

Nonlinear Intensity Invariant Local Image Descriptors

PhD THESIS

submitted in partial fulfillment of the requirements of

Doctor of Technical Sciences

within the

Vienna PhD School of Informatics

by

Sajid Saleem

Registration Number 1128888

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: A.o.Univ.-Prof. Dipl.-Ing. Dr.techn. Robert Sablatnig

External reviewers:

Prof. Dr. Vaclav Hlavac, Czech Technical University in Prague, Czech Republic.

Prof. Dr. Thomas Pock, Graz University of Technology, Austria.

Wien, 24.09.2014

(Signature of Author)

(Signature of Advisor)

Declaration of Authorship

Sajid Saleem

Guglgasse 8/3/9B, 1110 Vienna, Austria.

I hereby declare that I have written this Doctoral Thesis independently, that I have completely specified the utilized sources and resources, and that I have definitely marked all parts of the work - including tables, maps, and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

(Place, Date)

(Signature of Author)

Acknowledgements

First of all, I would like to express my deepest gratitude to my doctoral advisor Prof. Dr. Robert Sablatnig for his extensive advise, guidance, cooperation, support, motivation, and patience from the beginning to the end of my PhD studies. He has contributed tremendously to my growth as a dedicated researcher.

I would like to thank the Vienna PhD School of Informatics, Austria, who provided the financial and the knowledge support during my whole PhD studies.

I would like to thank Prof. Dr. Vaclav Hlavac and Prof. Dr. Thomas Pock for reviewing my thesis.

I am deeply thankful to Dr. Abdul Bais for motivating me and increasing my interest in Computer Vision during my studies for a master's degree in Electrical Engineering. This really helped me to understand Computer Vision deeply, provided opportunity to gain research experience, and to develop my skills.

Many thanks to all my colleagues at the Computer Vision Lab of the Vienna University of Technology, Austria, in particular Fabian Hollaus, for their friendship, their suggestions, and for sharing their knowledge and skills. Thanks to all the people I met in these three years, which made my research experience more enjoyable.

Finally, I feel immensely grateful towards my parents, who always stood behind me, motivating, and pushing me to strive for more, especially my mother whose unwavering confidence and prayers kept me on the right path of hard work and dedication to achieve my aims and to reach thus far.

Most importantly, I would like to thank my brother who has always been there for me providing understanding, patience, encouragement, and support.

Abstract

An interesting problem in Computer Vision is the construction of local image descriptors. It deals with the description of intensity patterns within image patches. Image patches are local image regions centered at feature points. The description of such image patches helps in establishing correspondences between the feature points of two or more images of the same scene under intensity, scale, rotation, and affine changes. Such correspondences are used in a wide range of applications, such as image matching, image retrieval, object tracking, and object recognition.

This thesis presents new methods for the construction of local image descriptors in order to establish feature point correspondences under nonlinear intensity changes. Nonlinear intensity changes occur in multispectral imaging or when a scene is acquired under variable lighting conditions. Background noise and degradation in ancient document images also cause nonlinear intensity changes. Nonlinear intensity changes affect the performance of the state-of-the-art local descriptors, such as Scale Invariant Feature Transform (SIFT) and result in a low matching performance in image-to-image and image-to-database matching tasks. To cope with these problems, the new methods proposed in this thesis use novel image features, which are obtained by combining the strengths of image gradients, Local Binary Patterns, and illumination invariant edge detectors. These features are read from image patches by using the SIFT-like feature histogram schemes to construct five new local descriptors, which are: Local Binary Pattern of Gradients, Local Contrast SIFT, Differential Excitation SIFT, Normalized Gradient SIFT, and Modified Normalized Gradient SIFT.

To evaluate the performance of new descriptors, experiments on five different image datasets are performed. The performance of new descriptors are compared with that of SIFT and seven other state-of-the-art local descriptors. In the case of image-to-image matching, ground truth homographies between the pairs of images are used and the number of correct descriptor matches is counted for the performance comparison. In the case of image-to-database matching, a nearest neighbor based descriptor matching strategy is used and the recognition rates for two different tasks are computed. These tasks are Scene Category Recognition (SCR) and Optical Character Recognition (OCR).

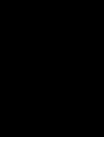
The experimental results show that the new descriptors obtain on average 0.5% to 12.8% better performance than SIFT in image-to-image matching. In the case of SCR, they obtain on average 1% to 5% better scene recognition rates than SIFT, whereas in the case of OCR, they demonstrate on average 1.1% to 6.7% better character recognition rates than SIFT.

Contents

1	Introduction	1
1.1	Intensity Changes Problem	1
1.2	Challenges in the Construction of Local Descriptors	2
1.3	Thesis Contribution	5
1.4	Evaluation Methodology	6
1.5	Thesis Structure	6
2	Related Work	7
2.1	Feature Points	7
2.1.1	Feature Point Detection	7
2.1.2	Feature Point Description	9
2.2	Intensity Invariant Local Descriptors	11
2.2.1	SIFT	11
2.2.2	GLOH	12
2.2.3	SURF	12
2.2.4	CS-LBP	13
2.2.5	DAISY (Dense Descriptor)	14
2.2.6	LIOP	15
2.3	Nonlinear Intensity Invariant Local Descriptors	17
2.3.1	GOM-SIFT	17
2.3.2	OR-SIFT	18
2.4	Summary	18
3	Methodology	21
3.1	Harris Laplace Points	21
3.2	Harris Laplace Regions	22
3.3	LBPG	23
3.3.1	Image Gradients	23
3.3.2	LBPG Feature	24
3.3.3	LBPG Feature Histograms	24
3.3.4	LBPG Descriptor	24
3.3.5	LBPG Example	25
3.4	LC-SIFT	25

3.4.1	LC Magnitudes	26
3.5	DE-SIFT	26
3.5.1	DE-SIFT and LC-SIFT Examples	26
3.6	NG-SIFT	27
3.6.1	Normalized Gradients	27
3.6.2	Proposed SIFT-like descriptor construction scheme	28
3.6.3	NG-SIFT Example	29
3.7	MN-SIFT	29
3.8	Comparison	30
3.9	Summary	32
4	Image Matching	33
4.1	Image Matching Framework	33
4.2	Evaluation Criteria	34
4.2.1	Precision – Recall Curve	35
4.2.2	AUC	36
4.2.3	Number of Correct Nearest Neighbor Matches	36
4.2.4	Overlap Error	36
4.2.5	Repeatability	37
4.3	CAVE Multispectral Image Dataset	37
4.3.1	Experiments on CAVE Dataset	40
4.3.2	Effect of Overlap Error	44
4.4	RWHI Image Dataset	48
4.4.1	Experiments on RWHI Dataset	49
4.4.2	Effect of Overlap Error	53
4.5	MSD Dataset	56
4.5.1	Experiments on MSD Dataset	57
4.5.2	Effect of Overlap Error	58
4.6	RGB-NIR Scene Dataset	62
4.6.1	Experiments on RGB-NIR Scene Dataset	62
4.7	Summary	67
5	Scene Category Recognition	73
5.1	Experimental Setup	73
5.2	Image Dataset	73
5.3	Cross Validation	74
5.4	Testing and Training Images	74
5.5	Description of Testing and Training Images	74
5.6	Naive Bayes Nearest Neighbor Classifier	75
5.7	Experimental Results	75
5.7.1	SCR based on Blue–NIR band images	75
5.7.2	SCR based on Red–NIR band images	77
5.8	Summary	79

6	Optical Character Recognition	83
6.1	Challenges in Optical Character Recognition	83
6.2	Proposed Method	85
6.2.1	Test Image	86
6.2.2	Dense SIFT Descriptors	86
6.2.3	Training SIFT Descriptors	86
6.2.4	Nearest Neighbor Distance Map	87
6.2.5	Non-Minima Suppression	87
6.2.6	Classification	89
6.3	Experimental Results	89
6.3.1	Experiments on Single Character Images	89
	Dataset-1	90
	Dataset-2	91
6.3.2	Experiments on Document Images	91
6.4	OCR Using Descriptors Proposed	92
6.4.1	Experiments on Single Character Images	93
6.4.2	Experiments on Document Images	93
6.5	Summary	94
7	Conclusion	97
7.1	Summary of Contributions	99
7.2	Future Research	99
7.2.1	Feature Point Detection	99
7.2.2	Image Registration	100
7.2.3	Document Image Analysis	100
A	List of Symbols	101
B	Definition of Terms	103
	Bibliography	105



Introduction

Local descriptors have proven to be effective for a wide range of applications, such as image registration [68, 71], image segmentation [21], object recognition [63], image retrieval [73], and biometric authentication [29]. Local descriptors represent the intensity patterns within image patches. Image patches are local image regions centered at feature points, which differ from their immediate neighborhood [39]. Local descriptors help in establishing correspondences between the feature points of two or more images of the same scene under scale, rotation, intensity, and affine changes [39, 40]. Several methods have been proposed for the construction of local descriptors [38]. Each method computes the local descriptors differently, but aims to make the descriptors invariant to scale, rotation, intensity, and affine changes between the images [19, 49, 63, 69]. This thesis presents new methods for the construction of local descriptors in order to establish feature point correspondences under intensity changes.

1.1 Intensity Changes Problem

There are generally two types of intensity changes: (i) monotonic [19, 69], and (ii) nonlinear [68, 71]. Monotonic intensity changes occur between the gray scale images of the same scene. They occur due to variable lighting conditions during the scene acquisition. They affect the intensity values, but result in almost similar intensity histograms. For instance, in Figure 1.1, the intensity histograms of two images I_1 and I_2 of the same scene, are shown [38]. Each intensity histogram shows the number of pixels against 256 intensity values. Some minor intensity changes between I_1 and I_2 can be seen, despite such monotonic intensity changes the intensity histograms of I_1 and I_2 look quite similar.

In contrast to monotonic intensity changes, nonlinear intensity changes occur between monochrome images of the same scene. Each of the monochrome image is taken with a sensor (filter) sensitive to a different wavelength [7]. A collection of several monochrome images of the same scene is referred to as a multispectral image [70], which is a data cube, consisting of two spatial dimensions plus one wavelength dimension [46]. The wavelength dimension consists of several

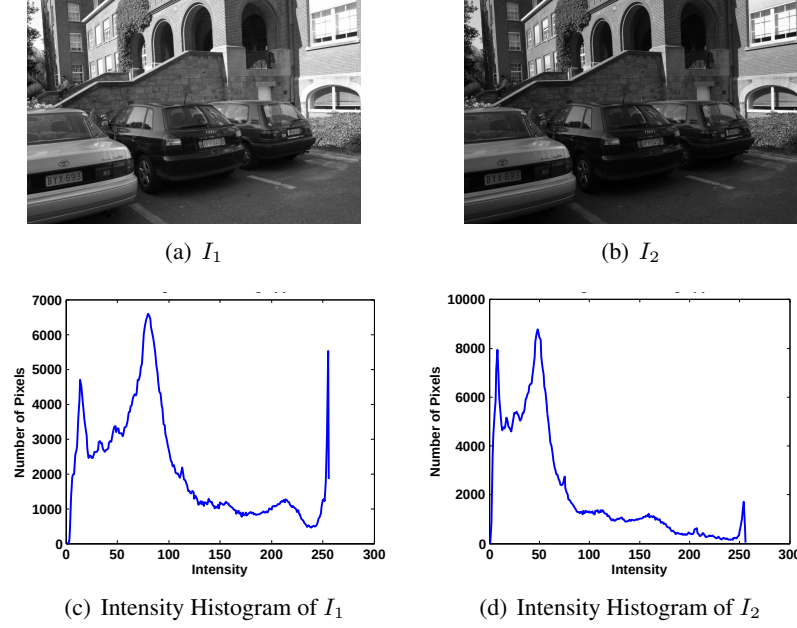


Figure 1.1: shows monotonic intensity changes between the images of the same scene.

discrete wavelength bands. A Red-Green-Blue (RGB) image in this context is a special type of a multispectral image. It consists of three bands. Each band of the multispectral image holds an image of the same scene, containing pixel intensities, which correspond to scene based sensory responses [7]. These responses are sensitive to objects in the scene as well as wavelength and cause nonlinear intensity changes between different band images of the same scene.

In Figure 1.2 two images I_{VS} and I_{IR} of the same scene are shown to illustrate the nonlinear intensity changes [1]. Due to scene acquisition in Visible Spectrum (VS) and Infra-Red (IR) bands [1], the scene contents in these images look different. The intensity histograms of I_{VS} and I_{IR} are also shown in Figure 1.2. It can be seen that large changes in the scene appearance as well as pixel intensities occur in the case of nonlinear intensity changes in contrast to monotonic intensity changes. This makes the task of image-to-image and image-to-database matching based on intensity values, intensity histograms, and local descriptors, more challenging and difficult than monotonic intensity changes [1, 53].

1.2 Challenges in the Construction of Local Descriptors

This thesis presents new methods for the construction of local descriptors in order to overcome the effects of nonlinear intensity changes between the images. Local descriptors represent the intensity patterns within image patches. Therefore, the intensity-based methods [28, 64, 69] for the construction of local descriptors, can not be used. This is due to large changes in the intensity values, which occur in the case of nonlinear intensity changes, as shown in Figure 1.2. The gradient based methods [19, 33] seem to be more suitable, because they use intensity dif-

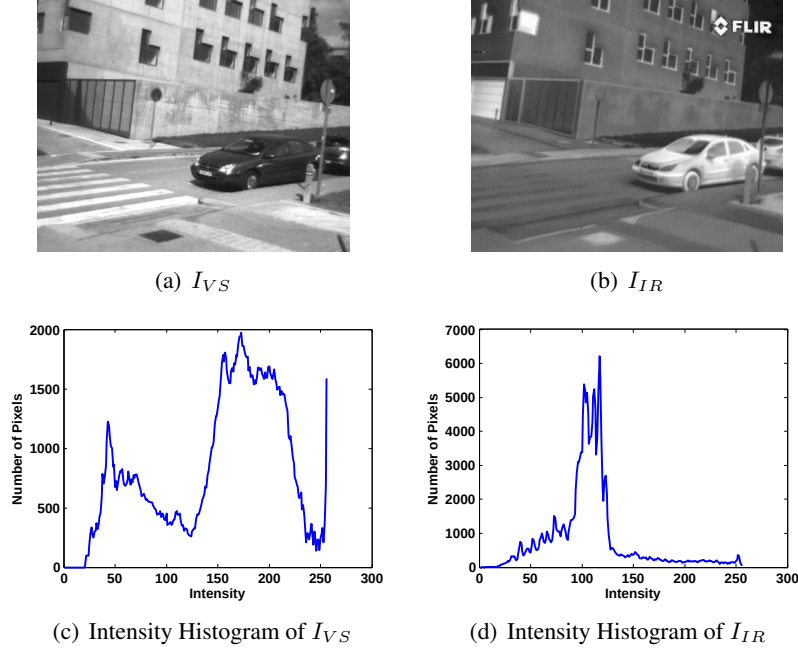


Figure 1.2: shows nonlinear intensity changes between the images of the same scene when VS and IR bands are used.

ferences (i.e., image gradients) as image features instead of intensity values for the construction of local descriptors. A performance evaluation of local descriptors in [38] shows that gradient based methods demonstrate better performance under intensity changes than the intensity based methods. Moreover, the performance evaluation shows that Scale Invariant Feature Transform (SIFT) [33] obtains better performance than other gradient based methods, such as Shape Context [3], Steerable filters [14], and Moment invariants [38]. Based on this performance evaluation [38], SIFT is used as a state-of-the-art method in this thesis for both descriptor construction and performance evaluation.

SIFT demonstrates better performance against intensity changes due to its spatial feature histogram scheme, which divides each image patch around the feature point into 4×4 location bins and then, constructs a histogram of oriented gradients for each location bin with the help of Gaussian weighted gradient magnitudes. SIFT concatenates such feature histograms and builds a description vector of length 128 for the image patch.

To illustrate the robustness of SIFT to intensity changes, two image patches R_1 and R_2 are shown in Figures 1.3(a) and (b), respectively. These image patches depict the same scene contents, but under monotonic intensity changes. In spite of that, the SIFT descriptors of R_1 and R_2 , which are shown in Figures 1.3(a) and (b), respectively, look quite similar. The elements of the SIFT descriptors are shown with line segments [65] inside 4×4 location bins [33]. The length of a line segment corresponds to the magnitude of a SIFT descriptor element. By comparing the line segments in corresponding location bins, it can be seen that the SIFT descriptors of R_1 and R_2 look quite similar, despite the monotonic intensity changes.

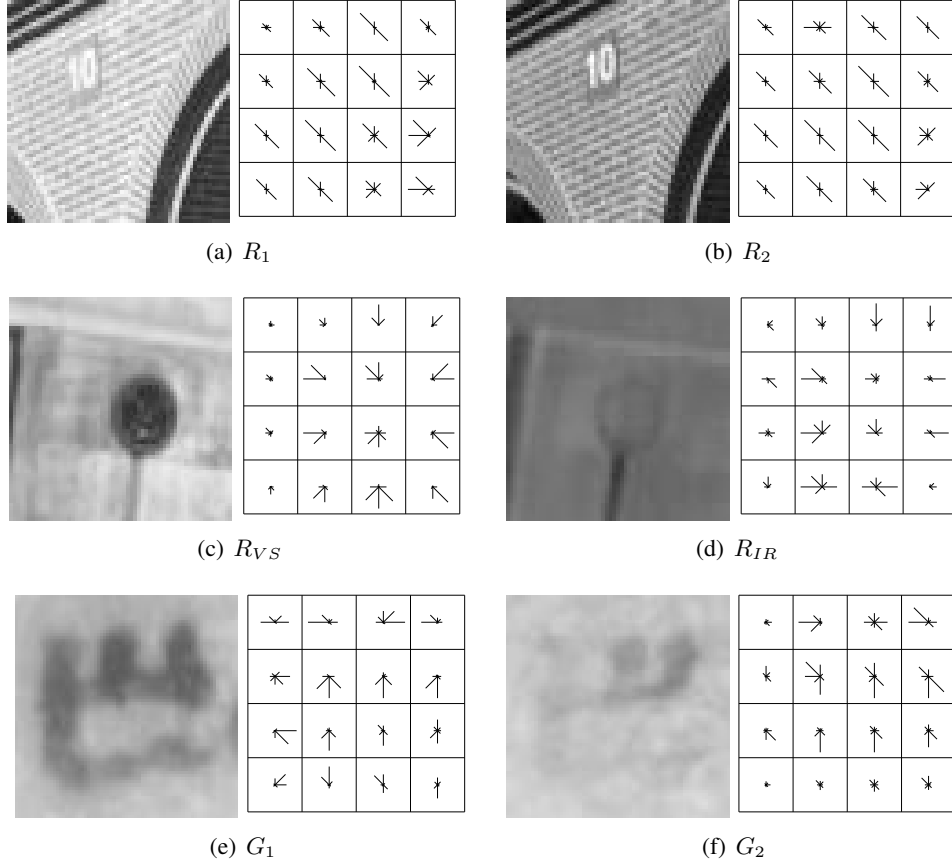


Figure 1.3: shows the effects of intensity changes on SIFT descriptor [33]. The elements of the SIFT descriptors are shown as line segments inside 4×4 location bins [65]. First row shows monotonic intensity changes, whereas second and third rows show nonlinear intensity changes between the image patches.

Now consider two image patches R_{VS} and R_{IR} in Figures 1.3(c) and (d), respectively. These image patches depict the same scene contents in VS and IR bands. Therefore, nonlinear intensity changes can be seen, which affect the appearance of the scene contents as well as the intensity values and result in dissimilar SIFT descriptors, as shown in Figures 1.3(c) and (d). It can be noted that SIFT efficiently overcomes the monotonic intensity changes between the corresponding image patches, but it suffers from nonlinear intensity changes. This problem motivates us to design new methods for the construction of local descriptors in order to overcome the effects of nonlinear intensity changes.

Note that, degradation in ancient document images also cause nonlinear intensity changes. For instance, in Figures 1.3(e) and (f) two image patches G_1 and G_2 , respectively, are shown. These image patches depict the same Glagolitic character [36], but due to degradation (noise and fading ink) the SIFT descriptors of G_1 and G_2 also look different.

1.3 Thesis Contribution

This thesis makes the following contributions:

- A performance evaluation of eight different state-of-the-art local descriptors under non-linear intensity changes [49]. The local descriptors used in this performance evaluation are:
 1. SIFT [33]
 2. Gradient Location Orientation Histogram (GLOH) [38]
 3. Speeded Up Robust Features (SURF) [2]
 4. Center Symmetric Local Binary Patterns (CS-LBP) [19]
 5. DAISY [58]
 6. Local Intensity Order Pattern (LIOP) [69]
 7. Gradient Orientation Modification (GOM)-SIFT [71]
 8. Orientation Restriction (OR)-SIFT [68]

These state-of-the-art local descriptors are further investigated in order to design new methods for the construction of local descriptors to achieve better performance against nonlinear intensity changes.

- A new method for the construction of local descriptors is proposed, which is referred to as Local Binary Pattern of Gradients (LBPG) [51]. This new method combines the strengths of SIFT and CS-LBP to overcome the effects of nonlinear intensity changes.
- Two different modifications to SIFT are proposed. These modifications are based on two different illumination invariant edge detectors, i.e., Local Contrast (LC) [56] and Differential Excitation (DE) [8]. The edge detectors provide illumination invariant edge responses, which are used in the SIFT method instead of gradient magnitudes to boost the contributions of local edges in the construction of local descriptors. This results in two new local descriptors, which are: LC-SIFT [52] and DE-SIFT [52].
- A modified SIFT-like descriptor construction scheme is proposed. The scheme proposed computes two new local descriptors by making use of Normalized Gradients (NG) and Modified Normalized (MN) gradients as image features. These new local descriptors are: NG-SIFT [53] and MN-SIFT [49].
- A new method for the recognition of Glagolitic characters is proposed [50]. It is based on Dense SIFT [63] algorithm and designed for degraded ancient manuscripts, named Missale Sinaiticum, containing historical documents written in the 11th century [36]. Due to parchment aging and degradations, the recognition of Glagolitic characters in these manuscripts, is a challenging task. The method proposed deals with such problems by making use of Dense SIFT algorithm and the descriptors proposed (LBPG, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT). It obtains better recognition rates than another

method [10], which has been used for the recognition of Glagolitic character in similar manuscripts.

1.4 Evaluation Methodology

Experiments on five different image datasets are performed and the performance of the descriptors proposed and the state-of-the-art local descriptors is measured for three different tasks, which are:

- Image Matching [38]
- Scene Category Recognition (SCR) [5]
- Optical Character Recognition (OCR) [10]

For image matching and SCR, experiments on multispectral image datasets [1, 5, 7, 70], are performed, whereas for OCR, a gray scale image dataset of Missale Sinaiticum, is used. In each task, the performance of descriptors proposed (LBP, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) is compared with that of the state-of-the-art local descriptors (SIFT, GLOH, CS-LBP, SURF, DAISY, GOM-SIFT, OR-SIFT, and LIOP).

