Colour Local Feature Fusion for Image Matching and Recognition

Tony Marrero Barroso, B.Eng.

January 2016

School of Electronic Engineering
Faculty of Engineering and Computing
Dublin City University

Supervised by Prof. Paul F. Whelan

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ ID No.: 54355687

Candidate

Date: 15th January 2016

Abstract

Tony Marrero Barroso

Colour Local Feature Fusion for Image Matching and Recognition

This thesis investigates the use of colour information for local image feature extraction. The work is motivated by the inherent limitation of the most widely used state of the art local feature techniques, caused by their disregard of colour information. Colour contains important information that improves the description of the world around us, and by disregarding it; chromatic edges may be lost and thus decrease the level of saliency and distinctiveness of the resulting grayscale image. This thesis addresses the question of whether colour can improve the distinctive and descriptive capabilities of local features, and if this leads to better performances in image feature matching and object recognition applications. To ensure that the developed local colour features are robust to general imaging conditions and capable for real-world applications, this work utilises the most prominent photometric colour invariant gradients from the literature. The research addresses several limitations of previous studies that used colour invariants, by implementing robust local colour features in the form of a Harris-Laplace interest region detection and a SIFT description which characterises the detected image region. Additionally, a comprehensive and rigorous evaluation is performed, that compares the largest number of colour invariants of any previous study. This research provides for the first time, conclusive findings on the capability of the chosen colour invariants for practical real-world computer vision tasks. The last major aspect of the research involves the proposal of a feature fusion extraction strategy, that uses grayscale intensity and colour information conjointly. Two separate fusion approaches are implemented and evaluated, one for local feature matching tasks and another approach for object recognition. Results from the fusion analysis strongly indicate, that the colour invariants contain unique and useful information that can enhance the performance of techniques that use grayscale only based features.

Acknowledgements

I wish to firstly thank Prof. Paul Whelan for the opportunity of pursuing this PhD and giving me all his continuous support and guidance throughout all these past years. I also want to also acknowledge Dr. John Mallon for all the help and guidance given to me for the first years of my PhD.

It would have been a completely different experience without the friendship and help of my fellow colleagues: Aubrey, Brendan, Michele, Stephen, Sean, Tarik, Trish and Vincent, and all my other friends specifically from the Mechanical Dept. (Ahmed, Aymen, David, Esco and Saba) and from the Optics group (Arsalan, Colm, Daniele, Desi, Fernando, Josue and Vidak). Their presence and interactions with me have truly made my PhD journey enjoyable.

I feel an immense gratitude for the enormous hands on help and guidance given to me by Dr. Ovidiu Ghita during the last years of my PhD, his efforts have been invaluable in the completion of this PhD. I also wish to thank my family and my girlfriend Silvia, for all their unconditional encouragement and patience.

Finally, I would like to thank the Irish state and specifically IRCSET (the Irish Research Council for Science, Engineering and Technology) funded by the National Development Plan, for generously funding my PhD studies.

Publications Arising

Conference Papers

Towards Real-Time Stereoscopic Image Rectification for 3D Visualisation. T. Marrero Barroso, A. K. Dunne, J. Mallon and P. F. Whelan. *Asian Conference for Computer Vision (ACCV) Workshop on Application of Computer Vision for Mixed and Augmented Reality*, Queenstown, NZ, 2010.

Enhancing SURF Feature Matching Using Colour Histograms. T. Marrero Barroso and P. F. Whelan. *In Proc. Irish Machine Vision and Image Processing Conference (IMVIP)*, Dublin, Ireland, pp 111-112, 2011.

Evaluating the Performance and Correlation of Colour Invariant Local Image Feature Detectors. T. Marrero Barroso, O. Ghita and P. F. Whelan. *In Proc. IEEE International Conference on Image Processing (ICIP)*, Paris, France, pp 5751-5755, 2014.

Internal Presentations

Real-Time Stereoscopic Image Rectification with Distortion Minimisation. T. Marrero Barroso, J. Mallon and P. F. Whelan. *Faculty Research Day*, Dublin City University, 2009. (poster)

Local Image Features. T. Marrero Barroso, O. Ghita and P. F. Whelan. *Annual Rince Research Day*, Dublin, 2014. (presentation)
(Awarded with the 1st Prize for Best Presentation)

Planned Journal Submissions

Color Invariants for Local Feature Matching and Object Class Recognition. T. Marrero Barroso, O. Ghita and P. F. Whelan. *IEEE Transactions on Circuits and Systems for Video Technology*.

Contents

List of Figures

List of Tables

List of Acronyms

ALOI Amsterdam Library of Object Images

BOVW Bag-of-Visual-Words

BRISK Binary Robust Invariant Scalable Keypoints

HL Harris-Laplace

DoG Difference of Gaussian

HOG Histograms of Oriented Gradients

HSI Hue-Saturation-Intensity

LIC Light Invariant Colour

LBP Local Binary Pattern

LoG Laplacian of Gaussian

MSER Maximally Stable Extremal Regions

OCS Opponent Colour Space

PCA Principal Component Analysis

RANSAC Random Sampling Consensus

SIFT Scale Invariant Feature Transform

SLAM Simultaneous Localisation and Mapping

SURF Speeded up Robust Feature

VOC Visual Objects Challenge

1 Introduction

Human visual attention is determined largely by two complementary psycho-neural mechanisms: environment-driven bottom-up saliency and knowledge-driven top-down guidance (Itti et al., 2005). This thesis is concerned with the bottom-up saliency mechanism of representing the visual form, and in computer vision, one of the most successful ways to achieve it is with local invariant image features (Schmid et al., 2005). Local features have become a vital part of modern computer vision solutions, and amongst the most popular and researched topics of the field. Their importance, versatility and maturity can be seen in the many recent applications that utilise image features. These include image retrieval (Arandjelovic and Zisserman, 2013), object recognition (Biagio et al., 2014), action recognition (Oneata et al., 2013), tracking (Takacs et al., 2013), reconstructing camera views and 3D reconstruction (Irschara et al., 2012, Frahm et al., 2010) and visual odometry (Newcombe et al., 2011).

In the field of computer vision, a local image feature is the name given to a vector that represents a region of an image. This vector contains information about the location of the region in the image, its size, and a fingerprint that describes certain aspects of the region. The extraction of local features is a technique comprising a detector and a descriptor. The detector finds local image regions that are deemed salient (interesting) and stable, which are then characterised by the descriptor with numerical signatures for subsequent feature matching tasks. The field has produced a significant number of these features in the last two decades, and they have proven very successful in their tasks as they can be made robust to varying imaging conditions such as scale, occlusion, rotation and perspective changes. Figure 1.1 provides an illustration of the feature extraction concept, applied on three images with different imaging conditions.

The detector in the example, is able to identify the same local image region of the scene (denoted by the green squares), in images taken with a different scaling, rotation and viewpoint. A descriptor must also be able to characterise those three regions with a similar descriptor, in order to identify them as pertaining to the same region of the scene.

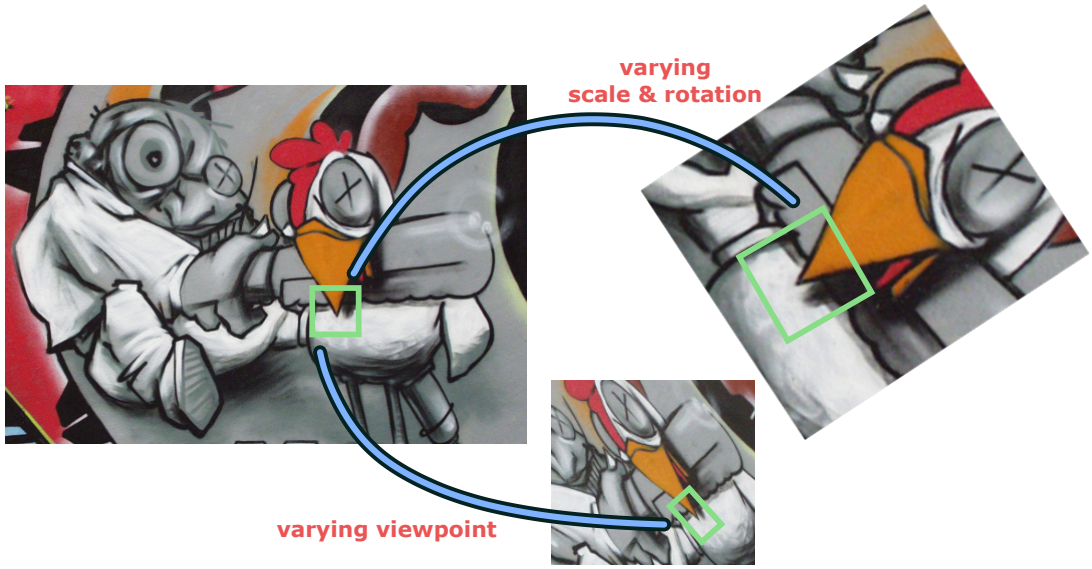


Figure 1.1: Illustration of the local feature extraction concept under varying imaging conditions, where a green square denotes the area of the detected local feature.

Despite the advances made each year, local invariant features inherently still have substantial practical limitations. These limitations became apparent at the early stages of this research, where the work focused on 3D stereo visualisation (Marrero Barroso et al., 2010) and 2D to 3D video conversion. Local features were being utilised to find point correspondences in stereo image pairs that viewed a scene from different viewpoints. The correspondences were necessary in order to estimate a Fundamental matrix (F) (Hartley and Zisserman, 2004), which describes the transformation between the coordinate frame of references of the two images. This matrix can subsequently be utilised to perform stereo rectification to align the image pair horizontally (analogously to the alignment of human eyes). The feature matching process however, requires the fitting of a geometric model in order to reject point mismatches that occur since not all of the descriptors are sufficiently unique. The matching can thus associate a local region of the scene from one image, to a different region of the scene in the corresponding second image. Experimenting with feature matching

revealed that it was normal for 40-50% of the features to be inaccurately matched. This conflict of geometric association between the two images results in an inaccurate F matrix, in which slight perturbations lead to significant changes in the resulting rectified stereo images. The geometric model that constrains the matching process is aided by a sampling algorithm that selects the most optimal set of matches that best fit the model. Random Sampling Consensus (RANSAC) (Fischler and Bolles, 1981, Hartley and Zisserman, 2004) was the chosen sampling algorithm, which randomly iterates between sets of matches until a particular threshold of model fit is reached. The F matrix model depends on correlating real world 3D points to their projected 2D positions on the two camera reference frames, thus different models can be obtained depending on which sub-set of 2D points are chosen.

Experimentation showed that different valid F matrices could be estimated from the same stereo pair by running the RANSAC algorithm multiple times. Essentially indicating, that utilising a geometric model to reject incorrect feature matches is not the most optimal strategy to employ. This research thus began exploring ways of relying less on RANSAC-like techniques to perform accurate feature matching, and turned towards making the local features themselves more unique and distinct. The features that were being used were the Scale Invariant Feature Transform (SIFT) (Lowe, 2004) and Speeded up Robust Feature (SURF) (Bay et al., 2008), which have been in practice sufficient for many computer visions tasks as they are inherently robust and distinct. However, just like the other state of the art mainstream feature approaches, they were designed to be used only on grayscale intensity information. In order to develop more distinct local features, the most apparent research direction to pursue was seen to be investigating the role of colour. All subsequent work on stereo rectification ceased and the focus turned to developing colour local features; the overall goals of this research thus became:

- Finding out why colour was not incorporated in the most popular local feature techniques.
- Exploring how colour could be used for local feature extraction.
- Determining what benefits, if any, colour information would provide.

1.1 Why Colour?

Vision is one of the most vital sensory mechanisms for intelligent machines and living organisms. In evolutionary terms, organisms developed monochromatic vision long before adapting to perceive colour. As they became better adapted to their environment, living organisms expanded the range of wavelengths of light that they were sensitive to by increasing the types of photoreceptors present in the eye. This enabled them to extract more information from the world in various specific ways. Bees for example, are sensitive to ultraviolet (UV) light which aids them in finding nectar in flowers (Michener, 1974). One of the eyes with the most types of photoreceptors in the animal kingdom belongs to the mantis shrimp, which has 12 different photoreceptors and enable the shrimp to see UV and even distinguish polarised light. Though the exact reasons for their complex visual system is not properly understood, scientists believe that UV detection can help localise transparent fish on coral reefs, and enable fluorescent bio-signalling during mating rituals (Mazel et al., 2004). Colour perception is also an important aspect of primate vision, which is believed to have developed when primates began eating coloured fruits and vegetables. Research into biological vision showed that visual form is perceived only from luminance (Marr, 1982). However experiments show that objects in coloured scenes are more easily detected, identified and remembered than in monochromatic scenes (Geisler, 2008). Other studies suggest that colour in fact, is processed together with luminance by the same neural pathways to achieve a united and more robust representation of the visual world (Gegenfurtner, 2003, Gegenfurtner and Kiper, 2003).

The use of colour in computer vision for extracting bottom-up saliency via local features is not as common or developed as in the natural kingdom. The most popular state of the art features used for general tasks are predominantly based on shape information using luminance intensity, and ignore colour information despite the wide availability of colour cameras. Ignoring colour is due to various practical difficulties, such as the extra computational cost of processing the colour channels. The main reason that colour is not widely used though, is due to the difficulty in achieving colour constancy and invariance under a wide range of imaging conditions, as the measured colour values vary

significantly, especially with illumination changes. Colour is however a very important quality for describing the world around us, and grayscale image conversion has a number of undesirable side-effects for local feature extraction. It is obvious that by disregarding colour values chromatic edges can be lost and thus decrease the level of saliency and distinctiveness of the converted image (Van de Weijer et al., 2006a). For example after grayscale conversion, an edge between a blue and a green region would contain the same saliency as an edge between two shades of grey. In other cases the difference between colours may not be visible at all. Grayscale conversions that map colour vectors to scalars based only on luminance intensity are particularly prone to loosing chromatic saliency. This is because isoluminant image regions (regions with the same RGB vector magnitude), will all share the same grayscale intensity value after the conversion.

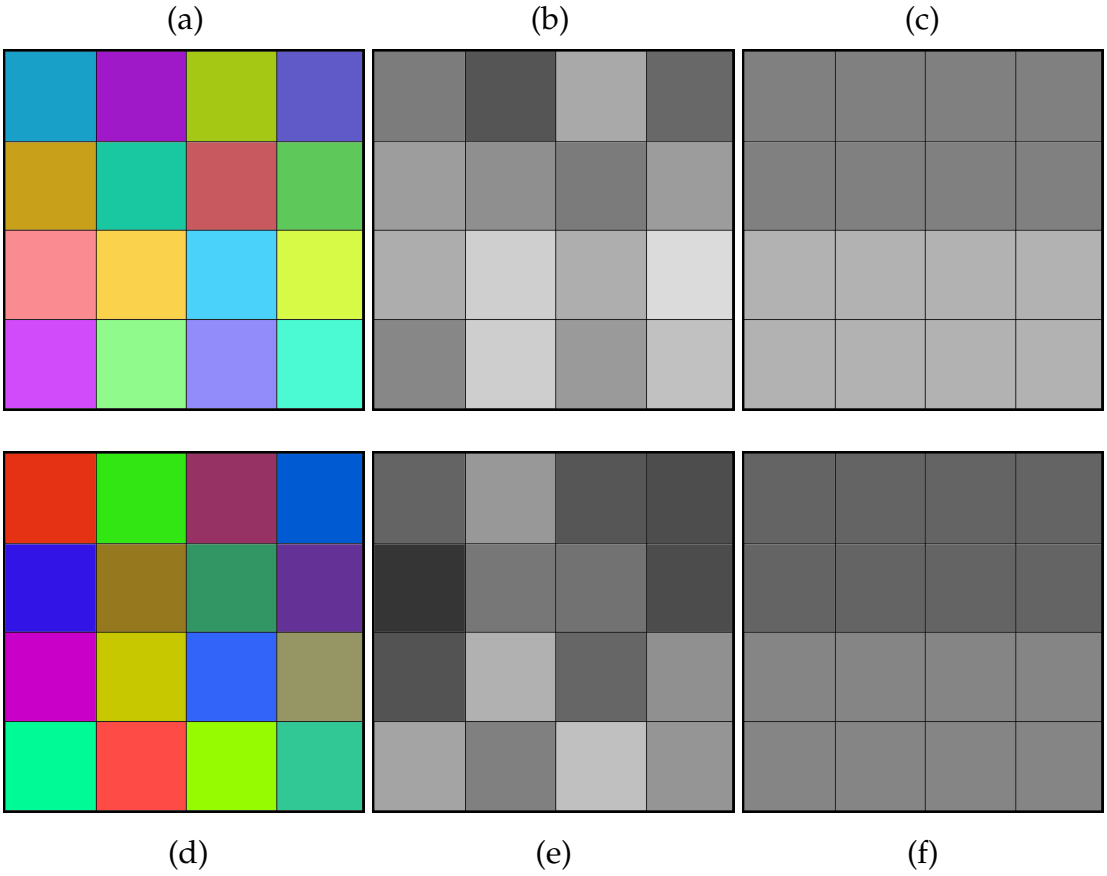


Figure 1.2: Example of a loss of image saliency when converting two colour images (a, d) to grayscale. Matlab's *rgb2gray* function is used for images (b) and (e). A standard luminance intensity method is used for the conversions (c) and (e).

Examples of this isoluminant loss in chromatic saliency can be seen in Figures 1.2 and 1.3. In Figure 1.2 two coloured chessboard grids are converted with two methods, Matlab's *rgb2gray* function (middle column), and a method based on luminance intensity (right column). Figure 1.3 shows the colour conversion to intensity of the painting *Impressionist Sunrise*, in which the painter Claude Monet deliberately makes the sun and the background isoluminant. It is clear to see in both figures how two very distinct colours can be mapped to the same or similar shade of gray. Colour to grayscale conversion is essentially a dimensionality reduction problem, of which there are many different types (Benedetti et al., 2012), however since all of them map a 3 dimensional colour vector to a scalar value, a loss of information will always take place. Despite the difficulty of harnessing colour information, if done appropriately the improved discriminative performance can justify the extra computational costs incurred.

Results from the literature indicate that colour feature detection improves results when using luminance descriptors (Stöttinger et al., 2009), and using colour for both detection and description can also provide a further improvement (Van De Sande et al., 2010, Vigo et al., 2010a, Krylov et al., 2012). Another positive example can be found in a study performed by the author, that utilised colour histograms to improve the matching performance of SURF descriptors (Marrero Barroso and Whelan, 2011). The study found that the use of colour resulted in performance gains of up to 9%.

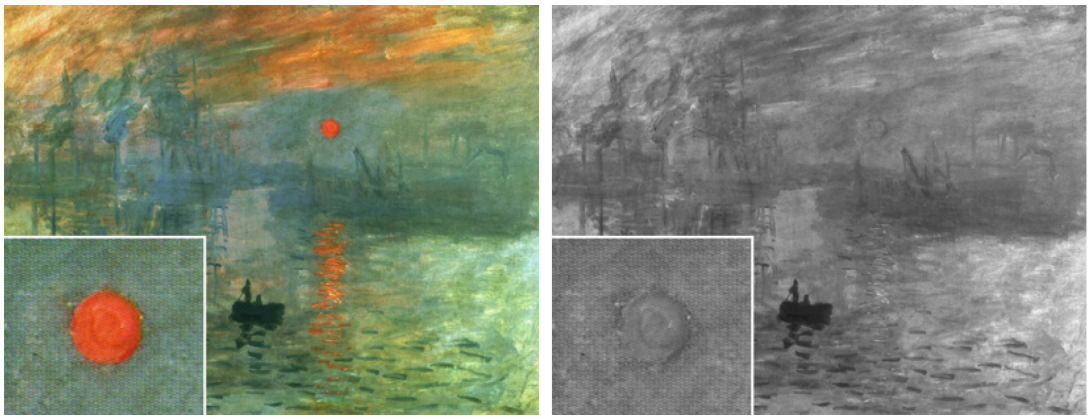


Figure 1.3: Grayscale conversion of the painting *Impressionist Sunrise*, by Claude Monet. Courtesy of www.artcyclopedia.com

1.2 Scope of the Research

This research has been conducted focusing on local colour image features, and the validation study is performed on local image feature matching and object class recognition tasks. Numerous interest point detectors have been proposed in the literature that are robust to different image distortions. Local point detectors can be divided into four main branches that detect corners, blobs, edges and affine regions. Detectors perform well when they are compatible with the structures or imaging distortions present in the image, so there is not one type of detector that is superior in all cases. For the applications of feature matching and image recognition, corner and blob detectors are generally used due to the abundance of those features in image and video data, and they will be the focus of this research. The most successful luminance and colour-based local image detectors are gradient-based, and rely on scale-invariant corner and blob detection, like the well known Harris-Laplace (Mikolajczyk and Schmid, 2001) and Laplacian of Gaussian (LoG) (Lindeberg, 1998) detectors. Similarly, the most prominent photometric colour invariants proposed in the literature for mitigating the constancy problem across varying imaging conditions also calculate image derivatives. The scope of this research will not include colour constancy approaches, whose aim it is to adjust the original image prior to subsequent processing by correcting the colour of the light source present in the image. Section 2.4 provides more details on the justification for this approach, but in summary it is mainly due to the uncertain benefits of using standard colour constancy algorithms for the selected applications. For all the previous mentioned reasons, the colour feature extraction (detection and description) developed in this research employs only gradient-based techniques and ignores statistical methods of achieving constancy.

Many implemented colour features in the literature have been limited in terms of their robustness to different imaging conditions. For this reason the developed local colour features in this research should comply with but not limited to the following criteria (Tuytelaars and Mikolajczyk, 2008) that apply to all local image features (grayscale or colour):

1. **Photometric Robustness** - The features should be robust to photometric

variations such as changes in exposure and lighting direction, shadow, shading and specularities.

2. **Geometric Robustness** - Invariance of the features is also needed with respect to geometrical changes, such as viewpoint, zoom, and orientation variations.

3. **Generality** - The features must be applicable for various applications such as matching, retrieval and classification. Additionally, they should be robust to variations in image quality and different types of camera acquisition.

The previous requirements improve the repeatability of a feature detector, which means that the same region of a scene can be detected repeatedly in images which vary in imaging conditions. Additionally, those requirements also increase the discriminative capacity of a feature descriptor to assign a local region with a numerical descriptor that is robust to variations across different imaging conditions. To validate the colour features for those requirements, appropriate datasets have been chosen on which to perform the feature matching and object class recognition. Four different feature matching datasets are used (see Section 2.5.3) which contain all of the aforementioned imaging variations, along with being acquired with different camera hardware. To ensure greater generality, the object class recognition is carried out on a vision community standard dataset of 9,963 images, and employing the well-known Bag-of-Visual-Words (BOVW) recognition approach (Sivic and Zisserman, 2003). The evaluation framework used here for both applications is the same as the standardised methods from the literature. The goal here is to study colour features and compare them with their grayscale counterpart under the same set of testing conditions. For this reason, the developed framework has made all the different evaluated local features compatible, by implementing them with the same source-code base.

1.3 Objectives and Approach

The motivation for using colour for local feature extraction in this research is to obtain features with increased levels of distinctiveness and repeatability with respect to grayscale-based counterparts. Numerous colour invariant models

have been proposed in the literature that aim to robustly use colour in practical computer vision tasks. Nevertheless, colour feature detection and matching have not been evaluated sufficiently in the literature and have largely been disconnected. The literature contains numerous scattered studies, that in some cases focus on niche applications or only merely prove theoretical concepts. There have been no clear indicators on the best practices or techniques for general-purpose local colour feature extraction. Therefore, this research set out to address the literature's lack of a substantial evaluation and comparison of the most prominent colour invariants. The overarching goal, is to provide insights that would facilitate the integration of colour into mainstream local feature extraction techniques aimed for general applications.

In this work, colour is applied throughout the entire local feature extraction process (detection and description). Utilising a compatible geometric colour-shape approach in both phases of the process is an important difference to much of the literature, as locating a region with a colour detector increases the saliency that a colour descriptor can then extract from it. The features of this study are implemented with a scale-invariant colour Harris-Laplace (HL) detector using the most promising colour invariant gradients from the literature, and for the description, colour invariant SIFT (Lowe, 2004) is used in order to use the local features robustly in their tasks. SIFT is a descriptor composed of a histogram representing the orientations of the image gradients within the local image region identified by the HL detector. The approach taken in this research has been:

1. The implementation of a Harris-Laplace detector and algorithm parameters optimisation. All the development was performed in Matlab.

2. Adapting photometric colour invariants in order to be used with the developed HL detector.

3. Implementing a SIFT descriptor algorithm, compatible with the used gray-scale and colour invariant gradients.

4. Selecting appropriate datasets and evaluation frameworks to robustly test the local features.

5. Evaluating and comparing the grayscale and colour invariant features within the same testing framework.

6. Analysing the correlation between the colour and grayscale features.

7. Studying feature extraction fusion techniques to utilise both luminance and colour information, for image feature matching and recognition tasks.

1.4 Summary of the Research

At this point in the thesis, a brief summary of the work will allow for the research contributions to be better understood. The broader results from the feature detection experiments demonstrate why colour is not preferred in the literature when tackling general vision tasks. Luminance (grayscale) features proved to be the overall best performer when taking into account all the different imaging distortion types. The colour invariants are more suited in scenarios that contain illumination variations, but even so there is only one colour invariant that performs clearly better than luminance under illumination distortions. In the case of the feature descriptor matching experiments, which tested the ability of the gradients to generate robust and distinct SIFT descriptors, grayscale obtained the second best results for general imaging conditions. However the majority of the colour gradients performed poorly.

The colour invariants in general performed better as descriptors than as detectors. Luminance, in one of the illumination varying datasets, in fact performed second worst in terms of matching score, which is essentially a measure of how distinct the descriptors are. The matching study thus proves the negative effects of loosing chromatic information when converting to grayscale. The overall conclusions of the substantial feature detection and matching evaluation performed in this research, is that there are only two colour gradient types that should be considered to be used alongside the grayscale. The findings can serve future works in proceeding in a more clear direction when it comes to local colour feature extraction.

Since the feature detection and matching results indicated that a grayscale technique should generally be used unless dealing with varying illumination conditions, a correlation study was carried out to investigate if colour could

contribute to enhancing grayscale-based techniques. This is a novel study that obtained positive results indicating that the colour features are largely uncorrelated with luminance and that each colour invariant produces local features that are unique (compared to all other colour and grayscale features). Significant numbers of these unique features produce correct correspondences, therefore when taking all correct unique and common correspondences into account, their sum is greater than the correspondences achieved by only using the luminance-based features. These results motivated an investigation into feature fusion extraction techniques, that aimed at selecting a set of features from an image by utilising grayscale and colour gradients in an appropriate manner. Fusion techniques were proposed for two applications; image feature matching and image recognition.

The feature matching fusion study, did not obtain favourable results for the proposed fusion techniques that focused on selecting the strongest subset across all grayscale and colour HL points. The fusion failed to consistently select the most appropriate set of interest points that would guarantee overall better performance. Further analysis was performed in order to investigate the reasons for the non-optimal fusion. The investigation concluded that the standard method for ranking the HL points is inadequate to select the best set from the same extraction technique. This insight into the ranking of HL points, is another novel contribution arising from this work. Its implication for the implemented fusion techniques, is that the ranking metric lacks enough information to qualitatively decide which colour or grayscale points should be selected for the optimal extracted feature set from an individual image. The fusion study in the scope of local feature image matching, thus concludes that despite the possibility of a positive feature fusion, it cannot be achieved with the standard ranking metric utilised for HL.

For the object class recognition fusion experiments, the fusion strategy focused on the SIFT descriptors of the features rather than the strength of the HL points. In the implemented approach, descriptors extracted from grayscale and multiple colour invariants are pooled together to represent the visual vocabulary of the BOVW pipeline. The selection of which descriptors form part of the visual vocabulary are dictated via K-means clustering, therefore the inadequate metric previously used to rank the HL points does not impact

the fusion technique. Luminance proved to be the best overall approach when evaluating the recognition performance of individual features. However 40% of the 20 different object classes in the used recognition dataset, obtain a better result with methods other than luminance. In terms of the fusion recognition experiments, the proposed feature fusion scheme consistently obtains better recognition results than using grayscale SIFT descriptors extracted with the standard dense sampling approach. This is caused by the higher levels of colour saliency that can be extracted from an image, since unique information is combined from different grayscale and colour descriptors.

1.5 Contributions

There are two main aspects of the contributions arising from this work; comprising a substantial evaluation of photometric colour gradient invariants for image feature matching and image class recognition applications, and a grayscale and colour feature fusion investigation aiming at conjointly utilising the best types of features from various gradient types. The overall evaluation and comparison of the invariants that is performed here is more comprehensive than previous works, namely due to utilising more types of invariants and testing on a more extensive collection of image datasets. The aim of this work is to develop image features that are able to perform robustly in general computer vision applications, therefore apart from evaluating on multiple datasets they are here also evaluated with standardised metrics under the presence of typical imaging distortions.

In terms of the feature fusion work, this research investigates a new concept for fusing colour features for feature matching and image recognition. The literature has utilised colour and grayscale information together for recognition tasks (focusing on the feature descriptors), but has ignored fusion approaches for local feature matching. In this work a correlation analysis for feature matching is performed that shows the level of redundancy in the extracted features between the colour and grayscale-based techniques. The study uncovers the extent of the useful information that colour can provide to the feature extraction process. It indicates that there is a strong potential for developing a feature fusion extraction approach for local feature matching, in which the best features

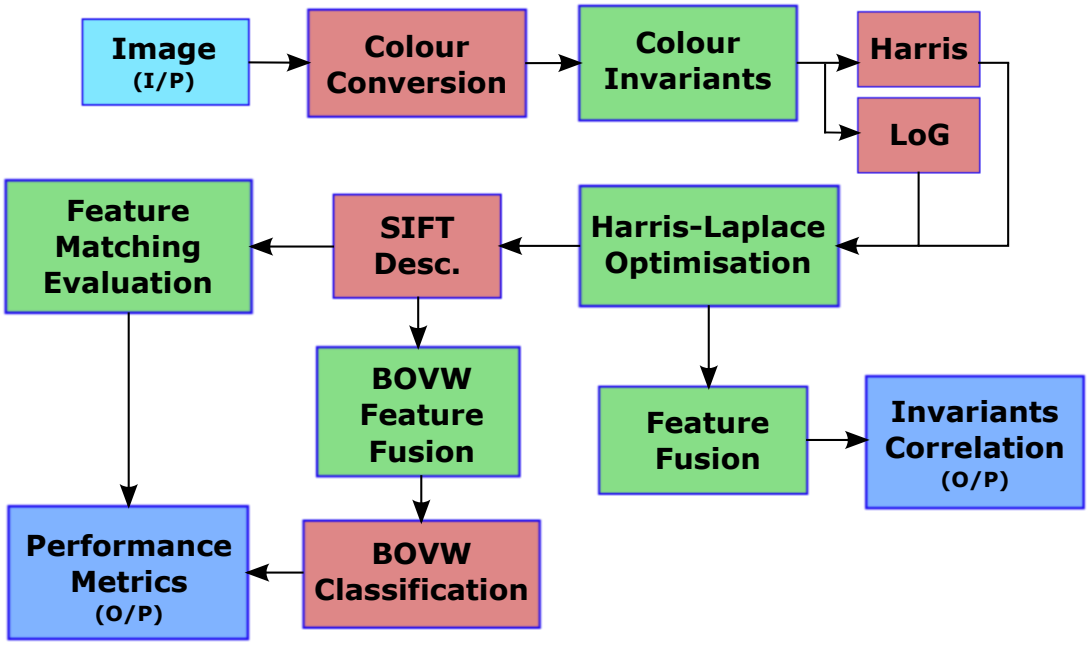


Figure 1.4: Flowchart of the main research activities. Contributions are in green, existing methods in red and input/outputs are in blue.

from grayscale and colour-based techniques can be conjointly extracted from an image.

The second part of the fusion work of this research focuses on image recognition, the recognition fusion strategy results in an overall set of descriptors (from multiple gradient types) that are more salient than a set extracted only from grayscale information. The proposed recognition feature fusion techniques obtain superior recognition results than the standard method used in the literature, which comprises of a dense random sampling grayscale-based feature extraction. This is due to the combination of features from multiple colour invariants which can each provide unique distinct descriptors, and thus increase the information content and complexity of the visual vocabulary of the BOVW framework. A diagram of the main research pipeline is shown in Figure 1.4 with areas of contribution denoted in green, existing methods in red, and inputs (I/P) and outputs (O/P) in blue. The specific individual set of research contributions are as follows:

1. This research has selected to extract local image features from colour invariants that are amongst the most promising and under-evaluated colour

invariant gradients from the literature. Most of those invariants are utilised for local image features here for the first time.

2. In this work the strategy employed for the local feature extraction, uses colour in both the detection and description phases of the process. Whereas the majority of previous works conducted disjointed studies, mainly using colour in the description phase only.

3. All the colour invariants that are evaluated here, are implemented as local features that are robust to the standard imaging conditions studied by the field. Whereas certain colour invariants have previously only been evaluated for edge or non-scale invariant corner detection.

4. A comprehensive feature matching evaluation is performed in this research, which utilises four different local image feature matching datasets with substantial numbers of images. Moreover, they contain the imaging distortions that are most widely tested for in the literature (i.e. scale, viewpoint, blurring, JPEG compression and illumination). This approach is in contrast to previous studies that use a single dataset, fewer number of image examples, or evaluate a limited variety of imaging distortion conditions. Having more datasets ensures this evaluation is unbiased to particular conditions contributed by the hardware of the acquisition, and a greater number of image examples improves the statistical significance of the reported results.

5. The feature matching evaluation framework that is implemented here, also addresses further limitations of many previous works, by evaluating colour features more rigorously using the metrics that are employed in the testing of state of the art luminance features. Furthermore, the image class recognition study that is carried out by this research is performed on one of the most popular recognition datasets of the community which is also particularly challenging for colour-based approaches.

6. A novel local feature correlation analysis is performed here, which is based on actual experimental results in order to investigate the potential benefit of using colour and luminance together. It reports the number of unique correct points that each gradient type can generate and identifies which gradient

types could be used conjointly to obtain a better overall feature matching performance.

7. This research proposes feature fusion techniques for feature matching tasks that focus on the strength of the HL corner points. The investigation indicates that even though there are potentially significant numbers of unique features available for fusion, the proposed technique fails to select the most appropriate set of points from each feature type (colour or luminance) using the standard ranking of HL points. The contribution that arises, is the discovery that the metric used to determine the strength of a corner (which is the standard HL ranking method), is not a sufficient indicator for the robustness of the point, and thus not suitable to differentiate between the best subset of grayscale and colour features.

8. A second feature fusion strategy is proposed, applied to image class recognition and focusing on fusing the information from the SIFT descriptors. The performed evaluation demonstrates that the two proposed fusion techniques outperform standard grayscale-based recognition approaches. This successfully proves the concept of using complementary grayscale and colour information together for local feature extraction.

1.6 Thesis Outline

The literature review is presented in Chapter 2, where the relevant previous studies are discussed in the areas of grayscale and colour feature detection, and grayscale and colour feature description. The focus of the chapter is not on implementation details or background theory, but on a conceptual high-level discussion and comparison of the state of the art. The necessary background theory will be provided in the relevant contribution chapters, to make the chapters self-contained and more easily understood. Chapter 2 has one section (Section 2.5) related to the background, which covers the datasets used in this research and the evaluation framework and metrics employed. That section supports the narrative of the research contributions and is necessary to compare the work carried out in this work with that of the literature. The chapter ends with a summary of the negative aspects of the literature and the necessity for this research, it highlights how this work addresses the limitations of the prior art

and the differences between it and the work performed here.

Chapter 3 presents the colour adaptation of grayscale local features. It outlines the Harris-Laplace algorithm, the implementation utilised in this work, and the optimisation study that was carried out to find the optimum set of algorithm configurations. The colour adaptation examines the photometric invariant theory utilised in this work, and outlines how the theory is implemented to obtain colour local image features. The evaluation of these features is then covered in Chapter 4, which details the experiments carried out on local image matching. The first experiments look at the performance of the various detectors, examining their ability to reliably detect the same local regions of a scene across varying imaging conditions. A correlation study then follows, that examines the data from the detection experiments and quantifies the level of effective correlation between the luminance and colour invariants. The third set of experiments evaluate the colour descriptors in a feature matching task. The number of correctly matched descriptors and the matching score are presented. The last part of the chapter, is a study on colour feature fusion for local feature matching. Two fusion strategies are outlined along with the experimental results.

Chapter 5 provides the last set of contributions, which examine the object class recognition aspects of the research. The chapter discusses relevant background information on BOVW, and outlines the recognition pipeline that is implemented in this work. Two experimental studies are presented, the first evaluates the recognition performance of individual gradient types. In the second experiment, the proposed fusion techniques for BOVW are evaluated and compared with the standard random grayscale dense sampling approach used in the literature. The thesis ends with Chapter 6, which summarises the contributions of the research and outlines some possible directions for future work.

Literature Review 2

Since colour is an important component for the distinction between objects, a large body of work has been proposed in the literature for utilising colour information in recognition and classification tasks. However the high variability of colour image values from a scene under varying imaging conditions, necessitates a solution to the colour constancy problem. This has impeded the full utilisation of colour for general unconstrained applications. Colour has had more success and attention in the field of image retrieval, especially in tasks that do not require a solution to be robust to occlusions or varying scales and imaging viewpoints. Such problems have traditionally been addressed with colour description approaches that neglect the geometrical characteristics of objects. They globally extract zero-order image representations like colour histograms and employ colour spaces that are perceptually uniform (e.g. CIE Lab) to suppress the level of variation of the colour values. For more general recognition tasks however, a local feature extraction method is necessary in order to provide robustness and invariance to geometrical variations such as translation, rotation, scaling, and affine/projective transformations. Local features are generally based on geometrical invariant approaches and extracted around highly informative regions like corners or blobs.

The feature extraction process involves two parts: The detection phase where local salient image regions or keypoints are identified (location and scale) and the description phase, where each detected region is characterised with a discriminative numerical signature (e.g. a histogram descriptor). Most of the geometrical invariant local feature extraction approaches in the literature are based on grayscale intensity information, as colour adds another layer of difficulty represented in the constancy problem. Colour also normally increases the

computational load of a vision algorithm, and unless utilised appropriately, the added benefits of using colour can be negligible. If colour is utilized appropriately however, the improved discriminative performance can justify the extra computational costs incurred.

Colour counterparts have been available to the community for some time although not widely used in general applications. In the context of local feature extraction, features should also be invariant with respect to photometric variations such as illumination direction, illumination intensity, illumination colour, and shadows and highlights. Various colour photometric invariants have thus been proposed to maximize the robustness to these variations. However, most of them have not received sufficient attention, making their contribution not fully explored in the context of local features. In works that use colour for local feature extraction, the majority apply colour in the description phase. Such colour descriptions can be either geometric approaches, or non-geometric based. The latter are usually used in conjunction with grayscale-based geometric descriptors.

In the past few years, BOVW approaches have significantly advanced object class recognition results (Chatfield et al., 2011). Most methods use the well-known SIFT descriptor (Lowe, 2004), thus they are based on local intensity shape (geometric) information. It has been shown that colour can also be a very useful cue within the BOVW framework for some image classification tasks (Van De Sande et al., 2010, Burghouts and Geusebroek, 2009, Vigo et al., 2010a, Van de Weijer and Schmid, 2006b). However, the efficient combination of multiple image cues is still an open problem because the relevance of each individual cue (colour, shape, texture, etc.) is highly dependent on the importance of colour in the data set (Khan et al., 2009). For example, colour information is necessary for discriminating football players from two different teams. On the other hand shape information is more essential to separate yellow bananas from yellow apples, while both types of cues are required to discriminate between types of flowers. The performance gain obtained by colour ranges from gains of up to 20% (Van de Weijer and Khan, 2013), on colour-dominant flower and sports datasets, to only a few percent on the shape-dominant PASCAL Visual Objects Challenge (VOC) data set. Colour can be applied to the BOVW framework in

two stages. Firstly, feature detection can be enhanced by choosing highly informative colour regions. Secondly, the feature description phase which typically focuses on shape, can be improved with a colour description of the local feature. Although both approaches have been shown to improve results, the combined merits have not been widely evaluated in the literature.

Only a small minority of previous works have utilised colour in both the detection and description phases (Abdel-Hakim and Farag, 2006, Gossow et al., 2010, Krylov et al., 2012), some use colour for feature detection but grayscale intensity for the description (Stöttinger et al., 2012, Vigo et al., 2010a). However, most approaches localise interest points first with a grayscale detector and then apply a colour descriptor, such studies have been applied to image feature matching (Van de Weijer and Schmid, 2006b, Burghouts and Geusebroek, 2009, Fan et al., 2009, Van De Sande et al., 2010, Jalilvand et al., 2011, Krylov and Sorokin, 2011, Song et al., 2013), image retrieval (Van de Weijer and Schmid, 2006b) and object class recognition (Burghouts and Geusebroek, 2009, Van De Sande et al., 2010, Khan et al., 2009). Other object class recognition studies apply colour description on densely sampled regions (Bosch et al., 2008, Wengert et al., 2011, Chu and Smeulders, 2012, Zhang et al., 2012). This thesis focuses on applying colour to both the detection and description of local image features, and the remainder of this review chapter will outline the most relevant works from the literature that utilise colour in the context of local feature extraction for the application of image matching and recognition.

Section 2.1 lays the foundation for local feature detection and the reasoning behind the direction that was taken for the development of the colour detector used in this research. Section 2.2 outlines the various colour interest point detectors that have been implemented in the literature, explaining their benefits and shortcomings. Section 2.3 briefly discusses the trend of local grayscale descriptors, covering the most important descriptors of the state of the art, and the justification for choosing the descriptor utilised in this research. In Section 2.4, the colour descriptors proposed in the literature are discussed. Section 2.5 explains the feature matching evaluation framework employed in this research, and all the datasets that are used. The chapter then ends in Section 2.6 with a recapitulation of the literature, putting it in context with the work carried

out in this research and highlighting the necessity for its more comprehensive evaluation of colour invariant local features.

2.1 Luminance Detectors

The Harris detector was introduced in 1988 and it has become the most widely used corner detector. It locates corners reliably using the eigenvalues of the second-moment matrix (also called the structure tensor or auto-correlation matrix), but its use is limited since it is not scale invariant (Mikolajczyk and Schmid, 2001). To have scale-invariance in local features, Lindeberg (1998) proposed a more unified concept for automatic scale selection. This approach detects blob-like structures at their characteristic scales using the circularly symmetric scale-normalised LoG operator. It obtains a scale-space stack representation by successively filtering the original input image with LoGs having Gaussian derivative kernels of varying standard deviations. By searching for local maxima across scale-space in an image stack of LoG responses, it is then possible to find the characteristic scale of an interest point.

Mikolajczyk and Schmid combined this scale-space approach with the Harris corner detector and refined it to create two robust and scale-invariant feature detectors, the Harris-Laplace (Mikolajczyk and Schmid, 2001, 2004), and the Hessian-Laplace (Mikolajczyk and Schmid, 2004). The Harris-Laplace uses the scale-adapted Harris measure to locate corners of various sizes. The Hessian-Laplace applies the determinant of the Hessian matrix to detect blob-like structures. For both detectors, the scale normalised LoG response determines the scale of the extracted interest points.

Matas et al. (2004), introduced the Maximally Stable Extremal Regions (MSER) detector, which extracts homogeneous intensity regions with a watershed-like segmentation algorithm. MSER was amongst the top performing detectors identified by Mikolajczyk et al. (2005b), working best for structured scenes that facilitate segmentation. It is amongst the fastest of the affine-invariant detectors and has mostly been used in wide baseline matching. MSER provides relatively few blobs but it is robust to geometric transformations, however it performs poorly under lighting distortions (Mikolajczyk and Schmid, 2005), and for object class recognition (Mikolajczyk et al., 2005a).

The trend during the last number of years, has been to create more computationally efficient detectors using essentially the same theoretical foundations as the early pioneering work. By increasing the complexity of the algorithms, it has advanced their performance and robustness to imaging distortions. Lowe (2004) proposed a more efficient approach for blob detection that approximates the LoG operator with the Difference of Gaussian (DoG). Using the DoG detector can significantly accelerate the detection without causing a substantial reduction in accuracy. This work introduced the well known SIFT approach, and arguably it has been to date the most robust local image feature that has been developed. Continuing the trend of improving computational efficiency, Bay et al. (2008) proposed a fast Hessian detector that detects blob-like objects more efficiently than DoG, their detection-description approach is named SURF. SURF uses integral images and haar-wavelet operators to approximate the determinant of the Hessian which, similarly to the LoG, also estimates the characteristic scale of a blob (Lindeberg, 1998). The accuracy of the scale estimation of these three methods (LoG, DoG and Hessian) largely depends on the selection of the scale sampling rate (Lowe, 2004).

A more recent corner detector, popular for real-time applications was proposed by Rosten et al. (2010). It is named the FAST detector, and builds on similar concepts to the classic SUSAN (Smith and Brady, 1997) corner detector. This type of detector has high spatial precision and low computational costs, but it lacks scale invariance and is less robust to high viewpoint distortions. Various other works have improved on its scale invariance capability, including AGAST (Mair et al., 2010) and Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011). Despite their real-time capabilities, these detectors are less compatible for a colour adaptation. Compatible approaches utilise raw multichannel colour gradients that can then be further combined and manipulated to introduce photometric invariance.

Despite its introduction two decades ago, the theory underpinning the Harris-Laplace detector is still intensively researched to achieve better robustness and invariance to image distortions. However it has been less popular than more recent techniques, that not only are computationally more efficient but also employ more sophisticated algorithms for point localisation. Despite this, in comparative studies involving different detectors (Mikolajczyk and Schmid,

2004, Mikolajczyk et al., 2005b), the performance of the Harris-Laplace was comparable to the top performers, with a good balance between localisation accuracy, repeatability scores and number of points extracted. The Hessian-Laplace (Mikolajczyk and Schmid, 2004), detects blobs instead of corners and has in general more stability and higher repeatability than the Harris-based counterpart. This is so because using the determinant of the Hessian reduces the detection of elongated ill-localised structures (Bay et al., 2008). They both detect similar locations however, and some approaches prefer to use the Harris-Laplace for visual recognition tasks (Zhang et al., 2007, Stöttinger et al., 2012), as the Hessian generates additional interest points that reduces the distinctiveness of the overall set of extracted features and can thus lead to a decreased probability of good matches. Second-order derivatives are also necessary for the Hessian, which could lead to inaccurate point localisation when using colour data which is inclined to contain more noise. For the aforementioned reasons, the Harris-Laplace was chosen for the development of the colour detection in this research.

2.2 Colour Detectors

The most successful grayscale intensity local image features are gradient-based, and rely on scale-invariant corner and blob detection. For this reason the majority of previous works, along with this research, focus on colour feature detection with gradient-based approaches. In the case of colour detectors, the most stable and robust to illumination variations as shown in Gouet and Boujemaa (2001), have been based on the colour Harris introduced by Montesinos et al. (1998). That approach replaces the intensity gradients in the second-moment matrix, with summations of squared RGB gradients. Van de Weijer et al. (2005) extended the colour Harris by proposing a set of photometric variants and quasi-invariant gradients which are less susceptible to shadows and specularities. Their evaluation focused on the accuracy of Canny edge detection when using their invariant gradients. Van de Weijer et al. (2006a) then introduced two full-invariants, which were evaluated by measuring the stability of non-scale-invariant Harris corner detection when subjected to varying levels of Gaussian noise.

Faille (2005) proposes a colour Harris corner detector based on the *m*-colour ratios of Gevers and Smeulders (1999) which is invariant to specularities, shading and colour illumination. However the method uses fixed scales for matching images under illumination distortions, making their interest points non-scale-invariant. Unnikrishnan and Hebert (2006) detect scale and rotation invariant points with the LoG operator using two illuminant invariant scalar functions, one invariant to a 3x3 perturbation of the RGB space and one invariant to independent scalings of the channels. Results showed that the version that is only invariant to independent scalings of the RGB channels is better under rotation and scale changes. A limitation of their study is the lack of an evaluation of their detector under viewpoint changes, and while they demonstrate that their colour LoG can work better than an intensity LoG detector, they do not perform a comparative evaluation.

Forssén (2007) proposes a colour extension of the MSER detector (Matas et al., 2004). It achieves higher repeatability rates than the original MSER for blurring and viewpoint distortions, but it is not evaluated for illumination changes. A similar approach is proposed by Penas and Shapiro (2009), that utilises the HSV colour space for their MSER adaptation and evaluates it for image matching and object classification. The authors obtain a gain in precision with respect to MSER for the Caltech 256 dataset (Griffin et al., 2004), but their image matching evaluation on the Oxford dataset (Mikolajczyk, 2004) is unclear. Their results show if a homography could be estimated between image pairs, but do not evaluate the repeatability or the number of correct point correspondences obtained, which has become the standard method for evaluating detectors. Furthermore, this is another study in which robustness to illumination changes is not examined.

Ming and Ma (2007) propose a multi-scale colour blob detector by substituting the derivatives in the Hessian matrix (for point localisation) by a weighted sum of RGB derivatives, and use the LoG for scale selection. Their weighting for each colour channel per pixel is the normalisation of the channel with the intensity. Their evaluation of the detection is not explained however, and only 5 images are used for their experiment. Shi et al. (2008) argue that with such an approach, the summation can lead to derivatives being cancelled out. They instead propose the use of quaternions to overcome this flaw, though this re-

quires the calculation of the eigenvalues from the quaternion Hessian matrix, which is computationally demanding (Le Bihan and Sangwine, 2003).

In the study of Gossow et al. (2010), the authors detect colour SURF points by analysing the determinant of the Hessian on separate colour channels as opposed to just the intensity channel like the standard SURF approach. Their approach also contains a colour description element which will be discussed further in Section 2.4. In another colour blob detector, Corso and Hager (2005) locate interest points in DoG responses obtained from three linear projections of RGB space. The technique essentially locates stable regions across scale-space that are homogeneous in colour content. Their technique provides scale, translation and affine invariance. The evaluation was performed on an image matching task and compared results when the images were subjected to changes in aspect ratios. The colour technique was inferior to an intensity-based approach on undistorted images, and only when the aspect ratio was halved did the colour achieve a marginally better performance.

Van de Weijer et al. (2006b) study colour derivative statistics from large image datasets and show that the distributions are dominated by a principal axis of maximum variation caused by the luminance intensity, and two minor axes of chromatic changes. This implies that changes in intensity are more probable to occur than chromatic changes and the authors argue that luminance intensity therefore has less information content. They estimate a 3x3 diagonal matrix that transforms an image to boost the effect of colour derivatives, and name their method colour saliency boosting. The boosting matrix parameters are estimated such that the original colour space of an image is rotated and aligned to the axis of major variation of the trained dataset. The approach transforms the original distribution to a more homogeneous one, aiming for intensity and chromatic changes to have more equal information content. The strength of the gradients of the boosted image relate to the saliency of the data, and the authors claim that this remapping increases the probability of detecting salient colour points.

Sebe et al. (2006) evaluate a Harris-Affine scale invariant colour detector using the saliency boosting of Van de Weijer et al. (2006b) on the Opponent Colour Space (OCS) and the m-colour ratio space (Gevers and Smeulders, 1999). They

perform feature detection experiments on 5 sequences of the Oxford dataset. The *m*-colour ratio detection performed worse than the opponent colour, and thus will not be considered for this research. The opponent colour results are mixed when compared with the grayscale detector. Under blurring distortions, the colour was 10% superior except for the most intense condition. Under all illumination variations however, the grayscale intensity performed better. Both grayscale and colour opponent detectors performed comparably for sequences with viewpoint or scale/orientation distortions.

Stöttinger et al. (2007) adapt the Harris-Laplace detector to utilise colour from various colour space representations: RGB, Hue-Saturation-Intensity (HSI) and OCS. Harris corners are identified by summing colour channel spatial derivatives, and their characteristic scale is estimated using the LoG operator on a boosted image. This boosted image is obtained via a Principal Component Analysis (PCA) of the covariance matrix of colour channel derivatives from the entire image (after a transformation to a chosen colour space), the final single-channel image results from the dot product of the principal eigenvector with the original 3-channel colour image. This work is continued by Stöttinger et al. (2009), where Harris-Laplace interest points obtained from the HSI colour space are named Light Invariant Colour (LIC) points, they also evaluate a colour boosted detector on the OCS. Their point detection experiments use the Oxford dataset, they run an image retrieval evaluation on the Amsterdam Library of Object Images (ALOI) dataset (Geusebroek et al., 2005), and lastly object class recognition is performed on the PASCAL VOC 2007 (Everingham et al., 2007) using grayscale SIFT descriptors and BOVW. The LIC points perform better on the detection repeatability compared to the grayscale Harris-Laplace. In the image retrieval experiments it also performs significantly better when 500 or less points are extracted per image. Recognition precision is comparable or marginally improved to the grayscale detection, but only half the amount of colour points are necessary. The final contribution in this line of work is presented by Stöttinger et al. (2012), though this is very similar to the previous work they carried out. In that study the LIC corners are obtained from summing saturation derivatives with hue derivatives which are weighted with the value of the saturation. While the corners are detected using colour information, the scale selection for the corners is estimated using the LoG operator on a boosted

single-channel image. It is not examined however, what role this boosting has on the performance of the extracted features, and how much can be attributed to the LIC colour invariant. The recognition experiments evaluate other descriptors other than SIFT and compare dense feature sampling to sparse LIC points. The recognition results show that dense sampling is the best performer overall but by a small margin, however using a grayscale Harris-Laplace detector gave comparable results to LIC. In the detection experiments using images under illumination variation, colour boosted points prove to be less repeatable than the LIC points, due to the saliency function being sensitive to luminance changes.

Vigo et al. (2010b) extend the LoG detector to utilise colour saliency by means of an adaptation to the colour saliency boosting of (Van de Weijer et al., 2006b). To estimate the boosting function for a particular image (and not from an entire dataset) the authors use Independent Component Analysis (ICA) on a covariance matrix of image derivatives. Their LoG detector uses RGB gradients and is evaluated on feature matching tasks on the Oxford dataset, with grayscale SIFT descriptors and the C-SIFT from Burghouts and Geusebroek (2009). Results indicate the colour points achieve an average improvement of 10% over a grayscale LoG detector. Vigo et al. (2010a) focus on improving image class recognition by using their colour boosted detector from Vigo et al. (2010b). Their recognition uses the Color Attention method (Khan et al., 2009), which allows to sample regions from an image in varying spatial concentrations by analysing the image's colour attention saliency map. The description step uses SIFT to build a shape vocabulary and colour names (Van de Weijer et al., 2009) with hue descriptors (Van de Weijer and Schmid, 2006b) for the colour vocabulary. Standard SIFT-based BOVW is performed and compared to using the colour attention method which combines both shape and colour information at the recognition stage. Results on the Flowers dataset (Nilsback and Zisserman, 2006) with SIFT BOVW, show that using only the boosted colour detector gives a 1% precision improvement over using only an intensity detector, and identical results when using both detectors simultaneously. The Colour Attention approach achieves a 16% improvement over BOVW, and using a combined intensity and boosted detection is 4% better than using only an intensity detector. On the PASCAL VOC 2009 experiments, using the combined detection is always superior to only boosted or intensity detection. The colour attention method is here again better

than BOVW, but by a lesser margin of 4%.

Despite various works applying a colour boosting function, minor improvement results are obtained in the BOVW recognition study of Vigo et al. (2010a). Results from Sebe et al. (2006) and Stöttinger et al. (2012) indicate that in image matching tasks the boosting is not robust to illumination distortions, which is an invariance that is of interest to this research. Furthermore, this boosting is global in nature and to estimate the boosting parameters it requires either the use of PCA to decorrelate the colour channels of an image or the derivative analysis of large datasets. This work focuses instead on evaluating raw local colour invariant gradients on their own merits, without using any higher level derivative statistics.

2.3 Luminance Descriptors

The keypoint detection is the fist step of the feature extraction process, which is required in order to then describe that image region with a feature vector. The descriptor vector/histogram must contain a representation that is robust and invariant to varying imaging conditions in order for the feature to be matched across multiple images. Although a good detector is important for allowing salient and potentially discriminative regions to be located, it is the descriptor which ensures that the feature can actually be utilised for practical vision tasks. Its vital importance has inspired a myriad of solutions in the literature, of which the most important in the context of local feature matching will be discussed here.

The most popular descriptor and arguably the most robust, is the SIFT proposed by Lowe (2004). It is based on obtaining Histograms of Oriented Gradients (HOG) from a grid surrounding the centre of the keypoint. The full dimensionality of the resulting descriptor after the concatenation of the HOGs is 128, and its discriminative power and robustness have made it the reference descriptor for the community. Since its introduction, the majority of the efforts have been to design a descriptor that performs comparatively to SIFT, but with a lower computational overhead and that can facilitate a faster matching algorithm. An example of the large SIFT-like family of descriptors that have been proposed, is the PCA-SIFT of Ke and Sukthankar (2004) that

reduces the descriptor to 36 dimensions using PCA. While the matching time is reduced, the time required to decrease the dimensions results in small gains in efficiency and a drop in the discriminative ability of the resulting descriptor.

One of the most popular descriptors after the introduction of SIFT, has been the SURF proposed by Bay et al. (2008). It cleverly approximates both the detection and description components of SIFT by applying Haar-wavelet filtering to obtain the HOGs and integral images to approximate the determinant of the Hessian needed to locate the keypoints. SURF's performance is comparable with that of SIFT, and much depends on their implementation and on the data on which they are tested. In general though, SIFT is regarded to be more discriminative and robust to imaging distortions as it has inherently less approximated calculations. SURF is however much faster than SIFT, although it still cannot be generally classified as a real-time descriptor when using standard hardware. Some improvements to SURF were proposed by Agrawal et al. (2008), with the Modified-SURF (M-SURF) descriptor and using the center-surround detector (CenSurE). M-SURF employs a two-stage Gaussian weighting scheme, making it more robust and better suited to suppressing descriptor boundary effects. Alcantarilla et al. (2012) introduce the KAZE feature, an adaptation of the SURF detection and description approach within a non-linear scale space framework. M-SURF is used for their description, and by applying non-linear diffusion filters they can obtain scale invariant features that improve the repeatability and distinctiveness of previous features that are based on the Gaussian scale space.

The novel description development of recent years (similarly to the detectors) has steered away from gradient-based techniques and focused on creating more compact and computationally efficient descriptors. This family of descriptors are comparison-based, and are derivations of the Local Binary Pattern (LBP) descriptor of Ojala et al. (1996). To generate an LBP-like descriptor, pixel intensities are compared from pairs of neighbourhood pixels around a centre pixel, the comparison results in a binary string. The size of the neighbourhood and the sampling strategy to generate the binary descriptor changes from method to method. The significance of binary descriptors is that their matching strategy is significantly faster than previous SIFT-like descriptors. Instead of utilising a distance metric like the Euclidean Distance to match descriptors from differ-

ent images, it is possible to use the Hamming distance (a bitwise XOR and a bit count operation) with a binary descriptor, which is computationally considerably more efficient. These runtime advantages make binary descriptors better adapted for real-time applications and mobile devices. Amongst the most well-known binary descriptors in the literature are: Binary Robust Independent Elementary Feature (BRIEF), Oriented Fast and Rotated BRIEF (ORB), Binary Robust Invariant Scalable Keypoints (BRISK), Fast Retina Keypoint (FREAK) and Local Difference Binary (LDB).

Calonder et al. (2010) propose the BRIEF descriptor, obtained via a Gaussian distributed random sampling of 512 pixel pair intensity comparisons. The resulting descriptor is not invariant to rotation or scale changes on its own, it has to be linked to information from the detector. The next development was by Rublee et al. (2011), which improved BRIEF's invariance to noise and rotation. Further improvements in scale and rotational invariance came with BRISK, proposed by Leutenegger et al. (2011) and which employs a concentric equally spaced circular sampling pattern. A more efficient retinal sampling pattern is used for FREAK (Alahi et al., 2012), this sampling sets a higher density on the centre of the region which then drops outwards exponentially similarly to the photoreceptors of the eye. One of the most recent descriptors is the LDB proposed by Yang and Cheng (2014), which improves both the matching accuracy and matching speed of previous binary descriptors.

