the saliency of color information and color invariances into account, improved retrieval of color images, being more stable to lighting and shadowing effects than using illumination correlated color information, is obtained. In an international benchmark the approach outperforms all other participants in 4 out of 20 classes using a fractional amount of features compared to other approaches.

The Gradient Vector Flow (GVF) has been used with one manually adjusted set of parameters to locate centers of local symmetry at a certain scale. This work extends this approach and proposes a GVF based scale space pyramid and a scale decision criterion to provide general purpose interest points. This multi-scale orientation invariant interest point detector has the aim of providing stable and densely distributed locations. Due to the iterative gradient smoothing during the computation of the GVF, it takes more surrounding image information into account than other detectors.

In the last decade, a great interest in evaluation of local visual features in the domain of images is observed. Most of the state-of-the-art features have been extended to the temporal domain to allow for video retrieval and categorization using similar techniques as used for images. However, there is no comprehensive evaluation of these.

This thesis provides the first comparative evaluation based on isolated and well defined alterations of video data. The aim is to provide researchers with guidance when selecting the best approaches for new applications and data-sets. A dedicated publicly available data-set of 1710 videos is set up, with which researchers are able to test their features' robustness against well defined challenges.

For the evaluation of the detectors, a repeatability measure treating the videos as 3D volumes is developed. To evaluate the robustness of spatio-temporal descriptors, a principled classification pipeline is introduced where the increasingly altered videos build a set of queries. This allows for an in-depth analysis of local detectors and descriptors and their combinations.

Acknowledgments

I am genuinely grateful to my supervisor Allan Hanbury. With your congeniality you always managed to keep my motivation and curiosity up. Special thanks goes out to my great mentors Nicu Sebe and Theo Gevers. The last four years have been such a great adventure, I wouldn't want to miss a bit. All my appreciation to Robert Sablatnig, who is leading one of the best equipped and social working environments ever; yes, the CVL. A good place to be.

Thanks to Michael Brandstötter and Martin Kampel for the experience to be part of the very dawn of the CogVis Ltd.; and for letting me be me throughout the years. It is awesome to see ideas become products.

Special kudos to all the nerds I wrote papers with! Without you, everything would be different. Rehanullah Khan, René Donner, Lech Szumilas, Jana Machajdik, Thomas Pönitz, Sebastian Zambanini, Bogdan Tudor Goras, Christian Liensberger, Marco Vanossi, Ivo Everts - you all are awesome, guys. Working alone is boring and pointless anyways.

Andi Müller, Markus Diem, Alex Dorfmeister, Flo Kleber, Angelika Garz, Stefan Fiel, Katharina Pois, Andi Zweng, Michi Smolle, Sigrid Elsinger, Peter Füreder, Naeem Bhatti, Jürgen Konetschnig, Roland Leitner, Georg Langs, Erich Birngruber, Thomas Matzke, Michi Hödlmoser, Martin Lettner, Adrian Ion, Yll Haxhimusa, Rainer Planinc, thanks for all the help, coffee, cigarettes, beers, parties, darts, table soccer, jokes and the good mood.

All the best to my family Sophie, Claudia and Toni - and my extended family Michi Englert, Michi Moser, Rita Wenzel and Lukas Großmeister. Kitz, quite a project we are having! I couldn't do it with anybody else. Finally, Sarah... you changed the whole game. Whatever tomorrow brings, we'll be there.

Contents

CHAPTER 1

Introduction

Interest points are the first stage of robust visual matching applications and build the state-of-the-art of visual feature localization. The interest point detectors allow for the reduction of computational complexity in scene matching and object recognition applications by selecting only a subset of image locations corresponding to specific and/or informative structures [Mikolajczyk and Tuytelaars, 2009]. The extraction of stable locations is a successful way to match visual input in images or videos of the same scene acquired under different conditions [Mikolajczyk et al., 2005b]. As evaluated in [Mikolajczyk and Schmid, 2004], successful approaches extracting stable locations rely on corner detection [Harris and Stephens, 1988; Mikolajczyk and Schmid, 2001], blobs like Maximally Stable Extremal Regions (MSER) [Matas et al., 2002], Difference of Gaussians (DoG) [Lowe, 2004] or detecting local symmetry [Loy and Zelinsky, 2003].

The majority of interest point extraction algorithms are purely intensity based [Harris and Stephens, 1988; Kadir and Brady, 2001; Mikolajczyk and Schmid, 2004]. This ignores saliency information contained in the color channels. It is known that the distinctiveness of color based interest points is larger, and therefore color is of importance when matching images [Sebe et al., 2006a]. Furthermore, color plays an important role in the pre-attentive stage in which features are detected [Itti et al., 1998; Sebe et al., 2006b] as it is one of the elementary stimulus features [van der Velde et al., 2004].

Fergus et al. [Fergus et al., 2003] point out that a categorization framework is heavily dependent on the detector to gather useful features. In general, the current trend is toward increasing the number of points [Zhang et al., 2007], applying several detectors or combining them [Mikolajczyk et al., 2006; Sivic et al., 2005], or making the interest point distribution as dense as possible [Tuytelaars and Schmid, 2007]. While such a dense sampling has been shown to be effective in object recognition, these approaches basically shift the task of discarding the non-discriminative points to the classifier [Cantu-Paz, 2002]. Further, dense sampling implies that a huge amount of data must be extracted from each image and processed. This is feasible when executed on a cluster of computers in a research environment. Nevertheless, there are environments in which the luxury of extensive computing power is not available. This is illustrated

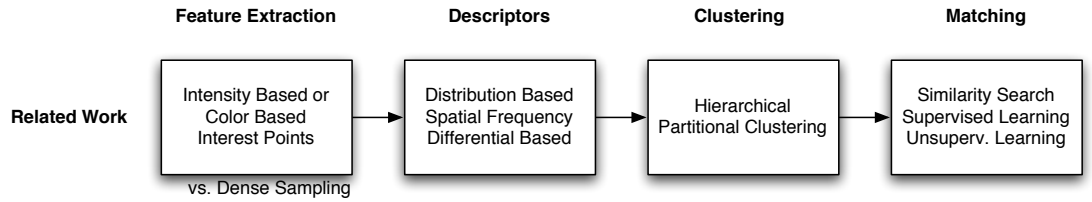


Figure 1.1: The main stages of image matching.

by the strong trend towards mobile computing on Netbooks, mobile phones and PDAs. With growing datasets, clustering and off-line training of features become infeasible [Schindler et al., 2007].

Therefore, there is a strong interest to exploit state-of-the-art classification and to focus on the extraction of discriminative interest points. Following the main idea of interest points to reduce the data to be processed, one important question is whether it is possible to reduce the number of interest points extracted while still obtaining state-of-the-art image retrieval or object recognition results. Recent work aims at finding discriminative features e.g. by performing an evaluation of all features within the dataset or per image class and choosing the most frequent ones [Turcot and Lowe, 2009]. All these methods require an additional calculation step with an inherent demand for memory and processing time dependent on the number of features.

In this thesis, color salient and invariant information is used for image point selection with the aim to achieve state-of-the-art performance while using significantly fewer interest points. The difference to prior work is that feature selection takes place at the very first step of the feature extraction and is carried out independently per feature. In contrast to feature selection that is taken care of by the classification step only, this method of regarding color saliency or color invariance provides a diminished amount of data for subsequent operations.

Techniques such as the bags-of-words approach are originally inspired by text retrieval. These have been extended to “2D” techniques on images and build the state-of-the-art in image matching. These approaches are now successfully carried out in both the spatial and the temporal domains for action recognition, video understanding and video matching (e.g. [Schüldt et al., 2004; Laptev et al., 2008; Duchenne et al., 2009; Junejo et al., 2010]).

The most successful approaches to video understanding and video matching use local spatio-temporal features as a sparse representation for video content. In these approaches, videos are treated as “3D” volumes and the detection and description stage is carried out in “3D" as well. Extracting features diminishes the data to be processed and aims to provide a sparse and robust representation of the video content. Common in these works is this first step of deciding on the regions to be described.

In this thesis, the following four main stages of visual matching applications are denoted. A diagram is given in Fig. 1.1.

Feature extraction is carried out with either global or local features. Global features lack invariance against occlusions and cropping but are a powerful tool in certain applications including image matching (e.g. [Torralba et al., 2008; Srinivasan and Sawant, 2008]) and provide a fast and efficient way of image representation. Local features are located upon either intensity

based or color based interest points. They undertake the task of deciding which parts of the visual input are used for further processing and which parts are discarded right away. Especially for the vast amount of data of video matching, the majority of the visual data can be disregarded right away (e.g. [Laptev and Pérez, 2007; Marszalek et al., 2009; Junejo et al., 2010]). Recently, dense sampling of local features achieved good performance especially for the bags-of-words approach and robust learning systems [Mikolajczyk et al., 2006; Tuytelaars and Schmid, 2007; van de Sande et al., 2009; Mikolajczyk et al., 2009].

Descriptors characterize the image information steered by the feature extraction. They are categorized in three classes: They describe the distribution of certain properties of the visual information (e.g. SIFT), spatial frequency (e.g. wavelets) or other differentials (e.g. local jets) [Mikolajczyk and Schmid, 2005]. For every feature extracted, a description is computed. A disadvantage is that the run-time increases with their number. However, efficient ways to calculate these descriptors exist, e.g. for features with overlapping areas of support, previously calculated results can be used.

Clustering for signature generation, feature generalization or vocabulary estimation assigns the descriptions into a subset of categories. There are hierarchical and partitional approaches to clustering. Due to the excessive memory and run-time requirements of hierarchical clustering [Jain et al., 1999], partitional clustering, such as the k-means, is the method of choice in creating feature signatures.

Matching summarizes the classification of the extracted features. Image descriptors are compared with previously learned and stored models. This is computed by a similarity search or by building a model based on supervised or unsupervised learning techniques. Classification approaches need feature selection to discard irrelevant and redundant information [Okada and Soatto, 2008; Dorko and Schmid, 2003; Jurie and Triggs, 2005]. It is shown that a powerful matching stage can successfully discard irrelevant information and better performance is gained with increased number of features [Tuytelaars and Schmid, 2007]. However, training and clustering are the most time consuming stages of state-of-the-art recognition frameworks. Clustering of a global dictionary takes several days for current benchmark image databases, becoming infeasible for online databases resulting in several billion features [Schindler et al., 2007]. Therefore, the goal is a feature selection within an earlier stage of this scheme. This thesis aims to develop and evaluate robust features to enable for a reduction of the numbers of features.

Krystian Mikolajzcyk gave his perspective on the state-of-the-art of local features in the opening talk of the feature evaluation benchmark at the CVPR 2009 [Mikolajczyk et al., 2009]. He is of the opinion that the most important future challenges for local features are the following and should be regarded for future research:

- **Memory requirements:** He suggests that scientists should focus on papers dealing with the "how do I reduce memory needs" problem, as this is very important for large scale problems.

- **Speed:** For many applications this is overrated, as most detector are not the bottleneck of a system. Typically, they can be parallelized easily. Therefore, improvements in the detection phase should be well motivated.

- **Coverage:** Classification tasks are in need of a reasonable spatial distribution of features. Unfortunately there is no metric for this yet.
- **Complementarity:** Regarding features used in an application, there is no objective quality measure. Even an "underperforming" feature in terms of repeatability is valuable, if complementary to other more robust features.
- **Geometric precision:** There is no sufficient evaluation for the actual precision of local features.

Following this outlook, this thesis addresses all of the challenges formulated: A reduction of the number of features helps to reduce the memory consumption. As correctly stated in the talk, the detection of features is typically not a bottleneck of a system. The time consuming steps are the subsequent operations. The most powerful way to reduce their calculation time is to select the best features and thereby reduce the work load beforehand. For the coverage measurement, this work proposes a simple coverage measurement for image features estimating the region covered by features in the image. Regarding video features, it proposes a robustness measure based on the relative coverage of the video.

The following Section 1.1 outlines the thesis' main objectives and research done in the field of image features whereas Section 1.2 focuses on the work done in the field of video features. Section 1.3 gives the scientific contributions presented in the thesis, where Section 1.3.1 summarizes the main points of improvements, and Section 1.3.2 describes the contributions in detail per field and publication. During the work on the thesis, several applications and research projects have been carried out. They are outlined in Section 1.4, where Section 1.4.1 focuses on the applications done with images and Section 1.4.2 on the video applications. Section 1.5 gives the structure of the thesis.

1.1 Image Features

Relevant to the color based interest point detection by [Stöttinger et al., 2007b] is the research of [van de Weijer and Gevers, 2005; van de Weijer et al., 2006]. They did preliminary work on incorporating color distinctiveness into the design of interest point detectors. In their work, color derivatives, which are used in both the detection of the interest points and the determination of the information content of the points, form the basis of a color saliency boosting function. Furthermore, the histograms of image color derivatives show distinctive statistical properties which are exploited in the color saliency boosting function. Therefore, in this thesis, computational methods are proposed to compute interest points, designed to allow a reduction in the number of salient points while maintaining state-of-the-art performance in image retrieval and object recognition applications. It achieves the ability to choose the most discriminative points in an image through including color information in the interest point determination process. To this end, a framework is presented for using color information to extract interest points and select a scale associated with each interest point. The aim is to select points based on expressive and concise properties of color information and distributions on a local scope.

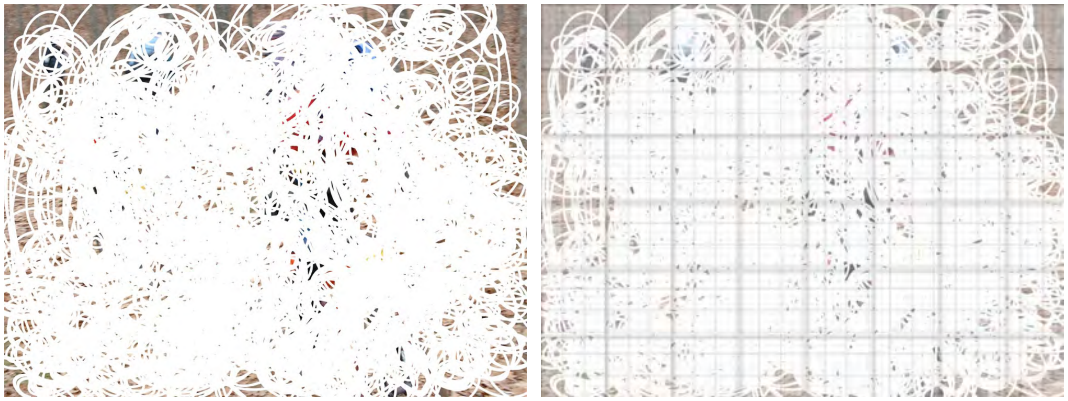
Therefore, the focus is on color models that have useful perceptual, salient and invariant properties to achieve a reduction in the number of interest points before the description phase. A reduced number of local features will yield a reduction of the computation of image descriptors. A method is proposed of selecting a scale associated to interest points, while maintaining the properties of the color space used, and to steer the characteristic scale by the saliency of the surrounding structure. Opposed to other color interest points used so far [Montesinos et al., 1998; Gouet and Boujemaa, 2001; Rugna and Konik, 2002; Faille, 2005; Gabriel et al., 2005; Abdel-Hakim and Farag, 2006; Unnikrishnan and Hebert, 2006; Forssén, 2007], the goal is to enhance an adapted multi-dimensional color Harris in conjunction with an independent scale selection maintaining the main properties of the chosen color space [Stöttinger et al., 2007a].

The sparse color interest point detector gives a more stable number of features and thus a better comparable image representation in computer vision applications. It is shown that by taking the saliency of color information and color invariances into account, improved retrieval of color images is obtained, thereby being more stable to lighting and shadowing effects than using illumination correlated color information. The gain in stability and distinction of these features is used for achieving a more sparse representation for object retrieval and categorization [Stöttinger et al., 2009b]. Moreover, object categorization using the well known PASCAL VOC dataset shows that the use of significantly fewer color salient points gives comparable performance to the best performing system in the 2006 challenge. An international categorization benchmark [Mikolajczyk et al., 2009] at the CVPR'09 evaluating 33 different features from the University of Amsterdam, TU Vienna, CMP Prague, ETH Zurich, EPFL Lausanne, University of Surrey, Stanford University, and the Harvard Medical School, showed that this sparse representation is equally representative to the best performing approach: While a dense representation gives insignificantly better results in the majority of the classes, the color representation outperforms the best performing approach in several classes with a fraction of the data used. The detailed evaluation is given in Section 4.4.

An example image is given in Fig. 1.2 where a natural image and the feature detection approaches from the experiments in this thesis is shown. White circles indicate the location and the scale of the detected features. The Harris Laplacian in Fig. 1.2(b) is chosen as the state-of-the-art baseline for comparison with [Zhang et al., 2007]. Fig. 1.2(e) and (f) show the selected color points using two perceptual approaches: Light invariant points give features based on the HSI color space, color boosted points do a statistical analysis of the entropy of a specific color before the detection. The approaches are given in more detail in Section 3.2.1.

[Donner et al., 2007] take the minima of the Gradient Vector Flow (GVF) [Xu and Prince, 1998] with one manually adjusted set of parameters to locate centers of local symmetry at a certain scale. This work extends their approach and proposes a GVF based scale space pyramid followed by a scale decision criterion to provide an approach for general purpose interest point detection. This multi-scale orientation invariant interest point detector has the aim of providing stable and densely distributed locations. Due to the iterative gradient smoothing during the computation of the GVF, it takes more surrounding image information into account than other detectors. Its stability against noise, blur, JPEG artifacts, rotation and illumination change makes it a promising approach for recognition tasks. For example, low quality images and videos in

(a) original (b) Harris Laplacian



(c) Harris Hessian (d) Harris Hessian + dense sampling

(e) light invariant points (f) color boosted points

Figure 1.2: VOC 2007 image number 337 as an example of a natural image. White circles indicate the location and the scale of extracted features. Only the visual data within white circles is used for succeeding operations in object categorization. (b) indicates the state-of-the-art [Zhang et al., 2007] whereas (c) shows a very dense extension [Mikolajczyk et al., 2006] and a combination with dense sampling (d) [van de Sande et al., 2009]. (e) and (f) visualize the proposed color points (see Section 3.2.1).

on-line applications and used in mobile computing suffer from such effects. Medical imaging also often deals with low contrast images. Therefore, this detection approach is promising for many applications in image matching in the field of mobile computing, medical imaging and dense sampling of noisy visual data [Stöttinger et al., 2008]. The approach is given in detail in Section 3.1.3, its evaluation is given in Section 4.2.

1.2 Video Features

Video understanding gains great attention in current computer vision research. With growing on-line data sources of videos, large private digital video archives and the need for storage and retrieval of surveillance videos, automated video understanding becomes necessary.

The most promising approaches for spatio-temporal features are extensions from successful features for images. There are spatio-temporal corners [Laptev and Lindeberg, 2003a], periodic spatio-temporal features [Dollár et al., 2005], volumetric features [Ke and Kanade, 2005] and spatio-temporal regions of high entropy [Oikonomopoulos et al., 2006]. Following [Mikolajczyk and Schmid, 2004], a stable representation which is invariant to lighting conditions, view point, quality of encoding, resolution and frames per second is desired. Recent work [Wang et al., 2009] evaluates spatio-temporal features on their matching performance on different datasets. They state that in the literature many experiments are not comparable as they differ in their experimental settings and classification techniques. However, until now there is no principled evaluation of the robustness of spatio-temporal features available: evaluation is done by measuring the overall performance of the application itself [Wang et al., 2009]. An evaluation of a complex framework only by its final performance does not give full insight into the performance of the chosen features. Subsequent operations (clustering, classification) are arbitrarily chosen and use empirically found parameters. Moreover, experiments in the literature are carried out with different classification algorithms tainting the experimental insights.

In this work FeEval[1] [Stöttinger et al., 2010c], a dataset for the evaluation of such features, is presented. For the first time, this dataset allows for a systematic measurement of the stability and the invariance of local features in videos. FeEval consists of 30 original videos from a great variety of different sources, including HDTV shows, 1080p HD movies and surveillance cameras. The videos are iteratively varied by increasing blur, noise, increasing or decreasing light, median filter, compression quality, scale and rotation leading to a total of 1710 video clips. Homography matrices are provided for geometric transformations. The surveillance videos are taken from 4 different angles in a calibrated environment. Similar to prior work on 2D images, this leads to a repeatability and matching measurement in videos for spatio-temporal features estimating the overlap of features under increasing changes in the data.

A way to evaluate the quality of these features independently of the framework or application is given. Every transformation on the videos denotes one *challenge* and is well defined. For the geometric cases all homography matrices are known. The change of noise, light, compression or frames per second are applied reproducibly according to the parameters given. The dataset consists of 30 original videos. Per video, 8 transformations are applied in 7 increasing steps, leading to a total of 1710 videos.

[1] `http://www.feeval.org`

Therefore, a new way for the evaluation of video retrieval approaches is proposed: The evaluation of detectors and descriptors is divided into two independent tasks. For detection, a repeatability measurement in 3D similar to [Willems et al., 2008] is proposed. For the descriptions a pipeline to identify the robustness of local spatio-temporal descriptions in a principled way is proposed. The performance of these two tasks are measured by their robustness under alterations of the visual input data. The original videos are used as ground-truth while it is observed to what extent the features change under the challenges.

Detectors are evaluated by treating every single detection as a 3D blob in a volume. Per challenge, the detections of the altered videos are projected back to the original ones. They are matched geometrically to observe their behavior in the challenge. Applying this on 30 classes leading to a total of 1710 videos, we receive a credible comprehension of the robustness in video classification per challenge. This allows to justify the choice of a certain feature based on the properties of the input video. It is shown that the robustness to noise, resolution and compression artifacts is highly varying for different features. It is shown that certain features remain stable, even when the video is so much altered that it is visually not appealing any more. Moreover, features are differently robust to preceding noise reduction and contrast. This is an important fact for preprocessing of video material.

Descriptors are evaluated in the same principled way. Robustness is denoted by the retrieval accuracy in the context of the challenges with varying state-of-the art detectors and descriptors.

What do we gain from this comparative evaluation? First, the causes for the different performances of state-of-the-art features on different data-sets are revealed based on the isolated alterations of the video data. Secondly, the choice of features for a new application can be derived by examining the respective video data and selecting those properties that are most important for the task at hand. From previous evaluations, the best performing features are known for certain data-sets. These results do not necessarily apply for new data. Moreover, the evaluation leads to more very practical insights for these tasks, e.g.: Can one reduce the numbers of frames per second for the retrieval in order to save time, space and memory? How lossy can one encode the videos until the retrieval application fails? Does it make sense to remove the noise or enhance the contrast beforehand? Is the description able to take care of different resolutions of the videos? Is rotation of videos challenging? This thesis gives detailed answers to these questions per detector and per descriptor and helps to optimize the many trade-offs.

1.3 Contributions

In the following, the scientific contributions of the thesis are presented. Section 1.3.1 gives an overview of the research topics and their relation to the various projects and applications given in Section 1.4. Section 1.3.2 describes the contribution of the main publications in more detail.

1.3.1 Summary of the Contributions

In this thesis computational methods are proposed to compute salient (interest) points in color images independent of the color space. Using color spaces that have useful perceptual, invariant and saliency properties, the goal of the interest point selection process is to reduce the

number of interest points while maintaining state-of-the-art performance in image retrieval and object recognition applications. An approach to use the GVF for general interest point detection, referred to as the GVFpoints, is proposed in Section 3.1.3. The approach introduces a complementarity way of finding well distributed dense features for image matching. The features are located *between* structure, finding very different features to corner and blob detectors.

A stable color scale selection method for the color Harris is proposed. This allows for the necessary discriminative points to be located and allows a reduction in the number of salient points resulting in an invariant (repeatability) and discriminative (distinctiveness) image description. Experimental results on large image datasets show that color based Harris energy gives a more reliable decision criterion for reducing features than the luminance based counterpart does. Further, the proposed color-based method obtains state-of-the-art performance, with the number of salient points reduced by half which is justified by a recent international benchmark. This reduction of the number of points allows subsequent operations, such as feature extraction and clustering, to run more efficiently. Moreover, the method provides less ambiguous features, a more compact description of visual data, and therefore a faster classification of visual data.

Exploring new ways for feature extraction, the non-minima suppression of the GVF magnitude is examined. Based on the GVF's properties it provides the approximate centers of blob-like structures or homogeneous structures confined by gradients of similar magnitude. It results in a scale and orientation invariant interest point detector, which is highly stable against noise and blur. These interest points outperform the state-of-the-art detectors in various respects. It is shown that the approach gives a dense and repeatable distribution of locations that are robust against affine transformations while they outperform state-of-the-art techniques in robustness against lighting changes, noise, rotation and scale changes. Extensive evaluation is carried out using a successful framework [Mikolajczyk et al., 2005b] for interest point detector evaluation.

Until now, no principled evaluation of spatio-temporal video features similar to those for image features has been done. This thesis presents FeEval, a dataset for the evaluation of such features. For the first time, this dataset allows for a systematic measurement of the stability and the invariance of local features in videos. Similar to prior work on 2D images, this leads to a repeatability and matching measurement in videos for spatio-temporal features estimating the overlap of features under increasing changes in the data. For the evaluation of the detectors, their repeatability is measured on the challenges treating the videos as 3D volumes. To evaluate the robustness of spatio-temporal descriptors, the work proposes a principled classification pipeline where the increasingly altered videos build a set of queries. This allows for an in-depth analysis of local detectors and descriptors and their combinations. It is shown that the features have different properties and behave differently under varying transformations (challenges). This helps researchers to justify the choice of features for new applications and helps to optimize the choice of input video in terms of resolution, compression, frames per second or noise suppression. All the extracted features are accessible on-line for further independent evaluation and applications.

An overview of the papers and their field of research is given in Tbl. 1.1. A full list of publications is given on-line[2].

[2] `http://www.caa.tuwien.ac.at/cvl/people/julianstoettinger/publications.html`

Topic	Section	Publication	Application/Project
Image Features	Sec. 3.2.1	**Do Colour Interest Points Improve Image Retrieval?** Julian Stöttinger, Allan Hanbury, Nicu Sebe and Theo Gevers, ICIP 2007.	MUSCLE NoE, object recognition showcase,
		Overview of the ImageCLEF 2007 Object Retrieval Task. Thomas Deselaers, Allan Hanbury, Ville Vitaniemirás A. Benczúr, Mátyás Brendel, Bálint Daróczy, Hugo Jair Escalante Balderas, Theo Gevers, Carlos Arturo Hernández Gracidas, Steven C. H. Hoi, Jorma Laaksonen, Mingjing Li, Heidy Marisol Marín Castro, Hermann Ney, Xiaoguang Rui, Nicu Sebe, Julian Stöttinger and Lei Wu, CLEF 2007.	presented at IBC 2007, ImageCLEF 2007
		The MUSCLE Live Image Retrieval Evaluation Event Allan Hanbury, Branislav Micusik and Julian Stöttinger, Live IR Event in conjunction with CIVR 2007.	Live image retrieval evaluation challenge CIVR 2007
		Ordnung in die Bilderflut - Arbeitsweise von Content-Based Image Retrieval Systemen. Julian Stöttinger and Allan Hanbury, MP Berlin 2009.	
	Sec. 3.1.3	**Evaluation of Gradient Vector Flow for Interest Point Detection.** Julian Stöttinger, René Donner, Lech Szumilas and Allan Hanbury, ISVC 2008.	
	Sec. 1.4.1	**Translating Journalists' Requirements into Features for Image Search.** Julian Stöttinger, Jana Banova, Thomas Pönitz, Nicu Sebe and Allan Hanbury, VSMM 2009.	APA-eszeve demonstrator
		Understanding Affect in Images. Jana Machajdik, Allan Hanbury and Julian Stöttinger, ACM MM 2010 - Grand Challenge	ACM MM exhibit
	Sec 2.4	**Scale Invariant Dissociated Dipoles** Marco Vanossi and Julian Stöttinger, AAPR 2010.	
Image Matching	Sec. 4	**Lonely but Attractive: Sparse Color Salient Points for Object Retrieval and Categorization** Julian Stöttinger, Allan Hanbury, Theo Gevers and Nicu Sebe, CVPRW 2009	CVPR 2009 benchmark
	Sec. 1.4.1	**Efficient and Distinct Large Scale Bags of Words.** Thomas Pönitz, René Donner, Julian Stöttinger, Allan Hanbury, AAPR 2010.	Commercial software product
		Efficient and Robust Near-Duplicate Detection in Large and Growing Image Data-Sets Thomas Pönitz and Julian Stöttinger, ACM MM 2010 - industrial exhibit.	ACM MM exhibit
Video Data-set	Sec. 6.1.2.	**FeEval - A dataset for evaluation of spatio-temporal local features** Julian Stöttinger, Sebastian Zambanini, Rehanulla Khan and Allan Hanbury, ICPR 2010.	Public data-set, Cosamed, MuBisA
Video Features	Sec 6.2	**Behavior and properties of spatio-temporal local features under visual transformations** Julian Stöttinger, Bogdan Tudor Goras, Nicu Sebe and Allan Hanbury, ACM MM 2010.	
		Systematic Evaluation of Spatio-temporal Features on Comparative Video Challenges Julian Stöttinger, Bogdan Tudor Goras, Thomas Pönitz, Nicu Sebe, Allan Hanbury and Theo Gevers, ACCVW 2010	
Video Classification	Sec. 1.4.2	**Skin Paths for Contextual Flagging Adult Videos** Julian Stöttinger, Allan Hanbury, Christian Liensberger and Rehanulla Khan, ISVC 2009.	Video-tag ZIT project

Table 1.1: Overview of the contributions and applications done in the course of this thesis.

1.3.2 Contributions in Detail

In the following, the contributions are described in more detail according to their publication at international conferences and chronologically ordered and grouped by topic.

Image Features and Matching

Setting up a collaboration with the Intelligent Sensory Information Systems group of the University of Amsterdam, the idea of scale invariant color interest points is presented orally at the ICIP 2007:

Do Colour Interest Points Improve Image Retrieval? [Stöttinger et al., 2007b] Julian Stöttinger, Allan Hanbury, Nicu Sebe and Theo Gevers, Proceedings of the 14th IEEE International Conference on Image Processing (ICIP 2007), San Antonio , Texas, September 16-19, 2007, pp. 169-172. [Stöttinger et al., 2007a]

The approach of **scale– and color-invariant interest points** is presented to an international audience for the first time.

- The paper shows that the color tensor can shift the interest points to more stable and distinct locations than luminance based methods.
- The color scale selection leads to a better stability under geometric transformations of objects.
- Using correlated color, boosted color or color invariant information, the method gains performance over luminance based methods.
- In retrieval scenarios, the approach shows to be more distinct and stable, which leads to a higher and more precise retrieval rate than reference implementations.

Looking out for new ways to extract stable features from images, collaboration with the Computational Image Analysis and Radiology group of the Vienna Medical University and the Automation and Control Institute of the Vienna University of Technology has been set up. Using the insights of [Donner et al., 2007] to localize centers of symmetrical structure of known sizes for sparse appearance models in medical imaging, it was shown in [Szumilas et al., 2007] that state-of-the-art object recognition can take advantage of symmetrical configuration of features. As a result of this collaboration, the paper has been presented orally at the ISVC 2008.

Evaluation of Gradient Vector Flow for Interest Point Detection Julian Stöttinger, René Donner, Lech Szumilas, Allan Hanbury, Proceedings of the 4th International Symposium on Visual Computing (ISVC), Las Vegas, Nevada, USA, December 1-3, 2008, pp. 338-348. [Stöttinger et al., 2008]

This paper shows that using the GVF for feature localization of scale– and rotational invariant interest points, the features remain more stable than state-of-the-art detectors.

- The first "general purpose" interest point detector based on the GVF, **GVFpoints**, is presented.
- The features give a rich and well-distributed representation for diverse and natural images.
- For the majority of the well known repeatability challenges, interest points based on GVF provide more stable locations than the well known and broadly used corner or blob detectors.
- The paper shows that the GVFpoints are practically invariant against linear and arbitrary lighting changes, rotation, noise, low contrast or heavy compression.

Following the promising results of the color invariant interest points from [Stöttinger et al., 2007a], the next research interest was found in the saliency of these features. Recent work shows – and the community is of the general opinion – that more interest points improve the classification and recognition performance of state-of-the-art systems. This is computationally costly and leaves a vast amount of data for the subsequent tasks. Therefore the next step was to propose computational methods for a reduction of the number of interest points, thus allowing for a more efficient classification. The work has been presented orally at the CVPR workshop on *Feature Detectors and Descriptors: The State of the Art and Beyond*:

Lonely but Attractive: Sparse Color Salient Points for Object Retrieval and Categorization Julian Stöttinger, Allan Hanbury, Theo Gevers and Nicu Sebe, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on Feature Detectors and Descriptors: The State Of The Art and Beyond, Miami, Florida, USA, June 20, 2009, pp. 1-8. [Stöttinger et al., 2009b]

Scale invariance is not only a very desirable property for local features, it is mandatory for image matching: As objects appear at many possible scales in natural images, an efficient and discriminative representation has to exhibit some sort of scaling or self-similarity. This hypothesis is intensified as there is strong evidence that the primate's early visual processing uses information in a scale invariant manner [Ruderman and Bialek, 1994]. Therefore, the proposed method includes the following fundamental contributions to image matching:

- Using color interest points allows applications in computer vision to take full advantage of colorful input data. Until now, stable corner detectors are either luminance only or are not scale invariant. It is shown that this limitation is overcome gaining efficiency and distinction in the subsequent operations.
- Incorporation of perceptual color spaces in local scale invariant features. The advantages of these color spaces are directly passed on to the representation of the features. Therefore, instability of luminance based local features due to changing shadowing, reflections, lighting effects and color temperature are implicitly addressed. Invariance to lighting

changes or the incorporation of a visual saliency function is achieved with one simple color transformation and is passed directly to state-of-the-art retrieval and categorization frameworks.

- Selection of discriminant features is typically done in the matching stage when the classification system builds its model. It is shown that with the proposed method, it is possible to perform this choice in the first step of the typical image matching pipeline making all subsequent operations faster. Moreover, this crucial step is conducted independently for every feature and image (e.g. without knowing the global feature space) and based on the local visual input only (e.g. no spatial inter-relation, ground truth or occurrence frequency of features is used).

- Runtime of every single step of image matching applications decreases with a more sparse representation of local features. Off-line procedures like building of a global dictionary are practically infeasible when the number of features in the training data goes into billions [Schindler et al., 2007]. The runtime of on-line procedures like quantization of features also depends on the number of features. There is a great interest in handling large datasets and making specific approaches simpler and more efficient. With the proposed method, the amount of data to be processed is reduced significantly by half for standard approaches.

- Higher dimensional data can be processed. The proposed representation of multi-channel information is not limited to any single color space.

First, the paper investigates different corner detection approaches and evaluates their repeatability under varying circumstances defined by Mikolajczyk [Mikolajczyk and Schmid, 2004], including different lighting changes, zoom and rotation, viewpoint, blur and jpeg compression. It is shown that the use of color invariants increases the stability and distinction of interest points in natural scenes under varying transformations. The sparse color interest point detector gives a stable number of features and thus a better comparable image representation in computer vision applications. By taking the saliency of color information and color invariance into account it is shown that improved retrieval of color images, thereby being more stable to lighting and shadowing effects than illumination based approaches, are obtained. Object categorization evaluation is performed using the well known PASCAL VOC dataset. It shows that the use of significantly fewer color salient points gives comparable performance to the best performing system in the 2006 challenge [Zhang et al., 2007]. The gain in stability and distinction of these features is used for achieving a sparse representation for object retrieval and categorization [Stöttinger et al., 2009b]. Using this representation the state-of-the-art in object categorization performance is maintained while dealing with half of the features [Mikolajczyk et al., 2009]. Reducing the number of interest points reduces the run-time of a recognition application at least linearly.

Video Features and Matching

In the following the contributions in the field of spatio-temporal features are given. The thesis proposes a new way to evaluate the quality of video features in an independent way from the

framework or application.

FeEval - A dataset for evaluation of spatio-temporal local features Julian Stöttinger, Sebastian Zambanini, Rehanulla Khan, Allan Hanbury, Proceedings of the 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, August 23-26, 2010, pp. 499-503 [Stöttinger et al., 2010c]

The paper provides the first database to evaluate extracted features for their stability and invariance in a spatio-temporal context, called *FeEval*. Every transformation denotes one *challenge* and is well defined. For the geometric cases all homography matrices are known. The change of noise, light, compression or frames per second are applied reproducibly according to the parameters given. The dataset consists of 30 original videos. Per video, 8 transformations are applied in 7 increasing steps, leading to a total of 1710 videos. A file server and a webpage was set up to provide the data-set to the community[3].

Unlike previous evaluations which concentrated on videos of basic human actions, the research aims to evaluate features under predefined transformations of arbitrary videos. This allows to justify the choice of a certain feature based on the properties of the input video. On a research visit at the University of Trento, a collaboration with the Department of Information Engineering and Computer Science and the University of Iasi, Faculty of Electronics, Telecommuncations and Information Technology has been set up leading to the following publication:

Behavior and properties of spatio-temporal local features under visual transformations Julian Stöttinger, Bogdan Tudor Goras, Nicu Sebe, Allan Hanbury, Proceedings of ACM Multimedia (MM), Firenze, Italy, October, 25-29, p. 1–4 [Stöttinger et al., 2010b]

The paper shows that the robustness to noise, resolution and compression artifacts is highly variable for different features.

- Certain features remain stable, even when the video is so much altered that it is no longer visually appealing. Moreover, features are differently robust to preceding noise reduction and contrast. This is an important fact for preprocessing of video material.
- For the reduction of frames, it is shown that the video can be equally represented by just using 12% of the original number of frames. This reduces the runtime and the memory requirements of such applications significantly.
- Spatio-temporal corner detectors are more affected by change of contrast and lighting than other detectors.
- Comparably with 2D features, Harris3D is more robust to scale changes than Hessian3D, although Hessian3D uses more scales in the evaluated approach.

Further experiments have been performed and the University of Amsterdam joined the collaboration. For the conclusive evaluation of video features, large scale experiments for video

[3] `http://www.feeval.org`

matching are performed.

Systematic Evaluation of Spatio-temporal Features on Comparative Video Challenges Julian Stöttinger, Bogdan Tudor Goras, Thomas Pönitz, Nicu Sebe, Allan Hanbury and Theo Gevers, Proceedings of the Tenth Asian Conference of Computer Vision (ACCV), Workshop on Video Event Categorization, Tagging and Retrieval, 2010, pp. 1-8. [Stöttinger et al., 2010b]

It proposes a new way for the evaluation of video retrieval approaches: The evaluation of detectors and descriptors is divided into two independent tasks.

- An efficient and straightforward repeatability measurement for spatio-temporal features which is able to evaluate hundreds of thousand of features is developed.
- It proposes a pipeline for the evaluation of spatio-temporal descriptions. By a reproducible alteration of the query sets, researchers are able to gain in-depth analysis of the robustness of state-of-the-art spatio-temporal features.
- The evaluation of detectors and descriptors is done independently allowing to choose the best features and their combinations for different data-sets.
- It showed to be worse to reduce noise in input data than to let the features take care of it on their own.
- Descriptions are more stable to changes of lighting and contrast than detectors are.
- The HOG3D descriptor is the best performing spatio-temporal descriptor today. It is only outperformed by the SURF3D descriptor in the challenges of compression, noise and median filtering. The high dimensionality of the HOG3D descriptor of 960 compared to 288 of the SURF3D descriptor is a drawback in terms of the complexity of all succeeding operations and should be considered when choosing the most appropriate descriptor.

1.4 Applications

During the three years of work for this thesis, several applications and projects have been developed. In this section an overview of these results is given. Details of the projects and the research done in the course of the projects are given in more detail in the references.

1.4.1 Image Features and Matching

Starting to work in the MUSCLE project[4], an EC-sponsored Network of Excellence, an online demonstrator was developed under the lead of Prof. Nicu Sebe in collaboration with the INRIA-IMEDIA group, Paris-Rocquencourt (Jaume Amores, Nozha Boujemaa). The proposed color interest points are estimated in less than 3 seconds and visualized on the fly. The system interacts

[4] `http://muscle.ercim.eu/`

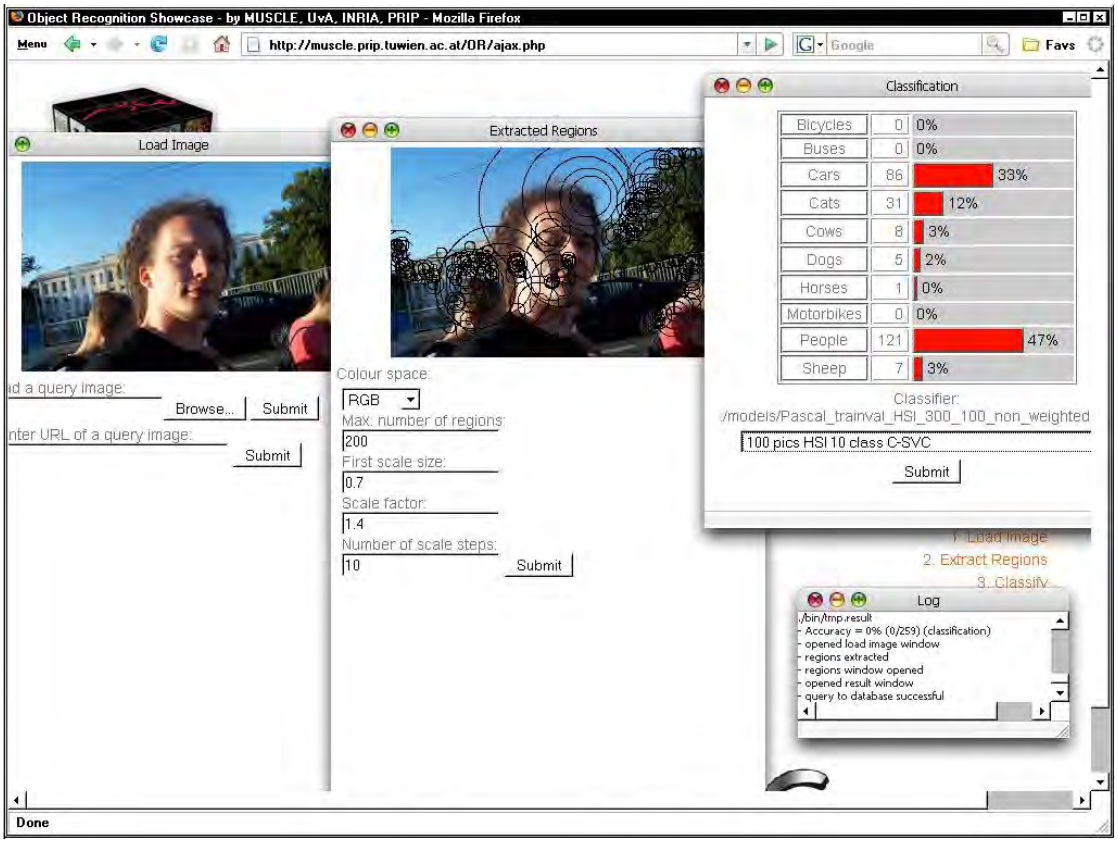


Figure 1.3: Screenshot of the MUSCLE object recognition showcase. Detection parameters and SVM models are adjustable by the user. The query image shows the some pedestrians, cars and the author in an urban environment. Classification correctly gives highest confidence to persons (0.47) and and cars (0.33). Follow-up classification confidences are cats (0.12), cows and sheep (both 0.03).

with a powerful server which makes it possible to classify uploaded images in less than 10 seconds to previous learned models of the VOC Pascal 2007 data-set.

The showcase has been presented successfully at the final meeting of the MUSCLE project in Paris, the *Medienproduktion Berlin 2007* [Stöttinger and Hanbury, 2009], the *International Broadcasting Conference (IBC) 2007* in Amsterdam and the *15th Summer School on Image Processing* in Szeged, Hungary. To the best of our knowledge, it is still the only on-line application which visualizes local detectors on the fly letting the user experiment with different parameters. With the extracted data, large SVM models are processed by a powerful server in the background making it possible to experience the changes in final classification performance when changing the detections of the query image; one of the main interests of this thesis. The showcase is

available online[5] and a screenshot is given in Fig. 1.3.