Image patches centered at Harris Laplace feature points [37] are used in image matching and SCR for the construction of descriptors proposed and the state-of-the-art local descriptors, whereas in the case of OCR, image patches centered at densely sampled feature points [63] are used. In the case of image matching, the number of correct descriptor matches is counted by using the ground truth homographies between the pairs of images for the performance comparison. Whereas for SCR and OCR, a nearest neighbor based descriptor matching strategy is used to obtain the recognition rates for the performance comparison.

1.5 Thesis Structure

The rest of this thesis is structured as follows. Chapter 2 presents related work. It focuses on methods that have been used for the construction of local descriptors. The methods that have been widely used against monotonic and nonlinear intensity changes are briefly described in order to understand how such methods compute the local descriptors to cope with the intensity changes and how new local descriptors can be computed to overcome the effects of nonlinear intensity changes. Chapter 3 briefly describes the construction of the proposed descriptors. It also presents the main ideas behind each descriptor proposed. Chapter 4 presents the experimental results for image matching. The results achieved on four multispectral datasets are reported and the comparison of descriptors proposed and the state-of-the-art local descriptors is presented. Chapter 5 presents the experimental results for SCR. Chapter 6 presents a new method based on Dense SIFT for the recognition of Glagolitic characters. It also presents the experimental results when descriptors proposed and the state-of-the-art local descriptors are used instead of Dense SIFT for the recognition of Glagolitic character. Finally, Chapter 7 concludes the thesis and gives directions for future research.

Related Work

This chapter presents an overview of the state-of-the-art methods, which have been used for the construction of local descriptors. These methods can be grouped into two different types: (i) intensity, and (ii) gradient based methods. The intensity based method [69] uses intensity values as image features for the construction of local descriptors, whereas the gradient based methods [2, 19, 33, 58] use the directional changes in the intensity, i.e., image gradients for the construction of local descriptors.

In section 2.1.1, first the feature point detectors are described. Then, the feature point descriptors are discussed in Section 2.1.2. The feature point descriptors are also referred to as local descriptors. Section 2.2 describes the intensity (i.e., illumination) invariant local descriptors. Finally, local descriptors invariant to nonlinear intensity changes are discussed in Section 2.3.

2.1 Feature Points

Feature points are distinct image locations, which differ from their immediate neighborhood [39]. They are also referred to as interest points [37] or keypoints [33]. They represent isolated image points, points lying on image edges, representing corner, or blob-like image structures. Generally, they are detected because of large image gradients around them [17, 40]. The description of image patches centered at such feature points help in establishing the correspondences between the feature points of two or more images under intensity, scale, rotation, and affine changes [33, 75].

2.1.1 Feature Point Detection

The detection of feature points can be traced back to the work of Moravec [41]. He uses a local image window and determines the average changes in the image intensity that result from shifting the window by a small amount in various directions for the corner detection. Harris and Stephens [18] revisit the Moravec's corner detection and use a smooth circular window in order to make the corner detection stable under noise and other small image variations. Such a corner

detection was initially used for stereo matching [41]. But in the later research, Harris shows its effectiveness for structure from motion recovery application [17]. Since then, it is known as Harris corner [17] and has been widely used for various other applications [75]. Although the process is known as a corner detection, but beside corners, it also detects isolated image points, which have large image gradients in all directions. Such image points are also referred to as interest points, feature points, or keypoints [2, 33, 38]. Several methods have been proposed for the detection of feature points in order to make the feature points invariant to intensity, scale, rotation, and affine changes [39].

The initial Harris corner detector [17] does not take into account the scale changes between the images. But later on, multi-scale Harris corner detectors have been developed [37, 39]. To this end, the scale space theory of Lindeberg has been widely used [31, 32]. Lindeberg shows an effective method for the estimation of an appropriate and consistent characteristic scale for detecting the feature points. Mikolajczyk and Schmid use the Lindeberg's scale space theory and detect Harris Laplace feature points [37]. They use a two step process. In the first step, they assess the Harris function [17] on multi-scale images and then, in the second step they remove all the Harris corners that do not show stability along the scale direction by using a Laplacian function [37]. They demonstrate good performance, i.e., repeatability with respect to scale, rotation, and intensity changes.

Lowe [33] also uses the scale space theory of Lindeberg. But he applies the theory in a different manner for the detection of SIFT keypoints. He uses a Difference of Gaussian filters instead of the Laplacian function and also a Hessian function [33] instead of the Harris function [17]. He shows that Difference of Gaussian filters with Hessian function detect blob-like image structures, which show good performance under scale, rotation, and affine changes for image matching and object recognition tasks [33, 34].

Mikolajczyk and Schmid revisit the detection of SIFT keypoints [33]. They use Hessian and Laplacian functions for the detection of Hessian Laplace feature points [40]. These points are similar to SIFT keypoints and represent blob like image structures. Moreover, Mikolajczyk and Schmid show that, with an affine adaptation Harris Laplace and Hessian Laplace feature points can be made robust to affine changes [40]. To this end, they estimate the affine shape of the feature points by modifying the points' scale and neighborhood until they converge to affine invariant points. They refer to affine invariant feature points as Harris Affine and Hessian Affine [40]. They show that such feature points resist large affine changes and also demonstrate good performance under scale and rotation changes.

Bay et al. [2] also revisit the SIFT keypoint detection [33]. They use integral images [55, 67] in order to speed up the detection process. They propose SURF features [2], which have been widely used [22, 25], especially for the real time applications [42]. Matas et al. [35] propose a completely different method, i.e., Maximally Stable Extremal Regions (MSER) in contrast to SIFT and Harris Laplace feature points. They show that MSER possess desirable properties, such as resisting continuous transformation of image coordinates, image intensities, and can be detected at near the frame rate. They use MSER in the wide-baseline stereo matching task and obtain good results under scale, intensity changes, out-of-plane rotation, and occlusion [35]. MSER have also been used in other visual tasks [9, 24, 38].

Rosten and Drummond propose a FAST feature point detector [47]. They show that SIFT [33]

and Harris Laplace [40] methods, although detect high quality feature points, but they are computationally intensive for the real-time applications. To meet the requirements of the real time applications, they propose a fast feature point detector [47], which uses a machine learning technique to first learn the intensity patterns around the feature points and then recognizes similar intensity patterns to detect the feature points at nearly the frame rate [47].

From a basic Harris corner detector [18] to a FAST corner detector [47], several methods have been proposed. These methods detect the feature points of desired quality and quantity in order to meet the application needs. Each method creates a list of feature points. The lists differentiates between the feature points based on their image location, scale, and orientation attributes. Although such attributes carry useful information, i.e., where these feature points are coming from, but are insufficient for establishing the feature point correspondences between the images. In order to deal with this problem, the description of the feature points have been widely used [38]. In the next subsection various methods for the construction of feature point descriptors are discussed in detail.

2.1.2 Feature Point Description

The description of feature points can be tracked back to the work of Zhang et al. [74]. They use correlation windows centered at Harris corners [17] for establishing the correspondences. But they get a large number of outliers due to their simple approach. They remove the outliers by making use of a fundamental matrix describing the geometric constraints between the images.

Torr [59] also uses a similar approach to establish the feature point correspondences between the images. He also uses the geometric constraints between the images to remove the outliers. However, the work of Schmid and Mohr for the description of the feature points is considered to be ground-breaking [54]. They use Harris corners [17] and describe their neighborhoods with rotationally invariant descriptors. They demonstrate good performance for object recognition, especially under occlusion and clutter.

The work of Lowe on the description of the SIFT keypoints [34] is also considered of great significance. Lowe uses image patches centered at SIFT keypoints for the descriptor construction. Unlike the construction of rotationally invariant descriptors [54], he uses the biological vision model of Edelman et al. [11], which shows that complex neurons in the primary visual cortex respond to a gradient at a particular orientation and spatial frequency, but the location of the gradient on the retina is allowed to shift over a small receptive field rather than being precisely localized [33]. Inspired by this biological vision model, Lowe filters the image patches centered at SIFT keypoints with directional gradients and spatially divides each image patch into 4×4 location bins. He computes a feature histogram of orientated gradients for each location bin by making use of the sum of Gaussian weighted gradient magnitudes. He concatenates the feature histograms over all the location bins to build a SIFT keypoint descriptor. He shows that such a method for descriptor construction, creates robust and distinct descriptors, which obtain high precision results in image-to-image matching and object recognition tasks under occlusion, scale, rotation, and affine transformations [33].

SIFT descriptor has been extensively used in a wide range of applications [2, 5, 38, 71]. Several modifications to SIFT have also been proposed in order to improve the SIFT robustness to intensity [19, 68, 71] and affine changes [72]. Mikołajczyk and Schmid compare the local

descriptors [28,33,64] and show that SIFT demonstrates the best performance for image matching [38]. They also propose an extended version of SIFT, i.e., GLOH, which is based on log-polar location bins instead of 4×4 location bins [33] of SIFT. They compute 272 dimensional description vectors for each image patch by using the log-polar location bins and then reduce the vector dimensions to 128 by making use of Principle Component Analysis (PCA) [23], in order to make the GLOH descriptors comparable to 128 dimensional SIFT descriptors. For PCA, they estimate the covariance matrix on 47,000 image patches and use the largest 128 eigenvectors for the GLOH descriptors [38].

Similarly, Ke and Sukthankar [23] propose a modified SIFT descriptor. They concatenate the horizontal and vertical gradients of each image patch to compute the descriptor instead of using the 4×4 location binning scheme of the SIFT. They refer to their method as PCA-SIFT [23]. To compute PCA-SIFT, they first resize each image patch to 41×41 pixels and read the image gradients from the resized patch at 39×39 locations to build $2 \times 39 \times 39 = 3042$ dimensional PCA-SIFT descriptor. They reduce the descriptor dimensions to 128 with PCA. They demonstrate better performance than SIFT and show that PCA-SIFT descriptor of dimension less than 128 also performs better than SIFT. This makes the descriptor matching less computationally intensive than 128 dimensional SIFT.

Bay et al. [2] also revisit the SIFT descriptor. They propose SURF descriptor as an alternative to SIFT. They use integral images [67] and the Haar wavelength responses instead of the image gradients to make the construction of SURF descriptors faster than SIFT. They show that 64 dimensional SURF descriptor outperforms 128 dimensional SIFT, which makes the descriptor matching less computationally intensive than SIFT.

Heikkila et al. also revisit the SIFT descriptor. They propose CS-LBP descriptor [20]. They use Local Binary Patterns (LBP) [44] scheme to compute the CS-LBP features for the construction of descriptors. They show that CS-LBP features possess properties similar to image gradients, but can be computed much faster than image gradients and demonstrate more robustness to intensity changes than image gradients. They read the CS-LBP features from the image patches by using the SIFT-like feature histogram scheme to build the CS-LBP descriptors. They demonstrate better results for image matching and object recognition than SIFT under intensity, scale, rotation, and affine changes. However, the CS-LBP method computes 256 dimensional description vectors, which make the descriptor matching more computationally intensive than SIFT.

Tola et al. also revisit the SIFT descriptor and propose a new local descriptor, which they refer to as DAISY [58]. They design DAISY for the dense matching applications [13, 27], which requires the construction of descriptors for densely sampled feature points [63]. They use circular location bins instead of 4×4 location bins of SIFT. Additionally, they replace the Gaussian weighted sums of gradient magnitudes used in the SIFT method, with the convolution of the image with several oriented derivatives of the Gaussian filters to expedite the descriptor construction. They show that such modifications to SIFT result in descriptors, which perform comparable to SIFT, but they make the descriptor construction much faster than SURF and SIFT. However, DAISY computes 200 dimensional description vectors and makes the descriptor matching computationally intensive than SIFT.

Wang et al. use a completely different method for the construction of descriptors [69]. They

use an intensity order based location binning scheme instead of cartesian (4×4) [33], log polar [38], and circular [58] bins. They obtain the intensity based location bins by sorting the pixel intensities of the image patch in the ascending order [12], then, equally quantizing the image patch into ordinal (i.e., location) bins according to the intensity orders. They compute LIOP features for each ordinal bin to compute a descriptor for the image patch. They demonstrate better performance under monotonic intensity changes than SIFT and CS-LBP. However, LIOP constructs 144 dimensional description vectors and makes the descriptor matching computationally intensive than SIFT.

The above discussion show that SIFT has been extensively used and several modifications to SIFT have been proposed, which either improve the performance of SIFT [19, 38, 72] or expedite the construction of SIFT to meet the needs of the real time applications [2, 6, 58]. In either case, SIFT has been used as an effective and efficient method for both descriptor construction and performance evaluation [8, 38, 72].

The main focus in this thesis is the construction of nonlinear intensity invariant local descriptors. Therefore, SIFT and its extended versions, such as GLOH, SURF, DAISY, and CS-LBP are of great importance. In the next section, these local descriptors are described in detail in order to understand how do they compute the descriptors robust to intensity (i.e, illumination) changes. A detailed overview of such local descriptors helps in providing the required insight to compute new local descriptors robust to nonlinear intensity changes between the images.

2.2 Intensity Invariant Local Descriptors

This section describes SIFT, SURF, GLOH, SURF, DAISY, LIOP, and CS-LBP descriptors in detail.

2.2.1 SIFT

SIFT is a gradient based descriptor [33]. It uses a spatial feature histogram scheme to describe the intensity patterns within the image patches. It divides each image patch into 4×4 location bins, as illustrated in Figure 2.1, where an image patch inside a bounding box is shown and a '+' sign depicts the location of the feature point. SIFT isolates this image patch from the image, as shown in Figure 2.1(b) and convolves the patch with kernels $[1, 0, -1]$ and $[1, 0, -1]^T$ to obtain the directional gradients F_h and F_v , respectively. Then, it computes the gradient magnitudes (Ω) and orientations (β) as follows:

$$\Omega = \sqrt{F_h^2 + F_v^2} \quad (2.1)$$

$$\beta = \text{atan2}(F_v, F_h) \quad (2.2)$$

Having computed the gradients, SIFT constructs a feature histogram of oriented gradients for each location bin. To this end, it quantizes the gradient orientations (between 0 and 2π radians) into eight different levels and constructs the feature histograms by using the sum of Gaussian weighted gradient magnitudes. The Gaussian weighting gives higher weightage to the gradient samples near the patch center than the patch boundary. SIFT also distributes such Gaussian

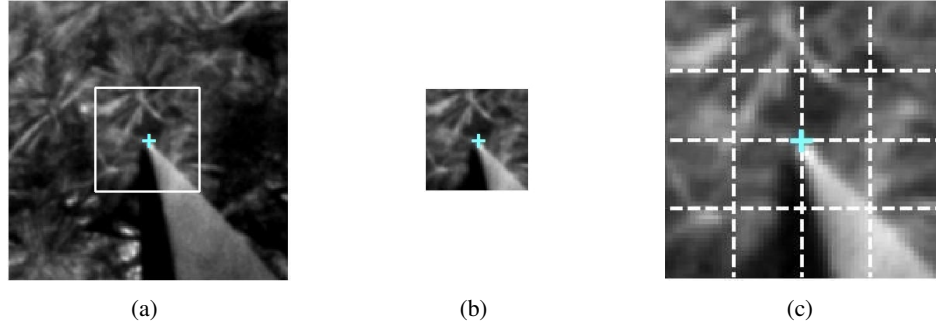


Figure 2.1: (a) An image patch centered at a feature point '+' is shown inside a bounding box, (b) the patch is cropped, and (c) divided into 4×4 location bins to compute a SIFT descriptor.

weighted gradient magnitudes among the adjacent feature histograms via a soft binning scheme in order to compensate the resulting descriptor for geometric deformations and feature point localization errors [33]. Finally, it concatenates the feature histograms over all the location bins to build a $4 \times 4 \times 8 = 128$ dimensional SIFT descriptor. It normalizes the descriptor to unit norm to overcome the effects of intensity changes, limits the larger descriptor elements to 0.2, and then re-normalizes the descriptor to unit norm.

2.2.2 GLOH

GLOH [38] is an extended version of SIFT. It uses a method based on log-polar location bins instead of 4×4 location bins of SIFT to compute the descriptors. Figure 2.2 shows 4×4 (cartesian) and log-polar location bins. In the case of log-polar location bins, three bins in the radial and eight in the angular directions are used, which result in total 17 location bins. The angular directions for the central bin, are not used. GLOH quantizes the gradient orientations into 16 different levels and computes a histogram of 272 bins for the image patch. The dimension of this histogram is reduced with PCA to 128. The covariance matrix for PCA is estimated on 47,000 image patches and the largest 128 eigenvectors are used for the GLOH descriptor [38]. A performance evaluation published in [38] shows better performance of GLOH than SIFT under intensity changes.

2.2.3 SURF

SURF [2] is a modified version of SIFT. It computes descriptors faster than SIFT and also demonstrates better performance than SIFT [2]. Like SIFT, it uses 4×4 location bins to capture the spatial information around the feature points for the construction of robust and distinct descriptors. Unlike SIFT, it uses Haar wavelet responses as image features. Figure 2.3 shows two dimensional Haar wavelet filters, which SURF uses to compute the Haar wavelet responses in horizontal (d_x) and vertical (d_y) directions. From each location bin, SURF reads Haar wavelet responses at 5×5 regularly spaced sample points. It also uses a Gaussian weighting of Haar wavelet responses to compensate the descriptor for geometric deformations and feature point

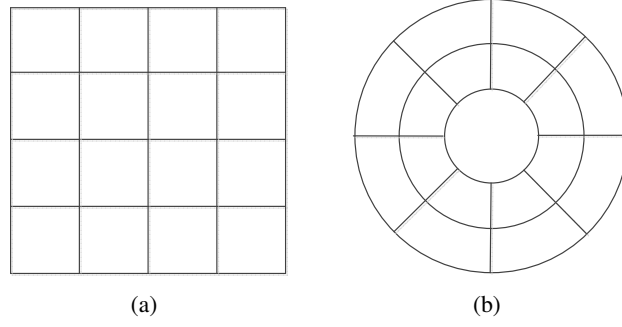


Figure 2.2: Location bins, which are used in the construction of (a) SIFT and (b) GLOH descriptors. Images from [38].

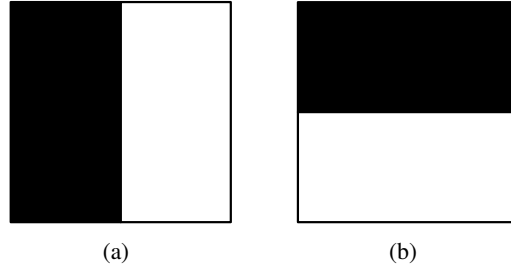


Figure 2.3: Two dimensional Haar wavelet filters are shown, which are used in the construction of SURF descriptors. These filters compute responses in (a) x and (b) y directions. The dark parts have the weight -1 and the light parts +1. Images from [2].

localization errors. Then, it constructs a four dimensional feature histogram v for each location bin as:

$$v = \{ \Sigma dx, \Sigma dy, \Sigma |dx|, \Sigma |dy| \} \quad (2.3)$$

where $\Sigma d.$ and $\Sigma |d.|$ represent the sum and the absolute sum of Haar wavelet responses. SURF concatenates feature histograms over all the location bins and builds a description vector of dimension 64. The performance evaluations of local descriptors in [22, 48], show better performance of SURF than SIFT under intensity changes.

2.2.4 CS-LBP

CS-LBP [19] is an extended version of SIFT. It computes descriptors faster than SIFT and also demonstrates better performance under intensity changes than SIFT [19]. Unlike SIFT, it uses a modified Local Binary Patterns (LBP) [44] scheme to compute the CS-LBP features and uses such features instead of image gradients in the SIFT like feature histogram scheme to construct the descriptors.

Equation 2.4 describes the computation of a CS-LBP feature for a central sample n_c , which is illustrated in Figure 2.4, where n_i represents equally spaced N samples at a radial distance

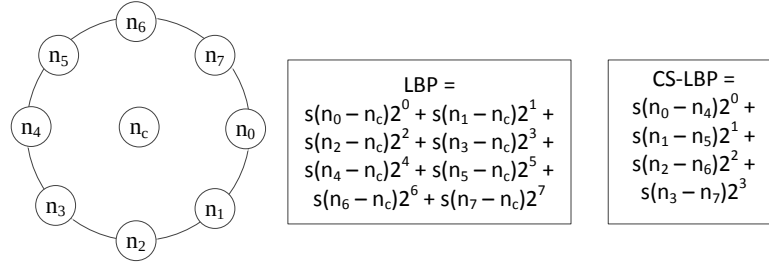


Figure 2.4: Illustration of LBP and CS-LBP for a neighbourhood of eight pixels. Images from [19].

of R from n_c . $s(z)$ binarizes the intensity differences between n_i and its center-symmetric neighbor $n_{i+(N/2)}$. The binarized differences are weighted and summed over all the center-symmetric pairs to obtain a CS-LBP feature for n_c , as described in Equation 2.4. This process is performed on each pixel of the image patch by using the pixel as a sample n_c in the CS-LBP method (Figure 2.4). Then, the CS-LBP features are read from the image patch by using the SIFT-like feature histogram scheme to obtain the CS-LBP descriptor.

Parameter N gives $2^{N/2}$ distinct center-symmetric local binary patterns and results in $4 \times 4 \times 2^{N/2}$ dimensional CS-LBP descriptors, where 4×4 represents the 16 location bins of SIFT. Heikkila et al. [19] show that $N=8$ and $R=2$ result in 256 dimensional CS-LBP_{2,8} descriptors, which demonstrate better performance than SIFT under intensity changes.

$$CS-LBP_{R,N} = \sum_{i=0}^{(N/2)-1} s(n_i - n_{i+(N/2)})2^i \quad (2.4)$$

$$s(z) = \begin{cases} 1 & z \geq 0.01 \\ 0 & \text{otherwise} \end{cases} \quad (2.5)$$

2.2.5 DAISY (Dense Descriptor)

DAISY [58] is a fast method for the construction of descriptors. It is designed for densely sampled feature points [13, 27]. It is inspired by SIFT. It achieves faster descriptor computation by replacing the sum of Gaussian weighted gradient magnitudes in the SIFT method with the convolutions of the image with several oriented derivatives of the Gaussian filters as described below [58]:

$$G_o^\Sigma = G_\Sigma * \left(\frac{\partial I}{\partial o} \right)^+ \quad (2.6)$$

where G_Σ represents a Gaussian kernel of a standard deviation Σ , o is the orientation of the derivative, the operator $(.)^+$ is such that $(a)^+ = \max(a, 0)$ and G_o^Σ represents the convolution result, which is also referred to as a convolved orientation map. Each orientation map of DAISY holds the values of the specific oriented derivative of the image, i.e., $G_o = \left(\frac{\partial I}{\partial o} \right)^+$.

DAISY uses 8 such orientation maps, i.e., $o_1, o_2, \dots, o_8$. It convolves each orientation map with Gaussian kernels of three different Σ values, i.e., $\Sigma_1 = 2.55, \Sigma_2 = 3\Sigma_1$ and $\Sigma_3 = 5\Sigma_1$. Then, it reads values from the convolved orientation maps at specific locations to construct the descriptors.