Many versions of implementations and comparisons exist of all the aforementioned descriptors, but in general it is clear to see the difference between binary and SIFT-like descriptors. BRISK's performance is comparable but not superior overall to SIFT or SURF (Leutenegger et al., 2011). FREAK performs slightly worse than BRISK while achieving better computational efficiency (Alahi et al., 2012), the paper's results also show SIFT to be overall better than SURF. Yang and Cheng (2014) compare LDB-64 and LDB-32 with ORB-32, BRISK-64, FREAK-64 and SURF-64 (descriptor versions of 32 and 64 dimensions), the matching performance of LDB was better on all the tested image-sets. Despite that performance, LDB was not compared with the full SURF or SIFT descriptor which comprise of 128 dimensions. In general it is evident that binary descriptors have closed the performance gap and are sufficient for most real-time applications, but the traditional SIFT-like descriptors are more robust, distinct, and still the

best choice for applications needing invariance for more challenging imaging distortions. This is one of the main reasons for why SIFT is the descriptor of choice for this research. Additionally all colour invariant methods have employed gradient-based SURF or SIFT descriptors, in order to be compatible with the colour invariants of the literature.

2.4 Colour Descriptors

The use of colour for description purposes has received more attention than using colour for detection, specially in more recent years. Traditionally there have been three ways to introduce colour into feature descriptors, with non-geometric, semi-geometric and full geometric chromatic information. The first approach consists of normalising the colour channels to provide a level of photometric invariance and creating a zero-order histogram descriptor from the resulting chromatic information (Van de Weijer and Schmid, 2006b, Van De Sande et al., 2010, Mojsilovic, 2005, Finlayson et al., 1998). Due to the weak spatial and geometric information content within such descriptors, the reported results show that shape-based geometric descriptors like SIFT are generally always superior in image matching or recognition tasks. The second (semi-geometric) approach, successfully applied in Van de Weijer and Schmid (2006b), Tang et al. (2012), and Diplaros et al. (2006), extracts two types of descriptors from an interest region, one containing shape/geometric information and the other the chromatic histograms mentioned previously. The shape information uses intensity information, to generally obtain SIFT descriptors. These two types of descriptors then get concatenated, and the performance of the resulting descriptors are superior to using only the colour histograms. Such a fusion strategy is particularly favoured for image retrieval and recognition rather than for image feature matching applications. This is due to the invariance to standard imaging distortions being diluted by a non-geometric colour component, this effect impacts less on the recognition. The third way of applying colour to feature descriptors, is to extract colour descriptors like SIFT which will then contain compatible spatio-geometric and chromatic information in the same representation. This is achieved by either applying the descriptor algorithm directly onto individual colour channels and then concatenating the resulting descriptors (Bosch et al., 2006, Van De Sande et al., 2010), or developing colour

invariant gradients from multiple channels which can then be compatible with SIFT (Abdel-Hakim and Farag, 2006, Burghouts and Geusebroek, 2009). This section focuses on previous works that employ semi or fully geometric methods of introducing colour into local feature descriptors.

Van de Weijer and Schmid (2006b) combine colour and shape information by concatenating histograms of hue or photometric invariant values with the grayscale SIFT descriptor. Hue is invariant to lighting geometry and specularities, however it is unstable near the gray axis (when saturation is low). Van de Weijer and Schmid (2006b) thus apply an error analysis to weigh the contribution of different hues while histogramming its values, it is based on how the certainty of the hue of a pixel is inversely proportional to its squared saturation. Their image matching evaluation is performed on three image-sets, two taken from the Oxford dataset (with viewpoint and illumination variations) and another image set with illumination variations. The matching results show that combining colour and shape is always significantly better than using the colour histograms alone. Compared to using grayscale SIFT on the viewpoint varying dataset, the colour hybrid descriptors perform 20% better. In the illumination varying sets however, colour improves the matching score by only 1-3%. They perform a further experiment of image classification on the Birds (Lazebnik et al., 2005) image dataset (6 classes, 600 total images) and a football (Van de Weijer and Schmid, 2006a) dataset (7 classes, 280 images). The classification results prove that adding non-geometric colour information to the SIFT descriptor improves its discriminative power, their approach is clearly superior to the grayscale SIFT in classification precision, and it is clearly better for this task than for image matching.

Abdel-Hakim and Farag (2006) propose a SIFT descriptor built using hue colour gradients based on a H-invariant from Geusebroek et al. (2001), which utilises the Gaussian Colour Model. They perform feature matching experiments on the ALOI dataset (Geusebroek et al., 2005); which contains images of objects under different illumination conditions. Feature matching results show that their CSIFT descriptor is more robust than the standard grayscale SIFT with respect to colour and photometric variations.

Burghouts and Geusebroek (2009) evaluate multiple colour invariants proposed by Geusebroek et al. (2001) for the purposes of feature matching (on the

ALOI dataset) and object class recognition (on PASCAL VOC 2006). The ALOI set comprises of 1,000 objects, imaged under different conditions: blurring, JPEG compression, illumination direction, viewpoint change and illumination colour. In their feature matching experiments, a non-standard and limited evaluation framework is employed. They use a grayscale Harris-Affine detector to find candidate local regions from each image of a set, from those only one region is manually selected (the most visible across all the varying viewpoint conditions of the set). Descriptors are then extracted for the selected singular regions across all imaging conditions of the set, and repeated for all the sets of objects in the dataset. The region at the first imaging condition, is matched against regions from other conditions across multiple other objects (including a region from the same object). The matching aim is to find the region belonging to the same object. Precision-Recall curves are calculated by matching the descriptors of each region with 100 or 500 other random regions from the rest of the dataset, in a 1000-fold cross validation. The regions are described by concatenating SIFT descriptors from different invariants and reducing the dimensionality of the overall descriptor to match the 128 dimensions of the original SIFT. The C-Invariant performs best overall for the matching tasks, in the object class recognition experiments it was compared against the standard grayscale SIFT and also achieved a better performance.

Various state of the art colour image descriptors are evaluated in the study of Van De Sande et al. (2010). The study uses the ALOI dataset to evaluate robustness to viewpoint and illumination changes and the PASCAL VOC 2007 for object class recognition. Amongst the descriptors tested were the HSV-SIFT used by Bosch et al. (2008), the colour moments used by Mindru et al. (2004), Hue-SIFT (Van de Weijer and Schmid, 2006b), C-invariant SIFT (C-SIFT) (Burghouts and Geusebroek, 2009) and Opp-SIFT, which is a concatenation of SIFT descriptors of the O1, O2 and O3 opponent channels of the OCS. Results showed that SIFT-based colour descriptors outperform histogram-based and moment-based descriptors. Additionally while the relative performance of the descriptors was data-specific, C-SIFT performed best in the recognition task, achieving a 4% gain in precision over standard SIFT.

Gossow et al. (2010) extend the SURF descriptor to the colour domain using the C and W colour invariants proposed by Geusebroek et al. (2001). The

descriptors are formed by concatenating SURF descriptors taken from different colour channels and invariants, and the optimum approach uses different invariant combinations for detection and description. Their feature matching is evaluated on a subset of the ALOI dataset and the Oxford dataset (Mikolajczyk, 2004), and results indicate that their best COLOR SURF candidate achieves better robustness to photometric distortions. The difference in the results are marginal however, and the implementation of their detector is not sufficiently explained.

Cui et al. (2010) propose the Perception-based Color SIFT descriptor (PC-SIFT), which provides geometrical and photometric invariance and is built using the perception-based colour space of Chong et al. (2008). They detect DoG interest points from each channel of the colour space, and the descriptors are formed from 3 dimensional colour gradients. Their feature matching experiments are carried out on the ALOI dataset and evaluate the robustness against colour lighting changes and illumination direction changes. Results show that PC-SIFT performed better compared with the CSIFT of Abdel-Hakim and Farag (2006) and grayscale SIFT. However their paper does not explain clearly how their SIFT descriptors are obtained from the 3 dimensional gradients.

Krylov and Sorokin (2011) propose a colour description extension of their grayscale keypoint extraction method based on Gauss-Laguerre circular harmonic functions. They utilise the invariant theory proposed by Geusebroek et al. (2001) to obtain colour descriptors. The results that are shown in their study are very limited however, and the usefulness of the approach is thus unclear. The authors follow up their work in the study (Krylov et al., 2012), which introduces a colour blob detector using the Hessian matrix. The detector sums weighted RGB spatial derivatives to obtain the gradients for the Hessian matrix, the weight per pixel for each colour channel derives from its 2nd order spatial derivative magnitude. Their evaluation dataset contains two parts, one with 28 image pairs that allow keypoints to be detected stably before and after grayscale conversion, and one part with 16 image pairs in which keypoints become less distinguishable after grayscale conversion. Detection results show a marginal improvement in the first dataset using the colour keypoints, and a significant improvement in the second set of images, but only after the number of extracted points per image is above 2000. In general, the authors claim that

using colour for both detection and description is the most beneficial strategy, as when using a grayscale descriptor the difference between using a colour or grayscale detector becomes minimal.

Jalilvand et al. (2011) use the C invariant from (Geusebroek et al., 2001) to create a colour SURF (Bay et al., 2008) descriptor. They utilise the ALOI dataset in their feature matching experiments and obtain better precision and recall results than the standard SURF. Fan et al. (2009) propose a colour-grayscale hybrid descriptor for local feature matching with one part composed of the standard grayscale SURF, and the other consisting of a colour histogram quantised from the YUV colour space with a Gaussian kernel. They employ different matching metric distances, the Euclidean distance matches the SURF and the Bhattacharyya distance matches the colour histogram. The colour histogram is only used to match the features that fail the matching threshold of the SURF descriptors. In a similar study, Geodemé et al. (2005) propose using a 3D colour descriptor based on colour moments, to filter out incorrect grayscale SIFT matches.

Song et al. (2013) propose compact local descriptors that encodes the colours in a region and their spatial distribution. These descriptors are designed to be robust to photometric variations that can be modelled by an affine transform in RGB colour space. They characterise each pixel of an interest region with 5 coordinates, two spatial (x, y) and three for the colour space (R, G, B), and approximate two affine transforms to go between the image and colour space and vice versa. Each of those transforms is used to generate two colour descriptors. For example for their ITC descriptor (image to colour), the pixels of elliptical interest regions are mapped to parallelograms in the colour space, and the colour descriptor is composed of the corner locations of these parallelograms. ITC has 36 dimensions and is robust to affine spatial distortions, CTI (colour to image) has 48 and is robust to photometric variations. They perform feature matching experiments on ALOI using the evaluation framework of Burghouts and Geusebroek (2009), and object class recognition on the Birds (Lazebnik et al., 2005), Flowers (Nilsback and Zisserman, 2006) and Football (Van de Weijer and Schmid, 2006a) datasets. Their approach is compared with RGB-SIFT, Hue-SIFT, Opp-SIFT, HSV-SIFT, C-SIFT and standard SIFT. The best results on the matching task are obtained by concatenating ITC and CTI, whereas the overall

top performer (best in 2 of the 3 datasets) in the recognition experiments is ITC. Their descriptors performed particularly well considering the significantly smaller size in their dimensions.

Apart from the aforementioned techniques that obtain illumination invariant image descriptors, there exists another approach in computer vision for solving the problem of varying illumination conditions across images. This field is called Colour Constancy, and focuses on estimating the colour of the light source of an imaged scene (Gijsenij et al., 2011). Once the light source is determined, the image can be corrected to appear as if it was taken under an ideal white canonical light source.

Although a significant body of colour constancy work exists in the literature, few have studied the effects of colour constancy correction for local image feature extraction applications. Kanan et al. (2010) study the effects of standard colour constancy algorithms on face and object recognition (using the ALOI dataset), and compare recognition rates by extracting SIFT features on various colour spaces with and without the colour constancy correction. Results on the ALOI dataset indicate that the results from the colour corrected descriptors improve the recognition accuracy by 2%, compared to using non-corrected RGB-SIFT descriptors. In a similar study by Joze and Drew (2010), that utilises the BOVW recognition approach on the PASCAL VOC 2007 dataset; the same observation is concluded regarding the comparison of standard RGB-SIFT and colour constancy corrected counterparts. That study however, found that C-SIFT and Opp-SIFT performed better than the colour constancy corrected techniques.

The constancy approach was thus not followed in this research, due to the uncertain benefits of using standard colour constancy algorithms, and because more complex constancy techniques require datasets calibration and performing statistical analysis similar to the colour boosting strategies. The concluding remarks of the literature review are given in Section 2.6, which summarise the differences between the state of the art and the work proposed in this research. They will serve to more easily highlight the contributions of this work compared to the literature and provide the reasoning for the direction taken here.

2.5 Evaluation Framework of Image Features

Mikolajczyk and Schmid (2005) propose a framework to evaluate the quality of local interest point detection and matching using robust metrics, where the first image of a sequence is matched in turn with images of varying levels of image distortions. This framework simulates the conditions of a real-world feature matching application significantly better than the one employed by Burghouts and Geusebroek (2009). In the framework of Mikolajczyk and Schmid (2005), all the extracted features from an image are considered in the matching process. Since the homographies between the first image of each sequence and all subsequent images are known, these are used to identify which feature correspondences are correct. To provide standardised results and employ a robust local feature extraction evaluation, this research follows the same framework as it has become the standard method in the field. Three metrics are used in the local feature matching evaluation, the repeatability index (Section 2.5.1), the matching score and precision-recall curves (Section 2.5.2).

The local feature matching evaluations of this research are carried out on four datasets: The Oxford dataset (Mikolajczyk, 2004) which has become the de facto database to evaluate local grayscale features. The Middlebury Stereo dataset which is amongst the most widely used in its field. The ALOI objects database, utilised predominantly for image retrieval and studies dealing with illumination varying conditions. The fourth one is the PHOS dataset (Vonikakis et al., 2012), comprising of sets of objects imaged under varying illumination conditions and rarely used within the literature.

Despite the benefits of his evaluation approach, Krystian Mikolajczyk mentioned at his opening talk of the CVPR 2009 feature benchmark (Mikolajczyk et al., 2009), that his framework contained a number of drawbacks. Firstly that larger regions have more of an advantage compared to smaller ones. Secondly, that a dense feature extraction will always obtain higher repeatability rates. The framework and datasets that are compatible with it also do not allow for evaluating the applicability of local features to increasingly general conditions. A suitable way to test for those scenarios, is to run a large scale recognition

experiment which is also performed in this research. For the object class recognition experiments, this study uses the testing framework from the PASCAL VOC (Visual Objects Challenge) (Everingham et al., 2007), and the 2007 version of the dataset. Since the PASCAL VOC's first appearance it has become the standard and most popular framework for testing recognition approaches (especially BOVW). Only in more recent years have other datasets and challenges started to become more important in the image recognition community, such as the IMAGENET challenge (Russakovsky et al., 2014) that contains millions of images. These types of datasets can be used with Deep Learning approaches (Bengio, 2009, Simonyan and Zisserman, 2014), that utilise Neural Networks and have become to represent the state of the art in object class recognition. This research is not interested in developing state of the art machine learning algorithms however. It focuses on colour invariant local feature extraction, and the BOVW approach to evaluate the PASCAL VOC is still one of the most common and valid recognition techniques that utilises local feature extraction.

2.5.1 The Repeatability Index

The repeatability index is a metric proposed by Mikolajczyk and Schmid (2005), to measure the ability of a local feature detector to reliably detect the same points within an image scene across varying imaging conditions. It is defined as the ratio between the number of correct corresponding points between two images and the total number of possible points available to be matched. This total number is obtained from the minimum number of points (out of the two images) that occur in the overlapping scene area common to both images.

$$repeatability = \frac{\#correspondences}{\#totalpoints} \tag{2.1}$$

In the study of Mikolajczyk and Schmid (2005), two regions are said to correspond if their areas overlap by more than 60%. Each region is expressed as an ellipse, and the overlapping areas are calculated by mapping (with a homography) an ellipse from the first image to the corresponding frame of reference of the second image. Feature detectors have higher repeatability

rates when they detect regions of more uniqueness, leading to fewer points being unmatched. A detector can have high repeatability with low numbers of correspondences, but optimum detectors should have both high numbers of correspondences and repeatability rates.

2.5.2 Precision-Recall

The typical matching strategy found in the literature uses a particular error distance metric to identify the nearest neighbours (NN) of the descriptors between two images. The first NN of a descriptor in the corresponding image is then selected as a match, if the error distance ratio of the first two neighbours ($NN1/NN2$) is below a certain matching threshold. This study utilises the Euclidean distance as the error metric. By varying that threshold it is possible to obtain precision-recall curves for the matching results, which are calculated with Equations 2.2 and 2.3. These curves convey a more detailed representation of the distinctiveness of the descriptors, as a good descriptor should obtain a high precision for all matching thresholds.

$$recall = \frac{\#correctmatches}{\#correspondences} \tag{2.2}$$

$$precision = 1 - \frac{\#falsematches}{\#correct + falsematches} \tag{2.3}$$

Apart from the precision-recall curves, this study also provides results for the matching score. This allows to measure the maximum correct matches that each colour invariant can potentially achieve without the need to optimise the matching strategy. To obtain the matching score, all the first NNs are picked as candidate matches, and since the homographies between the two matching images are known, it is possible to quantify which of the matches are correct. The matching score is here defined as the percentage of correct descriptor matches with respect to the correct number of correspondences from the detection stage.

2.5.3 Datasets

Oxford

Mikolajczyk's Oxford dataset[1] consists of image-sets with various distortions: blurring, zoom and rotation, JPEG compression, illumination and viewpoint changes. All sets are used here (7 colour sets), except the black and white set which is not relevant to this study. The 6 images in each Oxford sequence are subjected to increasing levels of distortions, examples of images contained in the 7 sets are shown in Fig.2.1. The image set *leuven* contains illumination changes, *bikes* and *trees* are subjected to blurring, *bark* to zoom and rotation, the set *Jpeg Compression* is compressed with increasing levels, and *wall* and *graffiti* contain viewpoint distortions.

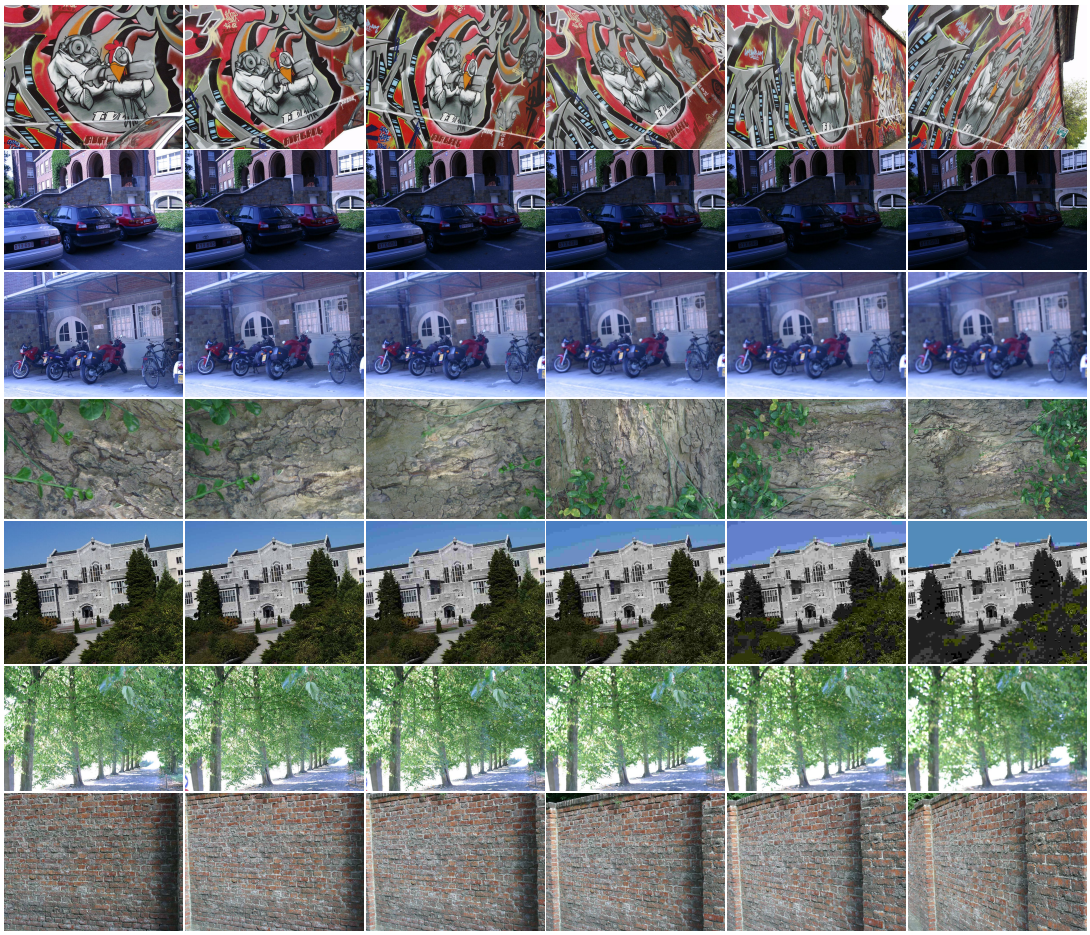


Figure 2.1: Oxford image sequences from the sets: *graffiti* (viewpoint), *leuven* (illumination), *bikes* (blurring), *bark* (scale orientation), *UBC* (JPEG compression), *trees* (blurring) and *wall* (viewpoint).

[1] www.robots.ox.ac.uk/ vgg/research/affine/

Middlebury

The Middlebury Stereo dataset[2] provided by Scharstein and Pal (Scharstein and Pal, 2007), consists of multiple sets of stereo images of natural scenes which vary in illumination conditions. In order to carry out the tests of this work, 5 sequences of 8 images (from different scenes) were compiled, which contain varying illumination but no viewpoint changes. The used image sequences are shown in Fig. 2.2.

Figure 2.2: Middlebury images sequences from the sets: *Art*, *Drumsticks*, *Dwarves*, *Moebius*, and *Monopoly*

[2]http://vision.middlebury.edu/stereo/data

Amsterdam Library of Object Images (ALOI)

The ALOI dataset[3] comprises of 1000 single objects (against a dark background) under supervised imaging conditions that include viewpoint changes, illumination direction changes and illuminant colour variations. For this research, 30 image sequences were selected of objects under 8 different illumination direction conditions, some examples are shown in Fig. 2.3. The 30 image sequences were selected by visual inspection, choosing the sets that were not too similar to each other and primarily those that contained visually richer scenes where the objects had sufficient levels of texture.

PHOS

The PHOS dataset[4] comprises of 15 sequences of multiple objects per scene (against a white background) under different illumination direction conditions from the same viewpoint. Here, 11 images per sequence were selected which visually appear to contain increasing levels of scene illumination, examples are shown in Fig. 2.4.

PASCAL VOC 2007

The PASCAL VOC (Visual Objects Challenge) has been for many years the most challenging dataset to perform object detection and classification tasks, only more recently have other more substantial datasets been introduced comprising up to a few million images. The VOC 2007 dataset[5] (Everingham et al., 2007) was chosen for this research, primarily because the majority of relevant colour studies utilised it (Biagio et al., 2014, Khan et al., 2013, Stöttinger et al., 2012, Khan et al., 2012, Zhang et al., 2012, Fernando et al., 2012, Van De Sande et al., 2010, Joze and Drew, 2010, Stöttinger et al., 2009, Khan et al., 2009), and that the recognition aspects of the VOC dataset have not changed significantly since 2007 as the numbers of classes has remained 20 and the latest version only has 1500 more images in total. Another reason for choosing the VOC dataset, is that it contains objects in real-world settings and allows to evaluate the colour features for general scenarios. Consequently the VOC is particularly challenging for colour approaches, since it contains many man-made objects that are mainly shape-dominant (Khan et al., 2009).

[3]http://aloi.science.uva.nl/
[4]http://utopia.duth.gr/ dchrisos/pubs/database2.html
[5]http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

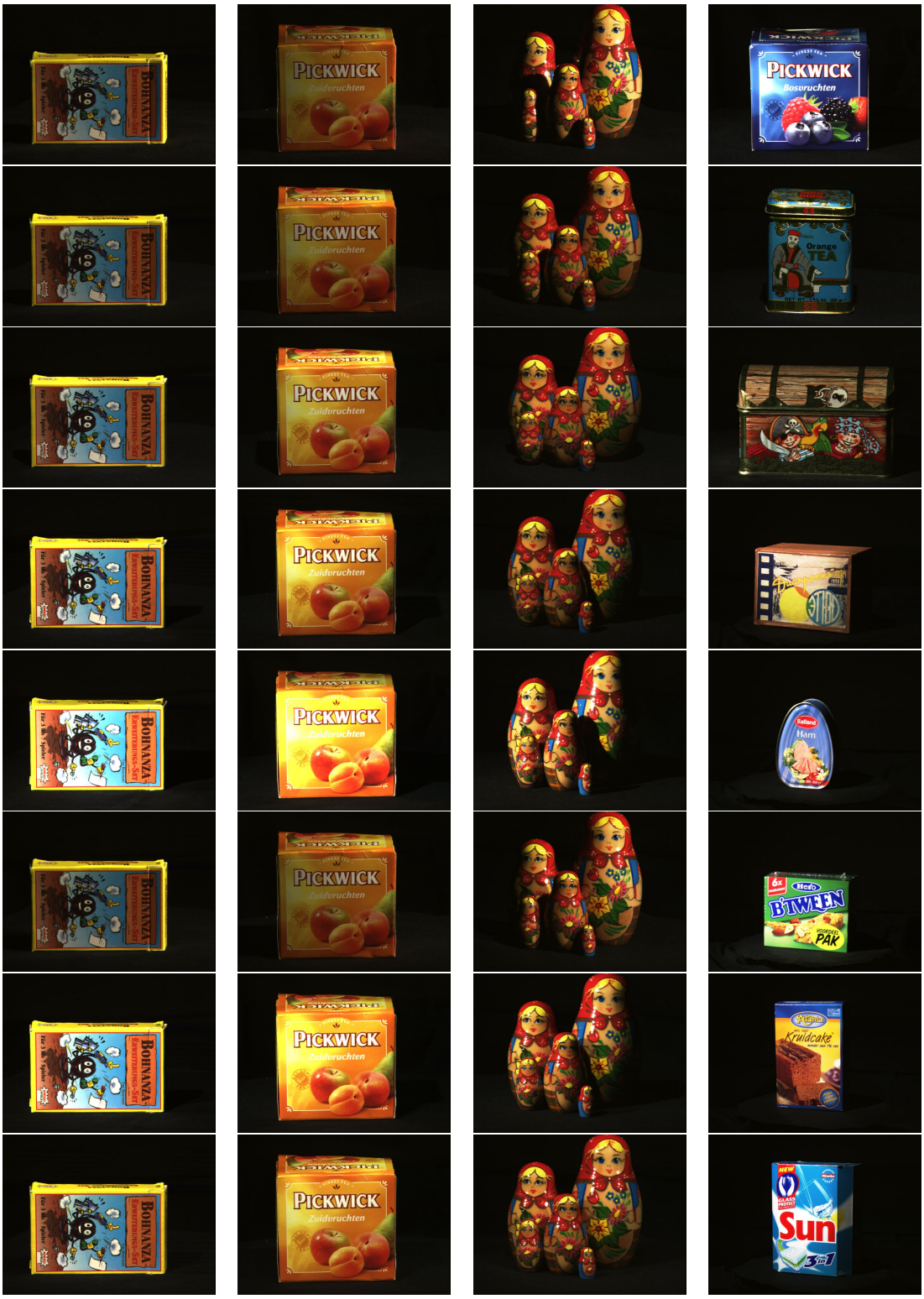


Figure 2.3: ALOI images sequences; the last column shows individual examples from 8 other sets.

Figure 2.4: Examples of image sequences from the PHOS dataset.

The focus of this research is on the image object class recognition task, and not the object detection, the set contains 20 object classes in total, comprising 5,011 training images and 4,952 test images, Figure 2.5 shows some examples of these. The VOC challenge stipulates to split the dataset into two parts, each containing approximately 50% of the images for the training/validation and 50% for the testing. The distribution of classes are equal across both parts, though the number of objects per class is not, i.e. there are more images of persons than of chairs etc. In total, the dataset contains 12,608 objects, and the assumption is that each test image contains at least one object of the class being searched for. The aim of the PASCAL VOC challenge is to query each image of the testing set, and detect the presence of all the 20 VOC classes in turn. The VOC challenge is evaluated by measuring the Average Precision (AP) that an algorithm obtains when querying for a particular class. The AP metric considers the ranked results from the retrieval, it is the average precision calculated at the position of all correct images from the ranked list. The first rank of this list is the image with the strongest probability of containing the searched class. The AP in geometric terms, is the area under the precision-recall curve. The final metric to evaluate the overall performance of an approach, is the mean Average Precision (mAP) from all of the 20 AP results.

Figure 2.5: PASCAL VOC 2007 image examples from all the 20 different classes.

In general the PASCAL VOC dataset is challenging for colour-based approaches. Apart from the many black and white, underexposed or blurred images, the difficulty lies in that the classes all contain similar intra-class geometric characteristics, but vary significantly in colour information. This is why it was the chosen dataset for this research, as it tests the ability of the colour invariants to function in a more general real world scenario.

2.6 Summary and Discussion

In this thesis, the colour invariants proposed by Van de Weijer et al. (2005, 2006a), and variations of those in Stöttinger et al. (2012) and Geusebroek et al. (2001), are utilised to create local image feature detectors and descriptors. The main reason why the invariants proposed by Van de Weijer et al. (2005, 2006a) are used in this work, is that they have not been applied before to local features despite showing promise in terms of their colour invariance theory. In the work of Stöttinger et al. (2012), the LIC gradients are used solely to locate corners, and a boosted image is used in the scale selection. This research utilises the LIC invariant in both corner detection and scale selection to isolate the benefits of the LIC invariant and evaluate it on its own merit. From the invariants proposed by Burghouts and Geusebroek (2009), only the C-SIFT has been previously used for recognition tasks. In this research, the top three types of invariants proposed by Burghouts and Geusebroek (2009) are implemented.

To achieve the 3 main goals of this research (outlined in the introduction of Chapter 1), all the aforementioned invariants need to be implemented and evaluated by addressing the limitations of their earlier adoptions. An important aspect of their evaluation, is for the invariants to be compared within the same robust testing framework, and ultimately discover if and where can they enhance feature matching and object recognition tasks. Here follows the main limitations to the studies carried out in the literature, which are then summarised in Table 2.1.

1 - Lack of Scale-Invariance

Some colour invariants were only implemented as edge or corner detectors (Van de Weijer et al., 2005, 2006a), or without scale invariance (Faille, 2005).

2 - Limited Distortions

Other studies did not test the robustness to all the standard set of varying imaging conditions. Unnikrishnan and Hebert (2006) did not test for viewpoint distortions and illumination changes were not taken into account in the study of Forssén (2007).

3 - Suboptimal Evaluation Framework

Not all local feature studies (Corso and Hager, 2005, Penas and Shapiro, 2009, Burghouts and Geusebroek, 2009, Van De Sande et al., 2010, Song et al., 2013, Krylov et al., 2012), have evaluated their approaches with the de facto standard robust framework of Mikolajczyk and Schmid (2005). For example, Burghouts and Geusebroek (2009) do not detect multiple regions per image and then match them across different distortion conditions. This does not robustly evaluate the tested colour invariants in a real world feature matching context. Several studies like those of Van De Sande et al. (2010) and Song et al. (2013) also followed the same approach.

4 - Inferior Image Data

Many previous studies were conducted using an inferior quantity of image data, compared to what is used in this research. A richer variety and greater number of images is necessary in order to test approaches under more realistic conditions. Examples of evaluations that were limited in this regard include those carried out by: Van de Weijer and Schmid (2006b), Ming and Ma (2007), Van De Sande et al. (2010), Krylov and Sorokin (2011), Krylov et al. (2012).

5 - Colour-biased Datasets

Some classification studies were only implemented on small-medium datasets which favoured colour-based techniques (Van de Weijer and Schmid, 2006b, Song et al., 2013).

6 - Few Datasets

The reliance of only testing on one dataset in the works of Sebe et al. (2006), Abdel-Hakim and Farag (2006), Vigo et al. (2010b), Cui et al. (2010), Jalilvand et al. (2011), limit the generality and confidence of their results.

7 - Colour Detection with Description

Only a few works have utilised colour in both the detection and description phases of their feature extraction (Abdel-Hakim and Farag, 2006, Gossow et al.,

2010, Krylov et al., 2012). As documented by Krylov et al. (2012), if a grayscale descriptor is used it will diminish the benefits of using a colour detector.

Table 2.1: Summary of the limitations in the literature.

Suboptimal Evaluation Framework	**Tests Lacking Scale Invariance**
Corso and Hager (2005); Jalilvand et al. (2011)	
Burghouts and Geusebroek (2009)	Faille (2005)
Van De Sande et al. (2010); Song et al. (2013)	Van de Weijer et al. (2005)
Penas and Shapiro (2009); Krylov et al. (2012)	Van de Weijer et al. (2006a)
Faille (2005); Abdel-Hakim and Farag (2006)	
Inferior Quantities of Data/Datasets	**Tests Lacking Illumination Invariance**
	Ming and Ma (2007); Corso and Hager (2005)
Ming and Ma (2007); Cui et al. (2010)	Forssén (2007); Krylov et al. (2012)
Van De Sande et al. (2010); Sebe et al. (2006)	Van de Weijer et al. (2005)
Krylov and Sorokin (2011); Song et al. (2013)	Van de Weijer et al. (2006a)
Abdel-Hakim and Farag (2006)	Penas and Shapiro (2009)
Vigo et al. (2010b)	
Jalilvand et al. (2011); Krylov et al. (2012)	**Tests Lacking Viewpoint Invariance**
Van de Weijer and Schmid (2006b)	Unnikrishnan and Hebert (2006)
Forssén (2007)	Van de Weijer et al. (2006a)
	Van de Weijer et al. (2005)

As can be seen, despite there being a substantial body of work in the literature, there has not been a comprehensive robust evaluation of colour features that provides sufficiently conclusive results. Prior to this study, there was not enough information on the performance of colour invariants to know which were suitable for colour detection, colour description, or for both. No study has solely evaluated and compared all the prominent colour gradient invariants together. As a result, colour has remained underused in modern feature extraction approaches, and determining the suitability of colour invariants for the task still remains unanswered in the literature. The work presented in this thesis addresses all the aforementioned issues of the literature.

The first three aforementioned issues are addressed by developing standard scale and geometric invariant local features, and using the testing framework of Mikolajczyk and Schmid (2005) with all the standard imaging distortions. Regarding Issues 4 and 6, a substantial amount of image data is used for the

evaluation and testing of features on 4 different image matching datasets and one large classification dataset. Additionally, using the shape-dominant PASCAL VOC dataset ensures this research is tested more rigorously and in a more general context than the approaches from issue no. 5. Finally, the last major difference between the literature, is that this study applies colour to both the detection and description phases of the feature extraction. In this way it is possible to know which invariants are more suitable for detection, description, or both.

Colour Invariant Local Image Features

3

This chapter outlines the colour invariant local features developed in this research. The features are detected with a colour adaptation of the Harris-Laplace and described with colour SIFT descriptors. Details of the HL detection algorithm are provided, along with results from the performed optimisation study. The colour space transformations employed for the colour feature extraction are explained here and visualised as 3D colour point clouds in order to perceive and compare their variations with changing illumination conditions. A brief account on the necessary photometric invariant background theory is given, before detailing how the colour invariants are directly obtained from the respective colour spaces.

3.1 Harris Corner Detection

The Harris-Laplace (Mikolajczyk and Schmid, 2001) has been one of the most widely used gradient-based detectors, and shown to be reliable under rotation, scale and illumination changes along with limited perspective deformations (Mikolajczyk and Schmid, 2004). It was an improvement upon the standard Harris corner detector, which is not repeatable when encountering images with scale changes as it only performs the detection using filters of one size. The scale-adapted Harris was introduced in order to address this issue, by convolving the original image with derivative kernels of varying sizes and thus locating differently sized corner-like structures. The scale-adapted Harris constitutes the first step in the Harris-Laplace algorithm. The Harris detector is based on the second moment matrix (also known as structure tensor or auto-correlation matrix), that is often used to describe local image gradient distributions. For an image, the scale-adapted structure tensor at position $\mathbf{x}$ is given by Equation 3.1:

$$H(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 G(\sigma_I) \otimes \begin{bmatrix} L_x^2(\sigma_D) & L_x L_y(\sigma_D) \\ L_x L_y(\sigma_D) & L_y^2(\sigma_D) \end{bmatrix} \tag{3.1}$$

The image gradients (L_x, L_y) are computed by convolution with the first derivatives of the Gaussian kernel with standard deviation σ_D (differentiation scale). These derivatives are then convolved with G(σ_I), the Gaussian kernel with standard deviation σ_I (integration scale). The eigenvalues of H at each image position measure the point's two principal signal changes in orthogonal directions. Corner-like structures and junctions will exhibit significant intensity variations in both directions, and in those cases both eigenvalues will be large. The Harris cornerness energy (Equation 3.2) is used to identify the points on the image that have corner-like characteristics. This measure ensures that points have greater cornerness energy if both eigenvalues are large. Following the original approach of Mikolajczyk and Schmid (2001), the factor k is set to 0.04 to achieve optimal results, and $3\sigma_D = \sigma_I$.

$$E(\mathbf{x}, \sigma_I, \sigma_D) \;=\; det\left(H(\mathbf{x}, \sigma_I, \sigma_D)\right) - k \cdot trace^2\left(H(\mathbf{x}, \sigma_I, \sigma_D)\right) \tag{3.2}$$

3.2 Characteristic Scale Selection

A local feature must have a scale associated with it so that a scale-invariant descriptor can correctly characterise the local image region. The scale essentially determines the neighbourhood size of the interest region around the spatial location of the interest point. Mikolajczyk and Schmid (2001) report that unlike the LoG, their multi-scale Harris responses (Equation 3.2) rarely attain maxima in 3D scale-space (2D spatial + 1D scale), which is required for the selection of a stable characteristic scale for the interest points. To achieve scale invariance, they proposed to estimate the characteristic scale based on Lindeberg's method (Lindeberg, 1998). This method uses the scale-normalised LoG response (Equation 3.3) to determine the stable scale for the local structures identified by the multi-scale Harris.

The LoG response is indicative of the similarity between the LoG kernel and the local image structure on which it is being convolved with. When the response results in a local 3D maxima across scales, then a characteristic scale

for that local structure exists at that location in scale-space. Examples of the profile of LoG responses across scale-space (at fixed image spatial locations) are shown in Figures 3.3 and 3.4 The scale-normalised Laplacian is obtained as follows:

$$|LoG(\mathbf{x}, \sigma_i)| \;=\; \sigma_i^2 |L_{xx}(\mathbf{x}, \sigma_i) + L_{yy}(\mathbf{x}, \sigma_i)| \tag{3.3}$$

where $L_{xx}(\mathbf{x}, \sigma_n)$ denotes the response at image location $\mathbf{x}$ of the convolution of the second derivative of the Gaussian (in the x-direction, with std. dev. σ_n) with the original input image. The response to this operator attains an extrema, when the size of the LoG kernel matches the size of the blob-like local image structure. It is more accurate to estimate the characteristic scale of blobs using the LoG operator as it has a certain affinity towards them due to its circular symmetry, but the LoG is also well suited to identify the scales of other local structures such as junctions, edges and corners.

3.3 Harris-Laplace Algorithm

An illustration of the developed Harris-Laplace detector algorithm is shown in Figure 3.1. To summarise the HL detector, Equations 3.1 and 3.2 are used to detect corners of various sizes and Equation 3.3 allows for a characteristic scale to be estimated for those corners. To achieve scale-invariance, two image stacks are constructed, a Harris energy stack and a LoG response scale-space stack. The LoG stack is obtained by convolving the input image multiple times with derivative kernels of increasing σ. Each scale provides an image of derivatives, that after applying Equation 3.3 becomes a LoG response image. A series of LoG images derived from all the scales used, then provides the 3D scale-space image stack. The Harris energy stack is similarly obtained, by applying Equation 3.1 followed by Equation 3.2 for successive varying scales. This complementary approach of the HL detector, provides the robust scale-invariance that is a characteristic of the LoG blob detector. Additionally due to the Harris corner measure, the HL can also detect more textured regions of higher variability and distinctiveness than many of the blobs detected by the LoG.

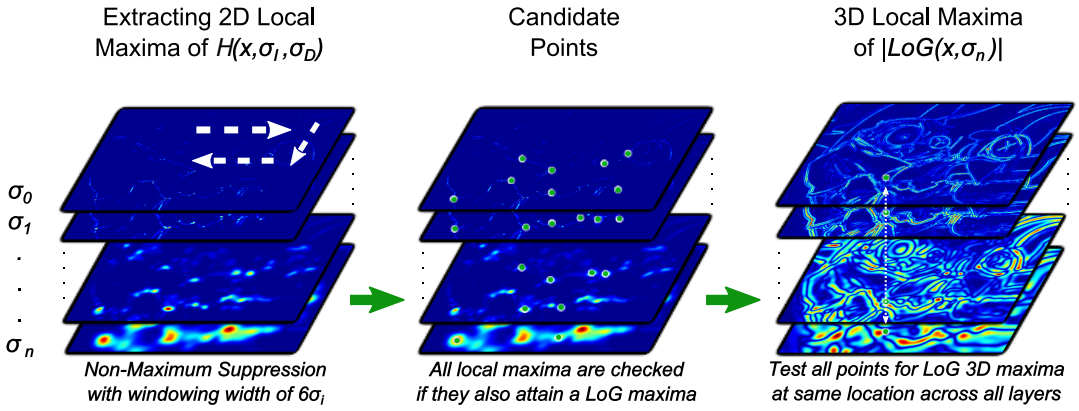


Figure 3.1: Diagram of the implemented Harris-Laplace algorithm.

In this research the number of scales being searched for is $n = 15$, it was chosen as it lies in between the range seen in the literature where 10 scales are used by Stöttinger et al. (2012) and 17 are used by Mikolajczyk and Schmid (2001). The integration scale σ_I, is varied at each layer of the scale-space stack with the following formula: $\sigma_i = t^i\sigma_o$, where $i = 0, 1, ... n$, t denotes the scale change factor and is set to $\sqrt{2}$ similarly to Mikolajczyk and Schmid (2001), σ_o is the initial scale and is set to 1. The derivative scale σ_D must be varied also at each layer of the scale-space stack with the relationship $0.333 * \sigma_i$. With known values for σ_I and σ_D for each scale level, Equation 3.1 is used to obtain a structure tensor for each location of the stack layers. Equation 3.2 will then generate an image of Harris corner energies for each scale in the stack. Interest points are detected in each scale by performing Non-Maximum Suppression (NMS) on the Harris energy stack in order to find 2D local maxima. For each layer in the stack, a local neighbourhood centred on each image pixel is compared. If that centre pixel has the maximum value within that neighbourhood, then it is deemed to be a local maxima. The size of this neighbourhood was varied in the optimisation study, described later in Section 3.4.

At this point the algorithm has identified all the local image structures at scales up to σ_n. The last part of the process involves searching for their characteristic scales. For each extracted corner (Harris energy local maxima), the LoG response value at its 2D image location across the $|LoG(\mathbf{x}, \sigma_n)|$ stack, is tested for a scale-space local maxima. The various methods for detecting the LoG maxima that were investigated, will be discussed in the optimisation study.

3.4 Algorithm Optimisation

In the original HL algorithm (Mikolajczyk and Schmid, 2001), local 3D maxima in the LoG stack are identified with the NMS method illustrated in Figure 3.2. When checking if a point $\mathbf{P}_{(x,y,\sigma_i)}$ is a local maxima, the value of that point is compared with the values of a 3×3 neighbourhood (W_i) centred at $\mathbf{P}_{(x,y,\sigma_i)}$, and encompassing the scales σ_{i-1}, σ_i and σ_{i+1}. To be considered a local maxima, the centre value must be higher than all the other neighbours and also be above a certain threshold T. Candidate $\mathbf{P}_{(x,y,\sigma_i)}$ points, are selected by performing a similar 2D NMS on the Harris stack for each scale.

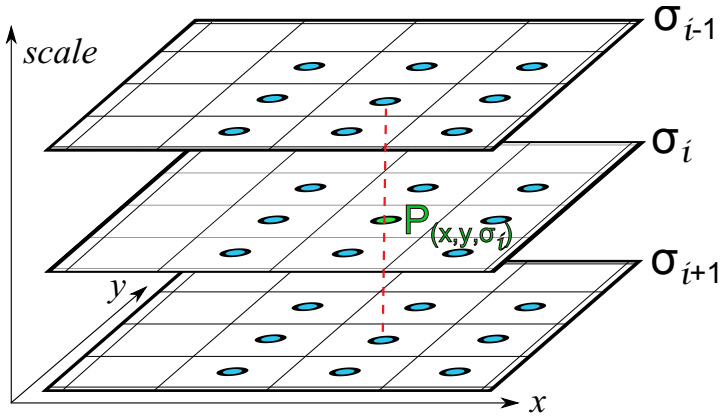


Figure 3.2: 3D local maxima Non-Maximum Suppression diagram of the original HL algorithm.

Since there are multiple variations of how to perform the NMS and the scale-space sampling, and various ways of obtaining spatial derivatives, it was necessary to perform an optimisation study to discover the most appropriate parameters for the specific algorithm that was implemented here. It is known that 99.73% of the data of a normal Gaussian distribution, lies within 3σ of the mean. The Gaussian derivative kernels employed here are designed to contain that 99.73%, therefore the actual width of the kernels in pixels is 6σ. In the optimisation experiments, the NMS was performed varying the W_i size from a 3×3 kernel to a $6\sigma_i \times 6\sigma_i$, the scale selection varied also with three different methods. The summary of the optimisation settings is shown in Table 3.1. There are 15 different HL settings, the parameter *LoG-NMS* refers to the half-window size of W_i for the NMS used in the LoG stack, *Harris-NMS* refers to the half-window size used in the stack of Harris energies, and the *LoG-Method* refers

Table 3.1: HL algorithm parameters for the optimisation study.

HL	(LoG-NMS, LoG-Method, Harr-NMS)	**HL**	(LoG-NMS, LoG-Method, Harr-NMS)
type1	2 , NeighPyr , 2	*type5*	2 , NeighPyr , 3
type2	3 , NeighPyr , 3	*type6*	2 , NeighPyr , 5
type3	5 , NeighPyr , 5	*type7*	2 , NeighPyr , σ-based
type4	σ-based , NeighPyr , σ based		
type8	3-Scales , orig. HL , 2	*type12*	NO-NMS , ALL-Scales , 2
type9	3-Scales , orig. HL , 3	*type13*	NO-NMS , ALL-Scales , 3
type10	3-Scales , orig. HL , 5	*type14*	NO-NMS , ALL-Scales , 5
type11	3-Scales , orig. HL , σ-based	*type15*	NO-NMS , ALL-Scales , σ-based

to the type of scale selection that was employed on the LoG stack. For the parameters in *LoG-NMS* and *Harris-NMS*, 2 equals a total size for W_i of 3×3, 3 equals 7×7, 5 equals 11×11, and σ-based means the size is $6\sigma_i\times6\sigma_i$ and varies according to the scale. For the setting 3-Scales of the *LoG-NMS*, the original Harris-Laplace method illustrated in Figure 3.2 is used both for scale selection and NMS.

The two other scale selection methods tested were *NeighPyr* which stands for Neighbourhood Pyramid, and *All-Scales*. In the method *NeighPyr*, for each Harris maxima location $\mathbf{P}_{(x,y,\sigma_i)}$, 2D LoG maxima are compared in a neighbourhood (W_i) centred at $\mathbf{P}_{(x,y,\sigma_i)}$ across all n scales. It was found that for W_i, a half-width of 2**i* worked best, the search area essentially forms a pyramid with the base at the biggest scale n of the LoG stack. For each scale, the biggest LoG response value of W_i is chosen (even if it's not the centre pixel) to represent the LoG response for scale i in the vector $\mathbf{R}(\mathbf{P}, i)$. This 1D vector represents the scale-space response profile for the point $\mathbf{P}_{(x,y,\sigma_i)}$ across all scales, examples of $\mathbf{R}(\mathbf{P}, i)$ plots are shown in Figures 3.3, 3.4 and 3.5. If $\mathbf{R}(\mathbf{P}, i)$ attains a local maxima at scale i, then the characteristic scale of the point is σ_i. If there is no local maxima then that point is rejected. In the case of the method *All-Scales*, the vector $\mathbf{R}(\mathbf{P}, i)$ is obtained by taking the LoG value at location $\mathbf{P}_{(x,y,\sigma_i)}$ for all scales, therefore no local 2D NMS is performed on the LoG stack.

As mentioned in Section 2.5, Mikolajczyk's evaluation framework is biased towards an approach that generates a dense clustering of points. To counteract this drawback, the penultimate step of this HL detector algorithm is to prune the final set of points by merging clustered points which overlap in area by

more than 90% (keeping the point with the highest Harris energy). Finally, the top N points with greatest Harris energies are chosen to be the final set of Harris-Laplace points for the image.

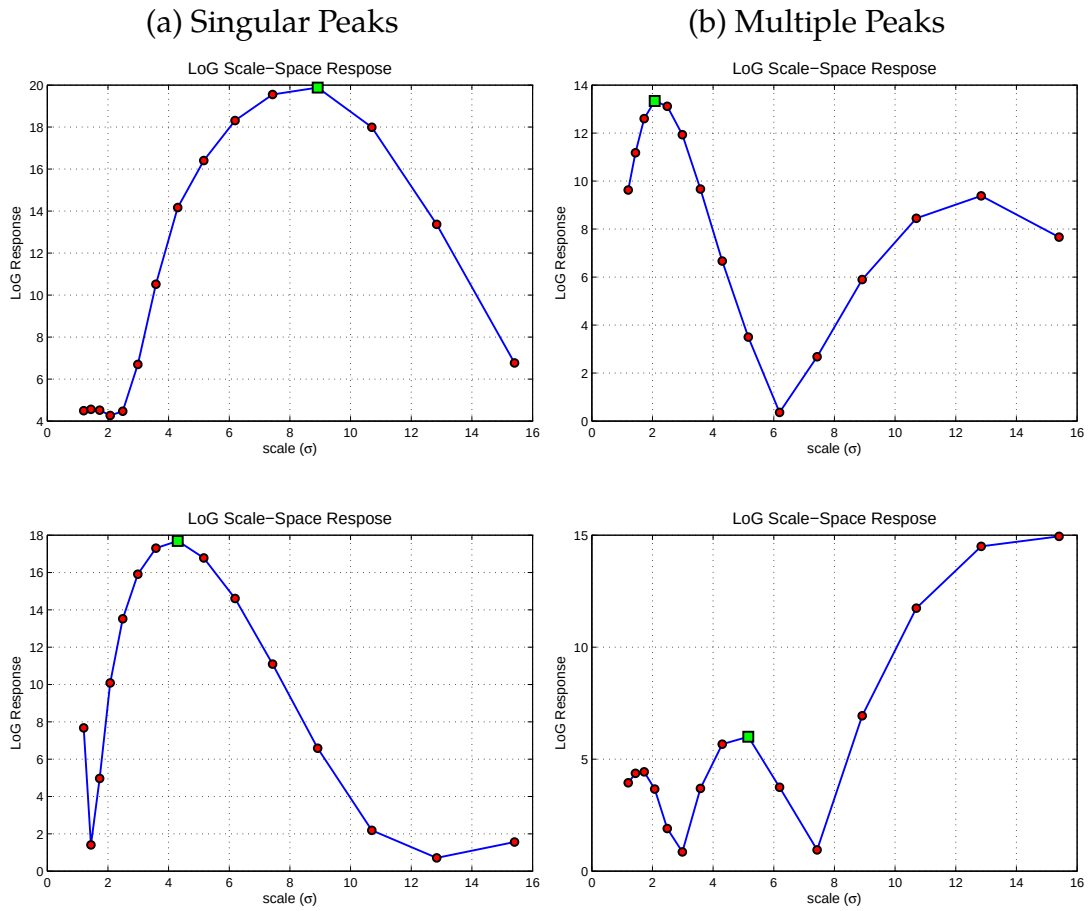


Figure 3.3: Log scale-space response plots with local maxima. Column (a) shows correct scale selection for profiles with only one maxima. Column (b) shows scale selection of the peak with the highest LoG response. The green square denotes the location of the estimated characteristic scale.

Figure 3.3 shows the LoG scale-space local maxima detection results (highlighted by the green square). The process of finding the peaks of the scale-space profile $\mathbf{R}(\mathbf{P}, i)$ in this research, uses first order difference information to identify the trend of the profile. The trend between two data points is given by the sign of their backwards difference, and a peak occurs when the trend of the profile changes from a positive to a downward one. In the case of flat segments that have a trend sign of 0, the trend of the profile is back-propagated and replaces the flat segment in order to have a trend composed only of +1's or -1's. The implementation of the peak detection algorithm used in this work, is Matlab's

findpeaks function. The parameter *minpeakdistance* of the Matlab function (which sets a limit to the proximity between two peaks) is set to 3, and this value was chosen from visual inspection of all the output peaks from one image. Figure 3.3a shows examples of when $\mathbf{R}(\mathbf{P}, i)$ contains only one local maxima, in the cases where there are more than one as shown in Figure 3.3b, the peak with the highest LoG response value is chosen. Those examples demonstrate the types of cases in which the implemented scale selection method performs successfully, however not every profile contains a local maxima, and the function *findpeaks* can at times output erroneous results. False positives are obtained in cases such as in Figure 3.4a, where a green square signifies the location of the incorrectly detected local maxima. The method can however, correctly identify scenarios where there are no local maxima such as the examples in Figure 3.4b.

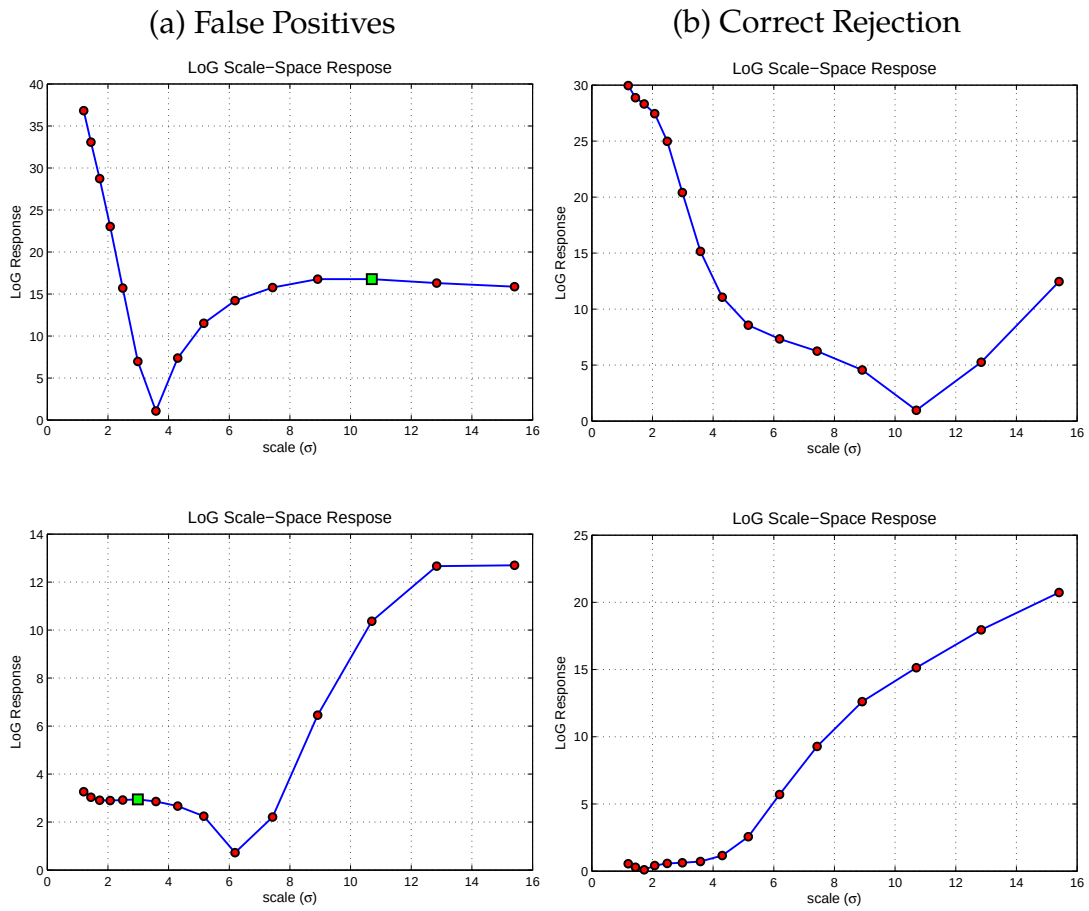


Figure 3.4: Log scale-space response plots without any local maxima. Column (a) shows cases where false positives are obtained from the scale selection. Column (b) shows where no peaks were identified, and thus the point is correctly rejected.

Apart from the disadvantages of not being able to consistently correctly detect or reject the local maxima of $\mathbf{R}(\mathbf{P}, i)$, the other unoptimised aspect of the implemented scale detection technique, is the varying levels of accuracy that can be achieved in estimating the characteristic scale of a point. The level of accuracy depends on the sampling density of the scale-space around the location of the peak. In Figure 3.5a for example, it can be seen that the sampling is dense and there is distinct visible location for the local maxima which the method accurately estimates. In Figure 3.5b, the peaks occur in the larger scales which are sampled with a higher separation, therefore the data point of a more accurate estimation of the peak does not exist. The scale sampling is performed according to the optimal guidelines of Mikolajczyk and Schmid (2001), and thus were not changed in this research. A more accurate scale could be achieved by simply increasing the scale sampling, but that would incur bigger computational costs.

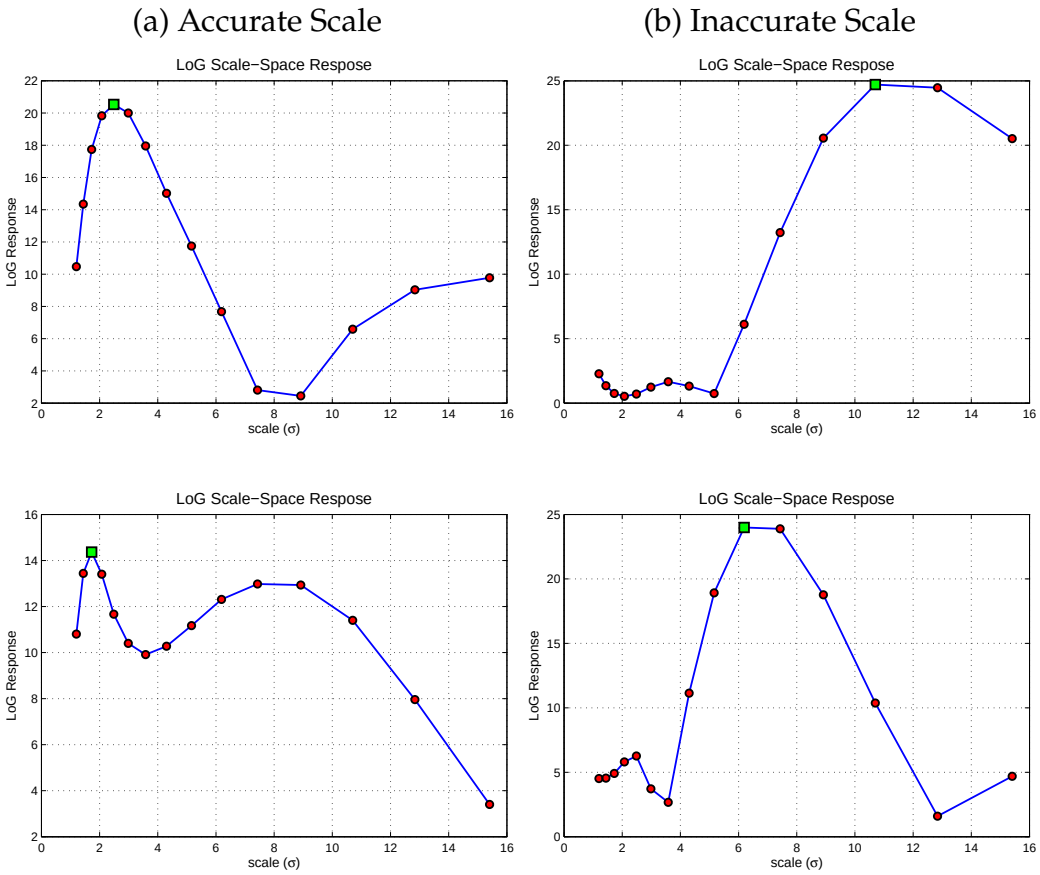


Figure 3.5: Accuracy of the scale estimation. Column (a) shows accurate scale estimations due to a dense sampling of the scale-space. Column (b) shows profiles in which the estimated peak occurs in a sparser sampled region, and the scale is thus less accurate.

This concludes the detailed explanation of the implemented HL algorithm and its various parameter settings. The last part of this section summarises the performance of the 15 different detector designs from Table 3.1, and the results are presented in Figures 3.6,3.7, 3.8 and 3.9.