The media asset industry's challenge is to store, manage and make large amount of data accessible and retrievable. The *Austrian Press Agency* (APA) is the biggest media provider in Austria[6]. The approach has been used in a collaboration project with the APA-IT[7] subcompany of the Austrian Press Agency. The aim was to find the best suited approaches of the state-of-the-art of image retrieval to satisfy the needs of journalists searching for press images. Journalists have access to huge databases of photographs for illustrating their articles. They would like on-line searches to give fast access to the most suitable photos satisfying a certain requirement. At present, the majority of search interfaces to these databases rely on pure text search, making use of meta-data entered by photographers when they upload their photos, or by archivists. The APA-IT was interested in adding the possibility of doing visual search to their current text retrieval system, and wished to find out which visual features would be the most suitable. One source for the set of image features is the work of [Machajdik and Hanbury, 2010; Machajdik et al., 2010]. An important prerequisite was that the end users should be able to tune the image similarity criteria based on his image needs. This implies that the features to be used for a specific search should be selectable, but also that the features should be intuitively understandable for the end users. The features should also enable the users to improve the returned results.

The main contribution of this project was to illustrate how taking advantage of user studies highlighting the user requirements can lead to the selection of suitable features in image search systems. The team approached this task by first carrying out an analysis of journalists' photo searching requirements by further analyzing the results of a published user study [Markkula and Sormunen, 1998]. These requirements were then mapped to suitable visual features. The emphasis was on identifying suitable and intuitive low level features, as these can be rapidly implemented in the existing text-based image search system. In contrast to selecting a set of generic image features, such an approach to feature selection based on user requirement analysis should lead to better acceptance and use of these features by the end users. The resulting application is shown in Fig. 1.4.

The results were presented orally at the VSMM conference in Vienna, Austria [Stöttinger et al., 2009a]. For local color features, nearest neighbor search shows promising results in searching for similar letters, texts or rigid objects. This can be used for example for searching for images with distinct and predefined logos. The project was regarded as very successful by the APA-IT which started to implement the best suited features into their software right away.

The project revealed another fact in this application: Probably the biggest challenge is the vast amount of press images. Additionally, the number of images in the data-set is growing every day. The main income of the APA comes from the copyright fees of sold press images. The journalists gain access to all the images, choose their favorite ones and use them in newspapers and other publications. Whenever an image appears in such a media, expenses are incurred. The newspapers are obligated to report every usage of an image. Nevertheless, the APA has to observe all the media in Austria manually to charge for the images. A great advantage would be

[5] `http://muscle.prip.tuwien.ac.at/OR`

[6] `http://www.apa.at`

[7] `http://www.apa-it.at`

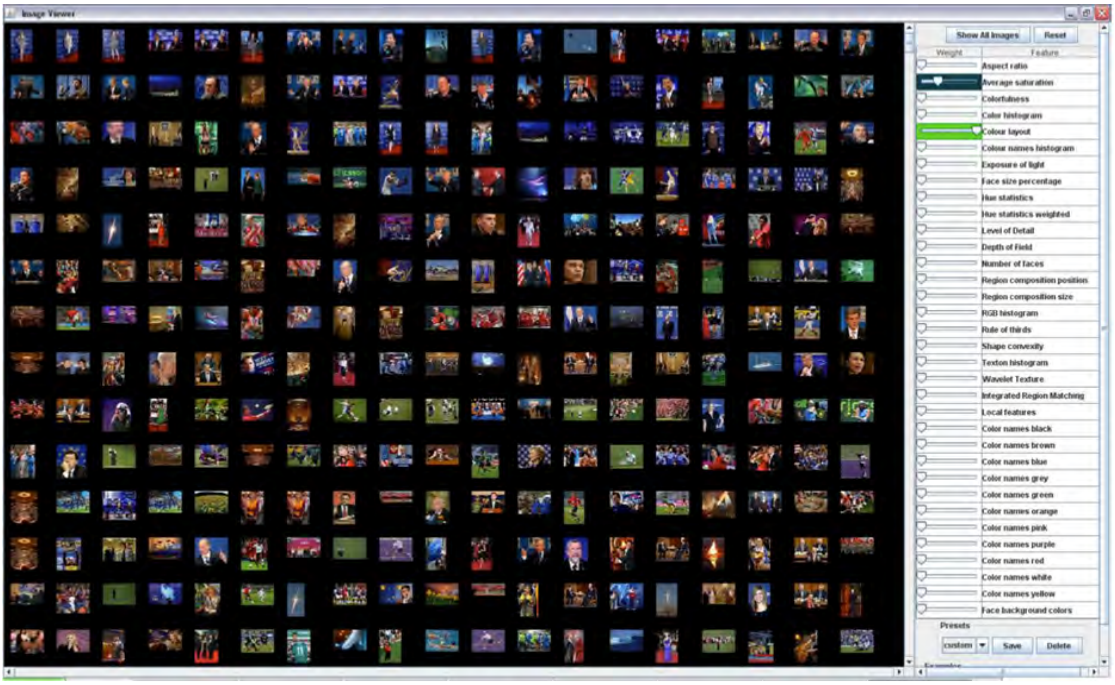


Figure 1.4: Screenshot of the APA demonstrator. Example journalists' search for images with red carpet on the bottom and blue background: The query image is a reasonably satisfying image (upper left). The journalist chooses to retrieve images with close color layout and similar average saturation.

gained from an application that is able to browse through all the published images of a day and compare them to all the images belonging to the APA. Because of the success of the prior project, a project between the TU Vienna, the CogVis Ltd.[8] and the APA-IT was set up to develop such an application. In the course of the development, knowledge and source code has been added from the CIR lab[9]. The CIR lab (Computational Image Analysis and Radiology) is an interdisciplinary research group of people from medicine, computer science and mathematics from different faculties, located at the Department of Radiology at the Medical University of Vienna. The results have been presented orally at the ÖAGM workshop in Zwettl, Austria [Pönitz et al., 2010] and is part of the industrial exhibit at the ACM MM conference in Florence, Italy [Pönitz and Stöttinger, 2010]. The application has been sold to a German media observer company which follows occurrences of advertisements in a large number of newspapers. Staff browse the newspapers and scan advertisements manually. The application counts similar advertisements from their steadily growing data-set. Fig. 1.5 gives an example use case. The company wants to track the advertisement campaign and therefore trains at least one representative advertisement. The application is able to *track* (find) replicas of the advertisement.

[8] `http://www.cogvis.at`
[9] `http://www.cir.meduniwien.at`

(a) query image (b) (c) (d) (e)

Figure 1.5: Media observations by visually tracking advertisements in newspapers. (a) shows the advertisement in the trained data-set. (b) - (e) give the successfully *tracked* images in decreasing similarity.

The main scientific challenge was the detection of image replicas in large scale image databases. In this context, a replica is denoted as a copy or reproduction of a work of art, especially one made by the original artist, or a copy or reproduction, especially one on a scale smaller than the original. In terms of computer vision the meaning of image replica is slightly different. Following [Maret et al., 2006] it refers not only to an exact copy of a given original image as replica, but also to modified versions of the image after certain manipulations, as long as these manipulations do not change the perceptual meaning of the image content. In particular, replicas include all variants of the original. These include images obtained after common image processing manipulations such as compression, filtering, adjustments of contrast, or geometric manipulations.

The problem of near duplicate detection is solved by local features and the bags-of-words approach already. Unfortunately, this does not hold for very large data-sets, where images become more and more similar to each other. Moreover, certain images tend to be similar to all the others as they contain almost every feature. This problem is defined as the *Kirschbaum* problem where images of small, non-repetitive texture (e.g. a close up image of a cherry tree in full bloom, but also water, grass or sand. An example is given in Fig. 1.6) tend to show up in every image query when using bags-of-words on large data-sets. We solve this problem by developing a dedicated classification technique outperforming standard approaches: To avoid the occurrence of ambiguous features in large image data-sets, the project aims to improve the approach of bags of visual features simply by the improving the distance measure between image signatures. It is done by developing a more specific decision criterion.

One major challenge using bags of visual features is the generation of a global codebook on large data-sets (e.g. [Jurie and Triggs, 2005], [Moosmann et al., 2006]). In the training phase, it is necessary to cluster all or as many features as possible of the given data-set. This means basically that every feature (up to 5000 per image in our case) has to be set in relation

Figure 1.6: The Kirschbaum problem: High frequency textures are problematic in conjunction with scale invariant local features. They tend to lose their distinction to other images in large databases.

to each other. The features in the complete training set have to be sampled in a representative way to reduce the amount of data for classification. The approach solves this task by iteratively approximating the desired result in linear time. The approach is called *kshifts* and is available online[10].

The developed application shows that it is possible to track a single image in large scale data sets. It is possible to distort and transform visual information in the form of cropping, blurring, and scaling just to mention a few. It is still possible to find the right images or very similar images in a reliable way. The runtime of the application meets the requirements for an industrial use as it is possible to track 1000 images per hour. One of the future improvements of the application could be the introduction of color description to the feature space – this would lead to more discrimination power of the actual image description, but will lead to drawbacks when aiming for the tracking of gray-scale images.

1.4.2 Video Features and Matching

In the course of the research, the step from image processing to video processing was simplified by getting assigned to a short-term project for color-based classification of videos by skin segmentation for an Austrian on-line portal provider. User generated content has become very popular in the last decade and has significantly changed the way we consume media [Cha et al., 2007]. With the international success of several *Web 2.0* websites (platforms that concentrate on the interaction aspect of the internet) the amount of publicly available content from private sources is vast and still growing rapidly [Cha et al., 2007].

The amount of video material being uploaded every day is too large to allow the operating companies to manually classify the content of every submitted video as appropriate or objection-

[10] `http://www.cogvis.at`

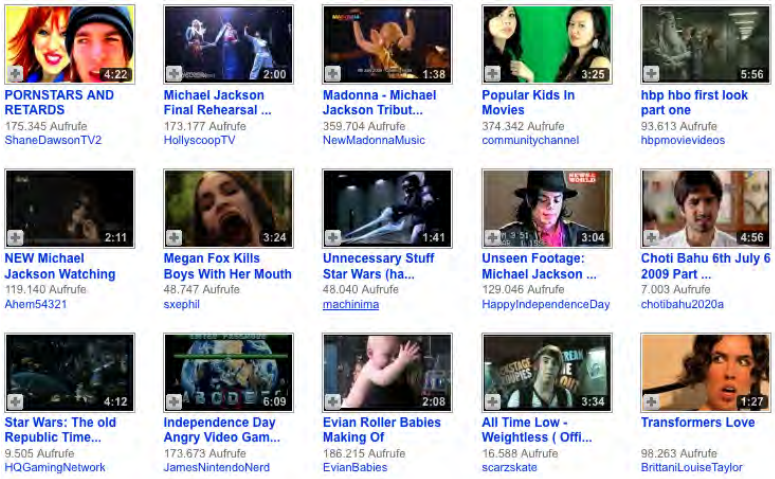


Figure 1.7: Most popular videos from youtube.com on July 4th, 2009. The most popular video on the top left is actually a teenage comedy with probably objectionable title and should thus not be rejected by an automated system.

able before publishing [BBC News, 2006]. The predominant methods to overcome this problem are to block contents based on keyword matching that categorizes user generated tags or comments. Additionally, connected URLs can be used to check the context of origin to trap these websites [Lee et al., 2007]. This does not hold true for websites like YouTube that allow uploading of videos. The uploaded videos are not always labeled by (valid) keywords for the content they contain (compare Fig. 1.7). As no reliable automated process exists, the platforms rely on their user community: Users flag videos and depending on this, the administrators may remove the videos flagged as objectionable. This method is rather slow and does not guarantee that inappropriate videos are immediately withdrawn from circulation. A possible solution for rapid detection of objectionable content is a system that detects such content as soon as it is uploaded. As a completely automated system is not feasible at present, a system that flags potentially objectionable content for subsequent judgement by a human is a good compromise. Such a system has two important parameters: the number of harmless videos flagged as potentially objectionable (false positive rate), and the number of objectionable videos not flagged (false negative rate). In the context of precision and recall of a classification application, these two parameters present a trade-off. For a very low false negative rate, a larger amount of human effort will be needed to examine the larger number of false positives. These parameters should be adjustable by the end-users depending on the local laws (some regions have stricter restrictions on objectionable content) and the amount of human effort available. A further enhancement to reduce the amount of time required by the human judges is to flag only the segments of videos containing the potentially objectionable material, removing the need to watch the whole video, or search the video manually. One reason why videos may be considered objectionable is due to explicit sexual content. Such videos are often characterized by a large amount of skin being visible in the frame, so a commonly used component for their detection is a skin detector [Lee et al., 2007;

Zheng et al., 2004]. However, this characteristic is also satisfied by frames not considered as objectionable, most importantly close-ups of faces.

Therefore the research considers the flagging of user-uploaded videos as potentially objectionable. The main contribution of this work is to introduce two uses of contextual information in the form of detected faces. The first is to use tracked faces to adjust the parameters of the skin detection model. As is shown in Fig. 1.7, user generated content contains many faces. Classification rules are developed based upon a prior face detection using the well known approach from Viola et al. [Viola and Jones, 2004]. This work builds on [Khan et al., 2008] where it is shown that more precise adaptive color models outperform more general static models especially for reducing the high number of false positive detections. In [Liensberger et al., 2009] it is shown that humans need contextual information to interpret skin color correctly. Their approach is extended by using a combination of face detectors: Frontal face detection and profile face detection is carried out in a combined tracking approach for more contextual information in the skin color representation.

The second use of face information is through the summarization of a video in the form of a path in a skin-face plot. This plot allows potentially objectionable segments of videos to be extracted, while ignoring segments containing close-ups of faces. It is shown that the properties of the skin paths give a reliable representation of the nature of videos. The proposed approach was kept algorithmically simple, and currently runs at over 30 frames per second. A high level of performance is required in such an application to cope with the large number of uploaded videos. The work has been presented orally at the ISVC 2009 in Las Vegas, Nevada [Stöttinger et al., 2009c] and is now supported by the city of Vienna to develop a market ready server - client application in the course of the *Zentrum für Innovation und Technology* (center for innvoation and technology, ZIT)[11] *Call Media Vienna* (CMV). The activities of ZIT encompass providing direct financial assistance to companies or making a technology-specific infrastructure available.

The Cosamed[12] project was granted as an initiation project for the follow up BENEFIT[13] project MuBisA[14]. The goal of the project is a reliable and automated computer vision system to enable an independent lifestyle for the elderly and disabled. In contrast to prior projects, the system relies solely on computer vision techniques. The main research interest was the evaluation of different state-of-the-art approaches to choose the most appropriate methods. The idea is that the daily lives of elderly people will not be affected and they can maintain their independent life longer. It aims for a robust fall detection of elderly people. In 2000, an estimated number of one million people were already supported and monitored by pendant or pull-cord alarms linked to a central control facility [Edwards, 2000]. This number of people in need is growing steadily. People suffering from Dementia, Parkinson's disease or Amyotrophia could take full advantage of an automated alarm system in their own homes to ensure an independent life for as long as it is possible. Since the detection set-up consists basically of Internet protocol cameras and a central storage and calculation server, the system is open and flexible towards other applications: Fire, smoke and water detection and assistance for medication are introduced

[11] `http://www.zit.co.at`
[12] `http://www.cogvis.at/cosamed`
[13] `http://www.ffg.at/benefit`
[14] `http://www.cogvis.at/mubisa/`

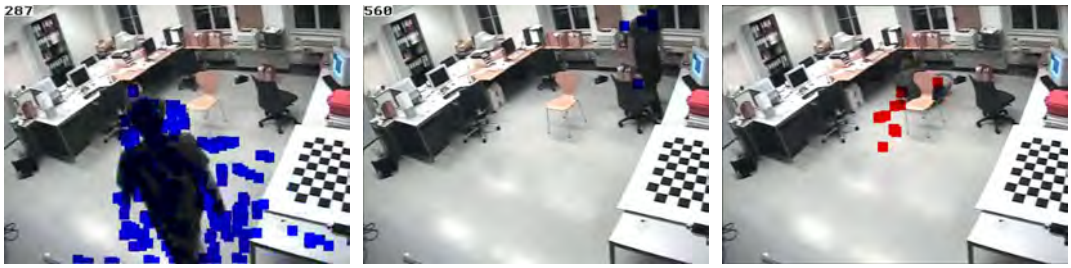


Figure 1.8: Results of the Cosamed project. Rectangles are placed on temporal centers of detected features. Blue are features of usual behavior, red gives emergency feedback classified by a Mahalanobis nearest neighbor classification of clustered local 3D jets.

in a second step.

As the most promising, but sophisticated and thus in terms of calculation time very expensive approach, the analysis of local video features has been chosen. The research investigates the use of spatio-temporal single scale interest points inspired by [Laptev and Lindeberg, 2003b]. Following their approaches, the extracted locations are described with local jets [Florack et al., 1996] extended to 3D in a straightforward way. These simple local descriptors are scale invariant, but change their information within rotation or perspective transformations. Additionally, the gradients are not fully illumination invariant. Nevertheless, it should give a valuable insight into the feasibility to use local features for such an application. Results showed that with a growing training data-set and sufficient processing power, such an approach is more flexible than approaches based solely on global movement. Example frames are given in Fig. 1.8. Blue rectangles denote local spatio-temporal features with their nearest neighbor in the *normal behavior* class, red ones with the highest similarity to previously trained *unusual behavior* class. Nevertheless, as in such an application calculation time is very important, for the subsequent project, global 2D features and calibrated 3D movement detection have been used successfully [Zambanini et al., 2010].

1.5 Structure of the Thesis

The thesis is organized as follows.

Chapter 2 gives the basic principles of the methods described in the thesis. Starting with the development of the color spaces used for feature extraction, the properties of the Gaussian kernel and the structure tensor are explained. A state-of-the-art of local description approaches for images and video is given. Finally, desired properties of local features are discussed.

Chapter 3 describes the most successful approaches for local feature detection in images. The chapter is divided into luminance based and color based approaches. For luminance based approaches, corner detectors, blob detectors, and symmetry based interest points are discussed.

Regarding symmetry based interest points, the proposed GVFpoints are described in detail. For these approaches, methods to estimate the scale in an affine invariant way are discussed. For color based approaches, corner and blob detectors are described. The proposed scale-invariant color points are described in detail.

Chapter 4 carries out extensive feature evaluation experiments on various image data-sets. First, the data-sets used are outlined. Starting with robustness experiments, the main properties of the proposed features are given. These are verified in image retrieval experiments under well-defined, artificial images and large scale object categorization experiments on natural images. The chapter concludes with a report of the international benchmark in conjunction with the CVPR'09 where the proposed features outperformed the state-of-the-art in 4 out of 20 classes using only a fraction of the number of features of the other approaches.

Chapter 5 gives the state-of-the-art of interest point detection in videos. First, the main concept of spatio-temporal features are given. Similarly to the previous chapter, the approaches are divided in luminance and color based approaches. Beginning with corner detectors followed by blob detectors, the relation of spatio-temporal features to their image counterparts are discussed. An approach without an image counterpart using Gabor filters is described. As there are no color based spatio-temporal features so far, the mathematical concepts to extend these approaches to color are outlined.

Chapter 6 follows the experimental set-up of the previous chapter and gives an outline of existing data-sets for evaluation of spatio-temporal features. A new data-set is described in detail allowing for a comprehensive and principled evaluation of video features in the same way as carried out for image features. The state-of-the-art of spatio-temporal features are extracted from this data-set and their properties and behavior are discussed. An efficient robustness measurement is suggested carried out on the proposed data-set. Finally, a video matching experiment is carried out proposing a new way of local feature evaluation using artificially altered videos.

Chapter 7 concludes the thesis. The main achievements of the work are discussed and a brief outlook for future work is given.

CHAPTER 2

Principles

This section gives the fundamentals of the approaches proposed in this thesis. It starts by deriving the two color spaces used for the extraction of salient points in Section 2.1. Subsequently, the basics for the feature extraction are given. The main properties of the Gaussian kernel are given in Section 2.2. The structure tensor is described in detail in Section 2.3. A brief survey of local image and video description is given in Section 2.4 to define the methods used in the experiments. To conclude the section, the concept of feature invariance is described in Section 2.5 with a discussion about the desirable perfect feature and its properties.

2.1 Color Spaces and Perception

Typical tv screens produce color by adding up the primary colors red, green and blue in seperate channels and combine them by varying the intensities per channel. The single color channels are denoted as $\mathcal{R}$, $\mathcal{G}$ and $\mathcal{B}$. Defined by the gamut range, the resulting mixtures in $\mathcal{RGB}$ color space can reproduce a wide but limited variety of colors. The relationship between the varying intensities of red, green, and blue light and the resulting color is unintuitive, especially for inexperienced users. Neither additive nor subtractive color models define color relationships the same way the human eye does [Berk et al., 1982].

Computer vision algorithms used on color images are a extensions to algorithms designed for grayscale images [Tuytelaars and Mikolajczyk, 2008]: As will be shown later in this thesis, each color component is separately passed through the same algorithm and combined afterwards in the simplest case. Because the $\mathcal{RGB}$ representation of an object's color appearance changes significantly changing with the amount of light being reflected by the object, image features in terms of those components make visual similarity infeasible to model [Stokman and Gevers, 2007]. In these terms, descriptions in cylindrical color spaces are more relevant [Cheng et al., 2001].

Therefore the representation of the color coordinates of $\mathcal{RGB}$ information in cylindrical coordinates is used. The coordinates are denoted as hue, saturation or chroma and lightness.

One drawback of this representation is the large choice of available transformations from an $\mathcal{RGB}$ space, e.g. HSV [Smith, 1978] , HSL, $HMMD$ [Manjunath et al., 2002], HSB and HSI [Gonzalez and Woods, 1992].

When the saturation is normalized by the lightness the transformation becomes unstable and noisy: The saturation for the HSV and HSL models is obtained by the percentages of the maximum saturation obtainable for a given lightness. This implies that in the lightness ranges where the saturation range is small, there a large variation is encountered.

[Hanbury, 2008] derives a cylindrical coordinate color space which overcomes this drawback and provides a way the above representations can be reduced to a unified model. In this thesis, a similar approach is used: The proposed interest points are derived in two color spaces encoding luminance and chroma information separately. The Opponent Color Space (OCS) is defined as

$$OCS = \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} = \begin{pmatrix} \frac{\mathcal{R}-\mathcal{G}}{\sqrt{2}} \\ \frac{\mathcal{R}+\mathcal{G}-2\mathcal{B}}{\sqrt{6}} \\ \frac{\mathcal{R}+\mathcal{G}+\mathcal{B}}{\sqrt{3}} \end{pmatrix}. \tag{2.1}$$

Compared to other OCS (e.g. [Plataniotis and Venetsanopoulos, 2000; Lambert and Carron, 1999]) color space definitions, this transformation gives rotated chromaticity axes and different normalization. This orthonormal transformation into OCS provides specular variance. As this color space can be motivated by simulating primate retinal processes, the opponent colors blue/yellow and red/green are the end points of the o_1 and o_2 axis of the color space. As primates do not see combinations of these colors (e.g. a "blueish yellow" or a "greenish red") it is argued that the co-occurrence of these opponent colors attracts the most attention. Therefore the largest distance is defined between them.

A polar transformation on o_1 and o_2 of the OCS leads to the HSI color space

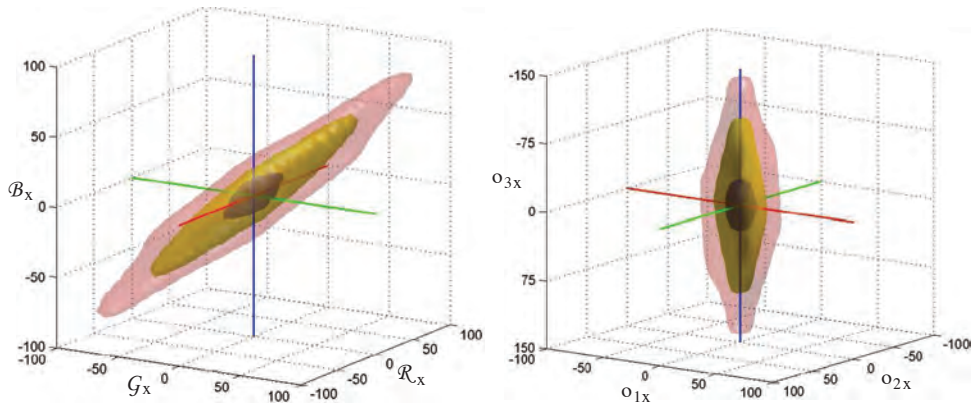
$$HSI = \begin{pmatrix} h \\ s \\ i \end{pmatrix} = \begin{pmatrix} \tan^{-1}\left(\frac{o_1}{o_2}\right) \\ \sqrt{o_1^2 + o_2^2} \\ o_3 \end{pmatrix}. \tag{2.2}$$

The derivative of the hue component h provides both the shading and the specular quasi-invariant [van de Weijer and Gevers, 2005], as it is both perpendicular to the shadow-shading direction and the specular direction. This means that those light effects should not change these coordinates. Obviously, this does not apply to every specular and shadowing effect in a natural scene. Further, small changes around the grey axis result in large changes in the hue.

The idea behind *color boosting* is to assign higher saliency to rare colors and thus a larger distance to more common colors [van de Weijer et al., 2006]. Following [Montesinos et al., 1998], they use the color jet of first order for a local description of color pixels to provide the local description v of a pixel in an image.

$$v = (\mathcal{R}\ \mathcal{G}\ \mathcal{B}\ \mathcal{R}_x\ \mathcal{G}_x\ \mathcal{B}_x\ \mathcal{R}_y\ \mathcal{G}_y\ \mathcal{B}_y)^T \tag{2.3}$$

where the indices x and y indicate the direction of the derivative of the color information. From information theory, it is known that the information content of an event is dependent on its



(a) Distribution of RGB derivatives (b) Distribution of OCS derivatives

Figure 2.1: Histogram of the distribution of derivatives in x direction of the RGB colors (a) and its transformation into OCS (b). The inner region covers 90%, the second 99% and the outer 99.9% of the total number of pixels of the 40000 pictures regarded. From: [van de Weijer et al., 2006]

frequency or probability. A color occurrence can be seen as an event in an image. As proposed in [van de Weijer et al., 2006], colors have different occurrence probabilities and therefore different information content $i(v)$ of a color description v:

$$i(v) = -log(p(v)). \tag{2.4}$$

where $p(v)$ is the occurrence probability of the descriptor v. Therefore, events which occur rarely are more informative. The information content of the descriptor v is approximated by estimating independent probabilities of the elements of v.

Looking for rare colors, statistics for the Corel Database containing 40000 color images showed that the three dimensional color distribution of derivatives of the color information was, as stated by the authors, remarkably significant (see Fig. 2.1). For all considered color spaces, one coordinate coincides with the axis of the maximum variation.

The color boosting transformation is obtained by approximating the surface of the three-dimensional color distribution obtained from a set of images by an ellipsoid. The ellipsoid is then transformed to a sphere, so that vectors of equal saliency lead to vectors of equal length. Gradient strength is so replaced by information content, so that higher gradient strength means higher saliency.

Traditionally, the derivatives of color vectors with equal vector norms have equal impact on the saliency function. The goal is to find a boosting function so that color vectors having equal information content have equal impact on the saliency function. This is a *color saliency boosting* transformation $g : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ such that

$$p(\mathbf{f}_x) = p(\mathbf{f}'_x) \leftrightarrow \|g(\mathbf{f}_x)\| = \|g(\mathbf{f}'_x)\|, \tag{2.5}$$

where $\mathbf{f}_x$ and $\mathbf{f}'_x$ are the derivatives in the x direction of two arbitrary color coordinate vectors $\mathbf{f}$ and $\mathbf{f}'$ of the form $(\mathcal{R}_x\ \mathcal{G}_x\ \mathcal{B}_x)^T$. The transformation is obtained by deriving a function describing the surface of the 3 dimensional color distribution, which can be approximated by an ellipsoid. The third coordinate of the color space is already aligned with the luminance, which forms the longest axis of the ellipsoid. The other two axes are rotated so that they are aligned with the other two axes of the ellipsoid. These derivative histograms can then be approximated by ellipsoids having the definition

$$(\alpha h_x^1)^2 + (\beta h_x^2)^2 + (\gamma h_x^3)^2 = R^2, \tag{2.6}$$

where $h_x^{[1..3]}$ is the transformation of a color derivative followed by the rotation to align the axes with those of the ellipsoid in the corresponding color space. To find the transformation g (compare Eq. 2.5), the ellipsoid is transformed to a sphere, so that vectors of equal saliency lead to vectors of equal length. The function g is therefore defined as

$$g(\mathbf{f}_x) = \mathfrak{M}h(\mathbf{f}_x), \tag{2.7}$$

which leads to a saliency boosting factor for each component of the corresponding color space. As stated in [van de Weijer et al., 2006] for the opponent color space, the diagonal matrix $\mathfrak{M}$ is given by

$$\mathfrak{M} = \begin{bmatrix} 0.850 & 0 & 0 \\ 0 & 0.524 & 0 \\ 0 & 0 & 0.065 \end{bmatrix}. \tag{2.8}$$

The shading and specular quasi-invariant HSI and the color boosted OCS color space are used to extract more salient interest points proposed in this thesis. They are referred to as *color invariant points* and *color boosted points*, respectively.

2.2 Properties of the Gaussian Kernel

Generally, Gaussian functions are of the class

$$f(x) = ae^{\frac{(x-b)^2}{2c^2}}, \tag{2.9}$$

where a, b and c are real positive constants and $e \approx 2.718$ denotes the Euler's number. The graph of such functions is a symmetric *bell curve* which is shown in Fig. 2.2 next to Carl Friedrich Gauss. e is the unique number such that the value of the derivative of the exponential function $f(x) = e^x$ at the point $x = 0$ is equal to 1. *Iff* $a = (c\sqrt{2\pi})^{-1}$, the integral of the function equals 1.

For two dimensions the Gaussian kernel G is defined by

$$G = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{2.10}$$

with the standard deviation σ as a parameter. This can be extended to any dimensionality. The standard deviation of G can be interpreted as a measure of its size. Additionally, for scale space

Figure 2.2: 10 Mark bill showing Carl Friedrich Gauss and the Gaussian bell curve.

implementations, it is related to the scale t by $t = \sigma^2$. The Fourier transform of a Gaussian function yields a Gaussian function and is thus non-negative. In the following, the main properties are given.

Normalization of the Gaussian kernel The term $(2\pi\sigma^2)^{-1}$ in front of the kernel is the normalization constant which results in unity for its integral over its full domain. Hence, the amplitude of the function decreases with increasing σ. In other words, blurring an image with this kernel gives equal average luminance (or response) for every scale. This is often referred to as average grey level invariance of the Gaussian kernel. These properties will be exploited to build a scale space [Lindeberg, 1998] for scale invariant interest points later in this section.

Separability of the Gaussian kernel Operations with the Gaussian kernel can be calculated per direction independently. For dimensions higher than one, the G can be expressed as the regular product of one dimensional kernels per dimension. For the two dimensional circular case

$$G = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{y^2}{2\sigma^2}}, \tag{2.11}$$

this can be simplified easily to Eq. 2.10. This allows for detectors to be run independently and in parallel per dimension. This is crucial as the runtime of a convolution increases with $O(\sigma^2)$ Therefore, estimation of convolutions with large σ are expensive. In order to apply this more efficiently, it is possible to re-use all the convolutions from one σ to a bigger one by cascading the operations:

Cascading of Gaussian kernels As the Gaussian kernel is a linear operation it allows for a cascade of convolutions $\otimes$ of smaller σ instead of using one bigger σ where the resulting new,

equivalent σ_n is

$$\sigma_n = \sqrt{\sigma_1^2 + \sigma_2^2 + ... + \sigma_N^2}, \tag{2.12}$$

leading to a Gaussian G_{σ_n} of size σ_n

$$G_{\sigma_n} = G_{\sigma_1} \otimes G_{\sigma_2} \otimes ... \otimes G_{\sigma_N} \tag{2.13}$$

for Gaussians of size σ_1 to σ_N. This results in a more efficient implementation of the algorithm especially for scale invariant approaches where convolutions with smaller σ are cascaded for larger scales.

Convolutions are commutative so that for every function of image intensities I

$$I \otimes G = G \otimes I \tag{2.14}$$

and the derivative of such a convolution satisfies

$$I' \otimes G = (I \otimes G)' = I \otimes G', \tag{2.15}$$

which will be used further on to estimate derivatives of I. An image derivative convolved with a Gaussian is therefore efficiently estimated by a convolution with the Gaussian derivative G_x in x direction and G_y in y direction.

2.3 The Structure Tensor

Interpreting visual data for its structural properties gives the need for an orientation estimation and a local structure analysis: Applications include texture analysis [Kass and Witkin, 1987], optical flow analysis [Bigün et al., 1991], shape analysis [Lindeberg and Gårding, 1993], fingerprint analysis [Weickert, 1998], segmentation of natural images [Rousson et al., 2003] and corner detection [Förstner and Gülch, 1987].

The straightforward way to get an idea of orientation is to use the gradient vector at a certain point and regard it as the orientation of this pixel. However, it is impossible to describe an orientation at a certain point just by means of this point itself. Consequently, the basic concept of the structure tensor is to describe the local structure of a point by describing its surroundings [Brox et al., 2006]. This description is given in a matrix array and became a very popular operator for robust orientation estimation.

Matrices and a tensor are both basically data arrays in a well defined arrangement of numerical entities. The main difference of a tensor to a matrix, is that the structure tensor describes physical quantities and is a geometric entity which thus lost its spatial and dimensional information. The rank of a tensor describes the dimensionality of the information it contains, or briefly, the number of indices that are needed to describe the elements of the tensor.

A very popular application for the structure tensor is the local optical flow estimation based on [Lucas and Kanade, 1981]. The approach is a two-frame differential method which assumes a locally constant flow. It looks for the direction in both the spatial and temporal domain which provides the least changes. It is shown for a structure tensor of a spatio-temporal image patch

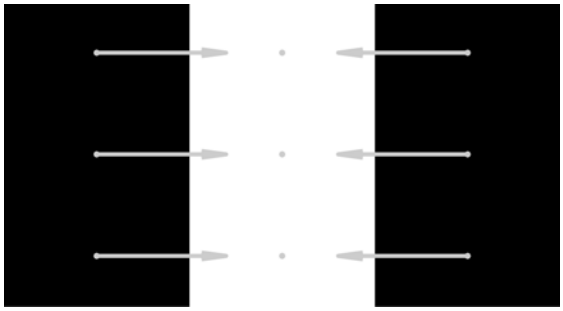


Figure 2.3: 3×3 pixel patch with thin vertical line and its gradients ΔI. $\sum \Delta I = 0$.

that this is the eigenvector to the smallest eigenvalue [Bigün et al., 1991]. The method is robust to noise and outliers in the local patch. It provides compared to other methods (e.g. compared to the global estimation method of [Horn and Schunck, 1981]) a relatively sparse vector field which fades out quickly across motion boundaries.

In texture analysis, the dominant orientation of an image patch is typically used as a feature for texture recognition and classification. As the structure tensor is robust to noise, the method also referred to as anisotropic diffusion is a local and scale invariant way to reduce noise in an image to reconstruct and enhance visual data. As a crucial part of this task is to let hard edges and corners remain stable, the orientational description of the structure tensor gives a reliable way to estimate this kind of structure [Perona and Malik, 1990]. This leads to the focus of this section: corner detection through the structure tensor using it both for estimation of location, scale and the description of local features in visual data.

The structure tensor enriches the information of one pixel by its gradient (for a first order approximation) and by the gradients of its surroundings in the radius σ weighted by a Gaussian kernel G (see Eq. 2.10).

Describing image information with gradients leads to certain drawbacks: The Gaussian convolution may cancel out fine structures in the image function regarded. On the one hand, it provides a representation which is robust to noise, but disregards thin edges and corners within the image. For a thin line, the two gradients perpendicular to the line will cancel each other mutually out, as the one is positive and the other one is negative (compare Fig. 2.3).

The solution of this problem is the outer product of the vectors. It allows the local description to be a symmetric positive semi-definite matrix of partial derivatives, which is often referred to as the initial matrix field J_0. The indices of L denote the orientation of the derivatives of I.

$$J_0 = \begin{pmatrix} L_x^2 & L_{xy} \\ L_{xy} & L_y^2 \end{pmatrix} \tag{2.16}$$

The three elements L are estimated by convolution and combining the image I and the derivatives of the Gaussian kernel $G_{x\sigma}$ and $G_{y\sigma}$ by

$$
\begin{aligned}
L_x^2 &= I^2 \otimes G_{x\sigma}, \\
L_y^2 &= I^2 \otimes G_{y\sigma}, \\
L_{xy} &= (I \otimes G_{x\sigma})(I \otimes G_{y\sigma})
\end{aligned}
\tag{2.17}
$$

where the subscript x or y indicates the direction of the gradient and kernel.

The structure tensor can be extended to multi-channel information, where every pixel is seen as a vector. This applies for example to color images. L is then a set of partial derivatives $L = (L_1, ..L_N)$ that is integrated by summing up all elements per color channel and direction.

$$
J_0 = \begin{pmatrix} \sum_{i=1}^{N} L_{ix}^2 & \sum_{i=1}^{N} L_{ixy} \\ \sum_{i=1}^{N} L_{ixy} & \sum_{i=1}^{N} L_{iy}^2 \end{pmatrix}. \tag{2.18}
$$

The structure tensor of a certain scale is defined by the size σ of G_σ. The elements of the structure tensor are convolved with a Gaussian kernel with the according size G_σ.

$$
J_\sigma = G_\sigma \otimes J_0. \tag{2.19}
$$

Note that this is not a smoothing of the image, but a smoothing of the structure information. This has an important effect on the subsequent interpretation of the predominant orientation within the patch.

Orientation Assignment The smoothing of the data in Eq. 2.19 increases the robustness of the tensor against noise. It makes the orientation assignment more reliable against artifacts in the visual data and other unwanted effects in images. Further, the convolution distributes the information about the orientation also to regions with gradients close to zero. In other words, it is also estimating orientation in areas "between" edges. The dominant orientation of a patch is obtained as the largest eigenvector $\mathbf{e}$ to the largest eigenvalue λ (also referred to as the characteristic value).

An eigenvector $\mathbf{e}$ of the tensor J is defined as the vector where the following equation holds:

$$
J\mathbf{e} = \lambda\mathbf{e}, \tag{2.20}
$$

where λ is a real scalar. In this context, the tensor is seen as an operator on the eigenvector, where $\mathbf{e}$ is left unchanged by J. To solve the equation, it is solved for λ and rearranged to $J\mathbf{e} - \lambda\mathbf{e} = 0$. By extending λ by the identity matrix $\mathbb{I}$ it follows

$$
(J - \lambda\mathbb{I})\mathbf{e} = 0. \tag{2.21}
$$

Therefore, by solving the polynomial determined by the characteristic equation $\det(J - \lambda\mathbb{I}) = 0$ the two non-negative eigenvalues of a structure tensor J of rank 2 are obtained.

$$
\lambda_{1,2} = \frac{L_x^2 + L_y^2 \pm \sqrt{L_x^2 - L_y^2 + 4L_{xy}}}{2}, \tag{2.22}
$$

This equation is a homogeneous system of 2 equations with 2 unknowns. This equation has a non-zero solution for **e** if and only if Eq. 2.21 is zero. The semi-definite symmetric tensor of order 2, J, has two non-negative eigenvalues which encode the quantity of the "interest" of a structure towards the corresponding orthogonal eigenvectors $\mathbf{e}_{1,2}$ [Köthe, 2003].

The difference of the eigenvalues $\lambda_1 - \lambda_2$ measures the edge strength in the patch. $2\lambda_2$ can be interpreted as a junction strength. Formally, these measures are contracted along the main diagonal by using the trace of the tensor $\lambda_1 + \lambda_2$. Therefore, rotational invariance is achieved.

$\mathbf{e}_1$ is a unit vector orthogonal to the major gradient edge with the orientation ψ given by

$$\psi = \frac{1}{2} \arctan \left(\frac{2L_{xy}}{L_x^2 - L_y^2} \right). \tag{2.23}$$

The eigenvalues give the extent of anisotropy or in other words, a *certainty* of the dominant orientation by their coherence c,,

$$c = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} \right)^2, \tag{2.24}$$

when $(\lambda_1 + \lambda_2) > 0$, otherwise, c is defined as 0.

These properties are predominantly used for local feature detection and description. As for detectors the run-time of an approach is a crucial issue, approximations of the measures above are used. For approximation, the explicit eigen-decomposition can be avoided, especially when the exact values of the eigenvalues do not have to be known, but just the relation between them.

$$\begin{aligned} \det(J) &= \lambda_1 \lambda_2 = L_x^2 L_y^2 - L_x^2 L_y^2 \\ \text{trace}(J) &= \lambda_1 + \lambda_2 = L_x^2 + L_y^2 \end{aligned} \tag{2.25}$$

Trace and determinant of the tensor can be calculated directly from the tensor elements. The relation r of λ_1 and λ_2 is then estimated by substituting $\lambda_1 = r\lambda_2$ and following

$$\frac{\text{trace}(J)^2}{\det(J)} = \frac{(\lambda_1 + \lambda_2)^2}{\lambda_1 \lambda_2} = \frac{(r\lambda_2 + \lambda_2)^2}{r\lambda_2^2} = \frac{(r+1)^2}{r} \tag{2.26}$$

This basic approach for the orientation assignment holds certain drawbacks. When the orientation in a patch is not homogeneous the Gaussian convolution integrates structures in an ambiguous way that leads to inaccurate estimations. There are several ways to make the estimation more robust including applying the Kuwahara-Nagao operator [Zenzo, 1986; Jahne, 1993; Bakker et al., 1999] or adaptive Gaussian windows [Nagel and Gehrke, 1998; Middendorf and Nagel, 2002].

2.4 Local Description

This thesis focusses on the detection, localization and scale selection of features in images and videos. Typically, after detection of an image patch, its appearance has to be described. In the following section a survey of the main concepts of local image description is given. The

thesis proposes a scale invariant extension to a descriptor based on dissociated dipoles [Vanossi and Stöttinger, 2010]. In Section 2.4.2 the most successful local spatio-temporal descriptors are described.

2.4.1 Local Description in Images

The simplest local image descriptor is a vector of image pixel intensities. However, the high dimensionality of such a description results in a high computational complexity for recognition. Techniques that use histograms to represent different characteristics of appearance or shape are the tool of choice for robust description.

SIFT The most popular descriptor is the one proposed in [Lowe, 1999]: The scale invariant feature transform (SIFT), which initially combined a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a histogram of gradient locations and orientations. The contribution to the location and orientation bins is weighted by the gradient magnitude. The quantization of gradient locations and orientations makes the descriptor robust to small geometric distortions and small errors in the region detection.

The feature vector is created by first computing the gradient magnitude and orientation at each image sample point in a region around the interest point location. These are weighted by a Gaussian window. The samples are accumulated into orientation histograms summarizing the contents over 4×4 subregions.

SIFT uses the location of a detected interest point and its scale for its description. For a position $\mathbf{x} = (x, y)$ and a scale σ, the image I is convolved with the Gaussian kernel of size σ

$$I_\sigma = I \otimes G_\sigma. \tag{2.27}$$

The gradient magnitude $m(\mathbf{x})$ and $\theta(\mathbf{x})$ is estimated for a scale σ and the position $\mathbf{x}$ by

$$\begin{aligned} m_{\sigma\mathbf{x}} &= \sqrt{(I_\sigma(x+1,y) - I_\sigma(x-1,y))^2 + (I_\sigma(x,y+1) - I_\sigma(x,y-1))^2} \\ \theta_{\sigma\mathbf{x}} &= \arctan\left(\frac{I_\sigma(x,y+1) - I_\sigma(x,y-1)}{I_\sigma(x+1,y) - I_\sigma(x-1,y)}\right) \end{aligned} \tag{2.28}$$

The orientations $\theta_{\sigma\mathbf{x}}$ are stored in a 36 bin histogram covering 10 degrees per bin. They are weighted by the corresponding magnitude $m_{\sigma\mathbf{x}}$ and smoothed by a 1.5 times larger Gaussian window $G_{\frac{3}{2}\sigma}$ than the previously used scale σ. The highest peak in this histogram builds the dominant direction in the gradients. If there is any other peak bigger than 0.8 times the maximum, another descriptor with that dominant orientation is built.