To understand the construction of a DAISY descriptor, let $h_\Sigma(u, v)$ be a feature vector made of values at a location (u, v) in the orientation maps:

$$h_\Sigma(u, v) = [G_{o_1}^\Sigma(u, v), G_{o_2}^\Sigma(u, v), \dots, G_{o_8}^\Sigma(u, v)]^\top \quad (2.7)$$

where $G_{o_1}^\Sigma, G_{o_2}^\Sigma, \dots, G_{o_8}^\Sigma$ are Σ -convolved orientation maps and the feature vector $h_\Sigma(u, v)$ holds the values of Σ -convolved orientation maps at a location (u, v) . DAISY normalizes this feature vector to obtain a unit norm vector $\tilde{h}_\Sigma(u, v)$. DAISY concatenates $\tilde{h}_\Sigma(u, v)$ vectors, which it obtains from 25 different locations centered at pixel (u_0, v_0) to build a DAISY descriptor $D(u_0, v_0)$:

$$D(u_0, v_0) = \begin{bmatrix} \tilde{h}_{\Sigma_1}^\top(u_0, v_0), \\ \tilde{h}_{\Sigma_1}^\top(l_1(u_0, v_0, R_1)), \dots, \tilde{h}_{\Sigma_1}^\top(l_N(u_0, v_0, R_1)), \\ \tilde{h}_{\Sigma_2}^\top(l_1(u_0, v_0, R_2)), \dots, \tilde{h}_{\Sigma_2}^\top(l_N(u_0, v_0, R_2)), \\ \tilde{h}_{\Sigma_3}^\top(l_1(u_0, v_0, R_3)), \dots, \tilde{h}_{\Sigma_3}^\top(l_N(u_0, v_0, R_3)) \end{bmatrix}^\top \quad (2.8)$$

where $l_j(u_0, v_0, R)$ is a location with distance R from (u_0, v_0) in the direction given by j when the directions are quantized into N values. DAISY uses $N=8$ directions with $R_1 = 2.5, R_2 = 3R_1, R_3 = 6R_1$ and $\Sigma_1 = 2.55, \Sigma_2 = 3\Sigma_1, \Sigma_3 = 5\Sigma_1$. These parameter settings result in $8 + 8 \times 3 \times 8 = 200$ dimensional DAISY descriptors [58].

Figure 2.5 shows 25 different locations as '+' signs, which DAISY uses for the descriptor construction. A line segment points one of the directions that DAISY uses. Circles represent the circular bins, whose radii are proportional to the standard deviations of the Gaussian kernels used in DAISY. It can be seen that the circular bins give less overlap between the adjacent bins than the location bins of SIFT and GLOH, which are shown in Figure 2.2. But such circular bins result in descriptors, which obtain a performance comparable to SIFT, SURF, and GLOH, but they make the descriptor construction faster than SIFT, SURF, and GLOH.

2.2.6 LIOP

LIOP is an intensity invariant local descriptor [69]. Unlike SIFT, GLOH, DAISY, and SURF, it uses an intensity order based location binning, which is obtained by sorting all the pixel intensities of the image patch in the ascending order [12], followed by equally quantizing the image patch into B ordinal bins based on the intensity order. Figure 2.6 shows the description of an image patch based on the LIOP method, where various colors inside the image patch depict different ordinal bins.

LIOP computes intensity invariant image features for each ordinal bin. To this end, it defines $\mathcal{P}^N = \{(p_1, p_2, \dots, p_N) : p_i \in \mathbb{R}\}$ as a set of N -dimensional vectors and Π^N as a set of all

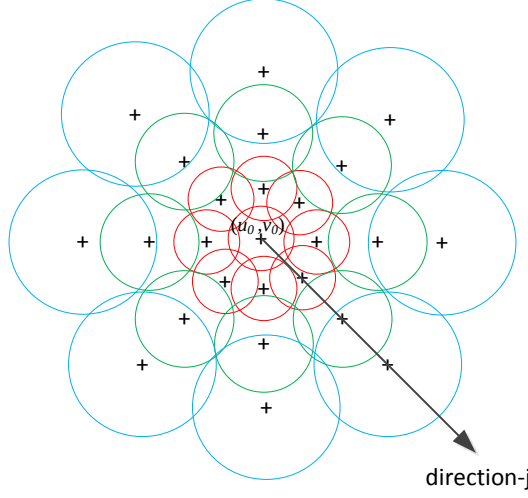


Figure 2.5: Circular bins are shown, which are used in the DAISY method. The radii of the circular bins are proportional to the standard deviations of the Gaussian kernels used in DAISY. To compute a descriptor at a pixel location (u_0, v_0) , DAISY reads values from 8 different orientation maps at 25 different locations, which are depicted in the illustration with '+' signs around (u_0, v_0) . Image from [58].

possible permutations of integers $\{1, 2, \dots, N\}$. Then, by making use of a mapping $\gamma : \mathcal{P}^N \rightarrow \Pi^N$, LIOP maps the N-dimensional vector to a permutation $\pi \in \Pi^N$ as follows [69]:

$$\gamma(P) = \pi, P \in \mathcal{P}^N, \pi \in \Pi^N \quad (2.9)$$

For instance, γ maps a vector $P = \{86, 217, 152, 101\}$ to a permutation $\pi = \{1, 4, 3, 2\}$ based on the ascending order of the elements of P . Then, by making use of another mapping ϕ , LIOP maps the permutation π to an $N!$ -dimensional feature vector $V_{N!}^i$ as:

$$\phi(\pi) = V_{N!}^i, \pi \in \Pi^N \quad (2.10)$$

where $V_{N!}^i$ is a vector, whose elements are all 0 except the i -th element, which is greater than 0. LIOP uses the above definitions and constructs $P(x)$ as an intensity vector, which contains the intensities of N neighboring samples of a pixel x . It then computes a $LIOP(x)$ feature vector as:

$$\begin{aligned} LIOP(x) &= \phi(\gamma(P(x))) \\ &= V_{N!}^i \end{aligned} \quad (2.11)$$

LIOP uses a circle of radius R around x to read the pixel intensities of N neighboring samples. It uses equally spaced neighboring samples in the anticlockwise direction. It repeats the same process on each pixel of the image patch to compute a LIOP descriptor as:

$$\begin{aligned} LIOP \text{ descriptor} &= (des_1, des_2, \dots, des_B) \\ des_j &= \sum_{x \in bin_j} LIOP(x) \end{aligned} \quad (2.12)$$

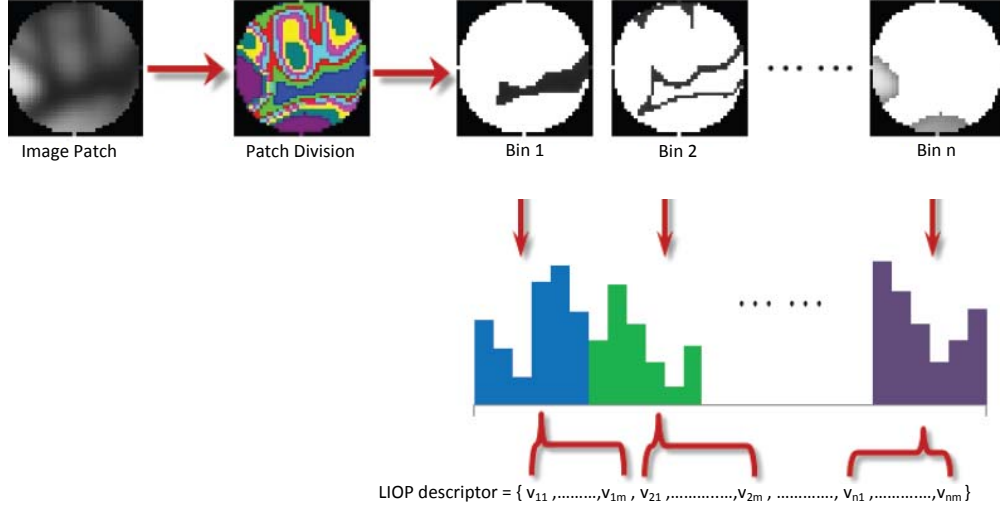


Figure 2.6: Illustration of the computation of the LIOP descriptor. Images from [69].

where B represents the number of ordinal bins and $\sum_{x \in \text{bin}_j} LIOP(x)$ is the linear sum of $LIOP(x)$ features of the j^{th} ordinal bin. LIOP concatenates descriptors (des_B) over all the ordinal bins to build a $N! \times B$ dimensional description vector for the image patch. The experimental results in [69] show that $N = 4$ and $B = 6$ result in LIOP descriptors, which demonstrate better performance than SIFT and CS-LBP under monotonic intensity changes.

2.3 Nonlinear Intensity Invariant Local Descriptors

This section presents nonlinear intensity invariant local descriptors. These descriptors are extended versions of SIFT and have been used to cope with the nonlinear intensity changes.

2.3.1 GOM-SIFT

GOM-SIFT [71] is an extended version of SIFT. It uses a Gradient Orientation Modification (GOM) as described in Equation 2.13. GOM modifies the gradient orientations (β) of each image patch and creates the restricted ones ϕ . GOM-SIFT uses such restricted gradient orientations instead of β in the SIFT method to construct the descriptors.

Yi et al. [71] show that nonlinear intensity changes affect the gradient orientations. Due to which the SIFT feature histograms of the corresponding image patches become different, which lead to dissimilar SIFT descriptors. To deal with these problems, they propose GOM and restrict the gradient orientations. They show that GOM improves the robustness of SIFT to nonlinear intensity changes and increases the performance by 7.04%.

$$\phi = \begin{cases} -\beta & \beta \in (-\pi, 0) \\ \beta & \beta \in [0, \pi] \end{cases} \quad (2.13)$$

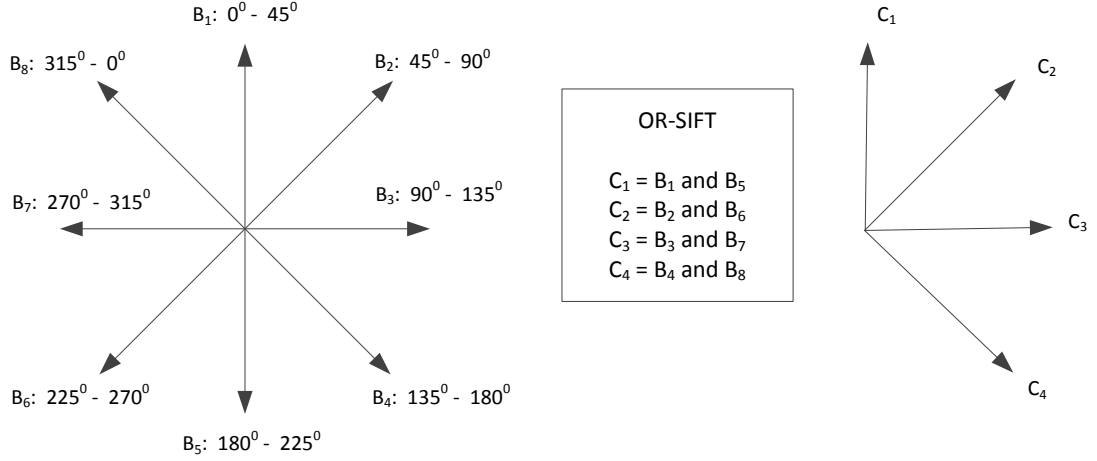


Figure 2.7: Orientation bins, which are used in the construction of SIFT (left) and OR-SIFT (right) descriptors. Images from [68].

2.3.2 OR-SIFT

Vural et al. [68] show that on one hand GOM [71] improves the robustness of SIFT, but on the other hand, it affects the rotation invariance of SIFT. This is mainly due to the orientation mapping, which is described in Equation 2.13. To deal with this problem, they propose an Orientation Restricted (OR) method [68], which first, computes the SIFT descriptors [33] and then combines the elements of the SIFT descriptors in the opposite orientation directions to obtain the OR-SIFT descriptors.

Figure 2.7 illustrates the OR method. $B_1, B_2, \dots, B_8$ are 8 quantized orientation bins of a SIFT feature histogram. A SIFT descriptor consists of 16 such feature histograms. In OR method, each feature histogram of the SIFT is transformed into an OR-SIFT feature histogram by combining the histogram bins of the SIFT in the opposite orientation directions. For instance, C_1 bin of OR-SIFT is obtained by summing the values in B_1 and B_5 bins of the SIFT feature histogram. Vural et al. show that such an OR method improves the robustness of SIFT and demonstrates better performance than SIFT and GOM-SIFT.

2.4 Summary

In this chapter an overview of the state-of-the-art feature point detectors and descriptors, were presented. The chapter described that how feature point detectors and descriptors have evolved over the time. It described the basic Moravec [41] and Harris [18] corner detectors and then the state-of-the-art feature point detectors, such as SIFT [33], Harris Laplace [37], and SURF [2]. Each detector creates a list of feature points, which contains information, such as location and the detection scale of the feature points. Such attributes hold useful information, i.e., where these feature points are coming from, but are insufficient for establishing the feature point cor-

respondences between the images. To address this problem, the description of feature points has been widely used.

The chapter described various methods for the description of feature points. It described the SIFT descriptor in detail. It also described the modified versions of SIFT, which have been widely used to improve the robustness of SIFT to image transformations, such as intensity (monotonic and nonlinear) changes [19, 38, 68, 71] or to expedite the construction of SIFT to meet the needs of the real time applications [2, 6, 58].

Methodology

This chapter presents the methodology used in the construction of new local descriptors. The new local descriptors are given below:

- LBPG [51]
- LC-SIFT [52]
- DE-SIFT [52]
- NG-SIFT [53]
- MN-SIFT [49]

In the following sections the construction of new local descriptors are described in detail.

3.1 Harris Laplace Points

Harris Laplace points [37] are used as feature points in this thesis to extract normalized image patches from the images for the construction of new as well as the state-of-the-art local descriptors. Mikolajczyk and Schmid [38] show that the ranking of the local descriptors is mostly independent of the feature point detector. Therefore, Harris Laplace detector is only used in this thesis. Harris Laplace detects corner like image structures [37], whereas Hessian Laplace and Difference of Gaussian detect blob-like image structures [33].

Harris Laplace points are invariant to scale and rotation changes. They are detected with a scale adapted Harris function as [18, 37]:

$$L(\mathbf{x}, s_n) = (s_n)^n G(s_n) * I(\mathbf{x}) \quad (3.1)$$

where $\mathbf{x} = (x, y)$ is a pixel location of a gray scale image $I(\mathbf{x})$ and $L(\mathbf{x}, s_n)$ is an n^{th} level image in the scale space, which is created through a convolution $(*)$ of $I(\mathbf{x})$ with a Gaussian kernel

$G(s_n)$. The term $(s_n)^n$ is used for scale normalization, where s_n represents the successive level of the scale-space representation [32], which is obtained as $s_n = k^n s_0$, where s_0 is an initial scale factor at the finest level of the resolution and k is a factor of scale change between the successive levels $n = \{0, 1, 2, \dots, m\}$ [37].

To detect the Harris Laplace points, the Harris function $F_H(x, s_n)$ is assessed at each scale space image as follows [37]:

$$C(x, s_n) = \begin{bmatrix} D_x D_x(x, s_n) & D_x D_y(x, s_n) \\ D_x D_y(x, s_n) & D_y D_y(x, s_n) \end{bmatrix} \quad (3.2)$$

$$F_H(x, s_n) = \det(C(x, s_n)) - \alpha \text{trace}^2(C(x, s_n)) \quad (3.3)$$

where $D_\bullet(x, s_n) = (s_n)^2 G(x, \tilde{s}) * L_\bullet(x, s_n)$ and $\bullet$ represents the directional gradients in x and y directions. The gradients are smoothed with a Gaussian kernel $G(x, \tilde{s})$. Then, local maxima detection is performed as follows [37]:

$$F_H(x, s_n) > F_H(x_w, s_n) \quad \forall x_w \in W \quad (3.4)$$

where W is an 8 neighborhood of a pixel x . To verify that each detected local maximum also form maximum in the scale direction, a Laplacian function $F_L(x, s_n)$ is computed as follows [37]:

$$F_L(x, s_n) = (s_n)^2 [L_{xx}(x, s_n) + L_{yy}(x, s_n)] \quad (3.5)$$

The Laplacian values in the adjacent scales are compared and the local maxima, which show stability along the scale direction are retained, and the non maxima are suppressed as [37]:

$$F_L(x, s_n) > F_L(x, s_{n-1}) \wedge F_L(x, s_n) > F_L(x, s_{n+1}) \quad (3.6)$$

Finally, a pixel x is declared as a Harris Laplace point if it simultaneously satisfies Equations 3.4 and 3.6.

Two thresholds t_h and t_l are used to reject weak Harris Laplace points i.e., the points that hold Harris and Laplacian values below t_h and t_l , respectively, as [37]:

$$F_H(x, s_n) < t_h \wedge F_L(x, s_n) < t_l \quad (3.7)$$

3.2 Harris Laplace Regions

Harris Laplace (HarLap) [38] regions are normalized image patches centered at Harris Laplace points. They are used in this thesis for the construction of new descriptors (LBPG, LC-SIFT, DE-SIFT, NG-SIFT and MN-SIFT) as well as the state-of-the-art local descriptors (SIFT, GLOH, SURF, DAISY, CS-LBP, LIOP, GOM-SIFT and OR-SIFT).

A method proposed in [38] is used for the computation of the HarLap regions. The method selects an image patch centered at each Harris Laplace point according to the detection scale (s_n) of the point. It resizes the selected patch to obtain a fixed size patch of 41×41 pixels and then, normalizes the intensity values of this fixed size patch to obtain the new intensities between 0 and 1. Such a resized and intensity normalized image patch is referred to as a HarLap region. Figure 3.1 illustrates the computation of a HarLap region, where an image patch centered at a Harris Laplace point ('+') is shown inside a bounding box. The image patch is resized and intensity normalized to obtain a HarLap region, which is shown in Figure 3.1(c).

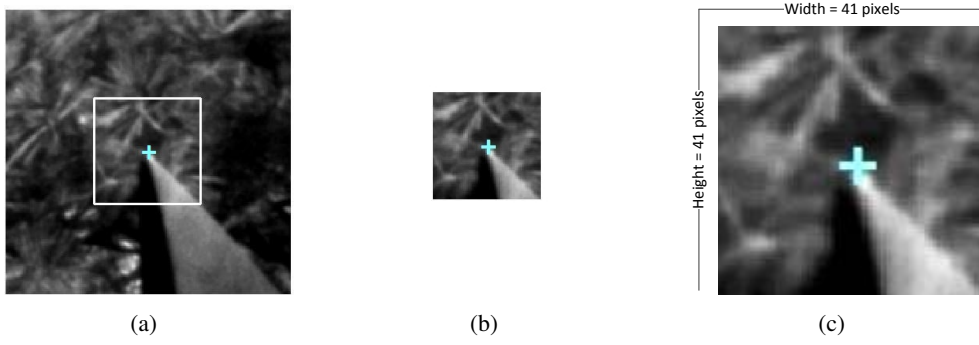


Figure 3.1: (a) An image patch centered at a Harris Laplace point is shown inside a bounding box, (b) the patch is cropped from the image, (c) resized and intensity normalized to obtain a HarLap region of constant size 41×41 pixels.

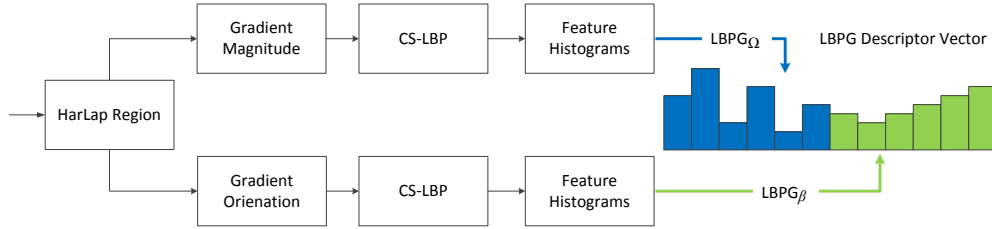


Figure 3.2: Illustration of the computation of the LBPG descriptor.

3.3 LBPG

LBPG [51] is a new method for the construction of local descriptors. It combines the strengths of CS-LBP [19] and SIFT [33] methods to overcome the effects of nonlinear intensity changes. The CS-LBP method overcomes the effects of intensity changes by making use of CS-LBP features. It uses such features instead of image gradients in the SIFT-like feature histogram scheme to construct the descriptors. It demonstrates good performance under monotonic intensity changes [19], but demonstrates a lower performance with regard to nonlinear intensity changes [51]. To cope with this problem, image gradients are used instead of image intensity [19] in the LBPG method to compute the CS-LBP features. The LBPG method is illustrated in Figure 3.2. In the following subsections, each block of the illustration is described.

3.3.1 Image Gradients

To compute the image gradients, LBPG convolves each HarLap region with $[-1, 0, 1]$ and $[-1, 0, 1]^T$ kernels and obtains directional gradients F_h and F_v , respectively. Then, it computes

the gradient magnitudes (Ω) and gradient orientations (β) as follows:

$$\Omega = \sqrt{F_h^2 + F_v^2} \quad (3.8)$$

$$\beta = \text{atan2}(F_v, F_h) \quad (3.9)$$

3.3.2 LBPG Feature

Having computed the gradients, LBPG applies the CS-LBP scheme [19] on image gradients (Ω and β), as described in Equations 3.10 and 3.11, respectively. This results in LBPG_Ω (magnitude) and LBPG_β (orientation) features, where Ω_i represents a gradient magnitude sample at a distance of $2R$ from its center symmetric neighbor $\Omega_{i+(N/2)}$. $s(z)$ binarizes the differences between center symmetric neighbors. The binarized differences are weighted and summed over all the center symmetric pairs to obtain a LBPG_Ω feature, as described in Equations 3.10. This process is performed on each Ω feature of the HarLap region by using the feature as a sample n_c (see Figure 2.4).

Similarly, the β features of the HarLap region are also processed in the same manner and LBPG_β features are obtained, as described in Equation 3.11, where β_i and $\beta_{i+(N/2)}$ are center symmetric β neighbors. Parameter N represents the number of samples and generates $2^{N/2}$ distinct binary patterns for both LBPG_Ω and LBPG_β features. In the LBPG method $N = 6$ is used and $R = 2$ is borrowed from [19].

$$\text{LBPG}_{\Omega, R, N} = \sum_{i=0}^{(N/2)-1} s(\Omega_i - \Omega_{i+(N/2)}) 2^i \quad (3.10)$$

$$\text{LBPG}_{\beta, R, N} = \sum_{i=0}^{(N/2)-1} s(\beta_i - \beta_{i+(N/2)}) 2^i \quad (3.11)$$

$$s(z) = \begin{cases} 1 & z \geq 0.01 \\ 0 & \text{otherwise} \end{cases} \quad (3.12)$$

3.3.3 LBPG Feature Histograms

Having computed the LBPG features, the HarLap region is divided into 4×4 location bins [33]. For each location bin, two feature histograms are computed, one by using the LBPG_Ω features and the other one by using the LBPG_β features of the location bin. The feature histograms are concatenated over all the location bins to build two descriptors, i.e., LBPG_Ω and LBPG_β , each descriptor is 128 dimensional.

3.3.4 LBPG Descriptor

Finally, LBPG_Ω and LBPG_β descriptors are concatenated and an LBPG descriptor of length $128 + 128 = 256$ is obtained. The descriptor is then normalized to unit norm. Larger value descriptor elements are limited to 0.2 and then the descriptor is renormalized to unit norm [33, 38].