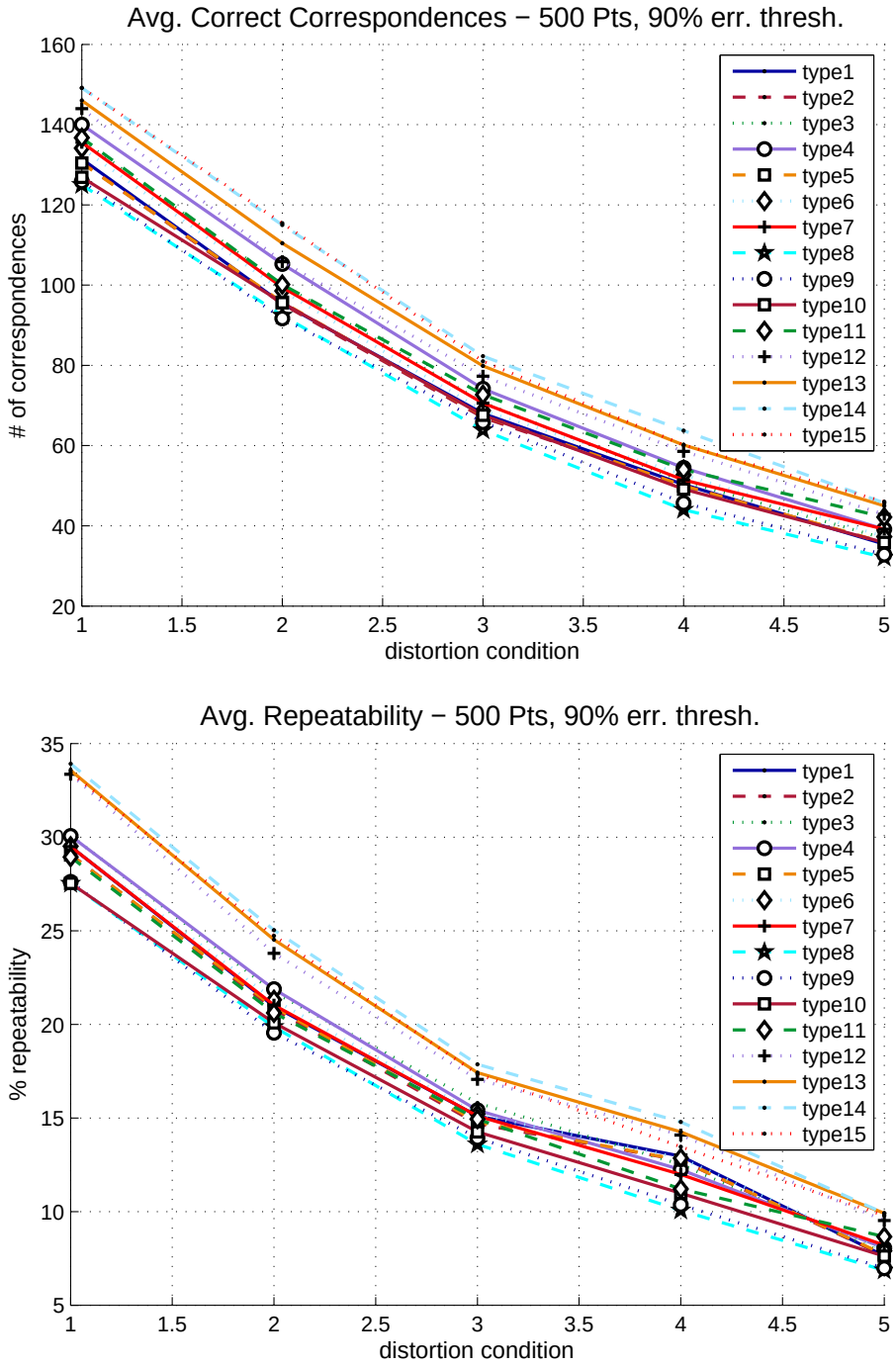


Figure 3.6: Summary of the optimisation study on the Oxford dataset with 90% error threshold.

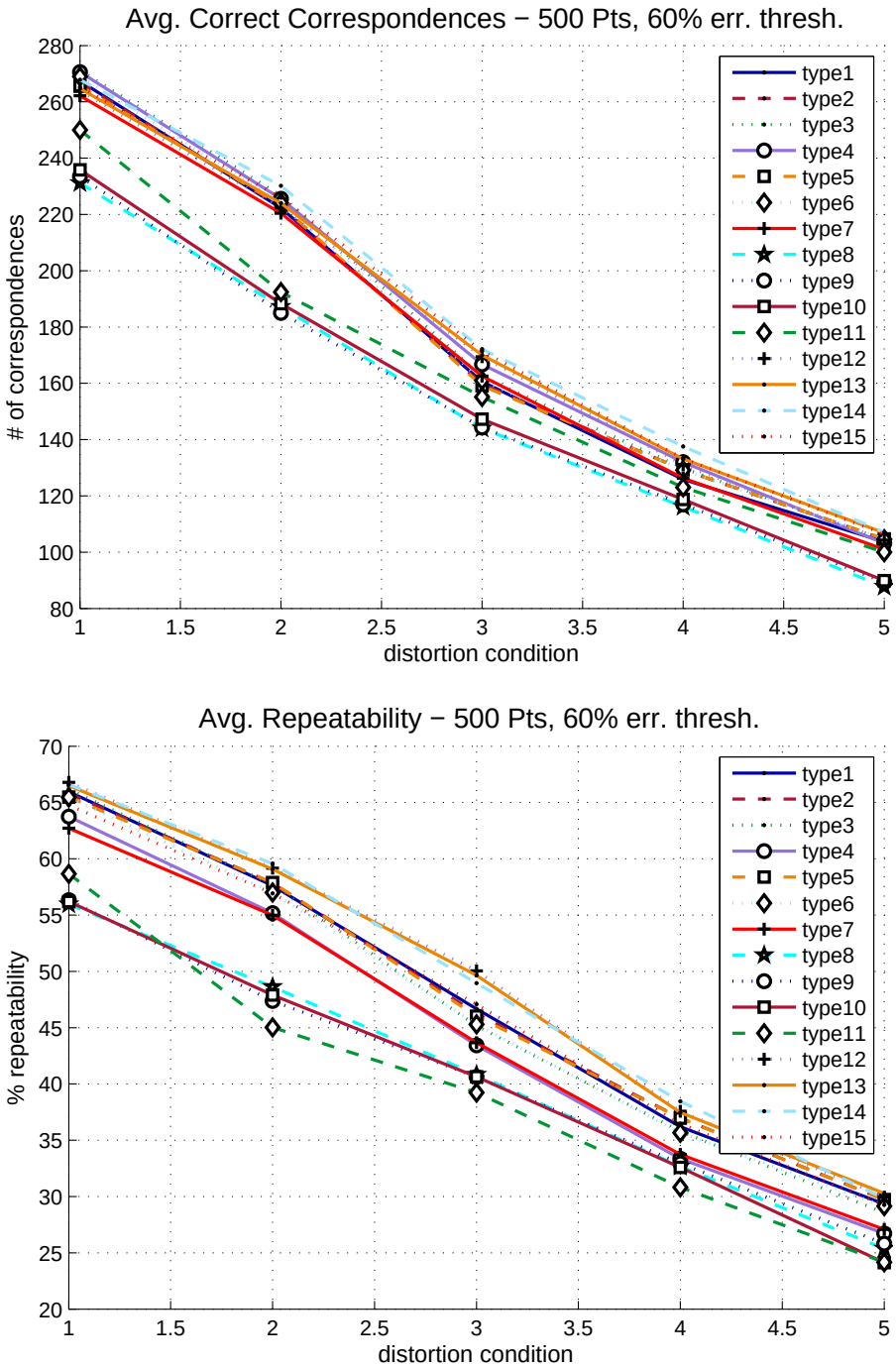


Figure 3.7: Summary of the optimisation study on the Oxford dataset with 60% error threshold.

To evaluate the detectors, a point correspondence experiment was carried out on all the imagesets of the Oxford and Middlebury dataset (Section 2.5.3). Only the grayscale information was utilised to extract 500 points from each image from the Oxford dataset and 300 points from the Middlebury. The test compared the number of correct point correspondences and the reliability of the detection. Figure 3.6 presents the results of the Oxford optimisation study

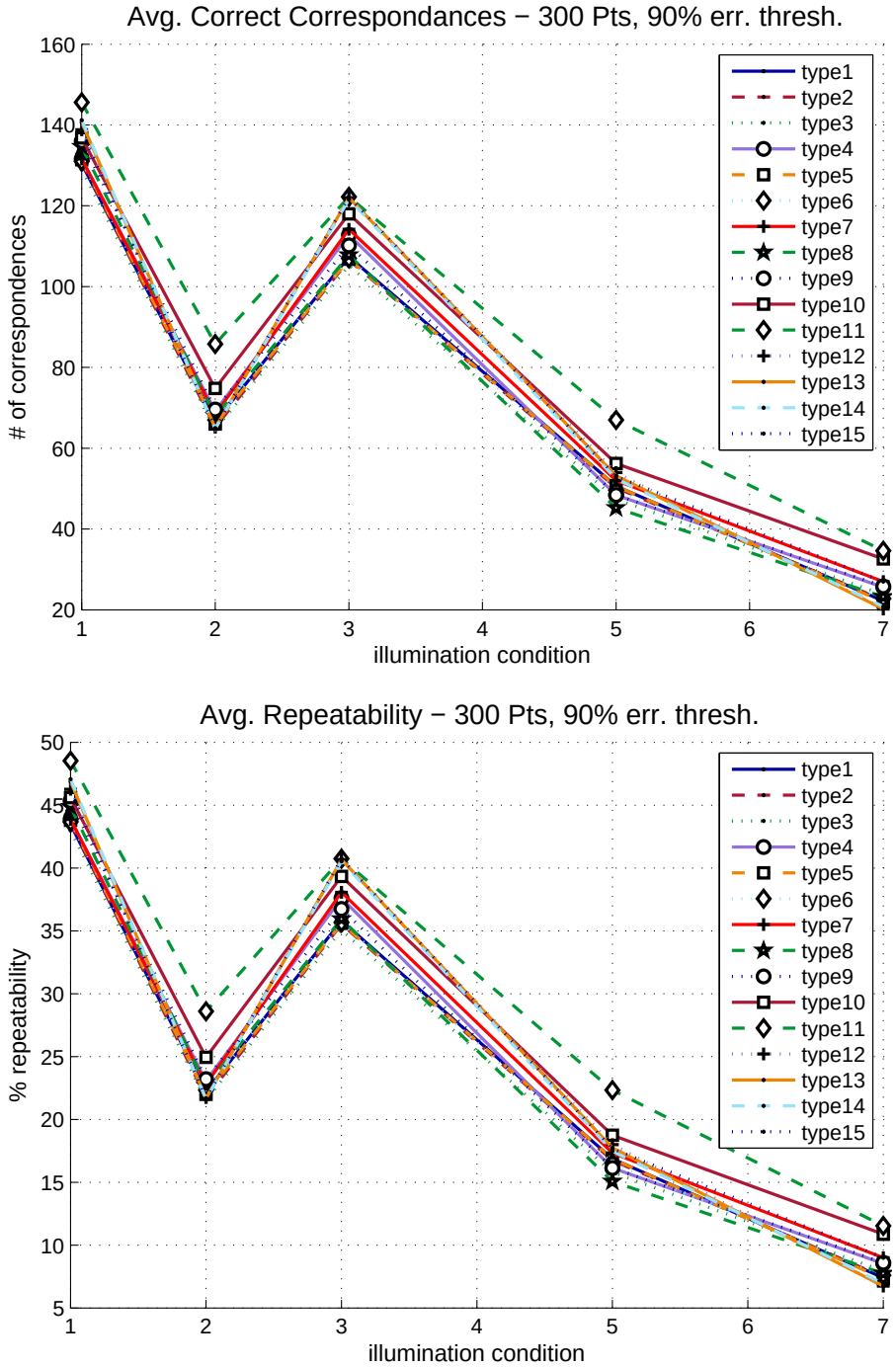


Figure 3.8: Summary of the optimisation study on the Middlebury dataset with 90% error threshold.

with a strict matching error threshold, for two corresponding points in separate images to be regarded as correctly matched, their overlapping areas must be more than 90%. Results with a laxer overlap threshold of 60% as has been the practice in most previous studies, are presented in Figure 3.7, and the performance nearly doubles in this case. The Middlebury results for both thresholds are shown in Figures 3.8 and 3.9. There are clear distinctions in performance

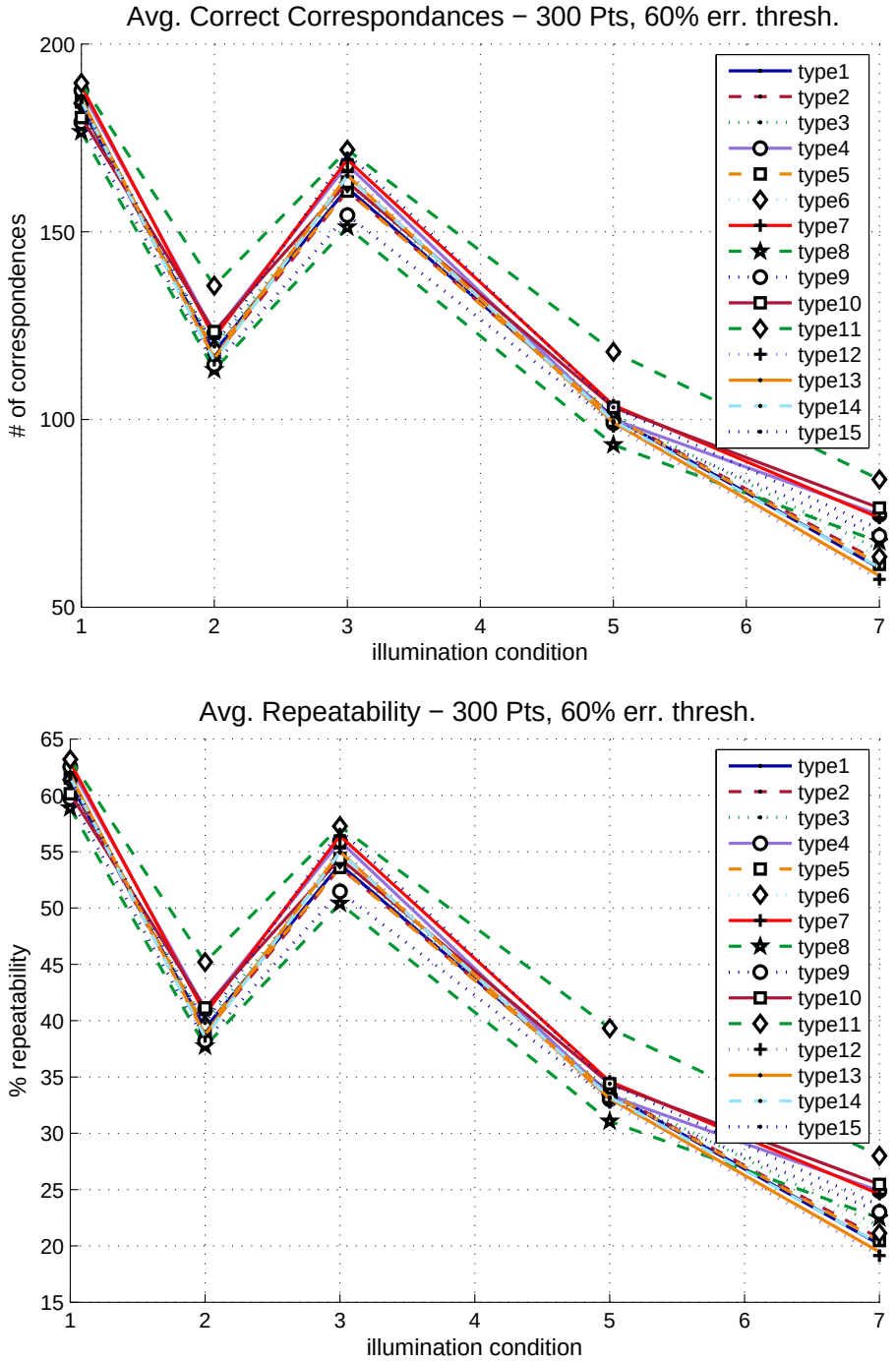


Figure 3.9: Summary of the optimisation study on the Middlebury dataset with 60% error threshold.

between the various detector types, and how their relative performance changes with a different threshold and with a different dataset. This is not surprising as it can be deduced that the various HL parameters would favour the detection of different types of corners, and the optimal HL algorithm is thus dataset dependant. There are some observations that can be made however, in order to select the parameters that will be used in this research. The original HL

algorithm (Mikolajczyk and Schmid, 2001) denoted as $type8$ in this experiment, is in fact amongst the worst performers and will thus not be considered. The best choice must perform well for both datasets and for both thresholds. There is not a significant difference among the top performers however, and many of them can be considered for this research. The best detector in the Middlebury dataset ($type11$), does not perform well in the Oxford set. The ones that perform well under all the results are $type14$ and $type15$, due to a balance of high repeatability and number of correspondences in both threshold settings. However, the detector chosen for the final HL algorithm used throughout this thesis is *type15*, as it varies the Harris-NMS windowing size with the scale σ (which makes more sense intuitively). To summarise, the most important result that arises from the optimisation study is that the original method proposed by Mikolajczyk and Schmid (2001) is clearly inferior. Additionally, although there is a minor difference amongst the top performing parameters, the actual optimised HL algorithm can only be dataset-specific.

3.5 Colour Photometric Invariants

Much of the relevant early work on colour was dedicated to global colour features for object recognition and image retrieval, Swain and Ballard (1991) for example used colour histograms for image description. Their method is not invariant to lighting geometry however, and they recommended the use of normalised RGB histograms to obtain the required invariance. Normalised histograms remain variant however to changes in the illuminant. A subsequent technique addressing that issue is the illuminant invariant indexing method of Funt and Finlayson (1995), which assumes a Lambertian reflectance model but still does not provide invariance to lighting geometry. Finlayson et al. (1998) combined aspects of the methods proposed by Swain and Ballard (1991) and Funt and Finlayson (1995), to propose an image indexing method invariant to both illuminant changes and shading. Full lighting geometry invariance was still not achieved as specularities needed also to be taken into consideration. Specular invariance was introduced by Gevers and Smeulders (1999) along with invariance to illuminant changes. That work on invariance was then extended by physical-based approaches that utilise the derivative structure of images

(Geusebroek et al., 2001, Van de Weijer et al., 2005).

Apart from being able to achieve invariance to light geometry and illuminant variations, using image derivatives allows photometric colour invariants to be used for edge and corner detection. It is these types of colour invariants which are of interest to this research, as the goal is to extract local salient features from an image, which in general require the manipulation of image derivatives. This thesis employs the colour photometric invariants proposed by: Van de Weijer et al. (2005, 2006a), Stöttinger et al. (2012), Geusebroek et al. (2001). They are derived from two illumination reflection models that are more sophisticated than the Lambertian model: the dichromatic reflection model (Shafer, 1985); and the Kubelka-Munk reflection theory of coloured bodies (Wyszecki and Stiles, 1982). The background theory of the two models and their derived photometric invariants will be detailed in the next section, while the colour space transformations needed to achieve these invariants are discussed in Section 3.5.3, and the actual implementation of the local colour features is described in Section 3.6.

3.5.1 The Dichromatic Reflection Model

Unlike the Lambertian reflectance model, which assumes that a surface reflects light with isotropic intensity in all directions, the dichromatic model accounts for optically inhomogeneous materials and splits the reflection of light from an object into a specular (interface) component and a diffuse (body) component (Shafer, 1985). At the air-surface boundary, some of the light is immediately reflected out as a specular component, the rest of the light is refracted into the material and partially absorbed and diffusely reflected out as the body colour component. The angles of incident and outgoing rays depend on the spatial refractive indexes of the carrier medium and the object material, these indexes are wavelength dependant. A useful assumption that can be made is having neutral interface reflectance, which treats all refractive indexes to be constant with respect to the light wavelength and thus simplifies the geometric modelling of the reflected light. With a second assumption of having an ideal white light source $\mathbf{c}^i$ with a smooth spectrum of equal energy at all wavelengths,

the RGB vector $\mathbf{f}= (R, G, B)^T$ at a particular image location, can be modelled as a weighted sum of two vectors:

$$\mathbf{f} = e(m^b\mathbf{c}^b + m^i\mathbf{c}^i) \tag{3.4}$$

$\mathbf{c}^b$ represents the colour of the body (i.e. the diffuse reflectance), and $\mathbf{c}^i$ the colour of the specular surface reflectance. Their scalar magnitudes are denoted by m^b and m^i, and e is the intensity of the light source. Lambertian reflection would apply for matte surfaces for which there would be no interface (specular) reflection (i.e. $m^i = 0$), and the model would thus simplify to:

$$\mathbf{f} = em^b\mathbf{c}^b \tag{3.5}$$

To obtain first-order photometric derivative information of an image, the spatial derivatives of Equation 3.4 must be computed. The spatial gradients of the RGB image vectors are then represented by:

$$\mathbf{f}_x = em^b\mathbf{c}^b_x + (e_x m^b + em^b_x)\mathbf{c}^b + (em^i_x + e_x m^i)\mathbf{c}^i \tag{3.6}$$

where, the subscript x indicates spatial differentiation, and as a known illuminant is assumed with neutral interface reflection, $\mathbf{c}^i$ is independent of x. The vector $\mathbf{f}_x$ represents the gradients in an image which are essentially made up of three causes: a body reflectance change in the direction $\mathbf{c}^b_x$, a shadow-shading change in the direction $\mathbf{c}^b$, and lastly a specular change in the direction $\mathbf{c}^i$. In the case of the shadow-shading component, $e_x m^b$ refers to changes in intensity which lead to a shadow edge, and em^b_x denotes a change in the geometry coefficient that causes a shading edge. Van de Weijer et al. (2005) estimate the direction of $\mathbf{c}^b$, by analysing the Lambertian reflection case of matte surfaces represented by Equation 3.5, which contains no specular components. They deduce that the shadow-shading component $\mathbf{c}^b$ has a direction parallel to $\mathbf{f}$ and coincides with $\hat{\mathbf{f}}$, where $\hat{}$ denotes a unit vector:

$$\hat{\mathbf{f}} = \frac{1}{\sqrt{R^2 + G^2 + B^2}} (R, G, B)^T \tag{3.7}$$

The specular direction $\mathbf{c}^i$ is where changes in the specular geometry coefficient occur, and is composed of em^i_x which represents changes in the angles

between the camera viewpoint, object and the light source; and a second component $e_x m^i$ which represents a shadow edge on top of a specular reflection. The unit vector $\hat{\mathbf{c}}^i$ is approximated by Van de Weijer et al. (2005) with a white light source:

$$\hat{\mathbf{c}}^i = \frac{1}{\sqrt{3}} (1,1,1)^T \tag{3.8}$$

The third relevant directional component from Equation 3.6 that influences the appearance of edges in an image, is perpendicular to $\mathbf{c}^b$ and $\hat{\mathbf{c}}^i$. Van de Weijer et al. (2005) name it the hue direction $\hat{\mathbf{b}}$ which is related to changes in body reflectance, this direction is used to form the shadow-shading-specular quasi-invariant, instead of the body reflectance direction $\hat{\mathbf{c}}_x^b$ due to its simpler calculation:

$$\hat{\mathbf{b}} = \frac{\hat{\mathbf{f}} \times \hat{\mathbf{c}}^i}{\left| \mathbf{f} \times \mathbf{c}^i \right|} \tag{3.9}$$

These three edge inducing directions $\hat{\mathbf{b}}$, $\mathbf{c}^b$ and $\hat{\mathbf{c}}^i$, are the basis of the photometric variant and invariant spatial derivatives proposed by Van de Weijer et al. (2005). Variants are obtained by projecting the image gradients $\mathbf{f}_x = (R_x, G_x, B_x)$ on the aforementioned directions. The shadow-shading variant is obtained by projecting $\mathbf{f}_x$ on the $\mathbf{c}^b$ direction, the specular variant by the projection with $\mathbf{c}^i$, and thirdly the shadow-shading-specular invariant is obtained by a projection on the hue direction $\hat{\mathbf{b}}$. The variants have invariant counterparts, which are estimated by Van de Weijer et al. (2005) as quasi-invariants, and by Van de Weijer et al. (2006a) as full-invariants. The shadow-shading variant $\mathbf{S}_x^V$ and invariant $\mathbf{S}_x^Q$ are calculated from Equation 3.10, where in the first expression the dot denotes the vector inner product, and the outer $\hat{\mathbf{f}}$ specifies the direction of the variant. The quasi-invariant (indicated by subscript Q) is obtained by subtracting the variant from the overall image derivative expression:

$$\mathbf{S}_x^V = \left(\mathbf{f}_x \cdot \hat{\mathbf{f}} \right) \hat{\mathbf{f}} \quad , \quad \mathbf{S}_x^Q = \mathbf{f}_x - \mathbf{S}_x^V \tag{3.10}$$

The invariant $\mathbf{S}_x^Q$ is the component of the image derivatives not caused by shadow-shading edges, it is mainly comprised of hue and specular edges. A similar approach is taken to obtain the variants and quasi-invariants of the other two image gradient-causing directions. Equation 3.11 shows the expression for

the specular variant and the quasi-invariant which is not affected by highlight edges.

$$\mathbf{O}_x^V = \left(\mathbf{f}_x \cdot \hat{\mathbf{c}}^i\right) \hat{\mathbf{c}}^i \quad , \quad \mathbf{O}_x^Q = \mathbf{f}_x - \mathbf{O}_x^V \tag{3.11}$$

The third expression is the specular-shadow-shading quasi-invariant $\mathbf{H}_x^Q$ shown in Equation 3.12, which is obtained by projecting $\mathbf{f}_x$ on the hue direction $\hat{\mathbf{b}}$. This quasi-invariant represents the true colour of a body, and is not affected by specular or shadow-shading edges. The calculation of the variant $\mathbf{H}_x^V$ does not follow the same procedure as $\mathbf{S}_x^V$ and $\mathbf{O}_x^V$ (i.e. simply projecting on $\hat{\mathbf{b}}$), since $\hat{\mathbf{b}}$ is in the direction of the invariant and not the variant. The variant $\mathbf{H}_x^V$ is obtained by subtracting the invariant derivatives from the image derivatives:

$$\mathbf{H}_x^Q = \left(\mathbf{f}_x \cdot \hat{\mathbf{b}}\right) \hat{\mathbf{b}} \quad , \quad \mathbf{H}_x^V = \mathbf{f}_x - \mathbf{H}_x^Q \tag{3.12}$$

Before ending this section on the invariants derived from the dichromatic reflection model, two other invariants proposed by Van de Weijer et al. (2006a) will be discussed. These are the shadow-shading full-invariant $\mathbf{s}_x$ and the shadow-shading-specular full-invariant $\mathbf{h}_x$, expressed in Equation 3.13. The full shadow-shading invariant is obtained by normalising the quasi-invariant $\mathbf{S}_x^Q$ by the image luminance intensity magnitude $|\mathbf{f}|$. Lastly, the full shadow-shading-specular invariant is obtained by by dividing the quasi-invariant $\mathbf{H}_x^Q$ by the saturation (s). This last invariant however, will not be used in this research as it is the hue derivative, which is unstable when the saturation is low and can produce very high gradients in locations where there may not be any.

$$\mathbf{s}_x = \frac{\mathbf{S}_x^Q}{|\mathbf{f}|} \quad , \quad \mathbf{h}_x = \frac{\mathbf{H}_x^Q}{s} \tag{3.13}$$

3.5.2 Kubelka-Munk Colour Model

The Kubelka-Munk reflection model of coloured bodies (Wyszecki and Stiles, 1982) is similar to the dichromatic model. It assumes isotropic scattering of the incident light on a material, and characterises the material by a wavelength scatter coefficient and a wavelength absorption coefficient. The reflected spectrum of light $E(\lambda, x)$, viewed by a camera is modelled as:

$$E(\lambda, x) = e(\lambda, x)\left(1 - \rho_f(x)\right)^2 R_\infty(\lambda, x) + e(\lambda, x)\,\rho_f(x) \tag{3.14}$$

where x denotes the image pixel position, λ is the wavelength of the light, $e(\lambda, x)$ refers to the illumination spectrum, ρ_f is the Fresnel reflectance, and $R_\infty(\lambda, x)$ is the material reflectivity. Geusebroek et al. (2001) utilise this model to propose four sets of photometric colour invariants. The first is the H invariant, designed for objects under imaging conditions of equal energy but uneven illumination. The second invariant set C, are valid for illuminations of equal energy but uneven illumination and assuming the objects have matte or dull surfaces. The third set W, are invariant for planar matte objects under equal energy and uniform illumination. The last invariant set N will not be considered in this research, due to being invariant to varying coloured illumination.

H Invariant:

In the case of the H invariant, assuming an equal energy illuminant across the imaged scene makes the spectral components of the light source $e(\lambda, x)$ to be constant with respect to the wavelength and only variable over the spatial location x. By differentiating Equation 3.14 once with respect to λ obtains the expression for the first spectral derivative E_λ shown in Equation 3.15, differentiating twice produces Equation 3.16. Under the aforementioned assumption only the term R_∞ is a variable of λ. All subsequent subscripts of λ and x will refer to a differentiation, spectrally in λ or spatially in x.

$$E_\lambda = e\left(1 - \rho_f\right)^2 \left(\frac{\delta R_\infty}{\delta \lambda}\right) \tag{3.15}$$

$$E_{\lambda\lambda} = e\left(1 - \rho_f\right)^2 \left(\frac{\delta^2 R_\infty}{\delta \lambda^2}\right) \tag{3.16}$$

Dividing Equation 3.15 by Equation 3.16 indicates that the resulting expression is a spectral derivative function of the surface reflectance term R_∞ only, as the terms $e\left(1 - \rho_f\right)^2$ are cancelled out. This reflectance property is denominated as the H invariant, which is related to the hue of the material. Summarising this invariant, for a surface with neutral interface reflection and under ideal white illumination, the ratio of the first and second spectral derivatives of the

incident light produce the H invariant. This invariant is independent of the light intensity, incident angle, view direction and of specular highlights.

$$H = \frac{E_\lambda}{E_{\lambda\lambda}} \tag{3.17}$$

In order for this spectral (chromatic) invariant to be useful for extracting local image features, it must be related to the spatial domain in order to obtain image derivatives. Differentiating Equation 3.17 up to the first spatial derivative and up to the second spectral order (Geusebroek et al., 2001), results in H_x which is the spectral-spatial H-invariant utilised in this thesis:

$$H_x = \frac{E_{\lambda\lambda}E_{\lambda x} - E_\lambda E_{\lambda\lambda x}}{E_\lambda^2 + E_{\lambda\lambda}^2} \tag{3.18}$$

C Invariant:

This invariant adds a further assumption to that of H that all surfaces are matte. For a matte surface there is no specular component thus $\rho_f(x) \approx 0$, which means $E(\lambda, x)$ can be modelled as: $E(\lambda, x) = e(x)R_\infty(\lambda, x)$. Differentiating this expression spectrally results in:

$$E_\lambda = e\left(\frac{\delta R_\infty}{\delta\lambda}\right) \tag{3.19}$$

The ratio of E_λ and E results in the C invariant, which depends only on a spectral derivative function of the surface reflectance R_∞ :

$$C = \frac{E_\lambda}{E} = \frac{e\left(\frac{\delta R_\infty}{\delta\lambda}\right)}{e(R_\infty)} = \left(\frac{1}{R_\infty}\right)\frac{\delta R_\infty}{\delta\lambda} \tag{3.20}$$

This means the C invariant depends on the camera view direction, and also neither on the intensity or the direction of the illuminant. Two separate spectral-spatial invariants can be derived from C, which are needed for the local feature extraction. Differentiating in the first spectral order and first spatial order obtains $C_{\lambda x}$, and with respect to the second spectral order results in $C_{\lambda\lambda x}$:

$$C_{\lambda x} = \frac{E_{\lambda x}E - E_\lambda E_x}{E^2} \quad , \quad C_{\lambda\lambda x} = \frac{E_{\lambda\lambda x}E - E_{\lambda\lambda}E_x}{E^2} \tag{3.21}$$

W Invariant:

The W invariant is calculated differently from H and C, in that it is built from the spatial derivative x instead of the spectral derivative λ. For this invariant, an ideal illuminant is assumed which is spatially uniform, along with a Lambertian surface reflectance (i.e. $\rho_f(x) \approx 0$). This illumination means that $e(\lambda, x)$ is treated as a constant e, and the expression for the incident light on the camera becomes: $E(\lambda, x) = eR_\infty(\lambda, x)$. Differentiating this expression spatially in x results in:

$$E_x = e\left(\frac{\delta R_\infty}{\delta x}\right) \tag{3.22}$$

The W invariant is then obtained by a ratio of E_x and E, which depends only on a spatial derivative function of the surface reflectance R_∞. Since it is a spatial derivative the invariant will be denoted with a subscript x as:

$$W_x = \frac{E_x}{E} = \frac{e\left(\frac{\delta R_\infty}{\delta x}\right)}{e(R_\infty)} = \left(\frac{1}{R_\infty}\right)\frac{\delta R_\infty}{\delta x} \tag{3.23}$$

This invariant expression describes an object's reflectance, independent of the intensity level of the illuminant. W_x is already a spatial invariant that can be used to extract local features, but two other invariants can be extracted by differentiating up to the first and second spectral orders:

$$W_{\lambda x} = \frac{E_{\lambda x}}{E} \quad , \quad W_{\lambda\lambda x} = \frac{E_{\lambda\lambda x}}{E} \tag{3.24}$$

In summary, this thesis uses the aforementioned three photometric invariants proposed by Geusebroek et al. (2001). H is shadow, shading and highlight invariant. C is an invariant to shadow and shading, and W is illumination intensity invariant. The highest level of invariance is associated with H, which in turn means less discriminative power. The feature matching and recognition experiments in this thesis will discover what level of invariance performs better. All the required theoretical background on the photometric invariants used in this thesis has now been covered, the next section describes the relevant colour spaces that are involved and how they relate to an estimation of the theoretical invariant expressions given in this section.

3.5.3 Colour Spaces

The previously discussed photometric components of the incident light seen by a camera, like shading, shadows and specularities, are all correlated in the common RGB colour model. Numerous colour transformations have been proposed in the literature to represent colour information in other models that have different characteristics and benefits. In this way certain photometric properties can be partially separated and distinguished. A colour space type that will not be considered in this research, referred to as a "perceptually uniform colour space" in which perceptual distances between two colours correspond to a Euclidean distance, has arguably been one of the most successful types of colour spaces in the field for certain applications. The most widely used colour spaces of this type are the CIE L^*a^*b and CIE L^*u^*v (Chong et al., 2008), they however have been extensively used primarily in segmentation applications and image retrieval techniques that employ zero-order histogram-based colour descriptors; and not for obtaining colour invariant gradients. No such works are reported in the comprehensive colour invariant review of Muselet and Funt (2013), and only one relevant colour study (Cui et al., 2010), has been found by the author that utilises a perceptually uniform colour space to obtain SIFT-like descriptors. However, as stated in Section 2.4, that work is not considered here due to insufficient detail in the publication about the descriptor's implementation.

The original works that propose the invariants used in this thesis derive their colour gradients from four colour spaces, which this research will thus also utilise: The Spherical Colour Space (SCS) (Van de Weijer et al., 2005), which is obtained from the RGB space with Equation 3.26. The Opponent Colour Space (OCS) shown in Equation 3.28 (Van de Weijer et al., 2005). The Hue Saturation and Intensity (HSI) of Equation 3.30 (Van de Weijer et al., 2005). The fourth colour representation is the Gaussian Colour Model (GCM) (Geusebroek et al., 2001), shown in Equation 3.31.

Spherical Colour Space:
This space is obtained by an orthogonal transformation from the RGB space. There are two colour channels in the SCS, θ and φ, and the magnitude of the light intensity is the third channel r which points in the shadow-shading direction. The derivative r_x is therefore the shadow-shading variant $\mathbf{S}_x^V$, and its quasi-

invariant counterpart $\mathbf{S}_x^Q$ exists in a perpendicular plane to r (Van de Weijer et al., 2005). This $\theta\varphi$-plane is calculated via:

$$\left|\mathbf{S}_x^Q\right| = r\sqrt{(\varphi_x)^2 + (\sin(\varphi)\theta_x)^2} \tag{3.25}$$

$$SCS = \begin{pmatrix} r \\ \theta \\ \varphi \end{pmatrix} = \begin{pmatrix} \sqrt{R^2+G^2+B^2} \\ \arctan(\frac{R}{G}) \\ \arcsin(\frac{\sqrt{R^2+G^2}}{\sqrt{R^2+G^2+B^2}}) \end{pmatrix} \tag{3.26}$$

Opponent Colour Space:

The opponent space results from an orthonormal transformation from RGB, which results in decoupling the specular information of the light. Channels $o1$ and $o2$ contain the chromatic information split into opponent red-green and blue-yellow components, $o3$ carries the achromatic luminance intensity information. The specular variant $\mathbf{O}_x^V$ is in the direction of $o3$ and the invariant $\mathbf{O}_x^Q$ is formed from a combination of $o1$ and $o2$.

$$\left|\mathbf{O}_x^Q\right| = \sqrt{o1_x^2 + o2_x^2} \quad , \quad \left|\mathbf{O}_x^V\right| = o3_x \tag{3.27}$$

$$OCS = \begin{pmatrix} o1 \\ o2 \\ o3 \end{pmatrix} = \begin{pmatrix} (R-G)/\sqrt{2} \\ (R+G-2B)/\sqrt{6} \\ (R+G+B)/\sqrt{3} \end{pmatrix} \tag{3.28}$$

HSI Colour Space:

The HSI colour space is formed from a polar transformation on the first two axes of the OCS. The intensity i is the same luminance channel as $o3$. h stands for hue, which characterises the colour of the light, s is the saturation and signifies the strength of the hue colour. With increasing s values, the colour becomes whiter, whereas with lower saturations the light colour appears more grey until it becomes black at zero saturation. In Section 3.5.1, the direction $\hat{\mathbf{b}}$ of the specular-shadow-shading quasi-invariant $\mathbf{H}_x^Q$, was described as being perpendicular to the shadow-shading direction $\mathbf{c}^b$ and the specular direction $\hat{\mathbf{c}}^i$.

This constraint is satisfied by the HSI colour transformation (Van de Weijer et al., 2005), and is why $\hat{\mathbf{b}}$ is named the hue direction. Therefore the hue derivative h_x represents the invariant $\mathbf{H}_x^Q$ as shown in Equation 3.29. The multiplication with the saturation, is needed as the hue is undefined on the grey-axis where the saturation is low, with this weighting unstable hue derivatives at low saturations will be suppressed.

$$\left|\mathbf{H}_x^Q\right| = s \cdot h_x \tag{3.29}$$

$$HSI = \begin{pmatrix} h \\ s \\ i \end{pmatrix} = \begin{pmatrix} \arctan(\frac{o1}{o2}) \\ \sqrt{o1^2 + o2^2} \\ o3 \end{pmatrix} \tag{3.30}$$

Gaussian Colour Model:

The explanation on Section 3.5.2 regarding the invariants of Geusebroek et al. (2001), dealt with the theoretical aspects of obtaining spectral-spatial photometric invariants. Those expressions explored the infinitely dimensional Hilbert space of spectra with an infinitesimally small spatial spacing. However, those spectral-spatial energies are in practice only measurable at a certain spatial resolution range and spectral bandwidth. The Gaussian Colour Model is used to probe the spatial and spectral dimensions at a selected spectral bandwidth, analogously to image Gaussian scale-space analysis. In order to model $E(\lambda)$, which is the energy distribution of the incident light seen by the camera, a Gaussian $G(\lambda_0, \sigma_\lambda)$ at spectral scale σ_λ positioned at λ_0, is the function that is used to probe the spectral-spatial space. The spectral energy distribution can then be approximated by a Taylor expansion at λ_0 in terms of the spectral derivative quotients $E(\lambda)$, $E_\lambda(\lambda)$ and ,$E_{\lambda\lambda}(\lambda)$.

Expressions can then be derived which allow these quotients to approximate the Hering basis (Hering, 1964) of human colour vision when being truncated at second order and assuming the parameters $\lambda_0 \approx 520$ nm and $\sigma_\lambda \approx 55$ nm. The Hering human colour basis is represented by the CIE 1964 XYZ colour space, which can be obtained from the RGB space via a linear transformation. This

allows for a direct linear transformation to be found that projects RGB values to the Gaussian Colour Model, shown in Equation 3.31. This transformation approximates the spectral derivative quotients of $E(\lambda)$,$E_\lambda(\lambda)$ and ,$E_{\lambda\lambda}(\lambda)$, the $\hat{}$ above the quotients in the equation denote that they are approximations:

$$GCM = \begin{pmatrix} \hat{E} \\ \hat{E}_\lambda \\ \hat{E}_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{3.31}$$

This model generates a three-channel image, with E containing the illuminant intensity, and the two other components containing opponent chromatic information. To obtain the spectral-spatial derivatives needed to implement the local feature invariants, the images $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$ can simply be convolved with Gaussian derivative filters to obtain the image gradients. Before detailing in the next section how all the aforementioned invariants are utilised to implement local image features, this section will close with illustrations of the colour distributions of the four relevant colour spaces along with the standard RGB space. Images from the same scene are chosen, but varying in illumination conditions. The colour distribution plots extracted from each image are 3D point clouds where each pixel in the image populates the 3D colour space with its colour. These plots will serve to compliment the discussed mathematical transformations and theory, by explicitly showing the visual differences between them.

Figure 3.10 shows the distributions of the RGB space. It is difficult to capture the essence of a dense 3D plot with a 2D representation, but the correlated nature of the RGB space can still be seen. The relative spatial relationships between the RGB colours across the various illumination conditions, vary more than in the other colour spaces. In contrast, the HSI plots of Figure 3.11 show how the colours are organised along the hue axis and maintain their relative positions in that axis across varying illumination.

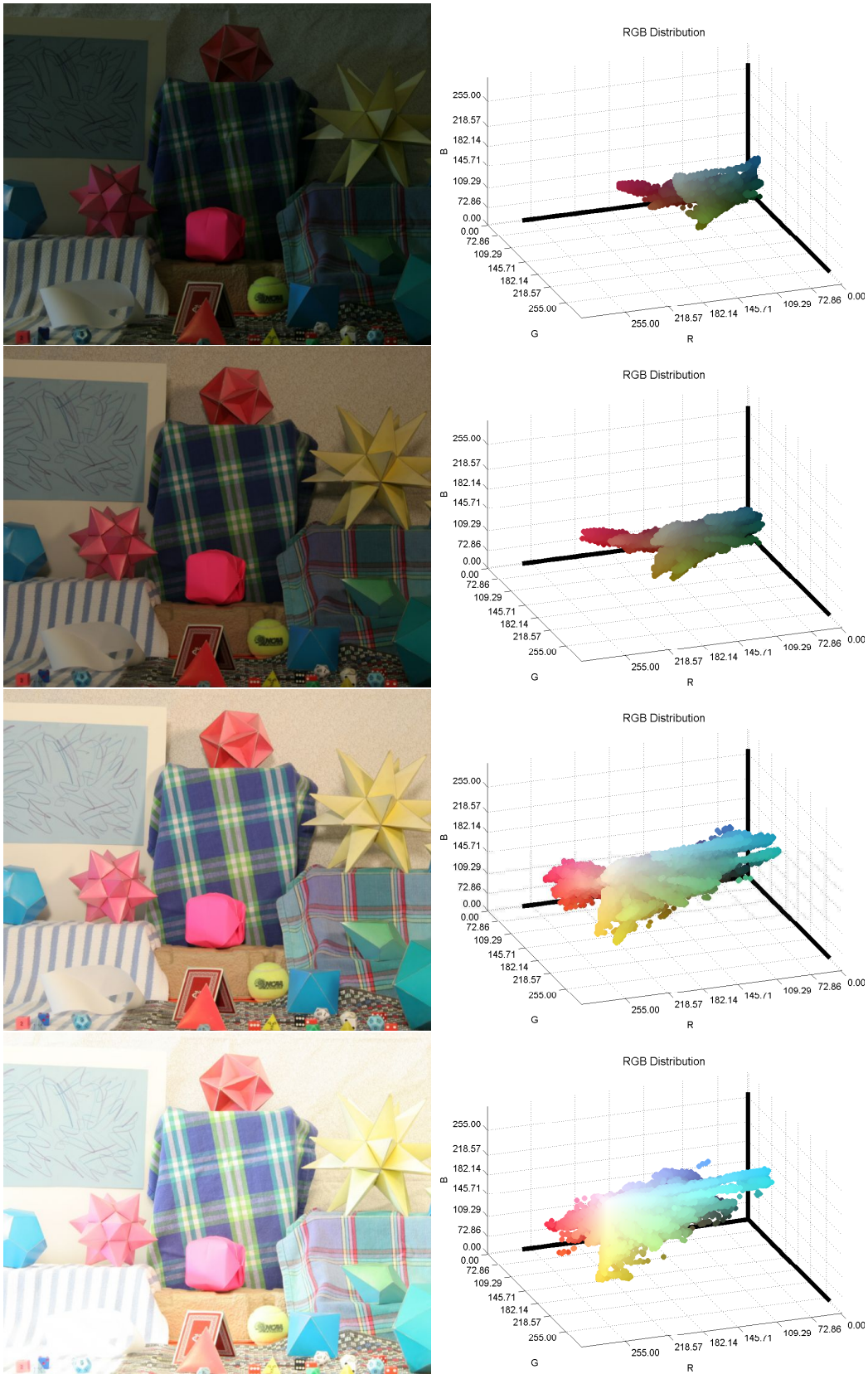


Figure 3.10: RGB colour space distributions across examples of the *moebius* scene.

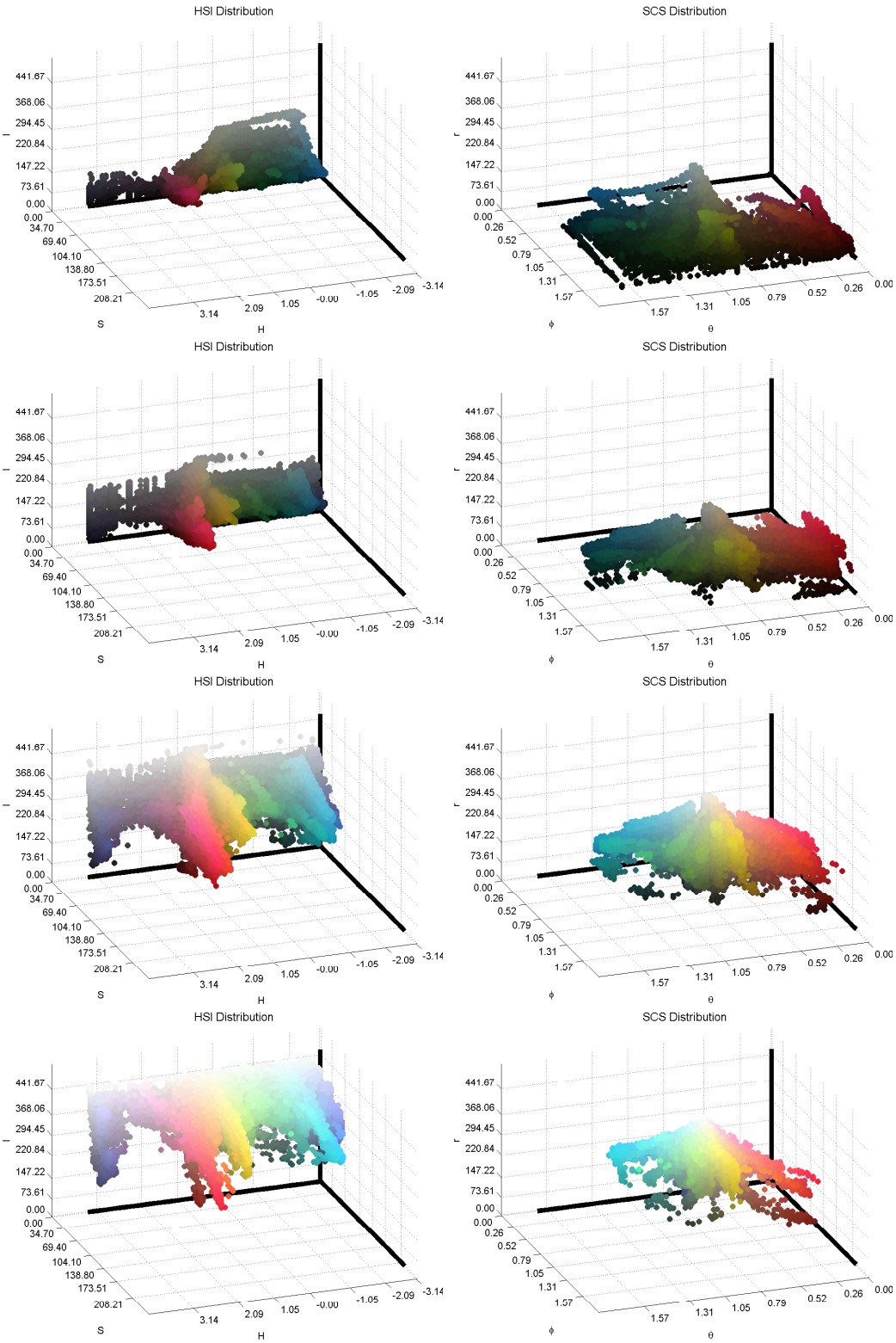


Figure 3.11: HSI colour space (left column), and spherical colour space (right column) distribution examples.

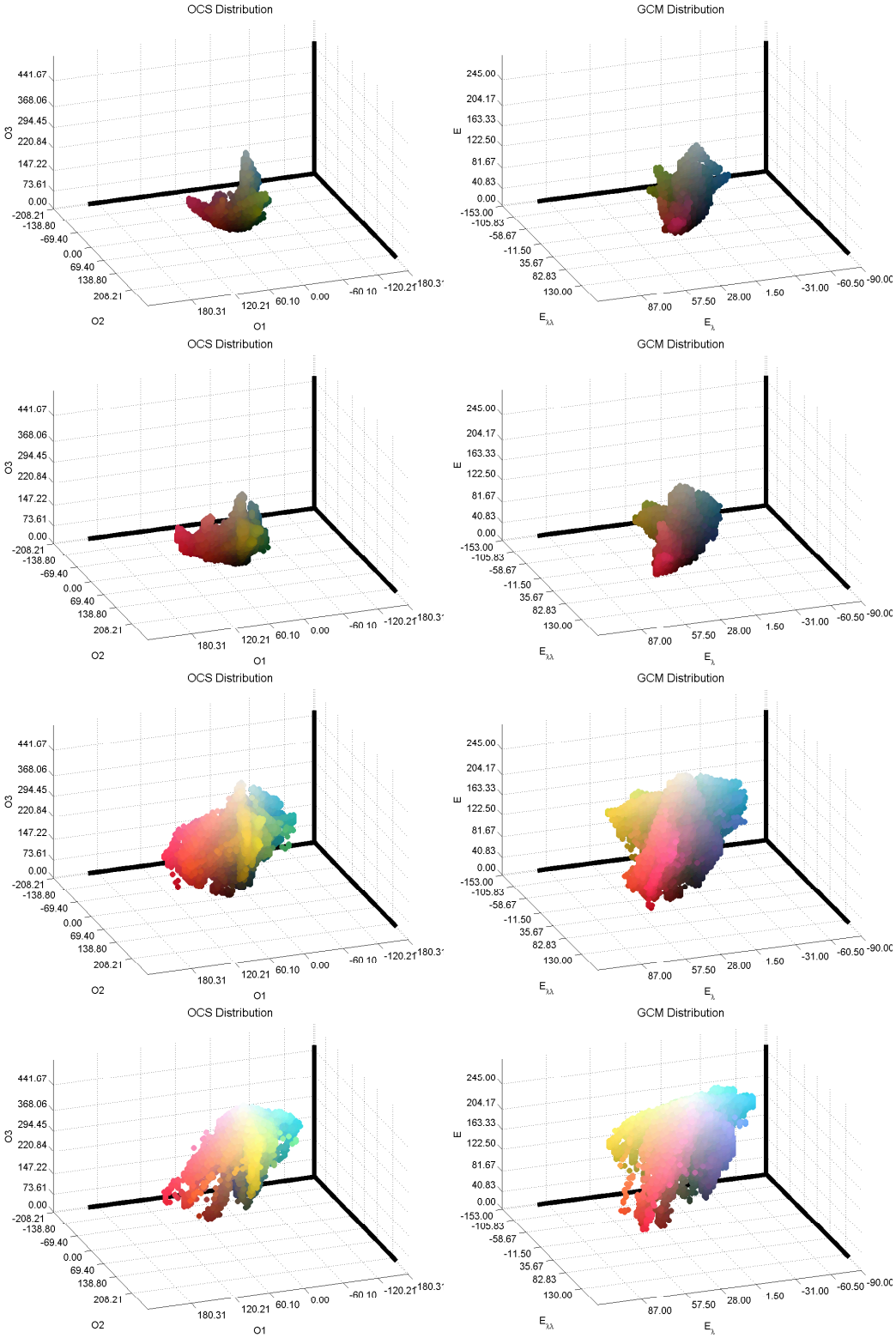


Figure 3.12: Opponent colour space (left column), and Gaussian colour space (right column) distribution examples.

3.6 Colour Invariant Features

The colour local image features developed in this research, are composed of a Harris-Laplace detection step followed by a SIFT description. It is straightforward to adapt these two techniques to colour, by identifying appropriate colour gradients to replace the grayscale gradients that those algorithms were designed to work with. In the case of the HL detector, the chosen colour invariants will be used for the first order L_x and second order L_{xx} image derivatives terms in Equations 3.1 (structure tensor) and 3.3 (LoG operator). The L_x derivatives are also used within the SIFT descriptor algorithm, which in this research is implemented by modifying the code by Kovnatsky (2010). That particular version of the SIFT algorithm is chosen for the adaptation, because it was implemented in Matlab and allows access to the image gradients with relative ease.

Table 3.2 summarises the implementation of all the nine chosen colour gradients, from their respective colour spaces. Since the HL and SIFT algorithms work with single-channelled gradients, the magnitude of the colour invariants are used in order to combine the gradients from multiple channels. The first five gradient types are from the works of Van de Weijer et al. (2005, 2006a). The shadow-shading quasi-invariant $\mathbf{S}_x^Q$ from Equation 3.25, is denoted as SS_{INV} in Table 3.2.

Table 3.2: Summary of the implementation of the colour invariants.

	SP_{INV}	$SPSS_{INV}$	SS_{F-INV}
L_x	$\sqrt{(o1_x)^2 + (o2_x)2}$	$\lvert h_x s\rvert$	$\sqrt{(\varphi_x)^2 + (\sin(\varphi)\theta_x)^2}$
L_{xx}	$\sqrt{(o1_{xx})^2 + (o2_{xx})^2}$	$\lvert h_{xx} s_x\rvert$	$\sqrt{(\varphi_{xx})^2 + (\sin(\varphi_x)\theta_{xx})^2}$

	SS_{INV}	$SPSS_{VAR}$	LIC
L_x	$r\sqrt{(\varphi_x)^2 + (\sin(\varphi)\theta_x)^2}$	$\sqrt{(i_x)^2 + (s_x)^2}$	$\sqrt{(h_x s)^2 + (s_x)^2}$
L_{xx}	$r_x\sqrt{(\varphi_{xx})^2 + (\sin(\varphi_x)\theta_{xx})^2}$	$\sqrt{(i_{xx})^2 + (s_{xx})^2}$	$\sqrt{(h_{xx} s_x)^2 + (s_{xx})^2}$

	C_{INV}	H_{INV}	W_{INV}
L_x	$\sqrt{\left(\frac{\hat{E}\hat{E}_{\lambda x} - \hat{E}_\lambda \hat{E}_x}{\hat{E}^2}\right)^2 + \left(\frac{\hat{E}\hat{E}_{\lambda\lambda x} - \hat{E}_{\lambda\lambda}\hat{E}_x}{\hat{E}^2}\right)^2}$	$\sqrt{\left(\frac{\hat{E}_{\lambda\lambda}\hat{E}_{\lambda x} - \hat{E}_\lambda \hat{E}_{\lambda\lambda x}}{\hat{E}_\lambda^2 + \hat{E}_{\lambda\lambda}^2}\right)^2}$	$\sqrt{\left(\frac{\hat{E}_x}{\hat{E}}\right)^2 + \left(\frac{\hat{E}_{\lambda x}}{\hat{E}}\right)^2 + \left(\frac{\hat{E}_{\lambda\lambda x}}{\hat{E}}\right)^2}$
L_{xx}	$\sqrt{\left(\frac{\hat{E}\hat{E}_{\lambda xx} - \hat{E}_\lambda \hat{E}_{xx}}{\hat{E}^2}\right)^2 + \left(\frac{\hat{E}\hat{E}_{\lambda\lambda xx} - \hat{E}_{\lambda\lambda}\hat{E}_{xx}}{\hat{E}^2}\right)^2}$	$\sqrt{\left(\frac{\hat{E}_{\lambda\lambda}\hat{E}_{\lambda xx} - \hat{E}_\lambda \hat{E}_{\lambda\lambda xx}}{\hat{E}_\lambda^2 + \hat{E}_{\lambda\lambda}^2}\right)^2}$	$\sqrt{\left(\frac{\hat{E}_{xx}}{\hat{E}}\right)^2 + \left(\frac{\hat{E}_{\lambda xx}}{\hat{E}}\right)^2 + \left(\frac{\hat{E}_{\lambda\lambda xx}}{\hat{E}}\right)^2}$

The specular quasi-invariant $\mathbf{O}_x^Q$ from Equation 3.27 is denoted as SP_{INV}. $SPSS_{INV}$ is the specular-shadow-shading quasi-invariant $\mathbf{H}_x^Q$ from Equation 3.29. Its variant counterpart $\mathbf{H}_x^V$ (Van de Weijer et al., 2005), called $SPSS_{VAR}$ in the table is not an invariant gradient but it is included here for a more complete evaluation and comparison of the invariants alongside the luminance intensity. The shadow-shading full invariant $\mathbf{s}_x$ (Van de Weijer et al., 2006a), given in Equation 3.13 is referred to in the table as SS_{F-INV}. The Light-Invariant Colour (LIC) gradient proposed by Stöttinger et al. (2012) is similar to $SPSS_{INV}$ but it has the additional saturation derivative. The last three invariants C_{INV}, H_{INV} and W_{INV} are based on the work of Geusebroek et al. (2001) and they use the Gaussian Colour Model. W_{INV} is invariant to illumination intensity, and in this implementation it is a summation of W_x from 3.23 with $W_{\lambda x}$ and $W_{\lambda\lambda x}$ from 3.24. C_{INV} is invariant to shadow and shading effects, formed from a summation of $C_{\lambda x}$ and $C_{\lambda\lambda x}$ from Equation 3.21. The final gradient type H_{INV}, is shadow, shading and highlight invariant, referred to as H_x in Equation 3.18.

The invariants developed by (Van de Weijer et al., 2005, 2006a) were adapted from their released code[6] (*Color Feature Detection I & II*). The LIC, SP_{INV}, SS_{INV}, $SPSS_{VAR}$ and SS_{F-INV} invariants are composed of square roots of two summed components, these colour components are scaled so an equal contribution for the gradient magnitude is provided by each. This implementation of the C and W invariants differs from the evaluation of Burghouts and Geusebroek (2009), in that here the invariants are composed of the combined gradient magnitudes of the separate $C_{\lambda x}$, $C_{\lambda\lambda x}$, W_x, $W_{\lambda x}$ and $W_{\lambda\lambda x}$ invariants. There is no scaling performed on these components before calculating the magnitude, as experiments revealed this deteriorated the results.

The term L_x in Table 3.2 refers to the L_x of the structure tensor (Equation 3.1) and the derivatives used for the SIFT descriptor, the x subscript denotes the first order derivatives in the x-direction. To obtain the second order derivatives for the LoG operator, the expressions for L_{xx} in Table 3.2 are used. The second order derivatives used here for C, H and W are different from the original work. Three different methods for each of the three invariants were implemented and tested in a point correspondence experiment on the Middlebury dataset. The

[6] http://cat.cvc.uab.es/ joost/software

first method was the original second order spatial derivatives, denoted as the set $\frac{\delta}{\delta_{ww}}$ in the work of Geusebroek et al. (2001). The second method simply involved convolving the L_x gradient images with a first order Gaussian derivative kernel. Lastly, the method shown in Table 3.2 replaces all first derivative terms from L_x with their second order counterparts. Results of the test are shown in Figure 3.13, and the methods denoted as C_{INV}, H_{INV} and W_{INV} in the figure prove superior due to having generated a greater number of correct correspondences. The reason for why the original second order invariants of Geusebroek et al. (2001) perform worse than the simplified variations proposed in this work, can be attributed to the complexity of the original second order invariant expressions. They contain many more partial derivative and second derivative terms, which introduce more instabilities than the invariants proposed here.