The descriptor aims 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. Due to the fixed number of bins, the resulting feature vector is then of length 128.

To reduce the effect of non-linear illumination change, the highest values of the normalized feature vector are cut off to a value of 0.2. 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.

The method provides local image features developed for reliable object matching. One of the points that have been improved with respect to the original approach is the extraction of the salient points [Dorko and Schmid, 2003; Stöttinger, 2008]. Although several enhancements of this descriptor have been made (e.g. PCA-SIFT [Ke and Sukthankar, 2004] based on [Fergus et al., 2003], GLOH [Mikolajczyk and Schmid, 2005], SURF [Bay et al., 2006]), the original method is still state-of-the-art in a general context of experiments (e.g. [Mikolajczyk et al., 2005b]), especially under heavy transformations.

PCA-SIFT The PCA-SIFT [Ke and Sukthankar, 2004] is introduced to achieve a more compact local descriptor than the original SIFT. 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 dimensions. 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 outperforms the SIFT descriptor in matching performance.

SURF Speeded up Robust Features [Bay et al., 2006] are 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 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 the scale space and the gradients. 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. 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.

Another approach are spin images [Johnson and Hebert, 1996] introduced for 3D object recognition in the context of range data. Their representation is a histogram of the relative positions in the neighborhood of a 3D interest point. The two dimensions of the histogram are distance from the center point and the intensity value. [Zabih and Woodfill, 1994] developed an approach robust to illumination changes. It relies on histograms of ordering and reciprocal relations between pixel intensities which are more robust than raw pixel intensities. This descriptor is suitable for texture representation but a large number of dimensions is required to build a reliable descriptor [Ojala et al., 2002].

Spatial-frequency techniques describe the frequency content of an image. The Fourier transform decomposes the image content into linear combinations of the basis functions. However, in this representation the spatial relations between points are not explicit and the basis func-

tions are infinite, therefore difficult to adapt to a local approach. The Gabor transform [Gabor, 1946] overcomes these problems, but a large number of Gabor filters is required to capture small changes in frequency and orientation. Gabor filters and wavelets [Vetterli, 1995] are frequently explored in the context of texture classification.

Geometric histogram [Ashbrook et al., 1995] and shape context [Belongie et al., 2002] implement a similar idea of histograms of gradient orientations for local description. Both methods compute a histogram of location and orientation for edge points where all the edge points have equal contribution to the histogram. These descriptors were successfully used for example for shape recognition of drawings for which edges are reliable features.

Differential descriptors use a set of image derivatives computed up to a given order to approximate a point neighborhood. The properties of local derivatives (local jet) were investigated by [Koenderink and van Doorn, 1987]. [Florack et al., 1991] derived differential invariants, which combine components of the local jet to obtain rotation invariance. Freeman and Adelson [Freeman and Adelson, 1991] developed steerable filters, which steer derivatives in a particular direction given the components of the local jet. A stable estimation of the derivatives is obtained by convolution with Gaussian derivatives.

[Baumberg, 2000] and [Schaffalitzky and Zisserman, 2002] propose to use complex filters. These filters differ from the Gaussian derivatives by a linear coordinates change in filter response space. Generalized moment invariants have been introduced in [Gool et al., 1996] to describe the multi-spectral nature of the image data. The moments characterize shape and intensity distribution in a region. They are independent and can be easily computed for any order and degree. However, the moments of high order and degree are sensitive to small geometric and photometric distortions. Computing the invariants reduces the number of dimensions. These descriptors are therefore more suitable for color images where the invariants can be computed for each color channel and between the channels.

Scale invariant dipole descriptor The 20-dimensional dipole descriptor is introduced by [Joly, 2007]. In the original approach, a fixed scale detection and description of the features is proposed. For a limited scale invariance, the input images are scaled down iteratively and the feature estimation is carried out on these smaller resolutions in parallel.

In the following, a basic scale invariance for this descriptor is developed [Vanossi and Stöttinger, 2010]. A decision on the characteristic scale and thus a scale invariance is achieved. The main idea is to use color salient points [Stöttinger et al., 2007a] for the location and scale estimation and pass their characteristic scale on to the subsequent description phase where the descriptor takes the scale as a parameter.

Like a simple edge detector, a dissociated dipole is a differential operator consisting of an excitatory and an inhibitory lobe and may be used at any orientation or scale. However, unlike a conventional edge detector, it allows an arbitrary separation between these two lobes, removing the correlation of inter-lobe distance and lobe size. Therefore, the dipole filters take advantage of having a flat variation at their center, providing a better robustness to localization errors, as shown in Fig. 2.4.

To achieve rotational invariance, each interest point is assigned an orientation following

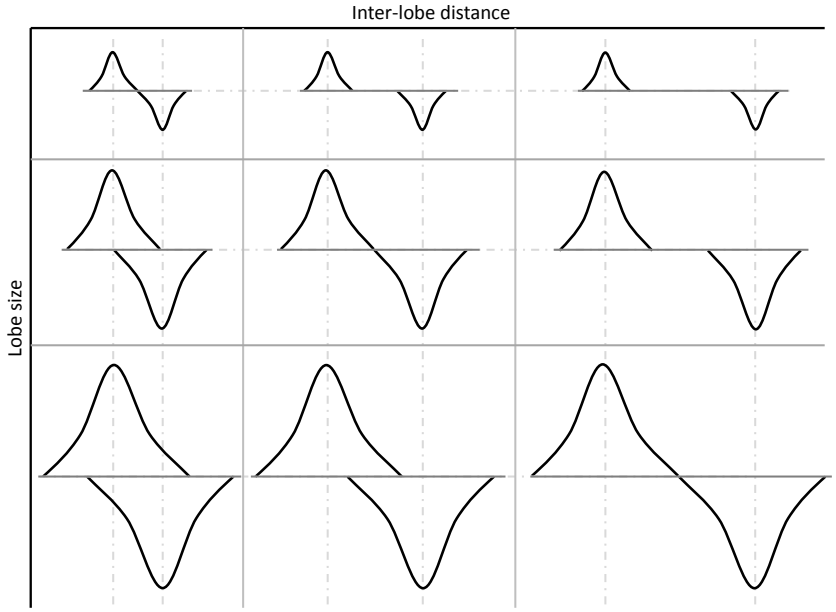


Figure 2.4: De-coupling of the parameters of inter-lobe distance and lobe size. From: [Balas and Sinha, 2003]

the SIFT approach: An orientation histogram is formed from the gradient orientations within a patch around $\boldsymbol{P}$ while the dominant direction θ_0 is estimated according to the highest peak in this histogram. Any other local peak that is within 80% of range is used to create a new interest point with the corresponding orientation. For better accuracy, the position of the peaks are interpolated by a parabolic fit on 3 histogram values.

Each dipole consists of a pair of Gaussian lobes, with standard deviation σ and a spatial separation of δ, and can be computed by the difference of two levels in the Gaussian scale-space. In order to be fully invariant to scale changes, the spatial separation δ_1 is set to the characteristic scale of the interest point and the size of the Gaussian window σ is defined by $\frac{\delta_1}{2}$. The descriptor is composed of two sub-vectors $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$. $\boldsymbol{F_1}$ is composed of 8 dipoles and has scale σ, while $\boldsymbol{F_2}$ has 12 dipoles with half scale.

First order dipoles: Let F be the vector composed of values in the scale space $I_{\mathbf{x}\sigma}$ in 12 directions at a distance δ_1 around the patch $\boldsymbol{P}$, its i-th component g_i being defined as $g_i = I_{\mathbf{x_i}\sigma_1}$ with $x_i = x_c + \delta_1 \cdot \cos(\theta_i)$, $y_i = y_c + \delta_1 \cdot \sin(\theta_i)$. where (x_c, y_c) is the center of patch $\boldsymbol{P}$ and

$$\theta_i = \theta_0 + (i-1) \cdot \frac{2\pi}{12} \tag{2.29}$$

First order vector $\boldsymbol{F_1}$ is then obtained by forming $\mathbf{D}_1$ = 8 dipoles from the components of F according to the linear relation $\boldsymbol{F_1} = \boldsymbol{A} \cdot F$ where

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.30}$$

Second order dipoles: $\boldsymbol{F_2}$ is composed of $\mathbf{D}_2$ = 12 dipoles at scale $\frac{\sigma}{2}$ and $\delta_2 = \frac{\delta_1}{2}$. They are computed along 12 orientations at a distance δ_1 around the interest point $\boldsymbol{P}$ according to:

$$\boldsymbol{F_2} = \begin{pmatrix} f_1^2 \\ .. \\ f_i^2 \\ .. \\ {f_{12}}^2 \end{pmatrix}, \text{where } f_i^2 = I_{\mathbf{x}_i^{'}\sigma_2} - I_{\mathbf{x}_i^{''}\sigma_2} \text{ and } \begin{bmatrix} x_i^{'} = x_c + (\delta_1 + \delta_2).\cos(\theta_i) \\ y_i^{'} = y_c + (\delta_1 + \delta_2).\sin(\theta_i) \\ x_i^{''} = x_c + (\delta_1 - \delta_2).\cos(\theta_i) \\ y_i^{''} = y_c + (\delta_1 - \delta_2).\sin(\theta_i) \\ \theta_i = \theta_0 + (i-1).\frac{2\pi}{12} \end{bmatrix} \tag{2.31}$$

In order to be invariant to affine luminance transformations, the two sub-vectors $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$ are normalized on a sphere by dividing them by their L_2-norms $||\boldsymbol{F_1}||_2$ and $||\boldsymbol{F_2}||_2$.

2.4.2 Local Description in Videos

The problem of a both compact and meaningful representation of events in spatio-temporal data has been the focus of recent research [Junejo et al., 2008; Laptev et al., 2008; Shechtman and Irani, 2007]. The straightforward extension to scale invariant spatio-temporal descriptors is to extend the concept of local jets [Koenderink and van Doorn, 1987; Florack et al., 1996] to the temporal domain and build an array of Gaussian derivatives j in all possible directions and their combinations in the corresponding scale of the detected interest point:

$$j = (L_x, L_y, L_t, L_{xx}, L_{xy}, ..., L_{tttt}), \tag{2.32}$$

where the elements L of the vectors are the Gaussian derivatives in a detected volume. They are precomputed by the spatio-temporal detector. Therefore the computational cost of this descriptor is very low. The order of the jets denotes the order of derivative which is part of the descriptor. Eq. 2.32 denotes a local jet of order 4. One drawback of this method is the lack of rotational invariance.

For making the jets robust to illumination changes, the local jets can be normalized to have values between -1 and 1 and to a standard deviation of 1:

$$j_{norm} = \frac{j - \text{mean}(j)}{\text{std}(j)} \tag{2.33}$$

[Laptev et al., 2007] build multi-scale jets by using all nine combinations of three spatial scales ($\frac{\sigma}{2}$, σ, 2σ and $\frac{\tau}{2}$, τ, 2τ) for every position.

To describe the detected patches by local motion and appearance, histograms of spatial gradients and optical flow accumulated in space-time neighborhoods of detected interest points are computed referred to as **HOG/HOF** [Laptev et al., 2008]. HOG results in a descriptor of length 72, HOF in a descriptor of length 90. For best performance they are simply concatenated. The descriptor size is defined by

$$D_x(\sigma) = D_y(\sigma) = 18\sigma, D_t(\tau) = 8\tau. \tag{2.34}$$

where t is the temporal direction from frame to frame of the input video. τ denotes the scale of the Gaussian in the temporal domain.

Each detected volume is subdivided into a $n_x \times n_y \times n_t$ grid of cells. Then normalized 4-bin histograms of gradient orientations (HOG) and normalized 5-bin histograms of optic flow (HOF) are computed. The approach is inspired by the SIFT descriptor. In the experiments in Section 6.4, the grid parameters $n_x, n_y = 3$, $n_t = 2$ are used as suggested by the authors. The binaries are available online[1].

Similarly, Willems et al. [Willems et al., 2008] proposed the **SURF3D** (ESURF) descriptor which extends the image SURF descriptor to videos. A video volume is represented by a 288 dimensional vector of weighted sums of uniformly sampled responses of Haar-wavelets. The volume is defined around a detected volume, with the standard value 3 as a factor to the initial scale

$$D_x(\sigma) = D_y(\sigma) = 3\sigma, D_t(\tau) = 3\tau. \tag{2.35}$$

With these parameters, the SURF3D describes less visual data than the HOG/HOF descriptor. The 3D patches are subsequently divided into $M \times M \times N$ bins, where M denotes the spatial dimensions and N the temporal one. For the feature vector v, each cell is represented by a vector of weighted sums

$$v = (\sum d_x, \sum d_y, \sum d_t) \tag{2.36}$$

of uniformly sampled responses of the Haar-wavelets d_x, d_y, d_t along the three axes.

In case rotational invariance is required in the spatial domain, the dominant orientation is estimated similar to [Bay et al., 2006]. For the spatio-temporal case, all Haar-wavelets used in this step are stretched out over the full τ. The binaries are available on-line[2].

The **HOG3D** [Kläser et al., 2008] is based on histograms of 3D gradient orientations efficiently computed using an integral video representation. Given a detected cuboid in the video volume, it is divided into $n \times n \times n$ subblocks. These n^3 subblocks form the set over which the histogram is built. For each of the subblocks the corresponding mean gradient is computed. It is subsequently quantized employing a regular polyhedron. With a fixed number of supporting mean gradient vectors, and by using integral videos for computing mean gradients of subblocks, a histogram can be computed for any arbitrary scale along x,y,t. This leads to a sparsely populated 960 dimensional vector. The descriptor is available on-line as an executable[3].

[1] `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`
[2] `http://homes.psat.kuleuven.be/~gwillems/research/Hes-STIP/`
[3] `http://lear.inrialpes.fr/software`

2.5 Invariance of Features

In computer vision one of the main challenges lies in classifying and recognizing objects from different views and lighting conditions. Depictions of natural scenes typically do not maintain their viewpoint, having rotational, perspective, projective and zoom changes between images of the same object.

Interest points have to focus on the same locations of an object, no matter from which point of view they are shown. When provided with stable interest points under these circumstances, local descriptors become more effective than using random features of an object [Fergus et al., 2003].

In this thesis robustness is used as a synonym for invariance. Regarding features, it is referred to as the ability to provide similar results under certain variation of the object or the change of the measurement. When large deformations of the appearance are encountered, the preferred approach is to model these mathematically if possible. The goal of invariant features is to develop methods for feature detection that are unaffected by these transformations. Typically, the term robustness is used in case of relatively small deformations. When it suffices to make feature detection methods imperishable to deformations, robustness is achieved. There is a trade-off between the accuracy of a feature and robustness. Common deformations of visual data are image noise, discretization effects, compression artifacts, and blur. For example, photometric changes can be mathematically modeled and invariance is achieved e.g. by using appropriate color spaces as described in Section 2.1. However, intensity based detectors and descriptors aim to be robust to these changes solely by making the approaches less sensitive by means of normalization and noise reduction.

Typically it is distinguished between the following different groups of invariance:

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 on the same location 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 the neighborhood affecting 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 there is change in the surroundings.

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 [Witkin, 1983], it is proposed that scale should be regarded as a continuous parameter for image representation, and [Lindeberg, 1994] showed that under some rather general assumptions, the Gaussian kernel. and its derivatives are the only possible smoothing kernels for scale space analysis. This *scale space theory* makes scale invariance possible as it allows the analysis of image data under varying resolution and size.

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

$$G(\Sigma) = \frac{1}{2\pi\sqrt{\det\Sigma}} e^{-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2}} \tag{2.37}$$

where Σ is the covariance matrix defining the affine transformation of the image. It is referred to as the *shape adapted Gaussian kernel*. This approach has four parameters to deal with, instead of one in the non affine cases. This leads to a more complex 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.

Lighting Invariance Robustness against illumination changes in natural scenes is desirable for interest points. Changes may occur because of change in lighting direction, lighting intensity or global illumination. These changes lead not only to a linear change of the luminance information, but introduce non-linear effects like shadows, highlights and shadow occlusions. For interest points, we desire stable and meaningful locations invariant to these effects [Quelhas, 2007].

Geometric Invariance Change in the viewpoint changes the object's appearance significantly. Due to occlusions, new parts of the objects are introduced or vanish from the picture. Geometric invariance aims for an optimal object representation being robust to these changes. Perfect geometric invariance would provide features of an object that do not change under different viewpoints. Projective transformations can be divided in three classes: affine, similarity, and Euclidean transformations. There are shape primitives that are invariant to these variations but suffer from noise and lighting changes leading to many small scale perturbations [Manay et al., 2006].

Properties of the Ideal Local Feature Following [Mikolajczyk and Tuytelaars, 2009], the perfect feature provides the following desirable properties:

Repeatability: Given two visual data of the same object or scene, recorded under different viewing conditions it should provide a high percentage of the features detected on the scene part visible in different instances. This measure is used extensively throughout the experiments with the provided comprehensive challenges. Repeatability is arguably the most important property of all. Moreover, a representation cannot provide perfectly repeated representation without understanding the scene: An object under heavy transformation changes its appearance heavily, so that without a deeper insight of the scene respective locations cannot be extracted.

Distinctiveness/Informativeness: The information patterns underlying the detected features should show a lot of variation, such that features can be distinguished and matched. This is debatable as for example symmetry based features do not necessarily extract the regions with the highest entropy but locate also homogeneous regions between symmetric structure. This can also lead to a robust representation as is proposed in Section 3.1.3.

Locality: The features should be local, so as to reduce the probability of occlusion. It would allow for simple models of the geometric and photometric deformations between two images or videos taken under different viewing conditions. This is often simplified by a local planarity assumption.

Quantity: The number of detected features should be sufficiently large, such that enough features are detected on every important object in the image. However, the optimal number of features depends on the application. It is broadly assumed that the number of detected features should be controllable over a large range by a simple and intuitive threshold. In this thesis, an alternative approach is evaluated in Section 4 where a fixed maximum number of more salient features ensures a smaller quantity but a better distinctiveness of features.

Density: The distribution of features should reflect the information content of the image to provide a compact image representation. A discussion of this property is given in Section 4.5.

Accuracy: The detected features should be accurately localized, in both locations of the matching, with respect to scale and possibly shape.

Efficiency: Preferably, the detection of features should allow for time-critical applications. Typically, the detection of features is not the bottleneck of the system. Most operations can be carried out in parallel and intermediate data can be used for the subsequent local description.

In summary, local features typically have a spatial extent, i.e., the local neighborhood of the detected location of a certain size. In contrast to segmentation, this can be any subset of an image. The region boundaries do not necessarily correspond to the changes in image structure. Also, multiple interest points may overlap, and "boring" parts of the image such as homogeneous areas can remain uncovered.

Ideally, one would like such local features to correspond to semantically meaningful object parts which remain stable on every instance of a picture of that object. In practice, however, this is unfeasible, as this would require high-level understanding and interpretation of the scene content, which is not available at this first stage of visual matching applications. Instead, detectors select local features directly based on the underlying image structure. The aim is to provide detectors that focus on salient and repeatable image patterns which allow the subsequent classification operations to perform as well as possible.

2.6 Summary

This chapter describes the basic techniques and tools for the detection and description of local features. As the basis for color based feature detectors, perceptual color spaces and their distinct properties are described in detail. As probably the most important concept in computer vision and used in all feature detectors in this thesis, the Gaussian kernel and its properties are given. This leads to a powerful mathematical model which relies on derivatives of the Gaussian kernels:

the structure tensor. It is used to interpret the structure of visual data and is used in both corner and blob detection.

For an image matching application, the detected locations have to be described in a robust and discriminative way. The main concepts for local image description and their extensions to spatio-temporal video description are described. An extension of the dipole descriptor is introduced. Based on color based local interest points, scale invariance is achieved. This allows for a better feature selection compared to the original approach.

The experimental chapters in this thesis aim to evaluate the robustness and performance of state-of-the-art local features. In this chapter, the different classes of feature invariance are presented leading to a discussion of the properties of a fictive perfect feature.

CHAPTER 3

Interest Point Detectors for Images

An image pattern which differs from its adjacent neighborhood is a local feature [Mikolajczyk and Tuytelaars, 2009]. Normally it is localized by a change of an observed image property or by observing several properties simultaneously. Nevertheless, it is not necessarily localized exactly on this change. The image properties commonly considered are intensity, color and texture. This chapter gives a survey of the main feature detectors for images. First, luminance based interest point detectors are explained in detail in Section 3.1. For color based interest points given in Section 3.2, their advantages and relation to the prior luminance based approaches are discussed.

3.1 Luminance Based detectors

In this section, the most successful approaches for detecting interest points on luminance information only are discussed. Corner detection approaches are desribed in Section 3.1.1. Section 3.1.2 gives a survey of blob detectors. Section 3.1.3 describes symmetry based methods, including the proposed GVFpoints.

3.1.1 Corner Detection

The main idea of using corners as an interest point detector is that corners provide the most stable and highest structured locations in an image being stable under geometric and lighting alterations. The problem of detecting corners can be seen as a problem of asking whether and how much an image function I is similar to itself when the position shifts from position (x, y) to position (u, v). This can be formalized as the sum of squared distances (SSD) [Moravec, 1977].

$$S(x, y) = \sum_{u,v} G(I(x + u, y + v) - I(x, y))^2 \tag{3.1}$$

(a) image 1

(b) image 6

Figure 3.1: Moravec corner detector on VOC Pascal image: Maxima of SSD of 1 pixel shift on image 1 (a) and image 6 (b) of the *graffiti* test-set.

G is the Gaussian kernel (Eq. 2.10) and can here also be substituted by a constant function. The maxima of S give the locations. An example is given in Fig. 3.1. It is known as the Moravec corner detector.

The Harris corner detector, introduced in [Harris and Stephens, 1988], provides a single scale corner measure overcoming these drawbacks. To become isotropic and be robust to shifts in any direction they approximate the function for small shifts by the first-order Taylor expansion:

$$S(x,y) = \sum_{u,v} G_\sigma (xX + yY + O(x^2, y^2))^2, \tag{3.2}$$

where the X and Y define the gradients in the patch in the corresponding direction, or in other words, the partial derivatives of the image function I.

$$\begin{aligned} X &= I \otimes (-1,0,1) = \frac{\partial I}{\partial x} \\ Y &= I \otimes (-1,0,1)^T = \frac{\partial I}{\partial y} \end{aligned} \tag{3.3}$$

This is can also be seen as an approximation of the convolution with first derivative of the isotropic 2D Gaussian Kernel $G_{x\sigma}$ and $G_{y\sigma}$ of $\sigma = 3$.

Following Eq. 3.2 for reasonable small shifts, S can be written as

$$\begin{aligned} S(x,y) &= L_x^2 x^2 + 2L_{xy} xy + L_y^2 y^2 \\ &= (x,y) M \begin{pmatrix} x \\ y \end{pmatrix}. \end{aligned} \tag{3.4}$$

where the L components are the elements of the structure tensor defined in Section 2.3 (compare Eq. 2.17) In the spatial domain the basis functions for the discrete Gaussian transform are

defined as the Gaussian and its derivatives with respect to the direction or dimension, respectively. The symmetric second moment matrix M describes the gradient distribution in the local neighborhood of a point giving a local structure tensor.

$$M = \begin{bmatrix} L_x^2 & L_{xy} \\ L_{xy} & L_y^2 \end{bmatrix} \tag{3.5}$$

where the subscript x or y indicates the direction of the gradient and kernel. [Harris and Stephens, 1988] state that one could use *for example* the Gaussian kernel to be more stable against noise.

The decision between corners and uniformness is defined as follows. The approach leads to a corner measure C_H, also referred to as Harris energy. It is based on the trace and determinant of the second moment matrix M using its eigenvalues λ_1 and λ_2.

Based on these values, a classification can be made: If they are near zero, there is 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.

$$\begin{aligned} C_H(M) &= \det(M) - \alpha \text{trace}^2(M), \\ \det(M) &= \lambda_1 \lambda_2 = L_x^2 L_y^2 - L_{xy}^2, \\ \text{trace}(M) &= \lambda_1 + \lambda_2 = L_x^2 + L_y^2 \end{aligned} \tag{3.6}$$

As the only constant, α indicates the slope of the *zero line*. For the original Harris border, 0 is the given border between corner and edge. The Harris detector remains stable up to a scale change of a factor of $\sqrt{2}$.

An extension of the Harris corner detector, a scale-adapted Harris detector *Harris Laplacian*, was introduced to achieve scale invariance as well as estimate local scale around the interest point by [Mikolajczyk and Schmid, 2002]. The main idea is to carry out the corner detection at multiple scales. When there are more candidates for corners at one location at multiple scales, an independent function decides on the *characteristic scale* of the structure. In the automatic scale selection, the term characteristic was chosen to the fact that the selected scale estimates the characteristic length of the corresponding image structures, in a similar manner as the term *characteristic length* is used in physics [Lindeberg, 1998].

The response of the Gaussian kernel decreases with higher scales in terms of the absolute value of the maxima as well as the number of maxima. To make the results of different scales comparable, scale normalization has to take place by the factor $\sigma 2$.

The second moment is convolved with a Gaussian kernel of the *integration* scale σ and is a constant factor l of the *differentiation* scale σ. The factor l is typically chosen to be 3. The second moment matrix M_σ of scale σ for the position $(\mathbf{x})$ is then

$$M_{l\sigma} = \left\{ \sigma^2 G_{l\sigma} \otimes \begin{bmatrix} L_{x^2\sigma} & L_x L_{y\sigma} \\ L_x L_{y\sigma} & L_{y^2\sigma} \end{bmatrix} \right\} (\mathbf{x}). \tag{3.7}$$

(a) image 1

(b) image 6

Figure 3.2: Harris Laplacian detector applied to the *graffiti* test-set. The size of the circles indicates the size (scale) of the kernel with the highest peak.

The scale space of the Harris function is built by iteratively scaling with a factor between 1.2 to $\sqrt{2}$. The result is a pyramid-like structure of C_H of multiple scales for the input image I. Every maximum of C_H is a candidate for an interest point.

To choose the characteristic scale of the candidates, the Laplacian of Gaussian (LoG) function Λ has been used to detect the characteristic scale automatically [Mikolajczyk and Schmid, 2001] it is found by using

$$\Lambda_\sigma = \sigma^2 |L^2_{x\sigma} + L^2_{y\sigma}|. \tag{3.8}$$

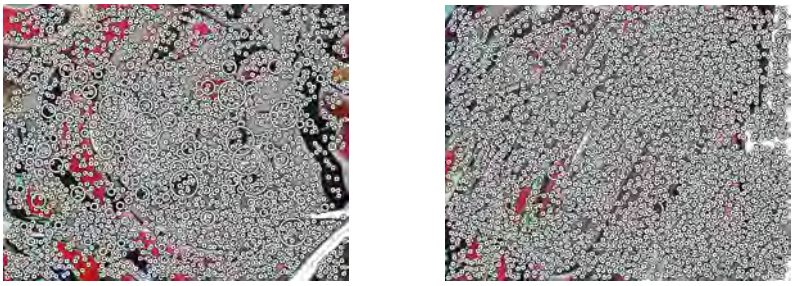
A characteristic scale of an interest point is found if both the Harris Energy and the Laplacian of Gaussian are extrema on that location and scale.

$$\nabla \Lambda_{\sigma_D} = \nabla M_{\sigma_D} = 0 \tag{3.9}$$

With this non-maxima suppression, the locations with their according scales are found. The affine invariant extension of the approach is described in Section 3.1.4. An example of the scale-invariant interest point detection is given in Fig. 3.2. The size of the white circles indicate the characteristic scale. As the approach aims to find the most stable scale of the local structure, the circles tend to fit in local structure. The graffiti test-set is described in Section 4.1.1.

3.1.2 Blob Detection

Blob detectors – based on the scale space theory introduced to computer vision by Witkin [Witkin, 1983] and extended by Lindeberg [Lindeberg, 1994] rely on differential methods such as Laplacian of Gaussians (LoG), difference of Gaussians (DoG) and Determinant of Hessian (DoH) [Lindeberg, 1998]. The result of blob detection using either LoG or DoG methods depends on the choice of scale sampling rate which is analyzed in [Lowe, 2004] using real images containing outdoor scenes, human faces, aerial photographs and industrial images.



(a) image 1 (b) image 6

Figure 3.3: DoG applied on image 1 (a) and image 6 (b) of the *graffiti* test-set.

Another technique within the class of blob detectors but unrelated to scale-space theory is Maximally Stable Extremal Regions (MSER) [Matas et al., 2002].

DoG

As demonstrated in [Lowe, 2004], the LoG can be approximated by the Difference of Gaussians (DoG) at reduced computational complexity. In the following, the implementation used in the evaluation is described. The features are localized by the difference between two Gaussian smoothed images of different σ. Subtracting a Gaussian kernel of bigger scale from one with smaller scale, the resulting kernel is an approximation of the Mexican Hat wavelet.

One level P_{σ_i} of the scale space is defined by

$$P_{\sigma_i} = G_{\sigma_i} \otimes I. \tag{3.10}$$

Every smoothing step is calculated with the 1D Gaussian function in the horizontal and vertical direction and with the parameter $\sigma_{i+1} = \sqrt{2}\sigma_i$. The image pyramid level D_i is computed as the difference of the P_{σ_i} and $P_{\sigma_{i+1}}$.

To detect the extrema, a pixel is compared to its 8 neighbors in one pyramid level D_i. If it is a maximum, it is compared to the 9 pixel neighbors in the adjacent levels D_{i-1} and D_{i+1}. This leads to an early diminished data-set, as the majority of the pixels are discarded right away in the first scale.

To make a candidate location $\mathbf{x}$ more accurate, [Brown and Lowe, 2002] developed a method to locate an interpolated location of the maximum. It allows location to be rejected when they are of low contrast or are localized along an edge. It uses the Taylor expansion of D at level i at the candidate position shifted to the origin.

$$D_{\mathbf{x}} \approx D_{\mathbf{x}} + D'_{\mathbf{x}} + \frac{1}{2} D''_{\mathbf{x}}, \tag{3.11}$$

where D' is the first derivative and D'' the second derivative at the location $\mathbf{x}$. The new location $\widehat{\mathbf{x}}$ is estimating

(a) image 1 (b) image 6

Figure 3.4: Hessian Laplacian applied on the *graffiti* test-set image 1 (a) and image 6 (b).

$$\widehat{\mathbf{x}} = -D''^{-1}D'. \tag{3.12}$$

If there is an offset $\widehat{\mathbf{x}}$ 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.

To discard points with low contrast, the newly estimated locations have to have a higher threshold on the DoG than 0.03. To discard edges and prioritize corners, the Hessian operator is used.

$$H = \left\{ \begin{bmatrix} L^2_{x\sigma} & L_{xy\sigma} \\ L_{xy\sigma} & L^2_{y\sigma} \end{bmatrix} \right\} (\mathbf{x}) \tag{3.13}$$

As described in Section 2.3, the eigenvalues of this structure tensor give a measurement of the curvature at the location. In case the determinant is negative, the point is discarded. In the original implementation, a threshold of $r = 10$ for $\lambda_1 = r\lambda_2$ is used as a threshold for a valid location. The approach is available online[1]. An example image is given in Fig. 3.3.

DoH

The determinant of the Hessian matrix (Eq. 3.13) gains great attention from researchers to extract scale invariant features (see Section 4.5). The distribution of features is similar to the Harris Laplacian representation, but much denser [Mikolajczyk and Schmid, 2004]. This gives an increased performance in various benchmarks, especially with powerful and robust classification techniques in combination with dense sampling of image patches.

They are complementary to their Harris-related counterparts, in the sense that they respond to a different type of feature in the image. Furthermore, this detector also responds to corner structures at fine scale. The returned locations, however, are more suitable for scale estimation than the Harris points due to the use of similar filters for spatial and scale localization, both based on second-order Gaussian derivatives [Tuytelaars and Mikolajczyk, 2008]. An example is given in Fig. 3.4.

[1] `http://www.cs.ubc.ca/~lowe/keypoints/`

(a) image 1 (b) image 6

Figure 3.5: Harris Laplacian and Hessian Laplacian locations applied on the *graffiti* test-set image 1 (a) and image 6 (b).

Recently, the combination of Harris Laplacian and Hessian Laplacian showed to give good performance in various object recognition tasks [Mikolajczyk and Uemura, 2008]. Example regions are shown in Fig. 3.5. In Section 4.5, combined detections are used in the approach with the best performance.

MSER

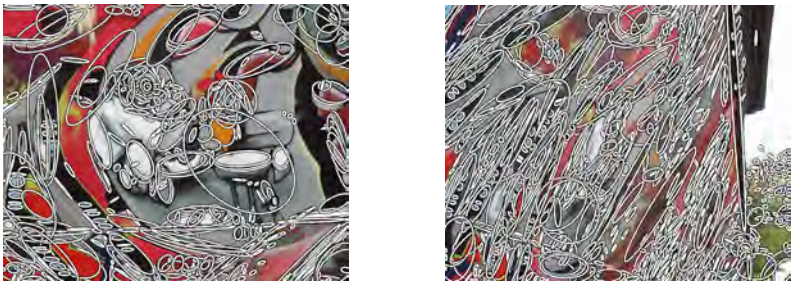
Maximally Stable Extremum Regions (MSER) [Matas et al., 2002] are obtained by a watershed like algorithm. Connected regions of a certain thresholded range are selected if they remain stable over a set of thresholds. The algorithm is efficient both in run-time performance and detection rate. The region priority or importance is measured in the number of thresholds where the region remains stable.

MSER have been used for wide baseline stereo matching as a region detector which is invariant to affine transformation of image intensities. The resulting regions are shown in Fig 3.6. Different to other detectors, MSER visually appear as ellipses are fitted in homogeneous regions of the image. They give a very robust and sparse distribution of features. However, this sparse representation is sensitive to parameter changes and lighting and shadowing effects and varying contrast, as is shown in the evaluation in Chapter 4.

An image region R is extremal if the intensity of all pixels $\mathbf{x} \in R$ shows to be higher than the intensity of boundary pixels $\mathbf{x}_b$ (adjacent to R) $I(\mathbf{x}) > I(\mathbf{x}_b)$. Region R is a contiguous image patch. For all pixels $R_\mathbf{x}$ in R a maximally stable region is found examining a shift Δ iff

$$\mathbf{x} = \frac{|R_{\mathbf{x}+\Delta}| - |R_{\mathbf{x}-\Delta}|}{|R_\mathbf{x}|} \tag{3.14}$$

is an extremum. Alternatively, the concept can be explained by a binary thresholding operation over different values of the intensity range of I. A threshold of zero results in no response at all. Increasing the threshold value, more and more of I responds until the threshold is the maximum intensity thus detecting the whole image. By analysis of the pixels that have the *longest* response over the set of thresholds, the regions are found.



(a) image 1 (b) image 6

Figure 3.6: Example of MSER locations applied on the *graffiti* test-set image 1 (a) and image 6 (b). The ellipses mark the most stable blobs.

The main advantage of the detector is its affine invariance and robustness to monotonic transformations of the contrast. Unfortunately, that does not apply for natural photometric effects. Contrary to what is stated many times in the internet (e.g.[2]) it is sensitive to non linear change of intensities as there are natural lighting effects like shadows. This is obvious as the approach examines intensities in a linear way and lighting effects change these in a non linear way. This can be overcome by incorporating color information which is described in Section 3.2.

3.1.3 Symmetry Based Interest Points

In this section, the detection of symmetry based interest points are described. The thesis proposes to use the extrema of the GVF for interest point extraction. An in-depth description is given here, evaluation of its robustness is given in Section 4.4.

The Generalized Symmetry Transform (GST) [Reisfeld et al., 1994] inspired the *Fast Radial Symmetry Transform* (FRST) [Loy and Zelinsky, 2002, 2003]. A pixel of the image contributes to a symmetry measure at two locations called negatively and positively affected pixels. The coordinates of negatively affected $\mathbf{p}_{-ve}$ and positively affected $\mathbf{p}_{+ve}$ pixels are defined by the gradient orientation at pixel $\mathbf{p}$ and a distance n (called *range* in [Loy and Zelinsky, 2002]) as follows:

$$\mathbf{p}_{+ve} = \mathbf{p} + \text{round}\left(\frac{\mathbf{g}(\mathbf{p})}{\|\mathbf{g}(\mathbf{p})\|}n\right), \quad \mathbf{p}_{-ve} = \mathbf{p} - \text{round}\left(\frac{\mathbf{g}(\mathbf{p})}{\|\mathbf{g}(\mathbf{p})\|}n\right) \tag{3.15}$$

The symmetry measure is a combination of orientation projection O_n and magnitude projection M_n maps, which are obtained through agglomeration of positively and negatively affected pixel contributions. Each positively affected pixel increments the corresponding element of the orientation projection map by 1 and magnitude projection map by $\|\mathbf{g}(\mathbf{p})\|$ while the negatively affected pixel decrements the map by these values:

$$O_n\left(\mathbf{p}_{+ve}(\mathbf{p})\right) = O_n\left(\mathbf{p}_{+ve}(\mathbf{p})\right) + 1, \quad O_n\left(\mathbf{p}_{-ve}(\mathbf{p})\right) = O_n\left(\mathbf{p}_{-ve}(\mathbf{p})\right) - 1 \tag{3.16}$$

[2] `http://en.wikipedia.org/wiki/Maximally_stable_extremal_regions`

(a) image 1

(b) image 6

Figure 3.7: Example of Loy symmetry points with a simple scale selection applied on the *graffiti* test-set image 1 (a) and image 6 (b). The size of the circles indicates the size of the range with a symmetry peak.

$$M_n\left(\mathbf{p}_{+ve}(\mathbf{p})\right) = M_n\left(\mathbf{p}_{+ve}(\mathbf{p})\right) + \|\mathbf{g}(\mathbf{p})\|, \qquad (3.17)$$
$$M_n\left(\mathbf{p}_{-ve}(\mathbf{p})\right) = M_n\left(\mathbf{p}_{-ve}(\mathbf{p})\right) - \|\mathbf{g}(\mathbf{p})\|$$

The radial symmetry measure at range n is a combination of normalized orientation and magnitude projection maps, additionally smoothed by a Gaussian kernel:

$$S_n = G_{\sigma_n} \otimes \left(\frac{M_n}{k_n}\right)\left(\frac{|O_n|}{k_n}\right)^{\alpha} \qquad (3.18)$$

where k_n is the scale normalization factor and α is the radial strictness parameter which allows to attenuate the symmetry response from ridges. The orientation projection map used for final calculations is thresholded using k_n. The symmetry measure can be also averaged over a set of ranges $N = \{n_1, ...n_K\}$ to achieve partial scale invariance:

$$S = \frac{1}{K}\sum_{n \in N} S_n \qquad (3.19)$$

An exhaustive discussion of the parameter choice and results are presented in [Loy and Zelinsky, 2003]. In the experiments these interest points are refered to as *Loy Points*. Two example images from the experiments are given in Fig. 3.7.

Gradient Vector Flow Based Interest Points

To detect points of high local symmetry GVF based interest points (GVFGVFpointspoints) are proposed. A similar approach has been previously used in [Donner et al., 2007] for localization of bone centers of known scales. The GVF [Xu and Prince, 1998], which yields a rotation invariant vector field, was originally proposed to increase the capture range of active contours. It is defined as the vector field $\mathbf{v}(x, y) = (u(x, y), v(x, y))$ which minimizes

$$\mathbf{G} = \int\int \mu(u_x^2 + u_y^2 + v_x^2 + u_y^2) + |\nabla f|^2 |\mathbf{v} - \nabla f|^2 dxdy \tag{3.20}$$

where f denotes the *edge map* of image I

$$f(x, y) = |(G_\sigma(x, y) * I(x, y))| \tag{3.21}$$

and the parameter μ gives the relation between the first smoothing term (compare with the classic optical flow calculation [Horn and Schunck, 1981]) and the second term. Its strengths include the ability to detect even weak structures while being robust to high amounts of noise in the image. When $|\nabla f|$ is small, the energy yields a very smooth field.

The field magnitude $|\mathbf{G}|$ is largest in areas of high image gradient, and the start and end points of the field lines of $\mathbf{G}$ are located at symmetry maxima. E.g. in the case of a symmetrical structure formed by a homogeneous region surrounded by a different gray level value the field will point away from or towards the local symmetry center of the structure.

The GVFpoints are thus defined as the local minima of $|\mathbf{G}|$. In contrast to techniques based on estimating the radial symmetry using a sliding window approach this will yield a sparse distribution of interest points even in large homogeneous regions. Increasing μ iteratively leads to an increasing smoothing of $\mathbf{G}$. As information is lost on local structure and $\mathbf{v}$ takes a gradually larger area into account, a rotation invariant scale space pyramid is built. For the experiments in Section 4.2.1, the parameters $\mu = 0.1$ and scale factor $s = 1.33$ are used. The scale factor is applied five times per image smoothing $\mathbf{G}$ for taking more area into account.

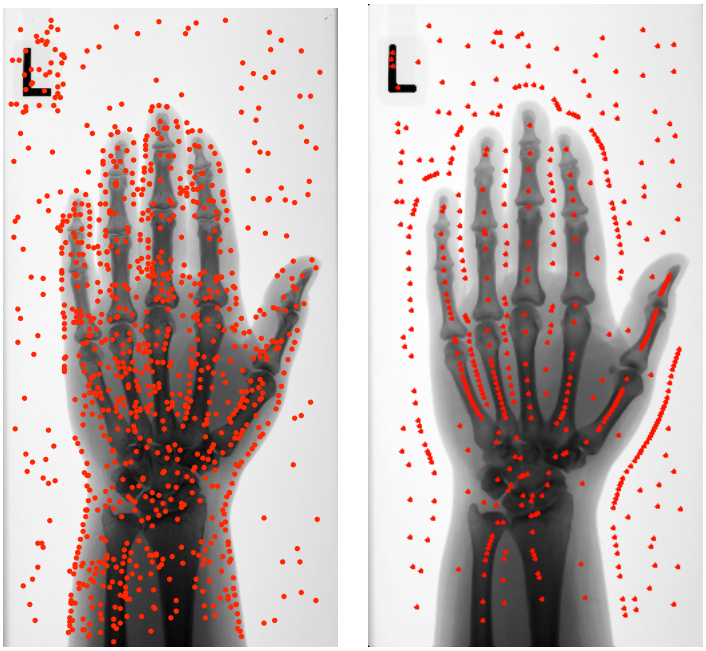
Example locations for the single scale approach [Donner et al., 2007] on medical data is shown in Fig. 3.8. The minima of the GVF distribute symmetrically on elongated structures. Examples on the *graffiti* test set of the proposed method are shown in Fig. 3.9: (d)-(f) show the distributions of the resulting points for increasing scale; (a)-(c) show the interest points on geometric transformations. Further examples are shown in Fig. 3.10.

3.1.4 Affine Invariant Estimation of Scale

To make features robust to affine transformations, every detected patch can be redefined in an affine invariant way: The transformation is determined that projects the intensity pattern of the point neighborhood to one with equal eigenvalues [Mikolajczyk and Schmid, 2004]. Note the parallels to the scale selection proposed in Section 3.2.1 and the color boosting [van de Weijer et al., 2006] which is used for color boosted points in the same section.

An affine invariant detector can be understood as a generalization of the scale invariant approaches to a non-uniform scaling and skew invariance. The main concept is to deal with different scaling factors in the orthogonal directions and without preserving the angles of the structure.