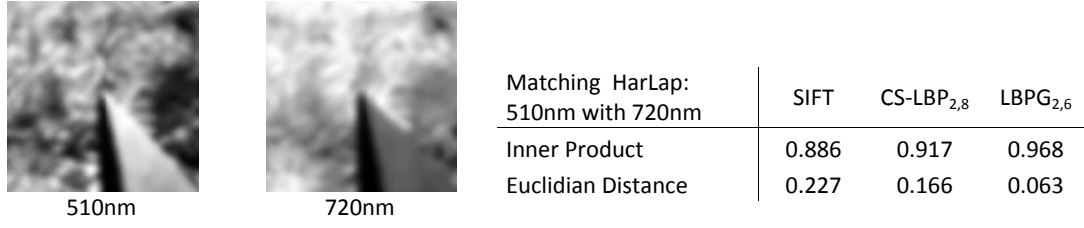


Figure 3.3: Comparison of SIFT [33], CS-LBP [19], and LBPG [51] descriptors based on a HarLap region pair, which depicts the same scene contents in 510nm and 720nm wavelength bands. The descriptor matching scores based on Inner product and Euclidean distance, show that LBPG outperforms SIFT and CS-LBP descriptors.

3.3.5 LBPG Example

To illustrate the robustness of LBPG method towards nonlinear intensity changes, two HarLap regions are shown in Figure 3.3. The regions depict the same scene contents in 510nm and 720nm (wavelength) bands, but the scene contents look different due to nonlinear intensity changes between the HarLap regions. LBPG, SIFT, and CS-LBP descriptors are computed for the HarLap regions and then these descriptors are compared with Inner Product and Euclidean distance. The descriptor matching scores, which are shown in Figure 3.3, show that LBPG obtains better descriptor matching scores than SIFT and CS-LBP descriptors, i.e., 0.968 (Inner product) and 0.063 (Euclidean distance). In the case of a perfect match, Inner product and Euclidean distance metrics provide matching scores of 1 and 0, respectively, for unit norm descriptors.

3.4 LC-SIFT

LC-SIFT [52] is an extended version of SIFT [33]. It is based on boosting the contribution of local edges in the SIFT method to overcome the effects of nonlinear intensity changes. Such an edge boosting is obtained by replacing the gradient magnitudes in the SIFT method with LC (Local Contrast) magnitudes [56].

LC-SIFT benefits from an observation that intensity changes between the images of the same scene occur due to varying lighting conditions, but such intensity changes little affect the image edges. The relative change in the intensity across the edges is preserved. If gradient magnitudes are used, they give dissimilar SIFT descriptors [19, 69]. This is because of gradient magnitudes, which depend on local intensity differences and such differences vary with intensity changes. Heikkilä et al. use the CS-LBP method [19] to cope with the effects of intensity changes on the gradient magnitudes. They binarize the intensity differences with a $s(z)$ function, as described in Equation 2.5 and compute the CS-LBP features. They use CS-LBP features instead of image gradients in the SIFT-like feature histogram scheme to compute the descriptors. In contrast to CS-LBP method, LC-SIFT uses LC magnitudes instead of gradient magnitudes to overcome the effects of intensity changes.

3.4.1 LC Magnitudes

To compute the LC magnitudes, an edge detector is used [56]. This edge detector computes the magnitude of edges invariant to image contrast and intensity changes. These edges magnitudes are referred to as LC magnitudes. The LC magnitudes for a HarLap region R are obtained as follows [56]:

$$LC(x, y) = \frac{R_{max}(x, y) - R_{min}(x, y)}{R_{max}(x, y) + R_{min}(x, y) + v} \quad (3.13)$$

where R_{min} and R_{max} are minimum and maximum pixel intensities, which are computed in a 3×3 window centered at each pixel (x, y) of R and v is an infinitely small positive number to avoid a division by zero. In fact, LC magnitudes are normalized edge magnitudes, which equally boost the contribution of weak and strong edges in the SIFT method and improve the performance of SIFT as well as demonstrate better robustness to intensity changes than CS-LBP features and gradient magnitudes based descriptors.

3.5 DE-SIFT

DE-SIFT [52] is an extended version of SIFT. It is inspired by LC-SIFT. It uses an edge boosting similar to LC-SIFT. This edge boosting is obtained by using an illumination invariant edge detector [8], which computes image edges for a HarLap region R as follows:

$$d(x, y) = \sum_{i=-1}^{i=1} \sum_{j=-1}^{j=1} R(x + i, y + j) \quad (3.14)$$

$$DE(x, y) = atan2(d(x, y) - 9R(x, y), R(x, y)) \quad (3.15)$$

where $d(x, y)$ is a local intensity sum, (x, y) represents a pixel location and $atan2$ limits the DE magnitudes between $-\pi/2$ and $\pi/2$ radians, which are then mapped with a proposed linear mapping, i.e., $DE(x, y) := \pi/2 + DE(x, y)$, between 0 and π radians, in order to compute the DE-SIFT descriptors by using positive DE magnitudes [52].

3.5.1 DE-SIFT and LC-SIFT Examples

To illustrate the robustness of LC and DE magnitudes to intensity changes, two HarLap regions are shown in Figure 3.4. The regions depict the same scene contents in 460nm and 720nm bands. The intensity changes between the regions can be seen, which more affect the gradient magnitudes than LC and DE magnitudes. As a result the descriptor matching scores, which are shown in Figure 3.4, suggest lower performance of SIFT than LC-SIFT and DE-SIFT. LC and DE magnitudes in contrast to gradient magnitudes, demonstrate more invariance and robustness to intensity changes. They equally enhance the weak and strong edges of the regions. This results in better descriptor matching scores of LC-SIFT and DE-SIFT than SIFT. In the case of a perfect match, the Euclidean distance and Inner product metrics give matching scores of 0 and 1, respectively, for unit norm descriptors.

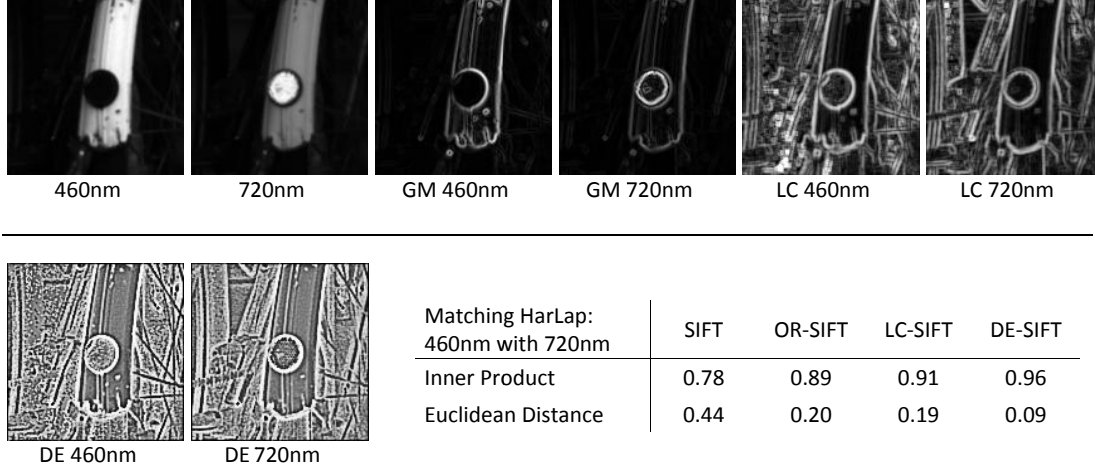


Figure 3.4: Comparison of Gradient Magnitude (GM), LC, and DE magnitudes based on a HarLap region pair. The pair depicts the same scene contents in 460nm and 720nm bands. LC-SIFT and DE-SIFT are based on LC and DE magnitudes, respectively. They demonstrate better descriptor matching scores than OR-SIFT [68] and SIFT [33].

3.6 NG-SIFT

NG-SIFT is an extended version of SIFT. It is inspired by LC-SIFT and DE-SIFT. It is based on Normalized Gradients (NG) to overcome the effects of nonlinear intensity changes. GOM-SIFT [71] and OR-SIFT [68] use restricted gradient orientations, to make SIFT robust to nonlinear intensity changes, whereas, LC-SIFT and DE-SIFT use LC and DE magnitudes, respectively, instead of gradient magnitudes in the SIFT method. In contrast to these local descriptors, NG-SIFT uses normalized gradients, which are fixed magnitude image features, i.e., either 0 or 1, and demonstrate more robustness to nonlinear intensity changes.

3.6.1 Normalized Gradients

To compute the normalized gradients for NG-SIFT, each HarLap region is first, convolved with $[-1, 0, 1]$ and $[-1, 0, 1]^T$ kernels to obtain the directional gradients F_h and F_v , respectively. The gradient magnitudes and orientations are computed, as described in Equations 3.8–3.9. Then, the gradients are normalized as:

$$\hat{F}_h(x, y) = \frac{F_h(x, y)}{\Omega(x, y) + v} \quad (3.16)$$

$$\hat{F}_v(x, y) = \frac{F_v(x, y)}{\Omega(x, y) + v} \quad (3.17)$$

$$\hat{\Omega}(x, y) = \sqrt{\hat{F}_h^2(x, y) + \hat{F}_v^2(x, y)} \quad (3.18)$$

$$\hat{\beta}(x, y) = \text{atan2}(\hat{F}_v(x, y), \hat{F}_h(x, y)) \quad (3.19)$$

where (x, y) is a pixel location and v is an infinitely small positive number, which is added to avoid a division by zero. Note that, the gradient normalization used in Equations 3.16–3.17, does not affect the gradient orientations, i.e., $\hat{\beta}(x, y) = \beta(x, y)$. However, after the normalization the magnitude of every gradient sample becomes $\hat{\Omega}(x, y) = 1$ except for $\Omega(x, y) = 0$, which results in $\hat{\Omega}(x, y) = 0$. Here, the idea is to avoid the modification of gradient orientations, because LC-SIFT and DE-SIFT demonstrate better performance than GOM-SIFT [71] and OR-SIFT [68] because they do not modify the gradient orientations, only replace the gradient magnitudes in the SIFT method with LC and DE magnitudes. NG-SIFT also uses the same idea and uses the normalized image gradients instead of image gradients to compute the descriptors.

3.6.2 Proposed SIFT-like descriptor construction scheme

To use the normalized gradients $(\hat{\Omega}, \hat{\beta})$ for the construction of NG-SIFT descriptors, a SIFT-like descriptor construction scheme is proposed. In the scheme proposed, each gradient sample contributes only a normalized gradient magnitude to its corresponding feature histogram bin, unlike SIFT, which uses a Gaussian weighted gradient magnitude for this purpose and also distributes such a weighted magnitude to adjacent histogram bins via a soft binning scheme [33]. This makes the scheme proposed simple and computationally less intensive than SIFT.

To further elaborate on the scheme proposed, Figure 3.5(a) shows an HarLap region H . The size of this region is 41×41 pixels and a location grid is used to divide it into 4×4 location bins. A location bin is denoted with $H_{r,c}$ ($r = 0, 1, 2, 3$ and $c = 0, 1, 2, 3$) and its pixels as $\{H(x, y) : x \in [l_c, u_c] \wedge y \in [l_r, u_r]\}$ where,

$$l_c = (41 - 1)c/4 \quad (3.20)$$

$$u_c = (41 - 1)(c + 1)/4 \quad (3.21)$$

$$l_r = (41 - 1)r/4 \quad (3.22)$$

$$u_r = (41 - 1)(r + 1)/4 \quad (3.23)$$

For instance, $H_{0,3}$ bin consists of $\{H(x, y) : x = 30, 31, \dots, 40 \wedge y = 0, 1, \dots, 10\}$ pixels.

To construct a NG-SIFT descriptor for H , the normalized gradients $(\hat{\Omega}, \hat{\beta})$ are computed by using Equations 3.16–3.19 and then, the $\hat{\beta}$ features are quantized as follows:

$$L(x, y) = \text{mod}(\lfloor \frac{\hat{\beta}(x, y)}{2\pi/8} + \frac{1}{2} \rfloor, 8) \quad (3.24)$$

where $\text{mod}(\cdot)$ performs a modular arithmetic, for instance, $\hat{\beta}(x, y) = -\pi/2$ after such a modular arithmetic becomes $L(x, y) = 6$. Then, a feature histogram $h_{r,c,t}$ is computed for $H_{r,c}$ bin as:

$$h_{r,c,t} = \sum_{x=l_c}^{u_c} \sum_{y=l_r}^{u_r} \hat{\Omega}(x, y) \delta(L(x, y) = t) \quad (3.25)$$

where $t = 0, 1, \dots, 7$ and $\delta(\cdot)$ is defined as:

$$\delta(z) = \begin{cases} 1 & \text{z is true} \\ 0 & \text{otherwise} \end{cases} \quad (3.26)$$

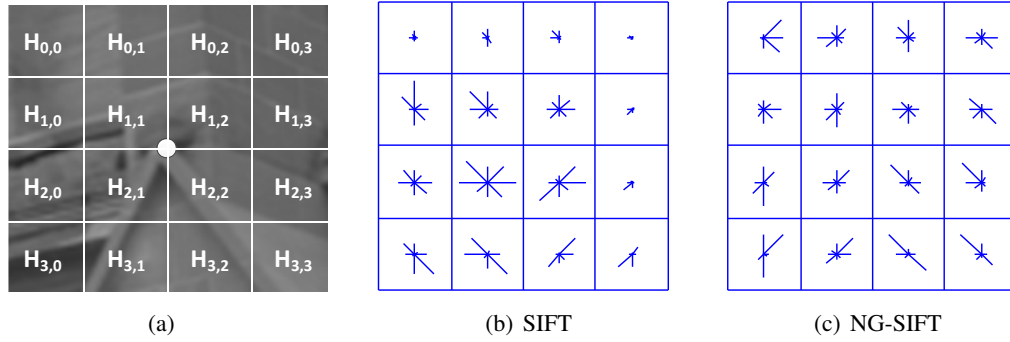


Figure 3.5: (a) Illustration of 4×4 location bins, which are used in the construction of NG-SIFT. The elements of SIFT and NG-SIFT descriptors of the HarLap region, which is shown in (a) are depicted with line segments [65] inside 4×4 location bins in (b) and (c), respectively.

Finally, the feature histograms are concatenated over all the location bins to obtain a 128 dimensional NG-SIFT descriptor for the HarLap region.

3.6.3 NG-SIFT Example

To illustrate the difference between SIFT and NG-SIFT, in Figures 3.5(b) and (c), SIFT and NG-SIFT descriptors of H (Figure 3.5(a)), respectively, are shown. The elements of SIFT and NG-SIFT descriptor are shown with line segments inside 4×4 location bins [65] for the sake of figure clarity. It can be seen that the elements of SIFT descriptor are larger if a location bin contains strong image edges, whereas for the same location bin, NG-SIFT shows invariance to image edges and their magnitudes due to the normalized gradients of the location bin (Equation 3.25). Here, the idea is to overcome the effects of nonlinear intensity changes, which affect the image edges, their magnitudes, and result in dissimilar SIFT descriptors for the corresponding HarLap regions belonging to different band images of the same scene, such as shown in Figure 3.6, where a HarLap region pair belonging to Blue and NIR band images is shown. It can be seen that the nonlinear intensity changes between the corresponding Harris Laplace regions of the Blue and NIR band images, more affect the SIFT descriptor than NG-SIFT.

3.7 MN-SIFT

MN-SIFT is an extended version of NG-SIFT. As described, NG-SIFT is based on normalized gradients, which are fixed magnitude image features i.e., either 0 or 1. Such binary magnitude features result in robust NG-SIFT descriptors; but on the other hand they degrade the performance of NG-SIFT on the images of textured scenes [53]. To achieve better performance on the images of textured scenes, while keeping the descriptors robust to nonlinear intensity changes, Modified Normalized (MN) gradients ($\bar{\Omega}, \beta$) are proposed. These MN gradients are used as image features instead of NG in the NG-SIFT method to construct the MN-SIFT descriptors. The

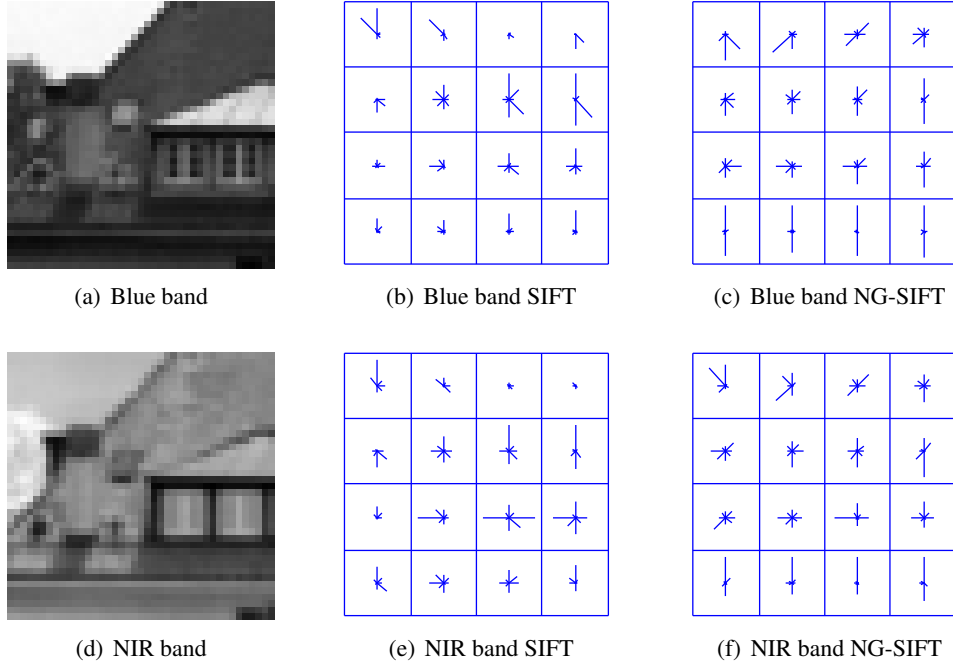


Figure 3.6: First row shows a HarLap region belonging to a Blue band image, SIFT, and NG-SIFT descriptors, whereas, the second row shows the corresponding NIR band image HarLap region, SIFT, and NG-SIFT descriptors.

MN gradients are computed as follows:

$$\bar{\Omega}(x, y) = \frac{\Omega(x, y) - \Omega_{min}}{\Omega_{max} - \Omega_{min}}; \quad \beta(x, y) = atan2(F_v(x, y), F_h(x, y)) \quad (3.27)$$

where $\Omega_{min} = \min\{\Omega(\bar{x}) : \bar{x} \in (x, y), 0 \leq x \leq 40, 0 \leq y \leq 40\}$ and $\Omega_{max} = \max\{\Omega(\bar{x}) : \bar{x} \in (x, y), 0 \leq x \leq 40, 0 \leq y \leq 40\}$ are minimum and maximum Ω values of the HarLap region. Ω and β features of the HarLap region are computed with Equations 3.8 and 3.9, respectively. Compared to $\hat{\Omega}$ (Equation 3.18), $\bar{\Omega}$ takes on values between 0 and 1. The experimental results [49] show that $\bar{\Omega}$ based descriptors (i.e., MN-SIFT) demonstrate more robustness to nonlinear intensity changes than SIFT, NG-SIFT, LC-SIFT, DE-SIFT, and LBPG on textured as well as structured scene images.

3.8 Comparison

This section presents the similarities and differences between the new descriptors (LBPG, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) and the state-of-the-art local descriptors (SIFT, GLOH, SURF, CS-LBP, DAISY, LIOP, OR-SIFT, and GOM-SIFT).

LBPG is a gradient based descriptor. It uses CS-LBP modified gradients, which are referred to as $LBPG_{\bar{\Omega}}$ and $LBPG_{\beta}$ features, in order to overcome the effects of nonlinear intensity

changes. To read such modified gradients from HarLap regions, LBPG performs 4×4 location binning of each HarLap region twice and computes LBPG_Ω and LBPG_β descriptors for the region. It borrows 4×4 location bins from SIFT. LBPG is a different method compared to CS-LBP, because it uses image gradients to compute the CS-LBP features, whereas CS-LBP method uses image intensity. It is also different from OR-SIFT and GOM-SIFT, which are modified gradient orientation based descriptors. LBPG computes 256 dimensional description vectors.

LC-SIFT is a gradient based descriptor. It uses LC magnitudes (Equation 3.13) in the SIFT method instead of gradient magnitudes to compute the descriptors. It uses 4×4 location bins. It quantizes the gradient orientations into 8 different levels and then, constructs the feature histograms by using the sum of Gaussian weighed LC-magnitudes. The LC magnitudes equally boost the contribution of weak and strong edges in the SIFT method, and demonstrate better performance than gradient magnitudes. LC-SIFT constructs 128 dimensional description vectors. Unlike OR-SIFT and GOM-SIFT, it is a modified gradient magnitude based descriptor.

DE-SIFT is a gradient based descriptor. It uses DE magnitudes, which are described in Equation 3.15, in the SIFT method instead of gradient magnitudes to compute the descriptors. It uses 4×4 location bins. It quantizes the gradient orientations into 8 different levels and then, by using the sum of Gaussian weighed DE-magnitudes, it constructs the feature histograms. DE magnitudes equally boost the contribution of weak and strong edges in the SIFT method and demonstrate better performance than gradient magnitudes. DE-SIFT constructs 128 dimensional description vectors. It is a modified gradient magnitude based descriptor, unlike OR-SIFT and GOM-SIFT.

NG-SIFT is a gradient based descriptor. It uses normalized gradients as image features and 4×4 location bins to compute the descriptors. It is based on a proposed SIFT-like descriptor construction scheme, which quantizes the gradient orientations into 8 different levels and uses the sum of normalized gradient magnitudes for the construction of feature histograms. NG-SIFT constructs 128 dimensional description vectors. It is a modified gradient magnitude based descriptor, in contrast to OR-SIFT and GOM-SIFT.

MN-SIFT is an extended version of NG-SIFT. It uses MN gradients, which are described in Equation 3.27, in the NG-SIFT method to compute the descriptors. It quantizes the gradient orientation into 8 different levels and uses sum of MN gradient magnitudes for the construction of feature histograms. It computes 128 dimensional description vectors.

SIFT is a gradient based descriptor. It uses 4×4 location bins, quantizes the gradient orientation into 8 different levels and uses the sum of Gaussian weighed gradient magnitudes for the construction of feature histograms. SIFT constructs 128 dimensional description vectors.

GLOH is an extended version of SIFT. It uses log polar location bins. It quantizes the gradient orientation into 16 different levels and computes the feature histograms by using the sum of gradient magnitudes. It computes 272 dimensional descriptors and then uses PCA to reduce the descriptor dimensions to 128. For descriptors proposed, log polar location bins and PCA are not used.

SURF is a Haar wavelet based descriptor. It uses 4×4 location bins. It uses Gaussian weighting similar to SIFT for the construction of descriptors. For the descriptors proposed, the Haar wavelet responses are not used. This thesis uses 128 dimensional extended SURF to evaluate the performance of SURF against nonlinear intensity changes.

CS-LBP is an extended version of SIFT. It uses a modified Local Binary Patterns scheme to compute the CS-LBP features. It reads such image features from the HarLap regions by making use of SIFT-like feature histogram scheme. It computes 256 dimensional description vectors.