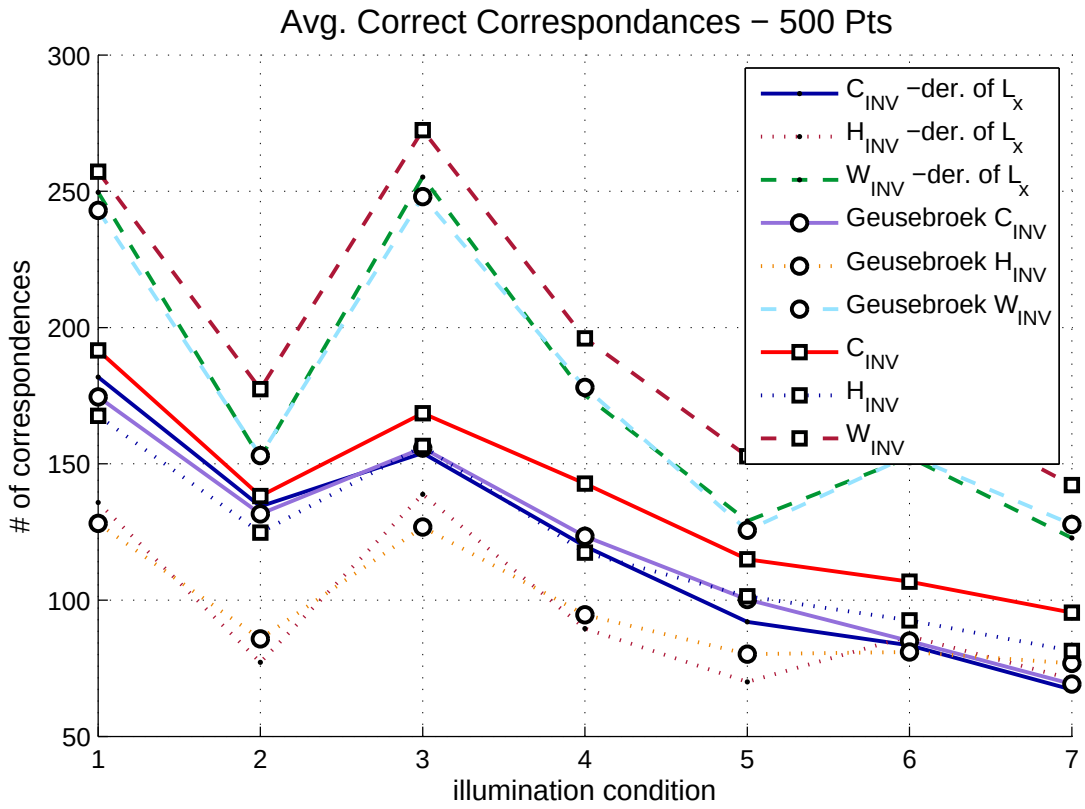


Figure 3.13: Variations of the second order spatial derivatives of the C, H and W invariants.

Figures 3.14, 3.15, 3.16 and 3.17 show the visual results of applying the various gradient types on two different images. The two images contain shadows, shadings (colour change within an object) and specularities, and it is easy to visually detect the extent of the invariance that each of the gradients can achieve. The visualisation is not a completely accurate comparison, as weak edges might not be visible, but it nonetheless highlights the main general differences between the gradients. In Figure 3.14, the specular invariant $|\mathbf{O}^{Q}_{xy}|$ can be seen to detect the shadows under the red ball and yellow ring and the shading transitions of the bottom yellow and green cubes, but the highlights on those same objects are less apparent than in the majority of the other gradients, especially compared to the luminance $|\mathbf{I}_{xy}|$ on Figure 3.15. In contrast, the shadow-shading invariant $|\mathbf{S}^{Q}_{xy}|$ in Figure 3.14, can be seen to be invariant to shadows and shadings but not to the highlights. Its full invariant $|\mathbf{s}_{xy}|$, detects the specularities and internal edges of each object to a lesser extent.

Shadows, shadings and highlights are not detected in the specular-shadow-shading invariant $|\mathbf{H}^{Q}_{xy}|$, only the boundary of each object is visible. Its variant counterpart $|\mathbf{H}^{V}_{xy}|$ detects most of the image gradients. The light-invariant colour invariant $|\mathbf{LIC}_{xy}|$ in Figure 3.15, provides a good balance of invariance to shadows, shadings and specularities, while detecting the boundaries of the objects. $|\mathbf{C}_{xy}|$ obtains good results in terms of invariance to shadows and shading, and $|\mathbf{H}_{xy}|$ behaves as expected by not detecting any shadows or specularities despite generating weaker gradients. The illuminant intensity invariant $|\mathbf{W}_{xy}|$, detects the object boundaries better than the intensity $|\mathbf{I}_{xy}|$. Even though $|\mathbf{W}_{xy}|$ is not invariant to shadow-shadings or specularities, its invariance comes into play when the scene varies in illumination intensity. In summary, all 9 colour gradients in Figures 3.14 and 3.15 behave as expected for that particular image. Other examples of the gradient's behaviour on a more complicated image, are shown on Figures 3.16 and 3.17.

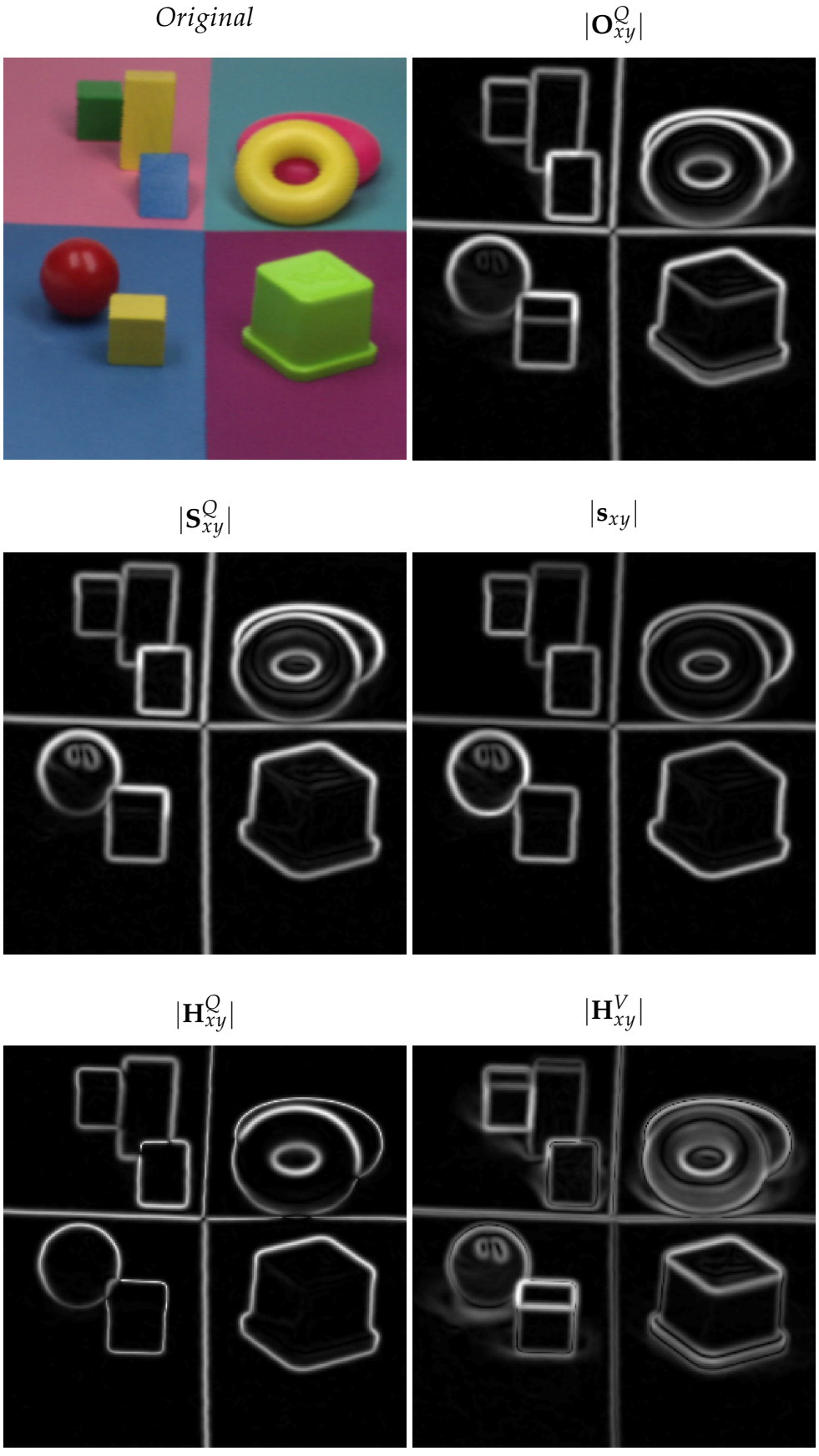


Figure 3.14: Visual examples of the colour invariant gradients.

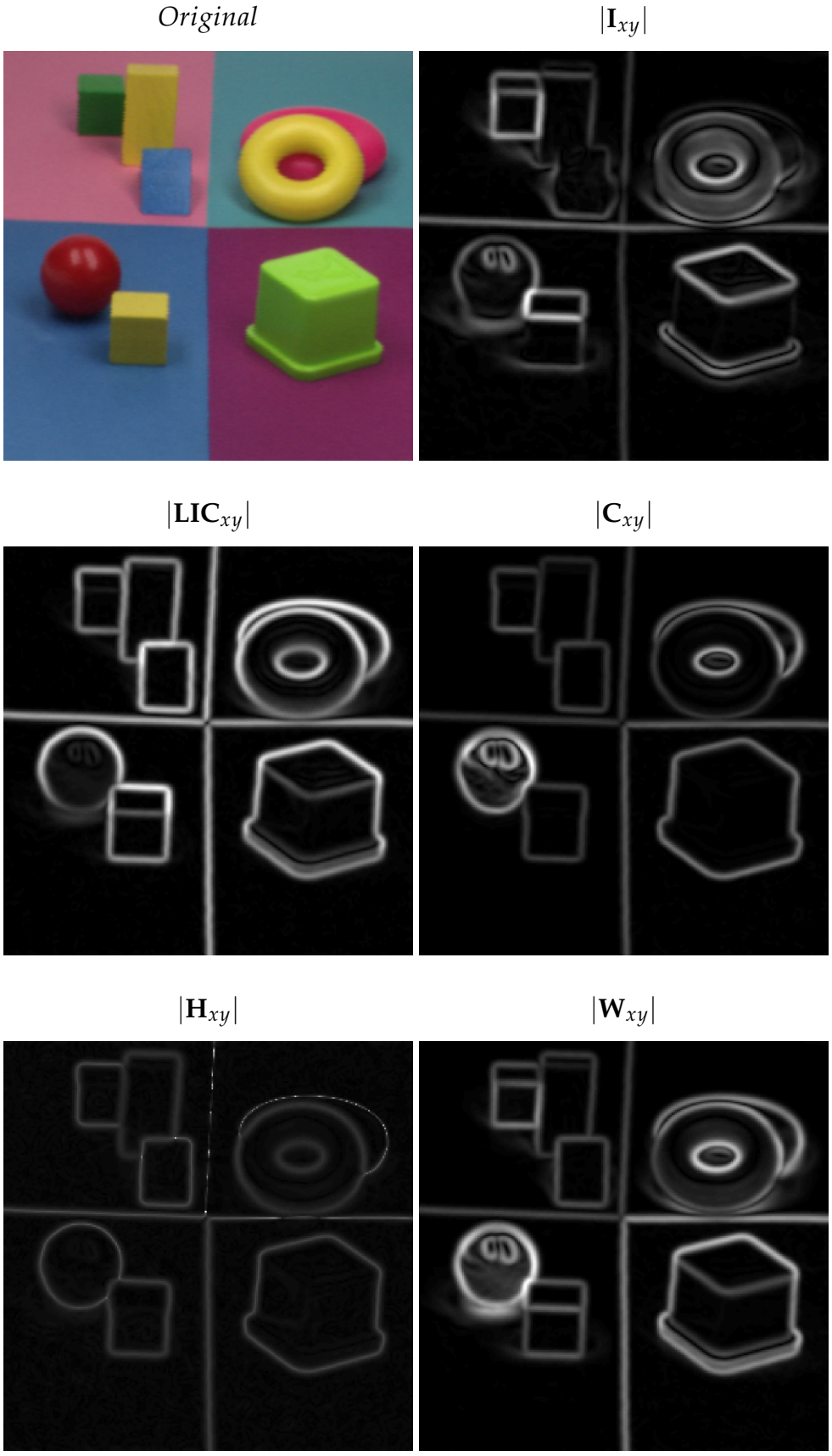


Figure 3.15: Visual examples of the colour invariant gradients.

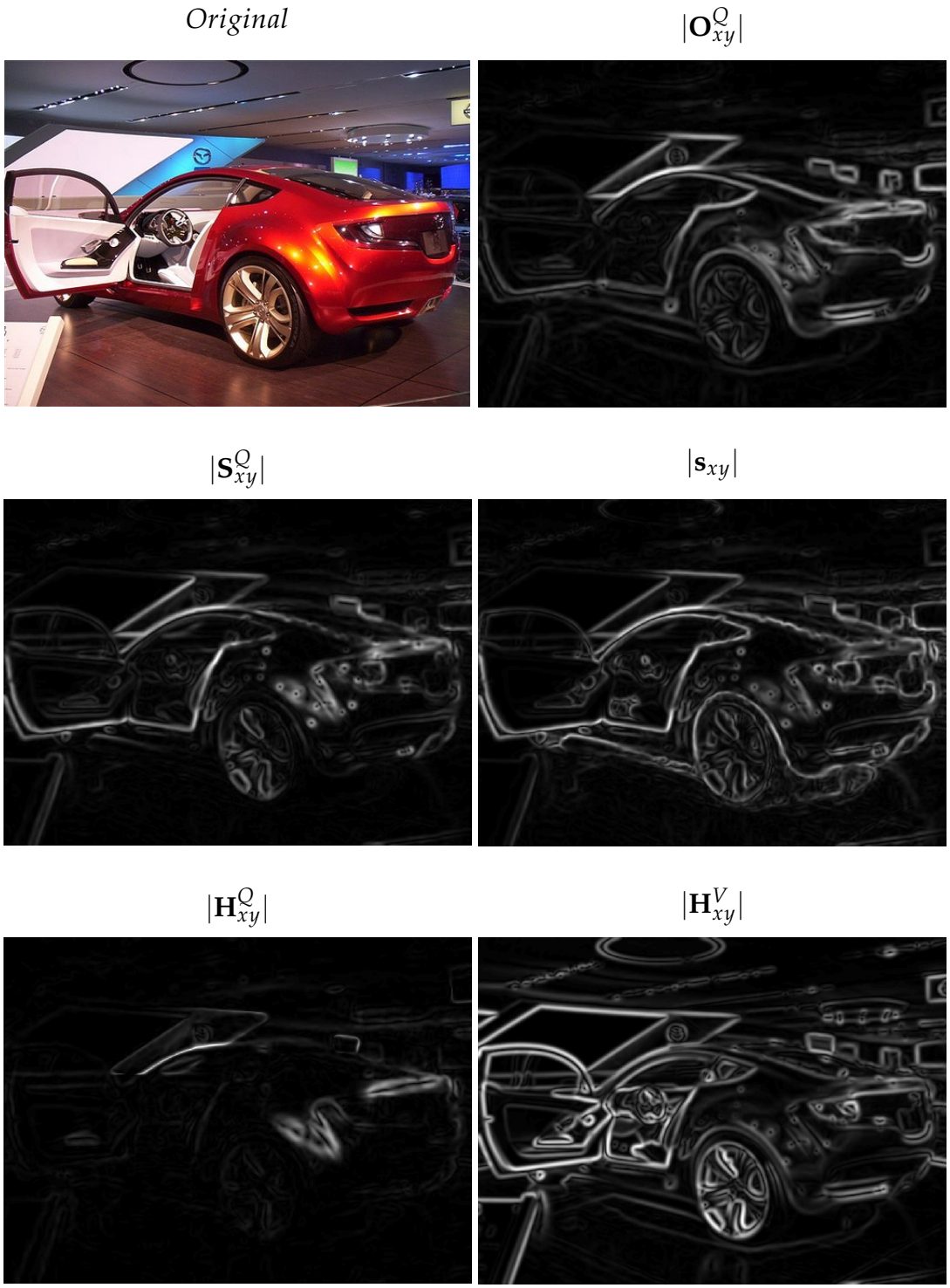


Figure 3.16: Visual examples of the colour invariant gradients.

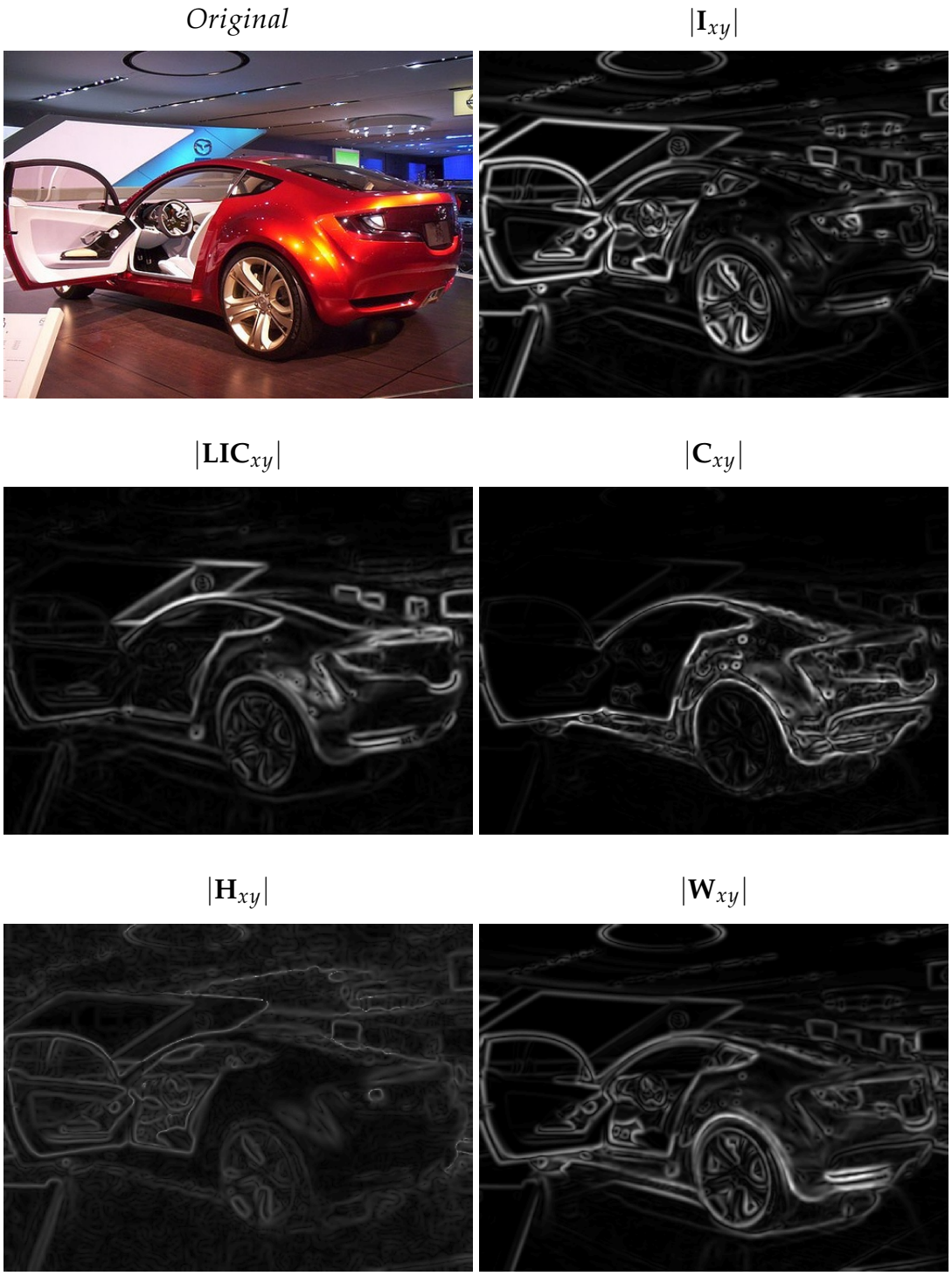


Figure 3.17: Visual examples of the colour invariant gradients.

3.7 Summary

This concludes this chapter, which has covered both background information and work contributions related to the Harris-Laplace detection algorithm and the photometric colour invariants. A detailed explanation of the implemented colour features is given in this chapter, which saw the optimisation of the HL algorithm parameters, the adaptation of the invariants from the literature and the visual results of their behaviour on two example images.

In the HL optimisation study, different Non-Maximum Suppression and characteristic scale estimation techniques were tested. Visual results of the scale estimation process were provided, along with the advantages and disadvantages of the implemented method. The optimisation study resulted in an optimum set of parameters which for the specific implementation of this research, proved to be better than the original method presented by Mikolajczyk and Schmid (2001). The colour invariant sections of this chapter, discussed the theory of obtaining three types of photometric colour invariants: shadow-shading, specular and illumination intensity invariants. The colour spaces necessary for obtaining these theoretical invariants were outlined, and their 3D distributions visually compared under varying illumination conditions.

A precise account of this research's implementation of the invariants is given, highlighting any differences with the original works and providing justifications for any changes. One such change regards the second order spatial derivatives of C, H and W; in which the final implementation method proposed here achieves better results than the original work by Geusebroek et al. (2001). The next two chapters give the account of the experimental results and the evaluation of the invariants, for feature matching in Chapter 4, and object class recognition in Chapter 5.

Feature Detection and Matching

4

The first part of this chapter presents the evaluation of the colour invariant gradients when applied to local feature detection and matching tasks, the second part presents the investigation of the colour feature fusion. The feature detection and matching experiments are performed on the four image matching datasets discussed in Section 2.5.3, and examine the number of correct point correspondences, the number of correctly matched descriptors, the detection repeatability and the descriptor matching score; which were explained in Section 2.5.

In the feature fusion investigation, a novel concept of a feature correlation analysis is presented that investigates the number of unique correctly detected features from each gradient type, and the similarity between the gradient types by calculating how many of the same features are mutually extracted between them. The last part of the fusion investigation, proposes fusion techniques that utilise multiple gradient types conjointly in the HL feature detection process.

4.1 Feature Extraction Visualisation

The visual output of the feature detection process is shown in Figures 4.2, 4.3 and 4.4. These figures show the extracted local regions for two images from the set *Moebius*. The images allow to see and compare the types of image structures that are deemed to be features by the various types of HL detectors, and show examples of how the selection of those features can change once the illumination condition is varied. The goal of the feature detection evaluation is to compare the first image of an image-set of the same scene with all other subsequent images containing different imaging conditions, and detect how many local features correspond at the same scene location in both of the images. In the

feature matching evaluation, the testing concept only differs in that it is the descriptors of each feature that are matched in both images, and a match is considered to be correct only if the features' locations are deemed to correspond to the same scene position. Examples of the feature matching are provided in Figure 4.1, where a subset of grayscale intensity feature matches are shown for image pairs of the image-sets *Graffiti* and *Art*. The descriptor regions are indicated by the red squares, and their main orientations (needed for the SIFT descriptor) are denoted by the internal line.

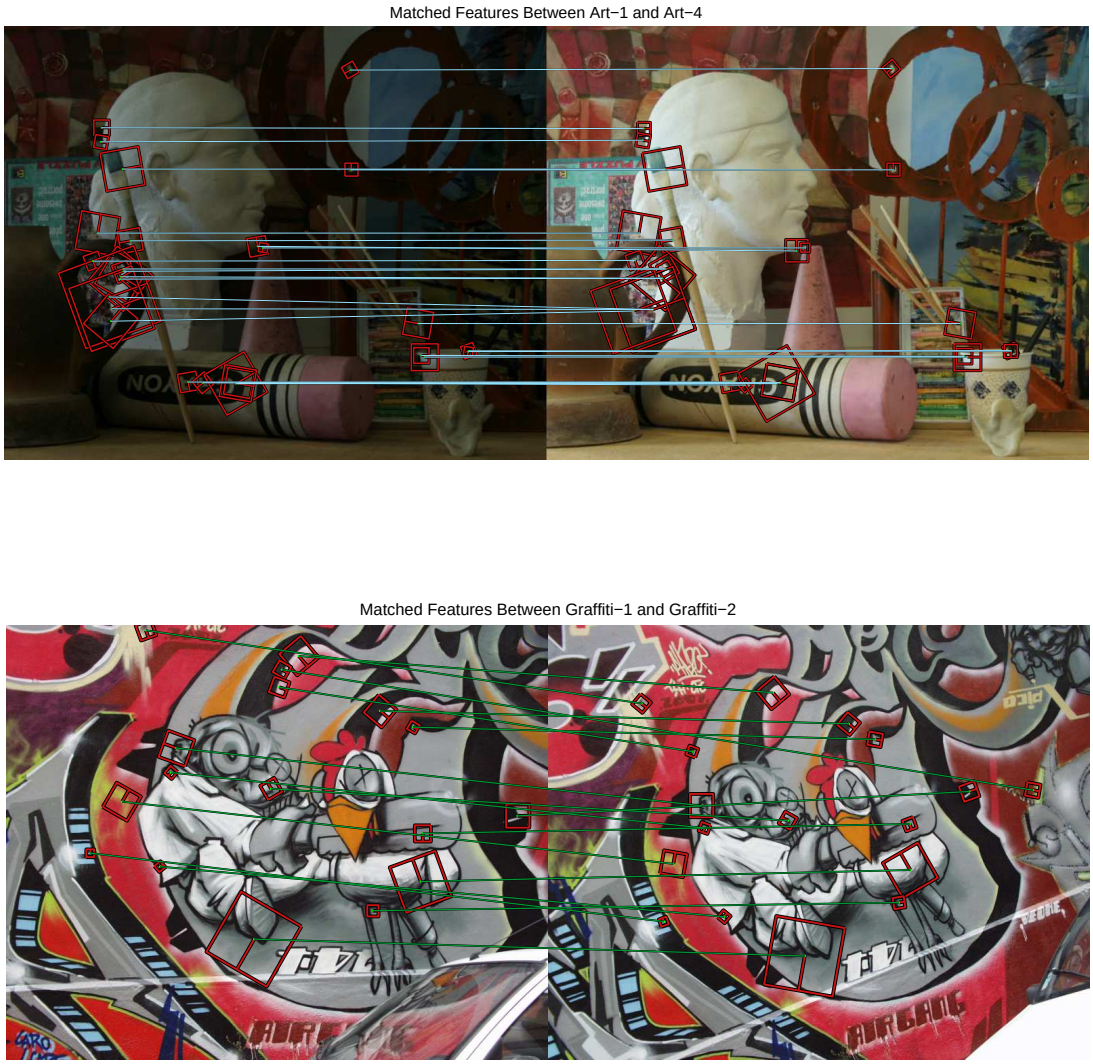


Figure 4.1: Illustration of local feature matching results of a subset of grayscale intensity descriptors.

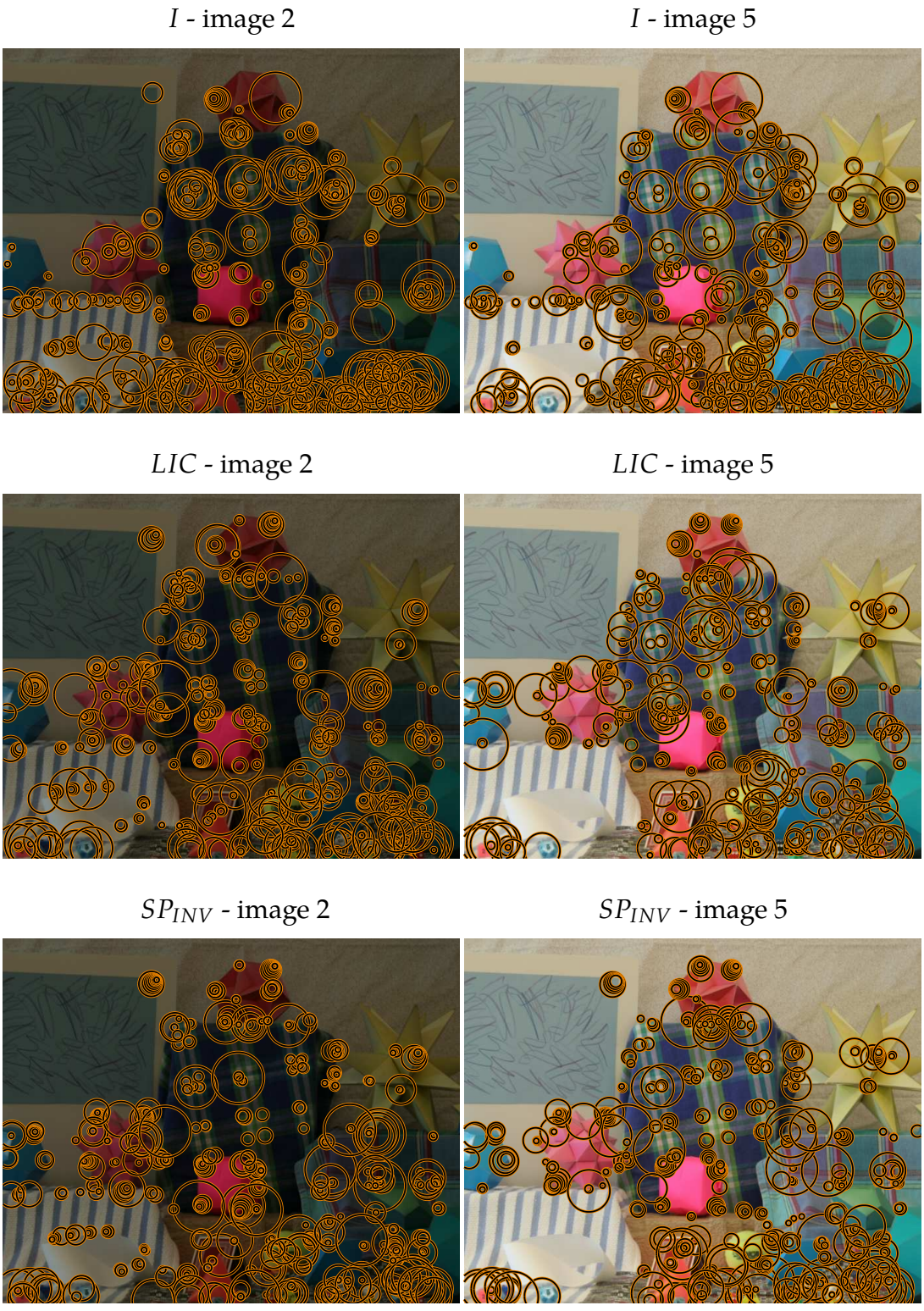


Figure 4.2: Visual illustration of the local feature extraction results on two different imaging conditions of the *Moebius* set, using three separate gradient types: Luminance, LIC and SP_{INV}.

Figure 4.3: Visual illustration of the local feature extraction results on two different imaging conditions of the *Moebius* set, using three separate gradient types: $SPSS_{INV}$, $SPSS_{INV}$ and SS_{F-INV}.

Figure 4.4: Visual illustration of the local feature extraction results on two different imaging conditions of the *Moebius* set, using three separate gradient types: C_{INV}, H_{INV} and W_{INV}.

4.2 Feature Detection Evaluation

Mikolajczyk and Schmid (2005) propose an approach to evaluate the quality of local interest point detection using robust metrics, the approach has become the most widely used evaluation framework for testing state of the art local features. As outlined in Section 2.5, the same evaluation method is followed in this work to provide standardised results. In the work of Mikolajczyk and Schmid (2005), two local features from separate images are considered to correctly correspond if the area of the scene that they describe overlaps by more than 60%. In this evaluation, a more strict threshold of 90% overlap is set to evaluate the localisation stability of the colour gradients more robustly.

Figures 4.5 and 4.6 show the summary of the results for the mean number of correct correspondences across all imaging distortion levels for all four datasets. Figures 4.7 and 4.8 present the results of the detection repeatability rates. For the presented results, different number of HL points were extracted from the images of each dataset according to the size of the images and image content. A study was performed in order to determine the appropriate number of HL features that would be extracted from the images. This study is presented in Appendix A, and the results show the number of correct correspondences and repeatability rates achieved by extracting varying numbers of HL points on the Middlebury dataset (Figures A.1, A.2, A.3, A.4 and A.5), and shows the repeatability rates achieved on the Oxford dataset in Figure A.6. The points were varied from 500 to a total of 3,000 in increments of 500, and the aim of the study was to select the parameter that achieved the highest repeatability rates. On the Middlebury set, the best parameter proved to be 500 points, as it performed better for 40% of the results and was equal to the top performers for another 40% of the results in Figures A.1, A.2, A.3, A.4 and A.5.

In the Oxford results of Figure A.6 it can be seen that extracting 500 to 2000 points per image achieves very similar repeatability rates. The chosen parameter for the Oxford dataset was thus 1000 points, since the Oxford images are larger than the Middlebury ones. As to the other datasets, 300 points were extracted from the ALOI dataset due to its images containing a texture-less background

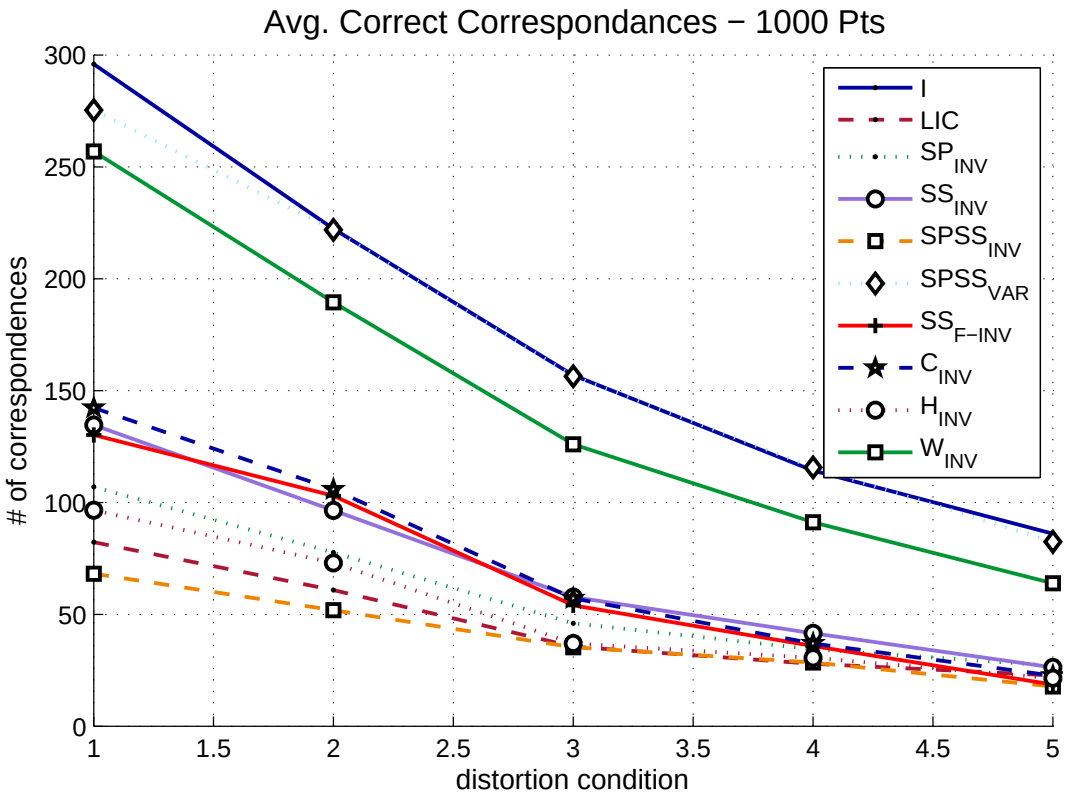
and being of similar size to the Middlebury images. The PHOS dataset also contains images with a texture-less background but it was decided to extract 500 points per image as the image sizes are larger. Visual inspection of the settings for the PHOS and ALOI HL extractions also contributed to the final choice of parameters, as the chosen number of extracted features adequately covered all of the objects in the images. In all the feature matching results of this chapter, the presented data points for each distortion level (x axis of the plots) is the average of all the results for that distortion level from all the image-sets of that database. The Oxford dataset contains 7 image-sets, Middlebury contains 5, ALOI 30 and PHOS contains 15 sets. The standard deviation plots of the results are presented in Appendix A.

Table 4.1 presents the overall results for the feature detection evaluation, combining the sum of all the correct correspondences and the repeatability rates from all the imaging distortions of all the datasets. The table allows for a high level comparison of the 10 different feature types. The W_{INV} colour invariant clearly obtains the superior overall performance, followed by the grayscale intensity I, and the specular-shadow-shading variant $SPSS_{VAR}$. These combined results however, are taken from datasets that are subjected to a majority of illumination distortions, for which the colour invariants have an advantage. In the case of the Oxford dataset which contains a variety of standard imaging distortions, the results in Figure 4.5a indicate that for general imaging conditions, I and $SPSS_{VAR}$ are in fact better gradients for local detection than the colour invariants.

Table 4.1: Cumulative sum of the feature detection evaluation results.

Metric	I	LIC	SP_{INV}	SS_{INV}	$SPSS_{INV}$	$SPSS_{VAR}$	SS_{F-INV}	C_{INV}	H_{INV}	W_{INV}
correspondences	3929	2289	3072	3041	1428	3885	2654	2630	2333	4056
cumulative % repeatability	706	438	589	576	270	699	500	511	424	744

(a) Oxford



(b) Middlebury

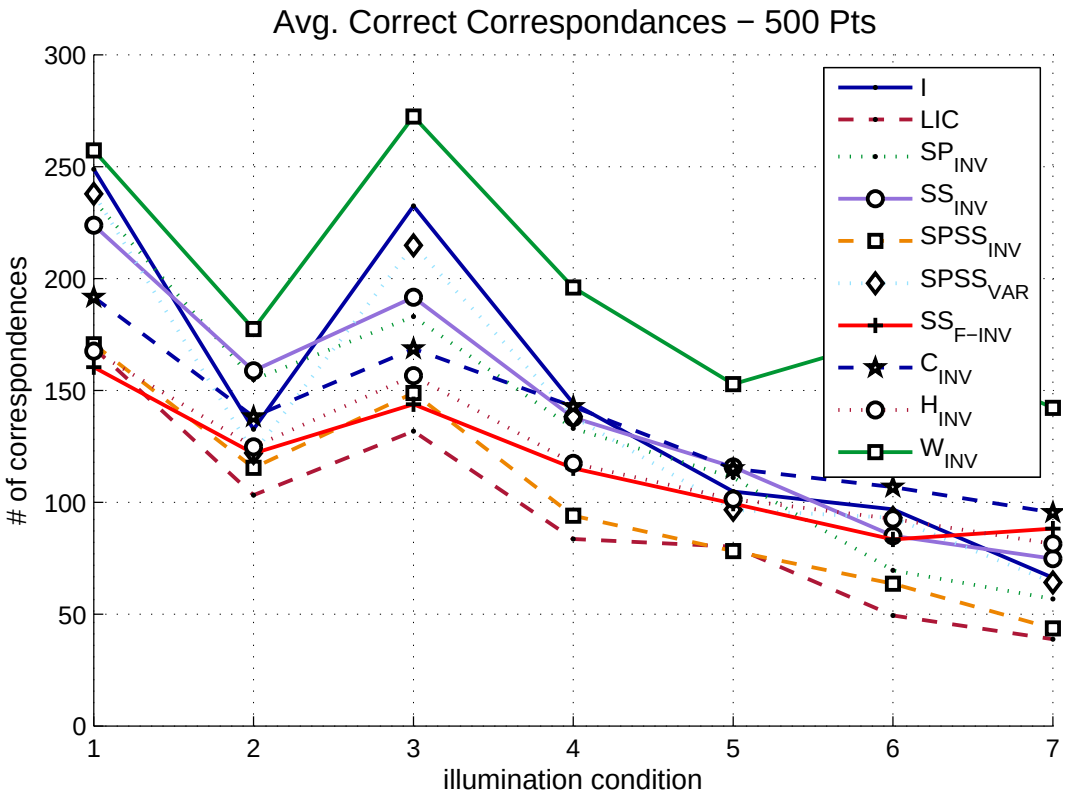
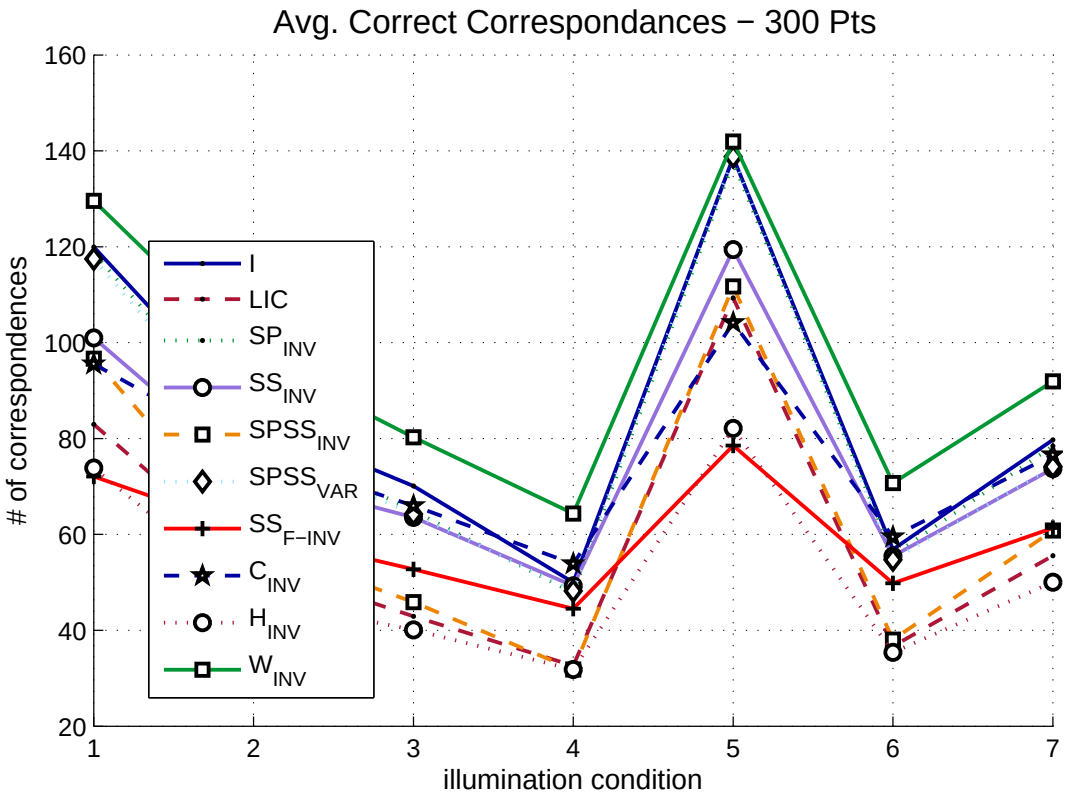


Figure 4.5: Summary of the correct correspondences analysis for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

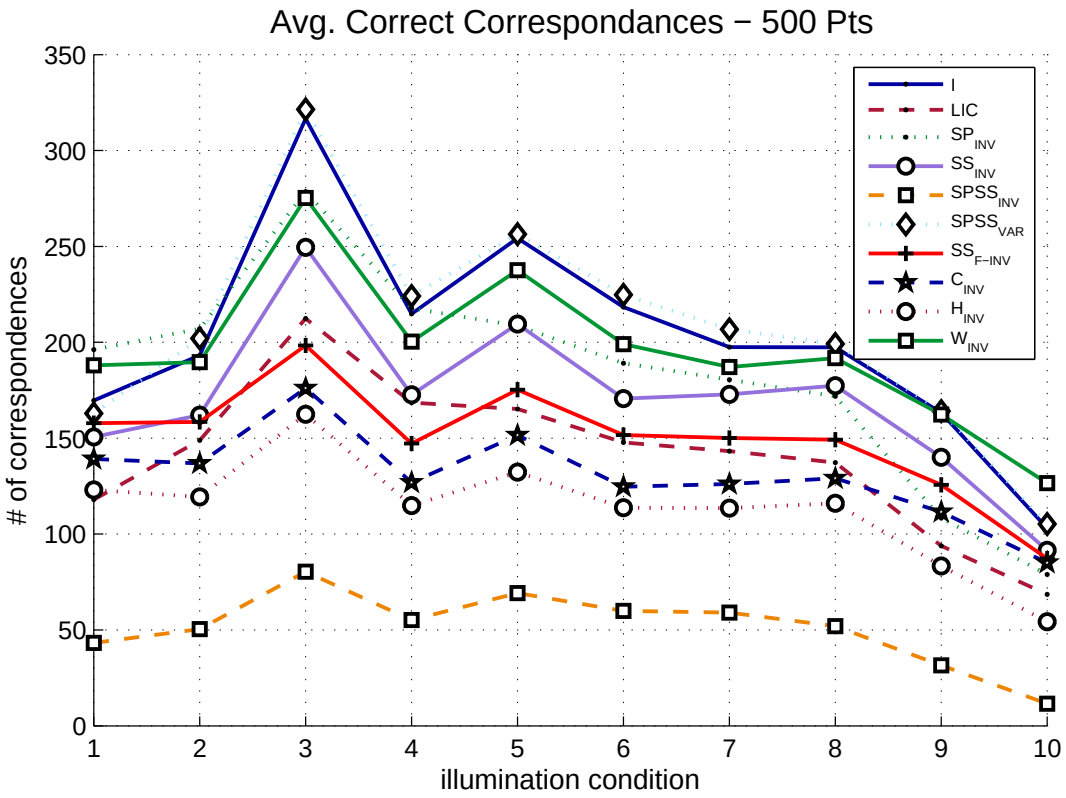
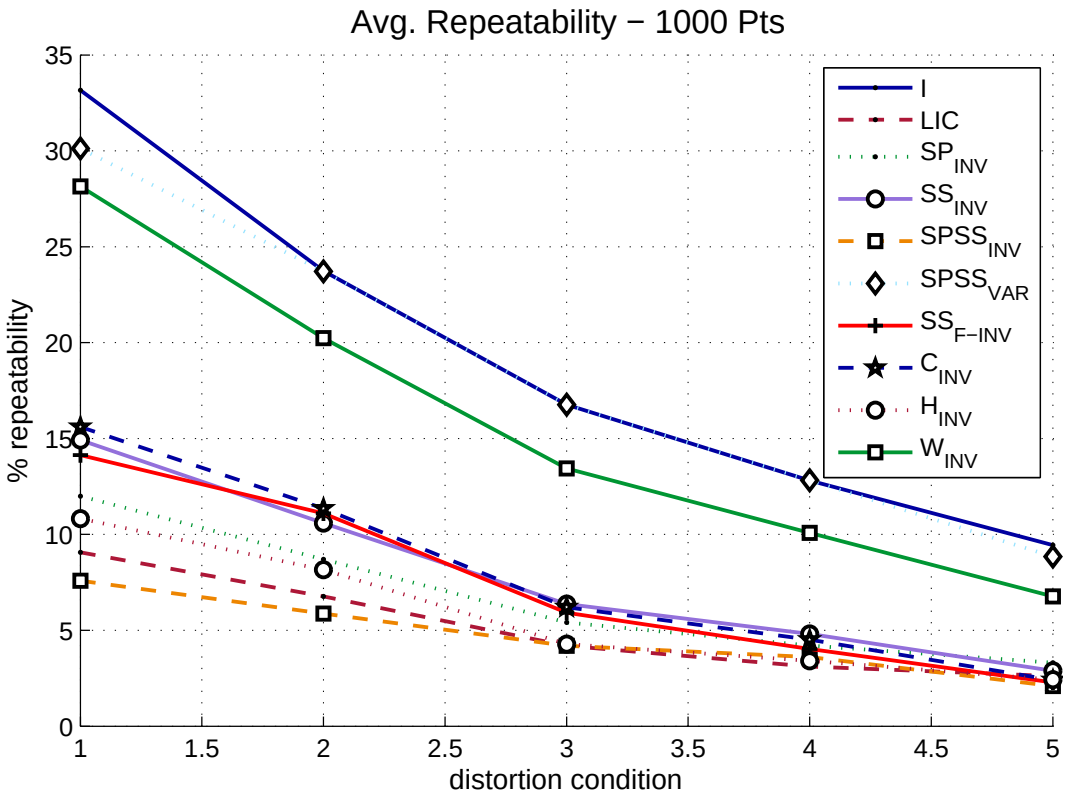


Figure 4.6: Summary of the correct correspondences analysis for the ALOI (a) and PHOS (b) datasets.

(a) Oxford



(b) Middlebury

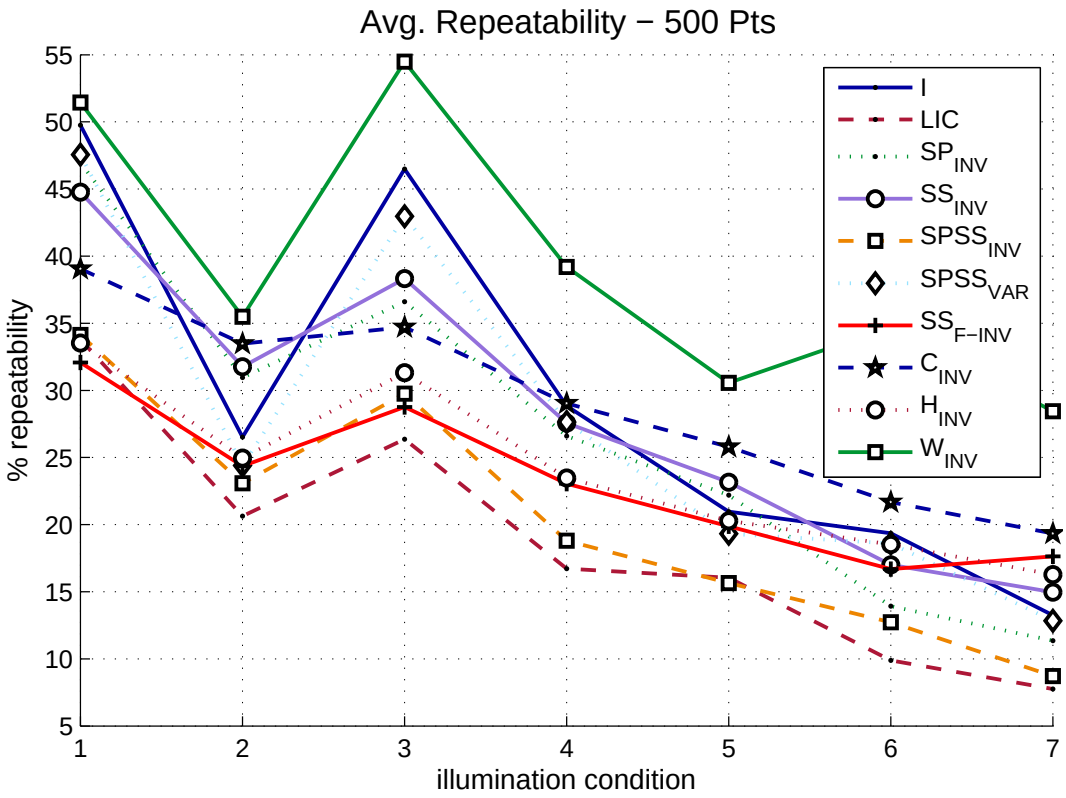
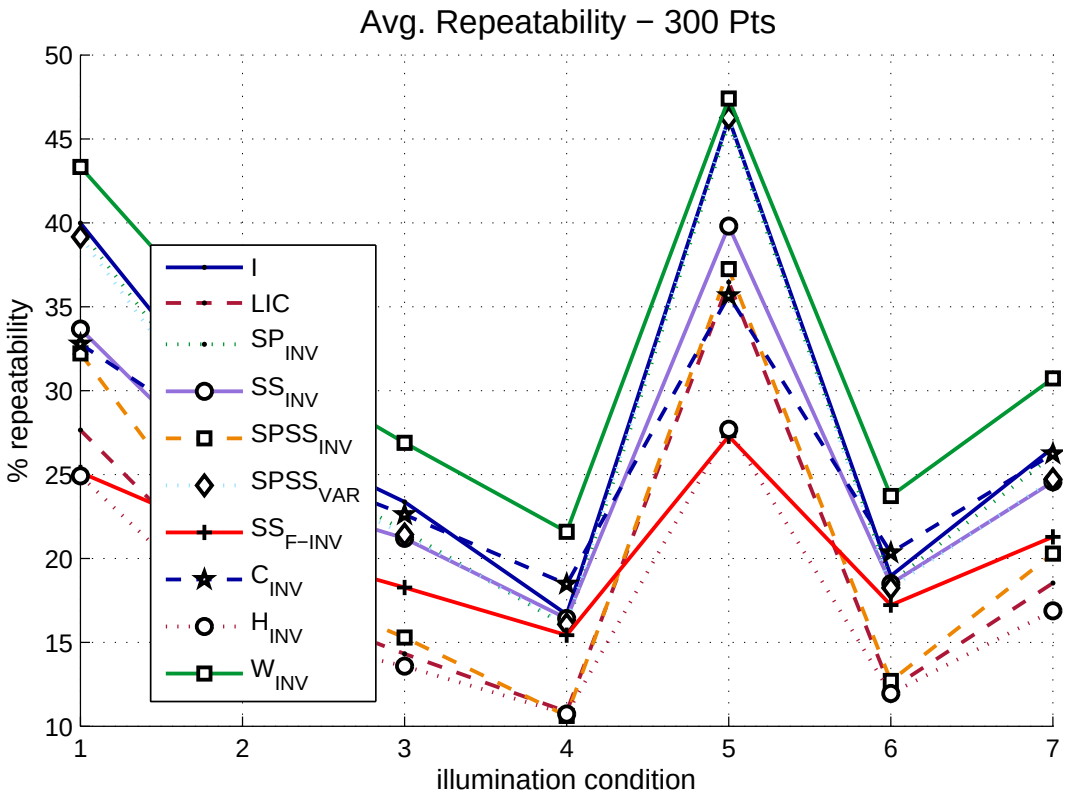


Figure 4.7: Summary of the repeatability analysis for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

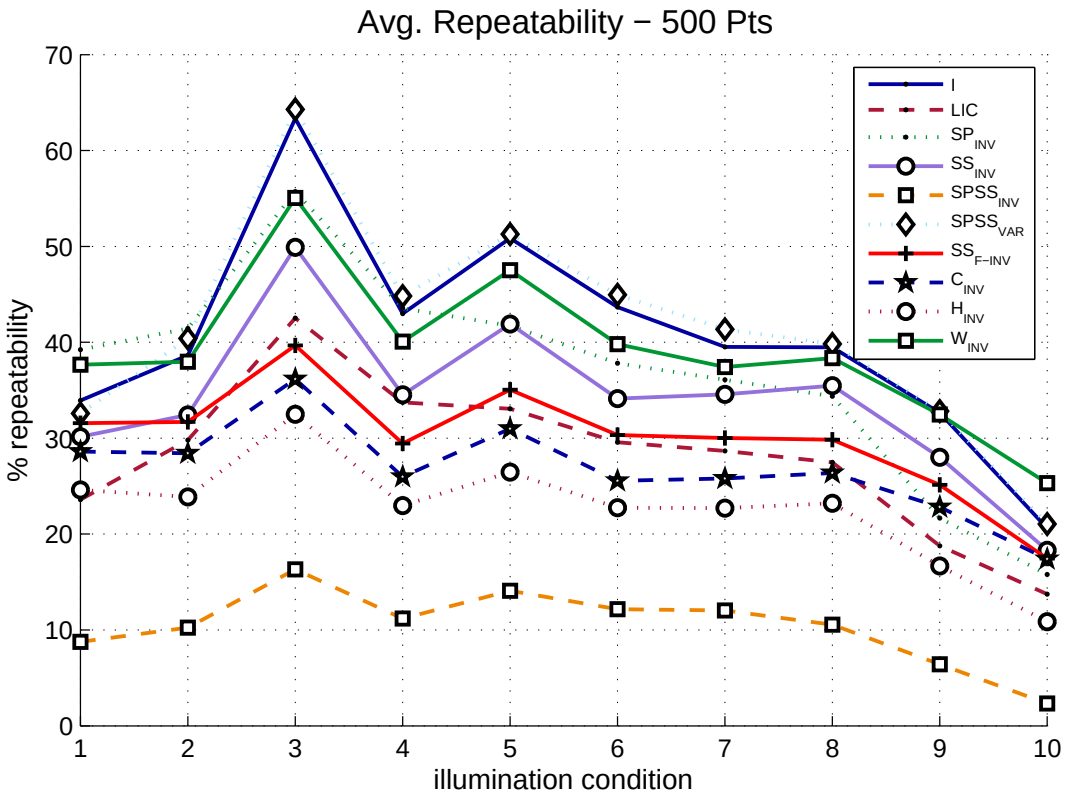


Figure 4.8: Summary of the repeatability analysis for the ALOI (a) and PHOS (b) datasets.

All colour invariants except for W_{INV}, perform significantly inferior to grayscale in the Oxford dataset as they prove to be less robust to imaging distortions, namely scale and viewpoint changes. These results demonstrate the necessity to evaluate colour invariants under a more general set of imaging conditions other than just illumination. In the results for the illumination varying datasets of Figures 4.5b and 4.6, only one colour invariant (W_{INV}) is overall superior to the grayscale intensity. The adequate robustness of luminance to general and varying illumination distortions can be attributed to the fact that the greatest image variations occur in the gray-axis of a colour space (Van de Weijer et al., 2006b), which are able to be detected more prominently by the grayscale intensity. This results in an extraction of abundant image gradients which help to mitigate the effects of imaging distortions.

The reported results of this substantial evaluation thus provide clear evidence to why grayscale is the preferred method in the literature for local feature detection. This evaluation also discovered that the W_{INV} invariant is the best performer for feature detection under illumination conditions, and has adequate robustness to general imaging conditions. Variations of the W_{INV} invariant have been ignored in the studies made by: Van De Sande et al. (2010), Abdel-Hakim and Farag (2006); though in the study of Burghouts and Geusebroek (2009) it proved to be amongst the top invariants when evaluated as a descriptor.

4.3 Feature Matching Evaluation

Two metrics are presented here for the local feature matching results, the number of correct matches and the matching score which essentially provides a qualitative result for the ability of the gradient types to generate distinct and robust descriptors. Table 4.2 presents the overall results for the feature descriptor matching evaluation, combining the sum of all the number of correct matches and the matching score (m-score) from all the imaging distortions of all the datasets.

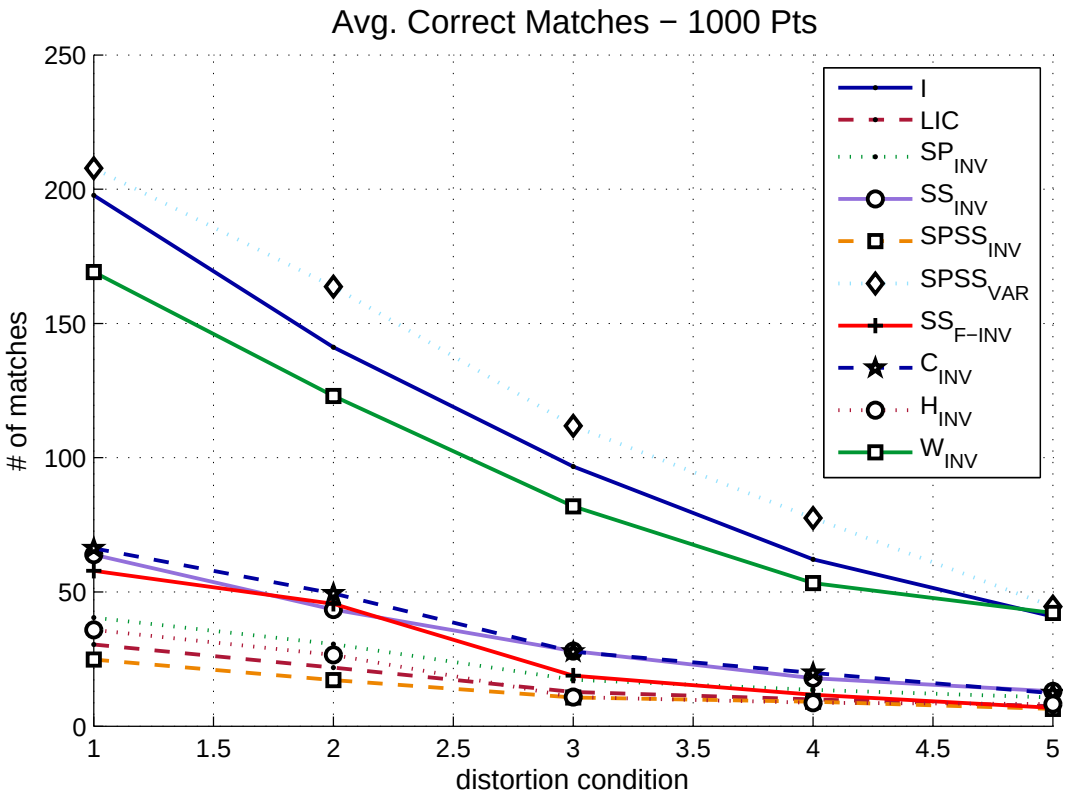
Table 4.2: Cumulative sum of the feature matching evaluation results.