The transformation can be applied to every location detected. The main steps are to analyze the structure, apply the transformation to get an affine invariant transformed structure and apply



(a) (b)

Figure 3.8: Comparison of the (a) interest points found by Difference of Gaussians (DoG) and (b) the locations found as minima of GVF magnitude. From: [Donner et al., 2007]

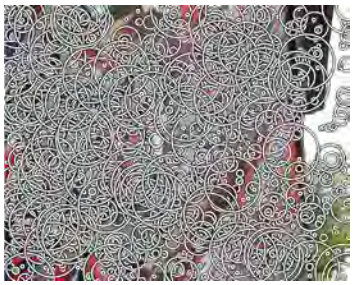
(a) frontal (b) 20 degree (c) 50 degree (d) 1st scale (e) 2nd scale (f) 3rd scale

Figure 3.9: GVFpoints on image details of the *graffiti* test-set. (a)-(c): GVFpoints under geometric transformation. (d)-(f): GVFpoints of the first three scales of one image detail.

the detector once again on that structure to redefine location and scale. This leads to a non uniform scale selection and thus to ellipses on the original image.

The transformation is given by the eigenvalues of the second moment matrix (compare Section 2.3). The patch is affine invariant when both eigenvalues are equal and is given by the square root of the second moment matrix $M^{\frac{1}{2}}$. The estimation of affine shape can be applied to any initial point given that the determinant of the second moment matrix is larger than zero and the signal to noise ratio is sufficiently large. Therefore this technique is able to estimate the shape of

(a) GVFpoints on image 1



(b) GVFpoints on image 6

Figure 3.10: GVFpoints applied to image 1 (a) and image 6 (b) of the *graffiti* test-set.

initial regions provided by the Harris-Laplace detector to ensure that the same part of the object surface is covered in spite of the deformations caused by the viewpoint change [Tuytelaars and Mikolajczyk, 2008].

In the literature, the resulting detectors are denoted as *Harris Affine* (shown in Fig. 3.11) and *Hessian Affine* (shown in Fig. 3.12). They are rarely used for image matching tasks as the estimation is time consuming and their performance is not as convincing as the mathematical concept promises. Repeatability does not go up for all geometrical challenges and decreases for some (e.g. example for uniform scale change) [Mikolajczyk and Schmid, 2004]. The assumption that a transformed patch is detected in affine variant way and is then described in an affine invariant way fails in extreme cases. Even when an affine transformation of the image would have been handled by the non-uniform scale selection perfectly, the initial detection is done in the scale invariant manner only, and therefore does not have to be detected in the first place.

The GVFpoints can be made affine invariant in a straightforward manner: After detecting the interest points the orientation $\alpha_i \in [0, 2\pi[$ of the local region surrounding the interest point can be estimated. Around each interest point rays $\mathbf{g}_\alpha^r$ at the 360 angles $\alpha \in [0, \ldots, 2\pi[$ at radii $r \in \{2, \ldots, 8\}$ are sampled from $\|\mathbf{G}\|$ using bilinear interpolation. The interest point i is then assigned the angle α_i which minimizes

$$\alpha_i = \underset{\alpha \in [0,2\pi[}{\operatorname{argmin}} \sum_r \mathbf{g}_\alpha^r. \tag{3.22}$$

The scale s_i of the region around the interest point is estimated by the mean distance of the interest point i to the two closest local maxima of $\|\mathbf{G}\|$ in the directions of α_i and $\alpha_i + \pi$ giving the two axes of the resulting ellipse.

3.2 Color based detectors

In the following, a survey of the approaches to use color information to detect local structure is given. The thesis proposes scale and color invariant interest points described in detail in

(a) Harris Affine on image 1

(b) Harris Affine on image 6

Figure 3.11: Harris Affine applied to image 1 (a) and image 6 (b) of the *graffiti* test-set.

(a) Hessian Affine on image 1

(b) Hessian Affine on image 6

Figure 3.12: Hessian Affine applied to image 1 (a) and image 6 (b) of the *graffiti* test-set.

Section 3.2.2. Color based blob detectors are given in Section 3.2.3.

Color based detectors do not have an excellent reputation in the research community. A recent survey [Tuytelaars and Mikolajczyk, 2008] states that most of the proposed approaches based on color are simple extensions of methods based on the intensity change. Color gradients are usually used to enhance or to validate the intensity change so as to increase the stability of the feature detectors but the pixel intensities remain the main source of information for feature detection.

In this thesis, the extraction of more salient and stable features is intended. The main question is if it makes sense to use better (and more expensive) features or simply to use more features. In this section, the algorithmic details are given. Experimental validation of the hypotheses is given in Chapter 4.

Predominantly used as color based interest points, the well known color Harris corner detector [Montesinos et al., 1998] is described as it is the basis for several approaches including the color interest points proposed in Section 3.2.2. The thesis presents an approach to incorporate color and scale invariances into the color tensor.

3.2.1 Corner Detection

Rugna et al. [Rugna and Konik, 2002] suggest a method to extract scale-invariant interest points based on color information for texture classification. They build a color Gaussian pyramid [Konik et al., 1996] and for every pyramid level and color channel, the original Harris energy is calculated. Features are selected based on their persistence through the pyramid. However, a scale selection based on the local structure is not achieved with this method. This method is independent of the color space used. The authors suggest the YUV or $CIELAB$ color space.

The most successful color features are based on the color Harris detector introduced in [Montesinos et al., 1998]. In image retrieval scenarios, they apply the fixed scale detector on gradually downsized images and use all of the detections extracted. This leads to multiple ambiguous features and no possibility to reuse precomputed results in the implementation. Therefore, it leaves the task of coping with ambiguous to the matching stage. They extend the second moment Matrix M for $\mathcal{RGB}$ information to

$$M = \left\{ \sigma^2 G_{l\sigma} \otimes \begin{bmatrix} \mathcal{R}_x^2 + \mathcal{G}_x^2 + \mathcal{B}_x^2 & \mathcal{R}_x\mathcal{R}_y + \mathcal{G}_x\mathcal{G}_y + \mathcal{B}_x\mathcal{B}_y \\ \mathcal{R}_x\mathcal{R}_y + \mathcal{G}_x\mathcal{G}_y + \mathcal{B}_x\mathcal{B}_y & \mathcal{R}_y^2 + \mathcal{G}_y^2 + \mathcal{B}_y^2 \end{bmatrix} \right\} (\mathbf{x}). \tag{3.23}$$

Instead of using just the intensity gradient, the gradient for each color channel is determined. These values are summed and averaged using a Gaussian kernel similar to Eq. 3.7. It is shown in [Gouet and Boujemaa, 2002] that at that time it has been the most stable interest point detector with respect to illumination changes, noise, rotation and viewpoint changes. It is successfully used in applications including object tracing [Gabriel et al., 2005], visual information retrieval [Rugna and Konik, 2002] and object-based queries [Gouet and Boujemaa, 2001].

[Faille, 2005] proposes a shadow, shading, illumination color and specularities invariant interest point localization which models the color information as terms C^R, C^G and C^B modeled as Lambertian and specular reflection. Derivatives of the invariants are incorporated in the Harris second moment matrix. It uses fixed scales for matching of images under varying lighting. No experiments under varying scales are done.

$$C^i = m_b(\mathbf{l},\mathbf{s})L^i S^i + m_s(\mathbf{l},\mathbf{s},\mathbf{v})L^i \tag{3.24}$$

for $i = \mathcal{R}$, $\mathcal{G}$ and $\mathcal{B}$. m_b and m_s express the geometric dependencies of these terms as a function of the light direction **l**, the surface normal **s** and the viewing direction **v**. S^i is the reflectance response and L^i is the illumination factor. Performing the logarithmic transformation $l^i = ln(1 + C^i)$ and convolving $(l^R - l^G)$ and $(l^B - l^G)$ with the derivatives of the Gaussian, M can be computed as:

$$M = \left\{ G_\sigma \otimes \begin{bmatrix} (l^i - l^G)_x^2 & (l^i - l^G)_x(l^i - l^G)_y \\ (l^i - l^G)_x(l^i - l_y^G) & (l^i - l^G)_y^2 \end{bmatrix} \right\} (\mathbf{x}). \tag{3.25}$$

This corner measurement relies on chrominance and focuses on real color edges which differ from specularity or illumination changes. In [Faille, 2005] the approach is evaluated with the shadow-shading invariant HSI approach that is used for the color invariant points described below.

[van de Weijer and Gevers, 2005] extend the color Harris approach to arbitrary color spaces and suggest two approaches: They develop a photometric quasi-invariant HSI color space providing a corner detector with better noise and stability characteristics compared to existing photometric invariants and a color boosting hypothesis for defining salient colors providing a visual saliency based detector. These two approaches provide a robust corner estimation under varying lighting and clutter for the quasi-invariant color space and a saliency measure based on the occurrence of colors in large datasets. In the following, these approaches are extended and a mathematically sound scale selection for color based interest point detection is built.

3.2.2 Scale Invariant Color Points

As suggested in [van de Weijer and Gevers, 2005], the second moment matrix can be computed using different color models. The first step is to determine the gradients of each component of the $\mathcal{RGB}$ color system. This is done using a convolution with a differentiation kernel of size σ. The gradients are then transformed into the desired color system. By multiplication and summation of the transformed gradients, all components of the second moment matrix are computed. Similar to Eq. 3.7, the results are scale normalized by $\sigma^2 G_{l\sigma}$ (see Sec. 2.2).

The symmetric second moment matrix M is a structure tensor describing the gradient distribution of the one channel image I in the local neighborhood of a point position $\mathbf{x}$:

$$M = \left\{ \sigma^2 G_{l\sigma} \otimes \begin{bmatrix} L^2_{x,\sigma} & L_{xy,\sigma} \\ L_{xy,\sigma} & L^2_{y,\sigma} \end{bmatrix} \right\} (\mathbf{x}), \tag{3.26}$$

In symbolic form, an arbitrary color space C is used with its n components $[c_1, \ldots, c_n]^T$. The elements for M are calculated more generally as

$$\begin{aligned} L^2_{x,\sigma} &= \sum_{i=1}^{n} c^2_{i,x,\sigma}, \\ L_{xy,\sigma} &= \sum_{i=1}^{n} c_{i,x,\sigma} c_{i,y,\sigma}, \\ L^2_{y,\sigma} &= \sum_{i=1}^{n} c^2_{i,y,\sigma} \end{aligned} \tag{3.27}$$

where $c_{i,x,\sigma}$ and $c_{i,y,\sigma}$ denote the components of the transformed color channel gradients at scale σ, and where the subscript x and y indicates the direction of the gradient. Similar to the luminance based Harris corner detector, the corner measurement $\mathfrak{C}$ is found based on the eigenvalues of M by (compare Eq. 2.22)

$$\mathfrak{C}_{\mathbf{x},\sigma} = \det(M_{\mathbf{x},\sigma}) - \alpha \cdot \text{trace}^2(M_{\mathbf{x},\sigma}). \tag{3.28}$$

The constant α indicates the slope of the *zero line*, i.e. the border between corner and edge.

With this definition, the corner measure can be calculated in any color space. This allows to estimate stable locations that are robust to noise, scale changes up to $\sqrt{2}$, translation and

rotation under arbitrary color spaces. For many computer vision tasks, it is crucial to provide a scale invariant feature as the size of objects and images can change drastically. Therefore a principled approach for saliency based estimation of the characteristic scale in arbitrary color spaces for local features is proposed.

Color Scale Decision

Using the elements of the structure tensor for arbitrary color spaces in Eq. 3.27, it is straightforward to extend the Color Harris to the Harris-Laplacian [Vigo et al., 2010] by applying Harris Corners and the LoG on different scales. Characteristic scale is then found when both functions are maximum.

Contrary to this extension, this thesis proposes a way to incorporate a global saliency measure in the process of scale selection while incorporating arbitrary color spaces. A comparison of these two methods is given in Section 4.2.2 showing the advantages of the proposed approach.

The *Principal Component Analysis* (PCA) aims for a set of the most representative projection vectors such that the projected samples retain the most information about the original samples [Delaca et al., 2005]. Having an s-dimensional sample space and aiming for a t-dimensional subspace for the m samples we define the scatter matrix S_t as

$$S_t = \sum_{i=1}^{m} (x_i - \overline{x})(x_i - \overline{x})^T \tag{3.29}$$

where m ist the number of samples, x_i a sample with the index i and $\overline{x}$ the sample mean.

As a result of the PCA, typically it is aimed for a lower dimensional subspace with the projected samples. Obviously, the distances between the samples themselves change depending on the scatter in relation to the projection vectors. In the scale decision these distances are used, to provide a saliency examination on a global scope (the whole image) for a local scale decision (the detected image patch).

The PCA analysis does not have a direct perceptual counterpart in the real-world, but tends to adapt itself to the color distribution of the image in a representative and robust way.

For a dimensional reduction, there are two other well known alternatives:

The Linear Discriminant Analysis (LDA) uses class membership information of the samples and tries to minimize the within-class scatter to find a linear combination of features which characterize classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. [Martinez and Kak, 2001] show that PCA provides a better representation when the training set is small and is less sensitive to a great variety within the training sets.

The *Independent Component Analysis* (ICA) separates a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian original samples. It estimates second and high-order statistics of the samples. The original samples are projected onto basis vectors which are as statistically independent as possible. There are several ways to estimate these vectors iteratively, details are given in [Bartlett et al., 2002].

However, the PCA is widely used for robust representations of visual information [Murase and Nayar, 1994] as it is the optimal linear transformation for keeping the subspace that has largest variance. The transformed color information of a pixel is considered as a sample in the processed image. This results in an inherently stable representation with respect to noise and occlusions [Wildenauer et al., 2002] but independent of its spatial properties. The principal components are of interest because they effectively summarize the dominant modes of variation on the different axes of the color spaces used. Colors which appear less often in an image are considered more salient (a use of the "saliency implies rarity" principle from [Kadir and Brady, 2001]). 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 prioritizing rare colors and not losing information on similar colors is achieved.

The proposed scale selection is carried out as follows. The input image is transformed to the same color space as used for the extraction of the Harris energy. The principal components of this transformed image I_C are computed. Its n channels of the m sample points $\mathbf{c}_j$ are reduced to a single channel $\widehat{I}$

$$\widehat{I}(\mathbf{x}) = \nu_\lambda I_C^T \tag{3.30}$$

by taking the dot product of the color information I_C and the corresponding principal eigenvector ν_λ of the scatter matrix

$$S_t = \sum_{j=1}^{m} (\mathbf{c}_j - \overline{\mathbf{c}})(\mathbf{c}_j - \overline{\mathbf{c}})^T \tag{3.31}$$

where $\overline{\mathbf{c}}$ denotes the sample mean, m is the number of pixels in the image and one sample $\mathbf{c}_j$ is a color vector of the pixel in I_C with the index j.

The Laplacian of Gaussian function is adapted to Λ to determine the characteristic scale [Mikolajczyk and Schmid, 2001]. Λ on position $\mathbf{x}$ of scale σ is defined by

$$\Lambda_{\mathbf{x},\sigma} = \left[\left(\frac{\partial^2 \widehat{I}}{\partial x^2} + \frac{\partial^2 \widehat{I}}{\partial y^2} \right) \otimes G_\sigma \otimes \Gamma_\sigma \right] (\mathbf{x}) \tag{3.32}$$

where Γ_σ is the circularly symmetric raised cosine kernel, which is defined for each location x_e, y_e as

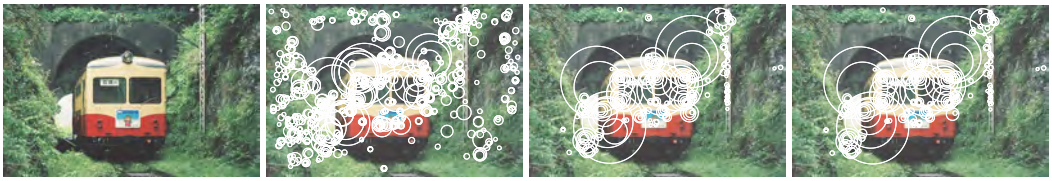
$$\Gamma_\sigma = \frac{1 + (cos(\frac{\pi}{\sigma}\sqrt{x_e^2 + y_e^2})}{3}. \tag{3.33}$$

A convolution with this kernel gives smoother borders than the Gaussian kernel G for scale decision and decreases noise effectively [Kenney et al., 2005]. For computational efficiency, Λ can be approximated by the sum of the independently computed values L_x^2 and L_y^2 of $\widehat{I}$:

$$\Lambda_{\mathbf{x},\sigma} = \left\{ \left[\sigma^2 | L_x^2(\mathbf{x},\sigma) + L_y^2(\mathbf{x},\sigma) | \right] \otimes \Gamma_\sigma \right\} (\mathbf{x}). \tag{3.34}$$

Corner locations shift as the scale changes. The smaller the scale change from one iteration to the next is, the more precise the location is estimated. As the Harris detector is robust to

(a) VOC 2007 nr. 395 (b) Harris Laplacian (c) HSI points (d) color boosted points



(e) VOC 2007 nr. 540 (f) Harris Laplacian (g) HSI points (h) color boosted points

Figure 3.13: VOC 2007 images and the interest points based on luminance, HSI and color boosted OCS information combined with the proposed scale selection.

scale changes up to the factor $t = \sqrt{2}$, this factor showed to be precise enough while giving good results [Lindeberg, 1998; Lowe, 1999]. The scale space of the Harris function is built calculating the Harris energy under varying σ. The number of different scales examined is of crucial importance 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.

Using scale levels $l_S = 1, 2, \ldots$ with a factor t from 1.2 to $\sqrt{2}$, the Harris energy is calculated at scales $t^s\sigma$ A *potential* characteristic scale of a possible region is found if both the Harris Energy and the Laplacian of Gaussian are extrema

$$\nabla\Lambda_{\mathbf{x},\sigma} = \nabla\mathfrak{C}_{H,\mathbf{x},\sigma} = 0. \tag{3.35}$$

With this non-maxima suppression, the locations with their corresponding scales are found. The following decision criterion gives the 3 times largest scale at the location, providing basis for *stronger* local description [Lazebnik et al., 2006]:

$$\widehat{R}_{\mathbf{x}} = \begin{pmatrix} \max\left[\widehat{E}_{\mathbf{x}}\right] \\ 3t^{\arg\max(\widehat{\Lambda}_{\mathbf{x},l})}\sigma \end{pmatrix}, \tag{3.36}$$

where, having chosen constants σ and t, the functions $\widehat{E}_{\mathbf{x}}$ give the scales of local maxima of $\mathfrak{C}_{H,\mathbf{x}}$ and $\widehat{\Lambda}_{\mathbf{x}}$ the scales of local maxima of $\Lambda_{\mathbf{x},\sigma}$. The local maxima of both functions are combined efficiently per scale level and location. The resulting locations are thus selected choosing the largest scale per location.

The resulting vector function $\widehat{R}_{\mathbf{x}}$ defines all candidates for interest points and the corresponding region size. The multi-channel Harris energy is the saliency property which is used for the decision of the retained interest points. Note that the characteristic scale is estimated independently of the scale in which the highest Harris energy occurs. The color boosting is related

to the proposed color scale selection. The underlying difference is that saliency values are learnt off-line for color boosting and the saliency is decided on-line per image in the scale selection.

In the experimental section, this approach is used with the HSI and the *color boosted OCS* described in Section 2.1. The resulting points are further denoted as *HSI points* and *color boosted points*. Example regions can be seen in Fig. 3.13. The luminance based Harris Laplacian detects many background features solely based on shadowing effects. The proposed features are able to reduce these number using two strategies: The HSI points use quasi-invariant color information in the shadow-shading direction. The color boosted points use color statistics to focus on rare colors. The focus toward salient regions is amplified by the proposed scale selection: In the PCA representation rare colors in the images have a higher probability to get selected as a feature.

3.2.3 Blob Detectors

[Unnikrishnan and Hebert, 2006] extract scale and illumination invariant blobs through color. They adopt the diagonal illumination model [Finlayson and Drew, 1993]. The model is based on the assumption that a diagonal transform is sufficient for the color representation. The prerequisite is a previously computed and a particular illumination effect tuned transformation matrix, which is first applied to the sensor outputs. They modify the LoG detector (compare Eq. 3.8) to

$$h_{diag} = \sigma^3 \prod_{I=\mathcal{R},\mathcal{G},\mathcal{B}} G_{xy,\sigma} \otimes c_i \tag{3.37}$$

where c_i denotes the color information elements of a pixel. The resulting regions are found by non-maxima suppression in the scale space pyramid. The simple multiplication of color channels is however debatable as the result becomes zero as soon there is no structure in one of the channels.

A more mathematically sound way is proposed in [Vigo et al., 2010]. They build on the same basis as the proposed approach in this thesis and use the elements given in Eq. 3.27. The **Color LoG** is then extracted applying Eq. 3.8. The main difference is that the elements of the color structure tensor are built on basis of the transformed derivative of the color information and summed up per direction. No cancelation or numerical problems are encountered for very small values.

For the **Color Hessian** detector, they use the determinant of the color structure tensor (compare Eq. 3.13) and approximate the determinant as given in Eq. 3.28. These two approaches preserve the chromatic variation of image structure which is not done when the color information is mapped to luminance.

[Donoser et al., 2006] propose a way to estimate MSER on color information. The image is transformed into $CIELUV$ Color Space. They aim for an ordering relationship between the colors in the three dimensional color space. They model the three dimensional feature space as a multivariate Gaussian distribution. It would be possible to use *Gaussian Mixture Models* (GMM) but they choose to use a single Gaussian distribution. As such it is defined by 3×1 mean vectors $\mu_{1..3}$ and the covariance matrix Σ. The matrix is symmetric and semi-definite and can therefore be defined with 6 values. The whole distribution is thus defined by 9 values. Two Gaussian distributions are compared by the Bhattacharyya distance [Bhattacharyya, 1943] using their covariance matrices $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$

$$\beta = \frac{1}{2} \ln \frac{|\frac{\mathbf{\Sigma}_1+\mathbf{\Sigma}_2}{2}|}{\sqrt{|\mathbf{\Sigma}_1||\mathbf{\Sigma}_2|}} + \frac{1}{8}(\mu_2 - \mu_1)^t \left[\frac{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2}{2}\right]^{-1} (\mu_2 - \mu_1) \tag{3.38}$$

The distance measures how *expensive* it is to discriminate between two probability distributions. A small distance is very expensive to divide, therefore they are very similar, bigger distances are more different.

Every color pixel is now processed in two subsequent steps: (1) the distribution is fitted for every pixel within an initialized region. (2) All pixels are ordered by their Bhattacharyya distance from the Gaussian of one region to the Gaussian that is fitted to the pixels within a kernel size of σ around each pixel. Bigger σ gives a smoother distance distribution. In case $\sigma = 1$ the Mahalanobis distance β_M from one feature point $\mathbf{x}$ to a distribution is used. Then, no neighborhood or color variance is incorporated in the distance.

$$\beta_M^2 = (\mathbf{x} - \mu)^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \tag{3.39}$$

These distances are used as pixel weights for building the connected graph for the subsequent MSER estimation as described in Section 3.1.2. The MSER are detected by the analysis of the connected graph of the image. In contrast to typical intensity values the weights are a continuous signal: The signal is normalized and evenly divided into n samples to get a discrete signal. For an improved sampling algorithm, e.g. a Lloyd-Max algorithm [Lloyd, 1982] could be used. For the decision on a stable region, Eq. 3.14 is used substituting the intensity values with the sampled weights.

3.3 Summary

This chapter gives the state-of-the-art of local feature detection in images. The detection approaches are divided in luminance based approaches and color based, or multi-dimensional approaches. Typically, the color based methods are extensions from the basic ideas of the single channel approaches. The approaches are arranged into three types of feature detectors: Corner, blob and symmetry based detectors.

Corner detectors give very stable locations under geometric and lighting changes. The local structure tensor gives an efficient way to detect corners in both single as multi-dimensional data on different scales.

Blob detectors rely either on the scale space theory and differential methods (DoH, LoG) or on segmentation algorithms (MSER). They can be used for the scale selection of detected corners. In this thesis, a new way of selecting robust scales in multi-dimensional data is proposed.

Symmetry based interest points detect locations within local symmetry. A new interest scale and rotation invariant point detector based on the GVF is proposed. It provides a dense and well distributed feature representation.

These approaches build the state-of-the-art of feature detection on images. The most successful image matching approaches rely on one or more of these detectors.

CHAPTER 4

Evaluation of Interest Points for Images

As the first step in local feature based image matching applications, local detections affect all the subsequent operations in an application. As images of different objects tend to be very different in their appearance, stable features which do not change the image's representation because of image alterations are desired. Throughout this experimental section, three different experiments are carried out. In every experiment, only the locations of the features are changed. In this way it is possible to isolate the impact of local feature locations in image matching. Fig. 4.1 shows an overview of the experimental set-up.

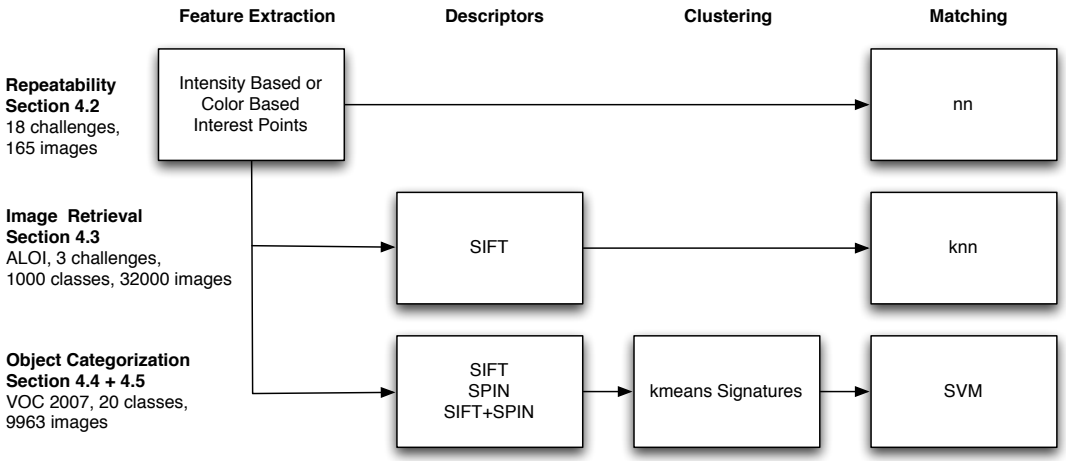


Figure 4.1: The main stages of image matching and the structure of the image experiments.

In Section 4.2 the robustness of the three proposed interest points is evaluated on a publicly available and well-known data-set. It provides a rotational and scale invariant interest point detector which gives a dense and well distributed representation of an image.

Image retrieval experiments are carried out in Section 4.3. It is shown that fewer but more informative interest points can lead to a gain in retrieval precision. A more robust color based local interest point detection increases matching performance significantly for simple classification techniques using fewer features as its luminance based counterparts. The image retrieval scenario shows an increased ability of color points to retrieve objects under lighting changes. The same local description and classification methods are used for all experiments while changing the interest points only. To show that the properties of sparse color interest points are highly desirable and crucial for a successful matching of local features in a computer vision task, an object retrieval scenario for 1000 objects, having 7 images per object under different lighting conditions is chosen. The most successful interest point detector, the Harris Laplacian detector, is evaluated and compared to the sparse color interest points. In this setup, the more distinct color interest points directly lead to a more precise retrieval performance.

Going into "real world" experiments, color interest points in a large public object categorization challenge are evaluated. The dataset provides a great inter-class variation and the images are not all colorful. State-of-the-art classification frameworks use a bags-of-words and multi-level classification schemes for this task. For more sophisticated and state-of-the-art classification schemes, it is possible to reduce the number of features and maintain state-of-the-art-results. It is shown that even with these sophisticated techniques, the first step of extracting salient interest points is crucial. The color interest points do not increase the performance of the classification system compared to other detectors. Nevertheless, due to their higher repeatability, up to 50% of the most salient interest points are enough to maintain the classification performance. Additionally it is shown that more meaningful points can improve the discrimination power between objects, even if the successive description phase is illumination based. For the object categorization experiment in Section 4.4, a state of the art framework is used, in which the PASCAL VOC challenge is run several times with different interest points in the first stage of the framework. For this large scale experiment, comparable results are obtained to the state-of-the-art interest point extraction algorithm but using fewer features.

Similar results are provided by an international benchmark on the same data-set, where a more sparse feature representation based on the proposed approach outperforms all other competitors in 4 out of 20 classes using only a fraction of their number of features. The benchmark results are given in Section 4.5.

In the experiments, the salient points based on the HSI quasi invariant coordinates are denoted as *light invariant points* and the salient points based on the color boosted OCS coordinates as the *color boosted points*. Both of them are referred to as *color points*. In Section 4.2, RGB points are added for comparison. As the state of the art reference, the *Harris Laplacian* as described in Section 3.1.1, and the latest implementation[1] evaluated in [Mikolajczyk et al., 2005b] is used.

All experiments are carried out with the same parameters $\sigma = 1$, $l = 10$, $t = \sqrt{2}$ as suggested in the literature and defined in Section 3.1.1. In case a subset of the provided points

[1] `http://www.robots.ox.ac.uk/~vgg/research/affine/`

is chosen, the points are ordered by their Harris energy (see Eq. 3.28). It is shown that the color based Harris energy gives a more reliable decision criteria for reducing features than the illumination based counterpart does. In the following, the data-sets used in the experiments are described in detail.

4.1 Data-sets

In this section, the three data-sets used in the experiments are described in detail. Starting with the most common data-set for evaluation of features in Section 4.1.1, the challenges for the local features are described. The image retrieval experiments are carried out on the Amsterdam Library of Object Images (ALOI) data-set. The data-set is described in Section 4.1.2. In Section 4.1.3 the VOC PASCAL 2007 data-set is presented, on which the experiments and the benchmark on object categorization are performed.

4.1.1 Robustness Data-set

[Mikolajczyk and Schmid, 2004] suggest a test for the quality of local features. They measure the repeatability of local features under different image transformations, denoted as *challenges*. These tests consist of a set of images, referred to as *test-set*, where one acts as the reference image and the other images show the same scene under predefined changes like blur, rotation, zoom, viewpoint change, JPEG compression or lighting changes. The data-set is available online[2]. An overview of the data-set is given in Tbl. 4.1.

The repeatability rate is defined as the ratio between the number of detected correspondences and the number of regions that occur in the area common to both images. Feature detectors tend to have higher repeatability rates when they produce a richer description. This is not true for certain extreme cases, but assuming a reasonable distribution of features of reasonable size, the chance of establishing a correspondence with the nearest neighbor is higher when the points are densely distributed. On the other hand, it is not always true that a lower number of regions automatically yields a lower repeatability rate. For a typical number and distribution of interest points, fewer points in two images are less likely to correspond. Therefore, for more regions the repeatability rate tends to rise.

In geometry, a homography is an invertible transformation from the real projective plane to the projective plane that maps straight lines to straight lines. Having a correct homography H between two images I_1 and I_2, a feature γ_1 at $\mathbf{x}$ in I_1 is repeated and thus robust if its projection $\mathbf{x}'$ in I_2 is within 40% of overlap error of the nearest other detected feature γ_2 in I_2. Interpreting the features γ as ellipses in the images, the overlap error ϵ_o is defined as

$$\epsilon_o = 1 - \frac{\gamma_1 \cap (H^{-1}\gamma_2 H)}{\gamma_1 \cup (H^{-1}\gamma_2 H)} \tag{4.1}$$

Between the reference images and all of the test images, the correct homography matrix is given [Mikolajczyk et al., 2005b]. The matrices between the reference image and the other

[2] http://lear.inrialpes.fr/people/mikolajczyk/Database

Challenge	Test-set	Nr. of images	Resolution	Color
Rotation	marseil	18	842×842	no
	monet	18	842×842	no
	new york	35	512×512	no
	van gogh	17	512×348	no
Zoom	belledonnes	4	760×555	no
	asterix	17	512×348	no
	crolles	10	760×555	no
	bip	9	768×574	no
	van gogh	17	512×348	no
	laptop	21	768×574	no
Rotation & Zoom	Boat	10	850×680	no
	east park	11	850×680	no
	east south	10	850×680	no
	inria	11	850×680	no
	inria model	11	850×680	no
	resid	11	850×680	no
	ubc	13	850×680	yes
	laptop	13	760×574	no
	ensimag	11	850×680	no
	bark	6	765×512	yes
	boat (Fig. 4.3)	6	800×640	no
Viewpoint	downtown	14	800×640	yes
	graffiti (Fig. 4.2)	6	800×640	yes
	graffiti2	9	800×640	yes
	graffiti4	15	800×640	yes
	graffiti5	9	800×640	yes
	graffiti6	9	800×640	yes
	graffiti7	5	800×640	yes
	ubc	13	800×640	yes
	inria	9	800×640	no
	bricks	6	1000×700	yes
Blur	bikes (Fig. 4.4)	6	1000×700	yes
	trees	6	1000×700	yes
Light	cars (Fig. 4.5)	6	921×614	yes
	fruits	8	512×512	yes
	graph	8	512×512	no
	mosaic	8	512×512	no
	movi	8	512×512	yes
	nuts	8	512×512	yes
	toy (Fig. 4.6)	20	512×512	yes
JPEG	ubc (Fig. 4.7)	6	800×640	yes

Table 4.1: Overview test-sets of the robustness data-set and it challenges.

images in a dataset are computed in a two step procedure. For the first step, a small number of point correspondences are selected manually between the reference and the other images. This can be done in a publicly available Java applet[3]. These correspondences are used to compute a rough approximation of the homography between the images, and the other image is warped by this homography so that it is aligned with the reference image.

For the second step, a standard small-baseline robust homography estimation algorithm is applied to the aligned image. With this accurate residual homography between the reference and aligned image (using hundreds of automatically detected and matched interest points), the

[3] `http://lear.inrialpes.fr/people/mikolajczyk/StereoVision/index.html`

composition of the two homographies (approximate and residual) gives an accurate homography between the reference and the other images.

In the challenges, the data-set provides similar challenges on different scenes. Therefore, it is possible to evaluate the impact of image alterations separately from the effect of scene changes. There are two main scene types: One scene type contains many homogeneous regions with distinctive and well defined edge boundaries (e.g. *graffiti* (Fig. 4.2), *buildings*), and the other contains highly repeated natural textures of different forms (e.g. *bark*).

Each of the test sequences contains at least 6 images with a gradual geometric or photometric transformation. All the images are of medium resolution of approximately 800 x 640 pixels. The challenges are produced in a "natural way" as the image alterations are produced by changing the camera position, zoom and focal length and not by artificially processing the images on a computer. An exception is the challenge of JPEG compression, of course. This sequence is generated varying the JPEG quality parameter from 40% to 2%.

In the challenges of viewpoint change test (Fig. 4.2), the camera position varies from a front view to one with significant foreshortening at approximately 60 degrees to the camera for planar scenes. The scale change (Fig. 4.3) and blur (Fig. 4.4) sequences are acquired by changing the camera zoom and focus. The scale changes towards a factor of four. The lighting changes are introduced by varying the camera aperture (see Fig. 4.5) or by changing the lighting direction.

The images are either of almost planar scenes or the camera position is fixed during acquisition, so that in all cases the images' relation is known and this mapping is used to determine ground truth matches for the feature evaluation.

In the following, some well known test-sets of the data-set are described in more detail.

The test-set *graffiti* (Fig. 4.2) is a broadly known test-set in the community and the highly textured wall has become almost an embodiment for feature evaluation itself. It shows a wall from different viewpoints until some of the background is seen as well.

The test-set *boat* (Fig. 4.3) consists of 6 images of highly textured, detailed view of a boat in the water with a low contrast background. It challenges zoom and rotation at the same time (similar to the FeEval video data-set in Section 6.1.2) and is often used in literature to verify the scale invariance of features.

In the test-set *bikes*, shown in Fig. 4.4, a natural scene with objects containing very small structures are successively blurred. Measuring the repeatability under increasing blur shows the ability to provide stable locations under diminishing visual information. Obviously, corner detectors suffer more from this challenge than other detectors. Detectors which do not rely on local gradients of fixed scale tend to be more stable than others. Therefore, it is also a challenge for the stability of the scale estimation.

The test-set *cars* (Fig. 4.5) depicts 6 images of two cars in front of a building. As the images are taken naturally, the images differ very slightly in their viewpoint. Choosing the darkest one of the data-set, the scene is of very low contrast. Proceeding to the other test images, more and more information appears. The challenge measures the stability against such changes.

The test-set *toy* (Fig. 4.6) provides 20 images from the same toy scenery. The global illumination does not change much, but the position of the direct lighting is successively changed over the images. The interest point detectors encounter different shadowing effects moving over the scenery.

(a) (b) (c)

Figure 4.2: Test-set *graffiti* depicts a painted wall under heavy viewpoint changes.

(a) (b) (c)

Figure 4.3: Test-set *boat* changes the viewpoint and the zoom level while rotating the scene.

The test-set *ubc* (Fig. 4.7) applies strong JPEG compression to a natural scene with color transitions and small textures. New colors are introduced and finally the most salient edge in the image between sky and roof is completely changed by artifacts for the last test image.

The JPEG format is broadly used in many image databases and the most common format used in the internet. Its lossy compression introduces artifacts to the images. To be stable against this noise improves the performance of an interest detector in many computer vision challenges, keeping in mind that the noise appears predominantly in locations of high frequency and structure.

Krystian Mikolajczyk stated the following drawbacks of his framework, at his opening talk to the CVPR 2009 feature benchmark [Mikolajczyk et al., 2009]:

(a) (b) (c)

Figure 4.4: Test-set *bikes* with different bikes getting more and more blurred.

(a) (b) (c)

Figure 4.5: Test-set *cars* provides a natural scene at different daytimes.

(a) (b) (c)

Figure 4.6: 6 out of 20 images of the test-set *toy*. It provides a natural scene under different lighting directions. Main challenge is the stability against shadowing effects.

(a) (b) (c)

Figure 4.7: Test-set *ubc* adds more and more JPEG compression artifacts to a natural scene.

- There is a strong bias towards dense responses. When a larger number of features is extracted, a better performance is achieved automatically. In case an approach returns all windows it will be perfect for most challenges
- Large regions have an advantage against smaller ones.
- Additionally, there is a bias towards (current) detector-friendly problems. As the scenes are selected for this purpose, standard detectors will perform well. This is a problem, as there are many image categories for which standard detectors are useless.
- Moreover, increasing generality (applicability) cannot be demonstrated on this dataset.
- For parameter tuning of certain approaches, there is a limited number of scenes so that the risk of over-fitting occurs.

Nevertheless, the repeatability measure is still used as the evaluation method for detectors. Predefined scene changes on visual data are the tool of choice in testing the robustness of visual features.

4.1.2 The Amsterdam Library of Object Images

The Amsterdam Library of Object Images (ALOI)[4] [Geusebroek et al., 2005] provides images of 1000 objects under supervised, predefined conditions on a dark background yielding a total of 110 250 images for the collection. Example images are shown in Fig. 4.8. Having a large scale data-set of objects without background clutter, it enables experiments precisely evaluating the feature representation on colorful objects only. With these images better conclusions can be drawn than on images under natural circumstances where a significant part of the extracted features are describing background information. The data-set applies transformations on the objects

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

Figure 4.8: Example images from the ALOI data-set.

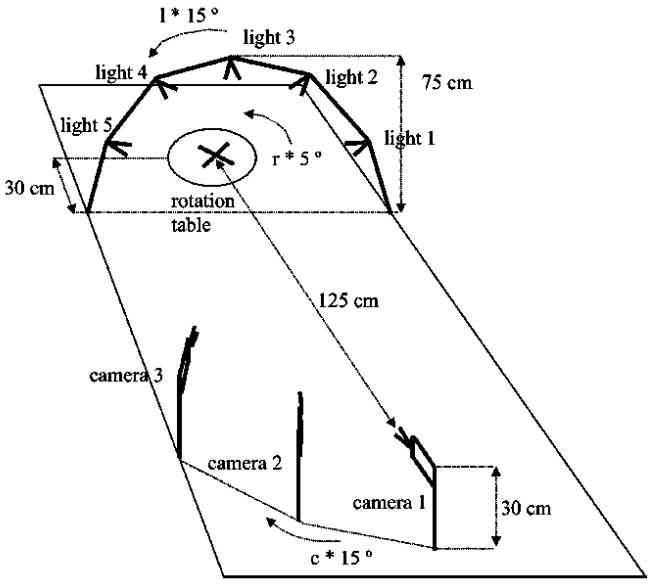


Figure 4.9: Set-up for capturing object images (from: [Geusebroek et al., 2005])

in a precise and well defined way. For the transformation, there are viewpoint transformations and varying lighting conditions. It was an inspiration for the FeEval data-set proposed in this work.

As shown in Figure 4.9, 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, well defined 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

a frame grabber, the Matrox Electronic Systems Ltd. CORONA II PCI frame grabber is used. All images are taken from a 124.5 cm distance, with 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 quasi hemispherical illumination, although restricted to a more narrow illumination sector than a true hemisphere. Illumination direction and illumination power is changed over the object. With these lighting changes, illumination based approaches suffer from instability and many ambiguous descriptions of shadowing effects.

- **Illumination Color** provides 12 configurations, all taken with all 5 lights turned on. Color temperature is successively increased from 2175K to 3075K.

- **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.

4.1.3 PASCAL VOC 2007 data-set

The color salient points are evaluated on the data-set from the PASCAL visual object classes (VOC) 2007 challenge[6]. The PASCAL VOC challenge is an annual benchmark of recognition challenges. For the first challenge in 2005, the data-set consisted of 4 classes of 3787 images and is growing every year in both number of images and number of classes. The data-set of 2007 is the latest one which contains a public annotation of all the images. Since then, no public ground truth of the test data is released.

This dataset consists of 9963 images, where 5011 images form the annotated training set (trainval). The test set (test) contains 4952 images which are used to evaluate the performance of the framework. The data has been split into approximately 50% for training/validation (trainval) and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and test sets. The number of objects in one image is not fixed, the whole dataset contains 12 608 objects. Twenty classes of object are annotated with ground truth. Example images are shown in Fig. 4.10. The classes are denoted as

- Person: person

[5] `http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html`
[6] `http://www.pascal-network.org/challenges/VOC/voc2007/`

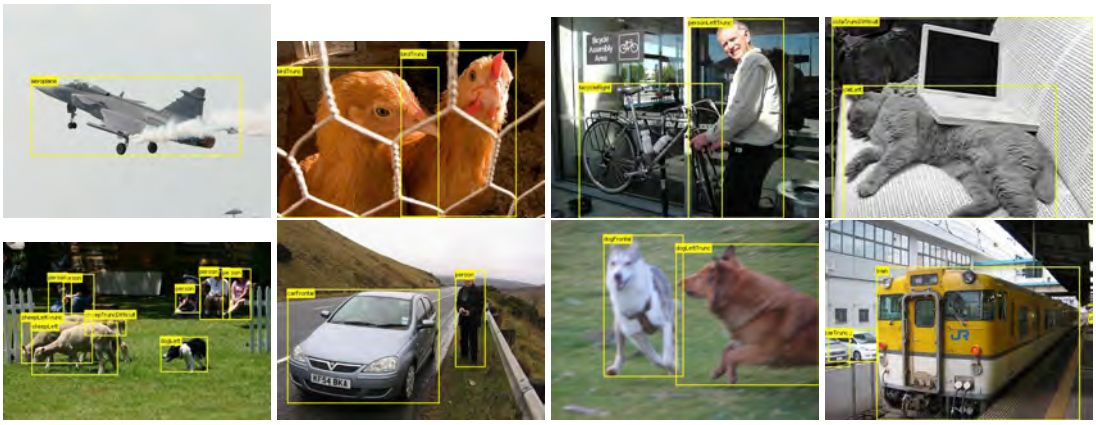


Figure 4.10: Annotated sample images from the VOC 2007 dataset.

- Animal: bird, cat, cow, dog, horse, sheep
- Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
- Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

The challenge is that for each image the presence or absence of the classes is asked. The assumption is that there is at least one but possibly more objects in each test image.

There are two competitions in this task: One where the trainval data-set is allowed for training only, the other one, where every training data is allowed, except the test data-set. The *detection* task aims for each of the classes the bounding boxes of each object of that class in a test image (if any) are predicted. For all following experiments, the classification task is evaluated with the provided trainval data-set for learning.

Tbl. 4.2 shows statistics of the classes in the data-set. *Img* denotes the number of images belonging to that class, *obj* denotes how many of those objects appear in the dataset.