DAISY uses image gradients and circular location bins to compute the descriptors. DAISY computes 200 dimensional description vectors. For descriptors proposed, circular location bins are not used.

LIOP is an intensity based descriptor. It uses intensity based location bins. LIOP computes 144 dimensional description vectors. For descriptors proposed, intensity based location bins are not used.

GOM-SIFT is a modified version of SIFT. It uses a gradient orientation modification, as described in Equation 2.13 to cope with nonlinear intensity changes. It uses 4×4 location bins and uses the sum of Gaussian weighed gradient magnitudes for the construction of feature histograms. It computes 128 dimensional description vectors.

OR-SIFT is a modified version of SIFT. It uses a method based on two steps to deal with the nonlinear intensity changes. In the first step it computes SIFT descriptors and then, in the second step it combines the elements of the SIFT descriptors in the opposite orientation directions to compute the OR-SIFT descriptors. It computes 128 dimensional description vectors.

3.9 Summary

In this chapter, five new local descriptors (LBG, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) were presented. The new descriptors are based on novel image features, which are used in the SIFT-like feature histogram schemes to compute the descriptors in order to overcome the effects of nonlinear intensity changes.

In the case of LBP, such features are obtained by applying the CS-LBP scheme on image gradients and then, using such features in the SIFT-like feature histogram scheme to compute the descriptors. In the case of LC-SIFT and DE-SIFT, LC and DE magnitudes are used instead of gradient magnitudes in the SIFT method to compute the descriptors. LC-SIFT and DE-SIFT are modified gradient magnitude based descriptors, in contrast to modified gradient orientation based GOM-SIFT and OR-SIFT descriptors.

In the case of NG-SIFT, normalized gradients are used as image features in a proposed SIFT-like descriptor construction scheme to compute the descriptors. Due to the fixed magnitude nature of the normalized gradients, i.e., 0 or 1, NG-SIFT demonstrates a lower performance on the images of textured scenes. To deal with this problem, MN-SIFT is proposed, which uses modified normalized gradients and demonstrates better performance than NG-SIFT on textured as well as structured scene images.

Image Matching

This chapter presents the experimental results for image matching. Image matching is a task, which is performed in this thesis between the pairs of images under nonlinear intensity changes by using the local descriptors. Two different descriptor matching strategies are used [38]: (i) nearest neighbor, and (ii) distance threshold. Each descriptor matching strategy computes descriptor matches between the pairs of images in order to evaluate the performance of the descriptors proposed (LBP, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) and to compare their performance measures with that of the state-of-the-art local descriptors (SIFT, OR-SIFT, GOM-SIFT, CS-LBP, GLOH, DAISY, and LIOP). Mikolajczyk and Schmid use the same descriptor matching strategies to evaluate the performance of local descriptors for image matching [38], but on gray scale images and under monotonic intensity changes. Similarly, Heikkilä et al. [19] and Wang et al. [69] use the same descriptor matching strategies to evaluate the performance of local descriptors on gray scale images under monotonic intensity changes. In contrast to [19, 38, 69], this chapter presents the performance evaluation of the local descriptors on the images of different wavelength bands, i.e., under nonlinear intensity changes. The objective is to find out the best local descriptor among the state-of-the-art local descriptors for image matching under nonlinear intensity changes and then use it as a reference to evaluate the performance of the descriptors proposed.

4.1 Image Matching Framework

Figure 4.1 illustrates the image matching framework used in this thesis. Mikolajczyk and Schmid [38] propose this image matching framework for gray scale images. In this thesis it is extended to evaluate the performance of local descriptors on multispectral images, i.e., under nonlinear intensity changes. The framework has been extensively used for evaluating the performance of new local descriptors or when a comparison among the existing local descriptors is required [2, 19, 22, 69].

The framework shown in Figure 4.1, performs image matching between pairs of different band images of the same scene. Four different multispectral image datasets are used, which are:

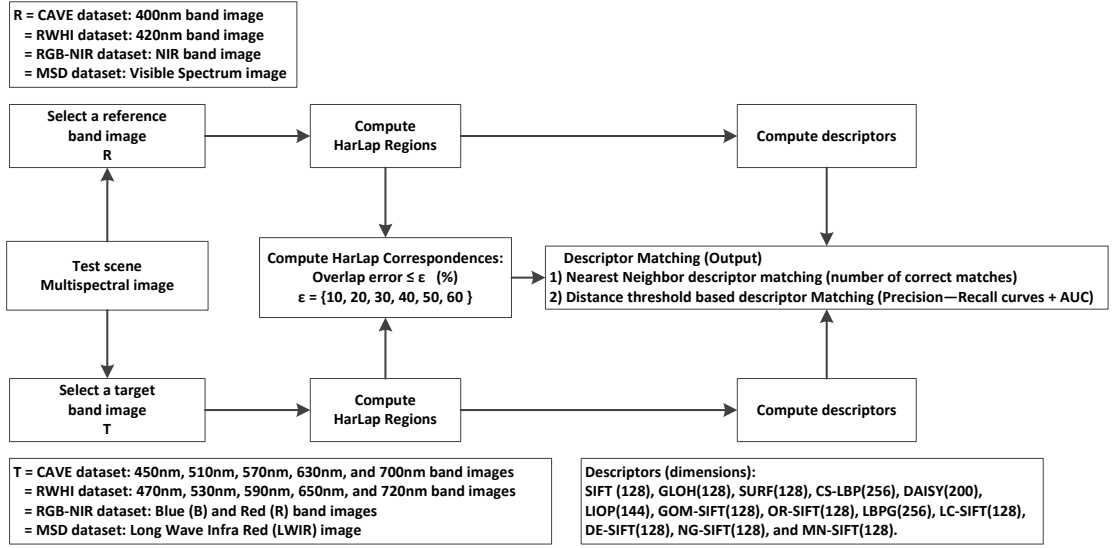


Figure 4.1: Framework for the performance evaluation of the local descriptors for image matching under nonlinear intensity changes.

CAVE [70], RWHI [7], RGB-NIR [5], and MSD [1]. The total number of multispectral images in these datasets is 659. The multispectral images depict various indoor and outdoor scenarios. These four datasets are selected in order to make the performance evaluation independent of the dataset bias [60].

Image matching begins with choosing a band image of a scene as a reference image and then performing its image matching with different band images of the same scene. HarLap regions are used for the construction of local descriptors, as described in Section 3.2. Then, HarLap region correspondences are established by using the homographies between the images. Homographies also act as ground truth data to identify the correct and false descriptor matches. Descriptor matches are obtained with a nearest neighbor and a distance threshold based descriptor matching strategies [38]. Nearest neighbor descriptor matching strategy gives the number of correct nearest neighbor descriptor matches for performance comparison, whereas the distance threshold based descriptor matching uses a set of distance thresholds, and computes Precision-Recall curves for performance comparison. Mikolajczyk and Schmid show that feature point detectors, i.e., HarLap or Hessian Laplace do not affect the ranking of the local descriptors for image matching and any of these detectors can be used [38]. Therefore, in this thesis only HarLap detectors is used. HarLap detects corner like image structures, whereas Hessian Laplace detects blob like image structures, similar to SIFT keypoint detector [33].

4.2 Evaluation Criteria

To understand the image matching framework, let K be a homography between I_r (reference) and I_t (target) images. The framework uses K for establishing correspondences between the

HarLap regions of I_r and I_t with the help of an overlap error criterion [38]. The overlap error (ϵ) measures, how well two regions n and m correspond under K by using the ratio of intersection to union of the regions:

$$\epsilon = 1 - \frac{n \cap (K^T m K)}{n \cup (K^T m K)} \quad (4.1)$$

n and m are considered as a corresponding HarLap regions if overlap error between them is less than or equal to 50% as:

$$G(n, m) = \begin{cases} 1 & \epsilon \leq 50\% \\ 0 & otherwise \end{cases} \quad (4.2)$$

where G is a ground truth matrix, $n \in \{0, 1, 2, \dots, N - 1\}$ and $m \in \{0, 1, 2, \dots, M - 1\}$. The number of detected HarLap regions in I_r and I_t are denoted with N and M , respectively. Then, framework compares the description vectors with Euclidean distance ($\| \cdot \|$) as:

$$D(n, m) = \| D_n - D_m \| \quad (4.3)$$

where D_n and D_m are the description vectors of the n^{th} and m^{th} HarLap regions of I_r and I_t , respectively. Then, a matching matrix D_w is computed as:

$$D_w(n, m) = \begin{cases} 1 & D(n, m) \leq w \\ 0 & otherwise \end{cases} \quad (4.4)$$

where w is a distance threshold.

4.2.1 Precision – Recall Curve

The framework uses a set of distance thresholds to compute the D_w matrix (see Equation 4.4). To this end, it uses a nearest neighbor based descriptor matching strategy, which first ranks the nearest neighbors by sorting the Euclidean distances between the neighbors in the ascending order [38] and then selects a first distance threshold, which is 10th on the ranking list to compute the D_w matrix, then a second threshold as 30th on the list, and so on. Then, it creates a Precision–Recall curve for image matching by using the following measures:

$$Recall = \frac{N_c}{N_a} \quad (4.5)$$

$$Precision = \frac{N_c}{N_c + N_f} \quad (4.6)$$

where

$$N_a = \sum_n \sum_m G(n, m) \quad (4.7)$$

$$N_c = \sum_n \sum_m D_w(n, m) \times G(n, m) \quad (4.8)$$

$$N_f = -N_c + \sum_n \sum_m D_w(n, m) \quad (4.9)$$

where N_a , N_c , and N_f are the number of correspondences, number of correct, and number of false descriptor matches, respectively. In addition to these measures, following measures are also used.

4.2.2 AUC

AUC represents Area Under a Precision–Recall Curve (AUC). It is a single valued measure. It is computed with a definite integral of precision p as a function of recall r in the closed interval $r \in [0, 1]$:

$$AUC = \int_0^r p(r)dr \quad (4.10)$$

Precision–Recall curves compare the performance of the local descriptors based on the precision criterion with respect to recall, whereas AUC measure provides a single valued score for this performance comparison.

4.2.3 Number of Correct Nearest Neighbor Matches

The number of correct nearest neighbor matches are based on a nearest neighbor based descriptor matching strategy [38]. This matching strategy is different from the distance threshold based descriptor matching strategy, which uses a set of distance thresholds to compute the Precision–Recall curves and AUC scores for the performance comparison. The nearest neighbor based descriptor matching strategy computes a nearest neighbor for each description vector of the reference image in the target image, and then by making use of the G matrix (Equation 4.2), it counts the number of correct nearest neighbor matches. It is a one-to-one descriptor matching strategy in contrast to distance threshold based descriptor matching strategy, which computes several matches for each description vector of the reference image in the target image, and several of them may be correct if they fulfill the distance threshold based descriptor matching criteria (Equations 4.2–4.4).

4.2.4 Overlap Error

An overlap error threshold of 50% is used as a default setting for establishing correspondences between the HarLap regions (Equation 4.2). This threshold is borrowed from [38], where it is used for the performance evaluation of the local descriptors for image matching. In addition to 50% overlap error, 10%, 20%, 30%, 40%, and 60% overlap errors, are also used in this thesis. Note that, $\varepsilon\%$ overlap error means that the number of correspondences, the number of correct nearest neighbor matches, Precision–Recall curves, and AUC scores are computed between 0% and $\varepsilon\%$ overlap errors, where $\varepsilon = \{10, 20, 30, 40, 50, 60\}$.

Figure 4.2 illustrates six different overlap errors between HarLap regions. The reference image HarLap regions are shown in green and the projected ones in red. The projected regions are obtained by transforming the target image regions into the geometrical coordinates of the reference image by using the ground truth homography K . The overlap error comes from different size and position of the HarLap regions [39].

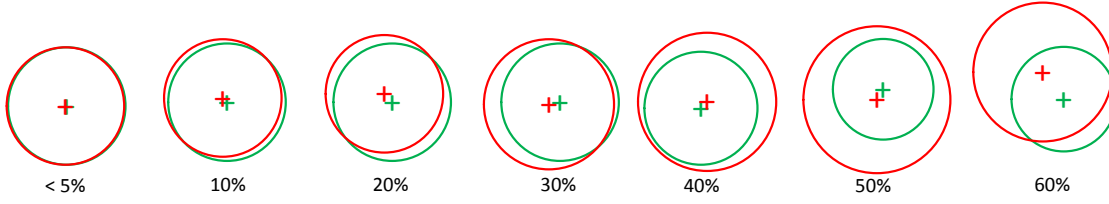


Figure 4.2: Overlap error ϵ . Examples of target regions (red) projected on the corresponding reference regions (green) with the ground truth transformation. Overlap error (%) in each case is shown below the regions. The overlap error comes from different size and position of the HarLap regions. '+' sign in each case depicts the region center (i.e., HarLap point). Images from [39].

4.2.5 Repeatability

Repeatability is a measurement for the performance evaluation of the HarLap detector. It is computed as the ratio of the number of HarLap region correspondences to the smaller of the number of HarLap regions between the pairs of images in the image matching.

$$Repeatability = \frac{N_a}{\min(N, M)} \quad (4.11)$$

where N_a is the number of correspondences, N and M are the number of HarLap regions in I_r and I_t images, respectively.

4.3 CAVE Multispectral Image Dataset

CAVE¹ image dataset consists of multispectral images of 32 different scenes [70]. Each scene is captured in 31 different discrete wavelength bands, ranging from 400nm to 700nm with a fixed step size of 10nm. To perform image matching on the images of the CAVE dataset, the 400nm band image per scene is used as a reference image and its image matching with 450nm, 510nm, 570nm, 630nm, and 700nm band images of the same scene is performed. This is to evaluate the performance of the local descriptors under increasing nonlinear intensity changes between the wavelength band images of the CAVE dataset. In the case of 400nm–450nm image matching (i.e., image matching between 400nm and 450nm band images), nonlinear intensity changes are relatively low than the ones in 400nm–720nm image matching.

CAVE Test Scenes

A subset of scenes from the CAVE dataset is shown in Figure 4.3. An image pair for each scene is shown. Each pair shows the same scene contents in 400nm (left) and 700nm (right) bands. The nonlinear intensity changes between the images of the same scene can be seen, which are due to the wavelength differences between the band images. They make the image matching based on local descriptors difficult.

¹<http://www.cs.columbia.edu/CAVE/databases/multispectral/>

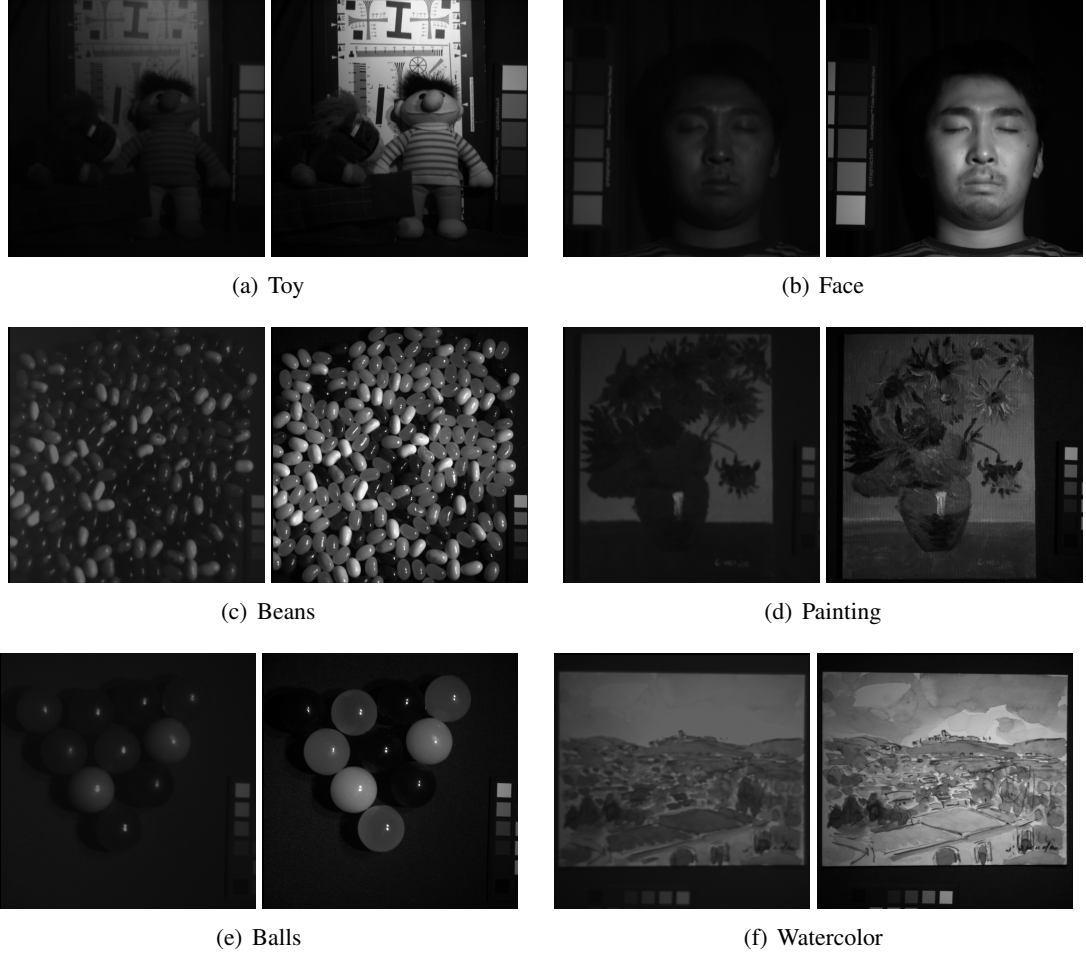


Figure 4.3: Samples from evaluated CAVE Dataset [70]. Six different scenes are shown in 400nm (left) and 700nm (right) bands.

Experiments on a CAVE Test Scene

This section presents the experimental results for the Toy scene of the CAVE dataset, which is shown in Figure 4.3(a). The experimental results described are: repeatability score, number of correct matches, Precision–Recall curves, and AUC scores. All the results are based on 50% overlap error, i.e., the number of correspondences, the number of correct matches, Precision–Recall, and AUC scores are counted between 0% and 50% overlap errors.

Repeatability: Figure 4.4(a) shows the number of HarLap regions detected and the number of HarLap region correspondences established between 400nm and 700nm band images of the Toy scene. The numbers of detected HarLap regions are 1023 and 1064, whereas the number of HarLap correspondences is 817. The repeatability score for this image matching is $100 \times 817 / \min(1023, 1064) = 79.9\%$ where $\min(1023, 1064) = 1023$ is the smaller number of

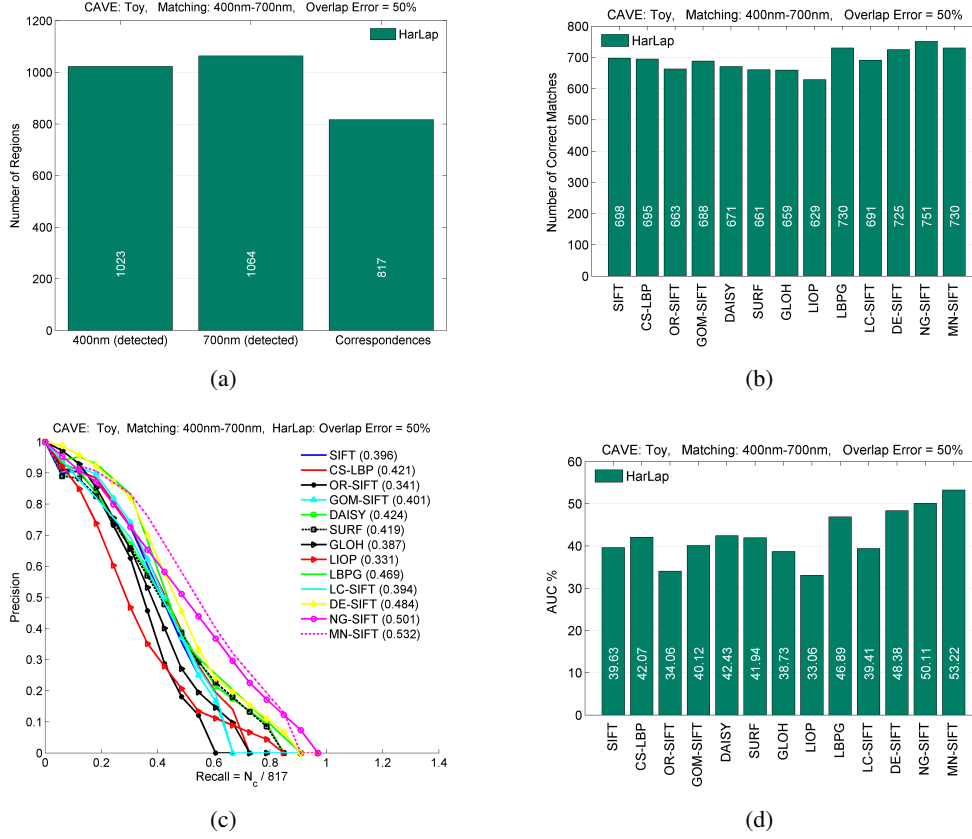


Figure 4.4: Results for image matching between 400nm and 700nm band images of the Toy scene of the CAVE dataset, which are shown in Figure 4.3(a). The last five descriptors in (b)–(d) are descriptors proposed.

HarLap regions.

Nearest neighbor matches: Figure 4.4(b) shows the number of correct nearest neighbor descriptor matches out of 817 correspondences. The highest number of correct matches is demonstrated by NG-SIFT. It obtains 751 correct matches, whereas LIOP obtains the worst number of correct matches, i.e., 629. SIFT obtains 698 correct matches and performs inferior to LBPG, DE-SIFT, and MN-SIFT, which obtain 730, 725, and 730 correct matches, respectively. LC-SIFT obtains a slightly lower number of correct matches than SIFT. Other local descriptors demonstrate inferior performance to SIFT.

Precision-Recall Curve: Figure 4.4(c) shows the Precision-Recall curves, which are based on 817 HarLap region correspondences. To understand these curves, consider a point (Recall, Precision)=(0.5,0.6). The number of correct and false matches obtained by MN-SIFT at this point are: $0.5 \times 817 = 409$ and $409/0.6 - 409 = 273$, respectively. Similarly, the number of correct and false descriptor matches for other local descriptors can also be computed in order to compare them with that of MN-SIFT. Precision-Recall curves show that MN-SIFT demonstrates

the best performance. It achieves an AUC score of 0.532 (shown inside parentheses in the figure legend) and outperforms SIFT and other local descriptors. DE-SIFT, LPBG, and NG-SIFT also obtain better AUC scores than SIFT. LC-SIFT obtains a slightly lower AUC score than SIFT. GOM-SIFT, CS-LBP, DAISY, and SURF obtain better AUC scores than SIFT, whereas OR-SIFT, GLOH, and LIOP perform inferior to SIFT. In Figure 4.4(d), the AUC% scores are plotted in order to better visualize the comparison among the local descriptors.

4.3.1 Experiments on CAVE Dataset

This section presents the experimental results for image matching when all the 32 scenes of the CAVE dataset are considered for 400nm–450nm, 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching. The experimental results shown in Figure 4.4 compare the local descriptors based on only image matching between 400nm–700nm band images of the Toy scene of the CAVE dataset.