Metric	I	LIC	SP_{INV}	SS_{INV}	$SPSS_{INV}$	$SPSS_{VAR}$	SS_{F-INV}	C_{INV}	H_{INV}	W_{INV}
matches	14,907	9,734	14,523	12,917	9,623	14,930	10,409	13,273	8,010	17,500
cumulative % m-score	13,404	13,814	15,004	14,190	13,388	14,108	13,640	14,849	12,196	14,576

Similarly to the table of detection results, Table 4.2 shows that the W_{INV} colour invariant is the best overall performer in terms of number of matches, followed by the specular-shadow-shading variant $SPSS_{VAR}$, and the grayscale intensity I. The description results however, differ from the detection results in that the matching score of the colour descriptors perform differently relatively to intensity, than their relative performance in the repeatability rates evaluation. The overall matching score results in Table 4.2, show that the intensity is in fact amongst the worst performers for generating distinct descriptors.

The descriptive quality of each invariant is represented by the matching score, which is a measure of how many of the extracted descriptors produce a correct match. The results here indicate that colour invariants are in general better descriptors than detectors. Intensity's relatively poor performance is an expected result even though the intensity is able to find sufficient numbers of gradients to detect (and thus obtain a high number of matches). However due to a lack of chromatic information, the intensity descriptors loose distinctiveness and obtain higher rates of mismatched features. Figures 4.9 and 4.10 present the summary of the results for the number of correct matches, showing the mean number of matches across all distortion levels for each of the four datasets.

(a) Oxford



(b) Middlebury

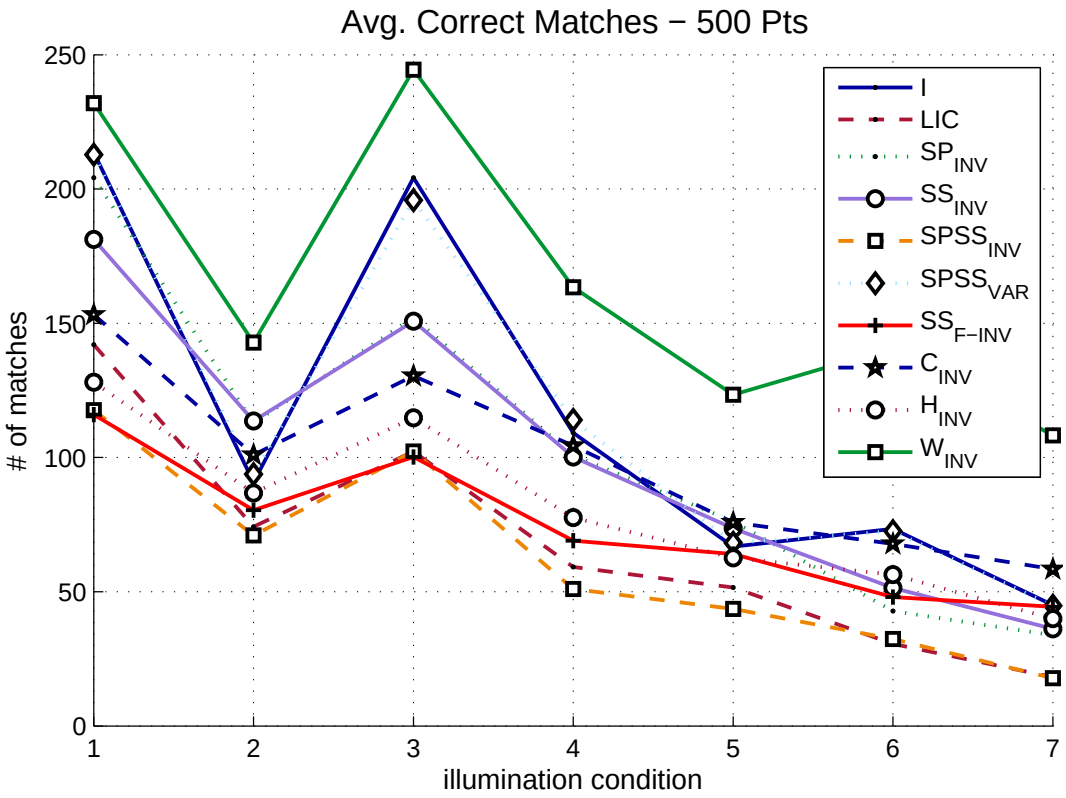
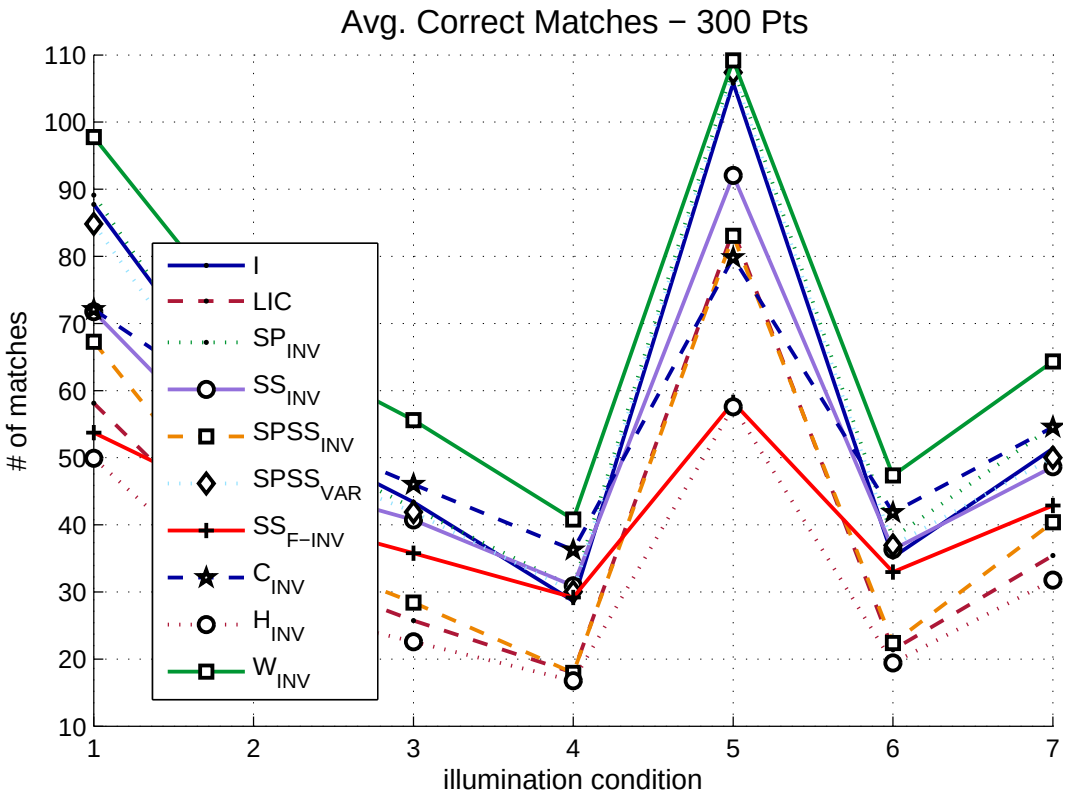


Figure 4.9: Summary of the number of correct feature matches for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

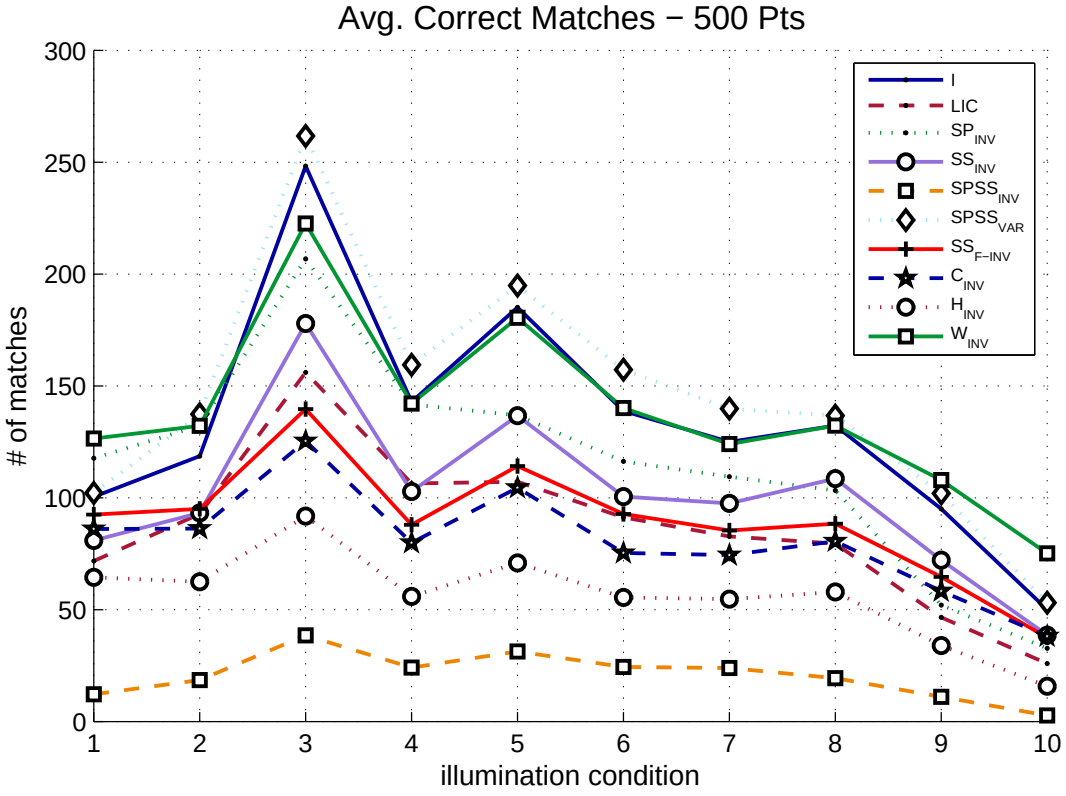


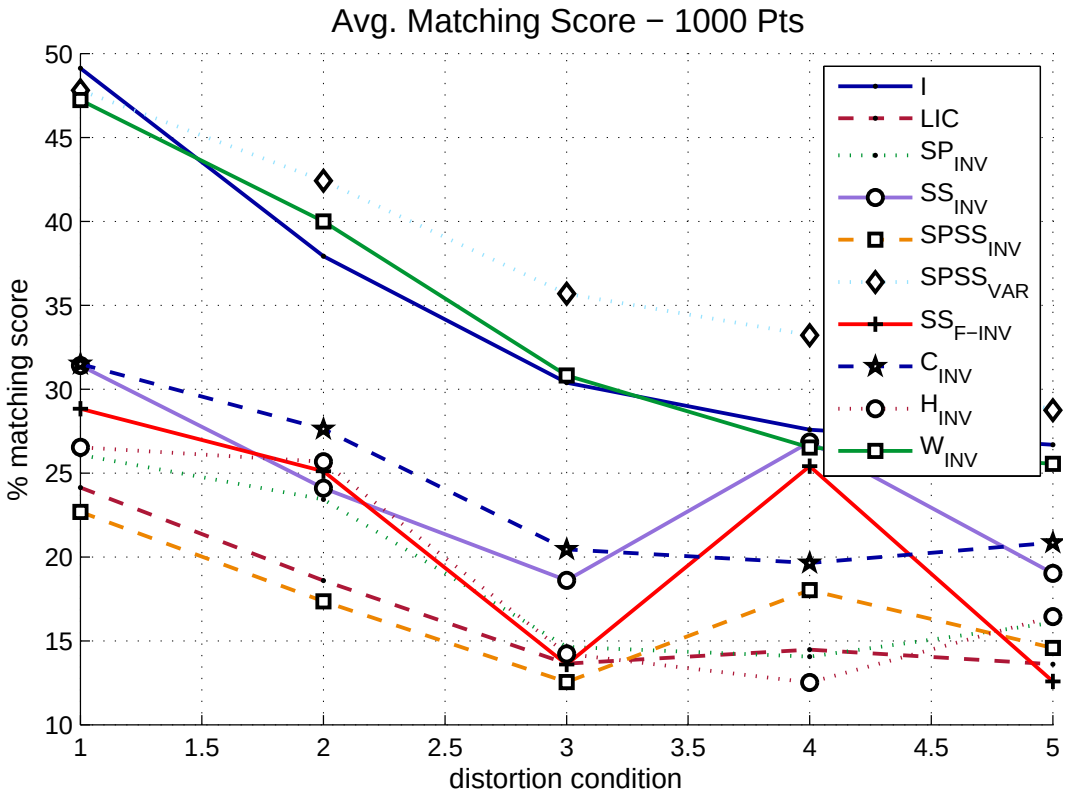
Figure 4.10: Summary of the number of correct feature matches for the ALOI (a) and PHOS (b) datasets.

Similarly to the detection, the results for the Middlebury, ALOI and PHOS datasets are not monotonically decreasing as in the Oxford dataset because the illumination varying image sequences are not organised to have increasing levels of the same distortion type. The Middlebury and PHOS image sequences are arranged from the darkest lighting to the most exposed lighting condition, although there is not a smooth transition of the same lighting effects from one image to the next (lighting direction and exposure vary independently). In the ALOI sequences, the first 5 images are obtained by sequentially changing the direction of one light source, and it is why results in Figure 4.10a decrease linearly for the first 4 matching results. The remaining 3 images are taken with multiple simultaneous light sources and thus the effects on the results are not linear.

The three best overall performers in terms of number of matches are W_{INV}, the grayscale intensity I and $SPSS_{VAR}$. The colour invariants are significantly inferior in the Oxford set compared to I and $SPSS_{VAR}$, except for W_{INV} which in fact performs comparatively to intensity and closes the margin as the distortions increase. A surprising result arises from the Oxford matching results in that $SPSS_{VAR}$ proves to be the best candidate, this is a significant finding as this gradient type has not been implemented as a local feature or evaluated previously in the literature.

For the illumination varying datasets of Middlebury, ALOI and PHOS, W_{INV} obtains the most number of feature matches overall. I and $SPSS_{VAR}$ however, perform comparably with the other colour invariants despite having limited photometric invariance. They perform well compared to the colour invariants as the number of matches heavily depend on the detection phase, essentially on being able to localise gradients of sufficient strength across different conditions. The worst performers are mainly those which contain the most level of invariance, such as $SPSS_{INV}$, H_{INV} and SS_{F-INV}. Too much invariance has decreased their distinctiveness by reducing the available colour gradients that are present in an image. The descriptive quality of each invariant is represented by the matching score results shown in Figures 4.11 and 4.12, and the precision-recall curves which are shown for completeness in Appendix B. The standard deviation of the matching results are also presented in Appendix B.

(a) Oxford



(b) Middlebury

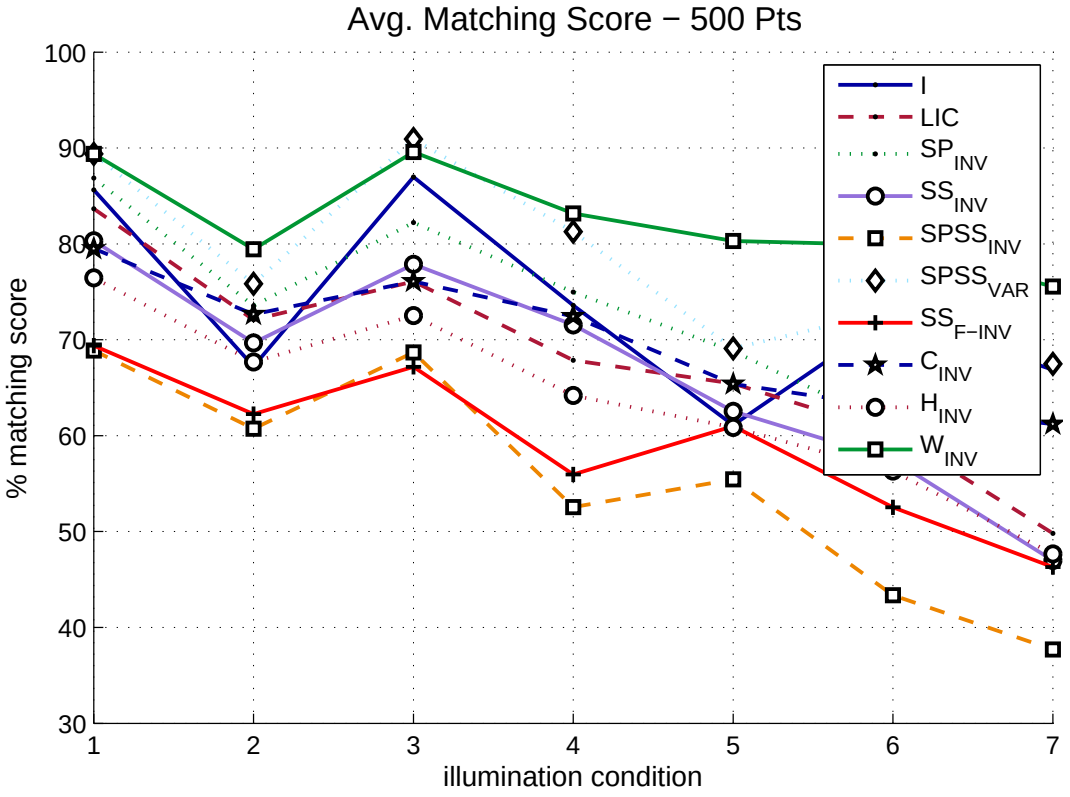
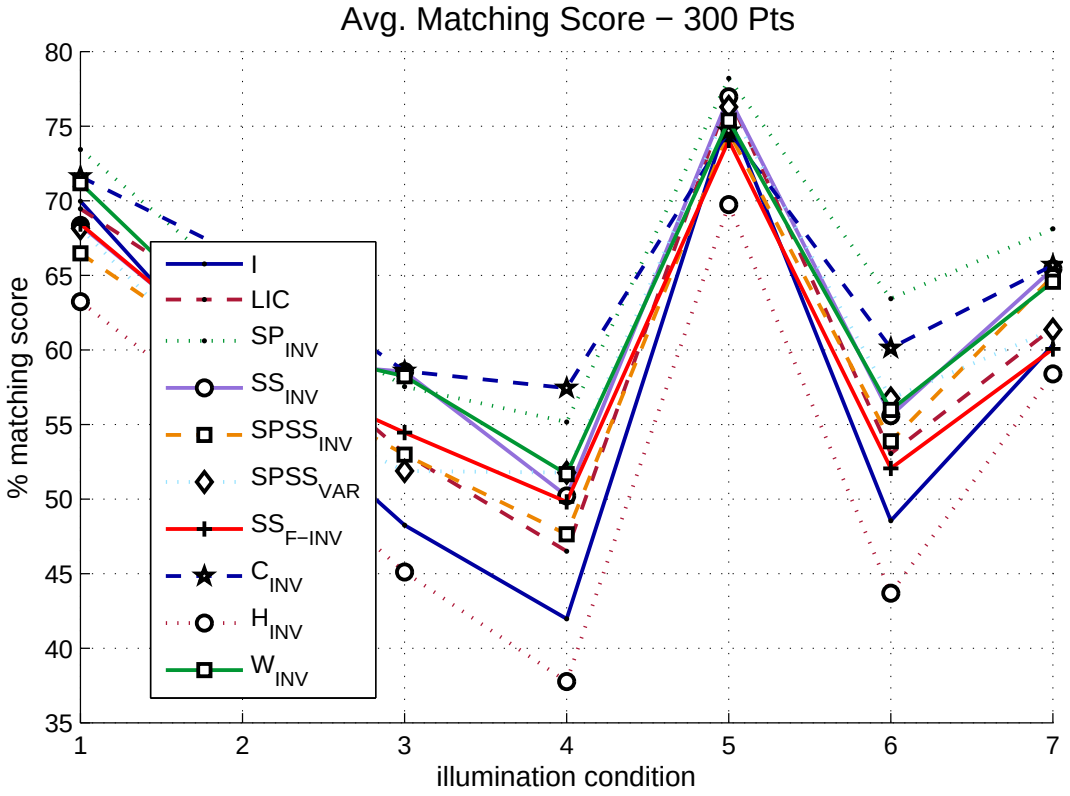


Figure 4.11: Summary of the matching score results for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

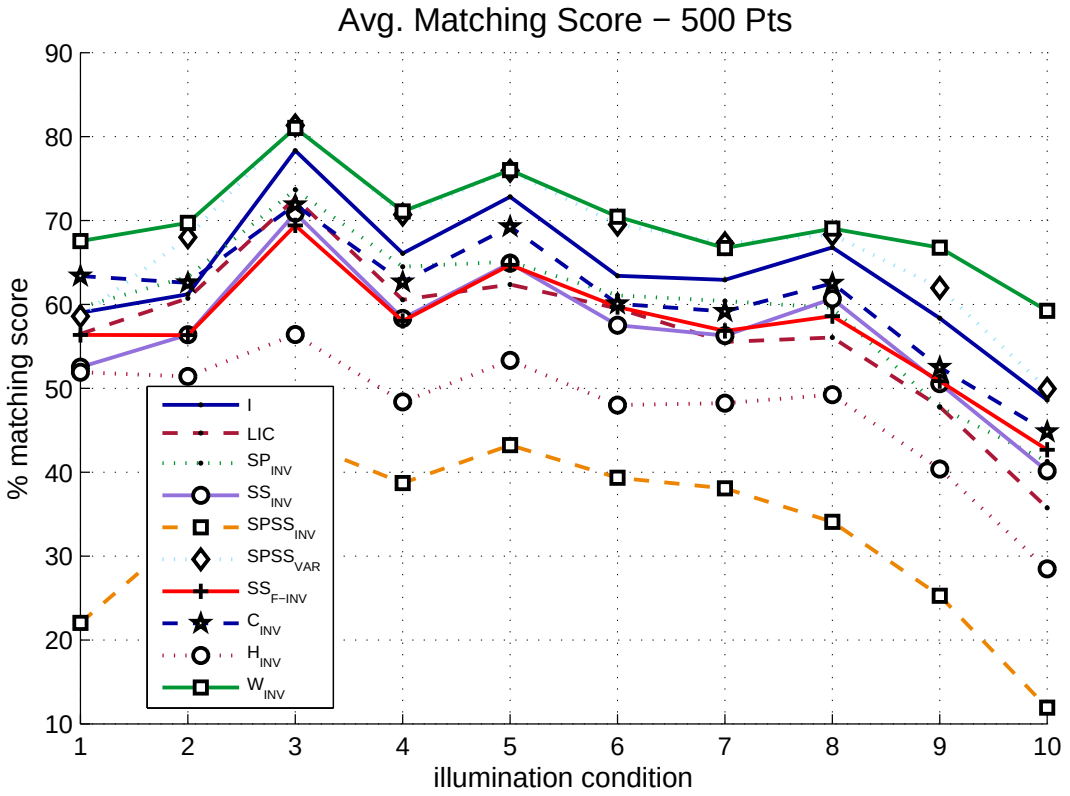


Figure 4.12: Summary of the matching score results for the ALOI (a) and PHOS (b) datasets.

The relative performance of the colour invariants with respect to intensity, is better for the matching score than for the repeatability rates of the detection study. This can be seen from the overall results in Table 4.2 and Figures 4.11 and 4.12. Intensity's relative matching score performance is worse than for the number of matches study, in the ALOI set it is in fact the second worst performer. $SPSS_{VAR}$ on the other hand fares better due to having the colour saturation component (refer to Table 3.2). Overall the matching score is high for all methods, and the difference in performance is smaller than in the number of correct matches study indicating most of the colour invariants have potential to be used as descriptors. The best three methods for the feature matching are I, $SPSS_{VAR}$ and W_{INV}, although other methods like SP_{INV} and C_{INV} are better suited for the ALOI dataset in terms of descriptor distinctiveness. This implies that for some colour invariants, it is best only to use them as descriptors in certain cases, either for only image recognition tasks or feature matching tasks where the features are detected with a different gradient, such as the grayscale intensity.

This concludes the local feature detection and matching evaluation of this research, which determined that grayscale features are the best candidate for general imaging conditions. As previously stated, the overall dominance of intensity-based features, can be largely attributed to the luminance axis containing the majority of the variation in the RGB-cube (Van de Weijer et al., 2006b), and the stability of the localisation of its gradients. For these reasons only under varying illumination conditions should colour gradients be considered for local feature matching. The overall performance of the majority of the tested colour invariants is insufficient to merit their independent adoption for general local feature matching tasks. However, they perform well as descriptors and in certain cases outperform grayscale features under illumination conditions, W_{INV} for example is generally always better than intensity in those cases. Due to colour invariants inherently containing different information than intensity, a study was required to investigate if colour could be utilised to enhance the intensity for feature extraction tasks. Thus in this way discover if a feature fusion extraction was possible, in which the most appropriate grayscale and colour gradients are utilised conjointly. The next section presents the feature fusion study that was carried out in this research for feature matching applications.

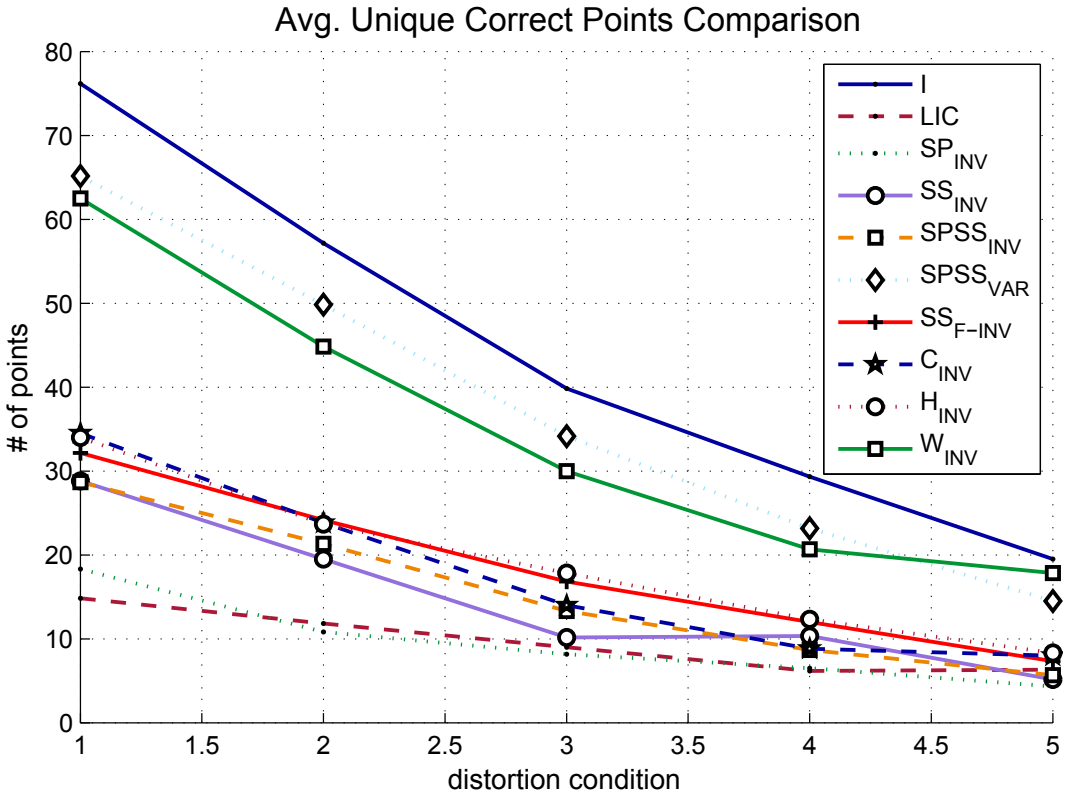
4.4 Feature Fusion for Image Feature Matching

This section is organised as follows: Section 4.4.1 presents the feature correlation study that was performed to investigate the number of unique features that each gradient can extract; which have the potential to be used in a feature fusion extraction approach. Section 4.4.2 outlines the proposed feature fusion strategies for local feature detection and the results of their evaluation. In Section 4.4.3, the ranking metric for obtaining the strongest HL corner points are examined, and their suitability to be used in an optimum fusion strategy is discussed.

4.4.1 Uniqueness and Correlation Analysis

The correlation of the gradients are here experimentally obtained by performing a feature detection experiment similar to the one presented in Section 4.2. The difference here is that the features from each detector type are also compared to all the other 9 gradient types in order to obtain the number of unique correct correspondences at each imaging condition of the image-sets. The approach is carried out as follows: The HL points from the first image of each image-set are compared in turn to the points from the same gradient type of all subsequent images of the set (at varying imaging conditions). The correct corresponding points for each imaging condition of the set, are then compared to the points from all other gradient types (at the same imaging condition). The same region overlap area error threshold of 90% is used here to determine a correct feature correspondence, as was done in the detection experiments of Section 4.2. After this comparison to all other gradient types, the HL points that do not find correspondences are thus unique. The number of HL points that correspond to other gradient types are used to calculate the level of correlation between them. The correlation between a gradient type A and B is calculated here by the percentage of correctly matched points that are common between them, from the lowest number of total correct points obtained from type A or B. The unique correspondence results are shown in Figures 4.13 and 4.14, which show the average number of unique correct points for all 4 datasets.

(a) Oxford



(b) Middlebury

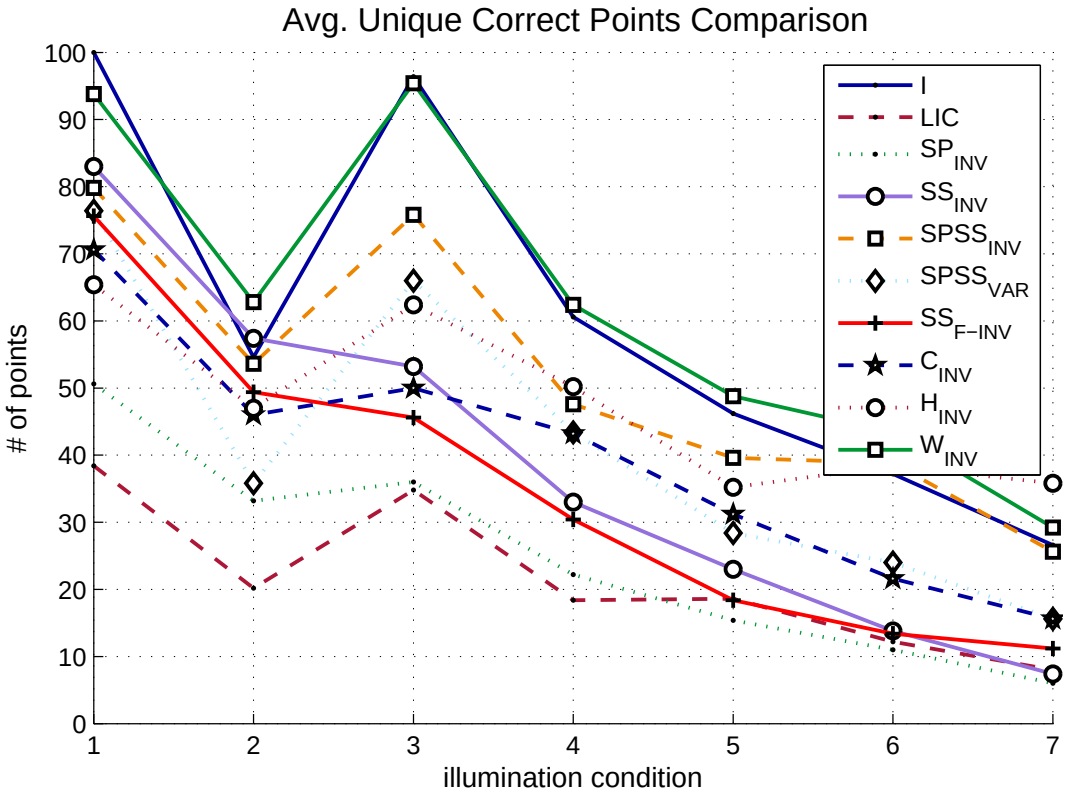
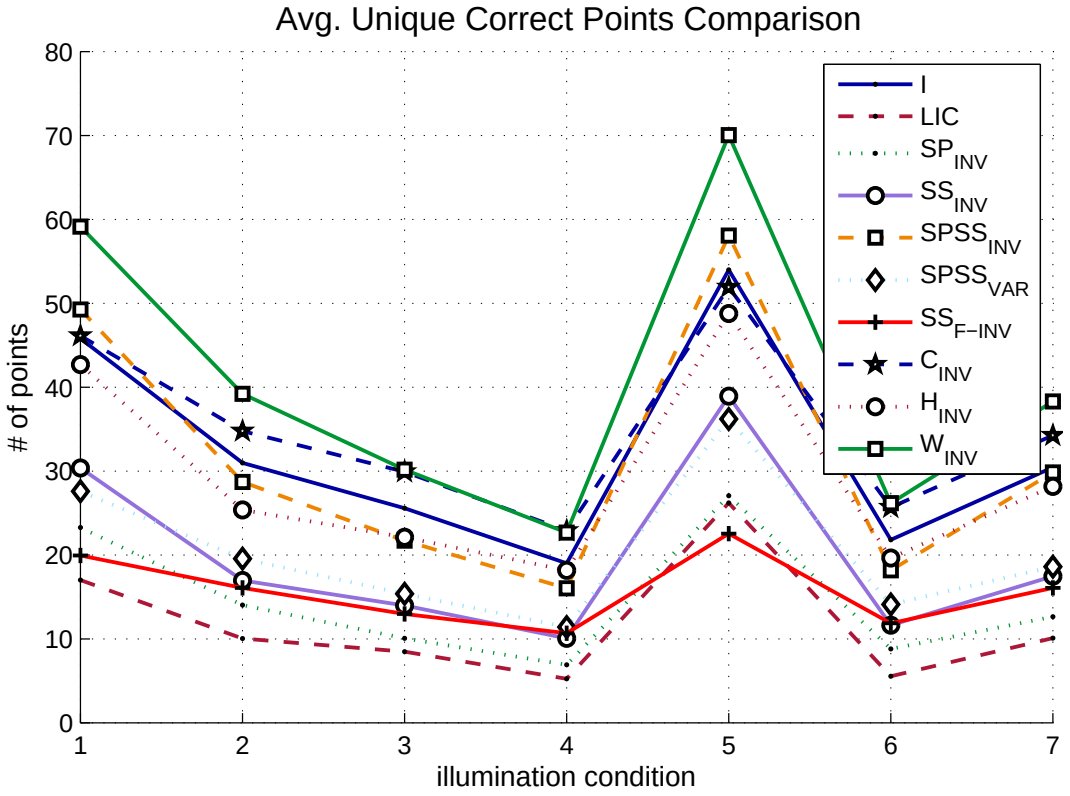


Figure 4.13: Summary of the unique correspondences analysis for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

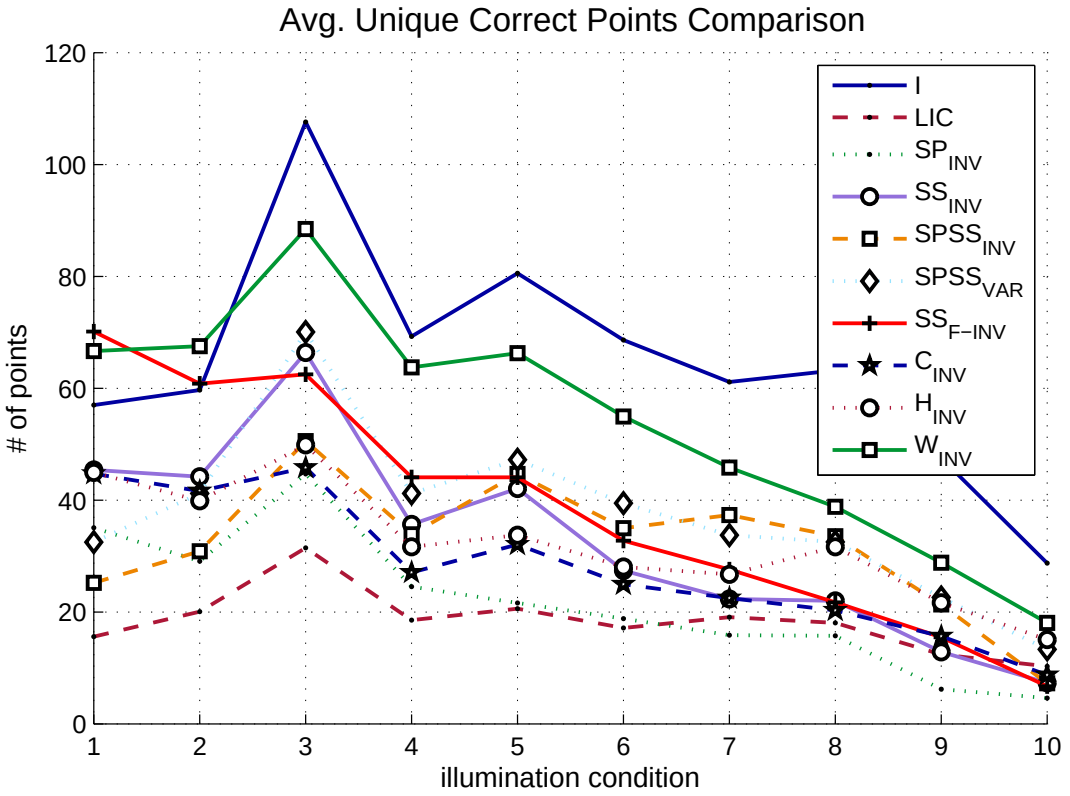
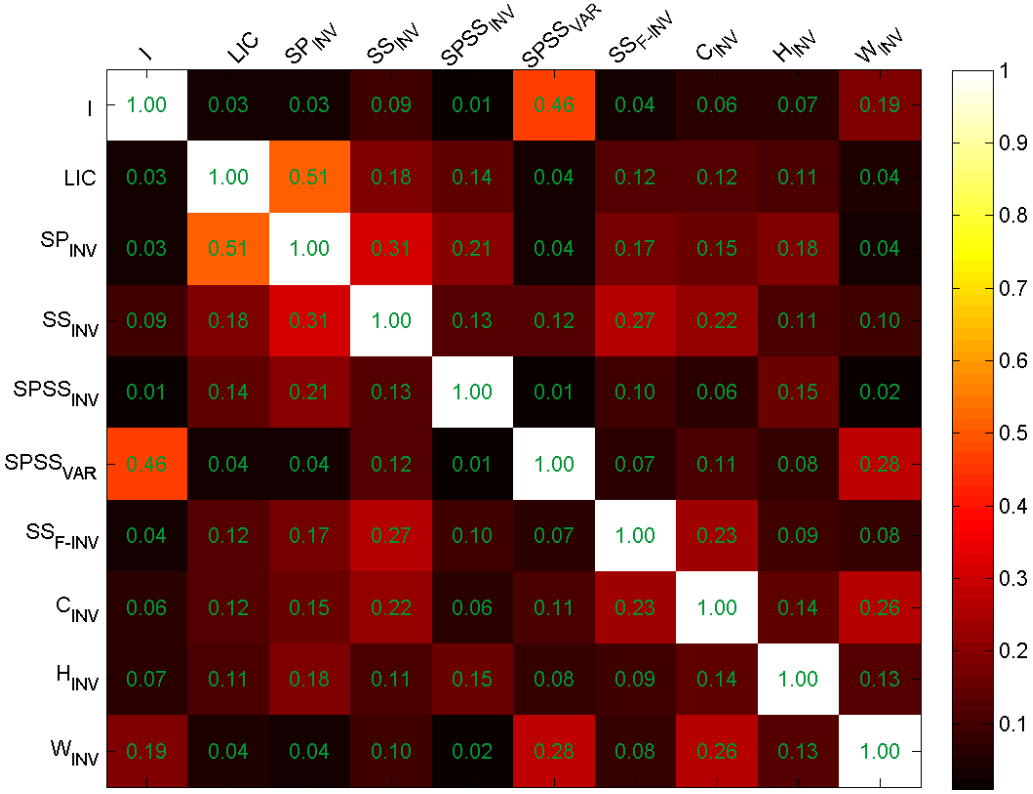


Figure 4.14: Summary of the unique correspondences analysis for the ALOI (a) and PHOS (b) datasets.

The average correlation matrix plots are presented in Figures 4.15 and 4.16. The standard deviation results of the number of unique points are shown in Appendix C. The results presented in Figures 4.13 and 4.14 prove that there is a substantial amount of useful information inherent in the colour invariants that are not shared by the grayscale intensity gradients. This can be deduced by the significant number of unique correct correspondences obtained by the colour gradients. The standard Oxford detection results of Figure 4.5a show that the intensity obtains 290 correct correspondences in the first distortion level and 80 for the last distortion. By analysing the unique correspondences of Figure 4.13a, it can be seen that the total number of unique correct points from all gradients types in the first distortion level amounts to 394, when adding to this the number of non-unique correct grayscale intensity points ($290\ \mathit{total} - 76\ \mathit{unique} = 214$), it results in a total of at least 608 potential correct correspondences if all 10 gradient types are utilised together in a fusion approach. Such an ideal fusion thus results in a 109% improvement upon the grayscale intensity's performance on the first distortion level of the Oxford image-sets. In the last distortion level, a fusion approach has the potential for improvement of 97%. When analysing the Middlebury unique correspondence results of Figure 4.13b and comparing them to the standard detection results of Figure 4.5b, the potential improvement of a feature fusion extraction is estimated to be 252% in the first distortion level, and 240% in the most severe illumination condition.

Results from all four datasets clearly indicate that the capacity to incorporate colour features alongside grayscale intensity for successful feature detection is thus significant. This implies that an appropriate feature extraction fusion approach could potentially select from an image a set of detected features that are more robust to distortions, more numerous and more unique than utilising only intensity information. The other set of results that back the previous conclusion, is the correlation matrices presented in Figures 4.15 and 4.16. The correlation results indicate that the colour invariants are highly uncorrelated to the intensity and are thus capable of positively influencing the performance of intensity if used conjointly. For each image, the correlation between two gradient types is calculated here as the percentage of the number of correct points that are common among the two detectors, out of the minimum number of total points extracted between the two detectors.

(a) Oxford



(b) Middlebury

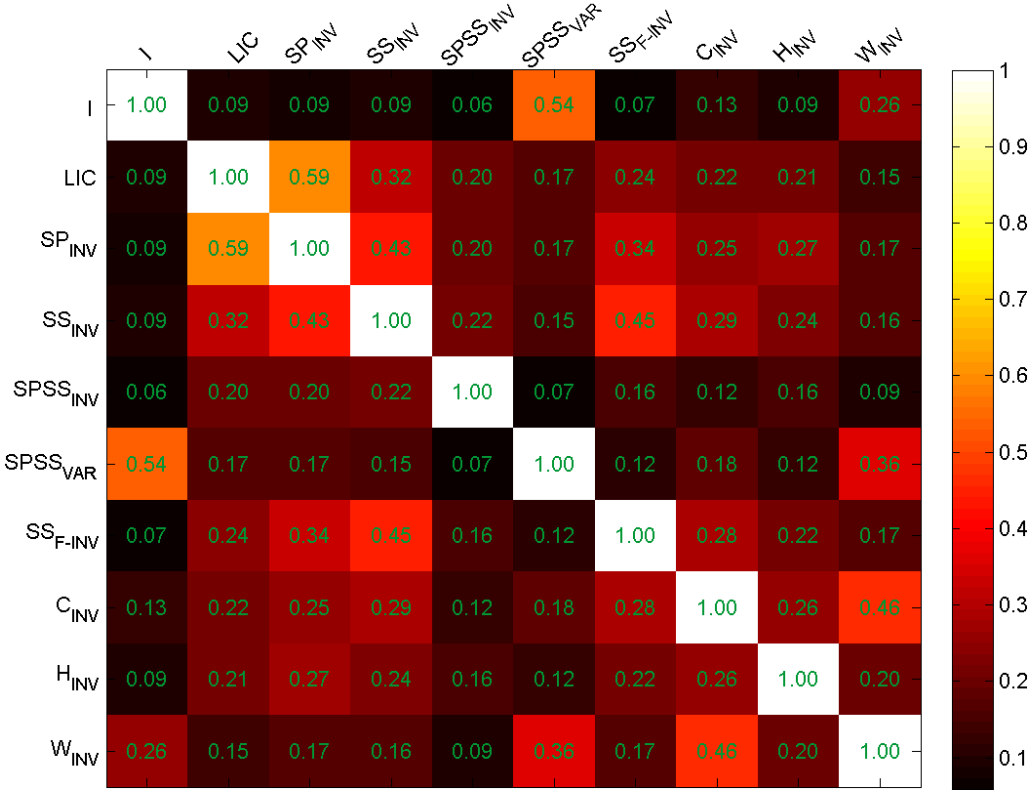
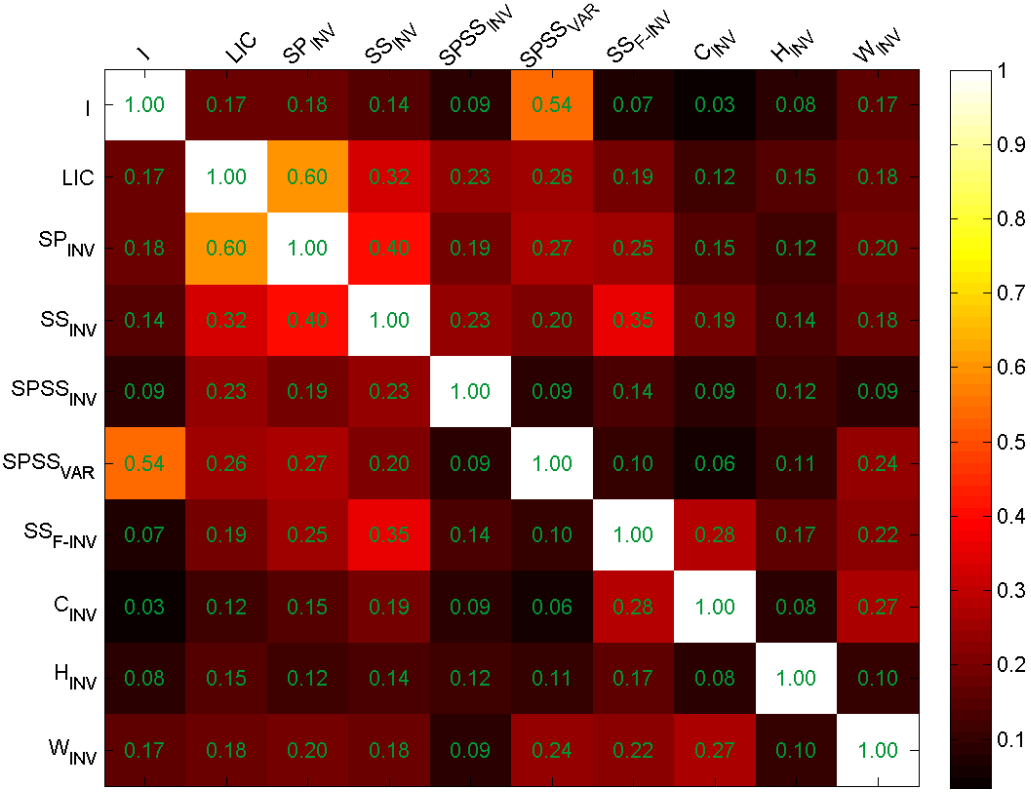


Figure 4.15: Summary of the correlation analysis for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

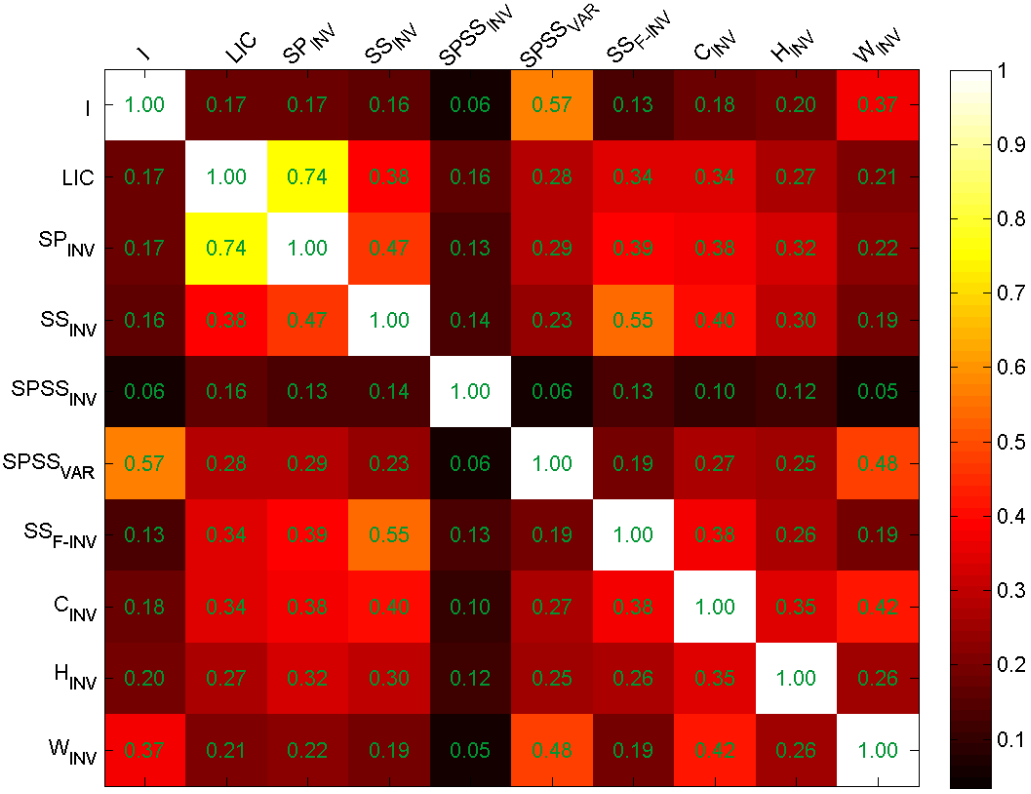


Figure 4.16: Summary of the correlation analysis for the ALOI (a) and PHOS (b) datasets.

That is why the correlation matrices appear symmetrical. The variant $SPSS_{VAR}$ is the most strongly correlated since half of its gradient composition contains an intensity component. W_{INV} is the second most correlated to intensity, which is expected as it performed the best out of all the colour invariants. While the exact correlations differ according to the specific dataset, overall the same relative correlations are apparent throughout all the results.

4.4.2 Fusion Strategies and Results

Two feature fusion techniques are proposed in this section and evaluated in a local feature detection experiment. The two techniques are based on selecting HL points from different gradient types based on the Harris cornerness energy strengths of the points (Equation 3.2). Traditionally in the literature, HL points have been selected by setting a threshold value to the Harris energies of the available points. In this research a fixed number of points from each image is extracted via a ranking of the strength of the Harris energy. In order to rank using this metric, certain normalisations must be performed before any fusion can take place. The different gradient types are scaled appropriately when extracting HL points from each gradient type. The magnitude of the highest response from each type then falls in range with all the others. This ensures that the Harris energies from the colour invariants are not weaker than the energies from the grayscale intensity and are therefore able to have an equal opportunity of being selected during the ranking.

The scaling factors for the gradients used to generate both the Harris and LoG image stacks, are obtained for each individual image using information from all the gradient types. For each gradient type the scaling process finds the highest LoG response and first derivative magnitude across all the scale space, then an individual scaling factor is given to each gradient type so that the maximum gradient magnitude achievable for each type is the same as all the other gradient types. Figure 4.17 shows the comparison of histograms of the scaled Harris energies from all gradient types for one image of the *Art* image-set.

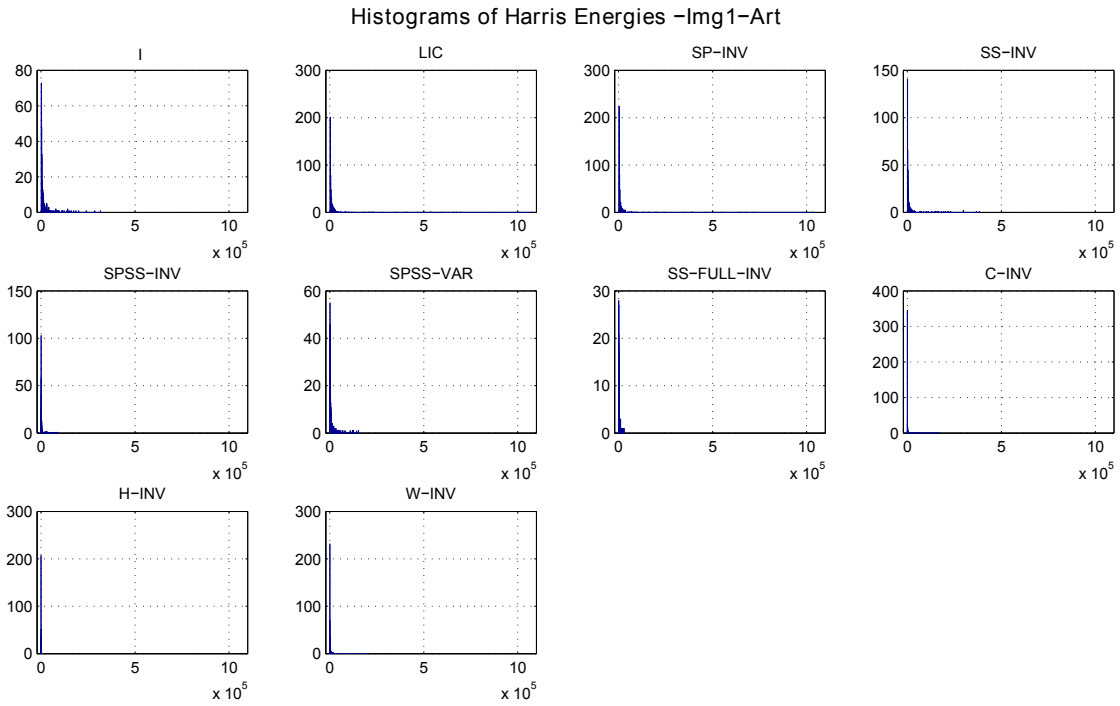


Figure 4.17: Histograms of the scaled Harris cornerness energy strengths of each gradient type.

However, these energies cannot be used together easily in a fusion technique as their values are non-linearly distributed and a sufficient overlap in the energies of the various gradient types is not guaranteed. Little detail can be distinguished in the plots of Figure 4.17 because a minority of the points have significantly greater energy values than the rest and their bin count it too low to be visible. In order to distribute the original Harris energies (H) more evenly to improve the fusion, the Harris energy distributions are stretched by taking their natural logarithm as shown in Equation 4.1.

$$H^s = \ln(H) \tag{4.1}$$

Figure 4.18 shows the effect of performing this operation. The stretching to H^s makes the energy distributions between the gradient types more compatible, which facilitates that all gradient types contribute to the fusion. The first fusion technique that is proposed and presented here, is named *Max H.E Fusion* and involves maximising the Harris energy of all image locations by combining multiple gradient types. The structure tensor shown in Equation 3.1 is composed of L_x, L_y and L_xL_y gradients components. *Max H.E Fusion* carries out

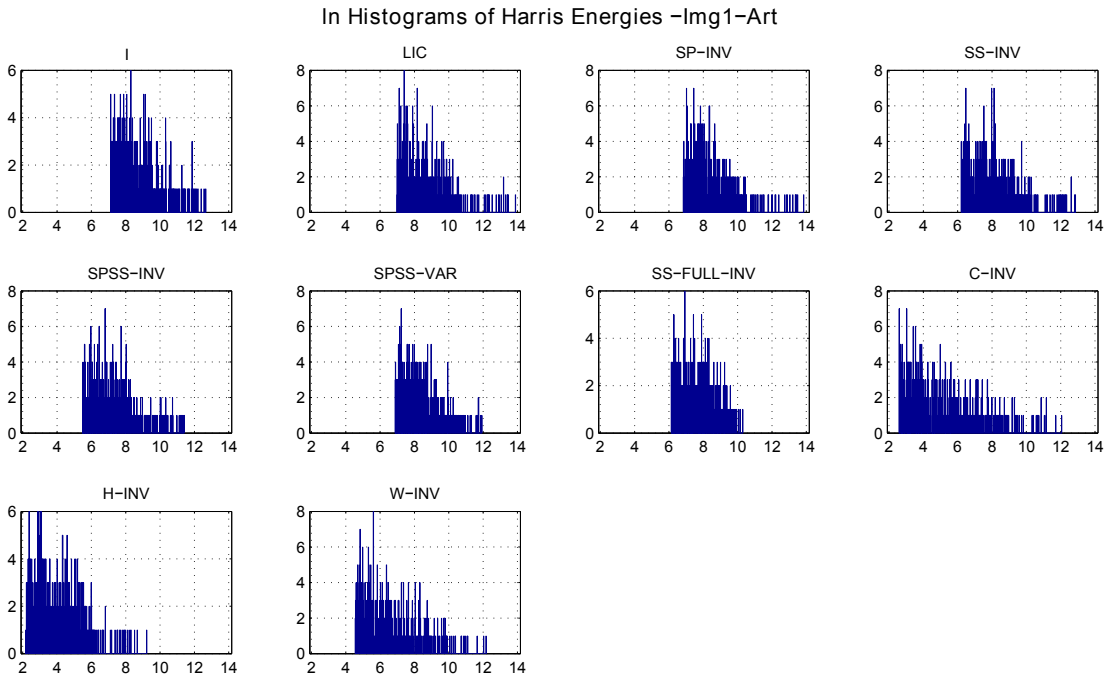


Figure 4.18: Histograms of the stretched Harris cornerness energy strengths, obtained by applying the natural logarithm of the original energy distributions.

permutations of Equation 3.1 with all possible combinations from different grayscale and colour gradients. The highest Harris energy that results from the permutations, is then assigned to that image location in the Harris energy stack. The reasoning for proposing this technique is that by utilising multiple gradients simultaneously (which vary in their strengths according to the imaging conditions) and optimising for a suitable metric, it increases the probability that under varying imaging conditions there will be a combination of gradients that ensures a strong response for the optimised metric (in this case the Harris energy). The second fusion technique is *H.E Ranked Fusion*, which pools together individually extracted HL points from different gradient types into a list of candidate points. The technique then ranks the points based on the Harris energy and selects the top N points, which could have been extracted arbitrarily from different gradient types. The evaluation results of the fusion techniques presented in Figure 4.19, were carried out on the Middlebury dataset. *Max H.E Fusion* $-$ *AFJ* is obtained by utilising the top three gradient types ($A = I, F = SPSS_{VAR}$ and $J = W_{INV}$), *H.E Ranked Fusion* $-$ *AFJ* selects the top HL points only from those three types, and *H.E Ranked Fusion* $-$ *ALL* selects the top 500 points from all of the 10 gradient types. In the other fusion types, the C denotes the invariant SP_{INV} and D refers to SS_{INV}.