An important fact is that the classes are not distributed evenly over the dataset. There are many more persons in the images than there are for example TV monitors. Nevertheless, their occurrence is correlated with persons, sofas and potted plants. On the other hand, there are classes which are not likely to show up on the same picture: Horses and TV monitors, motorbikes and sheep, aeroplanes and bottles are not likely to appear together in one classification image.

Generally, the data-set is not particularly advantageous for color based approaches, because there are many black & white and artistic near black & white pictures. Several home made snapshots (especially of persons and pets) are blurred, or over or underexposed. Although all approaches suffer from these effects, the color based approaches are more sensitive.

	train		val		trainval		test	
	img	obj	img	obj	img	obj	img	obj
Aeroplane	112	151	126	155	238	306	204	285
Bicycle	116	176	127	177	243	353	239	337
Bird	180	243	150	243	330	486	282	459
Boat	81	140	100	150	181	290	172	263
Bottle	139	253	105	252	244	505	212	469
Bus	97	115	89	114	186	229	174	213
Car	376	625	337	625	713	1250	721	1201
Cat	163	186	174	190	337	376	322	358
Chair	224	400	221	398	445	798	417	756
Cow	69	136	72	123	141	259	127	244
Diningtable	97	103	103	112	200	215	190	206
Dog	203	253	218	257	421	510	418	489
Horse	139	182	148	180	287	362	274	348
Motorbike	120	167	125	172	245	339	222	325
Person	1025	2358	983	2332	2008	4690	2007	4528
Pottedplant	133	248	112	266	245	514	224	480
Sheep	48	130	48	127	96	257	97	242
Sofa	111	124	118	124	229	248	223	239
Train	127	145	134	152	261	297	259	282
tvmonitor	128	166	128	158	256	324	229	309
Total	**2501**	**6301**	**2510**	**6307**	**5011**	**12608**	**4952**	**12032**

Table 4.2: Object distribution in the VOC Pascal 2007 data-set per class (from: [Everingham et al., 2007])

4.2 Robustness

In this section, the repeatability experiments as defined in Section 4.1.1 are carried out. In the following, GVFpoints are evaluated in comparison to state-of-the-art luminance based detectors in detail. In Section 4.2.2, it is shown that the use of color invariance and color boosting increases the stability of corner detections. Using a stable scale selection, state-of-the-art detectors can be outperformed in robustness to various challenges.

4.2.1 GVFpoints

In this section, a robustness evaluation of the GVFpoints is given. It is shown that they outperform current approaches for invariant interest point locations in several important tasks. Using the GVF for the extraction of interest points provides comparable or better results. The data-set is described in Section 4.1.1. The following test-sets are evaluated: *Graffiti* and *bricks* are used to evaluate viewpoint changes, the test-set *boat* to challenge zoom and rotation, the test-set

cars and *toy* to analyze changing in lighting condition, *bikes* evaluating increasing blur and *ubc* testing increasing JPEG compression.

For the challenges, the repeatability graphs and numerical results are given. The detailed numerical results consist of the means of the repeatability rate, number of correspondent regions, area covered by the features, standard deviation of the area and the number of interest points in the image give. The graphs the repeatability per image.

The most successful approaches for detecting interest points based on luminance are evaluated. For the most stable and broadly used corner detectors, the Harris Laplacian is chosen for its excellent performance in [Mikolajczyk and Schmid, 2004]. For the broadly used blob detectors DoG and MSER is selected. As their approach is related to the proposed approach, symmetry based interest points are evaluated using the Loy points. Details of the approaches are given in Section 3.1.

The histograms in Fig. 4.11 provide a summary view of the ranks of the individual algorithms. Each of the 91 reference image / test image pairs are treated as a separate experiment. For each of these, the algorithms are ranked according to their repeatability from 1 to 5. In 57.1% of the cases the GVFpoints exhibited the best performance (rank=1), while in 80.2% they performed either best or second best (rank$\leq$2). Harris Laplacian and Loy's symmetry points show far lower performance, with Loy performing worst (rank=5) in 47% of cases. MSER and DoG display mixed results: While showing leading performance in same cases they perform badly in others, exhibiting an average performance overall.

GVFpoints show to be repeatable under geometric transformation (Fig. 4.12). Elongated structures like the ones found in the *graffiti* test-set (see Fig. 4.2 and example features in Fig. 3.9) are centered precisely. This works also for MSER, having very well defined blobs on the wall. Therefore, DoG performs also better than the Harris Laplacian as the corners are heavily transformed during the challenge. For Loy points, no repeatability is found for the last two test images. Note the high number of GVFpoints compared to the other approaches because of the elongated structure of the blobs, which increases the repeatability rate. The statistics in Fig. 4.12 show that the GVFpoints give more than three times more features than the DoG do. The features with the largest scales are provided by the Harris Laplacian and the GVFpoints. The Harris Laplacian features have the highest standard deviation of the area covered by features. This means that many *different* scales are selected on the image.

On small, often repeated structures like in test-set *bricks*, GVFpoints are able to estimate correspondences for over 75% of all locations, even after 60 degree of viewport change of test image number 6. As shown in Fig. 4.13, the dense representation of the GVFpoints remains significantly more stable than other approaches. Contrary to the experiment on the *graffiti* test-set, Harris Laplacian features are more sensitive to change in the small repeated structure than DoG are. Symmetry based Loy points perform similarly to Harris Laplacian. GVFpoints gives more than double the number of features than DoG do, but also give the biggest features on this dataset.

The experiment on the test-set *boat* (see Fig. 4.14) shows that GVFpoints exhibit higher

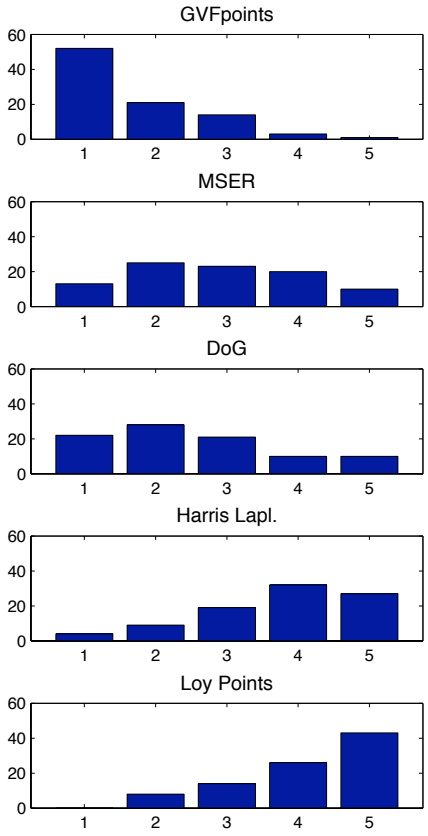
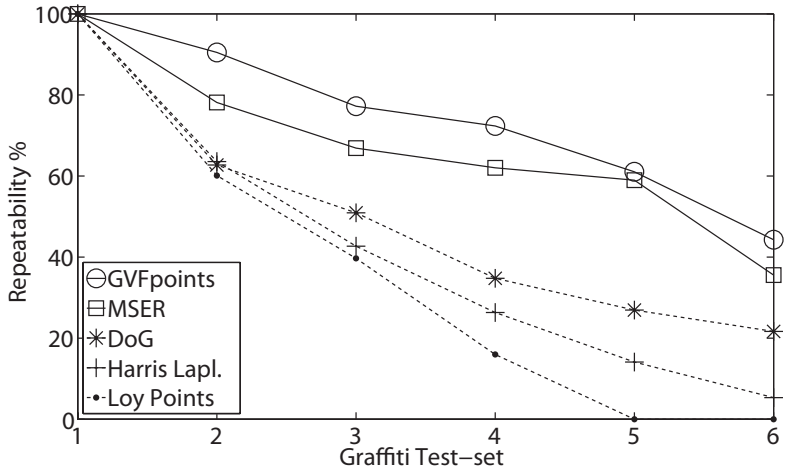


Figure 4.11: Histogram of the ranks of the compared algorithms. For each of the 91 test images the algorithms were sorted according to their performance, summing up the more detailed results presented in this section. Note that in 57.1% of the cases the GVFpoints exhibited the best performance (rank=1), while in 80.2% they performed either best or second best (rank$\leq$2).

repeatability at small details, being more invariant to rotational change than other approaches. The Harris Laplacian gives again the highest standard deviation of the area covered by feature providing on average the large scales. All approaches have a very similar loss in robustness from the reference image to the first test image. GVFpoints are significantly more robust.

As shown in Fig. 4.15 and Fig. 4.16, GVF based points are more stable against changing illumination than all other interest point detectors. Linear illumination change does not affect the GVF to the same degree as the other interest point detectors. However, this is only true until a certain degree of lighting change, as can be seen in the last image. In the last test image many new GVF minima are introduced and MSER provide slightly more stable points than GVF.

For changing lighting direction in Fig. 4.16, all interest point detectors have an immediate shift in their position due to different shadows. In contrast, the GVFpoints repeatability



	GVFp	MSER	DoG	HarLap	Loy
(rep)	74.2	66.9	49.5	42.0	36.0
corresp	7361.7	215.8	1984	1016.8	474.5
(area)	2496.2	1565.7	65.1	4490.5	684.4
std(area)	4273.5	3930.7	149.5	17926.1	1013.8
nr. pts	17349	533	5479	3116	1690

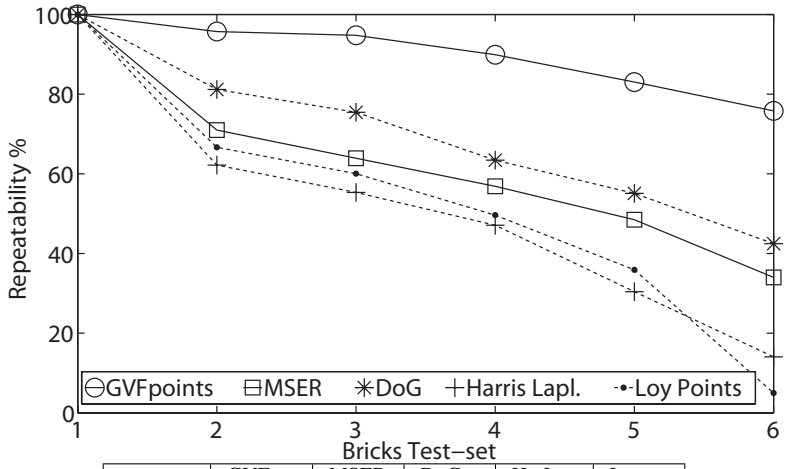
Figure 4.12: Repeatability experiment test-set *graffiti* challenging viewpoint transformation on images of a colorful wall.

decreases in comparatively small steps. DoG based interest points are more stable in the continuing change of lighting directions, outperforming the GVFpoints. Loy symmetry have almost the same mean repeatability rate as MSER in this context. Shading variant corner detectors are heavily dependent on shaded edges and perform therefore worse than the other approaches.

For heavy change of lighting, MSER provide slightly more stable locations than GVFpoints. Fig. 4.17 shows the GVFpoints are almost perfectly invariant to blur. This is reasonable as the GVF does not change its extrema under these changes whereas detectors based on e.g. edges are more dependent on high contrasts in the image. Almost 80% of GVFpoints stay stable over the whole data-set, while different approaches suffer increasingly from instability in these changes.

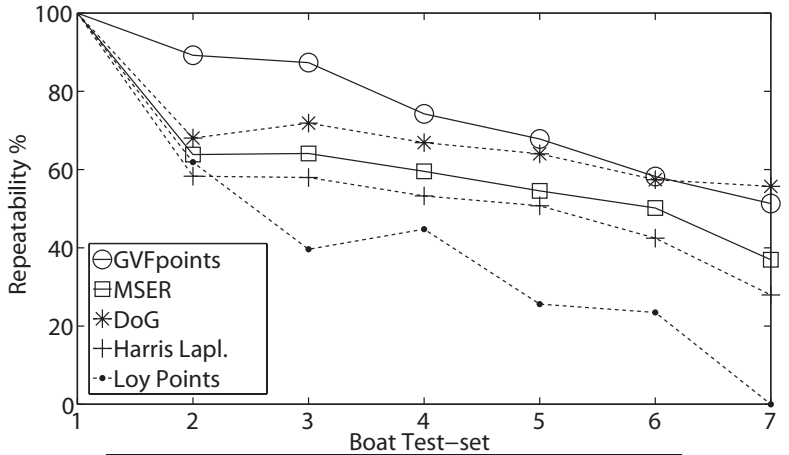
Local noise like the JPEG compression artifacts in test-set *ubc* are evaluated in Fig. 4.18. It is shown that the GVFpoints provide more stable locations to the point where the extrema are significantly shifted by the newly introduced structures. Surprisingly, MSER turn out to be very unstable to this kind of noise, whereas Loy points provide better results. Harris Laplacian points are obviously more stable and perform almost comparable to the DoG.

With these experiments, it is shown that the localization of features using the GVF can provide a dense and stable representation for image matching. Especially the robustness to JPEG artifacts, blur and rotation could make the features valuable e.g. in application in the field of mobile computing, where small and unsteady cameras suffer from these effects. Research question could if such a rich representation is necessary and feasible for matching application.



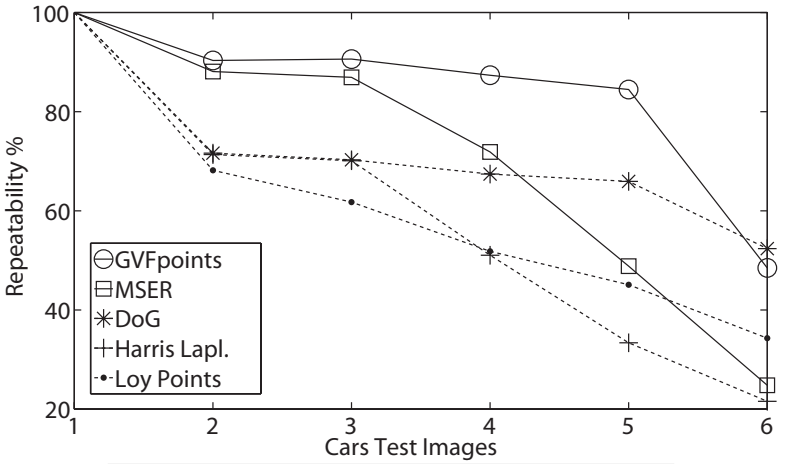
	GVFp	MSER	DoG	HarLap	Loy
(rep)	89.9	62.4	69.6	51.5	52.9
corresp	15828.8	1026	8270.8	3036.2	776.5
(area)	2481.8	507.8	52.8	1714.9	669.2
std(area)	4252.9	1459.3	65.0	2789.4	1007.6
nr. pts	28093	2363	13604	7599	1600

Figure 4.13: Repeatability experiment test-set *bricks* – viewpoint transformation on a highly structured plane. The scale invariant GVFpoints approach outperforms state of the art approaches.



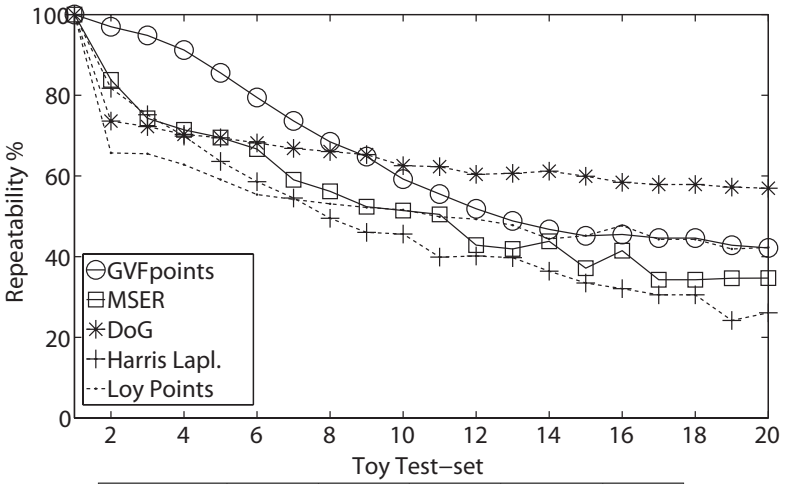
	GVFp	MSER	DoG	HarLap	Loy
(rep)	75.5	61.3	69.1	55.8	42.2
corresp	6990.6	497	4969.7	1628.4	502.3
(area)	2422.3	1083.8	51.6	3857.0	816.6
std(area)	4206.8	4320.7	33.9	11852.8	1077.8
nr. pts	17625	1524	13797	4961	1803

Figure 4.14: Repeatability experiment test-set *boat* – zoom and rotation of a boat with fine texture and a blurred background.



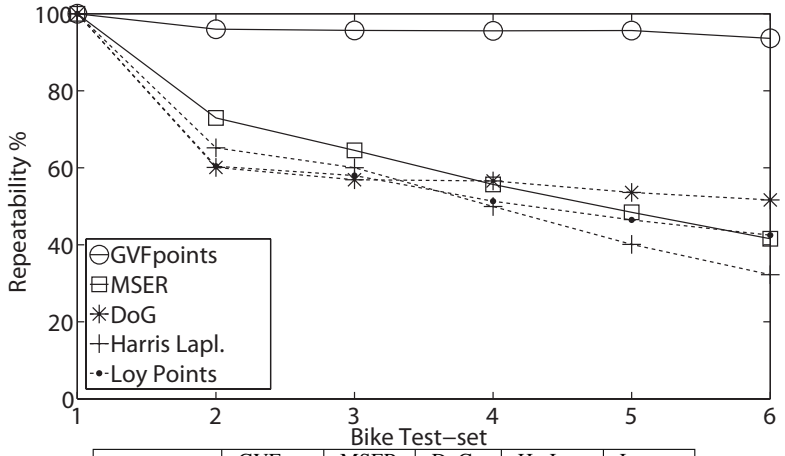
	GVFp	MSER	DoG	HarLap	Loy
(rep)	83.5	70.1	71.3	57.9	60.2
corresp	10467.2	90.2	8623	421.5	950.7
(area)	2369.7	1395.7	51.4	2345.5	576.4
std(area)	4168.9	3824.4	32.8	4916.2	934.0
nr. pts	14524	172	12477	813	2106

Figure 4.15: Repeatability experiment test-set *cars* challenging increasing lightness, or different daytimes, respectively.



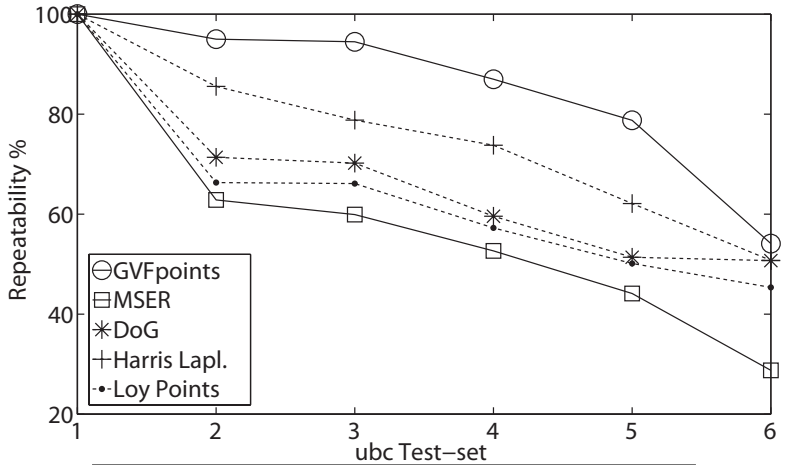
	GVFp	MSER	DoG	HarLap	Loy
mean(rep)	64.1	54.0	65.4	48.9	53.8
corresp	3804.2	56.3	3742.2	348.8	538.4
mean(area)	2493.1	350.8	54.0	3671.8	263.5
std(area)	4264.8	625.5	76.6	11198.4	621.3
nr. points	6995	117	5732	799	1238

Figure 4.16: Repeatability experiment test-set *toy* – changing lighting direction challenge



	GVFp	MSER	DoG	HarLap	Loy
mean(rep)	96.1	63.9	63.1	57.9	59.8
corresp	13795.7	193.2	6893	1096.7	898.3
mean(area)	2476.8	580.2	51.4	5894.8	669.2
std(area)	4248.7	1726.0	32.0	25058.1	1007.6
nr. points	19648	616	14382	2111	1600

Figure 4.17: Repeatability experiment test-set *bikes* – increasing blur challenge



	GVFp	MSER	DoG	HarLap	Loy
mean(rep)	84.9	58.1	67.2	75.2	64.2
corresp	11743.2	415.7	5160.2	1482.3	983.8
mean(area)	2403.7	451.8	51.4	2839.4	669.2
std(area)	4194.2	1708.6	36.8	8376.9	1007.6
nr. points	15388	770	8417	2053	1600

Figure 4.18: Repeatability experiment test-set *ubc* – increasing JPEG compression challenge

Here, the opposite question is investigated: Is it possible to reduce the number of features and maintain the state of the art?

In the following, a contrary approach is evaluated. The experiments aim to show that incorporating color invariance and color saliency into scale-invariant corner detection leads to a more robust and sparse representation than the state-of-the-art approaches.

4.2.2 Color Points

The repeatability experiments in this section evaluate the impact of color and a more robust scale selection for feature localization on the predefined challenges. It is shown that a more sparse distribution of interest points is able to maintain the same or better repeatability performance while obtaining a stable and comparably smaller number of interest points.

In the experiments, the interest points based on the quasi-invariant HSI are denoted as *HSI points* or *light invariant points* and the interest points based on the color boosted OCS as *color boosted points*. For both, they are refered to as *color points*. As the state-of-the-art reference, the *Harris Laplacian* is used. In the repeatability experiments, the *Hessian Laplacian* is evaluated since it provides a richer representation and the best repeatability rates. For the two luminance based approaches the implementation[7] is evaluated similar to [Mikolajczyk et al., 2005b]. All experiments with all four algorithms are carried out with the same parameters $\sigma_D = 1$, $l = 10$, $t = \sqrt{2}$ as defined in Section 3.2.

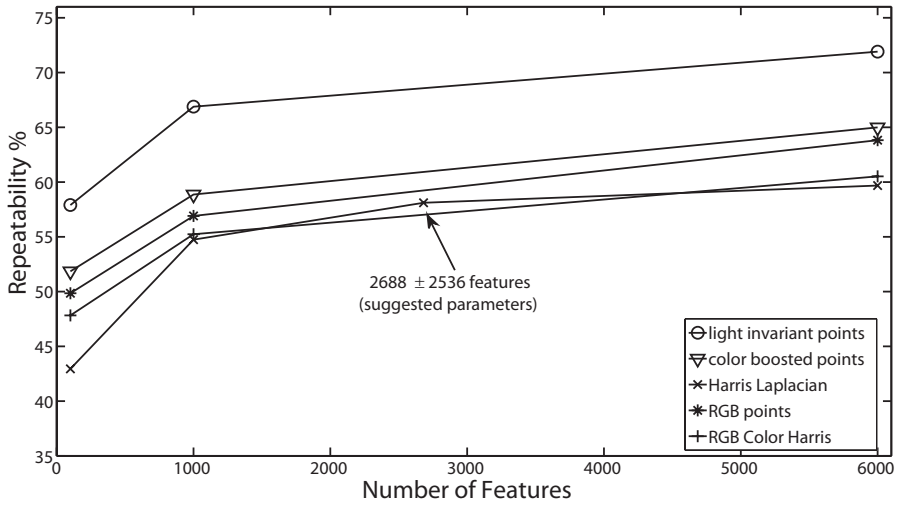


Figure 4.19: The mean repeatability rate of the 18 repeatability challenges per number of points.

The robustness data-set contains 18 test-sets with color images (compare Tbl. 4.1). With these test-sets, the gain in stability of the proposed color points are tested. The experiments aim to show that a stable scale selection, color invariance and color saliency improves the selection

[7] http://www.robots.ox.ac.uk/~vgg/research/affine/

	repeatability	nr. of regions
Harris Laplacian	29.2%	1416 ± 185
Hessian Laplacian	33.4%	430 ± 81
most salient light invariant points	31.0%	100
more light invariant points	51.2%	1000
dense light invariant points	69.6%	6000

Table 4.3: Averaged results from repeatability experiment on the 'nuts' test set with the number of extracted regions $\pm$ their standard deviation.

of robust features. In order to test this ability, the experiments are carried out under varying number of features. All features are selected based on their Harris energy, the approaches are evaluated by their ability to select the most stable interest points. The Harris Laplacian and the Hessian Laplacian have a previously fixed threshold on the Harris energy (referred to as *suggested parameters*). Obviously, this leads to a variable number of points for the images of the data-set based on their contrast. For the color points, the Harris energy gives a better saliency measurement and it makes sense to select a fixed maximum number of interest points. This helps in achieving a predictable density of the description which is more stable to contrast changes.

In Tbl. 4.4 the averaged results of the experiments are shown. Considering only the 1000 most salient locations, color points gain comparable results to the Harris Laplacian and the Hessian Laplacian detector. For lighting change test-sets only 100 interest points per image are enough to outperform greyscale based approaches.

In Fig. 4.19, the averaged results of the experiments with the colorful test sets are shown. The x-axis shows the number of features in the image. Starting with a maximum of 100 points, the number of features is increased up to 6000 points, when statistically all the pixels are covered by at least 10 features at once. This is denoted as a *dense distribution* of salient points. The approaches are evaluated for selecting salient points. For the Harris Laplacian, literature suggests a fixed threshold on the Harris energy. This leads to a variable number of points for the images of the dataset based on their contrast. This increases stability for certain challenges, but is a drawback for others such as varying contrast which happens, e.g. at lighting challenges. It is shown that 1000 color points reach Harris Laplacian performance with the suggested parameters (mean number of points: 2688 ([763,9191] $\pm$ 2536) and even outperforms its dense distribution of 6000 points per image.

Comparing light invariant points with Harris Laplacian, 100 light invariant points are enough to outperform the state-of-the-art. 0.1% of the dense distribution of the color points reach the state of the art providing equally stable locations. Tbl. 4.3 gives the results on the *nuts* test-set challenging varying light direction and shadowing effects.

Fig. 4.20 shows the mean repeatability over the five data-sets with varying lighting (*cars, fruits, movi, nuts, toy*). Increasing the number of Harris Laplacian points does not improve the repeatability against light changes significantly. In contrast, light invariant points remain more stable throughout the experiment. Generally, color boosted points prove to be less repeatable

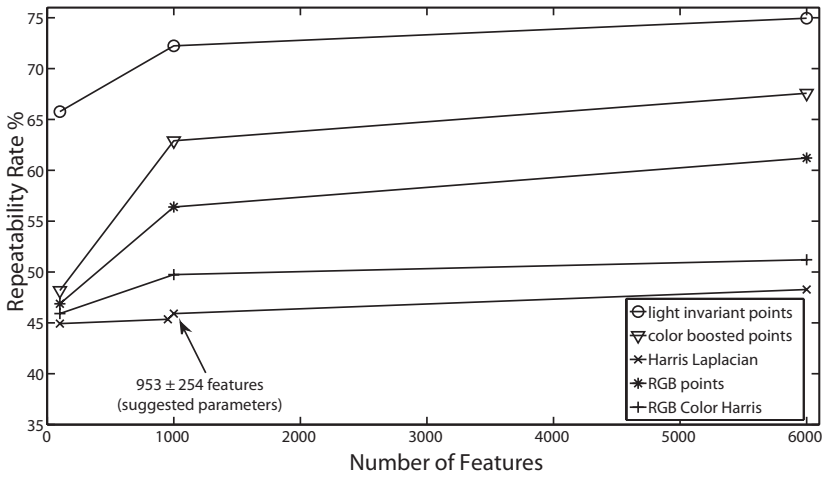


Figure 4.20: The mean repeatability rate of the 5 data-sets challenging lighting changes only.

than the HSI points, which is reasonable as their saliency function is variant with respect to illumination changes and focuses on the occurrence probability. The RGB points have the lowest repeatability among the color interest points tested, and are therefore omitted from the subsequent experiments.

These results show that the Harris energy of the color points gives a better saliency measurement for reducing features. Additionally it can be seen that it makes sense to select a fixed maximum number of salient points. This helps in achieving a predictable density of the description which is invariant to contrast changes. In the worst case this limitation changes the main focus of the interest points in a test-set and thereby reduces the repeatability. For the color salient points, a possible worst case scenario of these test sets is a zoom and rotation scenario: by changing the viewpoint and therefore the scene dramatically, the PCA scale selection should change its basis vectors significantly. Tbl. 4.5 shows the results for this challenge. As already stated, increasing the number of color points increases the repeatability for the color points, but even with a very sparse description, we have reasonable results outperforming the Harris Laplacian. A complete description is not necessarily a matter of quantity but of the reasonable distribution of the points.

4.3 Image Matching

This experiment evaluates the impact of different color spaces and scale selections for feature localization and selection in retrieval scenarios. The experiments are carried out on the ALOI data-set, described in Section 4.1.2. Each of the 1000 object images is queried once and aims

	repeatability	nr. of points
Harris Laplacian	$56.1 \pm 14.8\%$	3179
Hessian Laplacian	$67.7 \pm 16.5\%$	2446
HSI points	$68.3 \pm 12.0\%$	1000
color boosted points	$63.4 \pm 7.2\%$	1000

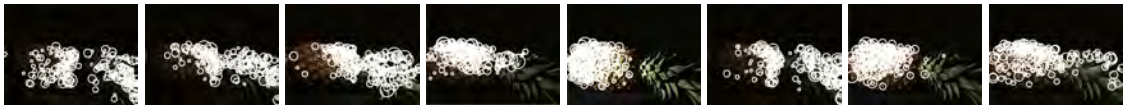
Table 4.4: Averaged results from repeatability experiments with the average number of extracted interest points.

	repeatability	nr. of points
Harris Laplacian	55.8%	4961
Hessian Laplacian	74.4%	5834
few color points	59.6%	100
more color points	63.7%	1000
dense color points	75.3%	6000

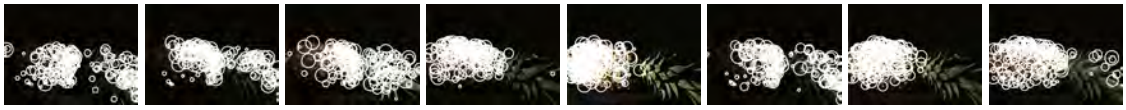
Table 4.5: Averaged results from repeatability experiment on the 'boat' test set (zoom+rotation) with the number of extracted points.

(a) ALOI object nr. 870 under different lighting conditions, query image on the right.

(b) Harris Laplacian



(c) HSI points



(d) color boosted points

Figure 4.21: Example of the interest points extracted for the image retrieval experiment.

for a perfect retrieval result which would be getting all the altered images of the same object as the best matches.

The interest point approaches evaluated provide the locations and scales for the subsequent calculation of SIFT descriptors. For the matching, the similarity between two images is determined by first calculating the Euclidean distances between each possible pair of normalized descriptors. The mean of the $N = 100$ smallest distances is then taken to be the distance between two images. Therefore, the only difference between the different retrieval tests is in the interest point extraction stage.

Three experiments are carried out. First, the data-set of changing illumination direction is chosen to select the best parameters for the subsequent experiments and test the impact of shadowing effects for the proposed approaches. Then, the image retrieval performance is carried out on the data-set showing rotated objects. Finally, the impact of changing color temperature on colorful objects is evaluated.

The part of the dataset that provides images under eight predefined illumination conditions for each object is used, where illumination direction and illumination intensity is varied. With these illumination changes, intensity based approaches suffer from instability and many ambiguous descriptions of shadowing effects. This experiment is carried out with 7000 images as ground truth set and 1000 query images, having thus seven true positives for every object class and query image.

In this experiment, large scale image matching is performed while the number of features is iteratively reduced. Starting with up to 22000 interest points for some of the images, the final minimum number of features is reduced down to ten features per image. The maximum number of N interest points implies that the N interest points with the largest Harris energies are extracted. For high N, probably less interest points than N are detected in an image. In this case all existing features are used. As can be seen from the average number of interest points in Table 4.6, several images in the database did not provide so many interest points. First all extractable interest points (up to 22117 maxima of the Harris energy per image) are used and then N is decreased (see Fig. 4.22). If fewer than N salient points are detected for an image, then all are used. Beginning with all extractable salient points (all of the up to 22117 maxima of the Harris energy per image) the number is reduced to $N = 1000, 500, 200, 100, 50, 10$.

The most similar images are estimated, considering them as ranked results for the retrieval evaluation. The precision and recall for the top 30 retrieved images is obtained and the mean of the resulting F1 score is plotted against the number of interest points in Fig. 4.22.

It is shown that there is a certain minimum number of features that is necessary to discriminate an object from 999 other objects. More importantly it can be seen that too many features make the description ambiguous. Fig. 4.23 shows a specific example of this decrease in performance with an increasing number of salient points. Object 225 is shown in Fig. 4.23(a): It is retrieved perfectly with the first 7 ranks being correct for 200 light invariant points. The next candidate with the 2nd best ranks is object 245 (Fig. 4.23(b)) for this set of parameters. This intuitively makes sense because the image contains similar texture.

With 200 light invariant points, object 225 does not appear in ranks 1–7 for queries using any of the other 999 objects. Taking all the 8775 features available, object 225 appears in 43 queries in the top 7 ranks, worsening the result significantly. For the query by object 225 itself,

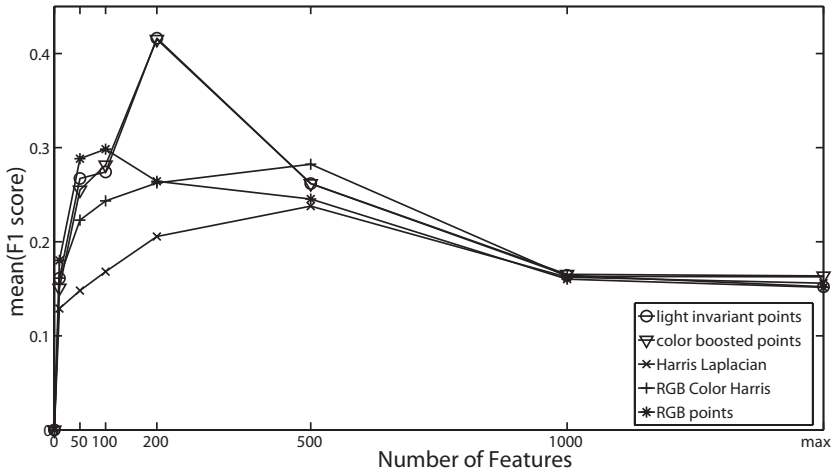


Figure 4.22: Mean F1 score of 30 ranks under different maximum number of features for changing illumination direction on the ALOI database.

(a) ALOI object 225 (b) ALOI object 245 (c) ALOI object 584

Figure 4.23: Sparse color points retrieve object 225 (a) perfectly with rank 8-13 going to object 245 (b). Dense points perform worse shifting object 584 (c) to rank 2-8.

it still ranks one correct candidate at the first rank, having the following 7 from object 584 (see Fig. 4.23(c)). As the only distinct features, the spikes at the border of object 225 and on the head of object 584 remain. The other features become more ambiguous the more points are considered. It is clear from Fig. 4.22 that a higher performance is achieved for a lower number of color salient points than for the Harris Laplacian points.

This problem can also be observed when image data-sets become larger and the discrimination between the images gets lost. In [Pönitz et al., 2010], this problem is referred to as the *Kirschbaum* problem. It is addressed in changing the nearest neighbor classification scheme in a way to discard ambiguous features. This experiment shows that this can also be achieved by selecting only the most salient and thus meaningful features in the process of feature localization.

Overcoming many problems of illumination changes, the color points remain more stable on the test images and thereby outperform all the other approaches with maximum number of 200 color points per image. Harris Laplacian reaches the best performance with a maximum of 500

	avg. precision	avg. nr. points
Harris Laplacian	0.52	381
color boosted points	0.82	192
HSI points	0.82	193

Table 4.6: Average of number of points extracted and the average precision in the image retrieval experiment.

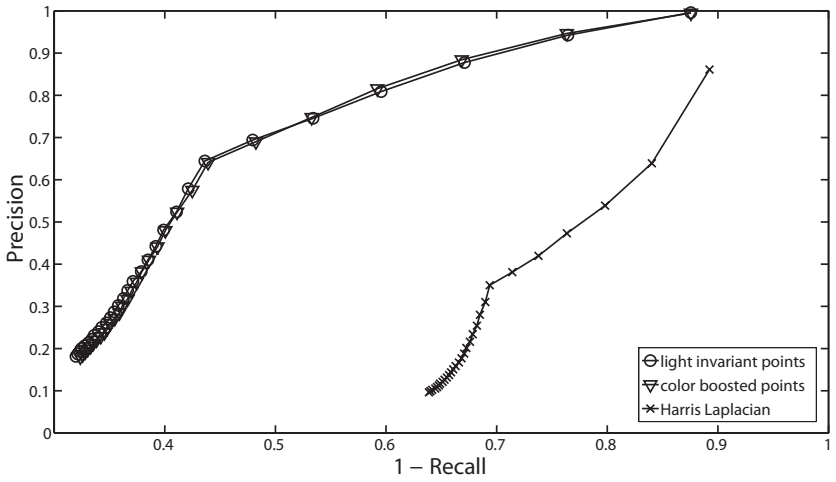


Figure 4.24: Best performing color points compared to suggested parameters of Harris Laplacian for changing illumination direction on the ALOI database.

points, which approximately coincidences with the suggested parameters of a fixed threshold providing 381 [12,8873] $\pm$ 393 Harris Laplacian points (Tbl. 4.6). On average, half of the color points are used to almost solve this retrieval scenario perfectly. Compared to the Harris Laplacian based approach, reducing the number of points to a half reduces the computational complexity significantly.

Suggested parameters denote the standard thresholding of the Harris Laplacian, *best performing* color points refer to best performing maximum number of 200 features per image (see Fig. 4.22). Going into detail on the best performing parameters, 30 matches are regarded and results are given in Fig. 4.24. One complete data-set of object number 870 is given in Fig. 4.21(a). The query image is shown on the very right side. Fig. 4.21(b) - (d) visualize the features extracted for the experiments in this section. It is shown that the color points disregard regions with unstable lighting effects. The color points disregard areas of shadow edges. Pure "shadow features" as they can for example be seen at the leaves of the object are not repeatable and ambiguous and therefore not desirable in an retrieval scenario.

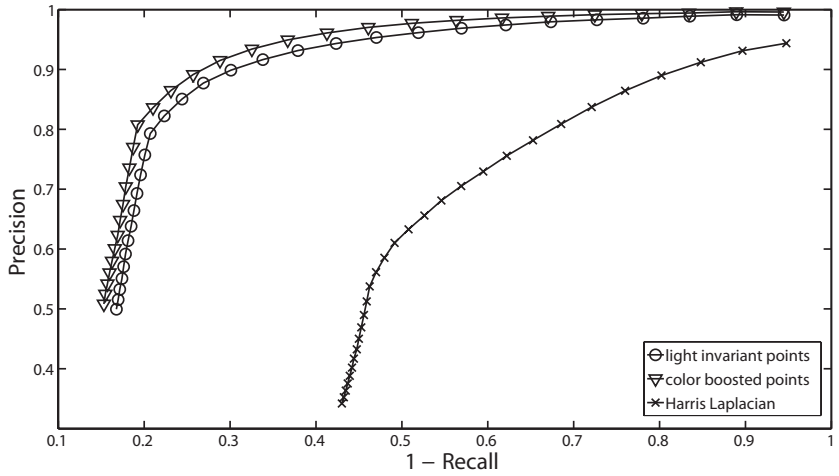


Figure 4.25: Best performing color points compared to suggested parameters of Harris Laplacian for object rotation on the ALOI database.

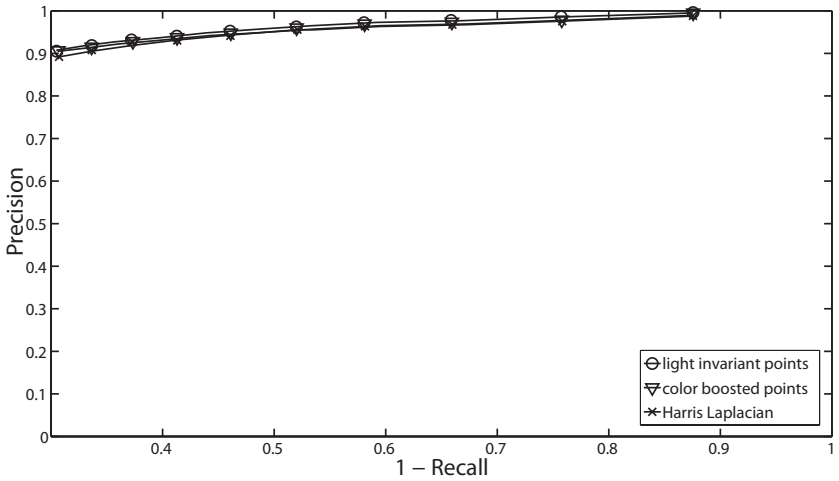


Figure 4.26: Best performing color points compared to suggested parameters of Harris Laplacian for changing color temperature on the ALOI database.

These insights also hold for the following experiment evaluating the stability under geometric transformations of objects. For each of the 1000 ALOI objects, 9 images are taken rotating the object 60° in both directions. From 5° to 30° and 355° to 330° 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 database of 18 000 images. Results are given in Fig. 4.25. Color points perform very similarly to each other but significantly better than Harris Laplacian features.

The ALOI provides predefined changes in illumination color. The best performing parameters are evaluated on the data-set of 18000 images. The changes of color temperature are not different enough to change locations of the detectors significantly. All approaches match almost

perfectly, the plot is shown in Fig. 4.26.

The difference in all the retrieval results is not only very significant because of the special colorful image data, the point is also emphasized by the use of a simple classification by the nearest neighbor. A more stable matching approach would gain and possibly compensate for retrieval performance (as done in the experiments in the next section), but the difference in the quality of the data would no longer be so obvious. Additionally, the approach has a runtime of $O(n^2)$, a reduction of the number of points changes the runtime of each query significantly.

These results only hold for this artificial data-set and the very simple and sensitive classification scheme. Nevertheless, the improved retrieval performance verifies that the proposed feature selection chooses the most discriminant features having single colorful objects on dark background under predefined conditions. It is shown that color based approaches increase the retrieval precision encountering lighting or geometrical changes of colorful objects. In the following, the experiments are extended to use natural images with background clutter and state-of-the-art feature classification for large scale object categorization.

4.4 Object categorization

The experiments in this section aim to demonstrate that state-of-the-art results can be obtained when using significantly fewer color salient points in object categorization.

One of the most successful approaches to object categorization is the bags-of-words in combination with SVM classifiers – the best performing methods at the PASCAL Visual Object Classes Challenge 2006 [Everingham et al., 2006] and later used variations on this approach. As a benchmark the algorithms that are evaluated in more detail by [Zhang et al., 2007] are used. This experiment is carried out in the same scheme with the same evaluation measures. The only difference is that the PASCAL VOC 2007 data-set is used (as described in Section 4.1.3) which contains the same images, but with additional images to make the data-set approximately of the double size and provides 20 instead of 10 object classes.

[Zhang et al., 2007] use a Harris Laplacian detector, a combination of SIFT and SPIN descriptors using a bags-of-words approach and an EMD Kernel SVM for classification. The workflow of the algorithm is shown in Figure 4.27. The best performing parameters of their evaluation is used. Image signatures consisting of 40 clusters of these descriptors are extracted. Clustering is done using the k-means algorithm. The earth mover's distance (EMD) [Rubner et al., 2000] showed to perform best for the task of estimating the similarity between image signatures. These image distances are incorporated into a one-against-all SVM classifier. The aim is to test the effect of using salient points obtained from luminance and color information on the categorization and calculation time performance[Stöttinger et al., 2009b].

Only the first step in the flowchart in Figure 4.27 is changed, all succeeding steps of the approach are carried out identically. An example showing the color points and Harris Laplacian points of an image from the VOC 2007 dataset is given in Figure 4.28. For this image, the color salient point detectors focus on the more colorful foreground objects. Fewer salient points are found in the background by the color salient point detectors than by the Harris Laplacian detector. As a consequence, in images where the Harris Laplacian approach provides up to 2500

Figure 4.27: Flowchart of the approach of Zhang et al. [Zhang et al., 2007] used in the object categorization experiment.