Repeatability: Figure 4.5 shows the average number of HarLap regions detected and the average number of HarLap region correspondences established between the various band images of the CAVE dataset. The average number of detected HarLap regions in 400nm images of the CAVE dataset is 838 ± 65 where 838 is an average value and 65 represents the standard deviation. Similarly, the average number of detected HarLap regions in 450nm, 510nm, 570nm, 630nm, and 700nm images are 765, 787, 804, 834, and 837, respectively. The detection of HarLap regions is performed independently, i.e., detection in one band image does not affect the detection in other band images of the same and other scenes. Therefore, Figure 4.5 shows similar number of detected HarLap regions in various band images of the CAVE dataset. However, the effects of nonlinear intensity changes on the average number of HarLap region correspondences can be seen. For instance, the average number of correspondences in 400nm–450nm image matching is 630 ± 58 (see Figure 4.5(a)), which is higher than the number of correspondences in 400nm–700nm image matching, i.e., 602 ± 50 (see Figure 4.5(c)). This shows a decrease in the number of correspondences with increasing wavelength differences (i.e., increasing nonlinear intensity changes) between the band images. Figure 4.5(d) shows such effects by plotting the average number of correspondences between 400nm and other band images of the CAVE dataset. The error bars (vertical lines) represent the average $\pm$ standard deviation scores.

Nearest Neighbor Matches: Figure 4.6 shows the average number of correct nearest neighbor descriptor matches obtained by the local descriptors when all scenes of the CAVE dataset are considered. It can be seen that all the local descriptors demonstrate better performance in the case of 400nm–450nm image matching (see Figure 4.6(a)) than 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching due to low nonlinear intensity changes between 400nm–450nm images. In order to better visualize and compare the effects of nonlinear intensity changes on the performance measures of the local descriptors, Figure 4.6(d) combines the average number of correct matches, which are shown in Figures 4.6(a)–4.6(c). This figure shows the best results for NG-SIFT. It obtains 539 ± 56 correct matches in 400nm–450nm image matching (shown at 450nm in Figure 4.6(d)), whereas it achieves 429 ± 47 correct matches in 400nm–700nm image matching. It can be seen that NG-SIFT demonstrates the highest number of correct matches with increasing wavelength (differences between the band images) than other local descriptors.

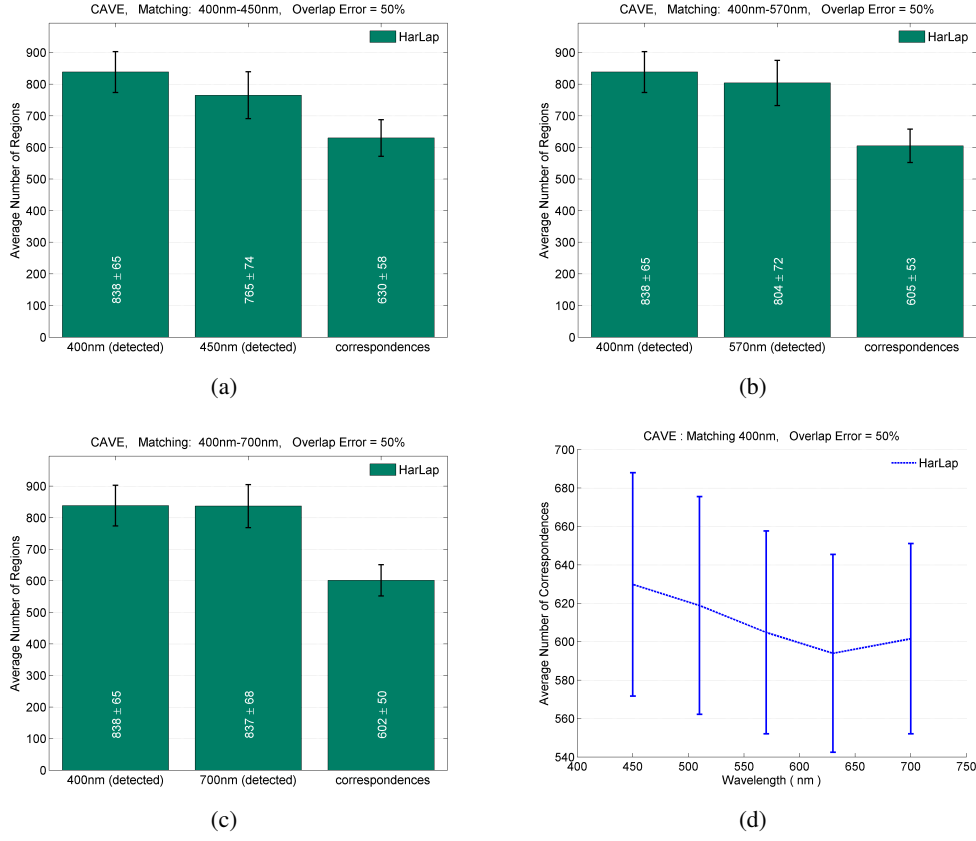


Figure 4.5: Average number of HarLap region correspondences when all scenes of the CAVE dataset are considered for 400nm–450nm, 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching. The error bars (vertical lines) represent the average $\pm$ standard deviation scores.

LBPG obtains the second best results. LBPG is an extended version of CS-LBP. It is proposed to improve the performance of CS-LBP with regard to nonlinear intensity changes. It is also an extended version of SIFT and uses CS-LBP modified image gradients in the SIFT method for the construction of local descriptors. This shows that the CS-LBP modified image gradients are more robust to nonlinear intensity changes than the image gradients and CS-LBP features, which are used in the construction of SIFT and CS-LBP descriptors, respectively.

MN-SIFT obtains the third best results. It is a modified version of NG-SIFT. It is proposed in order to improve the performance of NG-SIFT on textured scene images. The CAVE dataset consists of a fewer number of textured scene images; therefore, the comparison shows lower performance of MN-SIFT than NG-SIFT. Both NG-SIFT and MN-SIFT are based on normalized image gradients. The experimental results show that such normalized gradients are more robust to nonlinear intensity changes than image gradients.

LC-SIFT and DE-SIFT also demonstrate higher number of correct matches than SIFT. They

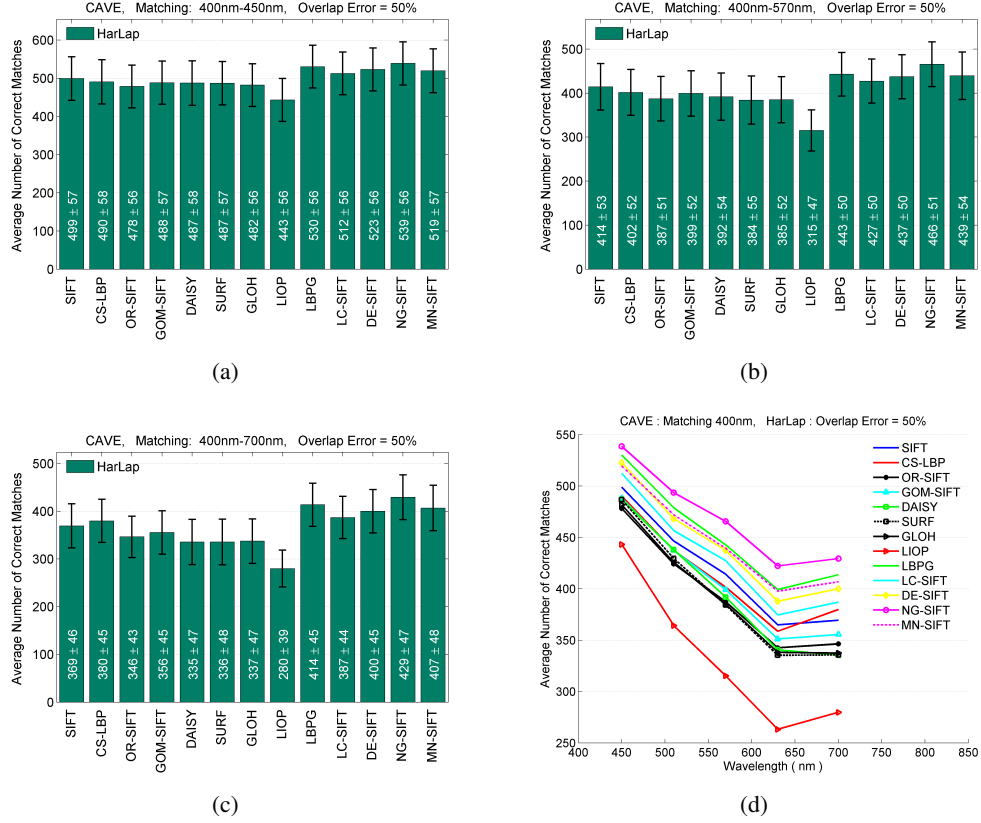


Figure 4.6: Average number of correct nearest neighbor descriptor matches when all scenes of the CAVE dataset are considered for 400nm–450nm, 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching at 50% overlap error.

are extended versions of SIFT. They use edge boosting in order to increase the contribution of local edges in the SIFT method. Their performances show that such an edge boosting gives higher number of correct matches than SIFT.

GLOH, CS-LBP, SURF, OR-SIFT, GOM-SIFT, and DAISY obtain lower results than SIFT. They are extended versions of SIFT, however, demonstrate inferior performance to SIFT with regard to nonlinear intensity changes. LIOP obtains the worst results, which suggests that the intensity based descriptors that LIOP computes, are less suitable for image matching under nonlinear intensity changes in contrast to gradient based descriptors. Among the gradient based descriptors, the descriptors proposed show the best performance. It shows that the modified gradients (i.e., LBPG, LC, DE, NG, and MN), which are proposed for the construction of descriptors, are more robust to nonlinear intensity changes than image gradients, CS-LBP features, and Haar wavelet responses, which are used in the construction of SIFT, CS-LBP, and SURF descriptors, respectively.

AUC: Figure 4.7 compares the local descriptors based on average AUC (%) scores. The average AUC scores are based on average Precision-Recall curves, which are shown in Fig-

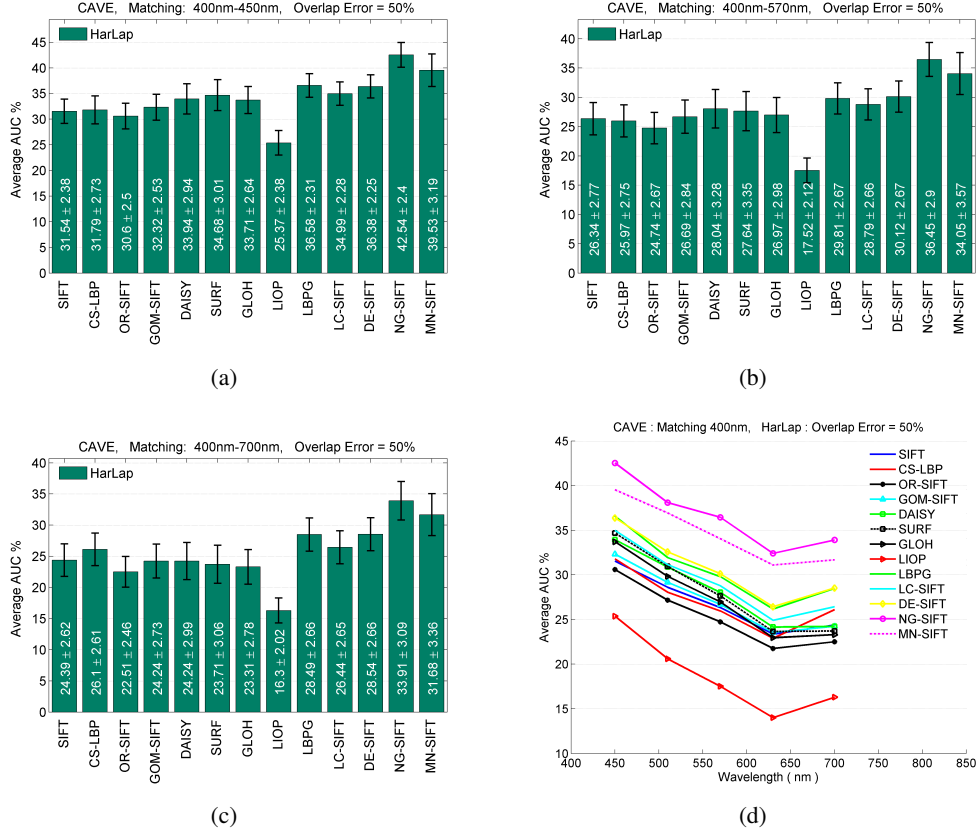


Figure 4.7: Average AUC(%) scores when all scenes of the CAVE dataset are considered for image matching at 50% overlap error.

ure 4.8. The Precision-Recall curves evaluate the performance of local descriptors based on the Precision criterion with respect to Recall. In the case of 400nm–450nm image matching (see Figure 4.7(a)), all the local descriptors demonstrate better performance than 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching.

The best AUC scores can be seen for NG-SIFT, which outperforms other local descriptors with respect to increasing nonlinear intensity changes (wavelength) between the band images, as shown in Figure 4.7(d). NG-SIFT obtains an average AUC (%) score of 42.54 ± 2.4 in 400nm–450nm image matching (shown at 450nm in Figure 4.7(d)), whereas it achieves an average AUC score of 33.91 ± 3.09 in 400nm–700nm image matching. This is due to high nonlinear intensity changes between 400nm–700nm images, but still it outperforms other local descriptors, even SIFT, which obtains an AUC score of 31.54 ± 2.38 in 400nm–450nm image matching. MN-SIFT achieves the second best average AUC scores. NG-SIFT and MN-SIFT are based on normalized image gradients, which are proposed as new image features for the descriptor construction to overcome the effects of nonlinear intensity changes. DE-SIFT obtains the third best AUC scores. LBP performs comparable to DE-SIFT. LC-SIFT also outperforms SIFT. SURF,

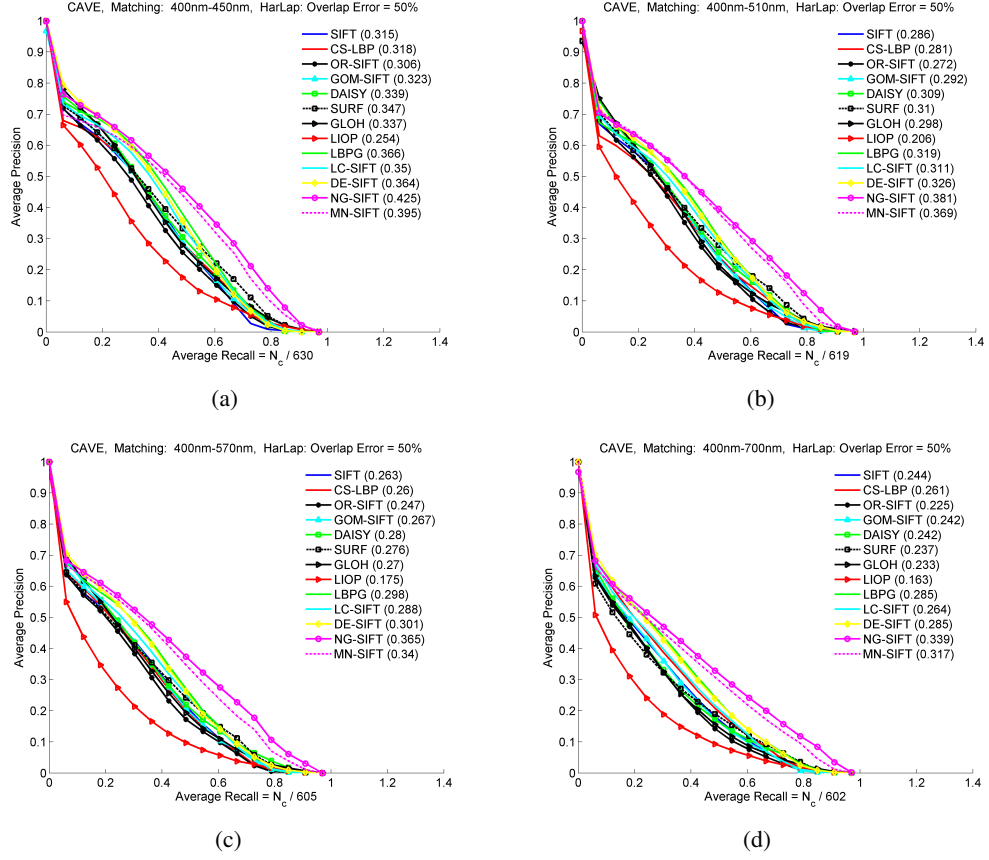


Figure 4.8: Average Precision-Recall curves when all scenes of the CAVE dataset are considered for image matching at 50% overlap error.

DAISY, GLOH, and GOM-SIFT achieve better AUC scores than SIFT. CS-LBP, OR-SIFT, and LIOP obtain lower AUC scores than SIFT. LIOP obtains the worst results.

4.3.2 Effect of Overlap Error

This section presents the effects of overlap error threshold on the performance measures of the local descriptors. The results described in this section are similar to the ones in Section 4.3.1, which are based on 50% overlap error, whereas in this section additional results are described when overlap error thresholds of 10%, 20%, 30%, 40%, and 60%, are used.

Overlap error is used to establish the HarLap region correspondences between the pairs of images in image matching. It creates the ground truth G matrix (Equation 4.2), which is used to identify and count the number correct descriptor matches (Equation 4.8). In Figure 4.2 six different overlap errors between the HarLap regions are illustrated. Mikolajczyk and Schmid [38] use an overlap error threshold of 50% for the performance evaluation of the local descriptors. To this end, Section 4.3.1 compares the local descriptors based on 50% overlap error, whereas

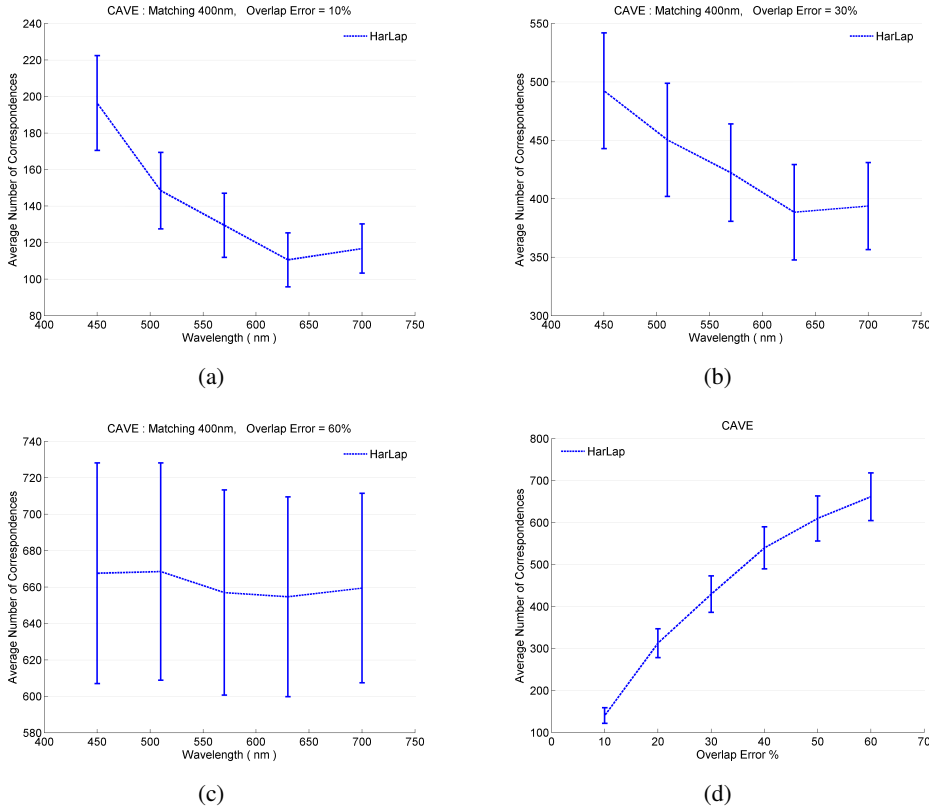


Figure 4.9: Average number of HarLap region correspondences when all scenes of the CAVE dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.5(d).

this section presents additional results under various overlap errors.

Repeatability: Figure 4.9 shows the effects of overlap error thresholds on the average number of HarLap region correspondences. The number of correspondences in 400nm–450nm image matching at 10% overlap error is 192 ± 33 , as shown in Figure 4.9(a) at 450nm. The same overlap error gives 117 ± 20 correspondences in 400nm–700nm image matching, which is far less than the number of correspondences in 400nm–450nm image matching due to high nonlinear intensity changes between 400nm–700nm images. The overall average number of correspondences in 400nm–450nm, 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching is 142 ± 23 , which is also shown in Figure 4.9(d) at 10% overlap error. This figure shows the average number of correspondences with increasing overlap errors and combines the results, which are shown in Figures 4.9(a)–4.9(c) and Figure 4.5(d) in order to better visualize the effects of increasing overlap error on the average number of HarLap region correspondences. The figure shows an increase in the number of correspondences with increasing overlap error. A large overlap error results in a large number of correspondences and also increases the number of correct nearest neighbor descriptor matches, as shown in Figure 4.10 (for the CAVE dataset).

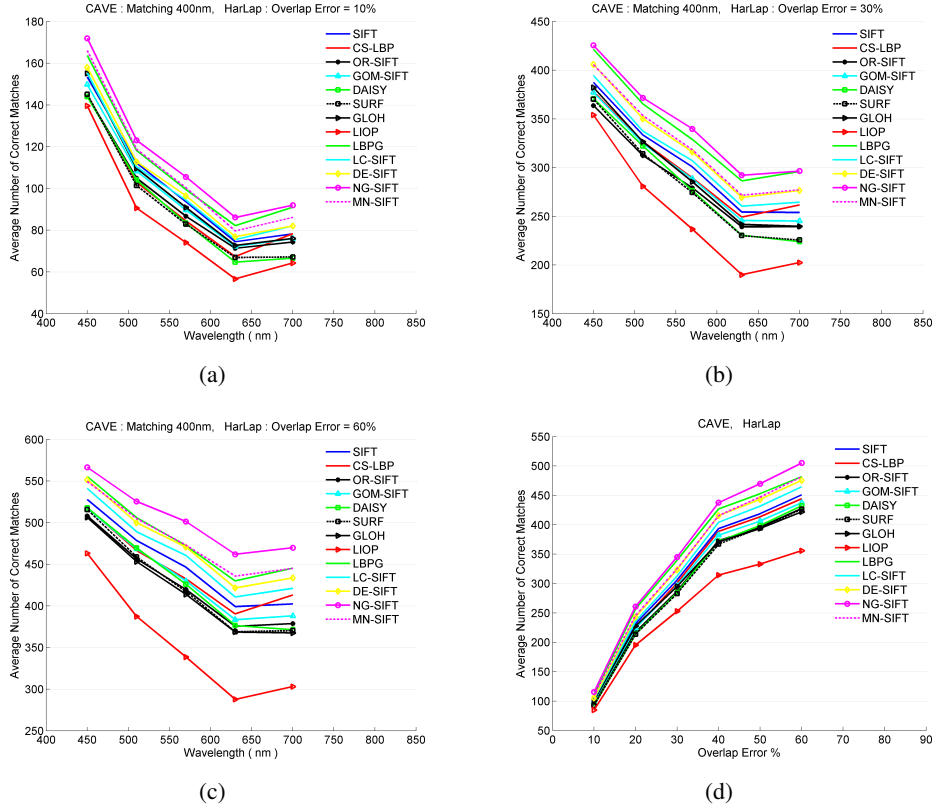


Figure 4.10: Average number of correct nearest neighbor descriptor matches when all scenes of the CAVE dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.6(d).