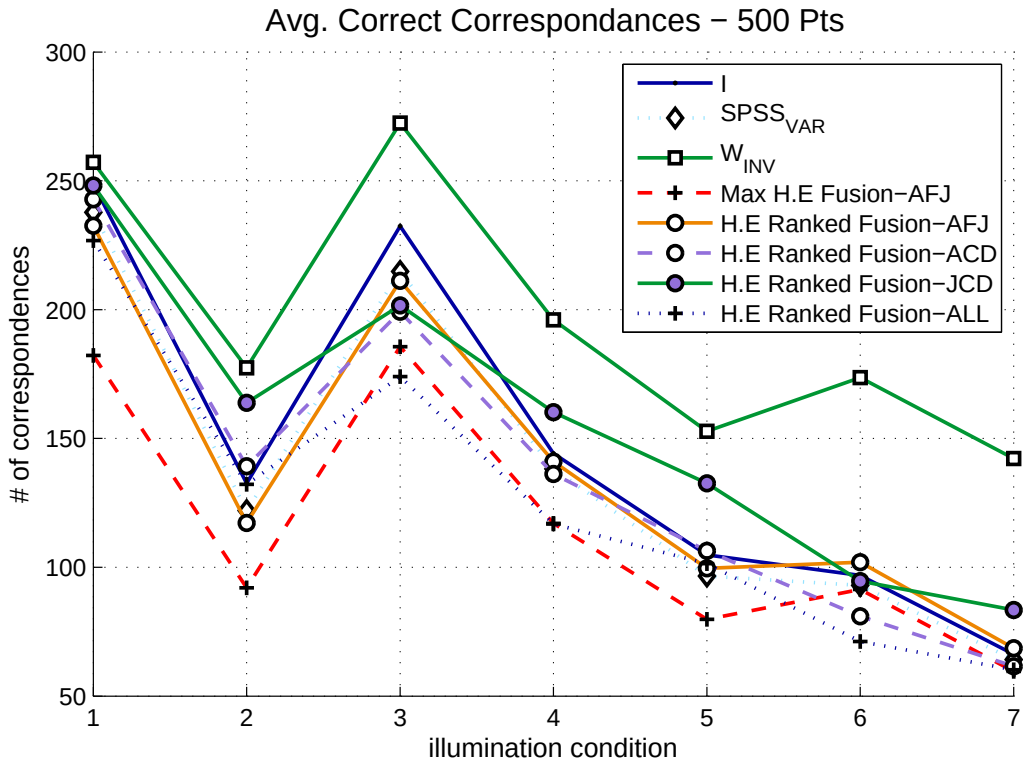


Figure 4.19: Comparison of the correspondences results of the proposed fusion techniques.

The fusion technique with the best performance is *H.E Ranked Fusion − JCD*. When considering all 10 gradients in the fusion of *H.E Ranked Fusion − ALL*, the overall performance deteriorates compared to when only three gradient types are considered. *Max H.E Fusion* is arguably the worst of the fusion techniques, as it performs particularly poorly on the first few distortion levels of the image-sets.

4.4.3 Analysing the Harris Energy as a Metric for Fusion

In order to investigate the reasons why the previously presented feature detection fusion techniques are not successful, a detailed study was performed on the effect that ranking HL points by the Harris energy has on the repeatability results of the detection process. This section outlines this investigation and the obtained results. From the fusion results in Figure 4.19, it can be deduced that the method *Max H.E Fusion* is not able to accurately capture the underlining corner structures of an image as it cannot locate the same corner locations of

the image scene across varying illumination conditions. This could be due to noisy L_x or L_y gradients, that when utilised for only one type of HL point extraction it reduces the probability of a strong corner being detected since both directional gradients have to be compatible to form a corner. However when a corner is formed by fusion from multiple gradient types, some noisy gradients from different gradient types could combine to generate a false reading for the presence of a strong corner.

In the case of the technique *H.E Ranked Fusion*, using 3 gradient types performed better than when the selection could choose amongst all the types. This result indicates that high corner energy HL points from the worst performing colour invariants were not positively contributing to the selection of the most optimum set of HL points for an image. The other observation that AFJ and ACD performed very similarly to I and $SPSS_{VAR}$, indicates that the colour points seem to be largely overshadowed in the fusion process. The analysis of the previous fusion study thus concludes that the Harris corner energies may not be the most appropriate metric to rank and select the best set of HL points amongst all the possible grayscale and colour candidates. The investigation in this section aims to prove if the aforementioned hypothesis is correct.

The HL ranking evaluation experiments were conducted on the Middlebury dataset for each of the 10 gradient types. The 500 HL points of each type, were separated into subsets in terms of their Harris corner strengths, after first stretching the distributions as outlined in the previous section. The points were organised into percentile subsets according to the relative strength of the Harris energies, with respect to the maximum value of the 500 points. The subsets of HL points were evaluated in a feature detection experiment, and the results of the repeatability rates are presented in Figures 4.20, 4.21 and 4.22. In the plots, the subset 90-100% for example, contains the HL points which have energy strengths ranging from 90% to 100% of the maximum value and thus are supposed to be the strongest corners of the 500 total points. The evaluation analyses the repeatability rates, in order to verify the robustness of the detection and the probability that the points will produce correct correspondences. Results indicate that the subset of points performs relatively differently across each gradient type.

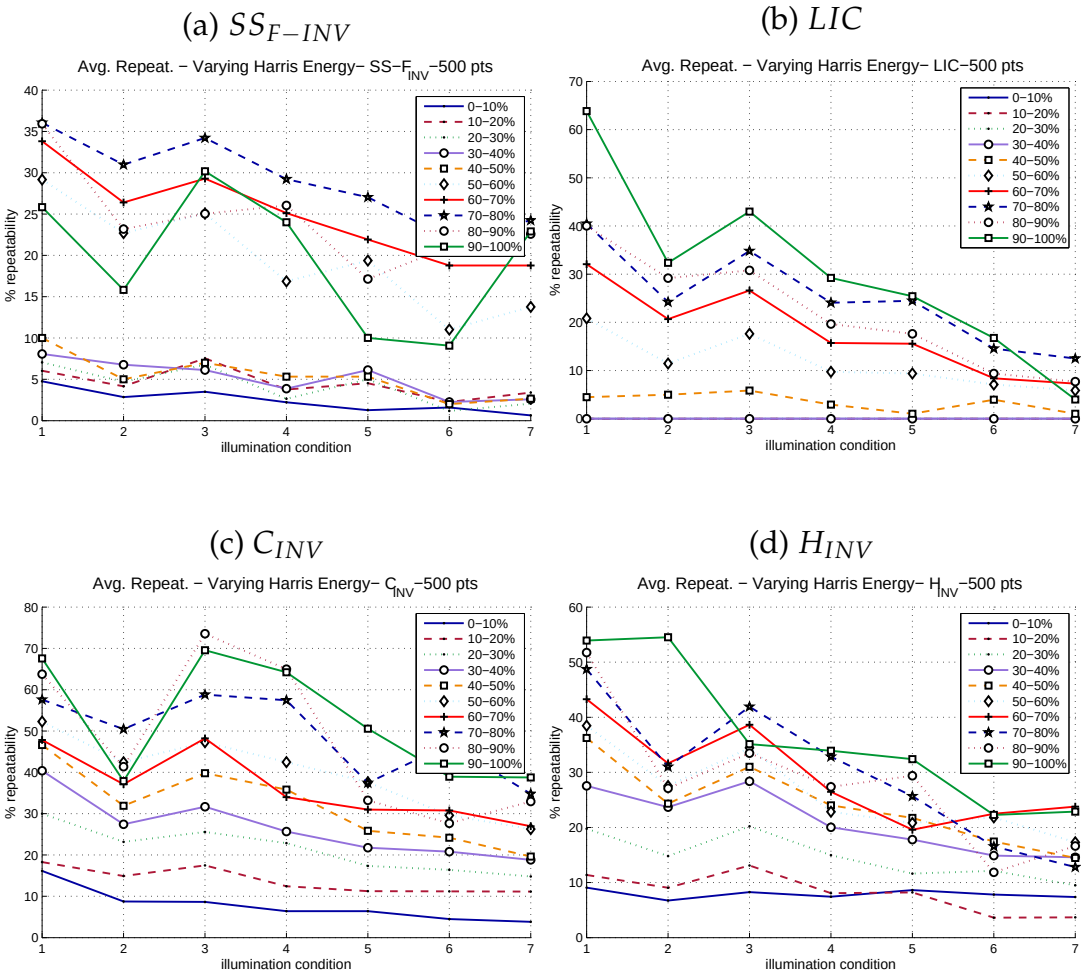
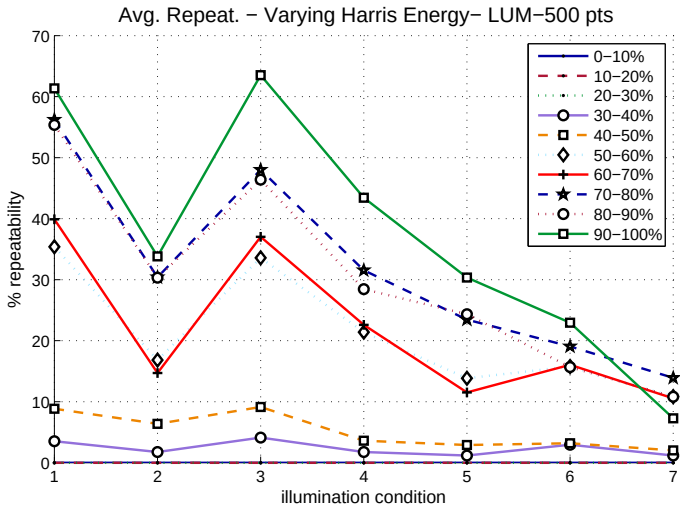
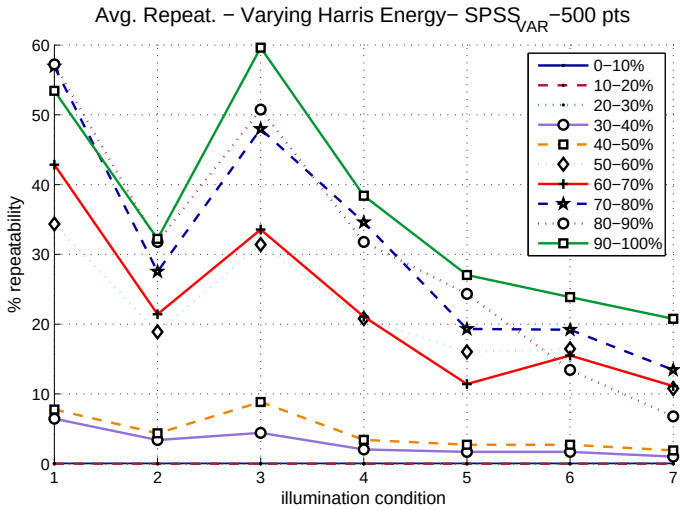


Figure 4.20: Detection repeatability of HL points of varying Harris energy ranges: (a) SS_{F-INV}, (b) LIC, (c) C_{INV} and (d) H_{INV} .

(a) I



(b) $SPSS_{VAR}$



(c) W_{INV}

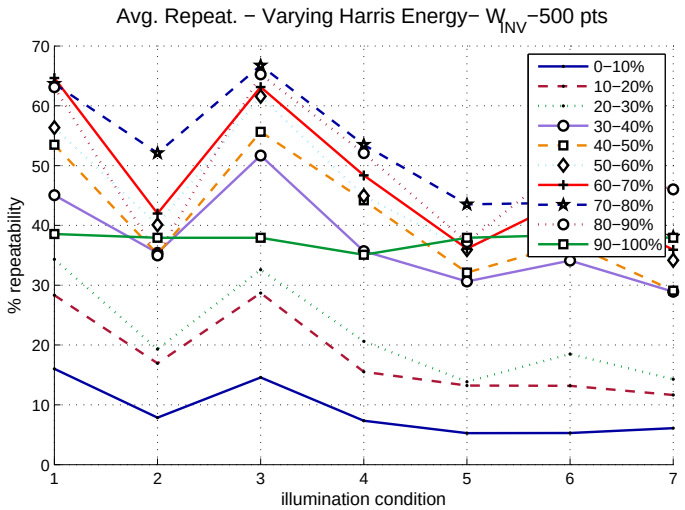
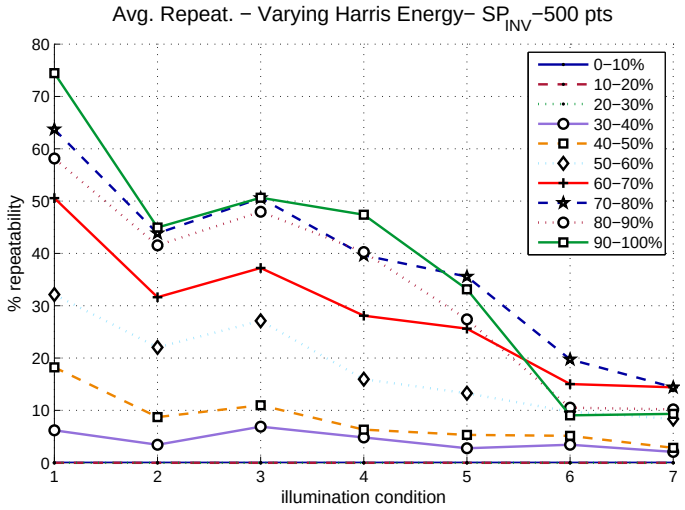
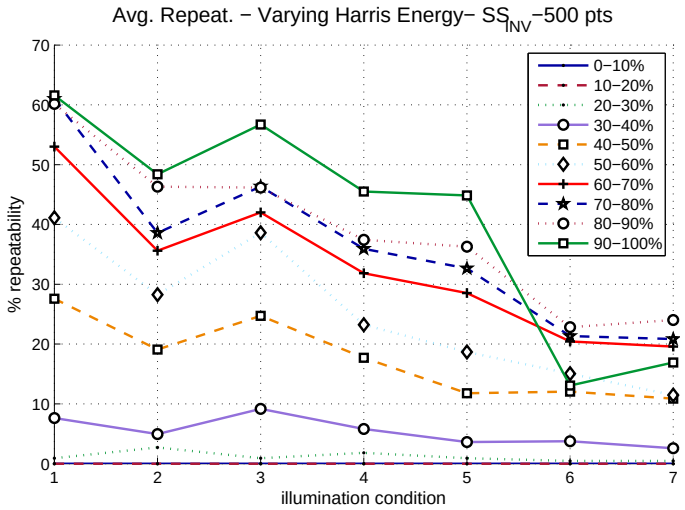


Figure 4.21: Detection repeatability of HL points of varying Harris energy ranges: (a) Grayscale luminance intensity, (b) $SPSS_{VAR}$ and (c) W_{INV}.

(a) SP_{INV}



(b) SS_{INV}



(c) $SPSS_{INV}$

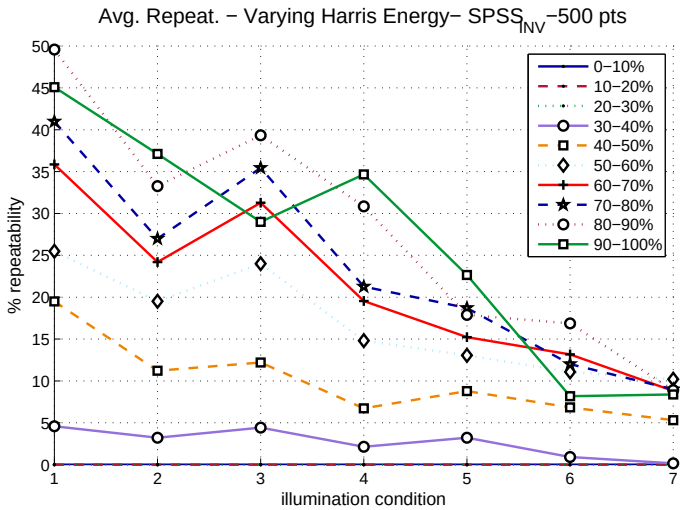


Figure 4.22: Detection repeatability of HL points of varying Harris energy ranges: (a) SP_{INV}, (b) SS_{INV} and (c) $SPSS_{INV}$.

The general trend is for the lowest 30% range of energies to obtain significantly lower repeatability rates, and for the top 30% to obtain the highest. This trend does not apply however in the case of W_{INV}, which is particularly problematic since it is the best performer from the colour invariants. This explains why $H.E\ Ranked\ Fusion - AFJ$ in Figure 4.19 performs similarly to the intensity, as few of the W_{INV} are actually selected. The reason for why the other $H.E\ Ranked\ Fusion$ types do not manage to consistently outperform the intensity, is that the Harris energy ranking evaluation shows that there is not a substantial difference in repeatability rates amongst the points selected in the top 30-40% range of energies. The gradient type that arguably performs closer to the ideal scenario is the grayscale intensity, but the probability of selecting a correct point in the 90-100% range is only higher than 50% in 2 of the 7 different imaging conditions. The intensity energy ranges of 70-90% obtain a probability of selecting a correct point higher than 50% only once. Therefore even for the grayscale intensity, the ranking via the Harris cornerness energy does not indicate a strong probability of selecting points that will be repeatable at the same scene location across varying imaging conditions. Other experiments that have been omitted from this thesis, performed the same analysis utilising the LoG response as the ranking metric for the HL points, and obtained similarly negative results. The study outlined in this section thus concludes that in order for a successful extraction fusion technique for local feature detection and matching tasks, a more appropriate ranking metric other than the Harris energy or LoG response must be found. Some possible further research directions for the ranking of the points are outlined in the future work section of 6.2.

4.5 Summary and Discussion

This chapter presented an evaluation of the performance of colour photometric invariants in the context of local feature detection and matching. This evaluation obtains more conclusive results on the performance of colour invariants for the aforementioned applications, when compared to previous studies in the literature. The main reasons for this are due to the utilisation of the more appropriate evaluation framework developed by Mikolajczyk and Schmid (2005), the testing on multiple datasets, and because the evaluation includes colour invariants that have never been implemented as local features before in the liter-

ature. Additionally, the testing was done on datasets containing the typical set of imaging distortions encountered in real-world applications. Thus this work has evaluated the colour invariants with the same level of rigour that state of the art grayscale intensity-based features have been evaluated with in the literature. Furthermore, this evaluation has implemented all features using the same code base and tested them within the same framework, ensuring an accurate comparison that provides more certainty than previous studies, on what role colour invariants have in the performance of local feature detection and matching tasks.

The detection results of Section 4.2, indicate that there are only three gradient types that consistently perform robustly across all imaging distortions (I, $SPSS_{VAR}$ and W_{INV}). In the Oxford dataset which allows the testing of all the distortion types, the gradients of the grayscale intensity and specular-shadow-shading variant $SPSS_{VAR}$ performed the best, followed by the colour invariant W_{INV} with an approximately 10% drop in the number of correspondences. The same colour invariant on the other hand, outperforms all other feature types in two of the three other illumination varying datasets used in the evaluation. All colour gradients apart from W_{INV}, perform poorly for general types of imaging conditions and it is thus recommended here to not use them individually for feature detection tasks. Despite not outperforming in all of the illumination varying tests, the grayscale intensity and $SPSS_{VAR}$ nonetheless perform adequately and at times comparatively with with the best colour counterpart. The results therefore validate why grayscale is generally preferred in the literature for local feature detection, and only under varying illumination conditions should the colour gradient W_{INV} be considered.

There is a limitation of the evaluation that must be mentioned, in terms of the extent of the generalisation that can be inferred from the feature matching results. This observation arises from the results of the HL optimisation study presented in Section 3.4. It was shown that although there was not a substantial difference amongst the top performers, the results indicated that the optimisation was dataset-dependant. Additionally, the optimisation was done using only luminance gradients. Therefore, it is possible to obtain a different set of optimum parameters for each gradient type on each of the evaluating datasets. Different parameters could potentially change the relative performance of the

gradients compared to what is presented in this thesis (in both detection and matching results). The purpose of the evaluation however is not to compare each optimised colour gradient with each other, it is instead to evaluate and compare them under equal conditions with an appropriate feature extraction algorithm. Furthermore, the optimisation study used the luminance as it represents the highest variability in the image data. Overall, the low variations in performance seen in the optimisation results of Section 3.4 and the consistently overall performance of the luminance, $SPSS_{VAR}$ and W_{INV} in the results of this chapter, convey confidence in that they would remain the top performers even if another set of HL parameters would be used. However it must be pointed out that the results are indeed biased but the impact of this inherent bias is unknown. Results are biased firstly towards the luminance, and secondly towards the Oxford and Middlebury datasets since only they were used for selecting the optimal HL parameters. The results for the PHOS and ALOI datasets, can therefore be regarded as a better representation of a general scenario application that does not utilise any data priors.

Section 4.3 presented the feature matching evaluation, which combines both detection and description matching and represents the most common real-world application for local image features. The matching results are thus more significant than the detection results and provide a greater indication for the usability of colour invariants. A surprising result arises from the Oxford matching results in that $SPSS_{VAR}$ proves to be the best candidate, this is an important finding as this overlooked gradient type has never been implemented as a local feature or evaluated previously in the literature. The best colour invariant in the Oxford results is W_{INV}, which performed comparatively to the intensity with a 13% drop in number of matches in the first distortion level and obtaining the same number of matches in the last distortion. For the illumination varying datasets W_{INV} again obtains the best results. It can thus be concluded that it has the best balance of invariance amongst all other colour types, in contrast to H_{INV} which contains too much invariance and fails to robustly detect enough gradients as can be seen in the visual examples of Figures 3.15 and 3.17.

Overall, $SPSS_{VAR}$ obtained only a slightly higher number of matches than intensity, but a substantially higher matching score. The matching score of grayscale intensity was in fact the second worst from all the 10 gradient types, confirming quantitatively that the conversion from colour to grayscale looses information and lessens the distinctiveness of the image data. Comparing accurately the results of this evaluation with the literature is not possible, as the evaluation frameworks and tested invariants generally vary. Burghouts and Geusebroek (2009) implement their own evaluation framework, but in their precision-recall descriptor matching results the W_{INV} was amongst the top performers and therefore shares a similarity with this work. The study of (Van De Sande et al., 2010) evaluates SIFT descriptors with a variant of C_{INV} and Abdel-Hakim and Farag (2006) uses only a variant of H_{INV}, they therefore both ignore W_{INV} and also use the framework of Burghouts and Geusebroek (2009).

In the work of Gossow et al. (2010), different variations of the C and W invariants are used simultaneously for SURF feature extraction and matching. Jalilvand et al. (2011) evaluate the H_{INV} and W_{INV} invariants also for SURF descriptor matching with the framework of Burghouts and Geusebroek (2009) on the ALOI dataset, and report that W_{INV} is the better invariant. The last comparison regards the version of the LIC gradient from the study of Stöttinger et al. (2012). The original LIC implementation used a boosted image to generate the LoG stack of the HL detector which resulted in improved performance in some situations. In this research LIC was used without the colour boosting in order to isolate the effect of the invariant itself, and the results show that it performs poorly under all the tested conditions.

Due to the relatively low number of correct feature correspondences and matches achieved by the majority of the colour invariants, a novel feature correlation analysis was devised and presented in Section 4.4. The correlation study investigates how colour and grayscale information can be utilised simultaneously for local feature extraction in the detection phase. Despite their overall inferiority when utilised individually, the correlation study obtained promising results and strongly indicates that colour invariants have a substantial potential to be used for feature fusion extraction, as they are uncorrelated to the intensity and generate a considerable number of correct unique point correspondences.

When analysing the overall performance of the feature detection, the proposed fusion techniques do not prove to be superior to all the other individual colour gradient features. However a detailed analysis on the standard ranking metric used in the literature for HL points, provides some answers as to why the proposed fusion techniques do not perform as expected. The analysis concludes that the Harris cornerness energy strength, is not an accurate metric for the ranking of HL points, and therefore should not be used in fusion techniques. The investigation on the ranking of HL points via the Harris energy is another completely novel aspect of this research which has been overlooked in the literature, and provides an interesting direction of future research for the field.

Object Class Recognition 5

This chapter is concerned with object class recognition, which is arguably amongst the most important application areas in computer vision (along with image feature matching), where local image features are used. The goal of object recognition is to determine what types/classes of objects are present in an image (e.g. building, person, car); whereas the goal of image retrieval (another important application) is to find the same image or the same object within a query image (e.g. The Colosseum), in a database of unknown images. Object recognition techniques must extract discriminative information from images and during the training phase be able to identify the information that is mutually shared amongst the same class of objects (intra-class similarity), and which in turn is unique and not similar to other classes (inter-class dissimilarity).

During most of the last decade, Bag-of-Visual-Words (BOVW) (Sivic and Zisserman, 2003) has arguably been the most successful approach in the area of image recognition (Chatfield et al., 2011), while also obtaining excellent results for object detection (Vedaldi et al., 2009) and image retrieval (Nister and Stewenius, 2006). The BOVW approach extracts the discriminative information from images with local image features, with the majority of techniques employing the SIFT descriptor (Lowe, 2004). BOVW represents a dataset of images with a visual vocabulary, where every *word* of said vocabulary is a feature descriptor (i.e. SIFT). Images are then encoded as a frequency histogram where each bin is represented by a visual word, the histograms of words are then used as the image descriptors for the training and classification phases.

More recently, with the popular resurgence of neural networks in the field of machine learning due to a rise in computational power and abundance of data, approaches like Deep Learning (Bengio, 2009, Simonyan and Zisserman, 2014)

have come to the forefront of image recognition and are now considered to be the state of the art in terms of recognition accuracy. Deep learning techniques do not use local image features, they instead learn the most appropriate set of low-level image structures for a particular class. As a consequence, deep learning relies on large volumes of image data, and the results are heavily influenced by the type of data that is trained on. Another disadvantage to this approach is the computational requirements and time needed to train a recognition system. Despite achieving the best recognition rates for image classification, deep learning is not ubiquitously used within the computer vision community, BOVW techniques are still commonly being researched in areas where computational power is limited or when the scale of the image database of the task does not comprise of millions of images.

Three representative examples of applications that utilise BOVW are: action recognition (Iosifidis et al., 2014), efficient visual search on mobile devices (Chen et al., 2014, Chen and Girod, 2014), and map loop closure in Simultaneous Localisation and Mapping (SLAM) techniques (Mur-Artal and Tardós, 2014). In this research, BOVW will serve as the evaluation framework for investigating if the colour invariants that were tested in the previous chapter, can positively contribute to object recognition tasks. The findings of this work will also have implications on other applications that utilise BOVW and local image features.

Recognition with BOVW has predominantly been based on extracting local shape information with grayscale intensity gradient information, although the use of colour has been proven to improve some image classification tasks (Van De Sande et al., 2010, Vigo et al., 2010a). The improvement in performance attributed to colour generally depends on the importance of colour in the data set. However, the best way to combine different image cues (colour, shape, texture, etc.) within the same recognition pipeline, remains an unanswered research problem. The efforts to finding an optimal fusion strategy have increased in recent years (Khan et al., 2011, Fernando et al., 2012, Khan et al., 2012). Two main approaches exist for combining colour and shape (geometric) information into the BOVW framework: early fusion and late fusion where the nomenclature depends on whether the fusion is performed prior or after the vocabulary generation.

Early fusion involves utilising colour and grayscale local image descriptors together and fusing them to create a single shape-colour vocabulary of the dataset. Late fusion approaches obtain separate image vocabularies and when an image is encoded to form a histogram of words, the final descriptor can then comprise of a concatenation of a shape histogram and a colour histogram, or the two descriptors can also be combined at the classifier stage. In late fusion, a parameter is commonly used to balance the relative contribution of colour and shape. The recognition of object classes that are colour-shape dependant (i.e. a red-white stop sign) are generally better served with an early fusion strategy; whereas classes which have colour and shape independence (i.e. cars, cats, dogs) are better represented with late fusion.

In this research, a standard BOVW recognition pipeline is implemented using the VLFeat open source toolkit Vedaldi and Fulkerson (2008). The aim is not to provide state-of-the-art results but rather to evaluate the colour invariant features and compare them under the same conditions. A sparse feature extraction technique versus a dense random sampling feature extraction will also be compared. Colour will provide its contribution to the recognition pipeline by being utilised in both the detection and description phases of the feature extraction process. Two recognition studies are performed here, the first evaluates all 10 gradient types individually, and the second is an early fusion approach which combines the best performing gradient types together. The proposed fusion technique relies solely on the discriminative power of the colour SIFT descriptors, and is thus not affected by the limitations of the Harris energies as discussed in the previous chapter.

Section 5.1 discusses the different methods of extracting features for BOVW recognition and Section 5.2 outlines the recognition pipeline that is used here. The evaluation details of the PASCAL VOC 2007 challenge are covered in Section 5.3 before presenting the individual recognition results in Section 5.4. The implementation details and results of the colour fusion technique is presented in Section 5.5, and the chapter ends with a summary and discussion in Section 5.6.

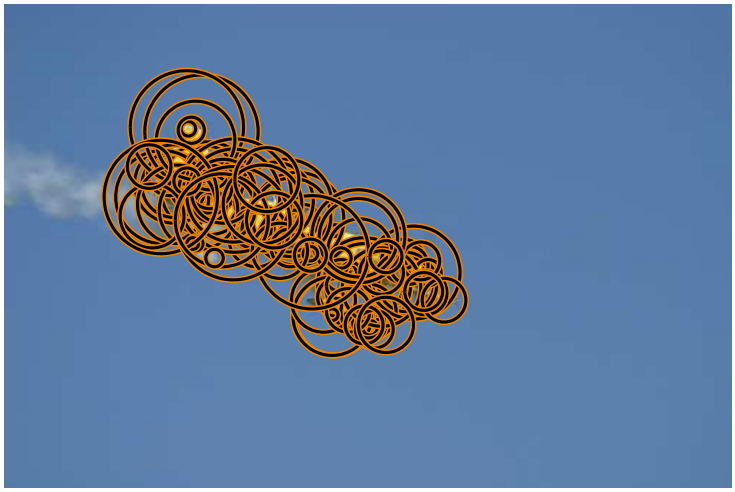
5.1 Feature Extraction

In the early years of BOVW-based recognition approaches, the image vocabulary of a dataset was obtained by extracting sparse local features in the standard way as outlined in Chapter 4. The goal was to find salient image regions using a detector and characterise those regions with a suitable descriptor. All grayscale-based descriptors were geometric like SIFT or SURF, whereas they could vary in the case of colour descriptors with some techniques utilising a mixture of grayscale SIFT with histograms of colours (Van de Weijer and Schmid, 2006b, Van De Sande et al., 2010) and others obtaining geometric colour descriptors by applying SIFT on individual colour channels (Bosch et al., 2006, Van De Sande et al., 2010). In terms of the detection strategy, it was found that results could be improved when combining several different types of detectors together (Mikolajczyk et al., 2006, Sivic et al., 2005).

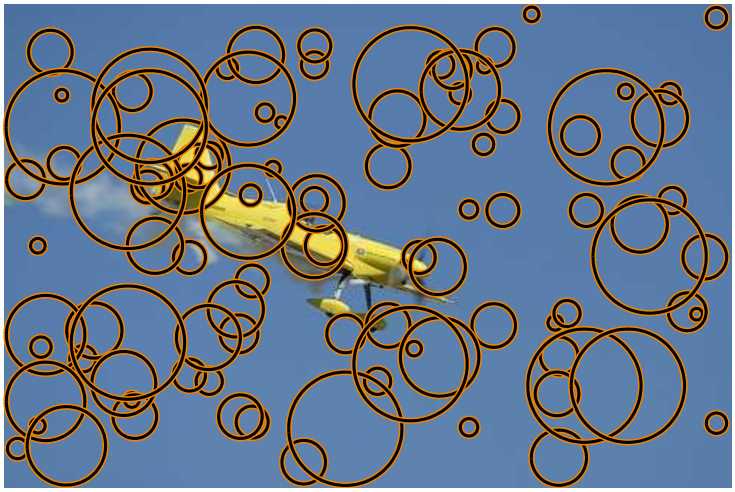
The current feature extraction approach is to disregard the initial detection step altogether, and to essentially extract all of the information from an image by covering the entire image with local image regions which are then characterised with a feature descriptor. This approach is called dense random feature extraction and involves making a grid of image positions, and centring on each position local image regions of various sizes in such a way that every area of the image is encapsulated by at least one descriptor. From a pool of all possible descriptors for an image, a random set of descriptors then gets selected. The current general trend is toward increasing the number of extracted descriptors from an image (Zhang et al., 2007, Nowak et al., 2006, Tuytelaars and Schmid, 2007). While dense feature sampling has been shown to obtain better results in image classification than sparse features (Nowak et al., 2006), the strategy essentially relies on the machine learning part of the pipeline for disregarding the non discriminative descriptors when formulating the visual vocabulary.

Figure 5.1 demonstrates the differences between the sparse and dense feature extraction approach used in this research. Figure 5.1a shows 100 sparse interest points detected using the developed HL algorithm, Figure 5.1b shows 100 random local regions selected from the pool of 1,000 total regions shown in Figure 5.1c.

(a) Sparse Interest Points



(b) Sparse Random Points



(c) Dense Pool of Points

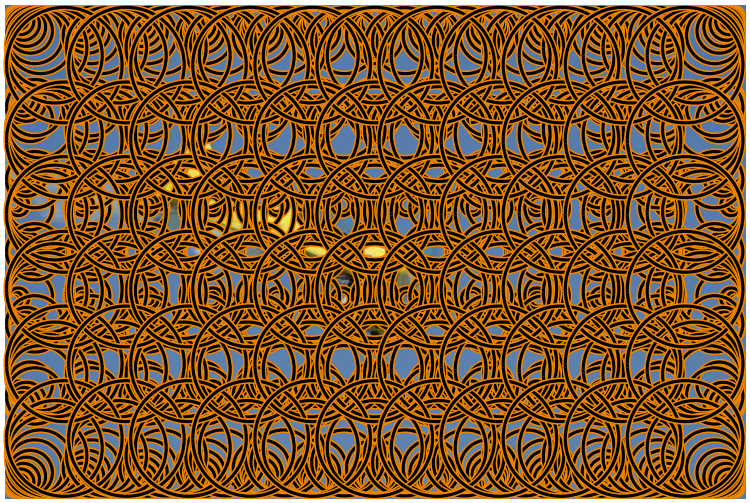


Figure 5.1: Local feature extraction approach comparison: (a) Top 100 sparse HL points, (b) 100 random point and (c) a pool of 1000 random dense points.

5.2 The Recognition Pipeline

The BOVW recognition pipeline that was developed for this research is derived from the standard code examples of the VLFeat toolkit (Vedaldi and Fulkerson, 2008). Figure 5.2 shows the diagram of the overall recognition process that is used here. Step 1 collects a representative subset of information from an image dataset, needed for the visual vocabulary generation. A number of local features are extracted from each image of the dataset, descriptors are then obtained from each local region and are accumulated in a large vector containing all the extracted descriptors. Step 2 generates the actual vocabulary by clustering (discretising) all the extracted descriptors from step 1 in the 128x128 dimensional SIFT descriptor space. This step is also referred to as vector quantisation, where for each cluster that is detected a new descriptor representing the information from the entire cluster is created by quantising the information of all the descriptors around the centre of the cluster. The quantised descriptors then become the words of the vocabulary, which are denoted as $W_1 ... W_N$ in Figure 5.2.

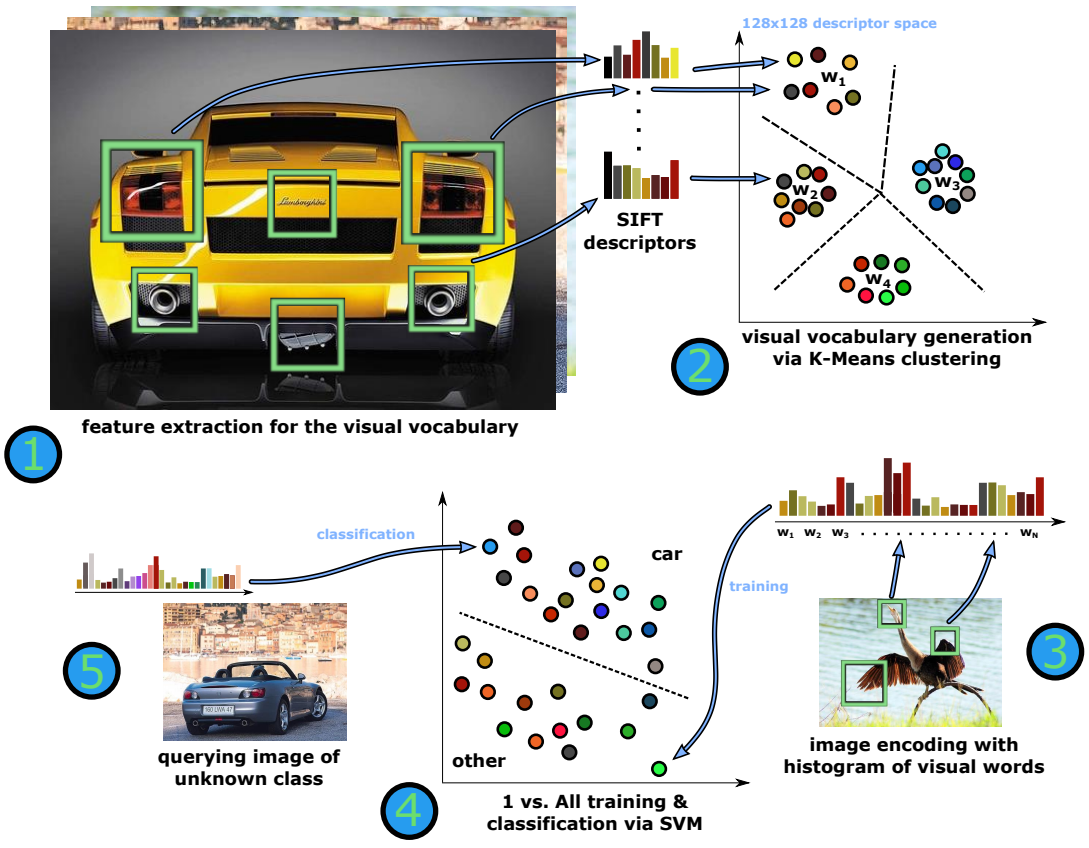


Figure 5.2: Diagram of the bag of words recognition pipeline.

In the experiments presented in this chapter, the clustering of the vocabulary is performed with the *K*-Means algorithm and the vocabulary size consists of 1,000 words. Therefore the histogram of words descriptor that is obtained for each image has 1,000 bins. This image descriptor is depicted in steps 3 and 5 of the recognition diagram. During the training phase of the recognition process, each image of the training dataset is encoded as a histogram over the visual vocabulary. Each of the feature descriptors extracted from an image are matched to one of the words of the vocabulary which increases that bin count of the histogram of words. The matching is performed by calculating the Euclidean distance between a descriptor and all the 1,000 words, and selecting the word with the smallest distance.

Since the training dataset is labelled (the types of classes in each image are known), it is possible to train a classifier for each class by feeding the classifier with known instances of histograms that pertain to that class and histograms that do not, this is depicted in step 4. A 1-vs-All Support Vector Machine (SVM) Shalev-Shwartz et al. (2011) classifier is used in this work with a linear kernel, which is able to obtain a linear hyperplane that separates the histogram of words descriptors into two categories; one that pertains to the class being trained for (e.g. car) and another category for all the other classes. During the classification stage in step 5, all the images from the evaluation dataset (also labelled) are encoded with histograms of words and are projected into a trained classifier to determine if it belongs to a particular class or not. Since all the dataset is labelled, it is possible to verify if a classification is correct.

The main focus of the implementation was to evaluate all the invariants within the same framework, prove the concept of the proposed feature fusion and utilise a well-known recognition toolkit to ensure the recognition experiments were performed correctly. Obtaining state of the art BOVW recognition results was not a goal of this work since this research is not concerned with developing machine learning techniques. This implementation is therefore not sufficiently optimised. In order to improve the results a number of tactics can be employed that may or may not necessarily affect the relative performance of the colour invariants with respect to the grayscale intensity. Furthermore the evaluation aims at investigating the role of colour features in their purest form, and not to

dilute their individual effect by employing more sophisticated machine learning techniques. As the study of Chatfield et al. (2011) demonstrates, each step of the recognition pipeline (feature extraction, image encoding, classifier method) and their specific tuning parameters substantially influences the recognition results.

One such technique that improves recognition results is Spatial Pyramidal Matching (SPM) (Lazebnik et al., 2006). This splits up the image into multiple segments of decreasing sizes, and then a histogram of words is generated for each individual segment, which are all then concatenated to form the final image descriptor. SPM is thus able to embed spatial information into the BOVW representation by essentially creating a form of local histograms of words. In the results shown in this research, only one image segment (the entire image itself) is used to generate the histograms of words.

Another optimisation method that could be employed is to generate a very large vocabulary size (e.g. 50,000 words), or to perform a more advanced vector quantisation with for example a Gaussian Mixture Model (GMM) and obtain a vocabulary of Fisher Vectors (Perronnin et al., 2010). The GMM does not only contain the descriptor located at the centre of a cluster, but also other statistical information like the mean vector and covariance matrix of the cluster which adds to the discriminative capacity of the vocabulary. The image encoding step can also be improved by performing a *soft-assignment* of each feature descriptor of an image to the histogram of words descriptor. The current implementation performs a *hard-assignment*, where one feature descriptor only contributes to the increment of a single word in the histogram. A soft-assignment is where a descriptor can contribute to the bin-count of more than one word, with the weight of the contribution depending on the Euclidean distance between the words and the feature descriptor. Finally, using a non-linear kernel for the SVM classifier will also tend to improve recognition results (Chatfield et al., 2011).

5.3 The PASCAL VOC 2007 Challenge

The PASCAL VOC (Visual Objects Challenge)[7] (Everingham et al., 2007), consists of two main tasks; object detection and object classification. This research focuses only on the classification task, which involves predicting if images from real-world scenes contain a particular class of object (car, bus, person etc.). The 2007 VOC dataset contains 20 object classes, 5,011 training images, 4,952 test images and is known to be predominantly shape dominant (colour has a lesser impact than geometric information). The challenge specifies the list of training and testing images that must be used for each of the 20 classes, so that the challenge is always performed with the same data in order to compare different techniques equally. Therefore, each object class contains two lists of image names, a list of names for all the images that will be used to train the classifier for that class, and a list for the images that will be used during the testing. Since the numbers of different object types are not equally distributed throughout the dataset (i.e. there are many more instances of persons than any other object class), the training and testing lists for a class vary in terms of the number of positive images (containing objects of the class) and negative images it contains. For the class *Bus* for example, there are 4996 training images of which 3.9% contain instances of a bus, and there are 4770 testing images with 3.6% positive image samples. Most of the other classes also train and test on approximately half of the entire dataset, maintaining equal proportions of positive samples in both training and testing at around 5%. Some examples of the proportion of positive samples are: $Aeroplane = 4.3\%$, $Car = 16\%$, $Cat = 7\%$, $Horse = 5.8\%$, $Motorbike = 4.7\%$, $Person = 70\%$, $Potted - plant = 4.7\%$ and $Train = 5.5\%$.

The PASCAL VOC challenge is evaluated with precision-recall curves for each object class, then obtaining an average precision (AP) per class and a mean average precision (mAP) for the final result of all the 20 classes. When a classifier must retrieve from a list of images those which contain a specific object class, the number of retrieved images are controlled by thresholding the scores from the SVM classification results. True positives (tp) are instances where an object was correctly classified in a retrieved image, and false positives (fp) are instances where the recognition incorrectly retrieved an image as pertaining to the class.

[7] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

False negatives (fn) on the other hand, are recognition instances where an image containing the object class was incorrectly not retrieved. The precision can then be defined as: $precision = \frac{tp}{tp+fp}$, and the recall as: $recall = \frac{tp}{tp+fn}$. Therefore precision measures the probability of accuracy of the prediction, and the recall measures the proportion of correct retrieved instances from all the total available positive instances in the database. The evaluation of the challenge is performed by ranking the retrievals by decreasing classifier scores, then calculating the precision and recall for the first ranked result, then the first two results, then the first three and so on until all the retrieval ranks are taken into account. The AP result for each class is obtained by averaging all the precision values of the precision-recall curve.

5.4 Recognition Results

The first set of recognition experiments evaluates and compares the individual suitability of each feature extraction method, i.e. the 10 different gradient type HL detectors and a random dense feature extraction approach. The overall mean average precision results (mAP) of each extraction type is shown in Table 5.1. These results are obtained with a visual vocabulary of 1,000 words, during the vocabulary formulation phase the top 100 features (ranked by Harris cornerness energy) were extracted per image, creating a total of 996,300 features that were clustered using K-Means. Various tests were carried out on a subset of the dataset which varied the number of words in the vocabulary from 1,000 to 6,000. A vocabulary of 1,000 words was chosen due to a small difference in the performance and because a larger vocabulary requires more computational time. The purpose of the experiments was not in maximising the performance however, but to evaluate the colour gradients in a comparable manner. The choice of extracting 100 features per image for the individual evaluation, partly arose out of memory constraint problems that occurred at the time of clustering all of the points extracted from the dataset. The maximum number of features that could be obtained per image was approximately 300, but 100 was selected in order to reduce the time required to perform the training. This number is however still significantly larger than in the study of van de Sande et al. (2010), who extract 20 features per image in their evaluation.

Table 5.1: Mean average precision results using 1,500 features per image.

Method	I	LIC	SP_{INV}	SS_{INV}	$SPSS_{INV}$	$SPSS_{VAR}$	SS_{F-INV}	C_{INV}	H_{INV}	W_{INV}	$DENSE$
mAP	23.85	16.82	17.92	17.96	13.63	22.22	16.91	18.25	13.63	22.47	22.17

During the image encoding phase, the number of features that are extracted per image was fixed at a maximum of 1,500. This number was chosen because the maximum number of HL points that could reliably be detected per image was approximately 1,500. However, not all methods were able to consistently extract that number of features for all the images.

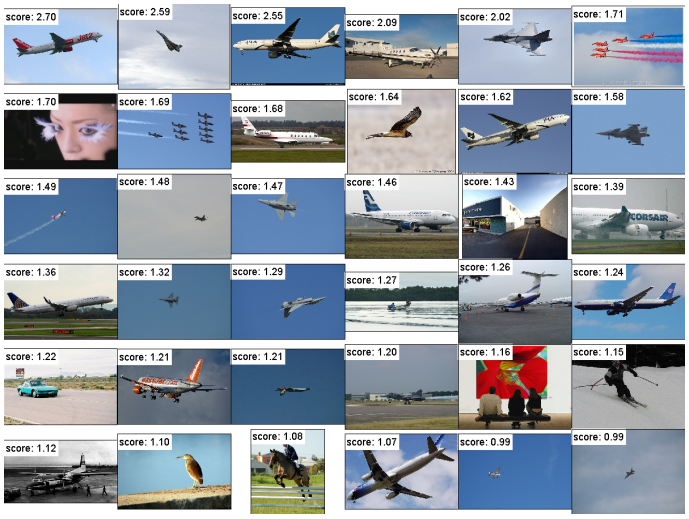
For the random dense extraction approach (denoted as $DENSE$ in the tables and plots), only grayscale intensity features are considered. The random sampler selects points from a pool of regions containing the same distribution of scales as the HL sparse detector (averaged from each image of the entire dataset). The final set of random points are generally uniformly spread throughout the image, visual results of the random feature extraction can be seen in Figures 5.1(b and c). The dense results are obtained with the same conditions as the sparse features, and the dense results shown are the mean of 6 repetitions. The individual AP results for each class are presented in Table 5.2.

Visual examples of the top ranking retrieved images (grayscale features) for 6 classes are shown in Figures 5.3, 5.4 and 5.5. Those figures clearly convey the output of the recognition process and demonstrate how, visually similar but different objects can be classified together, and in other cases how clearly different objects can still be interpreted to be in the same class. A useful observation that can be made from the visual results is that the background appears to have a strong impact on the recognition process. This is evident in the Aeroplane class for example, where in the dataset many images of this class contain a texture-less background of the sky. In the results of Figure 5.3a, many of the incorrectly retrieved images also contain a texture-less background; like the two bird images, the car, the skier and the woman. The last set of example results are presented in Figures 5.6 and 5.7, that show the precision-recall curves of the top 6 gradient types for 4 object classes.

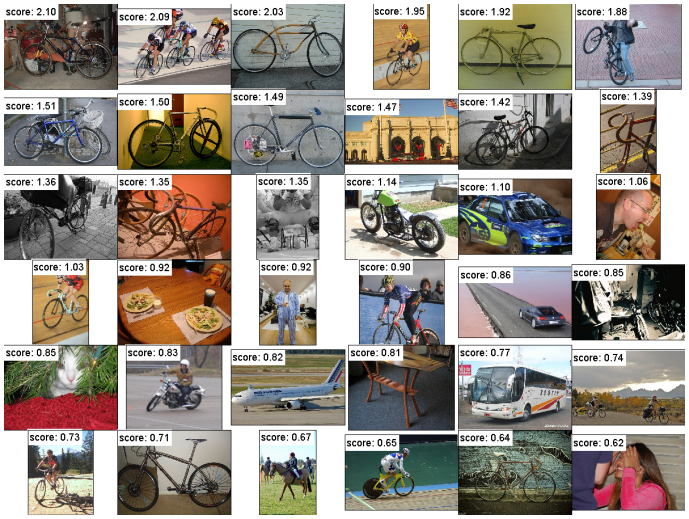
Table 5.2: Average precision results per class using 1,500 features per image.

Method	I	LIC	SP_{INV}	SS_{INV}	$SPSS_{INV}$	$SPSS_{VAR}$	SS_{F-INV}	C_{INV}	H_{INV}	W_{INV}	$DENSE$
Aeroplane	43.28	23.09	25.24	27.32	16.46	**46.93**	27.17	24.26	32.60	44.75	43.26
Bicycle	**28.35**	13.54	13.95	10.65	9.05	16.73	16.29	13.12	12.57	20.55	23.61
Bird	14.25	11.50	12.94	13.48	10.39	15.05	12.14	12.5	15.06	**16.22**	13
Boat	**33.74**	16.13	18.6	26.07	10.66	31.85	15.69	21.92	16.20	28.34	32.19
Bottle	8.77	7.32	8.8	8.23	7.24	9.47	6.45	8.01	8.42	**10.61**	8.43
Bus	19.54	12.99	13.40	16.37	7.96	21.68	12.77	13.73	11.87	**20.16**	16.88
Car	45.55	35.80	38.72	37.26	30.18	44.63	36.09	39.24	38.64	**47.16**	45.74
Cat	**19.29**	13.05	13.49	11.21	10.58	15.41	10.56	14.73	11.88	13.85	18.9
Chair	**28.51**	23.43	23.33	24.88	16.18	26.90	24.93	25.52	25.16	27.96	26.82
Cow	**11.68**	5.79	6.35	6.53	6.03	8.54	5.65	4.46	7.19	7.04	9.87
Dining Table	**15.07**	12.47	11.23	9.21	8	11.42	7.65	8.49	10.97	11.46	12.12
Dog	**15.32**	11.20	12.43	12.22	11.52	14.04	12.82	13.96	13.42	14.75	15.16
Horse	26.29	26.18	26.72	24.42	17.78	**27.90**	17.74	22.72	20.35	25.61	25.45
Motorbike	19.57	12.14	14.73	15.03	8.91	17.85	13.06	16.77	17.14	**25.14**	16.53
Person	60.45	58.45	59.47	57.88	55.51	59.36	57.89	**61.26**	60.83	60.36	57.87
Potted Plant	**8.18**	7.66	7.03	7.52	6.88	6.64	6.13	6.53	7.02	7.27	7.55
Sheep	**11.73**	8.71	8.54	6.56	8.31	10.13	9.76	7.04	7.52	7.57	8.69
Sofa	**17.06**	12.02	15.96	14.39	7.82	14.02	12.68	13.03	13.15	15.49	16.32
Train	**30.72**	13.60	15.42	16.46	13.15	27.34	18.27	22.16	21.28	29.49	29.54
TV Monitor	**19.74**	11.35	12.01	13.49	10.01	18.46	14.38	15.47	12.70	15.64	15.38
mAP	23.85	16.82	17.92	17.96	13.63	22.22	16.91	18.25	13.63	22.47	22.17

(a) Aeroplane



(b) Bicycle



(c) Cat

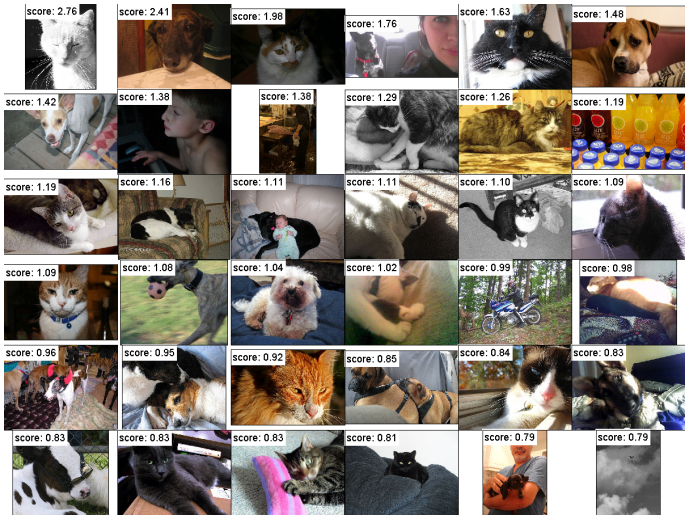
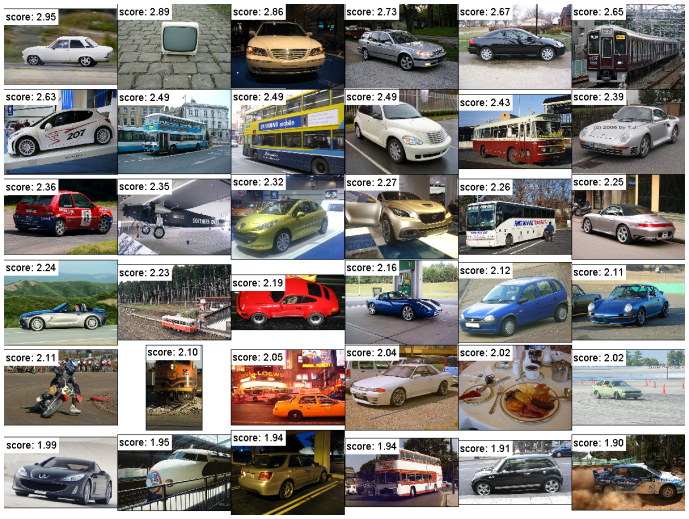
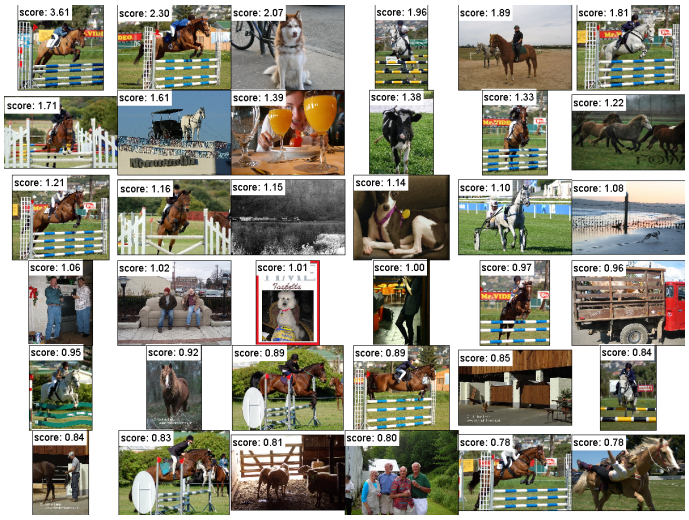


Figure 5.3: Examples of the top 30 ranked images of the classification results for the classes: a) Aeroplane, b) Bicycle and c) Cat.

(a) Car



(b) Horse



(c) Motorbike

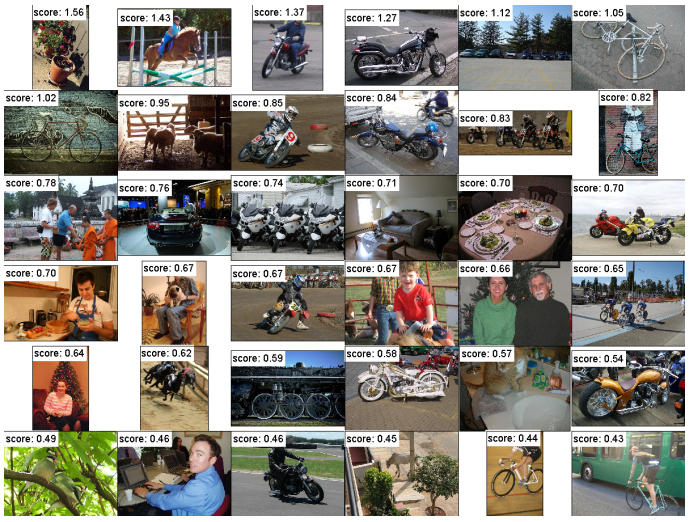
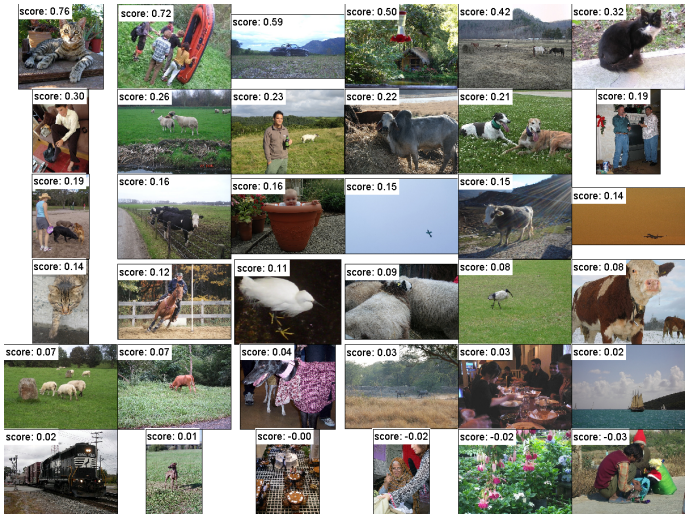
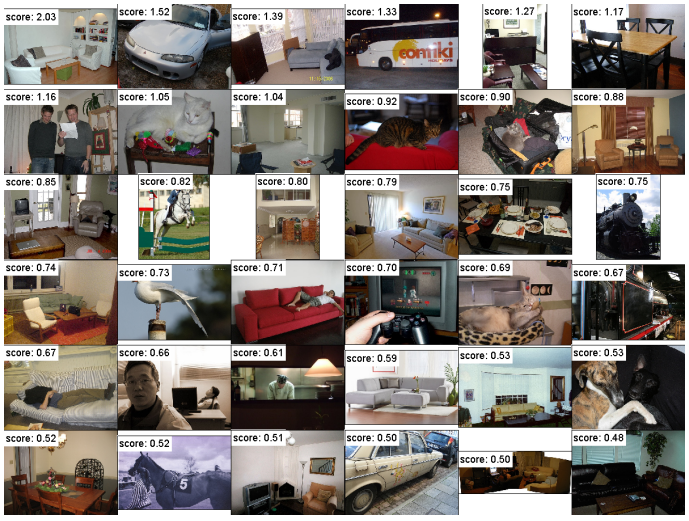


Figure 5.4: Examples of the top 30 ranked images of the classification results for the classes: a) Car, b) Horse and c) Motorbike.

(a) Sheep



(b) Sofa



(c) Train

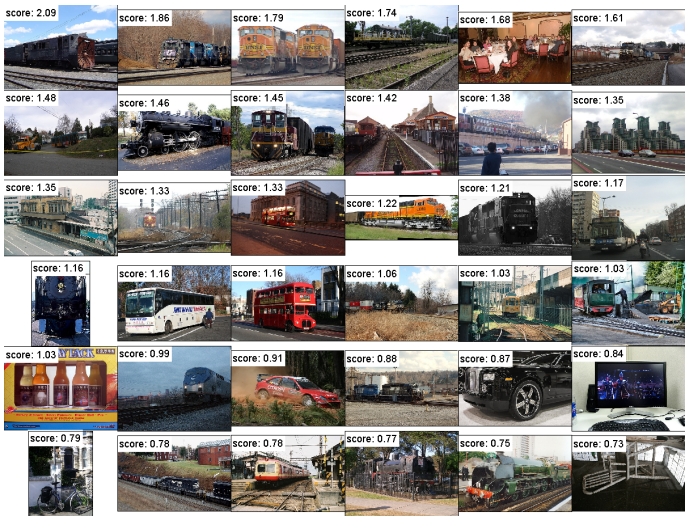


Figure 5.5: Examples of the top 30 ranked images of the classification results for the classes: a) Sheep, b) Sofa and c) Train.