(a) original (b) Harris Laplacian (c) HSI points (d) color boosted points

Figure 4.28: VOC 2007 image number 5221 and the interest points used in the object categorization experiment.

extracted points, the HSI features provide sometimes just 1% of that number. However, those points are more distinct and discriminative.

The quality of each single one-against-all classifier is tested by carrying out 10 fold cross validation on the VOC 2007 training set. This is referred to as the discrimination accuracy and gives a measure for the quality of the training data and the trained SVM model. The second column of Table 4.7 shows the discrimination accuracy for different numbers and types of salient points and descriptors, averaged over 20 one-against-all classifiers. No matter which description is fed into the SVM, the classifier manages to reach about 93% accuracy on this 2 class problem.

The accuracy when categorizing the test data into one of 20 classes is shown in the third column of Table 4.7. The results for the SIFT and SPIN descriptors used individually are given as mean $\pm$ standard deviation over all classes. The combination of the SIFT and SPIN descriptors as the final classification accuracy is given per class, with the mean over all classes shown in the last rows of the table. For each class, the best classification result is shown in bold. The table shows that the more sparse description is equally effective but much more efficient.

For this experiment, reducing the number of Harris Laplacian points by about 50% gives around 60% of the original categorization performance. This does not hold for color salient points: the 400 salient points with the highest Harris energy per image are kept and the performance of the richer description (800 points) is maintained.

Therefore it is argued that the color points are more distinct and discriminative, even when intensity based descriptors are used. It is shown that the use of color in the detection phase does not degrade the model and the description is as complete as for the best performing Harris

SIFT	discrimination	categorization	number of points
Harris Laplacian	93.12 ± 2.52%	79.5 ± 15.5%	771 ± 531
	92.41 ± 2.65%	54.6 ± 20.7%	387 ± 72
light invariant points	93.27 ± 2.17%	81.7 ± 10.6%	800
	93.54 ± 2.34%	80.9 ± 11.4%	400
color boosted points	93.49 ± 2.28%	83.0 ± 10.1%	800
	93.41 ± 2.44%	**83.1 ± 10.4%**	400
SPIN			
Harris Laplacian	92.95 ± 2.64%	66.2 ± 17.8%	771 ± 531
	92.19 ± 2.8%	38.4 ± 12.9%	387 ± 72
light invariant points	**93.16 ± 2.61%**	**68.9 ± 18.3%**	800
	93.08 ± 2.56%	68.3 ± 18.9%	400
color boosted points	93.13 ± 2.71%	68.8 ± 17.2%	800
	93.04 ± 2.62%	68.7 ± 16.4%	400
SIFT + SPIN			
Harris Laplacian	93.50 ± 2.4%	85.9± 9.9%	771 ± 531
	92.83 ± 2.75%	54,8 ± 18.5%	387 ± 72
light invariant points	93.52 ± 2.61%	**86.6 ± 8.7%**	800
	93.57 ± 2.37%	86.5 ± 8.5%	400
color boosted points	93.49 ± 2.65%	86.2 ± 8.9%	800
	93.47 ± 2.38%	86.4± 8.4%	400

Table 4.7: Discrimination accuracy of the classifier and the categorization accuracy of the challenge as average ± standard deviation over classes.

Laplacian detector. The classifier is able to discriminate between the given object classes equally well, while training on significantly fewer descriptors.

For certain classes, the categorization accuracy increases when fewer features are used: Colorful object classes like humans (especially the clothes) or cats seem to benefit from more sparse color points. Object classes where the objects typically cover a large part of the image (airplanes, busses) also show increased performance. Performance seems to decrease for small objects with low color and contrast as is for objects of small area without shadows or specular effects as it is the case for the class sofa.

The number of salient points are an indication for the runtime of the system. Every step of the object categorization (see Fig. 4.27) has to deal with only about half of the data as the state of the art does, which diminishes the runtime significantly. For building the image signatures, assigning 400 SIFT descriptors to 40 clusters takes on average 1.4 seconds including all I/O operations. Clustering the descriptors using the k-means algorithm is the most time consuming task in the approach. Its run-time is dependent on the number of descriptors given and the relation of the number of data points and centroids. Additionally, it gains complexity in dealing with many outliers [Leibe et al., 2006]. The software prototype uses iterative k-means which is 3.8 times faster using the more discriminative color points than using the Harris Laplacian points

	Harris Laplacian	HSI points	color boosted points
nr. of points	771	400	400
Avg. SIFT	79.5 ± 15.5%	80.9 ± 11.4%	**83.1 ± 10.4%**
Avg. SPIN	66.2 ± 17,8%	68.3 ± 18,9%	**68.7 ± 16,4%**
Aeroplane	89.2%	**94.9%**	81.3%
Bicycle	91.3%	**94.7%**	92.5%
Bird	92.1%	93.9%	**95.3%**
Boat	90.6%	**95.9%**	87.3%
Bottle	95.1%	**95.7%**	90.0%
Bus	82.7%	**83.8%**	82.3%
Car	76.8%	77.5%	**77.6%**
Cat	70.5%	**86.2%**	85.1%
Chair	72.2%	**87.0%**	75.1%
Cow	89.0%	84.9%	**90.8%**
Diningtable	88.2%	83.5%	**92.0%**
Dog	91.2%	86.9%	**93.2%**
Horse	86.0%	**92.1%**	88.9%
Motorbike	**93.6%**	91.1%	92.6%
Person	55.7%	58.0%	**59.7%**
Pottedplant	**93.5%**	87.5%	82.2%
Sheep	93.4%	89.4%	**95.8%**
Sofa	**92.7%**	82.5%	87.8%
Train	**90.0%**	82.5%	89.1%
Tvmonitor	83.4%	82.9%	**89.3%**
Average	**85.9±9.9%**	**86.5±8.5%**	**86.4±8.4%**

Table 4.8: Categorization accuracy and averaged number of interest points on the PASCAL VOC 2007 data-set.

in the clustering stage. There is a three times higher complexity in finding the positions and in scale of the color salient points. However, the calculation time is made up in the next steps of the object categorization framework. Having significantly fewer salient points, less local descriptors have to be calculated and clustered. Using half the regions means half the processing time in calculating the descriptors.

This experiment shows that the best performing object categorization approach from 2006 can be carried out with significantly less features. The following experiments use the same features and same data-set, but as a part of an international feature benchmark using the best classification system of 2009.

4.5 Feature Benchmark

In conjunction with the CVPR 2009 conference a workshop called *"Feature Detectors and Descriptors: The State of the Art and Beyond"* was held. The idea of the workshop was an in-depth evaluation and discussion about solved problems and future challenges of the community, solely focusing on local detectors and descriptors. Additionally, Krystian Mikolajczyk of the University of Surrey set up an international benchmark, where everybody was invited to extract new features from the VOC PASCAL 2007 data-set (see Section 4.1.3) and upload them to Surrey's machine learning framework.

The descriptors are clustered using the k-means into 4000 clusters. Image representation is built by a histogram of cluster occurrences. This is a 4000 bin histogram where the bins correspond to the clusters. Each bin contains the number of descriptors in the image that fall into the cluster corresponding to that bin. Euclidean similarity measure is used to compare descriptors. For kernel construction χ^2 distance and generalized RBF kernel is used. The extracted test features are matched to the clusters and the occurrence histogram is produced. Classification is performed using spectral regression kernel discriminant analysis (SRKDA) [Cai et al., 2007] and SVM kernel-fusion [Yan et al., 2009] with χ^2 distance measure. Details are given on the web-page[8].

The performance is measured using the Average Precision (AP). Whereas precision and recall are based on the whole list of images returned by the system, average precision considers ranked results. It is the average of precisions computed at the point of each of the correct images in the ranked sequence. Geometrically it is the area under the precision-recall curve. Mean Average Precision (MAP) denotes the mean of this metric over all queries.

This classification framework [Everingham et al., 2009] in combination with color features of [van de Sande et al., 2009] won the annual PASCAL challenge the last two times in a row and should therefore be regarded as probably the best classification system today. It was possible to submit detectors only, standard SIFT descriptors are then used as descriptors. For descriptors evaluation only, Harris Laplacian detections have been used as detection approach. The results are therefore comparable to the evaluation in the previous section.

The goal was to have an independent evaluation of features with the best performing classification framework. 33 different features are evaluated, the approaches are submitted from the following universities. Detectors only are submitted from three participants:

- CMP Prague (MSER in opponent chromatic space)
- University of Surrey (Harris Laplacian + Hessian Laplacian + DoG)
- TU Vienna (Sparse Color Points)

The following participants submitted descriptors, Harris Laplacian is used for localization:

- EPFL Lausanne (DAISY)
- ETH Zurich (SURF)

[8] `http://www.featurespace.org/`

class	400 HSI points	800 HSI points	performance gain
aeroplane	0.585912	0.662140	1.13
bicycle	0.234748	0.310271	1.32
bird	0.224675	0.293639	1.31
boat	0.448092	0.607609	1.36
bottle	0.128760	0.145417	1.13
bus	0.303697	0.371983	1.22
car	0.550822	0.628213	1.14
cat	0.393214	0.455901	1.16
chair	0.353188	0.442940	1.25
cow	0.196282	0.263788	1.34
diningtable	0.204528	0.314897	1.54
dog	0.294662	0.328828	1.12
horse	0.609964	0.702450	1.15
motorbike	0.337167	0.409013	1.21
person	0.707702	0.752396	1.06
pottedplant	0.101262	0.134227	1.33
sheep	0.161029	0.325316	2.02
sofa	0.215373	0.332121	1.54
train	0.505796	0.643607	1.27
tvmonitor	0.285369	0.368749	1.29
Average MAP	0.342112	0.424675	1.24

Table 4.9: Mean average precision results of the 2 submitted results of the TUVienna1 (400 HSI points) and TUVienna2 (800 HSI points) approaches in the CVPR local feature benchmark. Performance gain gives the ratio of the average precision of the two approaches.

- Stanford University (CHOG)
- Harvard Medical School (Ordinal SIFT)
- University of Amsterdam (Color histograms, Color moments, Color SIFT))
- University of Surrey (Color Descriptors)

Harris Laplacian and SIFT is used as a baseline, as done in the previous section. In the benchmark, the two parameter sets evaluated in the previous section using HSI color points were submitted. First submission was the approach from the previous section having maximum number of features of 400 HSI points (TUVienna1), the second approach providing up to 800 HSI points (TUVienna2).

With the state-of-the-art classification framework more HSI points lead to an increased performance of the categorization task. As can be seen in Tbl. 4.9, the double number of features

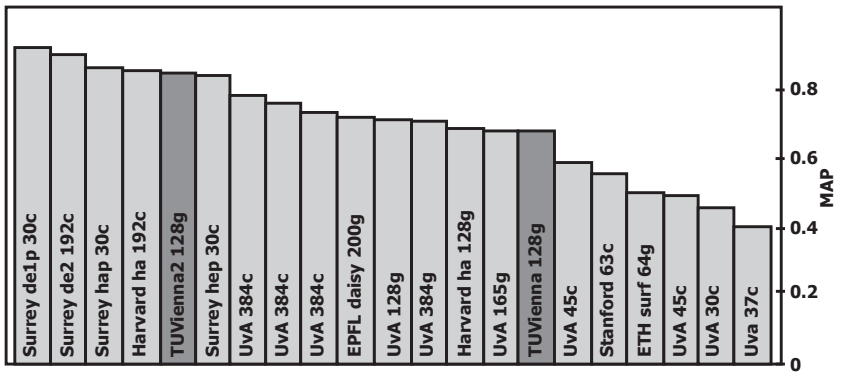


Figure 4.29: Best 21 approaches ranked per MAP.

increases the precision significantly. The third column gives the performance gain in terms of the ratio of the performance of the two approaches.

The precision rises for all the classes. This means that in contrast to the one-against-all SVM used in the previous section, the newest and best performing classifier today is able to take advantage of the additional features. Still, the results from the previous experiments holds: The better performing approaches use a significantly denser feature representation than TUVienna2. Therefore, the benchmark shows that those more sparse features perform equally well compared to other, more dense representations of features processing significantly less data.

All approaches are measured by the MAP and shown in Fig 4.29, the best performing sparse color points are ranked as the 5th best approach in the benchmark. The leading approach from University of Surrey uses a very dense distribution of features with a very low dimensionality of 30 for the local description. The numbers after the names of the approaches give the dimensionality of the descriptors. “g” denotes gray-level descriptors, “c” color descriptors. TU Vienna denotes the only approach using color in the stage of detection. The University of Surrey – and therefore the benchmark focuses on this issue – addresses the issue of efficiency and reducing data in using more, but lower dimensional features successfully in terms of overall performance.

The results in detail show that a maximum of 800 HSI points outperform all other approaches in 4 out of 20 classes (see Fig. 4.30) and are ranked second in 3 other challenges (see Fig. 4.31). This is remarkable as the leading approaches from University of Surrey use three detectors at once, providing a vast amount of features for the subsequent steps of the categorization framework. The proposed approach uses only a fraction of these features still obtaining state-of-the-art results.

In [Mikolajczyk et al., 2005a] another property of local features is evaluated: the *feature density*. An agglomerative clustering approach may produce a varying number of clusters, varying from one cluster for the whole data-set up to a level of one feature per cluster. If one detector

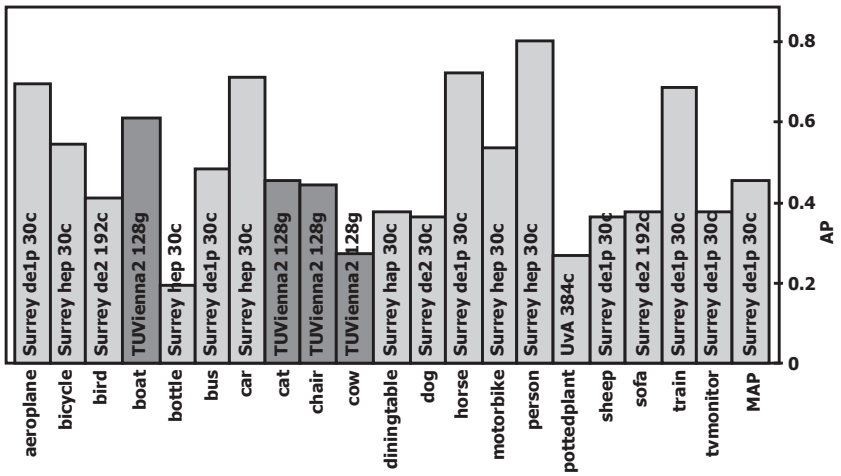


Figure 4.30: Overall results for rank 1 approaches.

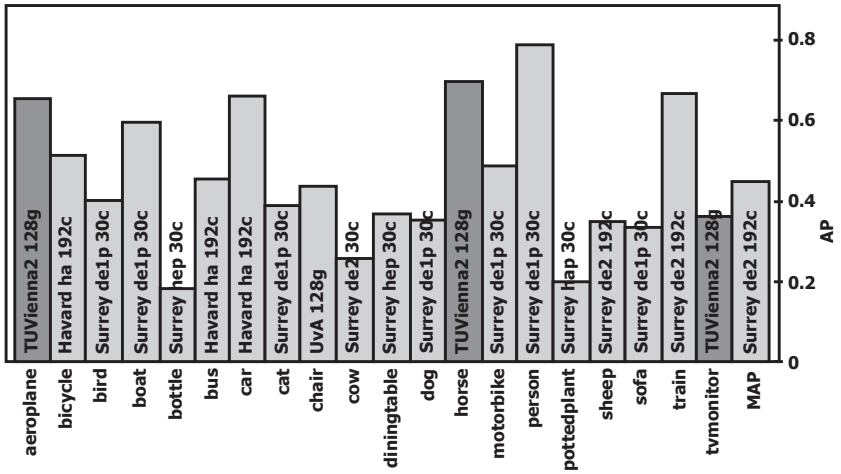


Figure 4.31: Overall results for rank 2 approaches.

provides less features than another and the same number of clusters for the same features is built, the sparse case will provide more clusters containing a single feature. Every evaluation is biased by the number of features and the number of clusters. To make the results comparable they refer to average density of features per cluster. Fig. 4.32 shows that the best performing approach and the proposed TUVienna2 provides significantly higher feature density than the other approaches, whereas the the more sparse TUVienna1 approach provides the second lowest feature density. As most of the approaches provide almost equal feature's density it is not clear what information this measure provides. However, it seems to be correlated with accurate categorization performance.

The benchmark showed that there is still room for improvement in the stage of classification

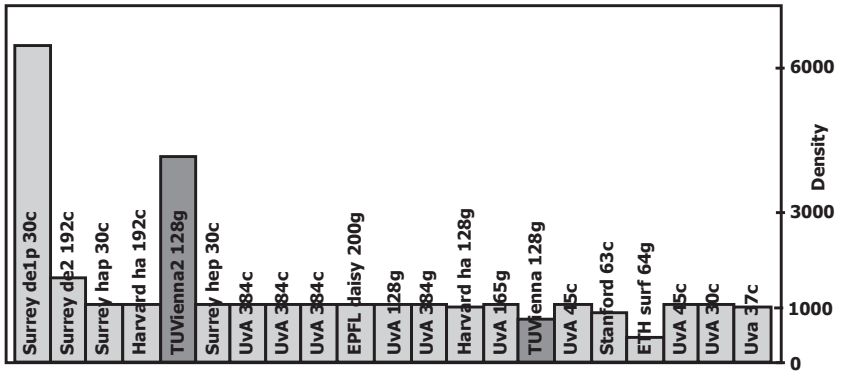


Figure 4.32: Density of features measured in singular clusters.

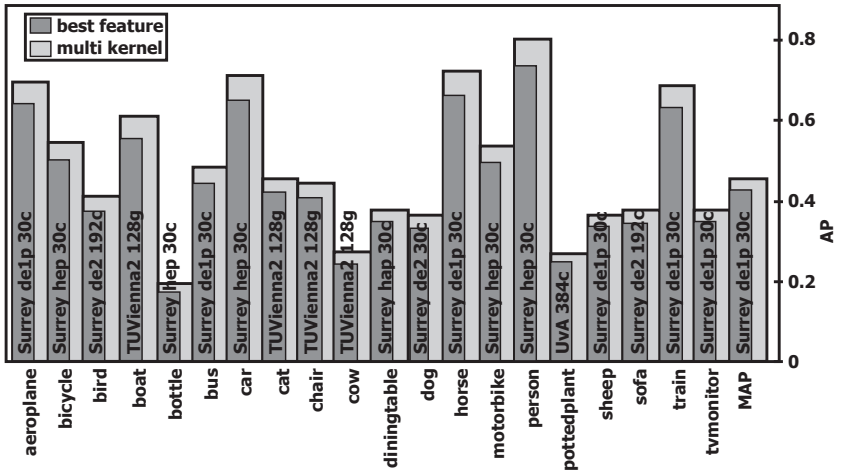


Figure 4.33: Comparison SVM classifiers.

using multi-class SVM. In Fig. 4.33 the difference between a recent SVM feature selection (best feature) and the current best performing multi-kernel fusion [Yan et al., 2010] is given. It can be seen that there is a gain of about 0.02 in precision for every class and every feature of the best performing approaches.

An in depth evaluation of the processing complexity of these features is missing so far in the literature. As stated in the benchmark overview, the best performing *Surrey de1p 30c* use the Harris Laplacian evaluated in the previous section, merged with the DoG and the Hessian Laplacian. As those additional detectors provide significantly more features than the Harris Laplacian (e.g. [Mikolajczyk et al., 2005a]) it suggests that this representation is at least 5 times denser than the TUVienna2 approach and 10 times denser than the TUVienna1 approach. In this benchmark, the sparse color points is the only approach that aims for a feature selection in the stage of detection. The trend of the community is to decrease the dimensionality of the descriptors. This is debatable, as the reduction of the dimension of local description gives many advantages

in subsequent operations as training and clustering, whereas the discrimination capabilities are diminished for large scale problems (255^{128} possible SIFT descriptors versus 255^{30} possible color descriptors of the University of Surrey). Selecting the best features before the description decreases the amount of data without this effect. Estimating the number of features of the winning approach *Surrey de1p* with 5 times more than TUVienna2, it uses 1.17 times more data than TUVienna2. Or, in overall amount of data, the TUVienna2 equals a descriptor length of about 25 in terms of description data.

With this benchmark it is shown that the feature localization has an impact on state-of-the-art classification schemes. Using more stable and salient features, the number of features can be reduced maintaining the *meaningful* features and thus state-of-the-art performance.

4.6 Summary

In this chapter, new approaches to localize image features are evaluated. GVF points are evaluated for their robustness under predefined challenges, showing that they give a dense and robust image representation. Especially for rotation, noise and JPEG artifacts, the approach is almost perfectly stable outperforming state-of-the-art approaches.

The proposed color points showed to outperform other interest points in several scenarios including robustness, image retrieval and object categorization. The overall trend in these experiments is that the more stable and robust the subsequent classification of features is, the less impact the feature localization has on the final performance. Nevertheless, also on the best performing system today, a trend is shown that more robust and salient features can maintain the performance with significantly less features than other approaches.

Extensive experimental results show that a sparser but equally informative representation, obtained by making use of color information, can be directly passed to current and successful image retrieval and object categorization frameworks, which then obtain state of the art results while processing significantly less data. When using color interest point detectors for object categorization with a one-against-all SVM classifier, the same performance is obtained using about half the number of color interest points compared to greyscale interest points. Such a reduction in the amount of data to be processed is useful in applications for which limited computing power is available.

On the latest international benchmark on local features, the color points showed that this assumption still holds for current categorization techniques: The approach is in the leading field of the benchmark with significantly less features than the best performing approach.

There is a strong trend towards decreasing the description data when scaling recognition problems to larger data-sets. The proposed approach is today the only one successfully trying to reduce the numbers of features in the stage of detection.

CHAPTER 5

Interest Point Detectors for Video

Spatio-temporal features aim to extract robust locations in videos. The main idea is to focus on stable spatial patterns as is done in images, but to extend this concept to the temporal domain to find salient patterns in a movement as well. The frames of videos are not regarded on their own, but the input video data is regarded as a volume. For the feature extraction, this means that the main *2D* concepts for images are extended to *3D*. A music video visualized as a volume V is shown in Fig. 5.1. The video V has the volumetric extension x denoting the width of the frames, y, the height of the frames and t, the number of frames.

In this sense, the extension of the Harris corner detector, the Harris3D detector [Laptev and Lindeberg, 2003a] searches for corners not only in a frame, but in video motion as well. The concept is illustrated in Fig. 5.2. As input there is a black and white video with a resolution of 50×50 pixels and 50 frames. The lower half is black. Within the 50 frames, the edge moves down and up again. The first frame is shown in Fig. 5.2a. The motion can be seen when the

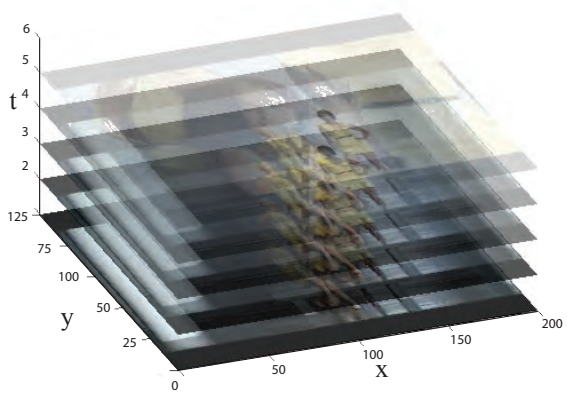
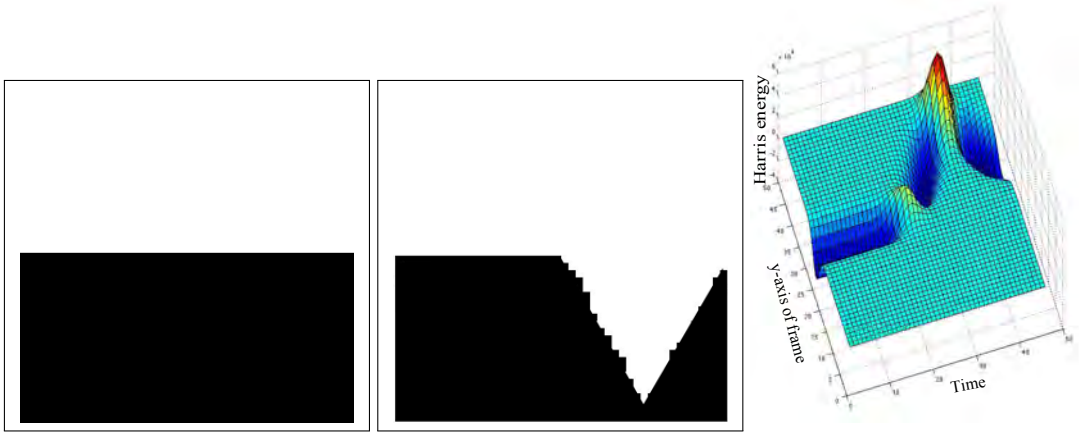


Figure 5.1: A video visualized as a layered volume of video frames.



(a) Frame 1 (b) Central y / t plane (c) Harris energy

Figure 5.2: Illustration of temporal corner detection. (a) shows the first video frame, lower half is colored black. Within 50 frames, the edge moves down and up again (b). (c) shows the y/t plane of the resulting spatio-temporal Harris energy ($\sigma = \tau = 3$).

(a) Walking cycle 1 (b) Walking cycle 2 (c) Walking cycle 3 (d) Walking cycle 4 (e) Walking cycle 5

Figure 5.3: Very simple walking cycle of a matchstick man. Corners without acceleration are marked green, corners with acceleration are colored blue.

central y/t plane is observed (see Fig. 5.2b). The spatio-temporal Harris energy gives then a volume of the same size as the video. Similar to Fig. 5.2b, the central y/t plane of the energy is given in Fig. 5.2c. For the edge without movement, the energy is negative, for uniform areas without any change in structure or movement, the values are 0. When the edge moves down, a small maximum is encountered for the first edge of the movement, for the full change in motion (*sharper* corner in the movement) the energy gives a higher maximum.

The idea is that for natural movements, the changes in a uniform movement (e.g. where the acceleration is non-zero) are the most stable location to describe a motion or a video. For the very simple walking cycle in Fig. 5.3, the torso and the head of the matchstick man remain stable moving uniformly from left to right. Therefore, only spatial corners are found (marked green). When there is acceleration taking place on the extremities, spatio-temporal corners are

extracted and marked blue. The concept is that those locations are the most stable ones and are enough to describe the motion. In the following section a survey of successful luminance based spatio-temporal features is given. Section 5.2 gives an outlook to concepts to extend those features to color.

5.1 Luminance Based Spatio-Temporal Features

In Section 5.1.1, the concept of spatio-temporal corner detection is described in detail. Section 5.1.2 gives the most successful spatio-temporal blob detector. Section 5.1.3 describes the Cuboid detector, the only detector which has no direct image counterpart.

5.1.1 Corner Detection

The **Harris3D** detector for videos [Laptev and Lindeberg, 2003a] extends the Harris corner detector [Harris and Stephens, 1988] for images (see Section 3.1.1). The authors compute a spatio-temporal second-moment structure tensor at each video point using independent spatial and temporal scale values σ and τ, a separable Gaussian smoothing function G, and space-time gradients L. Extending the scale space to the temporal domain, the temporal variance τ^2 is added to get

$$L_{\mathbf{x},\sigma^2,\tau^2} = \nabla G_{\mathbf{x},\sigma^2,\tau^2} \otimes V_{\mathbf{x}} \tag{5.1}$$

for the position **x** in the corresponding video volume V. The position in the video is defined by x and y in the spatial and t in the temporal domain. t typically refers to the frame number. The spatio-temporal Gaussian kernel is defined as

$$G_{\mathbf{x},\sigma^2,\tau^2} = \frac{1}{2\pi\sigma^4\tau^2} e^{-\frac{x^2+y^2}{2\sigma^2}-\frac{t^2}{2\tau^2}}. \tag{5.2}$$

It is separable and thus can be calculated for each dimension on its own and in parallel. This extension gives the structure tensor for every location and scale having

$$M = \left\{ G \otimes \begin{pmatrix} L_x^2 & L_{xy} & L_{xt} \\ L_{xy} & L_y^2 & L_{yt} \\ L_{xt} & L_{yt} & L_t^2. \end{pmatrix} \right\} (\mathbf{x}). \tag{5.3}$$

The final locations are extracted by applying

$$\begin{aligned} C_H &= \det(M) - k \cdot \text{trace}^3(M) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \end{aligned} \tag{5.4}$$

and extracting the positive maxima of the corner function C_H based on their eigenvalues. Compare with Eq. 3.28 for the relation to the detection in two dimensions only. The results can be efficiently calculated by the elements of the structure tensor:

(a) Harris3D (b) Hessian3D (c) Cuboid

Figure 5.4: The three approaches most successful approaches of spatio-temporal features on a Hollywood movie frame. Parameters are as given in Section 5.1.1 to 5.1.3, approaches are used in Chapter 6 and [Wang et al., 2009].

$$\begin{aligned} \det(M) &= L_x^2 L_y^2 L_t^2 + L_{xy} L_{yt} L_{xt} + L_{xt} L_{xy} L_{yt} - L_x^2 L_{yt}^2 - L_{xy}^2 L_t^2 - L_{xt} L_y^2 L_{xt} \\ \text{trace}(M) &= L_x^2 + L_y^2 + L_t^2 \end{aligned} \tag{5.5}$$

[Laptev and Lindeberg, 2003a] propose an optional mechanism for spatio-temporal scale selection. This is not used in the experiments, but the points are extracted at multiple scales based on a regular sampling of the scale parameters σ, τ as suggested by the authors. The original implementation[1] and its settings $k = 0.0005$, $\sigma^2 = 4, 8, 16, 32, 64, 128$, $\tau^2 = 2, 4$ with a detection threshold of 10^{-9} are used in Section 6.4.

A scale selection aims to detect features at various scales and to have a salient measure that gives local extrema when a meaningful and stable detection is reached. The Harris Laplacian approach can be extended in a straightforward way to reach a maximum when the image function builds the most perfect Gaussian blob at the current scale.

For the case of videos, the normalized LoG for videos treated as volumes is defined as

$$\Lambda^v = \sigma^{2a} \tau^{2b} L_x L_x + \sigma^{2a} \tau^{2b} L_y L_y + \sigma^{2c} \tau^{2d} L_t L_t \tag{5.6}$$

where the normalization parameters are $a = 1$, $b = \frac{1}{4}$, $c = \frac{1}{2}$, $d = \frac{3}{4}$. As typical applications and benchmarks for video matching do not challenge scale invariance with great variation yet, the suggested implementation without scale selection performs almost equally well [Laptev, 2005]. The limitations of this single scale approach are shown in the evaluation in Chapter 6.

5.1.2 Blob Detection

The **Hessian3D** detector [Willems et al., 2008] is the spatio-temporal extension of the Hessian blob detector [Lindeberg, 1998]. The saliency of a location is given by the determinant of the 3D Hessian matrix. It is related to the Harris3D approach but more efficient in the calculation of the features. The approach aims for an efficient detection and to provide a dense distribution of features. It is defined by the structure tensor Γ. Similar to the structure tensor given before in

[1] `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`

Eq. 5.3, the elements are defined by their location $\mathbf{x}$, the scale of the Gaussian derivative σ and the temporal scale τ (compare Eq. 5.2).

$$\Gamma = \left\{ \begin{pmatrix} L_x^2 & L_{xy} & L_{xt} \\ L_{xy} & L_y^2 & L_{yt} \\ L_{xt} & L_{yt} & L_t^2 \end{pmatrix} \right\} (\mathbf{x}) \tag{5.7}$$

Similar to the DoH detector explained in detail in Section 3.1.2, the locations are determined by the analysis of the structure tensor Γ. The saliency S of an interest point is given by its tensor determinant

$$S = |det(\Gamma)| \tag{5.8}$$

For the 3D case, it is not guaranteed that in case of S being positive, all eigenvalues are. Therefore, also *saddle points* are detected. These are stationary points of a function which are not local maxima. For the 2D case, a sufficient criteria for a saddle point is an indefinite or null Hessian matrix. That does not hold for the 3D case.

[Willems et al., 2008] state that for typical applications it does not matter about the property of the local features as long as they are robust, repeatable and reliable. Therefore they keep saddle points in the representation. In case these points should be disregarded, all eigenvalues have to be checked to be positive in advance.

Aiming for an efficient scale selection and to avoid the iterative optimization of the approach of [Laptev, 2005] the scale selection relies solely on S. The idea is that $S = det(\Gamma)$ reaches a maximum at the center of a perfect Gaussian blob $g(\mathbf{x}, \sigma_0)$ in n dimensions of size $\sigma_0 = [\sigma_{0,1}, \sigma_{0,2}, ..., \sigma_{0,n}]$. It can be shown that there is a relation between the scale of the elements of the structure tensor and the scale of actual size of the underlying structure g_{σ_0} which is defined as

$$\sigma = \sqrt{\frac{2}{n}}\sigma_0 \tag{5.9}$$

With this relation of structure in the image and scale in the tensor, a simultaneous localization and scale selection is possible. Therefore, local maxima are extracted from the 5D space $(\mathbf{x}, \sigma, \tau)$. Having a non-iterative procedure for scale-selection, a significant speed-up of the implementation is achieved.

For efficiency, box-filter operations are applied on an integral video structure on multiple scales. Therefore they do not have to be computed hierarchically but can be efficiently implemented by simply upscaling the box-filters [Ke et al., 2005]. Each octave is divided into 5 scales, with a ratio between subsequent scales in the range [1.2; 1.5] for the inner 3 scales. A non-maximum suppression algorithm selects the common extrema over space, time and scales: (x,y, t,σ, τ). For certain applications, some dimensions can be disregarded, e.g. τ when time invariance is not an issue.

In the evaluation the authors' implementation is used with the suggested parameters of 3 temporal and 3 spatial scales, a previous sampling of the video of every second pixel both in spatial and temporal dimension and a detection threshold of 0.001. The implementation is very

efficient and aims to provide a dense representation of the video. This is true for certain conditions but for the experiments in Section 6, the number of detections is comparable with Harris3D, while the detected scales tend to be bigger than the ones for Harris3D.

The **Maximum Stable Volumes** (MSVs), proposed by [Donoser and Bischof, 2006], are an elegant extension of the MSER detector [Matas et al., 2002] described in Section 3.1.2. The approach is an efficient concept of 3D segmentation and has not been used as spatio-temporal features for video matching yet.

The volume is analyzed as a component tree which was used for watershed segmentation in volumes [Couprie et al., 2005]. For a given volume , the component tree T has connected volumes P_i^ω of a certain threshold ω as nodes. Every node consists of a set of connected voxels $v \in T$ with

$$\forall v \in P_i^\omega, \forall u \in \text{boundary}(P_i^\omega) \rightarrow j(v) \geq j(u) \tag{5.10}$$

where u are the boundary voxels of P_i^ω and the function j gives the intensity of the voxel. There are levels of the component tree for every level of intensity. Moving up and down the tree within the levels, a binary thresholding is achieved.

The MSVs are defined as the connected volumes (nodes in T) with the highest stability. The stability criterion is similar to the 2D case (see Eq. 3.14) given by

$$\Psi_{MSV}(P_i^\omega) = \frac{|P_i^{\omega+\Delta}| - |P_i^{\omega-\Delta}|}{|P_i^\omega|} \tag{5.11}$$

where $|P_i^{\omega+\Delta}|$ denotes the cardinality of $P_i^{\omega+\Delta}$. The parameter Δ defines the stability (and thus the number of) the extracted features.

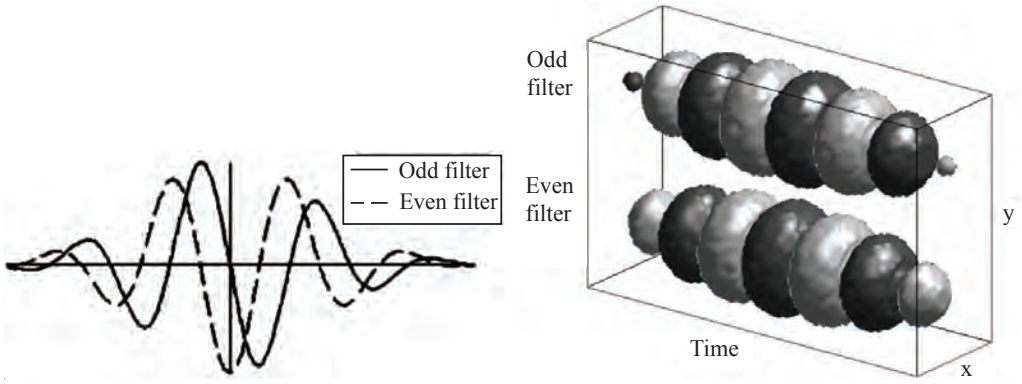
5.1.3 Gabor Filtering

The **Cuboid** detector is a set of spatial Gaussian convolutions and temporal Gabor filters [Dollár et al., 2005]. The authors state that the direct 3D counterparts to 2D detectors are inadequate and advocate an alternative approach. The Gabor filters give a local measurement focusing not only on local changes in the temporal domain, but prioritize repeated events of a fixed frequency. The function gives

$$R_{\sigma\tau\omega} = (I \otimes G_\sigma \otimes H_\tau^{ev})^2 + (I \otimes G_\sigma \otimes H_\tau^{od})^2 \tag{5.12}$$

where the 2D Gaussian smoothing is only applied in the spatial domain, whereas the two filters H^{ev} and H^{ov} are applied in the temporal domain only. H^{ev}, the *even filter* and H^{ov} the *odd filter* are the quadrature pair of 1D Gabor filters. The kernel are defined as

$$\begin{aligned} H_\tau^{ev} &= -\cos(2\pi\tau\omega)e^{-\frac{-t^2}{\tau^2}} \\ H_\tau^{ov} &= -\sin(2\pi\tau\omega)e^{-\frac{-t^2}{\tau^2}}. \end{aligned} \tag{5.13}$$



(a) Gabor kernel (b) Gabor filter in time and Gaussian kernel in space

Figure 5.5: The kernel of the Cuboid detector. (a) the two Gabor filter (b) the combination in time and space. From: [Niebles et al., 2006]

where $\omega = \frac{4}{\tau}$. The kernel is shown in Fig. 5.5. Fig. 5.5a gives the two Gabor filter. The filters are linearly separable. Fig. 5.5b gives the combination in time and space.

The authors suggest the use of a fixed scale approach. The final locations are found by extracting the maxima of R throughout the video. The set of functions is available on-line as a toolbox[2]. As suggested and used in previous evaluations, in the experiments $\sigma = 3$ and $\tau = 4$ are chosen.

The detector extracts features whenever variations in the image structure contains periodic motion. This approach is inspired by human and animal behavior analysis, where periodic actions are the most interesting: chewing, walking or a bird flapping its wings has periodic response and is thus more salient than translational motion. Nevertheless, spatio-temporal corners are detected with a lower priority than repeated temporal events. In the experiments in Chapter 6, this detector produces the most features and provides a significantly denser representation than the other detectors.

5.2 Color Based Spatio-Temporal Features

In 2005, the first important papers about local features for videos were published (e.g. [Dollár et al., 2005; Ke et al., 2005; Laptev, 2005]) and there is increasing research to be observed in the field of local feature based video matching. In this active community there is no interest in color-based local features so far. As for image features, video features aim for robustness against lighting and shadowing effects. In videos, change of lighting can be very strong on the same object. Imagine a person walking or driving along a boulevard with trees: The object stays the same, whereas the shadows on the object change quickly over time.

[2] `http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html`

In the worst case, these changes of lighting are considered as more salient than the actual object. Then, the shadowing effects are described and matched. In this scenario, incorporating color would increase the stability of the locations significantly. An extension to color features for video would make sense and can be achieved in a straightforward way. The main problems that have to be solved in the future are the computational costs: With growing number of dimensions that have to be processed whole extracting the features, the memory consumption and demand of computation rises exponentially.

In the following two promising extensions for incorporating color for spatio-temporal local features are presented.

5.2.1 Corner Detection

An extension of the Harris3D detector can be carried out using the approach of [van de Weijer and Gevers, 2005]. As stated before, the first step is to determine the gradients of each component of the RGB color system. The values have to be averaged by a Gaussian integration kernel with size σ. For the spatial-temporal case this thesis proposes to extend the elements to

$$\begin{aligned} L^2_{x,\sigma} &= \sum_{i=1}^{n} c_{i,x,\sigma}, \\ L_{xy,\sigma} &= \sum_{i=1}^{n} c_{i,x,\sigma} c_{i,y,\sigma}, \\ L^2_{y,\sigma} &= \sum_{i=1}^{n} c^2_{i,y,\sigma}, \\ L_{xt,\sigma} &= \sum_{i=1}^{n} c_{i,x,\sigma} c_{i,y,\sigma} c_{i,t,\sigma}, \\ L_{yt,\sigma} &= \sum_{i=1}^{n} c_{i,x,\sigma} c_{i,y,\sigma} c_{i,t,\sigma}, \end{aligned} \tag{5.14}$$

where an arbitrary color space C is used with its n components $[c_1, \ldots, c_n]^T$. $c_{i,x,\sigma}$ and $c_{i,y,\sigma}$ denote the respective components of the transformed color channel gradients at scale σ. The subscripts x and y indicate the direction of the gradient. It can be seen as an extension of the color based extension for the spatial case given in Eq. 3.27.

With this definition, the corner measure can be calculated in any color space and in the temporal domain at the same time. Applying Eq. 5.3 for the structure tensor and Eq. 5.5 to get the Harris energy in the scale σ gives a corner measure in any color space.

Scale Decision

As an extension for the saliency measure for deciding on the scale of a corner, the most straightforward way is to use Eq. 5.6 with the elements from Eq. 5.14. With this substitution a spatio-temporal extension to the Color Harris Laplacian from [Vigo et al., 2010] is achieved.

Estimating a global saliency measure for the input data as proposed in Section 3.2.2, the approach can be extended to the temporal domain in a straightforward way. Similar to the 2D counterpart, the PCA can be applied to reduce the dimensionality of the data. Research question will arise in the future if this representation holds for such high-dimensional data and makes sense in this context. The principal components of the color information of Video V_C are computed.

The scatter matrix is given by

$$S_t = \sum_{j=1}^{m} (\mathbf{c}_j - \overline{\mathbf{c}})(\mathbf{c}_j - \overline{\mathbf{c}})^T. \tag{5.15}$$

The dimensions of the information are reduced by taking the dot product of the color information I and the corresponding principal eigenvector ν_λ

$$\widehat{V} = \nu_\lambda V_C^T, \tag{5.16}$$

where $\overline{\mathbf{c}}$ denotes the sample mean, m is the number of pixels in the image and one sample $\mathbf{c}_j$ is a color vector of the pixel with index j. $\widehat{V}$ can now be treated as a single channel video and the scale selection methods described in the previous section can be applied (compare Eq. 5.6).

5.2.2 Blob Detection

Eq. 5.14 defines the elements of the spatio-temporal second moment Matrix Γ in any color space and any dimensionality. This allows to extend the most successful spatio-temporal blob detectors in a straightforward way.

As already stated in the previous Section 5.2.1, the extension to the **Color LoG3D** is to use Eq. 5.6 substituted with the elements from Eq. 5.14. Building a scale space of derivatives and using the scale normalized approximation of the LoG, an efficient scale invariant blob detector is developed. It can be processed on arbitrary color spaces and high dimensional data.

The same concept can be applied to develop the **Color Hessian3D**. The detector regards blobs and corners as salient points. It is possible to substitute the way the gradients are found in the same manner as the Color LoG3D: The Hessian features can then be extracted in the described way given in Eq. 5.7. This allows for stable features to be estimated without losing e.g. the chromatic information at the transformation from color to luminance. This will allow for more stable features to lighting variation than their luminance based counterparts are. The elements of the color structure tensor are built on the basis of the transformed derivative of the color information developed in Eq. 5.14 and summed up per direction.