Nearest neighbor matches: Figure 4.10 compares the local descriptors based on the average number of correct nearest neighbor descriptor matches under various overlap error thresholds. All the local descriptors demonstrate a low number of correct matches at 10% overlap error (see Figure 4.10(a)). MN-SIFT in 400nm–450nm image matching achieves 172 correct matches out of 192 correspondences (Figure 4.9(a)), whereas in the case of 400nm–700nm image matching, it obtains 90 correct matches due to high nonlinear intensity changes. The overall average number of correct matches obtained by MN-SIFT in 400nm–450nm, 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching is 118, which is shown in Figure 4.10(d) at 10% overlap error. This figure compares the local descriptors based on the average number of correct nearest neighbor matches with respect to increasing overlap error thresholds and combines the results, which are shown in Figures 4.10(a)–4.10(c) and Figure 4.6(d). The figure shows an increase in the number of correct matches with increasing overlap error thresholds. At 10% overlap error, all the local descriptors demonstrate similar performances. But at 60% overlap error, MN-SIFT demonstrates 502 correct matches and outperforms all the other local descriptors. LBPG and NG-SIFT achieves the second best results. DE-SIFT and

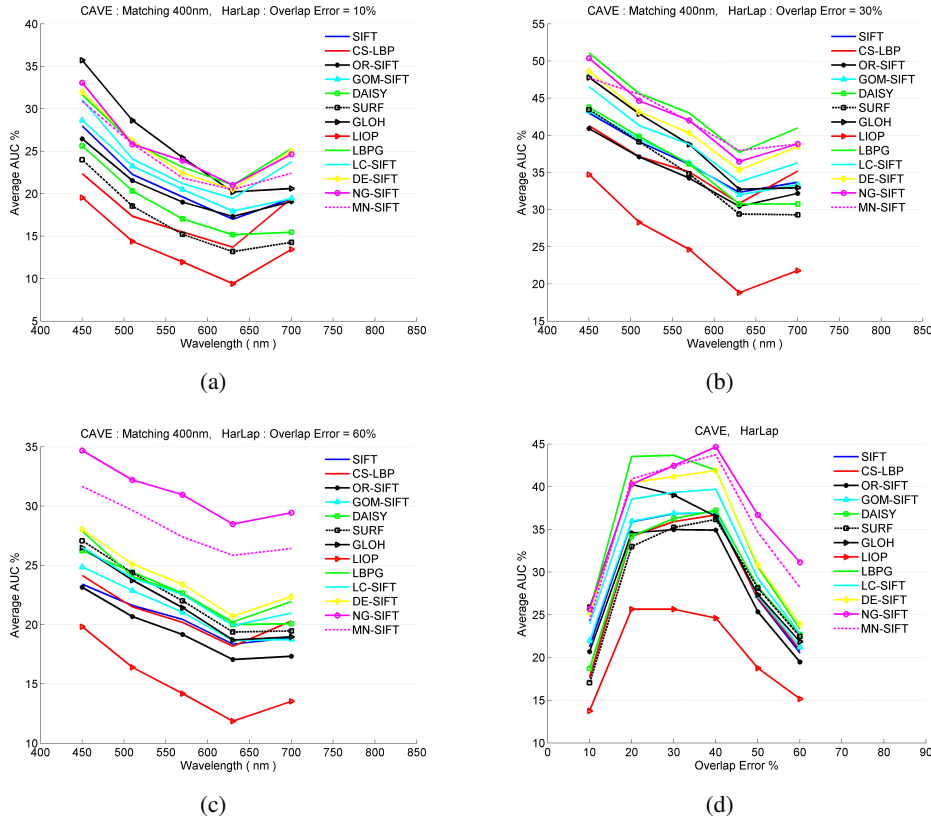


Figure 4.11: Average AUC (%) scores when all scenes of the CAVE dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.7(d).

LC-SIFT demonstrate better performance than SIFT, whereas SIFT obtains better performance than GLOH, CS-LBP, SURF, OR-SIFT, GOM-SIFT, and DAISY. LIOP demonstrates the lowest number of correct matches with increasing overlap error threshold.

AUC: Figure 4.11 compares the local descriptors based on average AUC (%) scores under various overlap error thresholds. The best AUC score is demonstrated by GLOH in 400nm–450nm image matching, at 10% overlap error. GLOH obtains an AUC score of 36%, but achieves AUC score of 20.6% in 400nm–700nm image matching, at 10% overlap error and performs 4.4% inferior to LBPG. The overall average AUC score achieved by LBPG in 400nm–450nm, 400nm–510nm, 400nm–570nm, 400nm–630nm, and 400nm–700nm image matching is 25.8%, which is shown in Figure 4.11(d) at 10% overlap error. This figure combines the AUC scores, which are shown in Figures 4.11(a)–4.11(c) and Figure 4.7(d) to compare the performance of the local descriptors based on average AUC scores under increasing overlap error thresholds. The figure shows low AUC scores for all the local descriptors at 10% and 60% overlap errors, whereas the local descriptors demonstrate the best AUC scores between 20% and 40% overlap errors. NG-SIFT obtains the best AUC score at 40% overlap error. MN-SIFT obtains the second best AUC

score. LBPG performs comparable to DE-SIFT and both outperform SIFT. LC-SIFT and GLOH also outperform SIFT. The comparison shows similar AUC scores for SIFT and GOM-SIFT. Both outperform CS-LBP, OR-SIFT, and DAISY. LIOP obtains the worst results.

The experimental results achieved on CAVE dataset show that all the descriptors proposed perform better than the state-of-the-art local descriptors for image matching. SIFT demonstrates better performance than other state-of-the-art local descriptors, i.e., OR-SIFT, GOM-SIFT, CS-LBP, GLOH, DAISY, and LIOP when average number of correct nearest neighbor descriptor matches are used for the performance evaluation, whereas GLOH demonstrates better performance when AUC score is used for the performance evaluation. Compared to SIFT and GLOH, the descriptors proposed demonstrate better performance in both average number of correct nearest neighbor descriptor matches and AUC scores based performance evaluation.

4.4 RWHI Image Dataset

Real-World Hyperspectral Image (RWHI)² dataset consists of multispectral images of 50 different indoor and outdoor scenes [7]. Each scene is acquired in 31 different discrete wavelength bands, ranging from 420nm to 720nm with a fixed step size of 10nm [7]. A subset of scenes from RWHI dataset is shown in Figure 4.12. An image pair for each scene is shown in 420nm (left) and 720nm (right) bands. Nonlinear intensity changes between the images of the same scene can be seen, which are due to the wavelength differences between the band images.

To perform image matching on the images of RWHI dataset, the 420nm image per scene is used as a reference image, and its image matching with 470nm, 530nm, 590nm, 650nm and 720nm images of the same scene are performed in order to evaluate the performance of local descriptors under increasing nonlinear intensity changes between the images of the RWHI dataset. In contrast to CAVE dataset, the images of the RWHI dataset possess low nonlinear intensity changes.

Experiments on a RWHI Test Scene

This section presents the image matching results for the imga6 scene of the RWHI dataset, which is shown in Figure 4.12(a). The experimental results described are: repeatability score, number of correct nearest neighbor descriptor matches, Precision–Recall curves, and AUC scores.

Repeatability: Figure 4.13(a) shows the number of HarLap region correspondences between 420nm–720nm images of the imga6 scene at 50% overlap error. The number of correspondences is 547 out of 1026 and 1003 detected HarLap regions in 420nm and 720nm images, respectively. The repeatability score for this image matching is $100 \times 547 / \min(1026, 1003) = 54.5\%$, where $\min(1026, 1003) = 1003$ is the smaller number of HarLap regions.

Nearest neighbor correct matches: Figure 4.13(b) compares the local descriptors based on the number of correct nearest neighbor descriptor matches in 420nm–720nm image matching of the imga6 scene. SIFT obtains 428 correct matches and shows lower performance than LC-SIFT, NG-SIFT, and MN-SIFT, which obtain 433, 437, and 446 correct matches, respectively. LBPG and DE-SIFT obtain 418 and 422 correct matches, respectively, and demonstrate lower

²<http://vision.seas.harvard.edu/hyperspec/>



Figure 4.12: Samples from the evaluated RWHI Dataset [7]. Six different scenes are shown in 420nm (left) and 720nm (right) bands.

performance than SIFT. MN-SIFT obtains the highest number of nearest neighbor matches, i.e., 446, whereas LIOP demonstrates the worst number of correct matches, i.e., 323.

Precision-Recall Curve: Figure 4.13(c) compares the local descriptors based on Precision-Recall curves. To understand these curves, consider a point (Recall, Precision)=(0.4,0.8). The number of correct and false matches obtained by MN-SIFT at this point are $0.4 \times 547 = 219$ and $219/0.8 - 219 = 55$, respectively. Similarly, the number of correct and false matches can also be computed for other local descriptors to compare them with that of MN-SIFT. The Precision-Recall curves show the best results for MN-SIFT. It obtains the best AUC score of 0.51 (shown inside parentheses in the figure legend) and outperforms SIFT and all the other local descriptors. NG-SIFT obtains the second best results. LC-SIFT, DE-SIFT, and LBPg perform inferior to SIFT. Except SURF, all the other descriptors show lower AUC scores than SIFT. LIOP obtains the worst results. In Figure 4.13(d) AUC% scores are plotted to better visualize the comparison among the local descriptors.

4.4.1 Experiments on RWHI Dataset

This section presents the experimental results when all scenes of the RWHI dataset are considered for 420nm–470nm, 420nm–530nm, 420nm–590nm, 420nm–650nm, and 420nm–720nm

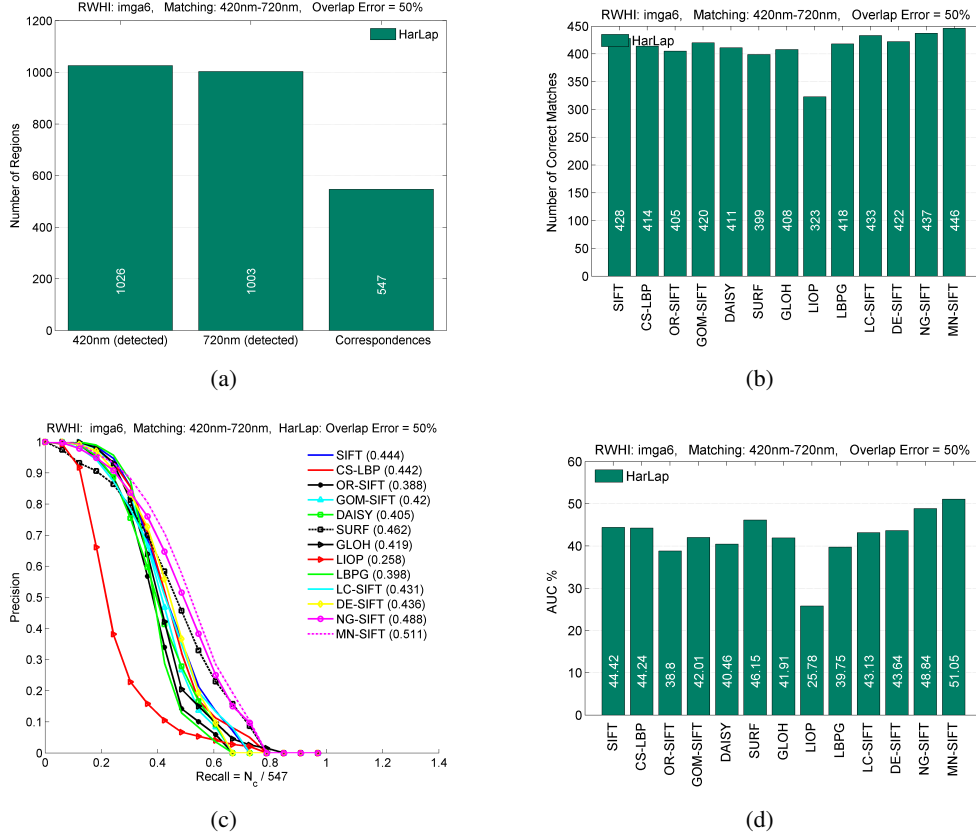


Figure 4.13: Results for image matching between 420nm and 720nm band images of the imga6 scene of the RWHI dataset, which are shown in Figure 4.12(a).

image matching at 50% overlap error. The RWHI dataset consists of 50 different scenes. Therefore, the results presented in this section are average values $\pm$ standard deviations.

Repeatability: Figure 4.14 shows the average number of HarLap regions detected and the average number of HarLap region correspondences established between the images of the RWHI dataset. The average number of detected HarLap regions in 420nm images is 1009, whereas in 470nm, 530nm, 590nm, 650nm, and 720nm images, they are 1031, 1024, 1028, 1029, and 1033, respectively. The detection of HarLap regions is independently performed, i.e., detection in a band image does not affect the detection in other band images of the same and other scenes. Therefore, Figure 4.14 shows similar average number of detected HarLap regions in various band images of the RWHI dataset. However, the effects of nonlinear intensity changes on the number of HarLap region correspondences can be seen. For instance, the average number of correspondences in 420nm–470nm image matching is 873 (see Figure 4.14(a)), whereas in 420nm–720nm image matching, it is 563 (see Figure 4.14(c)). This shows a decrease in the number of correspondences with increasing wavelength differences (increasing nonlinear intensity changes) between the band images. Figure 4.14(d) shows such effects by plotting the average

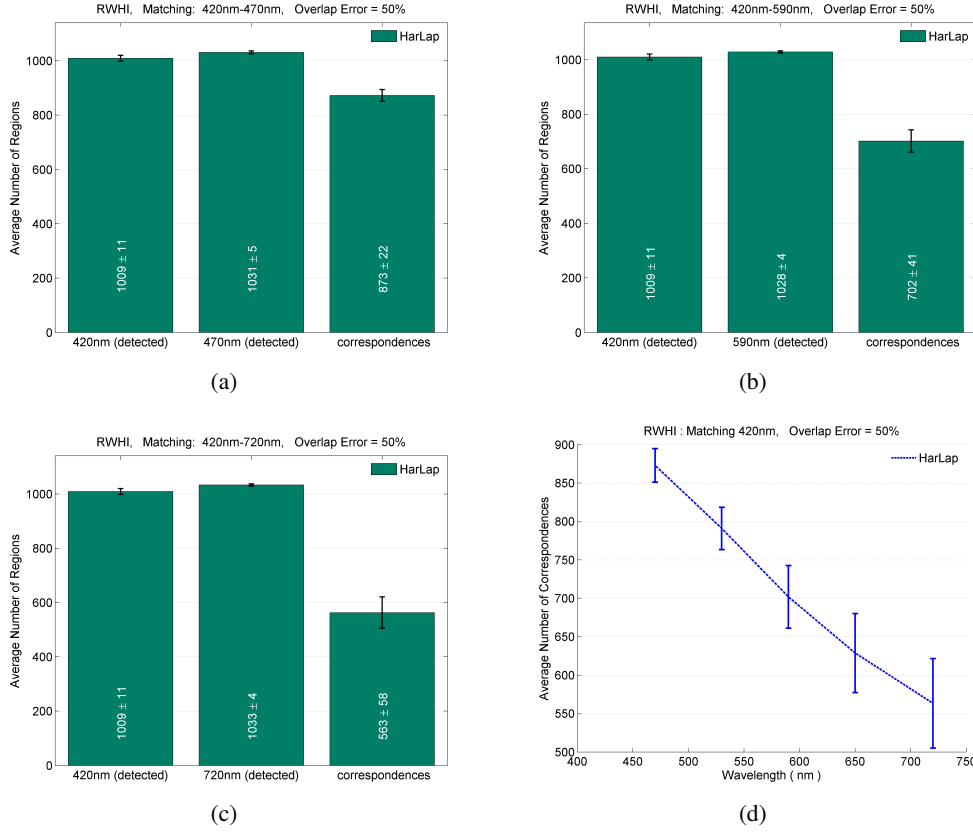


Figure 4.14: Average number of HarLap region correspondences when all scenes of the RWHI dataset are considered for 420nm–470nm, 420nm–530nm, 420nm–590nm, 420nm–650nm, and 420nm–720nm image matching at 50% overlap error.

number of correspondences with respect to increasing wavelength.

Nearest Neighbor Matches: Figure 4.15 compares the local descriptors based on average number of correct nearest neighbor descriptor matches. All the local descriptors demonstrate better performance in the case of 420nm–470nm image matching than others, due to low non-linear intensity changes. To better visualize the effects of nonlinear intensity changes on the performance of the local descriptors, Figure 4.15(d) combines the results, which are shown in Figures 4.15(a)–4.15(c). This figure shows that NG-SIFT demonstrates the best performance. It obtains 825 correct matches in 420nm–470nm image matching (shown at 470nm), whereas it obtains 454 correct matches in 420nm–720nm image matching.

MN-SIFT demonstrates the second best performance. LBP-G obtains the third best results. LC-SIFT and DE-SIFT obtain higher number of correct matches than SIFT. GLOH, SURF, OR-SIFT, GOM-SIFT, and DAISY obtain lower numbers of correct matches than SIFT, whereas CS-LBP performs comparable to SIFT. LIOP obtains the worst results.

AUC: Figure 4.16 compares the local descriptors based on average AUC (%) scores. These

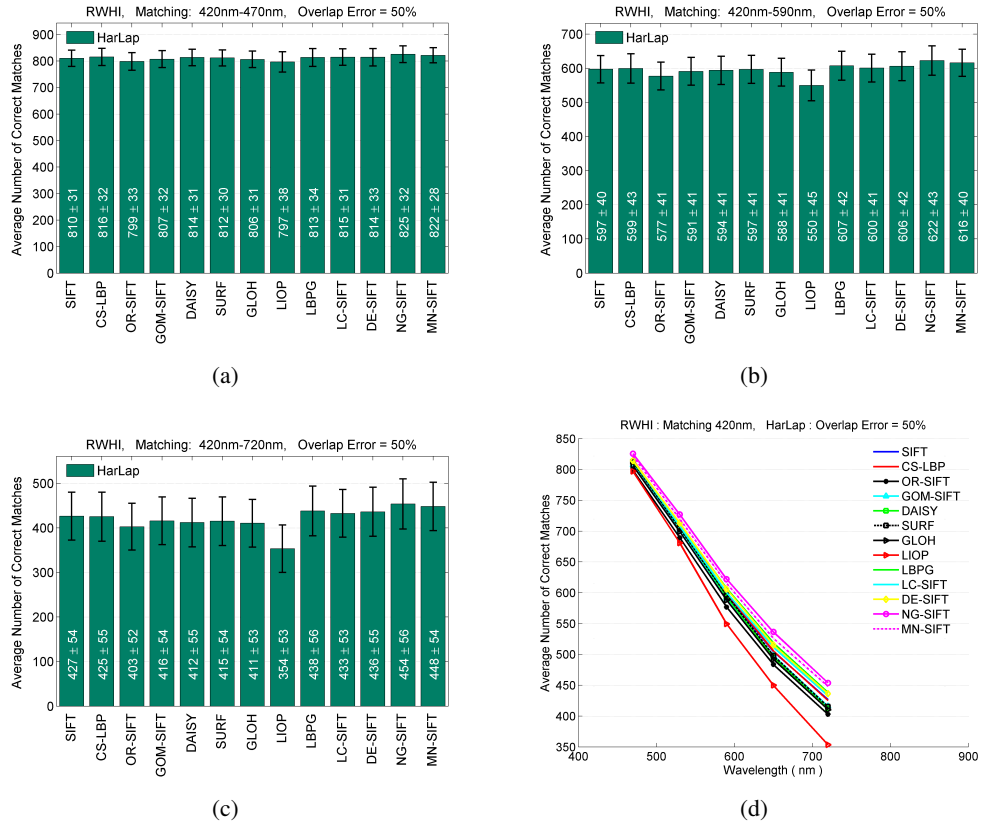


Figure 4.15: Average number of correct nearest neighbor descriptors matches when all scenes of the RWHI dataset are considered for image matching at 50% overlap error.

scores are based on the average Precision–Recall curves, which are shown in Figure 4.17. The curves compare the performance of the local descriptors based on average Precision scores with respect to Recall, when all scenes of the RWHI dataset are considered for image matching. The results show the best AUC scores for all the local descriptors in 420nm–470nm image matching compared to others.

NG-SIFT obtains the best AUC scores in 420nm–470nm image matching. It achieves an AUC (%) score of 53.93, whereas it gives an AUC score of 36.86 in 420nm–720nm image matching and performs inferior to its extended version, i.e, MN-SIFT, which achieves an AUC score of 37.40. This shows the effects of nonlinear intensity changes (i.e., wavelength differences between the band images) on the AUC scores of the local descriptors. To better visualize such effects, Figure 4.16(d) combines the average AUC scores, which are shown in Figure 4.16(a)–4.16(c). This figure shows the best performance of NG-SIFT and MN-SIFT with increasing wavelength. DE-SIFT and LBPG also demonstrate better performance than SIFT. But they perform comparable to SURF. LC-SIFT outperforms SIFT, but it obtains lower results than DAISY, CS-LBP, and SURF. LIOP obtains the worst results.

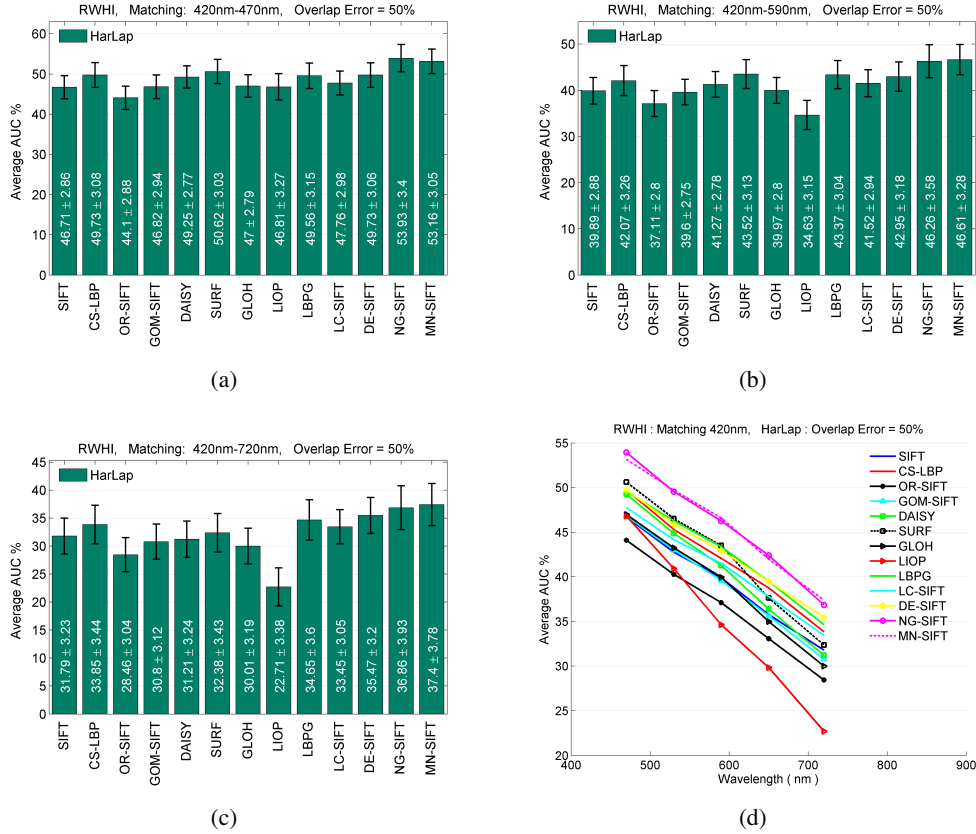


Figure 4.16: Average AUC (%) scores when all scenes of the RWHI dataset are considered for image matching at 50% overlap error.