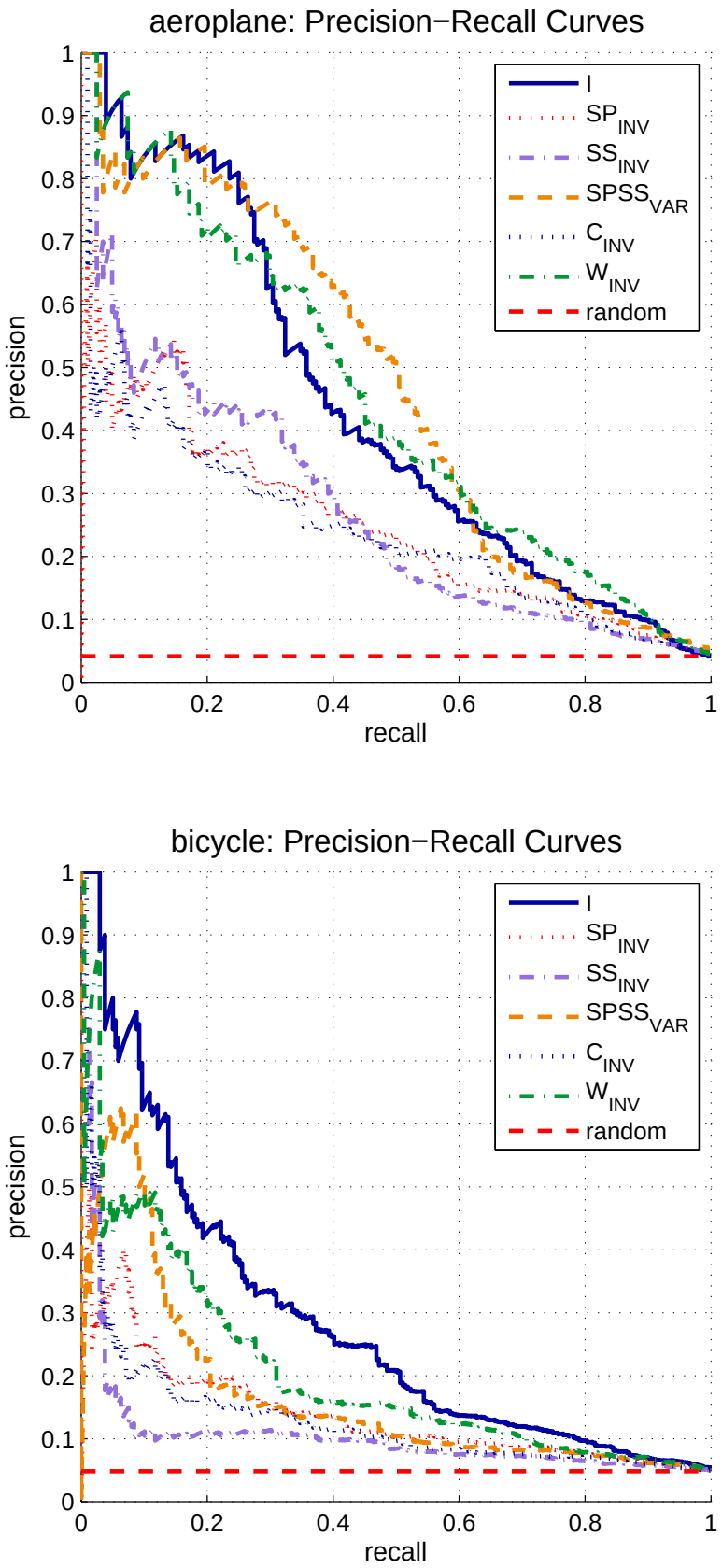


Figure 5.6: Precision-Recall curve examples for the classes Aeroplane and Bicycle.

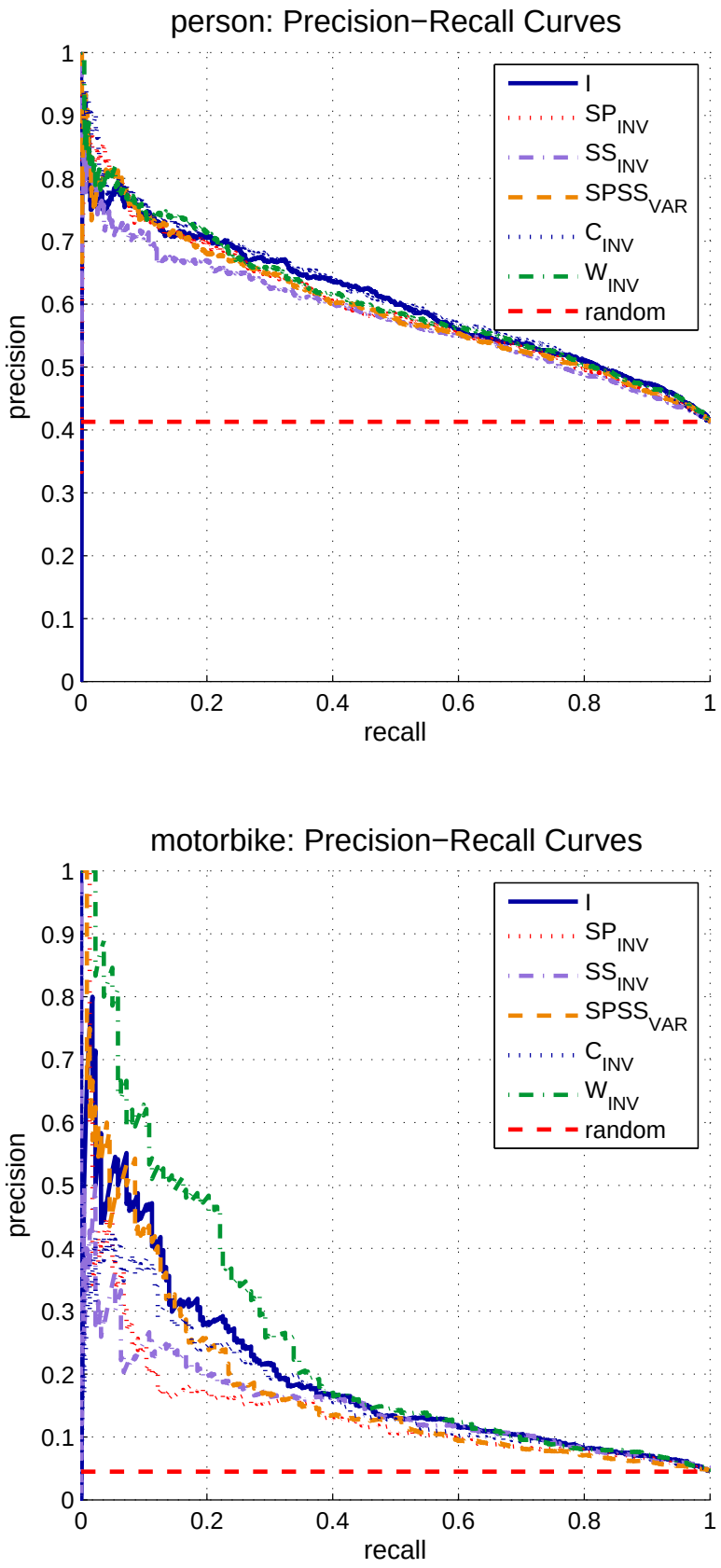


Figure 5.7: Precision-Recall curve examples for the classes Person and Motorbike.

Grayscale intensity proves to be the superior method overall, but there are 8 classes (40% of the challenge) that are better represented by other colour gradients. The fact that nearly half of the challenge can be improved by using colour is a major significant finding, that justifies why the main goal of this work was to investigate the impact of colour on local image features. This result thus clearly proves that colour can benefit image recognition tasks and that the tested colour gradients should be considered for local feature extraction. What is an entirely unique contribution of this research however, is that the best colour gradient types are shown to be $SPSS_{VAR}$ and W_{INV}, which have previously never been utilised for image recognition in the literature.

The random feature extraction performed comparatively to $SPSS_{VAR}$ and W_{INV} although it did not obtain the best results for any individual class. The usefulness of the random extraction is only apparent when the sampling of features is very dense. This study thus concludes that the proposed sparse colour feature extraction techniques have the potential to be used in a fusion approach to improve the recognition results of grayscale features. The next section of this chapter continues the evaluation by investigating if a BOVW feature fusion extraction can enhance the recognition rates of a grayscale only feature extraction.

5.5 Feature Fusion for Object Recognition

Dense random feature sampling has shown in the past to improve the results of BOVW recognition (Nowak et al., 2006). In general the current trend for BOVW image classification has been toward increasing the number of densely sampled points or combining several types of detectors. Even though dense sampling has been effective, these approaches essentially shift the task of discarding the non discriminative points to the machine learning algorithms and therefore diluting the impact that computer vision tools can make to the recognition process.

The goal of this section is to investigate if using a feature fusion extraction method with different colour invariant detectors selects more discriminative interest regions than a randomised dense sampling approach. The fusion would thus benefit image class recognition by providing a more salient set of descriptors to the machine learning algorithms (i.e. clustering and classific-

ation). This section evaluates the performance of a grayscale dense random feature extraction approach as the number of encoding features are increased, and compares it with two fusion extraction approaches (random and sparse), that incorporate features from grayscale and multiple colour gradient types. In the feature detection studies of Chapter 4, it is shown that the colour invariants extract a significant number of unique features that are correctly matched and can be of value in a feature fusion extraction approach for image feature matching tasks. This section tests the hypothesis that the same *uniquefeatures* concept from Chapter 4 holds in a BOVW framework, and if the different colour descriptors when combined will enhance the classification of object classes.

The proposed feature fusion strategy is implemented in its raw form to prove the hypothesis in a direct manner and focus the analysis on the local features themselves, limiting the role of more high-level information or machine learning. The fusion that is implemented is therefore in the category of early-fusion where the different feature types are combined prior to obtaining the vocabulary. The same number of features are extracted from different colour invariants, and all are used together for both the vocabulary formulation and the histogram of words encoding. BOVW results improve by increasing the number of extracted features, by combining the maximum 1,500 number of features from each colour invariant, it is possible to increase the effective number of features for encoding for the sparse fusion technique. A dense representation is thus obtained by fusing the results of separate sparse interest point detectors.

The feature fusion combinations are composed of the top performing extraction types from Table 5.1, and the clustering step is performed using approximately 50 features per image for each feature type. This parameter is chosen due to the memory constraints that were previously mentioned, where the maximum number of features that can be extracted per image is approximately 300. The maximum number of gradient types used simultaneously in the fusion is 7, therefore in that case each gradient type extracts 43 (300/7) features per image. In all the other fusion techniques the clustering is performed using 50 features, in order to maintain the fusion results comparable.

Results of the fusion experiment are shown in Figure 5.8, with the number of encoding points varying up to a maximum of 10,500 and equally distrib-

uted amongst the feature types in each fusion. A more detailed comparison is presented in Table 5.3, that contains the results per each object class of the top methods from Figure 5.8, at 6,000 encoding points. The experiment compares the standard dense random sampling approach using grayscale ($I\ DENSE$) and colour random dense approaches ($SPSS_{VAR}\ DENSE$ and $W_{INV}\ DENSE$) to various fusion extraction techniques. There are two random dense fusion techniques ($I + W_{INV}\ DENSE$ and $I + SPSS_{VAR} + W_{INV}\ DENSE$), that are obtained similarly to the other dense methods but each gradient type contributes equal number of descriptors to the final set of extracted features.

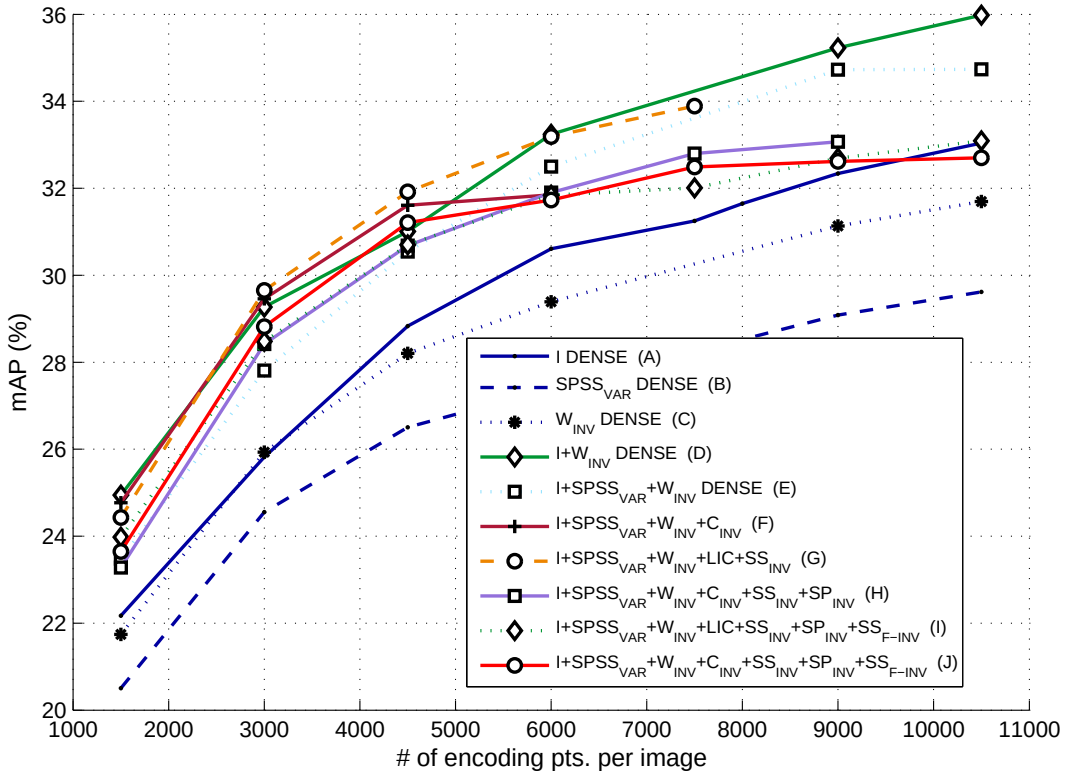


Figure 5.8: Results of multiple fusion methods with varying number of encoding points.

All the other fusion types combine the sparse features extracted with the HL detector. The most important result that arises from this experiment, is that the standard dense sampling (using intensity SIFT only) obtains consistently worse results compared to the proposed fusion combinations for the same number of encoding points. Only at the last parameter of 10,500 points does $I\ DENSE$ obtain comparable results, indicating it is a far less optimal solution to object recognition, especially in a system with constrained resources or real-time

performance requirements. The dense sampling can achieve better recognition rates than a single individual sparse feature extractor such as I, but only when the encoding points are increased above 2,000. The colour fusion random techniques were able to achieve the best precision rates of all the approaches, at the highest number of encoding points. The technique $I + W_{INV}\ DENSE$ on average also achieved comparable results to the best sparse fusion technique at the lowest number of encoding points.

$SPSS_{VAR}\ DENSE$ obtained the worst results out of the individual random dense approaches, which is likely to be one of the reasons for why $I + SPSS_{VAR} + W_{INV}\ DENSE$ did not perform better than $I + W_{INV}\ DENSE$, despite having an extra gradient type in the fusion. The best sparse fusion method is $I + SPSS_{VAR} + W_{INV} + LIC + SS_{INV}$ despite only being able to extract a maximum of 7,500 points. It is also better than the combination of the top 5 individual methods ($I, W_{INV}, SPSS_{VAR}, C_{INV}, SS_{INV}$). It can thus be deducted that the results for both the random fusion and sparse fusion techniques indicate that increasing the number of gradient types does not improve the overall recognition results, and that the fusion schemes need to be more sophisticated to optimise the impact of each colour invariant and avoid the inclusion of intra-class inconsistent points. Evidence of the lost potential of the fusion, can be seen in the results of Table 5.3, where each of the methods obtains the best results for at least one of the 20 classes. For example method J ($I + SPSS_{VAR} + W_{INV} + C_{INV} + SS_{INV} + SP_{INV} + SS_{F-INV}$), while overall performing the worst of the 7 methods in terms of mAP, still obtains the best precision for 3 of the classes. This indicates that fusing multiple gradients types can be beneficial, but due to its localised positive impact it suggests that the current fusion strategy cannot harness the full potential benefit of fusing all the colour invariants.

In the random dense fusion, the information from the different gradient types are combined via a K-Means algorithm, the sparse fusion also depends on the K-Means but prior to that the feature descriptors are selected with the HL detection, which favours regions with discriminative gradient information. The benefit of adding more discriminative regions to the BOVW vocabulary is apparent from the results, as the random extraction of $I\ DENSE$ performed worse than the sparse colour fusion. However, when analysing the results of

the random dense fusion techniques, the discriminative information added by the colour can be seen to be enough to compensate for the randomness of the extraction, as it improves the recognition rates compared to the standard approach of *I DENSE*. The comparative results of the sparse and random colour fusions, indicates that the background of an image and the regions of low texture can substantially influence the recognition results. In the object classes similar to *Aeroplane*, *Boat* and *Bird*, the same types background are frequently present in the images (e.g. sky and sea), and are thus an important cue for their recognition. The sparse HL feature detection focuses on textured regions, generally inside and around the boundaries of objects, whereas the random extraction can extract regions from every part of the image.

Table 5.3: Average precision results per class for the fusion techniques.

Method	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Aeroplane	**60.90**	60.29	58.32	58.27	51.15	54.90	55.21
Bicycle	35.10	30.40	34.95	34.64	34.95	35.77	**37.52**
Bird	**24.68**	22.40	20.56	23.92	23.13	22.09	20.23
Boat	47.71	50.15	47.22	**50.48**	50.44	47.55	46.97
Bottle	10.12	11.64	**12.93**	12.55	10.46	9.35	10.11
Bus	**35.40**	29.55	25.85	33.74	27.08	32.08	28.77
Car	**59.43**	58.32	56.38	54.86	53.71	54.48	52.29
Cat	25.42	**27.44**	22.57	24.66	23.49	25.60	22.42
Chair	**37.28**	35.79	34.83	36.60	36.18	35.74	35.11
Cow	15.46	14.83	12.76	**20.97**	14.82	11.68	15.16
Dining Table	19.86	**20.08**	16.52	18.61	18.26	16.46	14.68
Dog	19.38	15.89	18.47	19.43	17.84	19.11	**19.77**
Horse	48.46	46.00	50.63	50.02	**53.20**	50.08	51.35
Motorbike	**36.10**	35.61	30.73	34.42	32.19	33.32	31.13
Person	67.56	67.41	67.28	70.10	69.49	70.05	**70.95**
Potted Plant	9.59	8.48	9.01	**10.64**	9.23	9.92	9.58
Sheep	11.46	13.83	14.79	15.00	14.16	**16.46**	16.15
Sofa	24.84	**25.97**	23.11	21.61	23.57	20.76	21.44
Train	50.07	49.78	**50.29**	48.88	46.33	47.13	49.87
TV Monitor	25.89	25.97	**29.62**	24.38	28.28	24.12	25.77
mAP	33.24	32.49	31.84	33.19	31.90	31.83	31.72

From the obtained experimental results, when comparing the proposed sparse and dense colour fusions, the overall results are not improved by performing a HL detection step prior to obtaining the BOVW vocabulary or encoding the histogram of words. However from the results in Table 5.3, it can be seen that 11 of the classes (55% of the challenge) are better represented by sparse fusion techniques. Therefore the HL detection step does increase the distinctiveness of the visual vocabulary in certain cases. The HL detection apart from largely ignoring the background, may detect many of the same regions across different colour gradient types apart from also detecting unique regions. This duplication therefore can at times contribute useful additional colour descriptors, or at other times contribute to conflicting and noisy descriptors that at the moment of clustering via the *K*-Means, can result in a detrimental effect to the recognition.

5.6 Summary and Discussion

The work in this chapter has evaluated the performance of local features from colour photometric invariants with SIFT descriptors, in the context of Bag-of-Visual-Words (BOVW) object class recognition on the PASCAL VOC 2007 challenge. Colour is used in both the detection and description phases of the feature extraction process, such a strategy has not been widely investigated before in the literature. Two recognition experiments were carried out, one to show the potential benefits of individual colour invariants (Section 5.4) and another to test this potential via a feature fusion scheme (Section 5.5). The individual results showed that despite the intensity achieving the overall best performance, 40% of the VOC's 20 classes obtain better results with methods other than grayscale intensity. This was an important result and a key finding of the overall research, which clearly demonstrated the value of using the colour gradients and indicated that the best colour gradients could be used together to enhance the performance of grayscale features in a fusion approach. The second experiment evaluated raw feature fusion schemes that combine features extracted from separate colour invariants. The proposed sparse and dense feature fusion schemes obtained consistently better precision results than using random dense sampling with grayscale intensity SIFT descriptors. Overall the experimental results strongly demonstrate that the tested colour invariants improve the recognition results of the BOVW framework, and that a random

dense fusion using grayscale intensity I and W_{INV} descriptors, obtains the best recognition precision results.

Contrary to the standard dense sampling used in the literature that randomises the selection of regions, the aim of the proposed sparse fusion feature extraction was to select a dense representation with higher levels of colour saliency. This was based on the hypothesis of detecting and combining salient features from multiple colour invariants which could each provide unique informative descriptors, and thus increase the information content and complexity of the visual vocabulary of the BOVW pipeline. This hypothesis is supported by 55% of the results from Table 5.3, but for the overall results the random dense fusion techniques are the best performers. If all the best precision results for each class (taken from different techniques) were summed and averaged, it would provide an indication for the maximum mAP that could be obtained if the fusion was optimised. That mAP amounts to 35.00, which is 4.81% higher than $I + W_{INV}\ DENSE$, and 14.75% higher than $I\ DENSE$. The full extent of the increase in performance achievable by any future optimal fusion techniques, can not be predicted however.

The best sparse fusion technique ($I + SPSS_{VAR} + W_{INV} + LIC + SS_{INV}$) performs comparatively to the $I + W_{INV}\ DENSE$ for the same number of encoding points, however only two gradient types are needed to achieve the random fusion. Additionally, $I + W_{INV}\ DENSE$ obtained better results overall as it was able to increase the precision rates by extracting more points, whereas $I + SPSS_{VAR} + W_{INV} + LIC + SS_{INV}$ was limited to a maximum of 7,500 encoding points. The success of the dense random fusion extraction, indicates that a random sampling strategy with an automated machine learning algorithm like K-Means is sufficient to fuse the colour information and achieve a performance gain. The overall conclusion of this chapter, is that there is much scope for improving the way that the gradient types (especially I, $SPSS_{VAR}$ and W_{INV}) are used conjointly for object class recognition. In order to maximise the performance, the vocabulary formulation and image encoding steps must be optimised to utilise all the gradient types in a more complimentary manner. Another optimisation must also be performed at the feature detection stage, in order to combine elements from both the random sampling and the sparse region extraction.

Conclusions and Future Work 6

There is one major research question which this thesis has dealt with: how does colour information contribute to the local image feature extraction process? The motivation for this research derives from the observed limitations of state of the art local feature matching, specifically their high rates of feature mismatches that are partly due to the insufficient discriminative power of feature descriptors. One important commonality among the most widely used feature extraction techniques from the literature, is that they were designed to work specifically with grayscale intensity information. This research aimed to investigate, if the inclusion of colour information increased the distinctiveness of local features, and if this influenced the actual performance metrics of two important computer vision applications.

Colour feature detection and matching have not been sufficiently evaluated in the literature prior to this research. Numerous works proposed colour invariants for various applications but unlike many other areas in the computer vision field, the colour features area lacked comprehensive comparative studies that would allow for a certain level of maturity to be reached. Upon inspection of the literature, it was unclear how to proceed in order to increase the distinctiveness of local features. The literature studies which proposed colour invariants generally focused only on their own implementations or on a very limited comparison. Studies also generally evaluated applications and scenarios that were advantageous to their proposed technique but left many unanswered questions. It was important to address those issues in this research, and obtain a more definitive conclusion on the role of colour in local image feature extraction.

The scope of the research has been limited to colour techniques that only take into account the local gradient structure of images, in order to be compatible with the most successful grayscale-based feature extraction methods.

Consequently, numerous colour constancy and colour boosting approaches that utilise statistical techniques to achieve a level of colour invariance, were not regarded in this research. The contributions arising from this work can be grouped in two main categories: *evaluation* and *fusion*.

6.1 Contributions Arising

The first set of contributions in the *evaluation* category, are presented in Sections 4.2 and 4.3 where feature detection and feature matching experiments evaluate and compare the chosen colour invariant gradients. The second evaluation contribution is presented in Section 5.4, where the invariants are tested in an object recognition task. In the fusion category, Section 4.4 presents an investigation on colour feature fusion for local feature detection, and Section 5.5 presents the proposed BOVW feature fusion extraction for object recognition. A conclusion of the main contributions of this research is presented next.

6.1.1 Choice of Colour Invariants

This research was able to obtain a more definitive conclusion on the usefulness of colour invariants for feature matching and recognition, since it has evaluated and compared within the same framework a substantial number of colour gradient invariants. The work uses the biggest number of different colour gradient invariants out of all the studies found in the literature. Experimental results demonstrate that for most of the tested gradient types, there is not enough evidence to justify their individual use for feature matching or recognition.

Of note, is the discovery that $SPSS_{VAR}$ proved to be the best candidate for feature matching tasks under general imaging conditions (varying viewpoint etc.), achieving the best balance of correct number of detections and descriptor matches. Since that gradient type has never been implemented as a local feature in the literature, the finding highlights the necessity for the comprehensive evaluation performed in this work. In regards to the other gradient types, results indicate that grayscale intensity is the overall top individual performer for feature detection in general imaging conditions, and W_{INV} is preferred for both

detection and matching tasks when requiring invariance to changing illumination conditions. The grayscale intensity features performed comparatively to the colour invariants under certain illumination conditions, and was also the best performer for 60% of the object recognition challenge when used individually. Therefore in summary, colour can make a positive contribution in a limited range of scenarios, and should only be considered either individually when the feature matching application requires an invariance to illumination variations, or when colour is appropriately used alongside grayscale in image recognition applications.

6.1.2 Implementation of Robust Colour Features

Many of the tested colour gradients were only implemented as non scale-invariant corner detectors or edge detectors in the literature, which meant their applicability for general local feature extraction was unknown. In this work the gradient types were utilised to extract Harris-Laplace points which are considered to be robust local regions, invariant to a limited range of imaging distortions like scale, rotation and viewpoint variations. The experimental results in this research show that most of the colour invariants are not robust to geometric distortions or even sufficiently invariant to illumination variations, whereas in the literature the invariants obtained positive results in the tasks that they were tested on.

Local gradients impact different applications in various specific ways, and the discovery that most of the colour invariants were unsuitable for individual use for general feature extraction was possible due to the manner that the colour gradients were implemented in this research. In the case of the LIC invariant, the work of Stöttinger et al. (2012) did utilise it to extract robust HL points, however their use of colour boosted images in their generation of the LoG stack masked the actual impact that the invariant had on the results. In this work, the actual impact of the LIC invariant is evaluated and the results show that it in fact performs poorly.

6.1.3 Rigorous Evaluation

The previous contribution allows for the evaluation of this research to be valid in order to test the invariants' suitability for robust feature extraction in general computer vision applications. The evaluation also needs to be rigorous and comprehensive however, and this work achieves these aims due to three key aspects of the work: 1) Utilising sufficiently large and diverse datasets. 2) Applying standardised metrics and testing frameworks for feature matching like the one proposed by Mikolajczyk and Schmid (2005). 3) Conducting the most widely used object recognition challenge (PASCAL VOC).

This allows this work's results and conclusions to have greater generality and increase the probability that they reflect how the colour invariants will perform in real-world scenarios. Most colour gradient types (except W_{INV} and $SPSS_{VAR}$) performed poorly in the Oxford dataset, which contained the typical set of imaging distortions encountered in real-world applications.

In the Middlebury tests W_{INV} was clearly the best performer, achieving nearly 100% more feature matches in the last distortion level than the second-best gradient type. W_{INV} was also the best gradient for the ALOI dataset, although the performance margins were greatly reduced. Results were much more mixed for the PHOS dataset, which saw $SPSS_{VAR}$ perform better in the majority of the distortion conditions, closely followed by W_{INV} and the grayscale intensity. The feature matching experiments thus obtained different results for each of the tested datasets, highlighting the need to test on multiple datasets and showing that the grayscale intensity in fact, exhibits adequate robustness to illumination variations. With regards to the object recognition study, the invariants were tested on a large dataset of 9,963 real-world images, which are particularly challenging for colour-based approaches since objects of the same class for the most part share their physical shape but not their colour. Despite the dataset being biased to grayscale-based techniques, in the individual recognition tests colour proved that it can have a significant impact on recognition, since 40% of the classes obtained the best results with a colour feature extraction.

6.1.4 Colour Correlation

Since the experimental results strongly indicated that the majority of the colour invariants were not optimal to be used individually in general real-world scenarios, an investigation was performed to determine if colour could still contribute to the feature extraction process in a complementary capacity. This investigation comprised of a correlation analysis on the local features that were detected from all the different gradient types. The analysis quantified the similarity between the gradients by identifying how many of the extracted features were shared amongst the gradient types, and how many features were unique only to one gradient type. Experiments showed that each detector located substantial numbers of correct unique HL points, which if used conjointly; would significantly improve the results of the feature correspondence. Furthermore, the data indicated that the colour invariants were largely uncorrelated with the grayscale intensity, the invariant with the biggest correlation was W_{INV} with approximately 25%, and the variant $SPSS_{VAR}$ showed a correlation of 52%. The performed correlation and unique features study is a novel idea within the local image feature extraction domain, and has identified that colour can indeed be used to compliment and enhance the grayscale-based feature matching process.

6.1.5 Fusion for Feature Matching

The potential for using grayscale and colour gradients conjointly for feature matching applications, was tested in this research with a fusion technique focusing on selecting the best subset of HL points extracted with multiple gradient types. The metric used for the ranking of the points was the Harris cornerness energy, which quantifies the strength of the detected corners. The goal of the testing was to be able to locate repeatable HL points (those that would appear at the same scene location across varying imaging conditions). Despite the significant number of repeatable points available to be selected among the different gradient types, the proposed fusion techniques were not able to consistently select them. This prompted a study on the suitability of the Harris energy for the selection of the optimum points. Results indicated that generally, the repeatability rates of the points did increase if they had higher Harris energies, but the best probability of selecting a repeatable point (at the

highest range of Harris energies) was seen to be only around 50% on average. Therefore despite the general apparent trend that higher Harris energies provide more repeatable points, the ranking metric proved to be too imprecise to achieve consistent optimal fusion and is thus unsuitable to use on its own in fusion techniques.

Another element that destabilised the fusion, is that W_{INV} behaved differently to the other gradient types as there was a lot more uniformity in the performance achieved by different ranges of Harris energies. The ranking metric was therefore more unsuitable for W_{INV}, which is particularly problematic since it was the best performer from the colour invariants and the best candidate for fusion alongside I and $SPSS_{VAR}$. Similar conclusions were seen when the same ranking study was performed on the LoG response, which is a metric that can be used to rank SIFT features. The performed analysis on the ranking of the features has never been documented before in the literature, and it provides an important insight into the theoretical limitations of HL and LoG feature detection, which will aid future developments of the field.

6.1.6 Fusion for Object Recognition

Sparse and random BOVW colour feature fusion extraction techniques were proposed in Section 5.5 and evaluated on the PASCAL VOC recognition challenge. The fusion is based on an early-fusion methodology, and relies primarily on the *K*-Means algorithm to select an appropriate grayscale/colour hybrid visual vocabulary. The fusion therefore does not rely on the ranking of HL points using the Harris energy metric, and as a consequence the recognition results were consistently better with the colour fusion approaches than when using only the grayscale information for the feature extraction process. The preferred technique that was tested is $I + W_{INV}\ DENSE$, which is a random feature extraction utilising equal numbers of grayscale intensity and W_{INV} descriptors. The sparse and random fusion techniques achieved comparable results for the lowest number of encoding points, but $I + W_{INV}\ DENSE$ is not limited to any maximum number of points and was thus able to achieve better recognition rates above 7,500 points. Despite the positive results from the fusion, the overall analysis indicates that the employed fusion strategy has substantial scope for

improvement, as currently the *K*-Means clustering on its own is not sufficient to optimally fuse multiple gradient types. There is not enough control to optimally and automatically determine which descriptors should be considered during the clustering and which ones will cause a detrimental effect to the BOVW recognition.

6.2 Directions for Future Research

There are two main topics of this work that are worthy of extension and further investigation. These are focused on obtaining an improved strategy for fusing the colour and grayscale information for the two computer vision tasks that this work was evaluated on. The first direction that can be taken would address the ranking of HL points in order to obtain an appropriate fusion for local feature matching applications. The second direction would concentrate on an optimal colour feature fusion extraction for image recognition applications.

In regards to the first research direction, various ideas exist in previous studies that can be employed in the development of a better ranking strategy for HL points. Comer and Draper (2009) study the repeatability rates of Harris-Laplace points by ranking optimal subsets of points using the determinant and the first and second eigenvalues of the structure tensor (Equation 3.1). In another study, Lemuz-López and Estrada (2008) rank corner points using the angular difference between dominant edges, they introduce a new ranking metric by weighing the Harris energy with the angular difference.

Another type of metric that can be used to quantify the information content of local features relates to the saliency of the extracted regions. Kadir et al. (2004) propose a saliency-based detector based on calculating the entropy of the local regions. Other saliency measures utilise the property of local image jets (Schmid and Mohr, 1997, Montesinos et al., 1998), which are vectors containing different types of grayscale or colour information for each pixel location that include; colour values, image derivatives and their combinations. The saliency of the region can be said to be inversely proportional to the probability of the occurrence of the local-jets. Therefore if a region contains local-jet information that is rarely found anywhere else on the image, it will have a high level of saliency.

Elements from all the aforementioned studies could be utilised and tested in the development of a suitable HL point ranking approach, that would aim to improve the probability of selecting repeatable features using grayscale and different types of colour gradients. From all the experience gathered throughout this research on the implementation and characteristics of the Harris-Laplace detector, the author foresees particular challenges in developing a ranking metric that is sufficiently robust to scale variations. As mentioned in Chapter 3, the scale-adapted Harris corner is not as suitable for scale-invariance as the LoG response. This fact may decrease the likelihood of success for developing a ranking approach using elements of the structure tensor. The best strategy to pursue, is thus seen to involve a colour saliency measure that is able to achieve stability in the scale-space domain.

A problematic nature of colour saliency approaches however, relates to their reliance on certain amounts of global image information, i.e. statistical information from the whole or large parts of the image must be gathered in order to assign a local region with a saliency measure. Such an approach would thus to a certain extent, be contrary to the benefits of local features that should be located only with local information in order for them to be robust to different scales and geometric distortions. The reduction of the local nature of the extracted features would not be the only obstacle, as the colour saliency measure must also be stable and robust to varying imaging conditions and it is unclear how colour statistics would behave in those circumstances. An optimised ranking of colour points will therefore not be a trivial task to accomplish.

As to the future development of the BOVW feature extraction fusion technique, an interesting aspect of the work to focus on would be mitigating the way that each object class responds better to a specific gradient type. In the proposed BOVW fusion techniques each gradient type makes an equal contribution of descriptors to the fusion, this may work well as a proof of concept but it is worth revisiting to find out if a more adaptive approach can maximise the benefits of the fusion. One way to address the issue of a sub-optimal visual vocabulary, is to examine the generation of a specific vocabulary for each object class, as was done in the study of Fernando et al. (2012). During the vocabulary generation of each

class, a set of weights could also be obtained to quantify the level of suitability of each gradient type for the particular object class. The individual weights can be tuned similarly to an optimisation problem that employs a Gradient-Descent or Levenberg-Marquardt algorithm. The tuning for example, can involve a process of obtaining a visual vocabulary and evaluating it on a trained dataset; and iteratively vary the weights and repeat the process until the weights ensure the maximum precision performance for each class of the dataset. Apart from adaptively changing the relative contribution that each gradient type has on the generation of the visual vocabulary and the image encoding, another area that should be examined further is the integration of randomly sampled descriptors and the descriptors extracted with the sparse HL detector. Table 5.3 showed that 55% of the classes were better classified with a sparse fusion technique, and 45% were more suited to a random sampling approach.

Therefore the experimental evidence clearly indicates that both strategies should be used together in order to maximise the overall recognition performance. The author recommends that the aforementioned fusion approach, would be more suitable to general recognition applications that do not have prior knowledge of the objects that need to be identified. Such applications include robotic navigation, where the SLAM algorithm must be able to recognise if the robot has returned to a previous location and adjust the map information accordingly (loop-closure). In conclusion, the further developments that can be made to the colour feature fusion of this work, would largely depart from the computer vision discipline, as the most apt tools to solve the fusion problem involves machine learning and data analytics.

6.3 Concluding Remarks

The overall goal of this work was to investigate if colour could increase the discriminative capabilities of local image features, and induce a gain in performance for applications that rely on local feature extraction. This proved to be a considerable challenge with potentially major implications for the field since the use of local features has become so ubiquitous in computer vision. The research began with uncertainties as to which direction to pursue, and what type of colour information to consider for the creation of the features. Ultimately

the work centred around the evaluation of the most prominent colour invariant gradients found in the literature, and testing their suitability for general real-world scenarios by evaluating them on two important applications; image feature matching and object recognition. The general aims of this research were stated in Chapter 1 to be:

1. Finding out why colour was not incorporated in the most popular local feature techniques.
2. Exploring how colour could be used for local feature extraction.
3. Determining what benefits, if any, colour information would provide.

These aims have been successfully accomplished during the course of this research, via the conclusive answers that were obtained from the evaluation of the colour invariant gradients. With regards to aim no. 1, the experimental results clearly show that the majority of the colour invariant features perform substantially worse than the grayscale intensity features, for both feature matching and recognition. Grayscale information proved to be sufficiently robust even under illumination variations, and overall it was evident to see why the use of colour has not been more prominent in state of the art local feature extraction approaches. The implication to the field that arises from these findings, is in knowing with greater certainty which colour invariants should be considered in any future investigation related to either feature matching or recognition.

With regards to aim no. 2, two aspects of the experimental results provide answers to the question of how can colour still be used to enhance local features despite its unsuitability in general scenarios. The first aspect of the results is the correlation study and unique detection correspondence analysis that was presented in Chapter 4. Those results strongly indicate that the colour gradients are uncorrelated and have the capacity to detect substantial numbers of repeatable HL points that are unique with respect to each other. The colour gradients can therefore be used in conjunction with grayscale intensity to extract a more distinct and robust set of local features from an image. This is a significant contribution since the concept of applying a feature extraction fusion approach to image feature detection and matching applications is proposed here for the first time.

The second aspect of the results that addresses aim no. 2, is the object recognition results from Chapter 5. The concept of using colour and grayscale descriptors for recognition has already been studied in the literature, however the specific implication of the recognition study of this work is to show that the tested colour invariants also improve the recognition results of a BOVW pipeline. Since many of the colour invariants are evaluated here for the first time, their suitability for object recognition was not previously known. The data from both the individual and the fusion recognition experiments comprehensively prove that the recognition results can be improved, by employing a fusion extraction strategy utilising multiple types of colour invariants.

Finally, aim no. 3 has been for the most part answered by the experimental results of this thesis, although the full extent of colour's benefit could not be ascertained by the proposed fusion techniques. The fusion would have to be adequately optimised in order to consistently harness all the advantages of using colour; which were apparent in certain conditions of the evaluation. When using the colour gradients individually, W_{INV} proved to be the best candidate in feature detection and matching under illumination conditions, and obtained the best precision results for 20% of the classes of the recognition challenge. $SPSS_{VAR}$ was the preferred gradient for 15% of the recognition challenge and performed marginally better for feature matching under general imaging conditions. From the unique correspondence analysis of Section 4.4.1, results indicated potential improvements in detection results of 97-109% (under general imaging conditions) if a fusion strategy would be utilised, and improvements of 240-252% for scenarios under varying illumination conditions.

In the case of the BOVW fusion for recognition, it is harder to estimate the potential improvements that a fully optimised fusion technique would achieve. From the results presented in Table 5.3 however, the indicative maximum mAP that could be obtained by taking the best precision results per class, amounts to 35.00 which is 4.81% higher than $I + W_{INV}\ DENSE$ and 14.75% higher than $I\ DENSE$. These gains are substantial in the object recognition field where each year the state of the art results improve by only a few percentages. These research aims were able to be achieved by addressing the 7 main limitations of the prior art, which were stated in Section 2.6:

1 - Lack of Scale-Invariance

This limitation of some previous studies was addressed in this work by using the colour gradients for the Harris-Laplace detector, that extracts scale-invariant local interest regions which are also robust to a limited range of other imaging conditions such as varying rotation and viewpoint.

2 - Limited Distortions

The evaluation carried out in this research subjected the colour features to the same set of varying image conditions as the state of the art grayscale-based feature studies. This was done using the Oxford dataset, as opposed to other colour studies that were tested on datasets that only had images with varying illumination conditions.

3 - Sub-optimal Evaluation Framework

By using the robust metrics and testing framework of Mikolajczyk and Schmid (2005) and conducting the VOC recognition challenge, the evaluation carried out in this research was able to simulate more realistic scenarios and determine how the colour features would perform in more general real-world applications.

4 - Insufficient Data

The rigour of the evaluation of this work was further strengthened, by utilising datasets that contained sufficient number of images in order to obtain more statistically significant findings.

5 - Colour-biased Datasets

All the datasets used in this thesis for both the image matching and recognition experiments, did not specifically favour colour-based approaches. They were picked because they contained a large range of real-world objects and scenes that would provide a good representation of the conditions found in a natural real-life environment.

6 - Few Datasets

An additional aspect of ensuring that the statistical significance of the results would be maximised, is the use of multiple datasets for the image matching experiments. This ensured that the evaluation tested the colour features on images of varying quality and which were acquired with different camera hardware.

7 - Colour Detection with Description

This research used colour throughout the whole feature extraction process to detect and describe the local features with the same colour invariants. This made it possible to evaluate the suitability of the invariants separately for detection and description purposes. Additionally, in the recognition fusion experiments, using different colour gradients to detect local features meant that more varieties of features were able to be extracted than if only grayscale would have been used for the detection. This resulted in various sparse fusion techniques being able to obtain the best recognition for individual VOC classes.

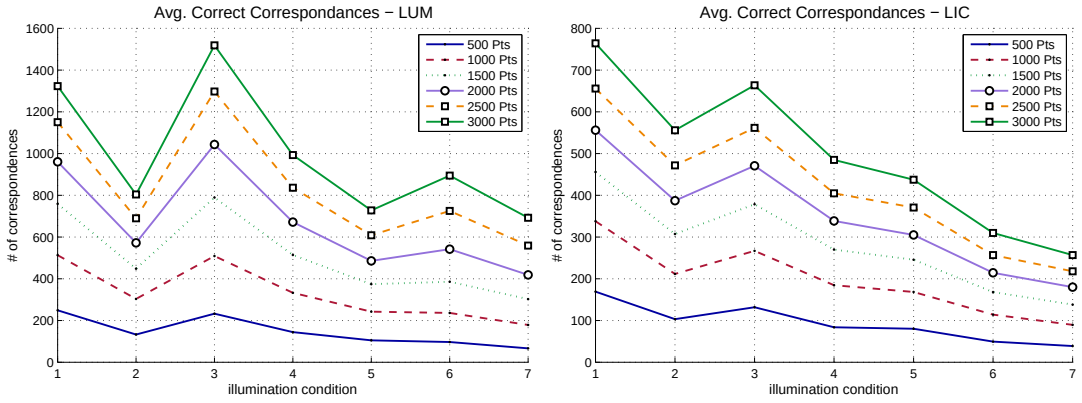
The answers to the aforementioned three general aims of the research, constitute the major contributions of this thesis to the computer vision field. The work has established that colour should generally only be used in a complimentary capacity to enhance the performance of grayscale features. This work proposes that this complementarity be accomplished via a colour feature fusion extraction approach, and advocates for future works to continue developing the integration of grayscale intensity and the best colour invariants that were tested. This recommendation is backed up by promising results which are able to prove that with the chosen colour invariants, a fusion approach improves the overall performance of object recognition applications. Furthermore, there is strong evidence to indicate that there is sufficient low-level image information in both image feature matching and recognition applications, that could be utilised by an optimum fusion approach. Such an approach would then be capable of consistently improving results under all types of object classes or imaging distortions.

The immediate practical implication of this research can be summarised as: For general feature matching applications, the local image feature implementation should utilise the $SPSS_{VAR}$ colour gradient. When the application requires robustness to varying illumination conditions, then the W_{INV} gradient should be used to construct the image features. Finally for object class recognition applications, this research shows that the evaluated colour invariants have the potential to significantly improve the performance of Bag-of-Visual-Words recognition techniques, although the optimal fusion between the grayscale and colour gradients remains an open problem.

Appendix A

This appendix contains supplementary results from the local feature detection experiments of Section 4.2. The presented results of Figures A.1, A.2, A.3 and A.4 show the number of correct correspondences and repeatability rates achieved by extracting varying numbers of HL points from the Middlebury dataset. More results showing the effect on the repeatability rates by varying the number of extracted points on the Oxford dataset are presented in Figure A.6. Figures A.7 and A.8 show the standard deviations of the detection results presented in Section 4.2.

(a) # corresp. I (b) # corresp. LIC



(c) % repeat. I (d) % repeat. LIC

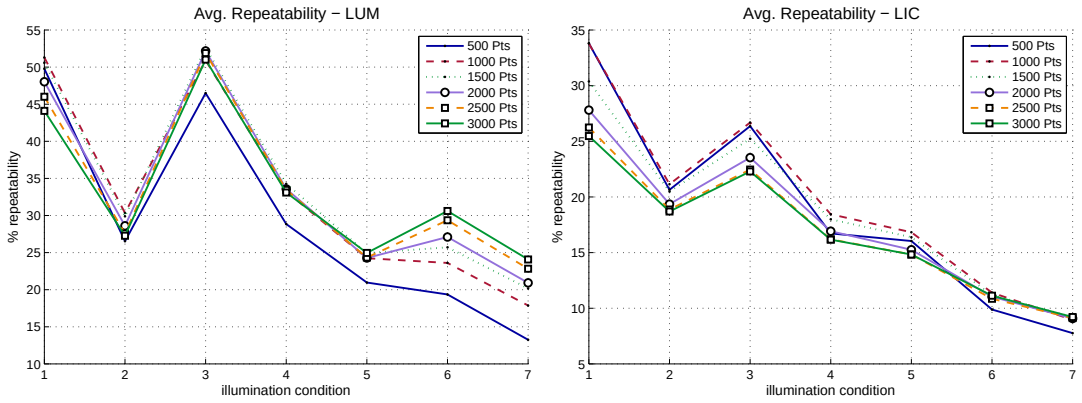
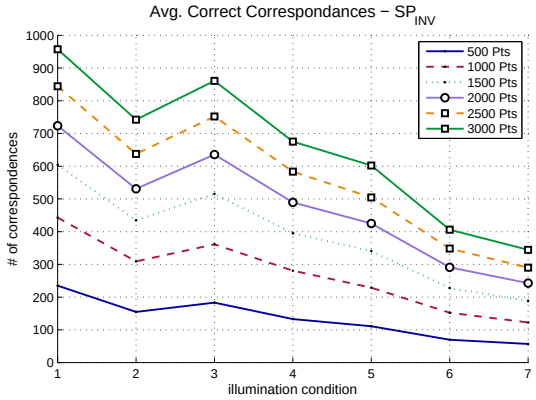
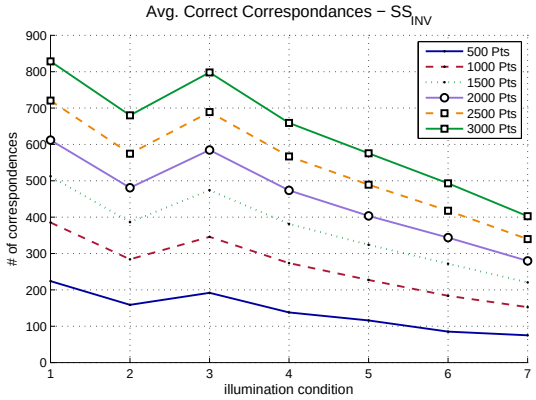


Figure A.1: Detection correspondences and repeatability results, varying the number of extracted HL points on the Middlebury dataset.

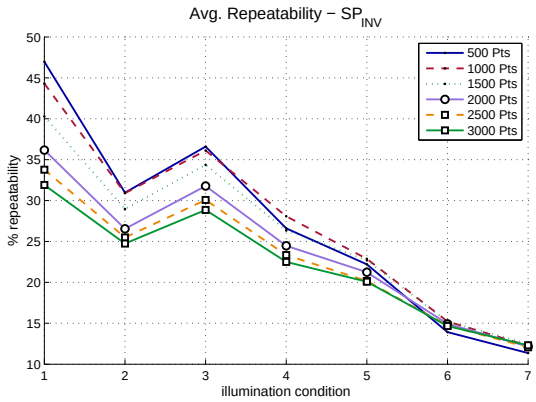
(a) # corresp. SP_{INV}



(b) # corresp. SS_{INV}



(c) % repeat. SP_{INV}



(d) % repeat. SS_{INV}

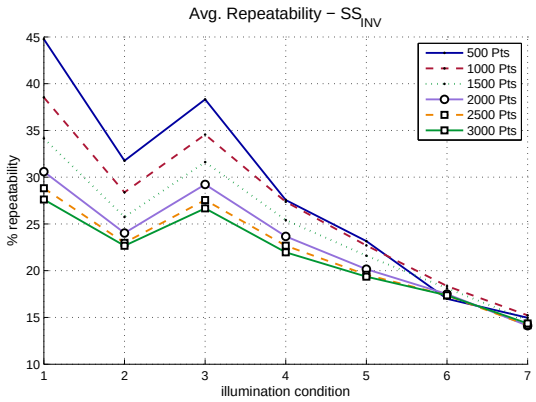
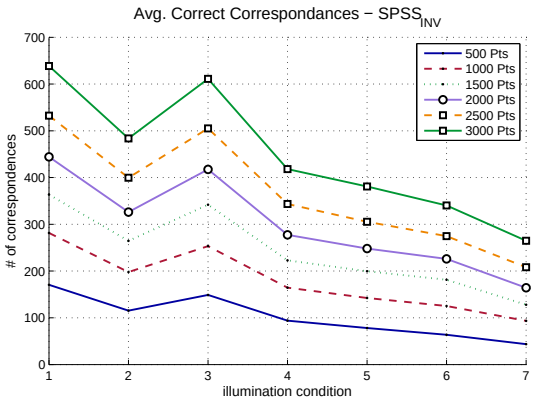
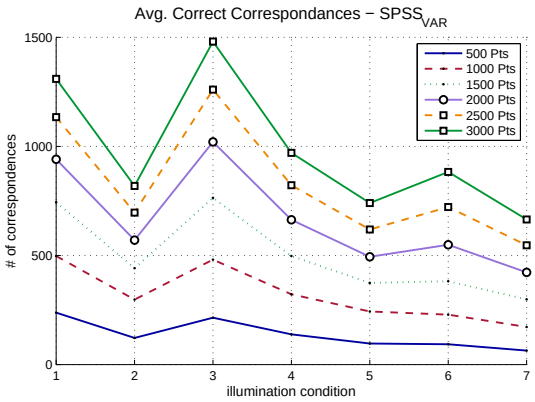


Figure A.2: Detection correspondences and repeatability results, varying the number of extracted HL points on the Middlebury dataset.

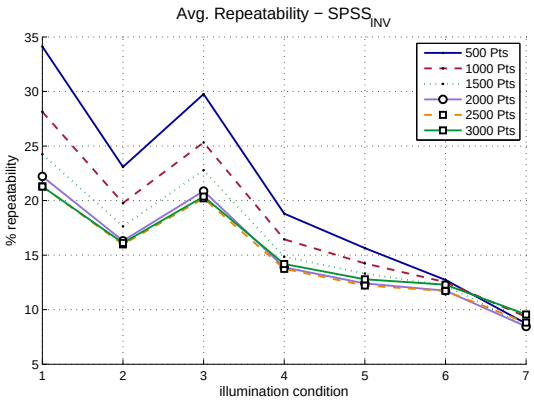
(a) # corresp. $SPSS_{INV}$



(b) # corresp. $SPSS_{VAR}$



(c) % repeat. $SPSS_{INV}$



(d) % repeat. $SPSS_{VAR}$

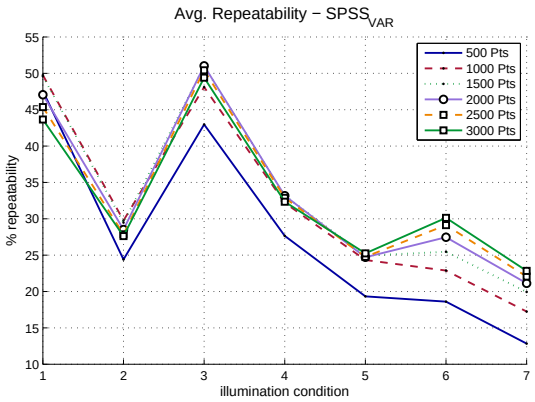
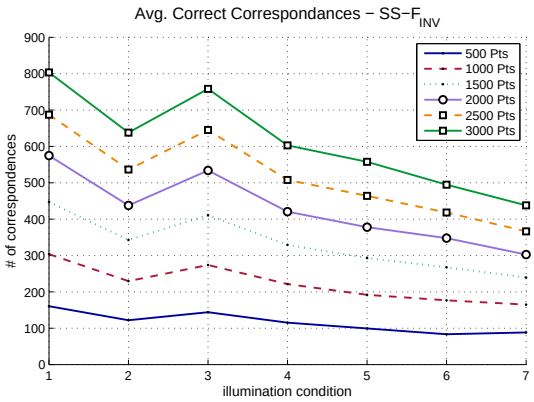
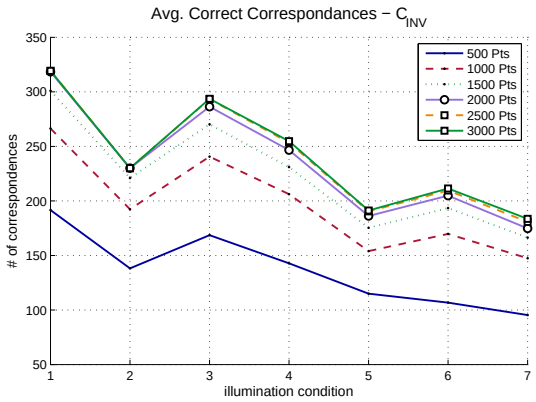


Figure A.3: Detection correspondences and repeatability results, varying the number of extracted HL points on the Middlebury dataset.

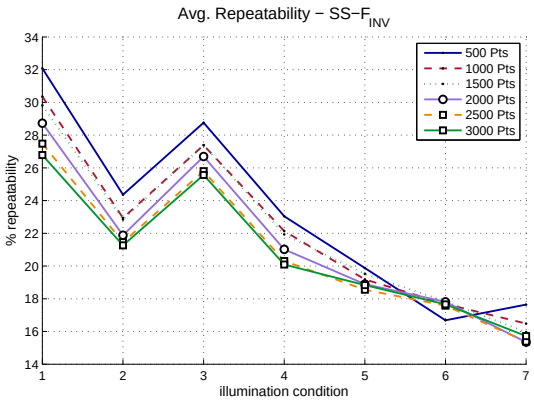
(a) # corresp. SS_{F-INV}



(b) # corresp. C_{INV}



(c) % repeat. SS_{F-INV}



(d) % repeat. C_{INV}

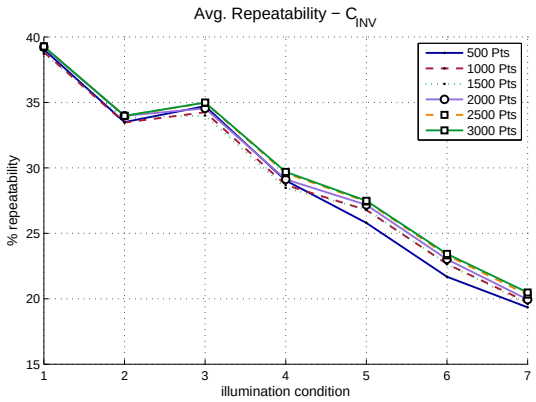
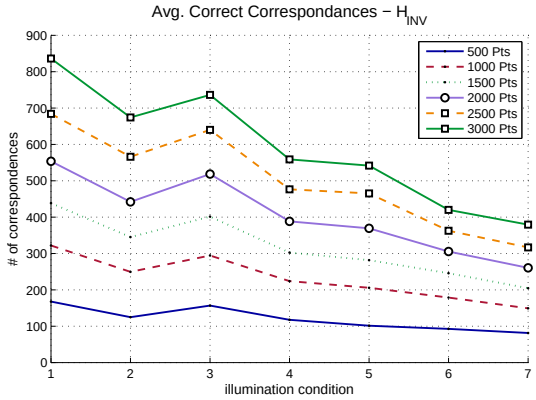
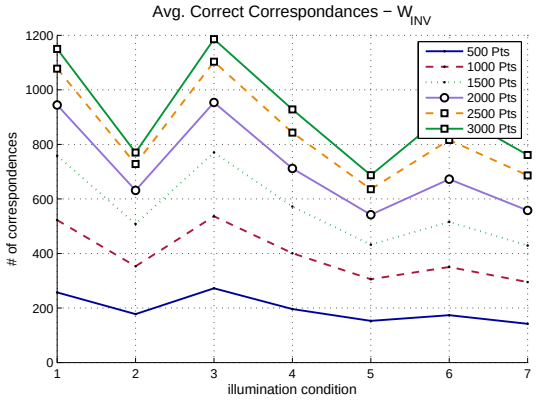


Figure A.4: Detection correspondences and repeatability results, varying the number of extracted HL points on the Middlebury dataset.

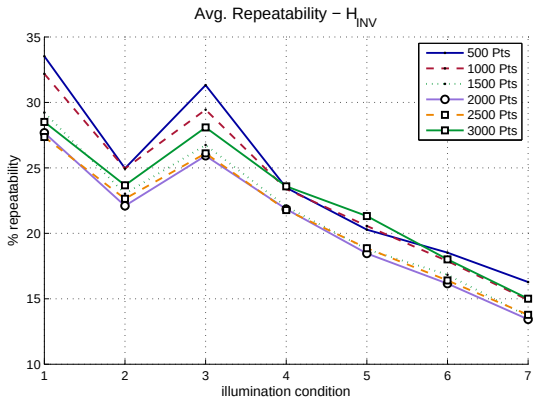
(a) # corresp. H_{INV}



(b) # corresp. W_{INV}



(c) % repeat. H_{INV}



(d) % repeat. W_{INV}

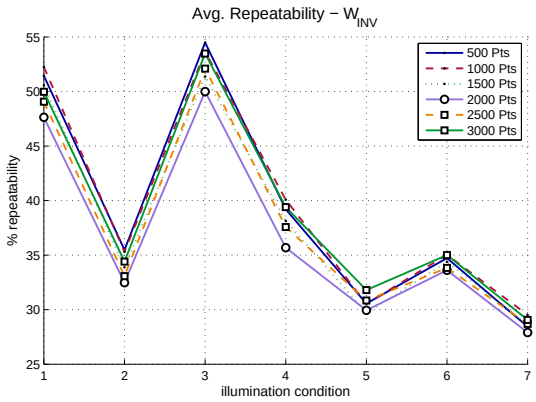


Figure A.5: Detection correspondences and repeatability results, varying the number of extracted HL points on the Middlebury dataset.