There are two extensions of the MSER approach. The Maximally Stable Volumes (MSV) [Donoser and Bischof, 2006] and the extension of the image features to color [Donoser et al., 2006]. These two approaches can be combined. Similarly as given in Section 3.2.3, the three dimensional feature space of colors can be modeled in a multivariate Gaussian distribution. To estimate a meaningful distance between two color pixels, the distribution is fitted for every pixel within an initialized region. To compare two distances, the Bhattacharyya distance given in Eq. 3.38 can be used for the local weights of the connected graph. Then, the regions can be extracted as described in Section 5.1.2.

To conclude, the community has not even started to develop color based spatio-temporal features. Observing the development of image matching in the last decade (e.g. from [Smeulders et al., 2000] to [Mikolajczyk et al., 2009]), a trend towards efficiency and large scale application is seen (e.g. [Chum et al., 2009; Wu et al., 2009]). It is very likely that we will observe the same for spatio-temporal features. As stated in various publications, color improves the performance of image matching (e.g. [Stöttinger et al., 2009b; van de Sande et al., 2009; Vigo et al., 2010]). As shown in [Stöttinger et al., 2010a], there is a strong connection between the behavior of image features and their 3D counterparts. Therefore the focus of future work will be to incorporate color into spatio-temporal features.

5.3 Summary

This chapter gives a state-of-the-art on spatio-temporal feature detection. Similar to the previous chapter, it is divided into luminance and color based approaches. For the luminance based approaches, corner and blob detectors build a straightforward extension of the previously described approaches in 2D. Gabor filter build a new concept for feature detection focusing on repeated events in videos. This is a very promising approach, which lets much room for improvement: The spatial structure does not influence the features significantly in the original approach. A combination of the advantages of a stable spatial detection and the Gabor filtering in the temporal domain might give improved performance in terms of robustness and stability.

Color based approaches are not proposed to the community so far. The chapter gives several theoretical models which could provide for improved detection in this very fresh field of computer vision. The main detectors are mathematically extended to use color information. Future work will include the development of more robust, color based spatio-temporal features.

CHAPTER 6

Evaluation of Interest Point Detectors for Videos

In this chapter, the evaluation of spatio-temporal features is carried out. FeEval is explained in detail, a data-set for the evaluation of such features. For the first time, this data-set allows for a systematic measurement of the stability and the invariance of local features in videos.

The evaluation of detectors and descriptors is divided into two independent tasks. Following [Wang et al., 2009], the best performing approaches for detection, namely Harris3D, Hessian3D, the Cuboid detector are chosen. For descriptors HOG/HOF, SURF3D (also referred to as *extended SURF*), and HOG3D are used for the evaluation on videos. The same parameters and the same implementations are chosen.

In the following section, an overview of popular video matching data-sets is given, with a more detailed description of the FeEval data-set in Section 6.1.2. In Section 6.2, properties and behavior of the evaluated features are given in detail. Detector robustness is evaluated in Section 6.3, followed by the large scale video matching experiments in Section 6.4.

6.1 Data-Sets

In this section we give an overview of existing action recognition data-sets. In the following, popular data-sets used in current feature evaluations are presented. In Section 6.1.2 the proposed data-set is described in detail.

6.1.1 Popular Video Data-sets

KTH actions data-set

The **KTH actions data-set** [Schüldt et al., 2004][1] provides videos of six human action classes: walking, jogging, running, boxing, waving, and clapping. Each action class is performed repeat-

[1] http://www.nada.kth.se/cvap/actions/

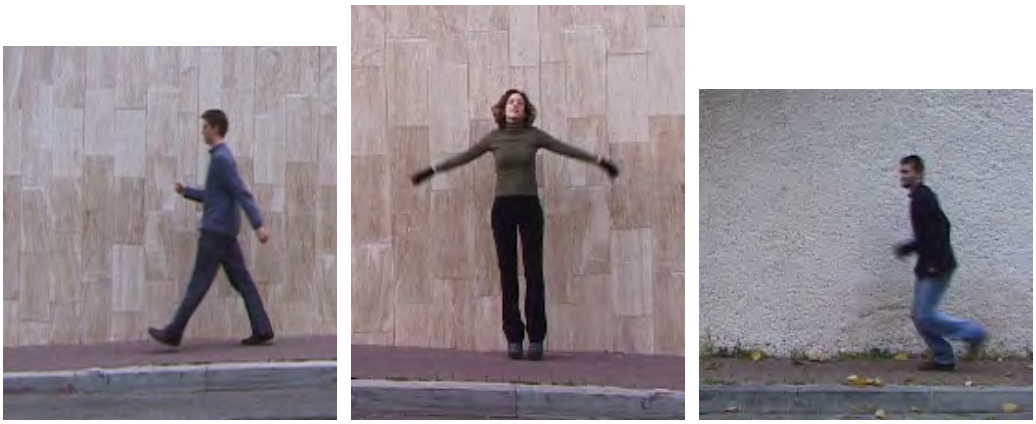
Figure 6.1: Example frames from the KTH actions data-set. From: [Schüldt et al., 2004].

edly by 25 persons. Alterations are defined on the videos as seen in Fig. 6.1: For every one of the six action classes, four different scenarios are recorded. *S1* provides videos recorded outdoors with one person performing the according action. *S2* records the same action with a different view-point, providing a scale variation. They are denoted as outdoors, outdoors with scale variation, outdoors with different clothes and indoors. All the resulting 2391 greyscale sequences were taken in front of homogeneous backgrounds with a static camera with 25fps frame rate at a resolution of 160×120 pixels and have a length of 4 seconds on average.

S3 and *s4* provide similar videos to *s1* but with the same persons wearing different clothes. All sequences are divided with respect to the subjects into a training set (8 persons), a validation set (8 persons) and a test set (9 persons).

The KTH data-set was released in 2004 and has become a popular data-set in the community providing a baseline for new approaches. Evaluations are for example given in [Schüldt et al., 2004; Dollár et al., 2005; Jhuang et al., 2007; Wong and Cipolla, 2007; Kläser et al., 2008; Laptev et al., 2008; Willems et al., 2008; Wang et al., 2009].

State-of-the-art approaches solve the standard challenge almost perfectly: The actions vary in their speed and in their spatial distribution of movements throughout the videos. Therefore the actions walking and boxing are well classified by most of the approaches. Jogging and running provides typically the most incorrect classifications.



(a) walking (b) jumping jack (c) running

Figure 6.2: Example frames from the Weizmann dat-set. From [Gorelick et al., 2007].

Weizmann Data-Set

The **Weizmann data-set** [Gorelick et al., 2007][2] provides 50 videos of nine people at a resolution of 180x144 pixels at 50 frames per second. Action classes are run, walk, skip, jumping jack, jump forward on two legs, jump in place on two legs, gallop sideways, wave with hands and wave one hand. Perfect retrieval results are obtained from various authors (e.g. [Blank et al., 2005]). Example frames are given in Fig. 6.2.

The data-set has been set up to treat the persons as silhouettes of a moving torso and protruding limbs undergoing articulated motion. For this scenario of limited variations and stable background, local spatio-temporal features seem to perform worse than specialized features as the *Poisson features*. Theses features give a function of the masked silhouette of the person. Additional features performing well in this evaluation are space-time saliency, *plateness* and *stickness* [Gorelick et al., 2007].

Recent publications [Junejo et al., 2010] show that perfect results are still not feasible for local spatio-temporal features. General approaches perform with an accuracy typically over 90% but with some wrong classifications.

UCF sport actions data-set

The **UCF sport actions data-set** [Rodriguez et al., 2008][3] contains ten different types of human actions with a great intra-class variety: swinging (on bar, pommel horse, floor), golf swinging, walking, diving, weight-lifting, horse-riding, running, skateboarding and kicking. It provides

[2] `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`

[3] `http://www.cs.ucf.edu/vision/public_html/`

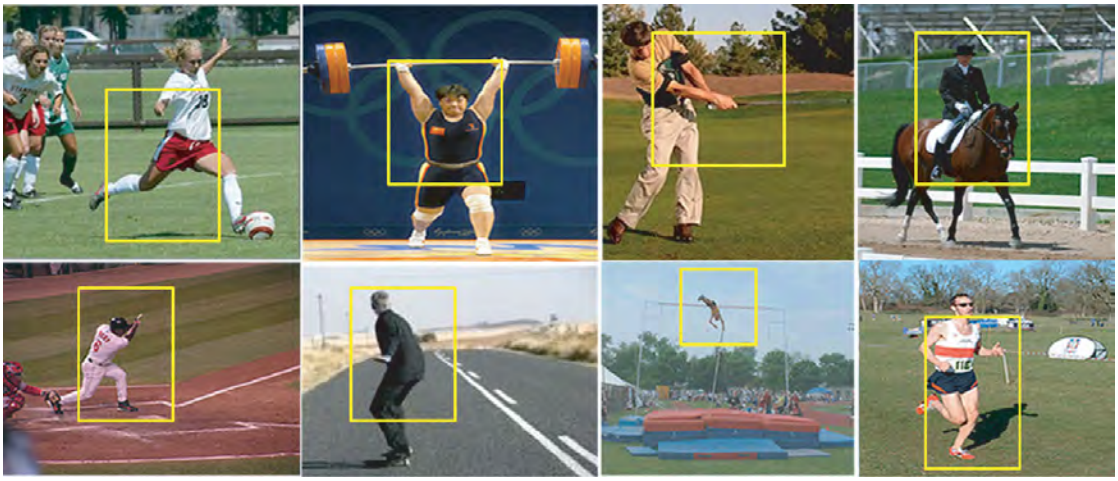


Figure 6.3: Example frames from the UFC sport actions data-set. From: [Rodriguez et al., 2008].

(a) eating (b) running (c) eating

Figure 6.4: Example frames from the Hollywood2 actions data-set. From [Laptev et al., 2008].

200 video sequences at a resolution of 720×480 pixels. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints.

Classification results for this data-set are lower than for the data-sets presented before. The authors give an overall mean accuracy of 69,2% as a baseline. Example frames and the annotated actions are seen in Fig. 6.3.

Hollywood2 actions data-set

The **Hollywood2 actions data-set** [Marszalek et al., 2009][4] has been collected from 69 different Hollywood movies. There are 12 action classes: answering the phone, driving a car, eating, fighting, getting out of a car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. There are 69 movies divided into a training set (33 movies) and a test set (36 movies) resulting in a total of 3669 sequences. Example frames are shown in Fig. 6.4

[4] `http://pascal.inrialpes.fr/hollywood2/`

Action samples are collected by means of automatic script-to-video alignment in combination with text-based script classification [Laptev et al., 2008]. Video samples that are generated from training movies correspond to the automatic training subset with noisy action labels. Based on this subset the authors constructed a clean training subset with action labels manually verified to be correct. There is a test subset with manually checked action labels available.

6.1.2 FeEval

The proposed data-set *FeEval* consists of 30 videos from HD TV shows, 1080p HD Hollywood movies of a resolution of 1920×1080 pixels per frame, and surveillance videos. Every video undergoes 8 transformations with successive impact, denoted as challenges. This leads to a data-set of 1710 videos each of about 20 seconds. All videos are encoded with the H.264 codec and stored in a .mov Quicktime container. The whole data-set has a size of about 34 Gigabytes (GB) and is available online[5].

10 videos are taken from two long running TV shows. An example is given in Fig. 6.6. The challenges are visualized in Fig. 6.5. Using TV show material has several advantages: It enables access to a vast amount of video content of a manageable group of people (the TV show cast) over the time of several years. Additionally, the actors also appear in other shows and movies, making large scale person detection and recognition experiments possible. Surveillance videos show 3 different persons in a calibrated environment. The persons enter the lab, fall onto the floor, get up and leave the scene again. Every scene is taken from 4 different angles, the scene is calibrated following [Svoboda et al., 2005] (see Fig. 6.7). The homography is available.

The 1080p HD movies are challenging because of their high resolution of 1920×1080 pixels and therefore the high demand of memory and processing power becomes an issue. An example is given in Fig. 6.8. Run-time and scale invariance of spatio-temporal features can be evaluated on the state-of-the-art of the home entertainment formats.

Every challenge consists of 7 levels. An overview is given in Tbl. 6.1. Geometric robustness of a feature is measured by estimating the repeatability of the features from the original video compared to the videos of the challenge. Description robustness is evaluated by the matching performance throughout a challenge [Mikolajczyk and Schmid, 2004].

- The **Gaussian blur** challenge applies increasing Gaussian blur per color channel. The kernel size is increased by 3 pixels at every level, beginning with a size of 3 pixels leading to 21 pixels for the 7th level.
- **Noise** adds random values to the video. Beginning with 5% noise in every frame, the challenge increases the amount of noise for every step by 5% up to 35%. At this point, more than a third of the original data is lost.

[5] `http://www.feeval.org`

(a) 1-blur-1.mov (b) 1-blur-4.mov (c) 1-blur-7.mov

(d) 1-noise-1.mov (e) 1-noise-4.mov (f) 1-noise-7.mov

(g) 1-dark-1.mov (h) 1-dark-4.mov (i) 1-dark-7.mov

(j) 1-light-1.mov (k) 1-light-4.mov (l) 1-light-7.mov

(m) 1-median-1.mov (n) 1-median-4.mov (o) 1-median-7.mov

(p) 1-comp-1.mov (q) 1-comp-4.mov (r) 1-comp-7.mov

(s) 1-scalerot-1.mov (t) 1-scalerot-4.mov (u) 1-scalerot-7.mov

Figure 6.5: Overview of *FeEval* data-set derived from original video 1.

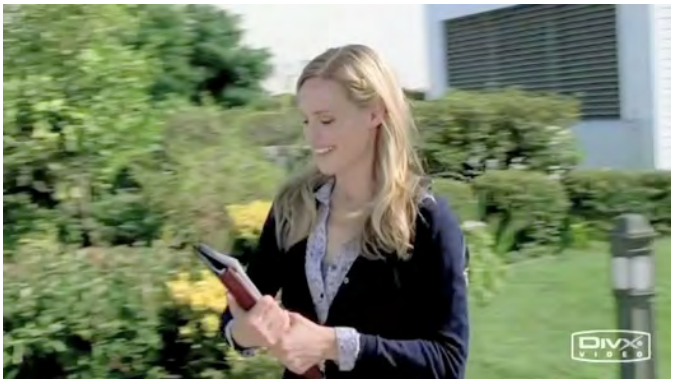


Figure 6.6: FeEval video 1, 624×352 HDTV show.

(a) 11.mov (b) 12.mov (c) 13.mov (d) 14.mov

Figure 6.7: Calibrated scene from 4 view points, homography is known.

Transformation	**Abbreviation**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Gauss σ in pixel	blur	3	6	9	12	15	18	21
Noise in %	noise	5	10	15	20	25	30	35
Darken: Lightness in %	dark	-30	-40	-50	-60	-70	-80	-90
Lighten: Lightness in %	light	30	40	50	60	70	80	90
Median Filter σ in pixel	median	2	3	4	5	6	7	8
H.264 quality	comp	60	50	40	30	20	10	0
Scale + Rotation in degree	scalerot	90% + 10°	80% + 20°	70% + 30°	60% + 40°	50% + 50°	40% + 60°	30% + 70°
Frames per Second	fps	20	15	13	10	7	5	3

Table 6.1: Video transformations for each of the 30 videos. Filename convention: "[number of video]-abbreviation-[number of column].mov"

- **Change of lighting** The videos are darkened and lightened by changing the lightness of the colors to simulate increasing and decreasing lighting conditions. The change of lighting is applied from ± 30% to ± 90% of the original lightness of the color pixels.
- The **median filter** is used to reduce speckle noise and salt and pepper noise effectively. The filter is applied with a kernel size from 2 pixels to 8 pixels.
- To test the effect of **increasing compression**, the H.264 quality is decreased from 60 to 0 leading to a video with strong JPEG artifacts and many wrong colors and edges.
- For evaluation of the invariance to **scale and rotation** the videos are increasingly shrunk to

Figure 6.8: FeEval video 27, 1080p HD movie.

a final size of 30% of the original size and rotated by 10% for every level. The homography matrices are straightforward to estimate and given on the web-page.

- To decrease the demand for storage space, surveillance videos are often handled with very few **frames per second**. For the challenge, the original 24 frames per second are reduced down to 3 frames per second.

With these definitions, a data-set to evaluate the robustness and invariance of spatial features against 8 challenges is built. The challenges of increasing blur, noise, change of lighting, median filtering, compression, scale and rotation and frames per second are inspired by previous evaluation of 2D features. A discussion of this framework is given in Section 4.5. For the first time, in a standardized way altered data is available for the evaluation of spatio-temporal features. For geometric transformations, homography matrices are provided. Furthermore, the videos have overlapping cast making it possible to evaluate action and person recognition under increasing transformation of the videos.

In contrast to existing data-sets, FeEval consists of videos of varying sources from surveillance cameras to high resolution 1080p HD movies. All of the videos are in color and display a large variety of persons, surroundings and lighting conditions. With this data-set of 1710 annotated videos, one problem of prior evaluation of spatio-temporal features is addressed: It allows for a principled evaluation on generalized data by measuring the geometric repeatability and the description robustness against well defined challenges.

Tbl. 6.2 shows statistics of the data-sets described in this section. It is shown that compared to other data-sets, FeEval provides challenging scenarios with many classes and is the first data-set with *full HD* videos of 1920×1080.

Name	# of videos	Classes	Instances	Resolution	year released
KTH action	2391	6 actions	25 persons	160×120	2004
Weizmann	90	10	9 persons	180×144	2007
UCF sport	182	20	15-35 scenes	720×480	2008
Hollywood	3669	12/10	n.a.	variable	2009
FeEval	1710	30/8	8x7	up to 1920×1080	2010

Table 6.2: Comparison to popular data-sets.

6.2 Feature Behavior

The experiments in this section aim to quantify the properties of the state-of-the-art spatio-temporal features described in Chapter 5. All experiments are carried out on the FeEval database described in Section 6.1.2. This section observes the features and their change over the challenges. It is an in-depth analysis of the representations themselves, not an evaluation of their robustness or matching performance. With these observations, a better understanding of the performance in the following experiments is achieved.

The approaches are chosen as follows. For the descriptors, the Harris3D (Section 5.1.1), Hessian3D (Section 5.1.2) and the Cuboid detector (Section 5.1.3) are used. Description of the extracted volumes is done by the HOG/HOF, SURF3D and HOG3D (all given in Section 2.4.2) is used. Throughout the experiments, the original implementations of the authors are used with the standard parameters. As stated before, these are the choices of [Wang et al., 2009].

Unfortunately, the implementation of Harris3D is only able to handle videos with a maximum resolution of VGA (640×480). The implementation of Hessian3D produces errors for about 30% videos with that high resolution. Therefore these videos had to be scaled down.

Every challenge starts with the original video which is then increasingly transformed. It is observed how the representation of detections changes for every transformation. This does not only give insight into the robustness of the feature but answers also the questions in the opposite direction: How can I alter my video in terms of noise reduction, compression and reduction of resolution and frames while being equally represented? What kind of videos do I have to provide to allow for a meaningful representation?

First, the *number of detections* and their *relative coverage* in the challenges is observed. The relative coverage $\mathcal{C}$ for a video V and its features $\gamma_{1..N}$ with the number of features N is defined as

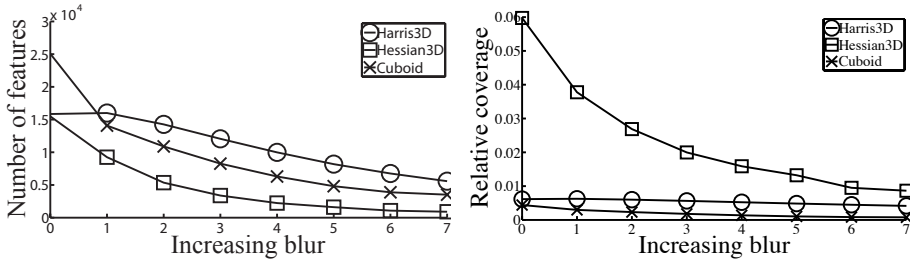
$$\mathcal{C} = \frac{\bigcup_{i=1}^{N} \mathcal{V}(\gamma_i)}{\mathcal{V}(V)}. \tag{6.1}$$

where $\mathcal{V}(V)$ is the volume of the video (*width* $\times$ *height* $\times$ *number of frames*) and $\mathcal{V}(\gamma)$ the volume of feature γ.

In relation to the number of detections, this gives an idea about the sizes of the extracted patches and thus within the challenge the robustness of the scale selection. For single-scale approaches, this measure is directly related to the number of features.

	Harris3D HOG/HOF	Hessian3D SURF3D	Cuboid HOG3D
mean number of features	13500	12400	51505
st. dev. of number of features	20445	19608	99534
maximum number of features	266706	96887	1033830
mean relative coverage	0,59%	4,8%	1,1%
st. dev. relative coverage	0.76%	3,9%	2,2%
mean descriptor entropy	6,78	4,14	2,39
st. dev. descriptor entropy	0,17	0,25	0,45

Table 6.3: Statistics of detections of FeEval.



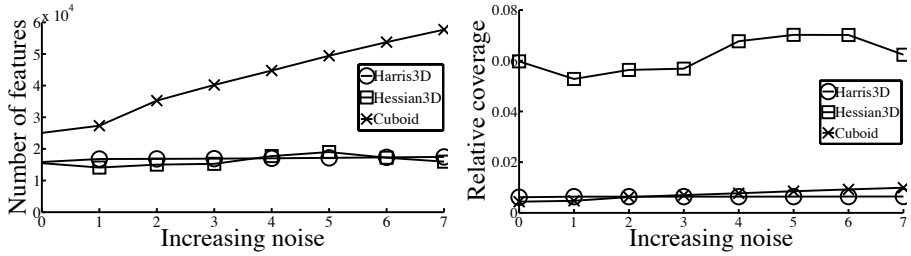
(a) Number of features (b) Relative coverage

Figure 6.9: Number of features (a) and their relative coverage (b) under increasing blur.

Throughout all the experiments it is observed that Harris3D and Hessian3D are very similar in their number of features. Tbl. 6.3 shows that the corresponding mean numbers of features and the standard deviations are comparable for these two methods. Still, the maximum number of features is 266706 for the Hessian3D and 96887 for the Harris3D. However, Hessian3D provides almost 10 times more relative coverage than Harris3D. The Cuboid detector gives a much higher number of small features on a single scale. In the following, the results per challenge are discussed.

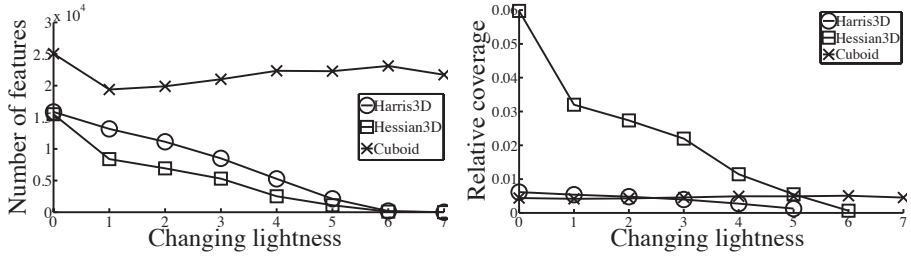
The **Gaussian blur** challenge applies increasing Gaussian blur per color channel. The kernel size is increased by 3 pixels at every level. Increasing Gaussian convolution can also be seen as down-scaling of the videos. The number of features decreases for all approaches linearly with increasing size of the Gaussian kernel (see Fig. 6.9a). This is reasonable and does not imply that the detectors suffer from instability against blur. The relative coverage of the features on the other hand (Fig. 6.9b) shows that the Harris3D is able to maintain its coverage (providing fewer, but larger regions) whereas the Hessian3D detector loses bigger blobs.

Noise adds random values to the video. Beginning with 5% noise in every frame, the challenge increases the amount of noise for every step by 5% up to 35%. At this point, more than a third of the original data is lost. The corner and blob detection extract their locations from



(a) Number of features (b) Relative Coverage

Figure 6.10: Number of features (a) and their relative coverage (b) under increasing noise.



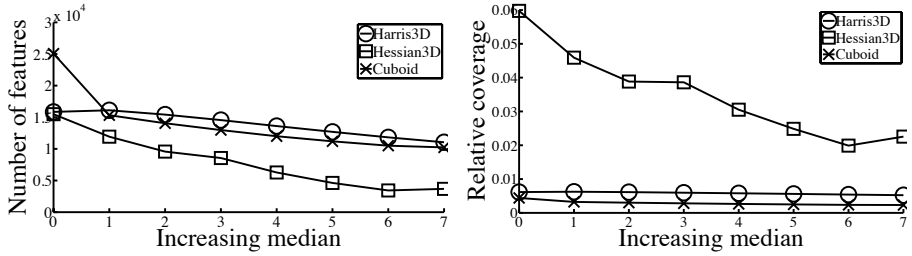
(a) Number of features (b) Relative Coverage

Figure 6.11: Number of features (a) and their relative coverage (b) under increasing change of lightness.

local structure tensors which are robust against noise. For the Cuboid detector, the temporal Gabor filter shows to be highly sensitive to noise (see Fig. 6.10a). It can be seen that the number of detections doubles within the challenge, leading to a linear increase of coverage. Harris3D maintains its coverage not changing the scales significantly (see Fig. 6.10b). Hessian3D becomes unstable in its blob detection giving varying coverage showing that the scale detection is less robust against noise.

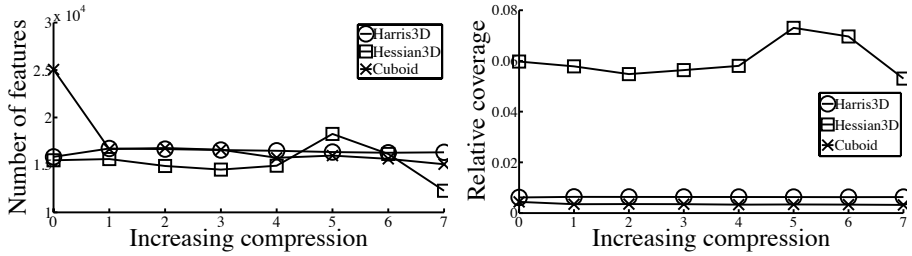
Change of lightness Videos are darkened and lightened by changing the lightness of the colors to simulate increasing and decreasing lighting conditions. Results are given in Fig. 6.11. Both Harris3D and Hessian3D suffer from decreasing contrast in the image. Both approaches are not able to detect any features in any of the very dark or light videos. As for blur, the Hessian3D is less robust to contrast change losing more relative coverage than the Harris3D. The Cuboid detector is not affected by the change of the light.

The **median filter** is often used to reduce speckle and salt-and-pepper noise effectively. We apply the filter with a kernel size from 2 pixels to 8 pixels. Visually, the videos are changed a lot at that point. The filter removes all fine structure and noise. Surprisingly, this does not affect the Harris3D detector. Its locations remain stable throughout the challenge (see Fig. 6.12).



(a) Number of features (b) Relative Coverage

Figure 6.12: Number of features (a) and their relative coverage (b) under increasing median filtering.



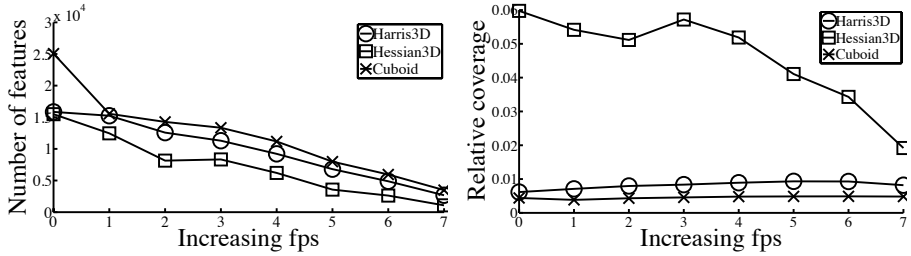
(a) Number of features (b) Relative coverage

Figure 6.13: Number of features (a) and their relative coverage (b) under increasing H264 compression.

To test the effect of **increasing compression**, the H.264 quality is decreased from 60 to 0 leading to a video with strong JPEG artifacts and many wrong colors and edges. The space requirements for the actual file are reduced to less than 25% from the high quality compression. All the detectors are stable to this condition where blob detection of the Hessian3D has the most variation. Cuboid detections change more in the initial step from no compression to the first step of the challenge as even slightest JPEG artifacts respond to the Gabor filtering (see Fig. 6.13a and 6.13b). But after this step, the performance remains constant.

To decrease the demand for storage space, surveillance videos are often handled with very few **frames per second**. For the challenge, the original 24 frames per second are reduced to 3 frames per second. All approaches provide a more sparse representation with fewer frames. Harris3D increases the relative coverage with fewer frames. Hessian3D is less robust to this change of data losing half of the relative coverage (see Fig. 6.14).

The descriptors are observed in their difference of information content over the whole data-



(a) fps (b) fps

Figure 6.14: Number of features in relation to their relative coverage under decreasing frames per second.

	Harris3D	Hessian3D	Cuboid
HOG/HOF (162 dims, text files)	30 GB	-	-
SURF3D (288 dims, binary files)	-	21 GB	-
HOG3D (960 dims, binary files)	17 GB	45 GB	**419 GB**

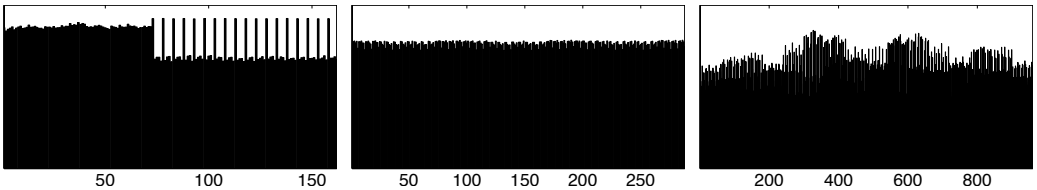
Table 6.4: Storage space in GB of the features on the 34 GB of compressed video data.

set. The entropy ϵ of descriptor d is measured by

$$\epsilon_d = \sum_{i=1}^{n} -\log(p(x_i))p(x_i) \tag{6.2}$$

where $p(x_i)$ is the occurrence probability for the value x_i within the n-dimensional descriptor d. The concatenated HOG/HOF descriptor with 162 dimensions has the highest entropy with the smallest standard deviation throughout the whole data-set. The HOG3D descriptor varies the most over the challenges. The 960 dimensions are comparably sparse and the description varies the most over the data-set. This can be seen in Fig. 6.15 where the distribution of non-zero data per dimension is given over the whole data-set.

The disk space needed depends on (1) the dimensionality of the descriptor, (2) the way the detector is encoded, and (3) the number of features. The HOG/HOF descriptor is stored in text files which takes more space than the binary output of the other implementations. In Tbl. 6.4 it is shown that the 30 GB of Harris3D + HOG/HOF descriptor produces approximately 17 GB of Harris3D+HOG3D data, having the same detection but almost 6 times more description information stored in the files. As the Cuboid detector is single scale and very small scales are suggested, it detects many features especially on noisy videos and videos of large resolution.



(a) HOG/HOF (b) SURF3D (c) HOG3D

Figure 6.15: Distribution of occurrences of non-zero values per dimension within the descriptors over 1710 videos.

6.3 Robustness

To evaluate the robustness of the three detectors Harris3D, Hessian3D, and Cuboid, their robustness or *repeatability* for each altered video with respect to its corresponding original video is measured. Each of the 30 original videos is regarded as a *boolean* 3D volume V_{o_i}, $i = 1..30$, sized according to the frame resolution and the total number of frames.

$$V_{o_i} = \begin{cases} 1 & \text{voxel is being detected by a feature} \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

Each of the m detected features $\xi_{c,1..m}$ in an altered video defines a cuboid in space. Per repeatability test, the cuboid $\xi_{c,j}$ is mapped to V_o to get its position and expansion in the original video's volume V_o denoted as $\xi'_{c,j}$. This is done by applying its homography matrix Ω

$$V_o \leftarrow \Omega * V_c \tag{6.4}$$

For the challenge of scale and rotation, the provided "2D" matrices are used. They are defined by the parameters given in Tbl. 6.1, as the alteration is per frame only and does not affect the temporal configuration. For the challenge of decreasing frames per second, it is regarded as a simple sampling in the temporal direction. Overlap ϱ of feature j is then defined by

$$\varrho = \frac{V_o \cap \xi'_{c,j}}{\upsilon(\xi_{t,i})} \tag{6.5}$$

where $\upsilon(\xi_{t,i})$ is the volume of the transformed feature's cuboid. The crucial parameter is the choice of the threshold when one feature is measured as repeated. The mean results on varying ϱ are given in Fig. 6.16. Similar to prior evaluations the feature is regarded as repeated when the overlap is more than 60% of the feature's size for further evaluation. The final repeatability score of a video is defined by the number of matched features divided by the total number of features in the challenge video. This forms a repeatability measurement for spatio-temporal features similar to image features in [Mikolajczyk et al., 2005b] and extended to video in [Willems et al., 2008]. The size of the features is not unified beforehand. In [Mikolajczyk et al., 2005b] the size of the features is set to 40 pixels, in [Willems et al., 2008] it is not given. Generally, the unification prevents a bias towards larger regions. It is done by setting the volume of every ξ to

	Harris3D	Hessian3D	Cuboid
mean repeatability	0.49± 0.05	**0.56±0.08**	0.15 ± 0.01

Table 6.5: Average results ± standard deviation of the repeatability experiments.

a predefined value. It is not done in this evaluation as the approaches provide only a very limited number of different scales and diminishes the effect of different detectors. The experiment aims to evaluate the robustness of the scale selection as well, which would be disregarded using fixed volumes for repeatability estimation.

The Harris3D provides detection from 6 predefined spatial and 3 temporal scales, whereas the Hessian uses 3 octaves in both spatial and temporal direction. The Cuboid detector is a single scale approach which does not detect spatial structure at all. Detections depend on the temporal Gabor filtering, the spatial structure is only smoothed by Gaussian blur. This explains the overall performance of the repeatability experiments where on average the Hessian3D detector outperforms the Harris3D detector, whereas the Gabor detector shows to be significantly less robust. The single-scale Gabor detector is not much affected by the change of the overlap criterium, as the large number of small features tends to be matched almost perfectly or not at all. This is of course different for the multi-scale approaches Harris3D and Hessian3D, where different sizes of features are matched. The bigger the features become, the more likely it is that they do not match perfectly. Therefore, the overlap criterium has more effect.

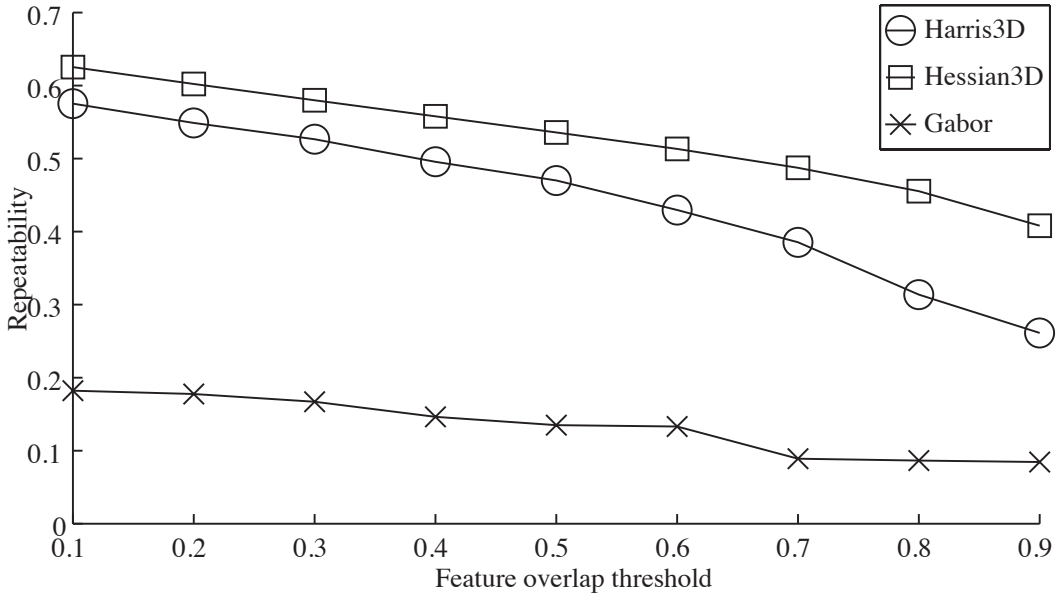


Figure 6.16: Mean repeatability results for the whole data-set over varying overlap ϱ.

Hessian3D has the best mean repeatability and performs best throughout the experiments. However, it provides a richer representation as its coverage is almost 10 times larger than Harris3D, thus making the probability for a geometrical match higher. Still, Harris3D performs comparably similar, which coincides closely to the evaluation of their 2D counterparts in [Mikolajczyk et al., 2005b]. In the following, the detailed results per challenge are given.

As it is shown in Fig. 6.17a, Harris3D and Hessian3D are almost equally robust to increasing blur. This also holds for increasing compression shown in Fig. 6.17b. The Cuboid detector is sensitive to Gaussian blur, which is reasonable as the approach only blurs in spatial dimensions, not using any derivatives. This holds also for increasing compression, where the Cuboid detector performs better, but also very sensitive. The two other detectors are very robust to increasing compression, showing similar results as evaluated on 2D images [Mikolajczyk et al., 2005b]. This is an important observation, since the spatio-temporal structure tensor has more degrees of freedom and might thus perform worse.

Hessian3D and Harris3D remain stable showing a repeatability around 0.60 throughout the challenge with up to 35% of noise in the video (Fig.6.17c). The Cuboid detector stays over 0.2 in the challenge. For increasing median filtering (Fig. 6.17d), Harris3D is equally robust as the Hessian3D for the first two images, then reducing the repeatability almost linearly. In contrary, Hessian3D increases repeatability from image 2 to 3 and 6 to 7 which is probably due to a bigger scales and thus bigger blobs being selected than for the Harris3D. The Cuboid detector is sensitive to median filtering.

In contrast to 2D detectors, the Harris3D and Hessian3D show to be very sensitive to change of lightness (see Fig. 6.17e and 6.17f). The number of features decreases rapidly with the decrease of contrast. This is the only challenge where the Gabor detector outperforms the other approaches in robustness at level 7.

The decrease of frames per second (see Fig. 6.17g) can be seen as scaling in the temporal domain. As the approaches are not scale invariant, they perform worse than their 2D counterparts. Hessian3D considers the most scales of the approaches evaluated, and remains stable until level 3, which is the reduction from 25fps to 13fps. Therefore the standard sampling rate of 2 for the Hessian3D approach can be easily set to 4 without a significant loss in performance, disregarding 50% of the data right away. For scale and rotation shown in Fig. 6.17h, Gabor and Harris perform poorly compared to the Hessian3D which is able to maintain a repeatability rate of 0.41 for a video scaled by a factor of 0.3 and rotated by 70°.

Based on these results, the following approach for dealing with noisy video data is proposed: Gaussian blur degrades the detections severely therefore it should not be used in pre-processing videos. Hessian3D on noise performs more robustly than on blurred data. Gabor detections are neither reliable on noisy nor on blurred data. When using the Harris3D detector, it is recommended to use the median filter to remove the noise in advance.

6.4 Video Matching

Research interest is to what extent state-of-the-art spatio-temporal descriptors maintain their robustness under alteration of their input videos. There is no one to one matching of local de-

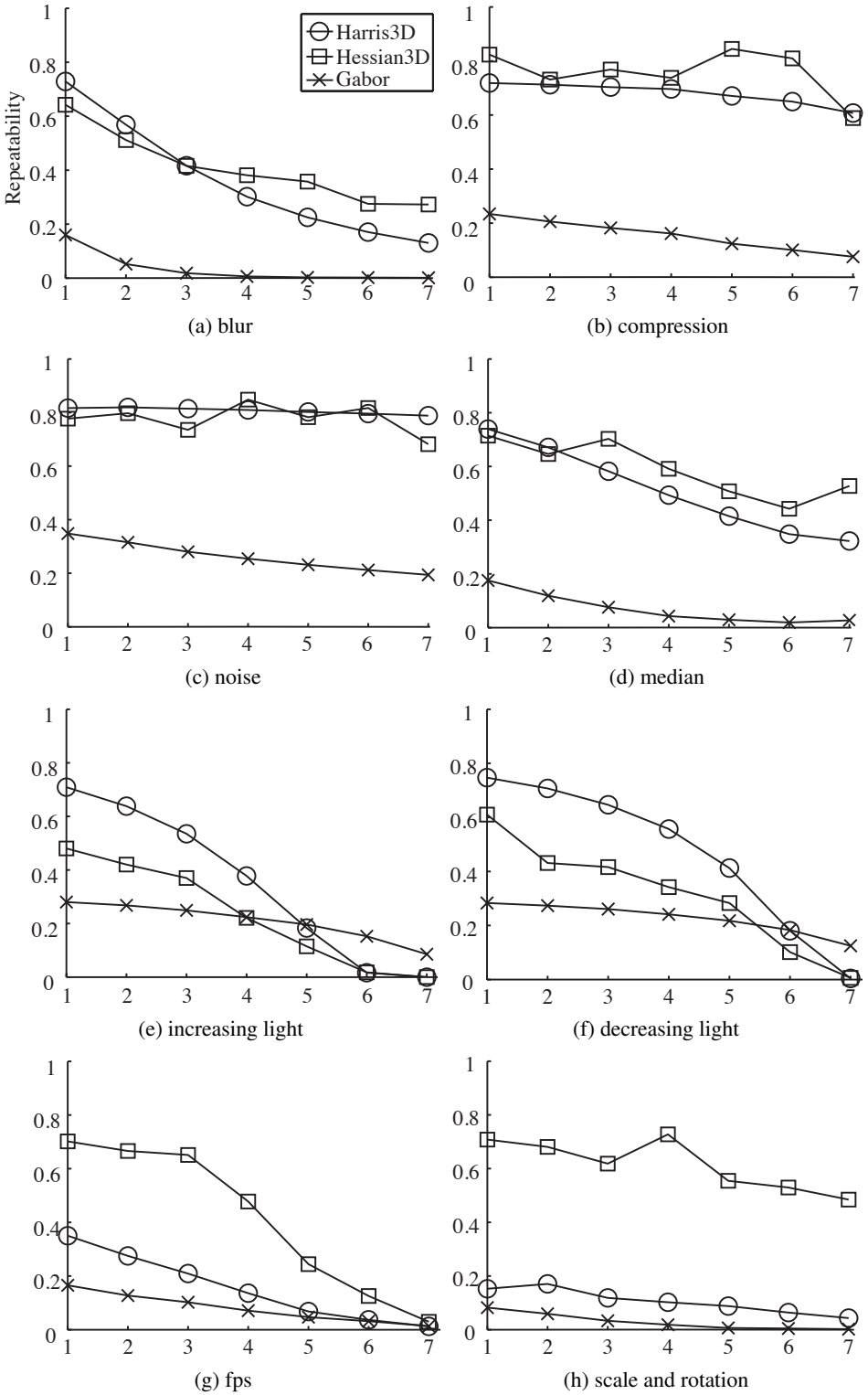


Figure 6.17: Mean repeatability ($\varrho = 0.6$) of 30 videos per challenge. Legend is shown in (a).

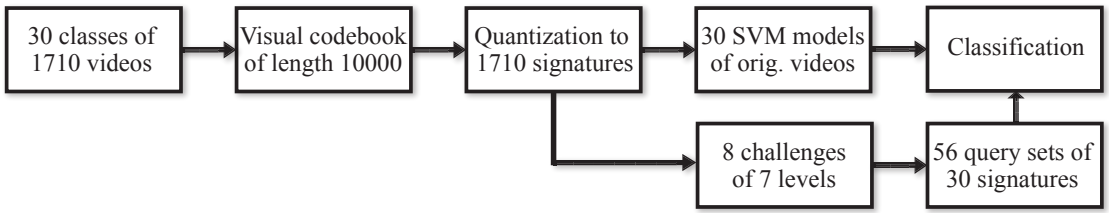


Figure 6.18: Experimental setup to test the description's robustness against visual alterations.

scriptors, as is usually done when evaluating local description (e.g. [Mikolajczyk and Schmid, 2005]). The aim is to test a descriptors' performance in a large scale video classification experiment where the training data consists of 30 original videos forming 30 classes of challenges. For the three descriptors HOG/HOF, SURF3D and HOG3D and the combination with the detectors the following set-up is carried out.

The flowchart in Fig. 6.18 gives the experimental workflow. In simple terms, the original videos build the training set and ground truth. For every altered video it is tested if it can be successfully matched with its original counterpart. The averaged results for this experiment are given in Tbl. 6.6, results per challenge and challenge step are given in Fig. 6.19 and Fig. 6.20.

One choice is the number of visual words to be used. [Wang et al., 2009] used 100000 random features to cluster 4000 visual words. Inspired by [Wang et al., 2008], a visual codebook of 10000 words is used to accommodate the approach to the large data-set and the great intra-class variations of the videos. It is formed by clustering all the features of the data-set with the *kshift* [Pönitz et al., 2010][6] algorithm. It is an iterative approximation of the k-means algorithm which has the same complexity, but needs only one iteration through the data-set for comparable results. In contrast to many other clustering implementations, the data-set can be larger than the memory. For every cluster center, it is only necessary to have the *next* feature in the memory, not the whole data-set. It is feasible to cluster 45 GB of 960 dimensional features within 20 hours using 2 X5560@2.8GHz processors (4 cores each).

A video's signature is built by quantizing its features to the codebook by the cluster center with the nearest Euclidean distance. For the training set, the 30 original videos with their normalized signatures of a length of 10000 each are used as ground truth classes. For every class, a linear one-against-all SVM model is trained equally weighting every class. For this set-up, the model is similar to a nearest neighbor classification.

The well known LibSVM library[7] is used with default parameters. For the 8 challenges with 7 levels, 56 test sets of equal size are built for the evaluation.

The experimental question is then until which alteration the description is still able to discriminate against the other videos and under which circumstances it fails. When an altered video is successfully classified as its original video, the description is regarded as robust to the alter-

[6] `http://www.cogvis.at`

[7] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

	Classification accuracy			Mean precision			Mean recall		
	Harris3D	Hessian3D	Gabor	Harris3D	Hessian3D	Gabor	Harris3D	Hessian3D	Gabor
HOG/HOF	23,57	-	-	19,40	-	-	23,57	-	-
SURF3D	-	39,52	-	-	40,46	-	-	44,80	-
HOG3D	**49,76**	37,96	34,75	**42,40**	38,80	28,15	**49,76**	42,20	35,30

Table 6.6: Overview of experimental results of descriptor evaluation.

ation. In this context, the classification performance according to the alterations gives then the descriptor robustness in the challenge.

The descriptors from [Wang et al., 2009] in combination with the detectors evaluated above are evaluated. Summary results are shown in Tbl. 6.6, where '-' denotes combinations that are not available as the author's implementation does not allow the input of other detectors. Results per challenge are shown in Fig. 6.19. In Fig. 6.20 results of the experiments using the HOG3D descriptor are given. The combination of Harris3D and HOG3D outperforms other approaches.

As already argued in the previous sections, Gaussian blur decreases the representation of the videos significantly. As seen in Fig. 6.19a, the classification accuracy approaches the prior probability of 3%. With these approaches, it is not feasible to match videos which are blurred by a Gaussian blur $\sigma > 3$ pixels. This is different for the HOG3D descriptor. For all detectors, there is a significant gain in classification performance, especially for the Harris3D+HOG3D raising to a mean accuracy of 54,76%. Fig. 6.20a shows that a more robust detector improves the representation of the video significantly.

Similar behavior is observed for change of lightness: For HOG/HOF and SURF3D, the classification accuracy goes down rapidly, whereas the HOG3D descriptor provides a stable description on data of varying contrast. Gabor+HOG3D outperforms these approaches in lighting changes (see Fig. 6.19e and 6.20e). When combining the detectors with HOG3D, a correlation with the repeatability experiments of changing lightness is observed. With a more stable descriptor, the more repeatable representation influences the classification performance: The most stable Harris3D outperforms other approaches until the level where the Gabor detector turns out to be more repeatable.

This does not hold for the fps challenge (see Fig. 6.19g and 6.20g). There is no correlation between detector robustness and classification performance. This suggests that none of the descriptors is scale invariant (in the temporal domain) to a satisfying extent. On the other hand, the loss of performance is coherent with the fact that for scaling most of the data to be described is lost. It can be deduced that for performance reasons, detectors can be applied on a reduced data-set but the local description has to be performed on full temporal resolution.

Descriptors are revealed to be more robust to increasing noise than the local detectors. The worst performing Harris3D+HOG/HOF reaches a mean accuracy of 51,43%. Hessian3D + SURF3D remains almost stable throughout the challenge (see Fig. 6.19c). HOG3D shows to be more robust than HOG/HOF (see Fig. 6.20c), but decreases the performance for the Hessian3D. It is shown that SURF3D is more robust to noise than HOG3D in this context.

Regarding noise reduction using the median filter (see Fig. 6.19d and 6.20d) performance decreases more compared to the noise challenge. HOG/HOF and HOG3D are sensitive to the

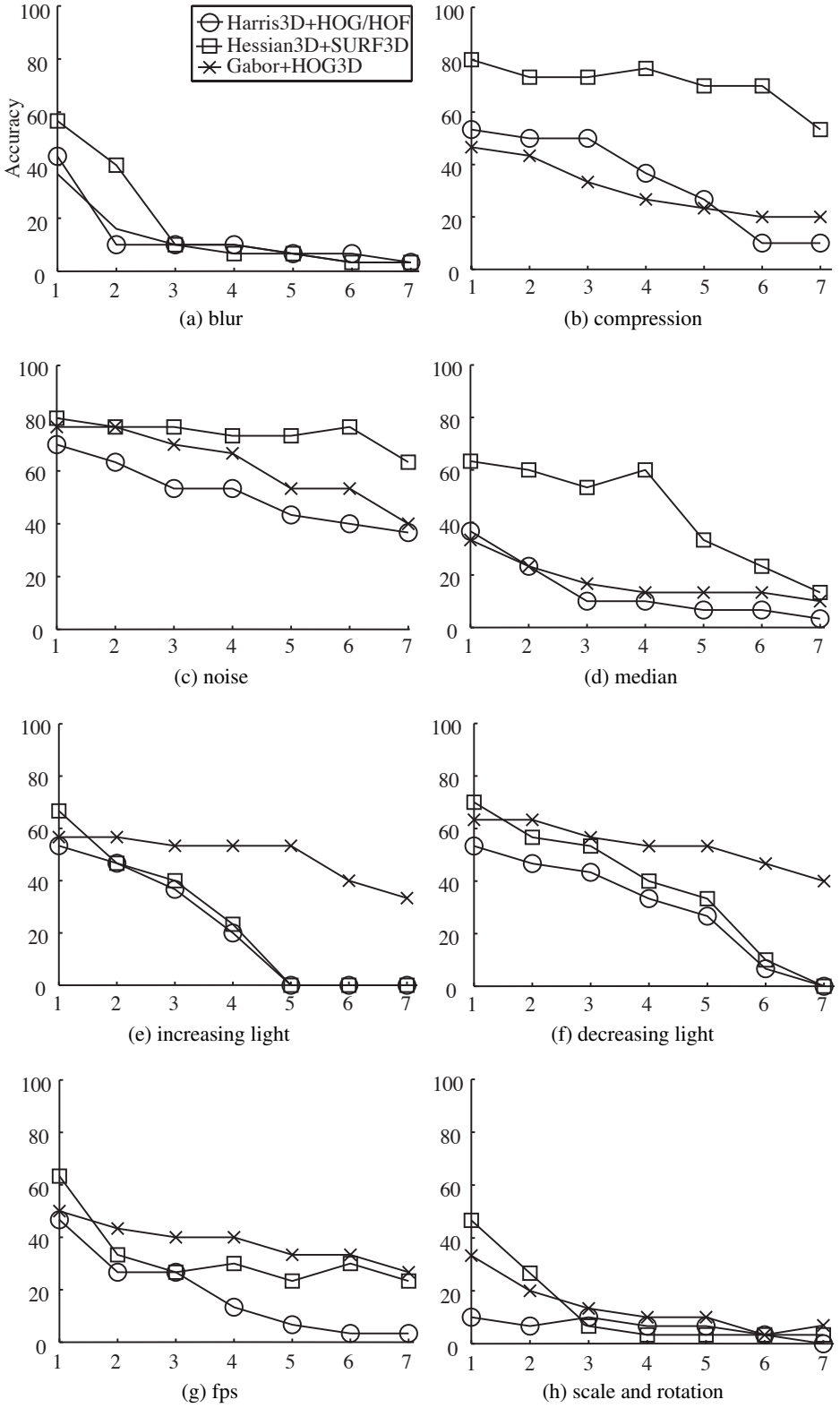


Figure 6.19: Classification accuracy per challenge. The legend is shown in (a).

	Detector Robustness			Descriptor Robustness		
	Harris3D	Hessian3D	Gabor	HOG/HOF	SURF3D	HOG3D
Gaussian blur	+/-	+/-	-	-	-	+/-
H.264 compression	+	+	-	-	+	+
Noise	-	+	-	+/-	+	+
Median Filter	+	+	-	-	+/-	+/-
Increasing lightness	+/-	+/-	+/-	-	-	+
Decreasing lightness	+/-	+/-	+/-	-	-	+
Frames per Second	-	+	-	+/-	+/-	+/-
Scale & Rotation	-	+	-	+/-	+/-	+/-

Table 6.7: Final suggestions based on the evaluation.

filtering, SURF3D performs similarly to the repeatability rate of its detector. Increasing compression does not affect the description performance of the HOG3D and the SURF3D descriptor. Even strong JPEG artifacts are described in a stable and discriminative way (see Fig. 6.19b and 6.20b). Considering that for level 7 of the challenge, the data is compressed up to 10% of the original file size, an additional compression for the classification task might help for certain applications, e.g. when the original data and the classification system are located apart from each other and data transfer is expensive or slow.

To visualize the essence of the evaluation, the results are shown categorized by simple votes according to the challenges: '-' denotes sensitivity, '+' robustness to the challenge. '+/-' refers to undecided decision or room for improvements in the algorithmic details of the approach. The votes are given in Tbl. 6.7.

6.5 Summary

In this section, the first comparative evaluation of spatio-temporal features using well defined visual challenges is carried out. Challenges are inspired by prior evaluation of local 2D image features. For detector robustness, comparable results for spatio-temporal features with their image counterparts are obtained. Generally, it showed to be worse to reduce noise in input data than to let the features take care of it on their own. For change of light, both the Harris3D and the Hessian3D are more sensitive than their 2D counterparts. Description is most stable using the HOG3D descriptor, outperformed by the SURF3D descriptor in the challenges of compression, noise and median filtering. The high dimensionality of the HOG3D descriptor of 960 compared to 288 of the SURF3D descriptor is a drawback in terms of the complexity of all succeeding operations and should be considered when choosing the most appropriate descriptor.

The authors of the evaluated approaches regard a full scale invariance of the detections as too computationally costly and suggest a limited multi-scale approach, or in the case of the Gabor detector, a single scale approach. Obviously, that leads to drawbacks challenging related alterations such as varying the number of frames per second and scaling down the video. Nev-

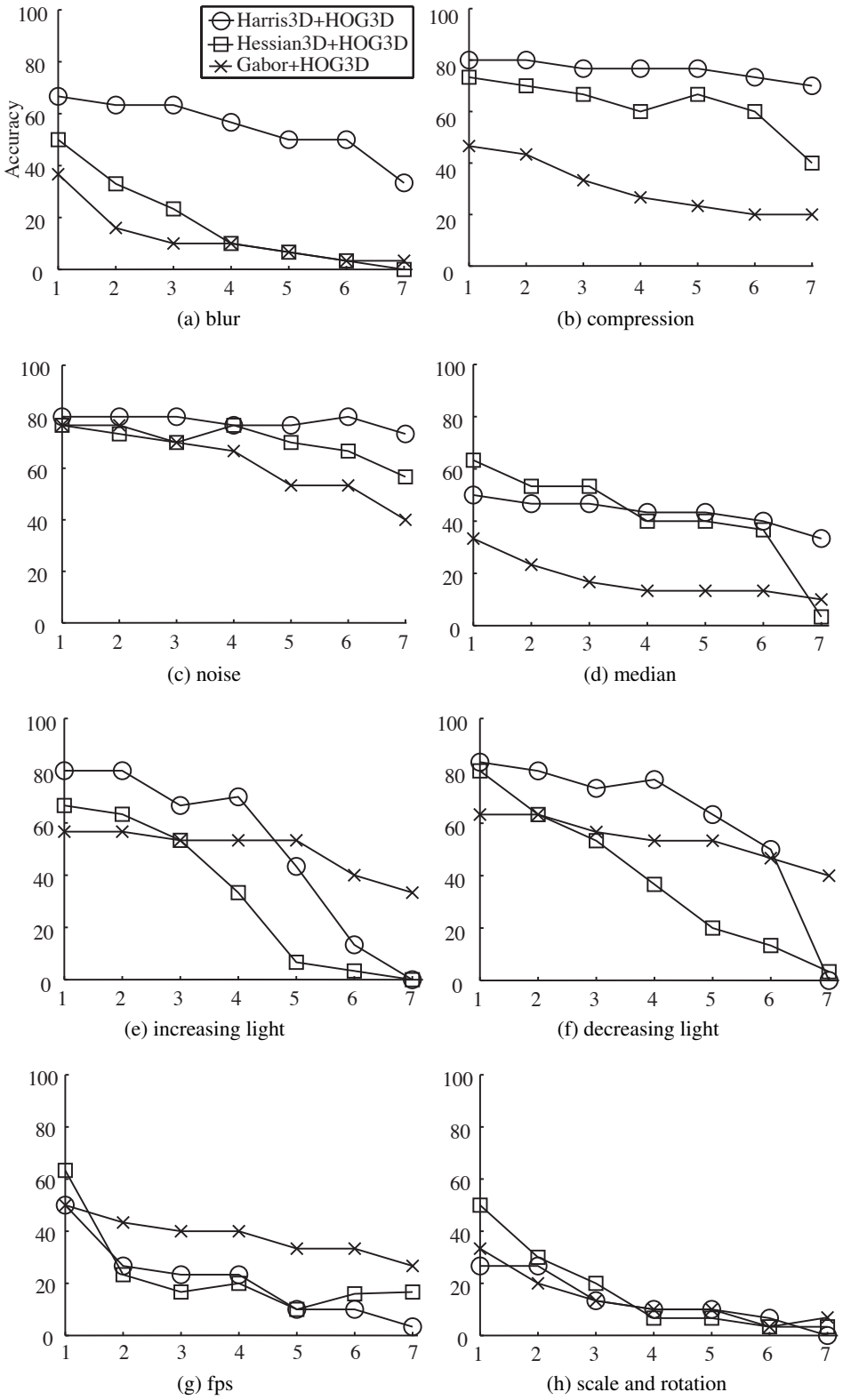


Figure 6.20: Classification accuracy per challenge. Legend is shown in (a).

ertheless, it is shown that detections can be carried out on a significantly smaller resolution than the original one, not changing the robustness of the representation significantly. Unfortunately, this is not possible for the stage of description. Here, all the data is important for a successful description. Nevertheless, computational costs can be reduced in the detection stage. For all approaches better results were obtained by letting the features take care of the noise in the data instead of removing the noise beforehand. More important is the contrast of the videos to the final classification performance. Varying the lightness of the videos changes the number and location of the features significantly. This should be taken care of beforehand, or the features should be improved to be more robust to these alterations similar to their 2D counterparts.

It is shown that the robustness to noise, resolution and compression artifacts is highly variable for different features. Certain features remain stable, even when the video is so much altered that it is no longer visually appealing. Moreover, features are differently robust to preceding noise reduction and contrast. This is an important fact for preprocessing of noisy video material where the evaluation shows that it is better to process noisy material for recognition than to reduce the noise beforehand.

CHAPTER 7

Conclusion

In this thesis, new approaches for visual feature localization and the evaluation of such visual features are presented. First, the use of Gradient Vector Flow for feature localization is proposed and evaluated. This leads to a dense and robust representation of grey-level images. Then, color invariance and color boosting is incorporated in a scale invariant interest point detector providing a higher stability than state-of-the art approaches. Extending the idea of a principle feature evaluation to videos, today's most successful spatio-temporal features are evaluated in a new and comprehensive way: FeEval, a dataset of 1710 videos, is set-up providing well-defined visual challenges for evaluation. Per challenge, the feature's behavior, robustness and matching performance is evaluated providing an in-depth analysis of the strengths and weaknesses of the different approaches.

Beginning with the GVFpoints, it is shown that interest points based on GVF provide more stable locations than the well known and broadly used corner or blob detectors. They give a rich and well-distributed description for diverse visual data. The main difference to other interest point detectors is that the GVF takes more surrounding image information into account than other detectors, due to the iterative gradient smoothing during the computation of the GVF. They provide almost perfect stability against local noise, blur and JPEG compression. This makes the proposed interest points well suited for many problems in computer vision like object detection, recognition, and categorization, image retrieval, baseline matching and object registration.

In this thesis, a principled approach to extract scale invariant interest points based on color invariance and color saliency is proposed. This allows the use of color based interest points for arbitrary image matching. Perceptual color spaces are incorporated and their advantages directly passed on to the feature extraction. Repeatability experiments show that with photometric invariants, stability is improved and color information increases the distinctiveness of interest points.

Using fewer features it is shown that comparable or better repeatability rate is obtained while using a more sparse representation. More discriminative features and a more sparse description of images for image matching is achieved. The current trend in using increasing numbers of interest points primarily combines multiple approaches to gather as much data as possible. This

development (1) leaves more responsibility to the classification stage to deal with ambiguous data, (2) lets the system deal with a vast amount of data which results in longer calculation times and (3) is contrary to the main idea of salient and repeatable interest points. Recent publications showed that using dense interest points is feasible and relying on robust classification systems very good results are achieved. In this thesis it is shown that a more sparse but equally informative representation can be directly passed to current and successful image retrieval and object categorization frameworks maintaining the performance with less data to be processed. This decreases the run-time of such systems significantly.

In the field of evaluation of spatio-temporal features, the thesis presents FeEval, a data-set to evaluate the robustness and invariance of spatial-temporal features against 8 challenges. These challenges include increasing blur, noise, change of lighting, median filtering, compression, scale and rotation and frames per second. For the first time, well-defined data for the evaluation of spatio-temporal features is available. For geometric transformations, homography matrices are provided. Furthermore, the videos have overlapping cast making it possible to evaluate action and person recognition under increasing transformation of the videos. In contrast to existing data-sets, FeEval consists of videos of varying sources from surveillance cameras to high resolution movies. All the videos are in color and display a grand variety of persons, surroundings and lighting conditions. It allows for a principled evaluation on generalized data by measuring the geometric repeatability and the description robustness against well defined challenges.

The 30 original videos are systematically transformed to test the robustness of the features. It is shown that contrast and lighting changes are an issue for corner and blob detection and should be taken care of in applications. For changes in scale, the scale selection of Harris3D performed more robustly than Hessian3D. Scaling in the temporal direction, e.g. reducing fps, does not change the representation of the videos significantly while reducing the amount of data to be processed to a minimum. When the number of frames is reduced to 12% of the original number, the detections are maintained, showing that this is a viable way of making feature detection more efficient. Lossy compression can be increased to a level where viewers would no longer be satisfied, but where features detected remain stable. This can help to reduce storage demand for buffering videos in classification tasks.

Future work will be directed into the development of robust spatio-temporal features using color information. Extraction of sparse features is even more important for video processing than for image processing, as the amount of data is much larger and the information tends to be more redundant.

Another future challenge will be the development of recognition methods that can handle huge data-sets. The described data-sets in this thesis are just a small fraction of the data-set we encounter every day on-line. Future work for image and video understanding has to deal with this huge amount of data.

Bibliography

Abdel-Hakim, A. E. and Farag, A. A. (2006). Csift: A sift descriptor with color invariant characteristics. In *CVPR*, pages 1978–1983.

Ashbrook, A. P., Thacker, N. A., Rockett, P. I., and Brown, C. I. (1995). Robust recognition of scaled shapes using pairwise geometric histograms. In *BMVC*, pages 503–512.

Bakker, P., van Vliet, L., and Verbeek, P. (1999). Edge preserving orientation adaptive filtering. In *CVPR*, pages 535–540.

Balas, B. J. and Sinha, P. (2003). Dissociated dipoles: Image representation via non-local comparisons. Technical report, MIT.

Bartlett, M., Movellan, J., and Sejnowski, T. (2002). Face recognition by independent component analysis. *Neural Networks*, 13:1450–1464.

Baumberg, A. (2000). Reliable feature matching across widely separated views. In *CVPR*, pages 288 – 303.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). SURF: Speeded up robust features. In *ECCV*, pages 346– 359.

BBC News (2006). MP urges YouTube violence debate. In *BBC Political News 2006/10/19*.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522.

Berk, T., Kaufman, A., and Brownston, L. (1982). A human factors study of color notation systems for computer graphics. *Commun. ACM*, 25(8):547–550.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. In *Bulleting of the Calcutta Mathematical Society*, volume 35, pages 99–110.

Bigün, J., Granlund, G. H., and Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow. *PAMI*, 13(8):775–790.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402.

Brown, M. and Lowe, D. G. (2002). Invariant features from interest point groups. In *BMVC*, pages 656–665.

Brox, T., van den Boomgaard, R., Lauze, F. B., van de Weijer, J., Weickert, J., Mrázek, P., and Kornprobst, P. (2006). Adaptive structure tensors and their applications. In *Visualization and Processing of Tensor Fields*, pages 17–47.

Cai, D., He, X., and Han, J. (2007). Efficient kernel discriminant analysis via spectral regression. In *ICDM*, pages 427–432.

Cantu-Paz, E. (2002). Feature subset selection by estimation of distribution algorithms. In *GECCO*, pages 303–310.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Int. Conf. Internet Measurement*, pages 1–14.

Cheng, H., Jiang, X., Sun, Y., and Wang, J. (2001). Color image segmentation: advances and prospects. *PR*, 34(12):2259 – 2281.

Chum, O., Perdoch, M., and Matas, J. (2009). Geometric min-hashing: Finding a (thick) needle in a haystack. *CVPR*, pages 17–24.

Couprie, M., Najman, L., and Bertrand, G. (2005). Quasi-linear algorithms for the topological watershed. *J. Math. Imaging Vis.*, 22(2-3):231–249.

Delaca, K., Grgic, M., and Grgic, S. (2005). Independent comparative study of pca, ica, and lda on the feret data set. *IJIST*, 15:252–260.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72.

Donner, R., Mičušík, B., Langs, G., and Bischof, H. (2007). Sparse MRF appearance models for fast anatomical structure localisation. In *BMVC*, pages 1045–1052.

Donoser, M. and Bischof, H. (2006). 3d segmentation by maximally stable volumes (msvs). *ICPR*, 1:63–66.

Donoser, M., Bischof, H., and Wiltsche, M. (2006). Color blob segmentation by mser analysis. In *ICIP*, pages 757–760.

Dorko, G. and Schmid, C. (2003). Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–641.

Duchenne, O., Laptev, I., Sivic, J., Bach, F., and Ponce, J. (2009). Automatic annotation of human actions in video. In *ICCV*, pages 1395–1402.

Edwards, N. (2000). Life-style monitoring for supported independence. In *BTTJV*, volume 18, pages 64–65.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2009). The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

Everingham, M., Zisserman, A., Williams, C. K. I., and Van Gool, L. (2006). The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

Faille, F. (2005). Stable interest point detection under illumination changes using colour invariants. In *BMVC*, pages 128–142.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271.

Finlayson, G. and Drew, M. (1993). Diagonal transforms suffice for color constancy. In *ICCV*, pages 164–171.

Florack, L., ter Haar Romeny, B., Koenderink, J., and Viergever, M. (1991). General intensity transformations and second order invariants. In *Scandinavian Conf. Image Analysis*, pages 338–345.

Florack, L. M. J., Romeny, B. M. H., Koenderink, J. J., and Viergever, M. A. (1996). The gaussian scale-space paradigm and the multiscale local jet. In *IJCV*, pages 61–75.

Forssén, P.-E. (2007). Maximally stable colour regions for recognition and matching. In *CVPR*, pages 184–190.

Förstner, W. and Gülch, E. (1987). A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ICFPP*, pages 281–305.

Freeman, W. T. and Adelson, E. H. (1991). The design and use of steerable filters. *PAMI*, 13:891–906.

Gabor, D. (1946). Theory of communication. *J. Inst. Electr. Engineering*, 3(93):429–457.

Gabriel, P., Hayet, J.-B., Piater, J., and Verly, J. (2005). Object tracking using color interest points. In *AVSS*, pages 159–164.

Geusebroek, J., Burghouts, G., and Smeulders, A. (2005). The Amsterdam library of object images. *IJCV*, 61(1):103–112.

Gonzalez, R. C. and Woods, R. E. (1992). *Digital Image Processing*. Prenitce Hall.

Gool, L. J. V., Moons, T., and Ungureanu, D. (1996). Affine/ photometric invariants for planar intensity patterns. In *ECCV*, pages 642–651.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *PAMI*, 29(12):2247–2253.

Gouet, V. and Boujemaa, N. (2001). Object-based queries using color points of interest. In *Workshop on Content-based Access of Image and Video Libraries*, pages 30–37.

Gouet, V. and Boujemaa, N. (2002). About optimal use of color points of interest for content-based image retrieval. Technical Report RR-4439, INRIA.

Hanbury, A. (2008). Constructing cylindrical coordinate colour spaces. *PR Letters*, 29(4):494–500.

Harris, C. and Stephens, M. (1988). A combined corner and edge detection. In *4th Alvey Vision Conference*, pages 147–151.

Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *AI*, 17:185–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259.

Jahne, B. (1993). *Spatio-Temporal Image Processing: Theory and Scientific Applications*. Springer.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *ICCV*, pages 1–8.

Johnson, A. E. and Hebert, M. (1996). Recognizing objects by matching oriented points. In *CVPR*, pages 684–689.

Johnson, A. E. and Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *PAMI*, 21(5):433–449.

Joly, A. (2007). New local descriptors based on dissociated dipoles. In *CIVR*, pages 573–580.

Junejo, I., Dexter, E., Laptev, I., and Pérez, P. (2008). Cross-view action recognition from temporal self-similarities. In *ECCV*.

Junejo, I., Dexter, E., Laptev, I., and Pérez, P. (2010). View-independent action recognition from temporal self-similarities. *PAMI*, page in press.

Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *ICCV*, pages 604–610.

Kadir, T. and Brady, M. (2001). Saliency, scale and image description. *IJCV*, 45(2):83–105.

Kass, M. and Witkin, A. (1987). Analyzing oriented patterns. *CVGIP*, 37(3):362–385.

Ke, Q. and Kanade, T. (2005). Quasiconvex optimization for robust geometric reconstruction. In *ICCV*, pages 986 – 993.

Ke, Y. and Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. In *CVPR*, pages 506–513.

Ke, Y., Sukthankar, R., and Hebert, M. (2005). Efficient visual event detection using volumetric features. In *ICCV*, pages 166 – 173.

Kenney, C., Zuliani, M., and Manjunath, B. (2005). An axiomatic approach to corner detection. In *CVPR*, pages 191–197.

Khan, R., Stöttinger, J., and Kampel, M. (2008). An adaptive multiple model approach for fast content-based skin detection in on-line videos. In *AREA workshop in conjunction with ACM MM*, pages 89–95.

Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 995–1004.

Koenderink, J. J. and van Doorn, A. J. (1987). Representation of local geometry in the visual system. *Biol. Cybern.*, 55(6):367–375.

Konik, H., Lozano, V., and Laget, B. (1996). Color pyramids for image processing. In *JIST*, volume 40, pages 535–542.

Köthe, U. (2003). Integrated edge and junction detection with the boundary tensor. In *ICCV*, pages 424–431.

Lambert, P. and Carron, T. (1999). Symbolic fusion of luminance-hue-chroma features for region segmentation. *PR*, 32:1857–1872.

Laptev, I. (2005). On space-time interest points. *IJCV*, 64(2):107–123.

Laptev, I., Capuo, B., Schultz, C., and Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *CVIU*, 108(3):207–229.

Laptev, I. and Lindeberg, T. (2003a). Interest point detection and scale selection in space-time. In *Scale Space Methods in Computer Vision*.

Laptev, I. and Lindeberg, T. (2003b). Space-time interest points. In *ICCV*, pages 432–439.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *CVPR*, pages 1–8.

Laptev, I. and Pérez, P. (2007). Retrieving actions in movies. In *ICCV*, pages 1–8.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 42–51.

Lee, J.-S., Kuo, Y.-M., Chung, P.-C., and Chen, E.-L. (2007). Naked image detection based on adaptive and extensible skin color model. *PR*, 40(8):2261–2270.

Leibe, B., Mikolajczyk, K., and Schiele, B. (2006). Efficient clustering and matching for object class recognition. In *BMVC*.

Liensberger, C., Stöttinger, J., and Kampel, M. (2009). Color-based and context-aware skin detection for online video annotation. In *MMSP*, pages 1 – 6.

Lindeberg, T. (1994). Effective scale: A natural unit for measuring scale-space lifetime. *PAMI*, 15:1068–1074.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *IJCV*, 30(2):79–116.

Lindeberg, T. and Gårding, J. (1993). Shape from texture from a multi-scale perspective. In *ICCV*, pages 683–691.

Lloyd, A. (1982). Least square quantization in PCM. In *Transactions on Information Theory*, volume 28, pages 129–137.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157.

Loy, G. and Zelinsky, A. (2002). A fast radial symmetry transform for detecting points of interest. In *ECCV*, pages 358–368.

Loy, G. and Zelinsky, A. (2003). Fast radial symmetry for detecting points of interest. *PAMI*, pages 959–973.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679.

Machajdik, J. and Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *ACM MM*, page to appear.

Machajdik, J., Hanbury, A., and Stöttinger, J. (2010). Understanding affect in images. In *ACM MM - Grand Challenge*.

Manay, S., Cremers, D., Hong, B.-W., Yezzi, A., and Soatto, S. (2006). Integral invariants for shape matching. *PAMI*, 28(10):1602 –1618.

Manjunath, B., Salembier, P., and Sikora, T. (2002). *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley.

Maret, Y., Nikolopoulos, S., Dufaux, F., Ebrahimi, T., and Nikolaidis, N. (2006). A novel replica detection system using binary classifiers, r-trees, and pca. In *ICIP*, pages 925–928.

Markkula, M. and Sormunen, E. (1998). Searching for photos - journalistic practices in pictorial IR. In *The Challenge of Image Retrieval*, pages 1–13.

Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *CVPR*, pages 2929–2936.

Martinez, A. M. and Kak, A. C. (2001). Pca versus lda. *PAMI*, 23:228–233.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *BMVC*, pages 384–393.

Middendorf, M. and Nagel, H.-H. (2002). Empirically convergent adaptive estimation of gray-value structure tensors. In *DAGM*, pages 66–74.

Mikolajczyk, K., Barnard, M., Matas, J., and Tuytelaars, T. (2009). Feature detectors and descriptors: The state of the art and beyond. In *Feature Workshop and Benchmark in conjunction with CVPR*.

Mikolajczyk, K., Leibe, B., and Schiele, B. (2005a). Local features for object class recognition. In *ICCV*, pages 1792–1799.

Mikolajczyk, K., Leibe, B., and Schiele, B. (2006). Multiple object class detection with a generative model. In *CVPR*, pages 26–36.

Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points. In *ICCV*, pages 525–531.

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector. In *ECCV*, pages 128–142.

Mikolajczyk, K. and Schmid, C. (2004). Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630.

Mikolajczyk, K. and Tuytelaars, T. (2009). Local image features. *Encyclopedia of Biometrics*, pages 939–943.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and can Gool, L. (2005b). A comparison of affine region detectors. *IJCV*, 65(1/2):43–72.

Mikolajczyk, K. and Uemura, H. (2008). Action recognition with motion-appearance vocabulary forest. In *CVPR*.

Montesinos, P., Gouet, V., and Deriche, R. (1998). Differential invariants for color images. In *ICPR*, pages 838–842.

Moosmann, F., Triggs, B., and Jurie, F. (2006). Fast discriminative visual codebooks using randomized clustering forests. In *NIPS*, pages 985–992.

Moravec, H. (1977). Towards automatic visual obstacle avoidance. In *IJCAI*, pages 584–590.

Murase, H. and Nayar, S. K. (1994). Illumination planning for object recognition using parametric eigenspaces. *PAMI*, 16(12):1219–1227.

Nagel, H. and Gehrke, A. (1998). Spatiotemporally adaptive estimation and segmentation of optical flow fields. In *ECCV*, pages 86–95.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, pages 1249–1259.

Oikonomopoulos, A., Patras, I., and Pantic, M. (2006). Kernel-based recognition of human actions using spatiotemporal salient points. In *CVPR*, pages 151–159.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987.

Okada, R. and Soatto, S. (2008). Relevant feature selection for human pose estimation and localization in cluttered images. In *ECCV*, pages 434–445.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *PAMI*, 12(7):629–639.

Plataniotis, K. N. and Venetsanopoulos, A. N. (2000). *Color Image Processing and Applications*. Springer.

Pönitz, T., Donner, R., Stöttinger, J., and Hanbury, A. (2010). Efficient and distinct large scale bags of words. In *AAPR*, pages 139–146.

Pönitz, T. and Stöttinger, J. (2010). Efficient and robust near-duplicate detection in large and growing image data-sets. In *ACCM MM, Industrial exhibit*, page to appear.

Quelhas, P. (2007). *Scene image classification and segmentation with quantized local descriptors and latent aspect modeling*. lo, LIDIAP.

Reisfeld, D., Wolfson, H., and Yeshurun, Y. (1994). Context free attentional operators: the generalized symmetry transform. *IJCV*, 14:119–130.

Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. *CVPR*, pages 1–8.

Rousson, M., Rousson, M., Brox, T., Brox, T., Deriche, R., Deriche, R., and Odyssee, P. (2003). Active unsupervised texture segmentation on a diffusion based feature space. In *CVPR*, pages 699–704.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121.

Ruderman, D. L. and Bialek, W. (1994). Statistics of natural images: Scaling in the woods. In *Physical Review Letters*, volume 73, pages 814–817.

Rugna, J. D. and Konik, H. (2002). Color interest points detector for visual information retrieval. In *Electronic Imaging*, volume 4672, pages 139–146.

Schaffalitzky, F. and Zisserman, A. (2002). Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *ECCV*, pages 414–431.

Schindler, G., Brown, M., and Szeliski, R. (2007). City-scale location recognition. In *CVPR*, pages 1–7.

Schüldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *ICPR*, pages 32–36.

Sebe, N., Gevers, T., Dijkstra, S., and van de Weijer, J. (2006a). Evaluation of intensity and color corner detectors for affine invariant salient regions. In *CVPRW*, pages 18–26.

Sebe, N., Gevers, T., van de Weijer, J., and Dijkstra, S. (2006b). Corners detectors for affine invariant salient regions: Is color important? In *CIVR*, pages 99–112.

Shechtman, E. and Irani, M. (2007). Matching local self-similarities across images and videos. In *CVPR*, pages 1 – 8.

Sivic, J., Russell, B., Efros, A. A., Zisserman, A., and Freeman, B. (2005). Discovering objects and their location in images. In *ICCV*, pages 370–377.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Smith, A. R. (1978). Color gamut transform pairs. In *SIGGRAPH*, pages 12–19. ACM.

Srinivasan, S. H. and Sawant, N. (2008). Finding near-duplicate images on the web using fingerprints. In *ACM MM*, pages 881–884.

Stokman, H. and Gevers, T. (2007). Selection and fusion of color models for image feature detection. *PAMI*, 29:371–381.

Stöttinger, J. (2008). *Local Colour Features for Image Retrieval*. VDM Verlag Dr. Müller.

Stöttinger, J., Banova, J., Pönitz, T., Sebe, N., and Hanbury, A. (2009a). Translating journalists' requirements into features for image search. In *VSMM*, pages 149 – 153.

Stöttinger, J., Donner, R., Szumilas, L., and Hanbury, A. (2008). Evaluation of gradient vector flow for interest point detection. In *ISVC*, pages 338–348.

Stöttinger, J., Goras, B. T., Pönitz, T., Sebe, N., Hanbury, A., and Gevers, T. (2010a). Systematic evaluation of spatio-temporal features on comparative video challenges. In *Workshop on Video Event Categorization, Tagging and Retrieval, in conjunction with ACCV*, pages 1–8.

Stöttinger, J., Goras, B. T., Sebe, N., and Hanbury, A. (2010b). Behavior and properties of spatio-temporal local features under visual transformations. In *ACM MM*, pages 1–4.

Stöttinger, J. and Hanbury, A. (2009). Ordnung in die Bilderflut - Arbeitsweise von content-based Image Retrieval Systemen. In *Prozessgestaltung in der Medienproduktion. Neue Geschäftsmodelle und Technologien für Mobile Protale und HD Broadcast*, pages 103–115. Gito Verlag.

Stöttinger, J., Hanbury, A., Gevers, T., and Sebe, N. (2009b). Lonely but attractive: Sparse color salient points for object retrieval and categorization. In *Workshop on Feature Detectors and Descriptors: The State of the Art and Beyond, in conjunction with CVPR*, pages 1 – 8.

Stöttinger, J., Hanbury, A., Liensberger, C., and Khan, R. (2009c). Skin paths for contextual flagging adult videos. In *ISVC*, pages 303 – 314.

Stöttinger, J., Hanbury, A., Sebe, N., and Gevers, T. (2007a). Do colour interest points improve image retrieval? In *ICIP*, pages 169–172.

Stöttinger, J., Sebe, N., Gevers, T., and Hanbury, A. (2007b). Colour interest points for image retrieval. In *Computer Vision Winter Workshop*, pages 83–91.

Stöttinger, J., Zambanini, S., Khan, R., and Hanbury, A. (2010c). FeEval - a dataset for evaluation of spatio-temporal local features. In *ICPR*, pages 499–503.

Svoboda, T., Martinec, D., and Pajdla, T. (2005). A convenient multicamera self-calibration for virtual environments. *PTVE*, 14(4):407–422.

Szumilas, L., Donner, R., Langs, G., and Hanbury, A. (2007). Local structure detection with orientation-invariant radial configuration. In *CVPR*, pages 121–128.

Torralba, A., Fergus, R., and Weiss, Y. (2008). Small codes and large image databases for recognition. In *CVPR*, pages 1–8.

Turcot, P. and Lowe, D. G. (2009). Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshop on Emergent Issues in Large Amounts of Visual Data*, pages 2109 – 2116.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280.

Tuytelaars, T. and Schmid, C. (2007). Vector quantizing feature space with a regular lattice. In *ICCV*, pages 1–8.

Unnikrishnan, R. and Hebert, M. (2006). Extracting scale and illuminant invariant regions through color. In *BMVC*, pages 124–138.

van de Sande, K., Gevers, T., and Snoek, C. (2009). Evaluating color descriptors for object and scene recognition. *PAMI*, 32:1582 – 1596.

van de Weijer, J. and Gevers, T. (2005). Edge and corner detection by photometric quasi-invariants. *PAMI*, 27(4):625–630.

van de Weijer, J., Gevers, T., and Bagdanov, A. (2006). Boosting color saliency in image feature detection. *PAMI*, 28(1):150–156.

van der Velde, F., de Kamps, M., and van der Voort van der Kleij, G. (2004). CLAM: Closed-loop attention model for visual search. *Neurocomputing*, 58:607–612.

Vanossi, M. and Stöttinger, J. (2010). Scale invariant dissociated dipoles. In *AAPR*, pages 155–162.

Vetterli, J. (1995). Wavelets and subband coding. *Prentice Hall*.

Vigo, D. A. R., Khan, F. S., van de Weijer, J., and Gevers, T. (2010). The impact of color in bag-of-words based on object recognition. In *ICPR*, pages 1549–1552.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *IJCV*, 57(2):137–154.

Wang, H., Ullah, M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC*, pages 127–138.

Wang, L., Zhou, L., and Shen, C. (2008). A fast algorithm for creating a compact and discriminative visual codebook. In *ECCV*, pages 719–732.

Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*. Teubner-Verlag.

Wildenauer, H., Melzer, T., and Bischof, H. (2002). A gradient-based eigenspace approach to dealing with occlusions and non-gaussian noise. In *ICPR*, pages 977–980.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, pages 650–663.

Witkin, A. P. (1983). Scale-space filtering. In *IJCAI*, pages 1019–1022.

Wong, S. F. and Cipolla, R. (2007). Extracting spatiotemporal interest points using global information. In *ICCV*, pages 1–8.

Wu, Z., Ke, Q., Isard, M., and Sun, J. (2009). Bundling features for large scale partial-duplicate web image search. In *CVPR*, pages 25–32.

Xu, C. and Prince, J. L. (1998). Snakes, shapes, and gradient vector flow. *TIP*, 7:359–369.

Yan, F., Kittler, J., Mikolajczyk, K., and Tahir, M. A. (2009). Non-sparse multiple kernel learning for fisher discriminant analysis. In *ICDM*, pages 1064–1069.

Yan, F., Mikolajczyk, K., Kittler, J., and Tahir, M. (2010). Combining multiple kernels by augmenting the kernel matrix. In *International Workshop on Multiple Classifier Systems*, pages 175–184.

Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV*, pages 151–158.

Zambanini, S., Machajdik, J., and Kampel, M. (2010). Early versus late fusion in a multiple camera network for fall detection. In *ÖAGM*, pages 15–22.

Zenzo, S. D. (1986). A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33(1):116 – 125.

Zhang, J., Marszałek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238.

Zheng, H., Daoudi, M., and Jedynak, B. (2004). Blocking adult images based on statistical skin detection. *ELCVIA*, 4(2):1–14.

Nomenclature

$ALOI$ Amsterdam Library of Object Images

AP Average Precision

$CHOG$ Colored Histogram of Gradients

CIR Computational Image Analysis and Radiology

CMP Center for Machine Perception

$CVPR$ Conference on Computer Vision and Pattern Recognition

$DAISY$ dense Descriptor Applied In Stereo vision

DoG Difference of Gaussian

DoH Determinant of Hessian

EMD Earth Mover's Distance

$EPFL$ Ecole Polytechnique Federale

$ESURF$ Extended Speeded Up Robust Features

ETH Eidgenössische Technische Hochschule

$FRST$ Fast Radial Symmetry Transform

$GLOH$ Gradient Location and Orientation Histogram

GMM Gaussian Mixture Models

GST Generalized Symmetry Transform

GVF Gradient Vector Flow

HD High Definition

$HDTV$ High Definition Television

$HMMD$ Hue Max Min Diff color space

HOF Histogram of Optical Flow

HOG Histogram of Oriented Gradients

$HOG3D$ Histogram of Oriented Gradients in 3D

HSB Hue Saturation Brightness color space

HSI Hue Saturation Intensity color space

HSL Hue Saturation Lightness color space

HSV Hue Saturation Value color space

$INRIA$ Institut National de Recherche en Informatique et Automatique

$JPEG$ Joint Photographic Experts Group

LoG Laplacian of Gaussian

MAP Mean Average Precision

$MSER$ Maximally Stable Extremal Regions

$MSER$ Maximally Stable Extremal Regions

MSV Maximum Stable Volume

$MUSCLE$ Multimedia Understanding through Semantics, Computation and Learning project

OCS Opponent Color Space

$PASCAL$ Pattern Analysis, Statistical Modeling and Computational Learning project

PCA Principal Component Analysis

PDA Personal Digital Assistant

RBF Radial Basis Function

RGB Red Green Blue color space

$SIFT$ Scale Invariant Feature Transform

$SPIN$ not an abbreviation - descriptor's name is chosen because the approach can be visualized by spinning a sheet of paper around the normal of the point [Johnson and Hebert, 1999]

$SRKDA$ Spectral Regression Kernel Discriminant Analysis

SSD Sum of Squared Distances

$STIP$ Space Time Interest Points

$SURF$ Speeded Up Robust Features

$SURF3D$ Speeded Up Robust Features in 3D

SVM Support Vector Machine

TU Technical University

URL Uniform Resource Locator

VGA Video Graphics Array

VOC Visual Object Class