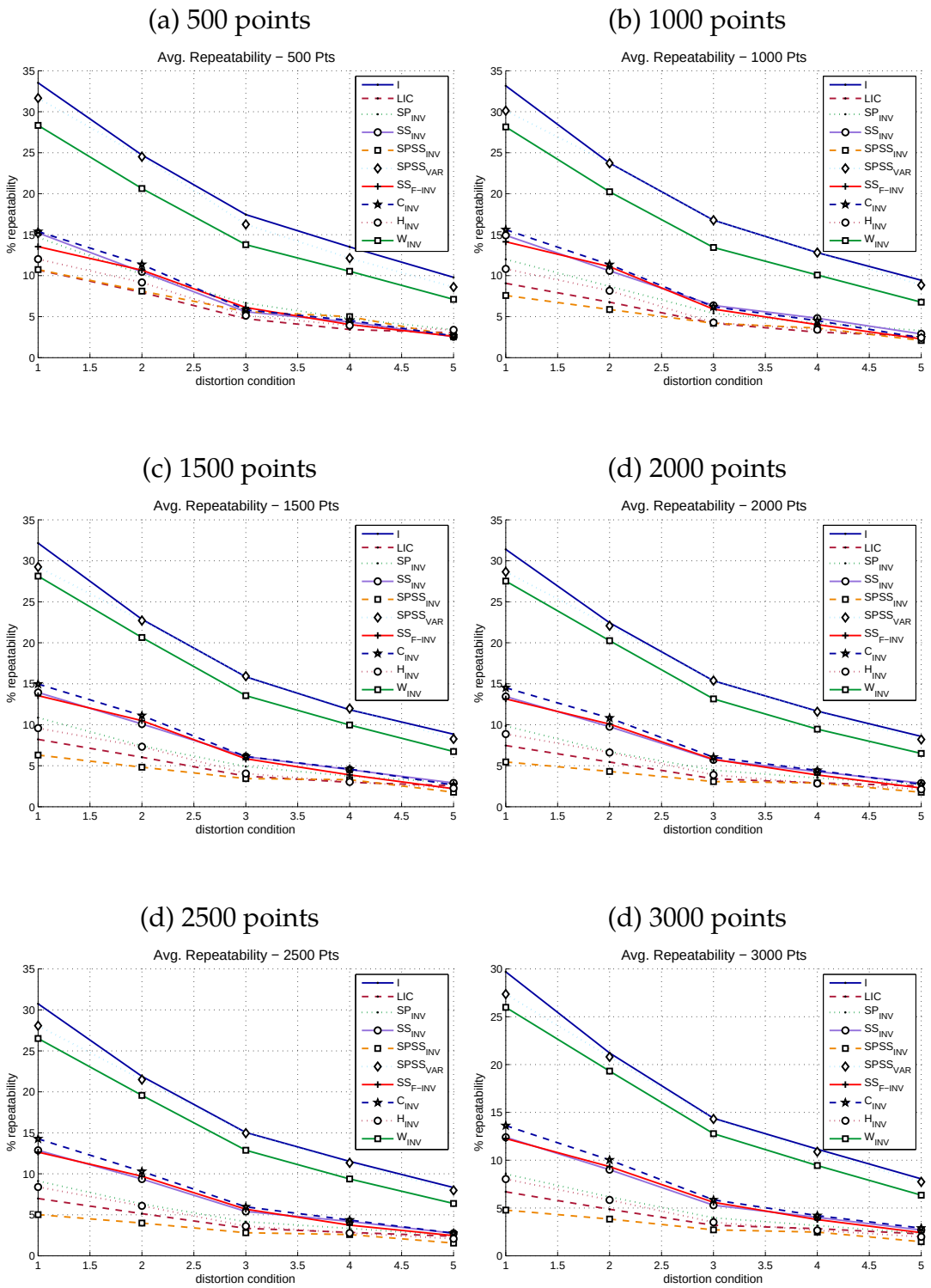
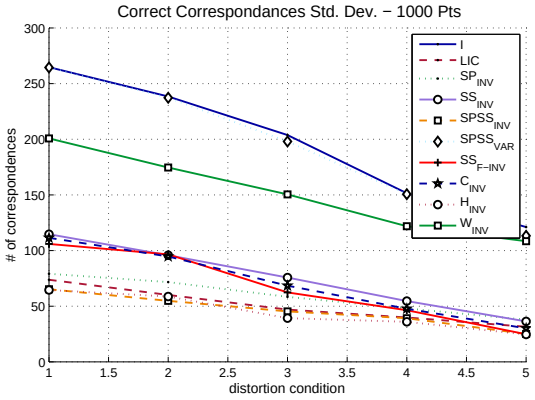
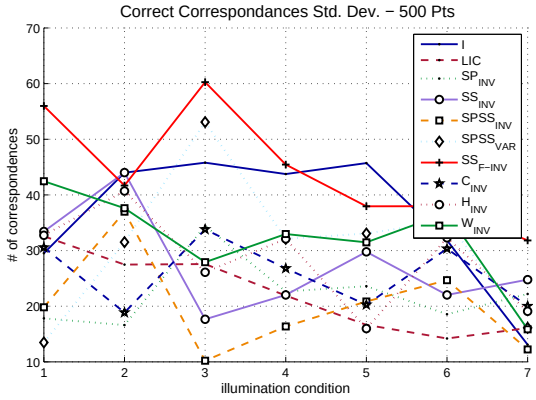


Figure A.6: Detection repeatability results, varying the number of extracted HL points on the Oxford dataset.

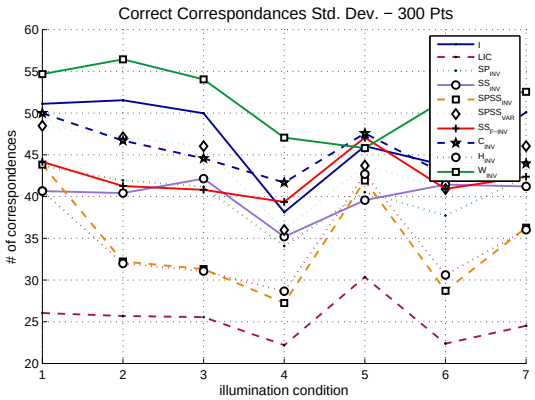
(a) Std. Dev. of # corresp. (Oxford)



(b) Std. Dev. of # corresp. (Middlebury)



(c) Std. Dev. of # corresp. (ALOI)



(d) Std. Dev. of # corresp. (PHOS)

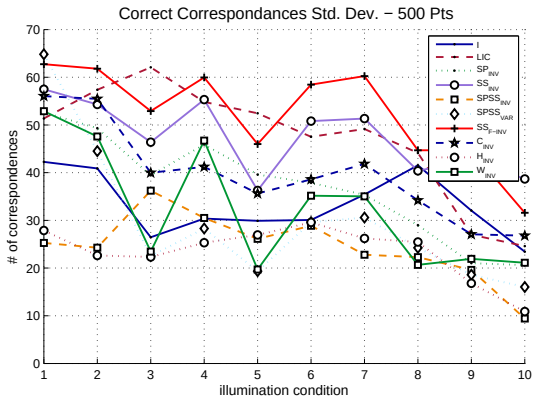
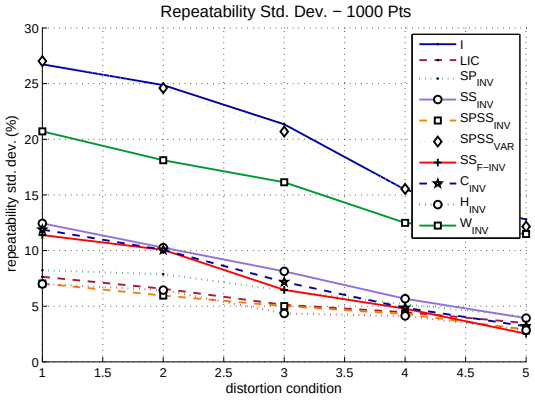
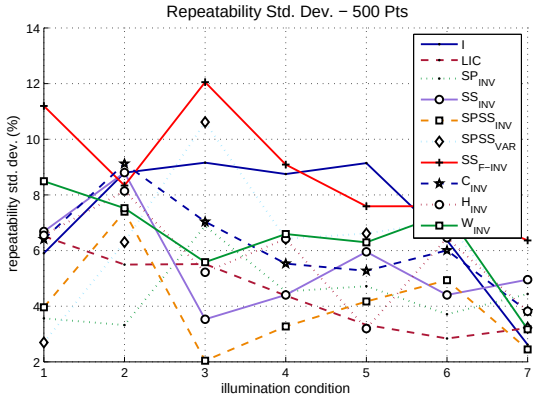


Figure A.7: Standard deviation of the detection correspondence results.

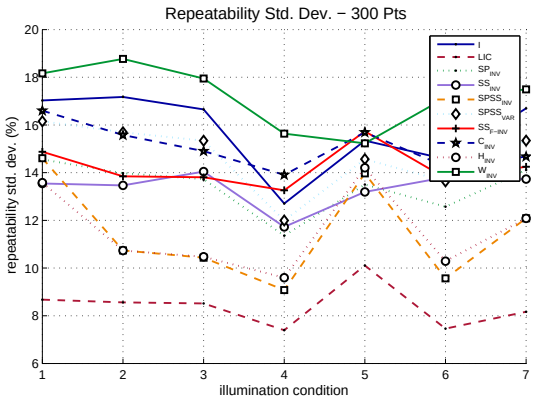
(a) Std. Dev. of % repeat. (Oxford)



(b) Std. Dev. of % repeat. (Middlebury)



(c) Std. Dev. of % repeat. (ALOI)



(d) Std. Dev. of % repeat. (PHOS)

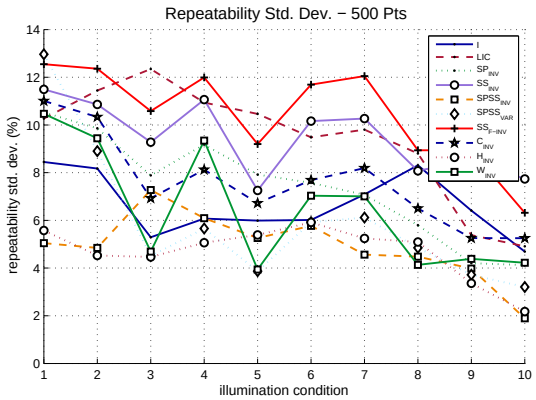
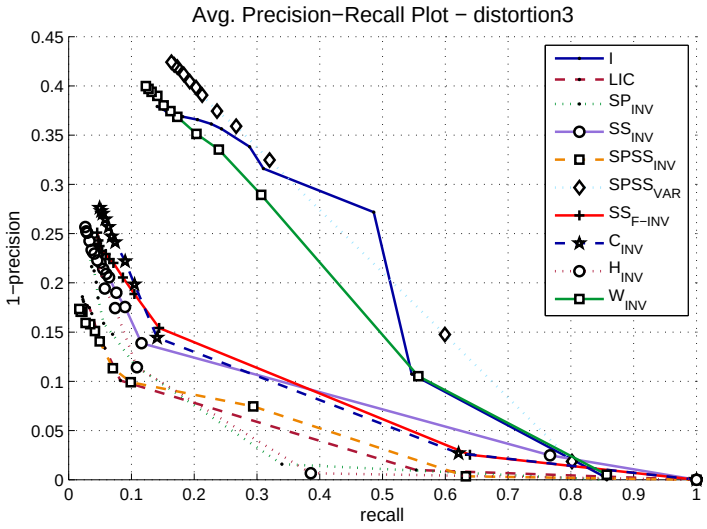


Figure A.8: Standard deviation of the repeatability results.

Appendix B

This appendix contains supplementary results from the local feature matching experiments of Section 4.3. Figures B.1, B.2 and B.3 show the precision-recall curves at two different distortion levels of the Oxford, Middlebury and ALOI datasets respectively. Each plot contains the average precision-recall data of the entire dataset, and two plots are shown for each dataset to compare the difference when the distortion level is increased (e.g. from Img-3 to Img-6 in Figure B.1).

(a) Oxford Img-3



(b) Oxford Img-6

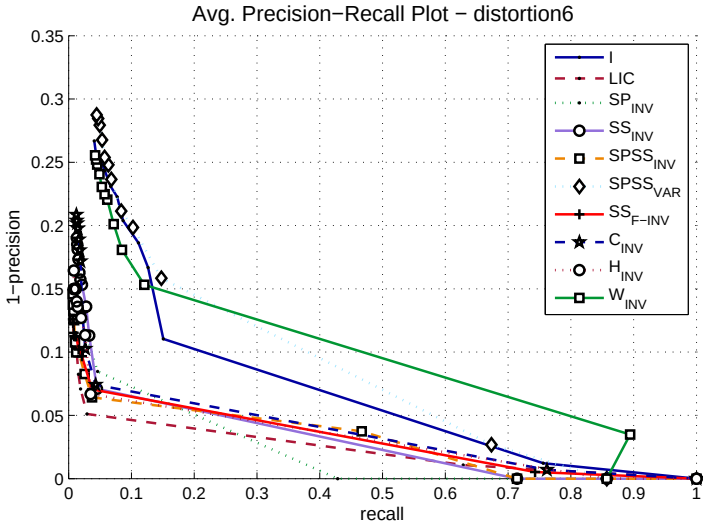
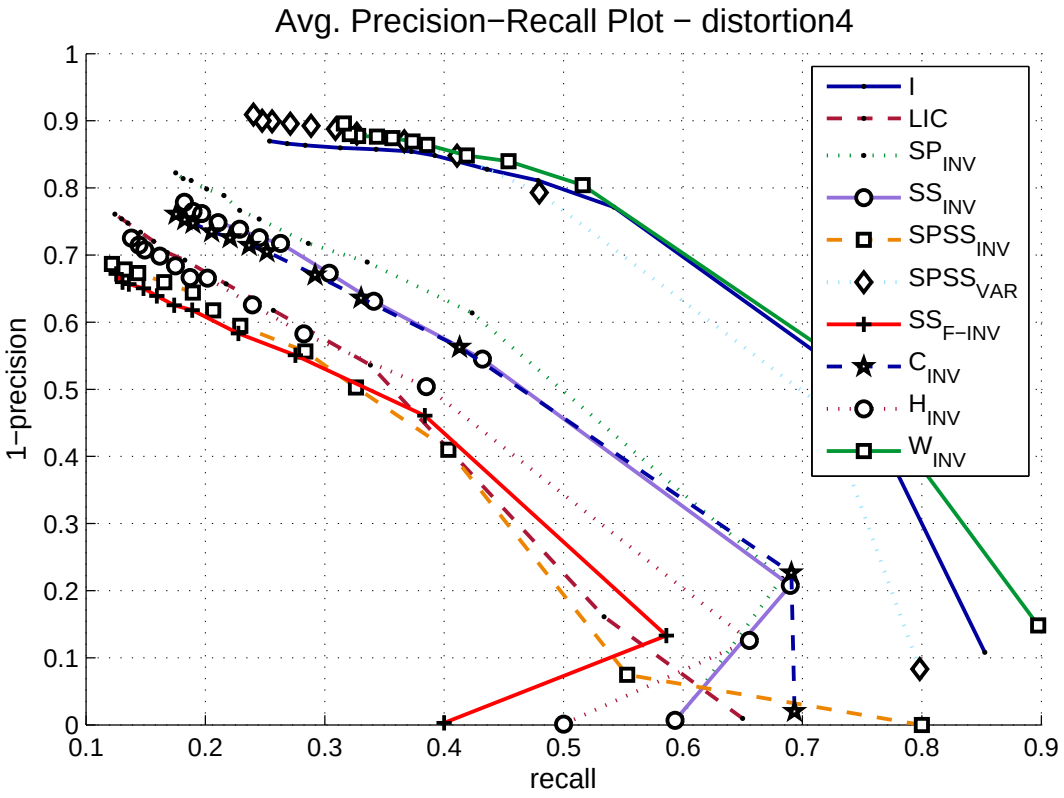


Figure B.1: Precision-Recall curves for the Oxford dataset

(a) Middlebury Img-4



(b) Middlebury Img-8

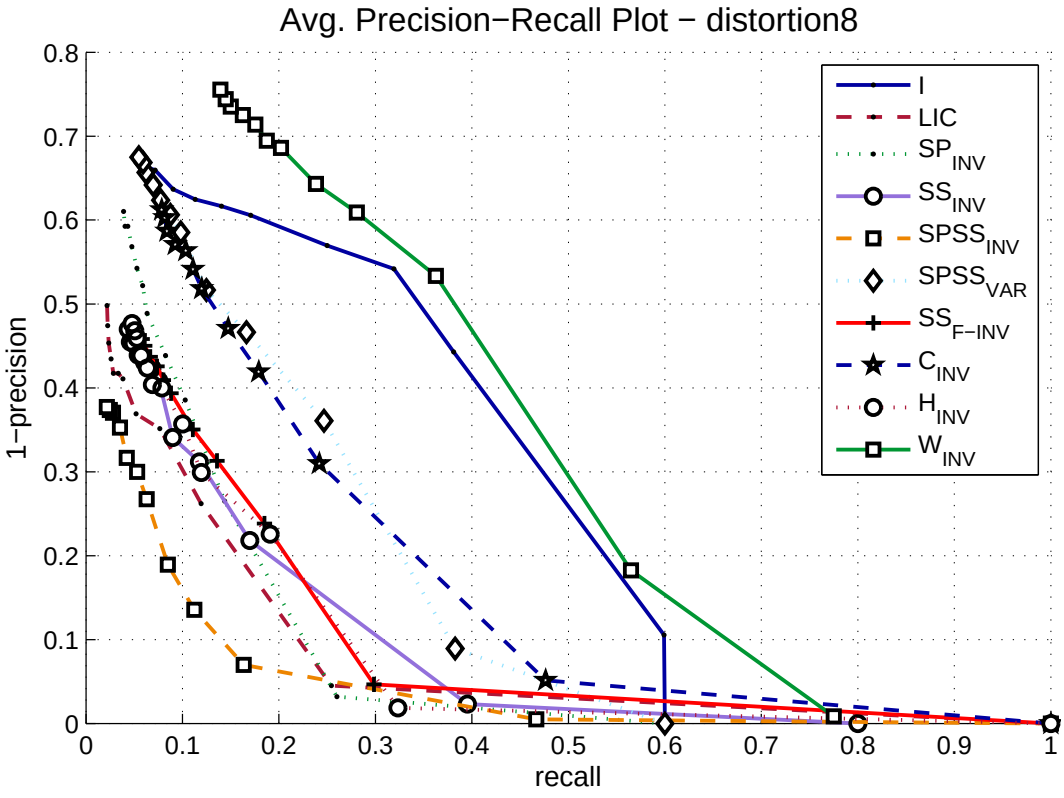
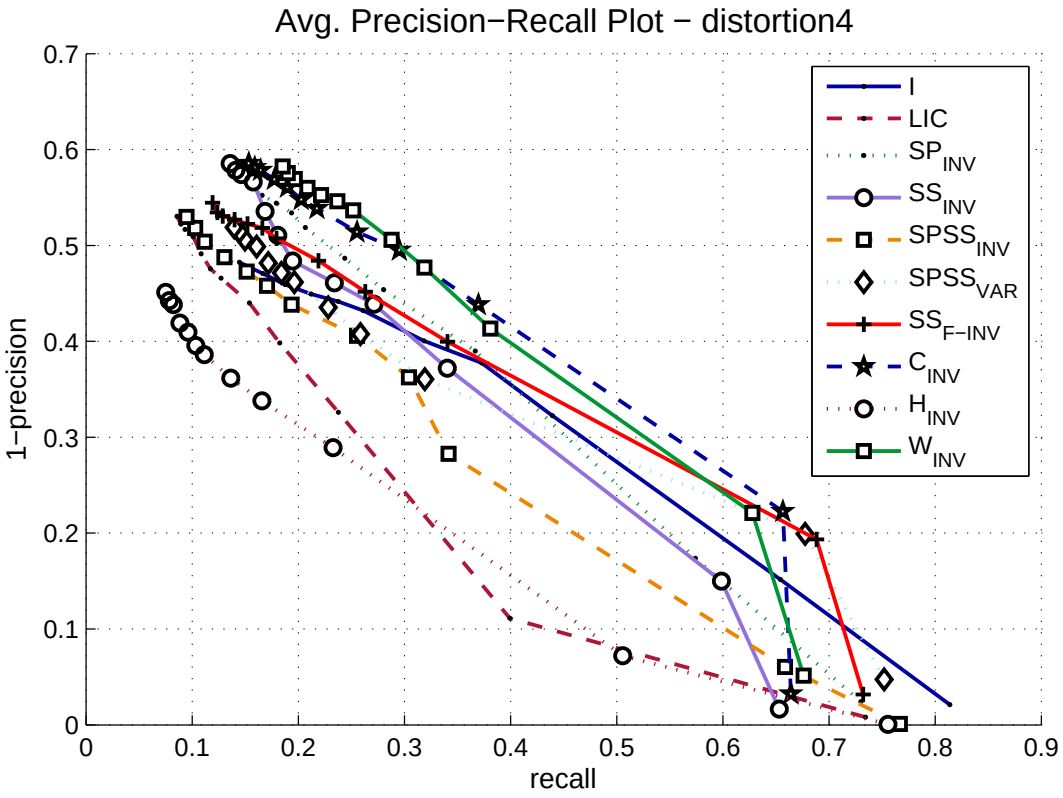


Figure B.2: Precision-Recall curves for the Middlebury dataset

(a) ALOI Img-4



(b) ALOI Img-8

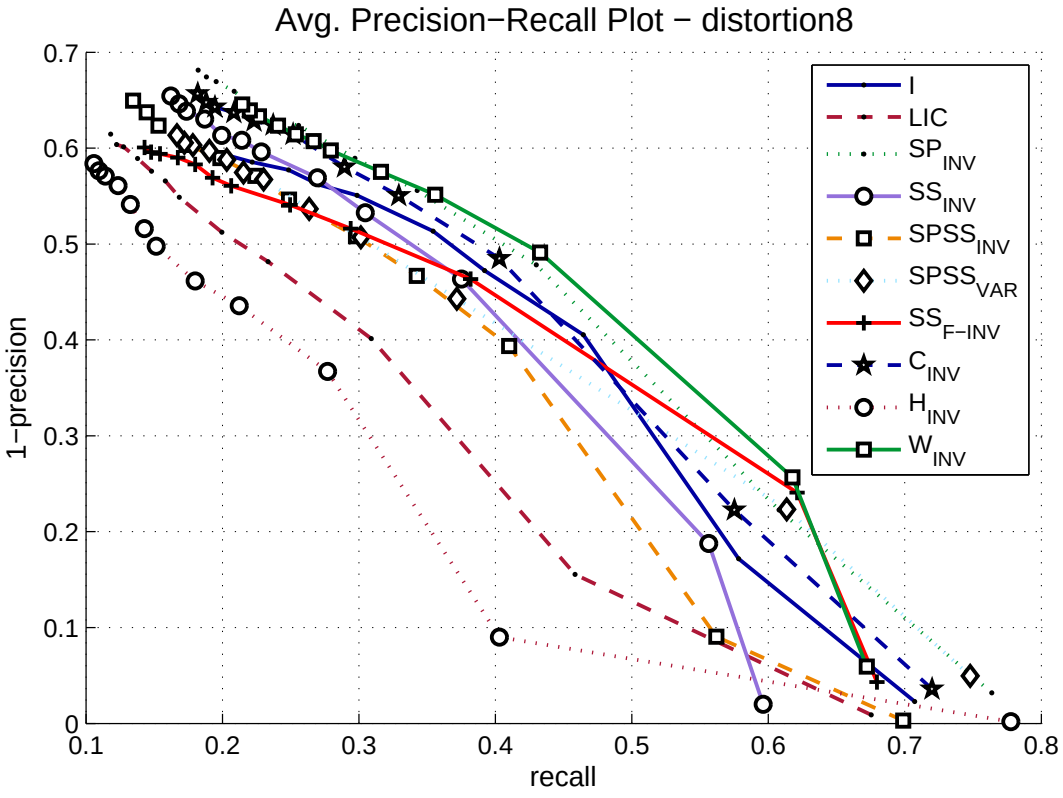
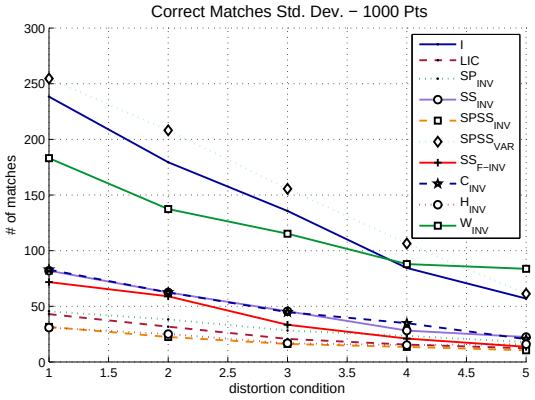


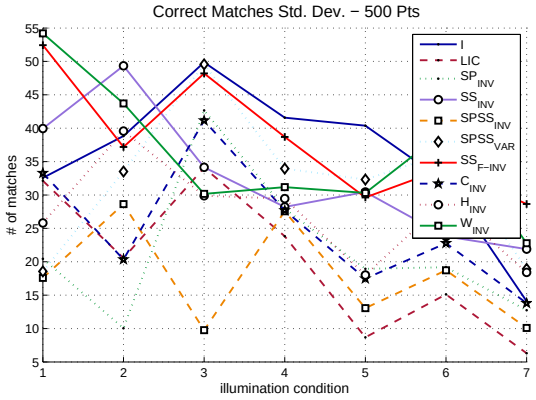
Figure B.3: Precision-Recall curves for the ALOI dataset

Figures B.4 and B.5 show the standard deviations of the matching results presented in Section 4.3.

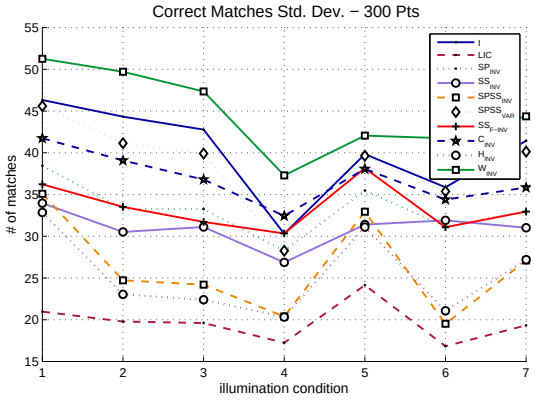
(a) Std. Dev. of # matches (Oxf.)



(b) Std. Dev. of # matches (Middl.)



(c) Std. Dev. of # matches (ALOI)



(d) Std. Dev. of # matches (PHOS)

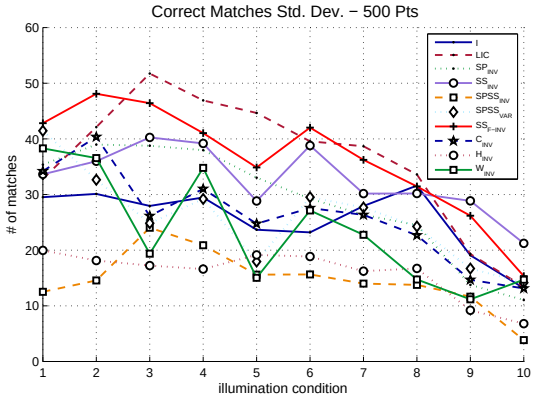
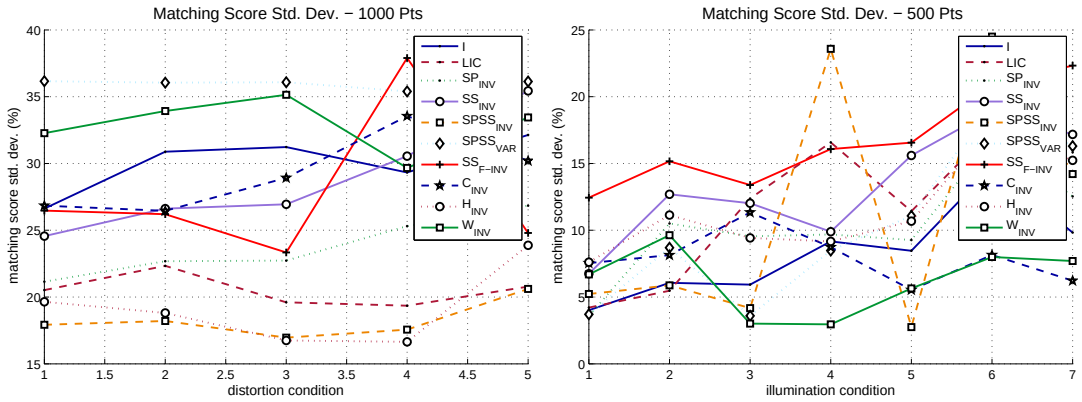


Figure B.4: Standard deviation of the number of correct matches.

(a) Std. Dev. of % match score (Oxf.) (b) Std. Dev. of % match score (Middl.)



(c) Std. Dev. of % match score (ALOI) (d) Std. Dev. of % match score (PHOS)

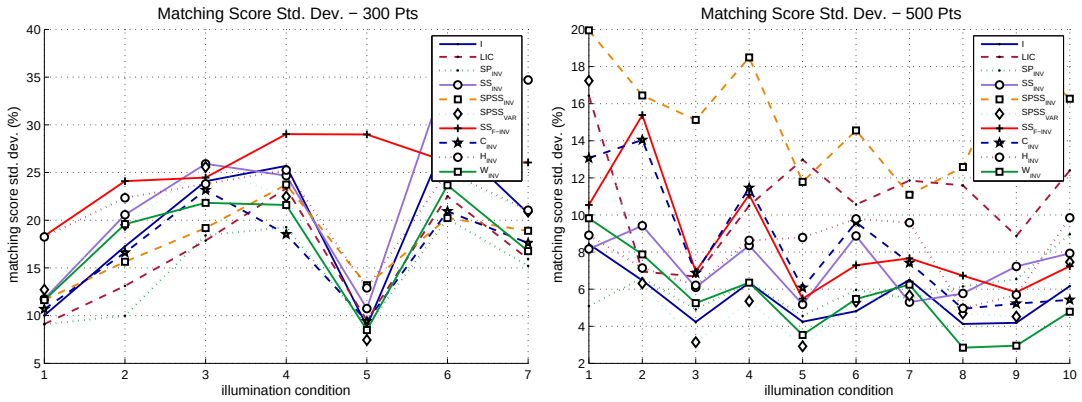
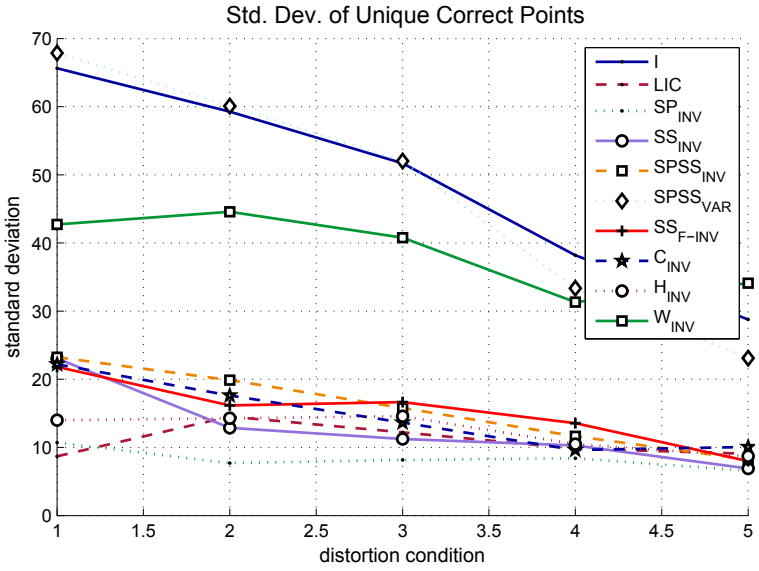


Figure B.5: Standard deviation of the matching score results.

Appendix C

This appendix shows the standard deviations of the results from the unique point correspondence experiments presented in Section 4.4.

(a) Oxford



(b) Middlebury

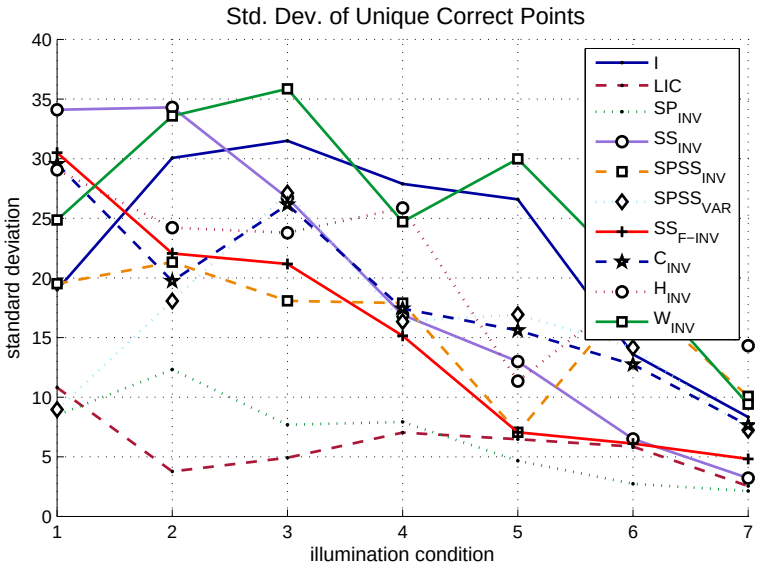
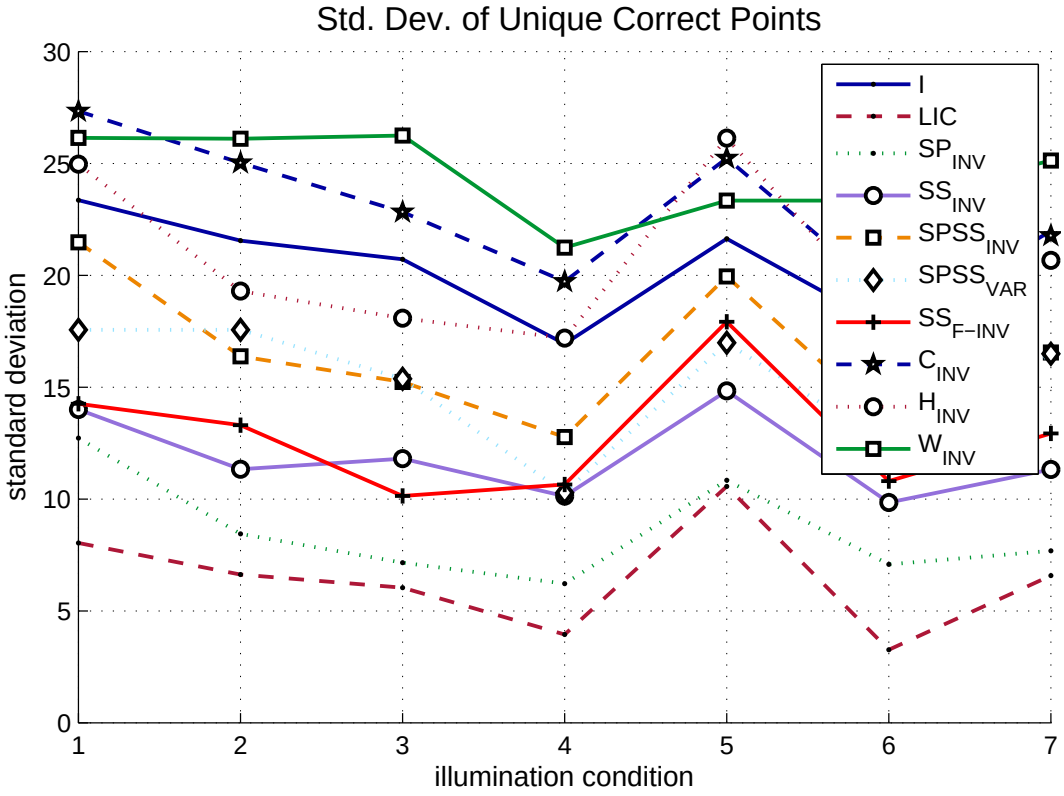


Figure C.1: Summary of the std. deviation of unique correspondences for the Oxford (a) and Middlebury (b) datasets.

(a) ALOI



(b) PHOS

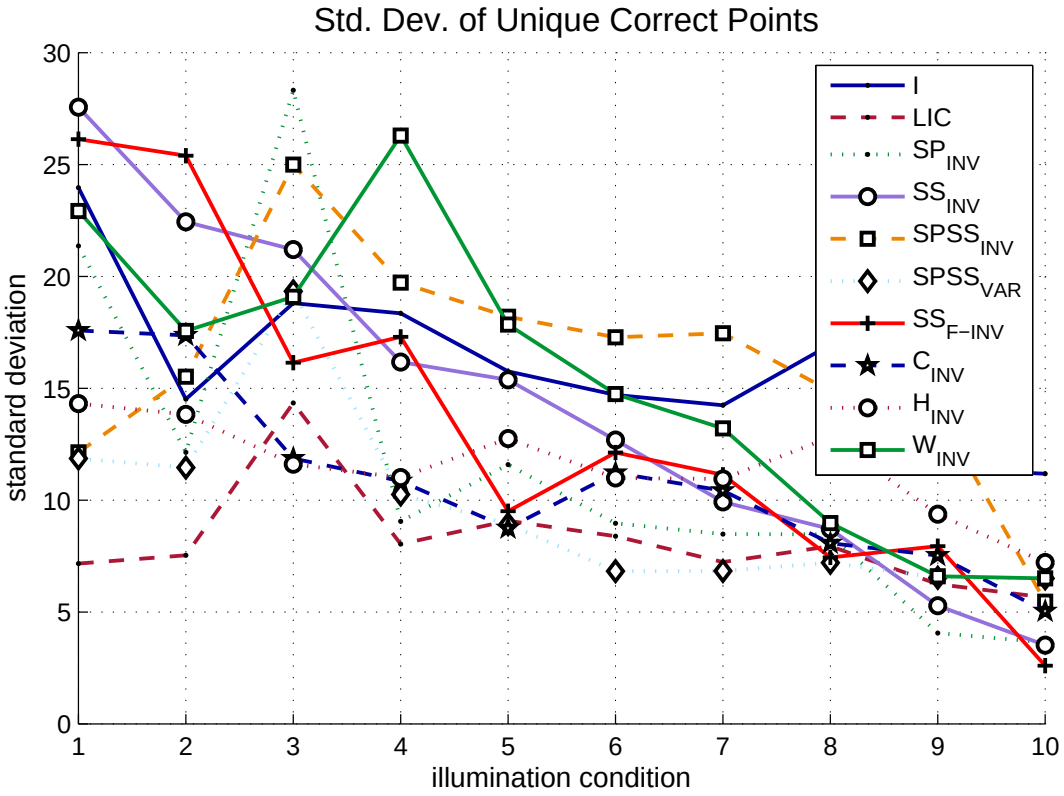


Figure C.2: Summary of the std. deviation of unique correspondences for the ALOI (a) and PHOS (b) datasets.

Bibliography

Alaa E Abdel-Hakim and Aly A Farag. Csift: A sift descriptor with color invariant characteristics. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1978–1983. IEEE, 2006. (Cited on pages 19, 31, 33, 46, 47, 97, and 122.)

Motilal Agrawal, Kurt Konolige, and Morten Blas. Censure: Center surround extremas for realtime feature detection and matching. volume 5305 of *Lecture Notes in Computer Science*, pages 102–115. Springer, 2008. (Cited on page 28.)

Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–517. IEEE, 2012. (Cited on page 29.)

Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *Computer Vision–ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 214–227. Springer, 2012. (Cited on page 28.)

R. Arandjelovic and A. Zisserman. All about vlad. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1585. IEEE, June 2013. (Cited on page 1.)

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. Elsevier. (Cited on pages 3, 21, 22, 28, and 34.)

Luca Benedetti, Massimiliano Corsini, Paolo Cignoni, Marco Callieri, and Roberto Scopigno. Color to gray conversions in the context of stereo matching algorithms. *Machine Vision and Applications*, 23(2):327–348, 2012. Springer. (Cited on page 6.)

Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. Now Publishers Inc. (Cited on pages 37 and 124.)

Marco San Biagio, Loris Bazzani, Marco Cristani, and Vittorio Murino. Weighted bag of visual words for object recognition. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 2734–2738. IEEE, 2014. (Cited on pages 1 and 41.)

Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification via plsa. In *Computer Vision–ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 517–530. Springer, 2006. (Cited on pages 30 and 127.)

Anna Bosch, Andrew Zisserman, and Xavier Muoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008. (Cited on pages 19 and 32.)

G. J. Burghouts and J-M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009. Elsevier. (Cited on pages 18, 19, 26, 31, 32, 34, 36, 45, 46, 47, 78, 97, and 122.)

Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *Computer Vision–ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 778–792. Springer, 2010. (Cited on page 29.)

Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. British Machine Vision Conference (BMVC)*, volume 2, page 8, 2011. (Cited on pages 18, 124, and 131.)

David M Chen and Bernd Girod. Memory-efficient image databases for mobile visual search. *IEEE MultiMedia*, 21(1):14–23, 2014. (Cited on page 125.)

Tao Chen, Kim-Hui Yap, and Dajiang Zhang. Discriminative soft bag-of-visual phrase for mobile landmark recognition. *IEEE Trans. on Multimedia*, 16(3):612–622, 2014. (Cited on page 125.)

Hamilton Y Chong, Steven J Gortler, and Todd Zickler. A perception-based color space for illumination-invariant image processing. *ACM Trans. on Graphics (TOG)*, 27(3):61–68, 2008. (Cited on pages 33 and 70.)

Dung Manh Chu and Arnold WM Smeulders. Color invariant surf in discriminative object tracking. In *Trends and Topics in Computer Vision*, pages 62–75. Springer, 2012. (Cited on page 19.)

H Thomson Comer and Bruce A Draper. Interest point stability prediction. In *Computer Vision Systems*, pages 315–324. Springer, 2009. (Cited on page 154.)

Jason J Corso and Gregory D Hager. Coherent regions for concise and stable image description. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 184–190. IEEE, 2005. (Cited on pages 24, 46, and 47.)

Yan Cui, Alain Pagani, and Didier Stricker. Sift in perception-based color space. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 3909–3912. IEEE, 2010. (Cited on pages 33, 46, 47, and 70.)

Aristeidis Diplaros, Theo Gevers, and Ioannis Patras. Combining color and shape information for illumination-viewpoint invariant object recognition. *IEEE Trans. on Image Processing*, 15(1):1–11, 2006. (Cited on page 30.)

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc 2007) results. `http://host.robots.ox.ac.uk/pascal/VOC/voc2007/`, 2007. Accessed: 2nd of July 2015. (Cited on pages 25, 37, 41, and 132.)

F. Faille. Stable interest point detection under illumination changes using colour invariants. In *Proc. British Machine Vision Conference (BMVC)*, pages 19–1, 2005. (Cited on pages 22, 45, and 47.)

Peng Fan, Aidong Men, Mengyang Chen, and Bo Yang. Color-surf: A surf descriptor with local kernel color histograms. In *International Conference on Network Infrastructure and Digital Content (IC-NIDC 2009)*, pages 726–730. IEEE, 2009. (Cited on pages 19 and 34.)

Basura Fernando, Elisa Fromont, Damien Muselet, and Marc Sebban. Discriminative feature fusion for image classification. In *Proc. IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 3434–3441. IEEE, 2012. (Cited on pages 41, 125, and 155.)

Graham D Finlayson, Bernt Schiele, and James L Crowley. Comprehensive colour image normalization. In *Computer Vision–ECCV'98*, volume 1406 of *Lecture Notes in Computer Science*, pages 475–490. Springer, 1998. (Cited on pages 30 and 62.)

M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. (Cited on page 3.)

P-E Forssén. Maximally stable colour regions for recognition and matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. (Cited on pages 23, 46, and 47.)

J-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *Computer Vision–ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 368–381. Springer, 2010. (Cited on page 1.)

Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995. (Cited on page 62.)

Karl R Gegenfurtner. Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, 4(7):563–572, 2003. Nature Publishing Group. (Cited on page 4.)

Karl R Gegenfurtner and Daniel C Kiper. Color vision. *Annual Review of Neuroscience*, 26(1):181, 2003. (Cited on page 4.)

Wilson S Geisler. Visual perception and the statistical properties of natural scenes. *Annual Reviews of Psychology*, 59:167–192, 2008. (Cited on page 4.)

T Geodemé, Tinne Tuytelaars, G Vanacker, M Nuttin, and Luc Van Gool. Omnidirectional sparse visual path following with occlusion-robust feature tracking. In *OMNIVIS Workshop, International Conference on Computer Vision (ICCV)*. IEEE, 2005. (Cited on page 34.)

J-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(12): 1338–1350, 2001. (Cited on pages 31, 32, 33, 34, 45, 63, 67, 68, 69, 70, 72, 78, 79, and 85.)

Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *Int. Journal of Computer Vision*, 61(1):103–112, 2005. Springer. (Cited on pages 25 and 31.)

Theo Gevers and Arnold WM Smeulders. Color based object recognition. *Pattern recognition*, 32(3):453–464, 1999. Elsevier. (Cited on pages 23, 24, and 62.)

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE Trans. on Image Processing*, 20(9): 2475–2489, 2011. (Cited on page 35.)

David Gossow, Peter Decker, and Dietrich Paulus. Extending surf to the color domain. In *Conf. on Colour in Graphics, Imaging, and Vision*, volume 5, pages 215–221. Society for Imaging Science and Technology, 2010. (Cited on pages 19, 24, 32, 46, and 122.)

V. Gouet and N. Boujemaa. Object-based queries using color points of interest. In *IEEE CVPR Workshop on Content-Based Access of Image and Video Libraries (CBAIVL)*, pages 30–36, 2001. (Cited on page 22.)

G. Griffin, AD. Holub, and P. Perona. The caltech 256. `http://www.vision.caltech.edu/Image_Datasets/Caltech256/`, 2004. Accessed: 21st of January 2015. (Cited on page 23.)

R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2nd. edition, 2004. (Cited on pages 2 and 3.)

Ewald Hering. *Outlines of a theory of the light sense*. Harvard University Press, 1964. (Cited on page 72.)

Alexandros Iosifidis, Anastastios Tefas, and Ioannis Pitas. Discriminant bag of words based representation for human action recognition. *Pattern Recognition Letters*, 49:185 – 192, 2014. Elsevier. (Cited on page 125.)

A. Irschara, C. Zach, M. Klopschitz, and H. Bischof. Large-scale, dense city reconstruction from user-contributed photos. *Computer Vision and Image Understanding*, 116(1):2 – 15, 2012. Elsevier. (Cited on page 1.)

Laurent Itti, Geraint Rees, and John K Tsotsos. *Neurobiology of attention*. Academic Press, ISBN: 9780123757319, 2005. (Cited on page 1.)

Ali Jalilvand, Hamidreza Shayegh Boroujeni, and Nasrollah Moghadam Charkari. Cwsurf: A novel coloured local invariant descriptor based on surf. In *Proc. International eConference on Computer and Knowledge Engineering (ICCKE)*, pages 214–219. IEEE, 2011. (Cited on pages 19, 34, 46, 47, and 122.)

Hamid Reza Vaezi Joze and Mark S Drew. Improved machine learning for image category recognition by local color constancy. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 3881–3884. IEEE, 2010. (Cited on pages 35 and 41.)

Timor Kadir, Andrew Zisserman, and Michael Brady. An affine invariant salient region detector. In *Computer Vision–ECCV 2004*, volume 3021 of *Lecture Notes in Computer Science*, pages 228–241. Springer, 2004. (Cited on page 154.)

Christopher Kanan, Arturo Flores, and Garrison W Cottrell. Color constancy algorithms for object and face recognition. In *Advances in Visual Computing*, pages 199–210. Springer, 2010. (Cited on page 35.)

Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–506. IEEE, 2004. (Cited on page 27.)

Fahad S. Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 979–986. IEEE, 2009. (Cited on pages 18, 19, 26, and 41.)

Fahad S Khan, Joost Weijer, Andrew D Bagdanov, and Maria Vanrell. Portmanteau vocabularies for multi-cue image representation. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 1323–1331. Curran Associates, Inc., 2011. (Cited on page 125.)

Fahad S. Khan, Joost van de Weijer, and Maria Vanrell. Modulating shape features by color attention for object recognition. *Int. Journal of Computer Vision*, 98(1):49–64, 2012. Springer. (Cited on pages 41 and 125.)

Rahat Khan, Joost Van de Weijer, Fahad Shahbaz Khan, Damien Muselet, Christophe Ducottet, and Cecile Barat. Discriminative color descriptors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2866–2873. IEEE, 2013. (Cited on page 41.)

Artiom Kovnatsky. Matlab file exchange code submission: Feature points in image, keypoint extraction. `http://www.mathworks.com/matlabcentral/fileexchange/29004-feature-points-in-image--keypoint-extraction`, 2010. Accessed: 14th of February 2015. (Cited on page 77.)

AndreyS. Krylov, DmitryV. Sorokin, DmitryV. Yurin, and EkaterinaV. Semeikina. Use of color information for keypoints detection and descriptors construction. In *Intelligent Science and Intelligent Data Engineering*, volume 7202 of *Lecture Notes on Computer Science*, pages 389–396. Springer, 2012. ISBN 978-3-642-31918-1. doi: 10.1007/978-3-642-31919-8_50. (Cited on pages 6, 19, 33, 46, and 47.)

AS Krylov and DV Sorokin. Gauss-laguerre keypoints descriptors for color images. In *Proc. Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2011. (Cited on pages 19, 33, 46, and 47.)

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Birds dataset. `http://www-cvr.ai.uiuc.edu/ponce_grp/data/`, 2005. Accessed: 26th of January 2015. (Cited on pages 31 and 34.)

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE, 2006. (Cited on page 131.)

Nicolas Le Bihan and Stephen J Sangwine. Quaternion principal component analysis of color images. In *Proc. IEEE International Conference on Image Processing (ICIP)*, volume 1, pages I–809. IEEE, 2003. (Cited on page 24.)

Rafael Lemuz-López and Miguel Arias Estrada. Ranking corner points by the angular difference between dominant edges. In *Computer Vision Systems*, pages 323–332. Springer, 2008. (Cited on page 154.)

Stefan Leutenegger, Margarita Chli, and Roland Yves Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2548–2555. IEEE, 2011. (Cited on pages 21 and 29.)

T. Lindeberg. Feature detection with automatic scale selection. *Int. Journal of Computer Vision*, 30(2):79–116, 1998. Springer. (Cited on pages 7, 20, 21, and 50.)

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004. Springer. (Cited on pages 3, 9, 18, 21, 27, and 124.)

Elmar Mair, GregoryD. Hager, Darius Burschka, Michael Suppa, and Gerhard Hirzinger. Adaptive and generic corner detection based on the accelerated segment test. In *Computer Vision–ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 183–196. Springer, 2010. (Cited on page 21.)

David Marr. Vision: A computational investigation into the human representation and processing of visual information. *WH Freeman and Company*, 1982. (Cited on page 4.)

Tony Marrero Barroso and Paul F. Whelan. Enhancing surf feature matching using colour histograms. In *Proc. Irish Machine Vision and Image Processing Conference (IMVIP)*, pages 111–112. IEEE, 2011. (Cited on page 6.)

Tony Marrero Barroso, Aubrey K. Dunne, John Mallon, and Paul F. Whelan. Towards real-time stereoscopic image rectification for 3d visualisation. In *Asian Conference for Computer Vision (ACCV), Workshop on Application of Computer Vision for Mixed and Augmented Reality (poster)*, 2010. (Cited on page 2.)

J. Matas, O. Chum, M. Urban, and T.áš Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. Elsevier. (Cited on pages 20 and 23.)

CH Mazel, TW Cronin, RL Caldwell, and NJ Marshall. Fluorescent enhancement of signaling in a mantis shrimp. *Science*, 303(5654):51–51, 2004. (Cited on page 4.)

Charles D. Michener. *The Social Behavior of the Bees: A Comparative Study*. Harvard University Press, 1974. (Cited on page 4.)

K. Mikolajczyk. Affine covariant regions resources. `www.robots.ox.ac.uk/~vgg/research/affine`, 2004. Accessed: 21st of January 2015. (Cited on pages 23, 33, and 36.)

K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 525–531. IEEE, 2001. (Cited on pages 7, 20, 49, 50, 52, 53, 57, 62, and 85.)

K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. Journal of Computer Vision*, 60(1):63–86, 2004. Springer. (Cited on pages 20, 21, 22, and 49.)

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 1615–1630, 2005. (Cited on pages 20, 36, 37, 46, 47, 91, 119, 151, and 159.)

K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1792–1799. IEEE, 2005a. (Cited on page 20.)

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. Journal of Computer Vision*, 65(1):43–72, 2005b. Springer. (Cited on pages 20 and 22.)

Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele. Multiple object class detection with a generative model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 26–36. IEEE, 2006. (Cited on page 127.)

Krystian Mikolajczyk, Mark Barnard, Jiri Matas, and Tinne Tuytelaars. Feature detectors and descriptors: The state of the art and beyond. In *Feature Workshop and Benchmark in conjunction with CVPR*, 2009. (Cited on page 36.)

Florica Mindru, Tinne Tuytelaars, Luc Van Gool, and Theo Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94(1):3–27, 2004. Elsevier. (Cited on page 32.)

Anlong Ming and Huadong Ma. A blob detector in color images. In *Proc. ACM International Conference on Image and Video Retrieval*, pages 364–370. ACM, 2007. (Cited on pages 23, 46, and 47.)

Aleksandra Mojsilovic. A computational model for color naming and describing color composition of images. *IEEE Trans. on Image Processing*, 14(5):690–699, 2005. (Cited on page 30.)

P. Montesinos, V. Gouet, and R. Deriche. Differential invariants for color images. In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, volume 1, pages 838–840. IEEE, 1998. (Cited on pages 22 and 154.)

Raúl Mur-Artal and Juan D Tardós. Fast relocalisation and loop closing in keyframe-based slam. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 846–853. IEEE, 2014. (Cited on page 125.)

Damien Muselet and Brian Funt. Color invariants for object recognition. In *Advanced Color Image Processing and Analysis*, pages 327–376. Springer, 2013. (Cited on page 70.)

R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2320–2327. IEEE, 2011. (Cited on page 1.)

M.-E. Nilsback and A. Zisserman. Flowers dataset. `http://www.robots.ox.ac.uk/~vgg/data/flowers/`, 2006. Accessed: 26th of January 2015. (Cited on pages 26 and 34.)

David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168. IEEE, 2006. (Cited on page 124.)

Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Proc. ECCV 2006*, volume 3954 of *Lecture Notes in Computer Science*, pages 490–503. Springer, 2006. (Cited on pages 127 and 141.)

Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996. Elsevier. (Cited on page 28.)

D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1817–1824. IEEE, Dec 2013. (Cited on page 1.)

Marta Penas and Linda G Shapiro. A color-based interest operator. In *Image Analysis and Processing–ICIAP 2009*, volume 5716 of *Lecture Notes in Computer Science*, pages 965–974. Springer, 2009. (Cited on pages 23, 46, and 47.)

Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3384–3391. IEEE, 2010. (Cited on page 131.)

E. Rosten, R. Porter, and T. Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32:105–119, 2010. (Cited on page 21.)

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: an efficient alternative to sift or surf. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011. (Cited on page 29.)

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. `http://image-net.org/`, 2014. Accessed: 13th of February 2015. (Cited on page 37.)

D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. (Cited on page 40.)

Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997. (Cited on page 154.)

Cordelia Schmid, Gyuri Dorkó, Svetlana Lazebnik, Krystian Mikolajczyk, and Jean Ponce. Pattern recognition with local invariant features. *Handbook of Pattern recognition and computer vision*, pages 71–92, 2005. (Cited on page 1.)

Nicu Sebe, Theo Gevers, Joost Van De Weijer, and Sietse Dijkstra. Corner detectors for affine invariant salient regions: Is color important? In *Image and Video Retrieval*, volume 4071 of *Lecture Notes in Computer Science*, pages 61–71. Springer, 2006. (Cited on pages 24, 27, 46, and 47.)

Steven A Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985. Wiley. (Cited on page 63.)

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127 (1):3–30, 2011. Springer. (Cited on page 130.)

Lilong Shi, Brian Funt, and Ghassan Hamarneh. Quaternion color curvature. In *Color and Imaging Conference*, pages 338–341. Society for Imaging Science and Technology, 2008. (Cited on page 23.)

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, `http://arxiv.org/pdf/1409.1556.pdf`, 2014. Accessed: 2nd March 2015. (Cited on pages 37 and 124.)

Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477. IEEE, 2003. (Cited on pages 8 and 124.)

Josef Sivic, Bryan C Russell, Alexei A Efros, Andrew Zisserman, and William T Freeman. Discovering objects and their location in images. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 370–377. IEEE, 2005. (Cited on page 127.)

Stephen M Smith and J. Michael Brady. Susan-a new approach to low level image processing. *Int. Journal of Computer Vision*, 23(1):45–78, 1997. Springer. (Cited on page 21.)

Xiaohu Song, Damien Muselet, and Alain Trémeau. Affine transforms between image space and color space for invariant local descriptors. *Pattern Recognition*, 46(8):2376 – 2389, 2013. Elsevier. (Cited on pages 19, 34, 46, and 47.)

Julian Stöttinger, Nicu Sebe, Theo Gevers, and Allan Hanbury. Colour interest points for image retrieval. In *Proc. of the 12th Computer Vision Winter Workshop*, pages 83–90, 2007. (Cited on page 25.)

Julian Stöttinger, Allan Hanbury, Theo Gevers, and Nicu Sebe. Lonely but attractive: Sparse color salient points for object retrieval and categorization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 1–8. IEEE, 2009. (Cited on pages 6, 25, and 41.)

Julian Stöttinger, Allan Hanbury, Nicu Sebe, and Theo Gevers. Sparse color interest points for image retrieval and object categorization. *IEEE Trans. on Image Processing*, 21(5):2681–2692, 2012. (Cited on pages 19, 22, 25, 27, 41, 45, 52, 63, 78, 122, and 150.)

Michael J Swain and Dana H Ballard. Color indexing. *Int. Journal of Computer Vision*, 7(1):11–32, 1991. Springer. (Cited on page 62.)

G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod. Rotation-invariant fast features for large-scale recognition and real-time tracking. *Signal Processing: Image Communication*, 28(4):334 – 344, 2013. Elsevier. (Cited on page 1.)

Jie Tang, Stephen Miller, Arjun Singh, and Pieter Abbeel. A textured object recognition pipeline for color and depth image data. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3467–3474. IEEE, 2012. (Cited on page 30.)

Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008. Now Publishers Inc. (Cited on page 7.)

Tinne Tuytelaars and Cordelia Schmid. Vector quantizing feature space with a regular lattice. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. (Cited on page 127.)

R. Unnikrishnan and M. Hebert. Extracting scale and illuminant invariant regions through color. In *Proc. British Machine Vision Conference (BMVC)*, pages 52–1, 2006. (Cited on pages 23, 46, and 47.)

K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010. (Cited on pages 6, 18, 19, 30, 32, 41, 46, 47, 97, 122, 125, and 127.)

K. E.A. van de Sande, T. Gevers, and C. G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32:1582–1596, 2010. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.154. (Cited on page 133.)

J. Van de Weijer, T. Gevers, and J-M. Geusebroek. Edge and corner detection by photometric quasi-invariants. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(4):625–630, 2005. (Cited on pages 22, 45, 47, 63, 64, 65, 70, 71, 72, 77, and 78.)

J. Van de Weijer, T. Gevers, and A. W. Smeulders. Robust photometric invariant features from the color tensor. *IEEE Trans. on Image Processing*, 15(1):118–127, 2006a. (Cited on pages 5, 22, 45, 47, 63, 65, 66, 77, and 78.)

Joost Van de Weijer and Fahad Shahbaz Khan. Fusing color and shape for bag-of-words based object recognition. In *Proc. International Conference on Computational Color Imaging (CCIW)*, pages 25–34. Springer, 2013. (Cited on page 18.)

Joost Van de Weijer and Cordelia Schmid. Soccer team dataset. `http://lear.inrialpes.fr/data`, 2006a. Accessed: 26th of January 2015. (Cited on pages 31 and 34.)

Joost Van de Weijer and Cordelia Schmid. Coloring local feature extraction. In *Computer Vision–ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 334–348. Springer, 2006b. (Cited on pages 18, 19, 26, 30, 31, 32, 46, 47, and 127.)

Joost. Van de Weijer, Theo. Gevers, and Andrew D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(1):150–156, 2006b. (Cited on pages 24, 26, 97, and 104.)

Joost Van de Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Trans. on Image Processing*, 18 (7):1512–1523, 2009. (Cited on page 26.)

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. Accessed: 2nd March 2015. (Cited on pages 126 and 129.)

Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 606–613. IEEE, 2009. (Cited on page 124.)

David Augusto Rojas Vigo, Fahad Shahbaz Khan, Joost Van De Weijer, and Theo Gevers. The impact of color on bag-of-words based object recognition. In *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, pages 1549–1553. IEEE, 2010a. (Cited on pages 6, 18, 19, 26, 27, and 125.)

David Rojas Vigo, Joost van de Weijer, and Theo Gevers. Color edge saliency boosting using natural image statistics. In *Conf. on Colour in Graphics, Imaging, and Vision*, number 1, pages 228–234. Society for Imaging Science and Technology, 2010b. (Cited on pages 26, 46, and 47.)

Vassilios Vonikakis, D Chrysostomou, R Kouskouridas, and A Gasteratos. Improving the robustness in feature detection by local contrast enhancement. In *Proc. International Conference on Imaging Systems and Techniques (IST)*, pages 158–163. IEEE, 2012. (Cited on page 36.)

Christian Wengert, Matthijs Douze, and Hervé Jégou. Bag-of-colors for improved image search. In *Proc. ACM International Conference on Multimedia*, pages 1437–1440. ACM, 2011. (Cited on page 19.)

Gunter Wyszecki and Walter Stanley Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*, volume 8. Wiley New York, 1982. (Cited on pages 63 and 66.)

Xin Yang and Kwang-Ting Cheng. Local difference binary for ultrafast and distinctive feature description. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(1):188–194, 2014. (Cited on page 29.)

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. Journal of Computer Vision*, 73(2):213–238, 2007. Springer. (Cited on pages 22 and 127.)

Jun Zhang, Youssef Barhomi, and Thomas Serre. A new biologically inspired color image descriptor. In *Computer Vision–ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 312–324. Springer, 2012. (Cited on pages 19 and 41.)