4.4.2 Effect of Overlap Error

This section evaluates the performance of the local descriptors under various overlap error thresholds, i.e., 10%, 20%, 30%, 40%, 50% and 60%, on RWHI dataset.

Repeatability: Figure 4.18 shows the effects of overlap error on the average number of HarLap regions correspondences. For instance, in 420nm–470nm image matching the average number of correspondences is 540 at 10% overlap error, whereas in 420nm–720nm image matching, the number of correspondences is 190 (see Figure 4.18(a)). The overall average number of correspondences in 420nm–470nm, 420nm–530nm, 420nm–590nm, 420nm–650nm, and 420nm–720nm image matching is 342, which is also shown in Figure 4.18(b) at 10% overlap error. This figure shows the average number of correspondences under various overlap error thresholds. The figure shows an increase in the number of correspondences with increasing overlap error thresholds.

Nearest neighbor matches: Figure 4.19 shows the effect of overlap error thresholds on the number of correct nearest neighbor descriptor matches achieved by the local descriptors. At 10% overlap error, all the local descriptors obtains a low number of correct matches (see

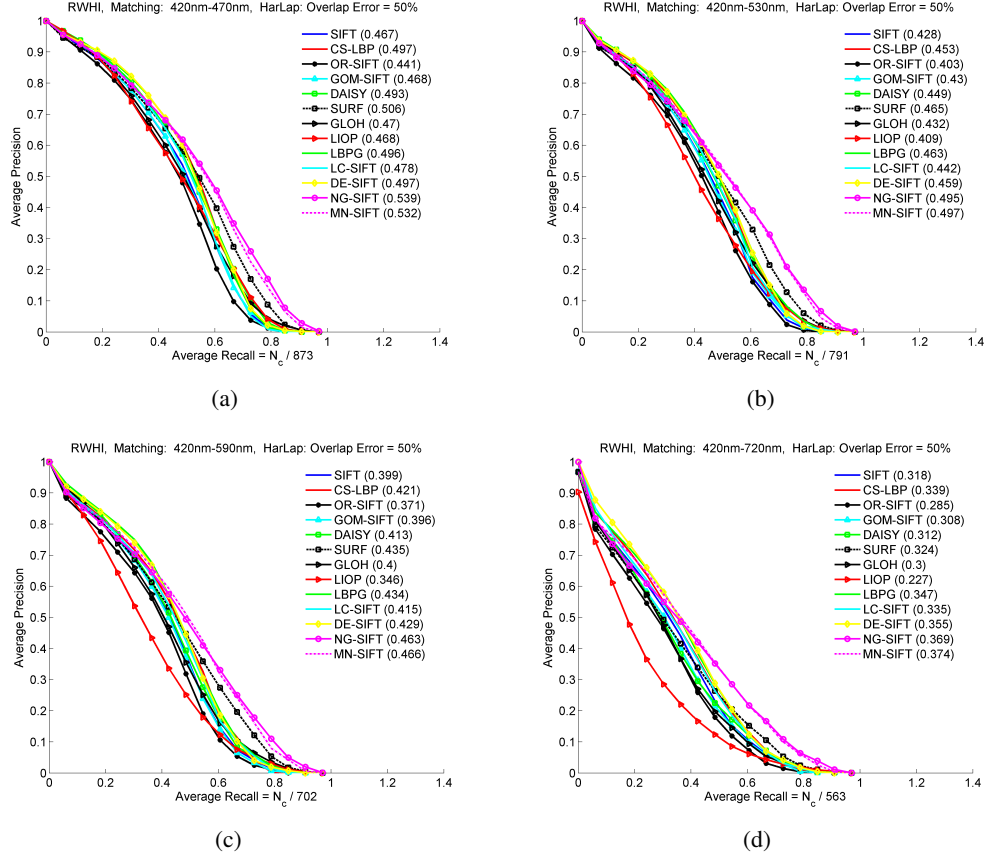


Figure 4.17: Average Precision-Recall curves when all scenes of the RWHI dataset are considered for image matching at 50% overlap error.

Figure 4.19(a)), for instance, NG-SIFT in 420nm–470nm image matching obtains 515 correct matches out of 540 correspondences (Figure 4.18(a)), whereas in 420nm–720nm image matching NG-SIFT gives 160 correct matches at the same overlap error, which are much less than 515. This performance decrease is due to high nonlinear intensity changes between 420nm–720nm images.

The overall average number of correct matches obtained by NG-SIFT in 420nm–470nm, 420nm–530nm, 420nm–590nm, 420nm–650nm, and 420nm–720nm image matching is 318 as shown in Figure 4.19(b) at 10% overlap error. This figure compares the local descriptors under various overlap error thresholds. The figure shows an increase in the average number of correct matches with increasing overlap error thresholds. A slight performance difference among the local descriptors can be seen at 10% overlap error. However, the performance difference becomes more clearer at 60% overlap error, where NG-SIFT and MN-SIFT obtain 653 correct matches and outperform others. LBP, LC-SIFT and DE-SIFT obtains higher number of correct matches than SIFT. CS-LBP, DAISY, GOM-SIFT, SIFT, and SURF demonstrate similar performances and outperform GLOH, LIOP, and OR-SIFT.

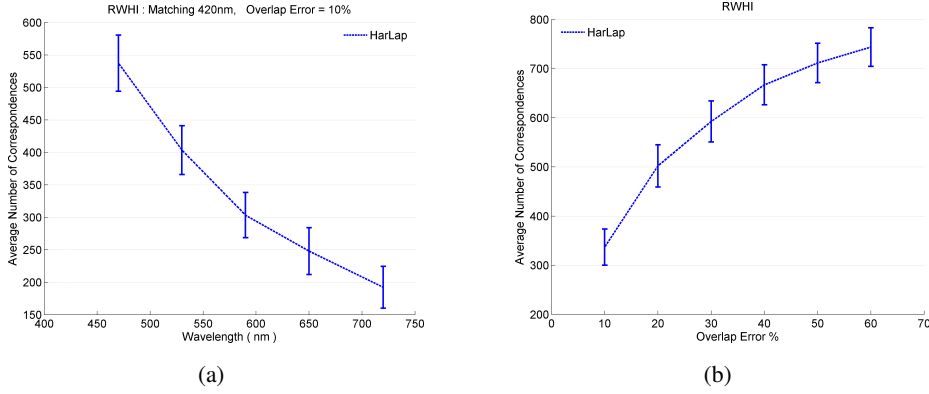


Figure 4.18: Average number of HarLap region correspondences when all scenes of the RWHI dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.14(d).

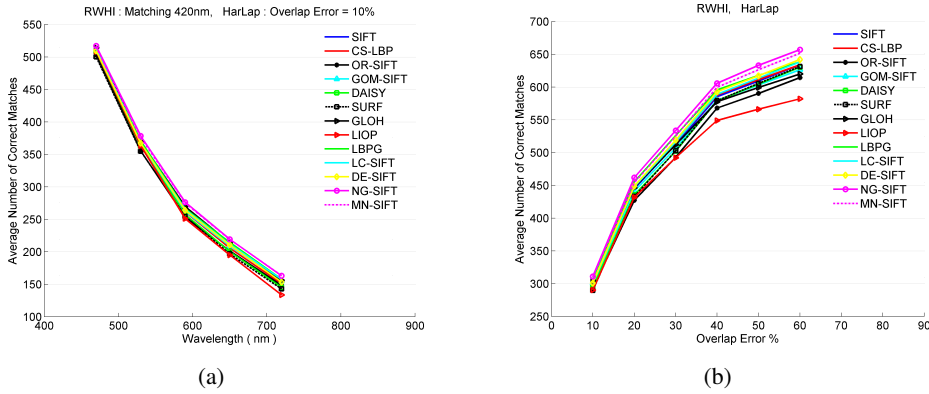


Figure 4.19: Average number of correct nearest neighbor descriptor matches when all scenes of the RWHI dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.15(d).

AUC: Figure 4.20 shows the effects of overlap error threshold on the AUC (%) scores of the local descriptors. In 420nm–470nm image matching, GLOH demonstrates the best AUC score of 80%, at 10% overlap error (see Figure 4.20(a)). In case of 420nm–720nm image matching, it obtains an AUC score of 45%. The overall average AUC score obtained by GLOH in 420nm–470nm, 420nm–530nm, 420nm–590nm, 420nm–650nm, and 420nm–720nm image matching is 62.8%, which is shown in Figure 4.20(d) at 10% overlap error. The figure compares the local descriptors based on average AUC scores under various overlap error thresholds. LBPG demonstrates the best AUC scores at 20% overlap error. NG-SIFT and MN-SIFT demonstrate the best AUC scores at 60% overlap error. They give an improvement of 7% over SIFT at 60% overlap error. LBPG, LC-SIFT, and DE-SIFT also outperform SIFT. LIOP obtains the worst

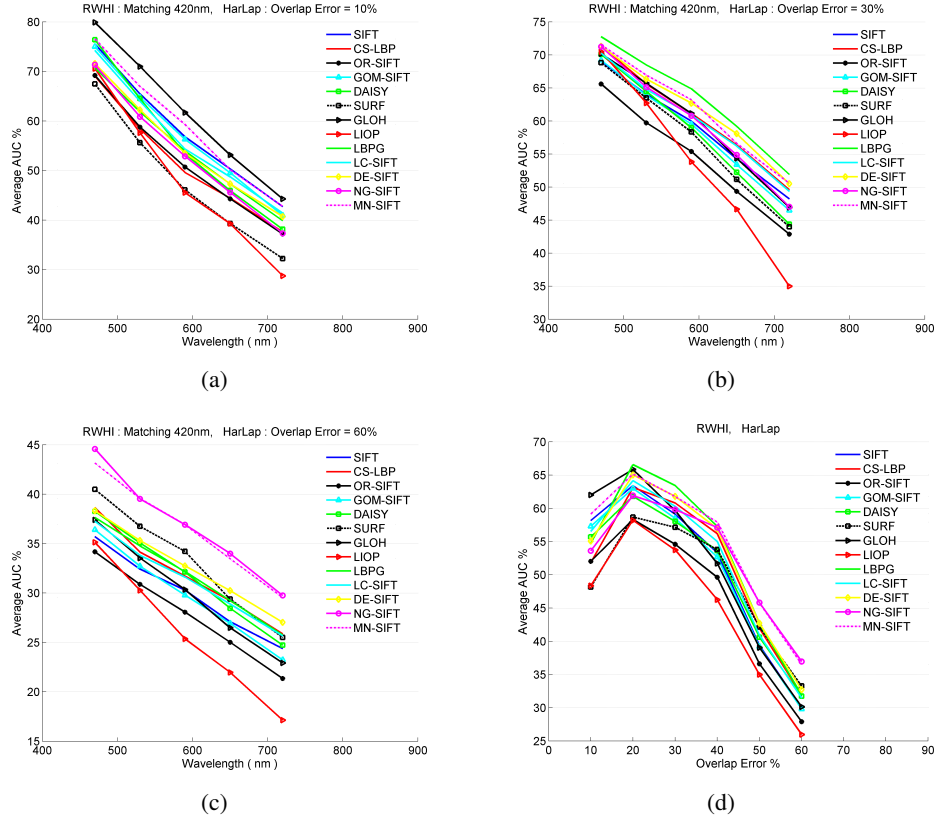


Figure 4.20: Average AUC scores when all scenes of the RWHI dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.16(d).

results.

4.5 MSD Dataset

Multimodal Stereo Dataset (MSD)³ consists of 100 VS–LWIR image pairs [1]. The pairs depict various urban scenario in Visible (VS: 400–700nm) and Long-Wave Infrared (LWIR: 800–1500 nm) bands. In contrast to CAVE [70] and RWHI [7] datasets, the images of MSD posses high nonlinear intensity changes, which can be seen from Figure 4.21, where VS–LWIR image pairs of three different scenes in VS (left) and LWIR (right) bands, are shown. Such nonlinear intensity changes also induce textural differences between the images and make the image matching based on local descriptors difficult.

³<http://www.cvc.uab.es/adas/projects/simeve/?q=node/2>

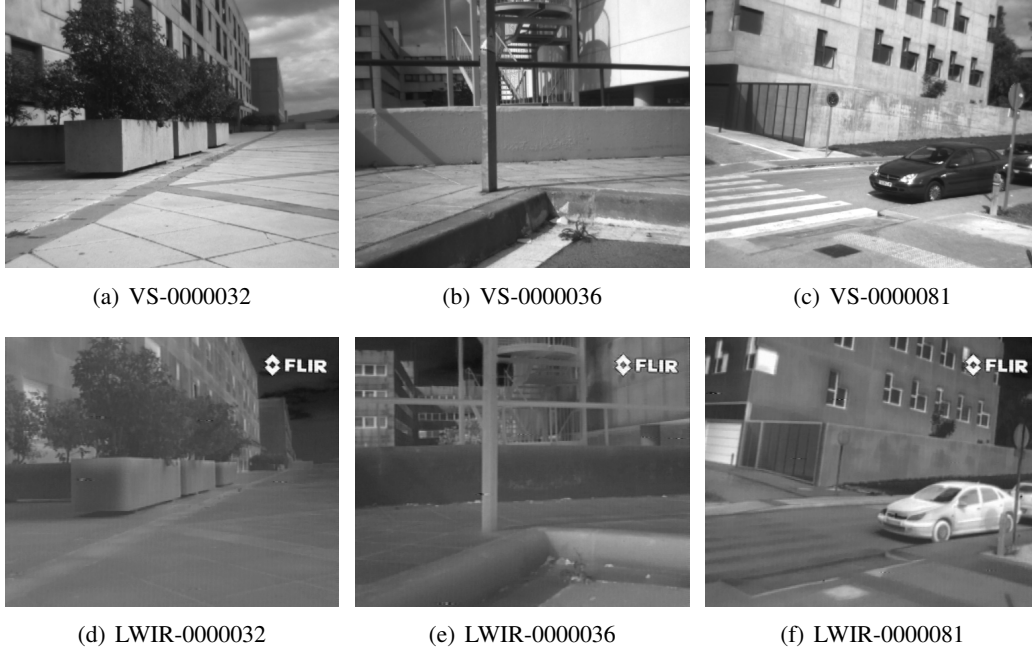


Figure 4.21: Samples from the evaluated MSD Dataset [1]. Three different scenes are shown. Each scene shows the same image contents in the Visible Spectrum (left) and Long Wave Infra Red (right) bands.

4.5.1 Experiments on MSD Dataset

This section presents the experimental results when all scenes of the MSD dataset are considered for VS–LWIR image matching at 50% overlap error. The dataset consists of 100 VS–LWIR image pairs. Therefore, the results presented in this section are average values $\pm$ standard deviations.

Repeatability: Figure 4.22(a) shows the average number of HarLap region correspondences in VS–LWIR image matching. The average number of detected HarLap regions in VS and LWIR images are 1002 and 1008, respectively. The detection of HarLap regions is performed independently, i.e., detection in a VS image does not affect the detection in the LWIR images of the same and other scenes. Therefore, Figure 4.22(a) shows similar average number of detected HarLap regions in VS and LWIR images. However, the effects of nonlinear intensity changes on the average number of HarLap region correspondences can be seen, i.e., the average number of correspondences in the case of VS–LWIR image matching is 581 and the average repeatability score is $100 \times 581 / \min(1008, 1002) = 58\%$ where $\min(1008, 1002) = 1002$ is the smaller number of detected HarLap regions.

Nearest Neighbor Matches: Figure 4.22(b) compares the local descriptors based on average number of correct nearest neighbor descriptor matches in VS–LWIR image matching. This figure shows the best performance of NG-SIFT, which obtains 120 correct matches and demonstrates 48-match improvement over SIFT. LBPG obtains the second best results. OR-

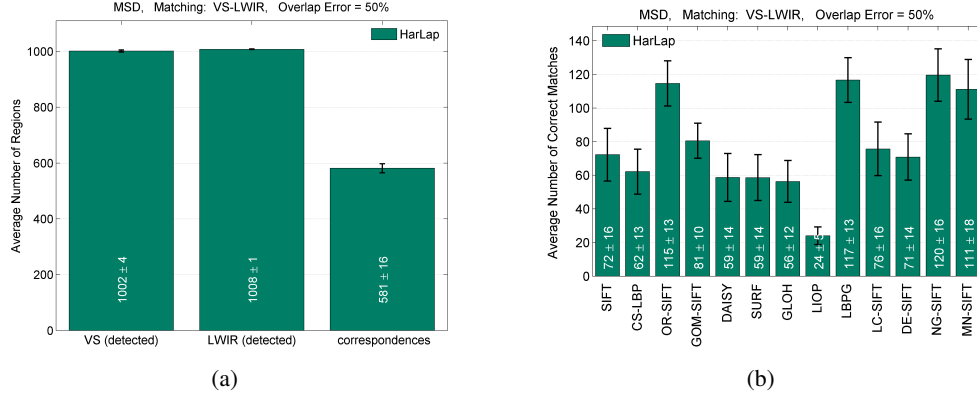


Figure 4.22: Average number of HarLap regions correspondences and average number of correct matches when all scenes of the MSD dataset are considered for VS–LWIR image matching at 50% overlap error.

SIFT achieves the third best results. OR-SIFT is a modified version of SIFT, which has been used to overcome the effects of intensity reversal and nonlinear intensity changes between the images [68]. The MSD dataset consists of a large number of VS–LWIR pairs with intensity reversal problem, therefore, OR-SIFT demonstrates better results than SIFT, MN-SIFT, LC-SIFT, and DE-SIFT. However, on CAVE and RWHI datasets, OR-SIFT performs inferior to these local descriptors. GOM-SIFT also outperform SIFT, but it performs inferior to OR-SIFT. GOM-SIFT has also been used to overcome the effects of nonlinear intensity changes and intensity reversal between the images.

MN-SIFT obtains the fourth best results. LC-SIFT and DE-SIFT achieves similar results and perform comparable to SIFT. GLOH, CS-LBP, SURF, and DAISY show lower results than SIFT, which show that they are less robust to nonlinear intensity changes between VS–LWIR images. LIOP obtains the worst results.

AUC: Figure 4.23(a) compares the local descriptors based on average AUC (%) scores. These scores are based on average Precision-Recall curves, which are shown in Figure 4.23(b). The best performance is demonstrated by MN-SIFT, whereas NG-SIFT obtains the second best results. LBPG performs comparable to OR-SIFT; but outperforms SIFT. DE-SIFT, LC-SIFT, and GOM-SIFT demonstrate better performance than SIFT. SURF, DAISY, and GLOH all obtain low AUC scores than SIFT. LIOP obtains the worst results, which shows that intensity based descriptors are less suitable for VS–LWIR image matching than the gradient based descriptors.

4.5.2 Effect of Overlap Error

This section presents the comparison of local descriptors based on VS–LWIR image matching under various overlap error thresholds. Section 4.5.1 describes similar results when 50% overlap error is used. This section presents additional results when overlap error thresholds of 10%, 20%, 30%, 40%, and 60%, are used.

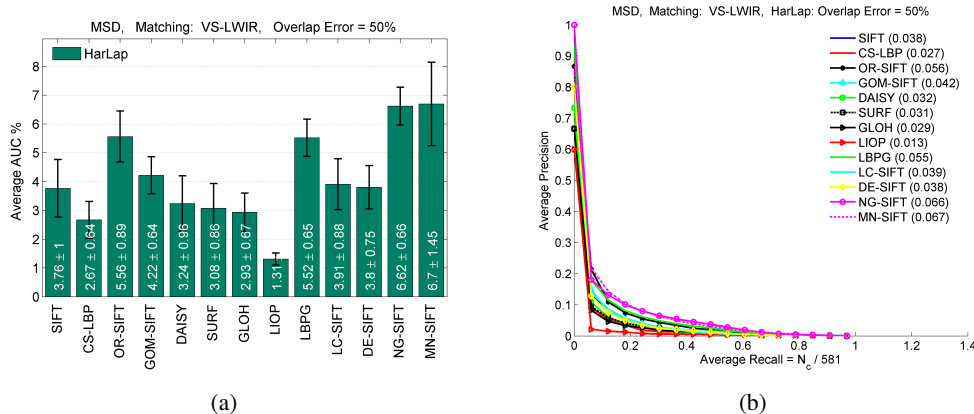


Figure 4.23: Average AUC (%) scores and Precision–Recall curves when all scenes of the MSD dataset are considered for image matching with 50% overlap error.

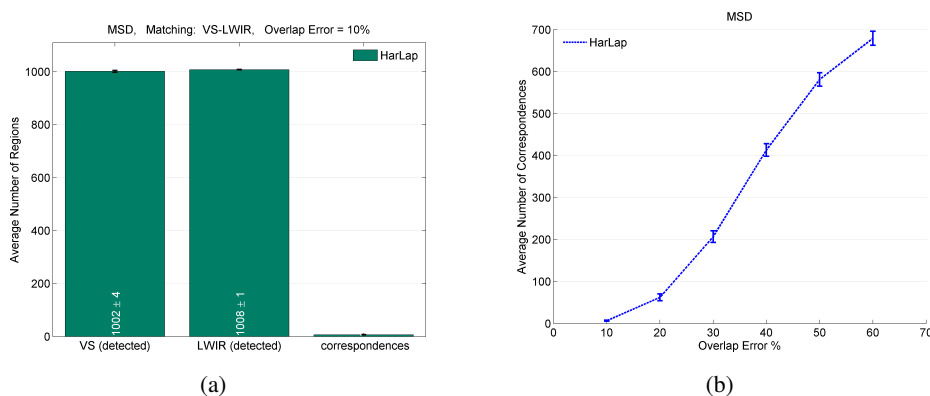


Figure 4.24: Average number of HarLap region correspondences when all scenes of the MSD dataset are considered for image matching under various overlap error thresholds. The results based on overlap error threshold of 50% are shown in Figure 4.22(a).

Repeatability: Figure 4.24 shows the average number of HarLap region correspondences under various overlap error thresholds. The average number of correspondences between VS–LWIR images is 7 at 10% overlap error, as shown in Figures 4.24(a) and (b). An increase in the number of correspondences with increasing overlap error thresholds can be seen in Figure 4.24(b). A larger overlap error results in a large number of correspondences between the VS–LWIR images and increases the number of correct nearest neighbor descriptor matches, as shown in Figure 4.25, for the MSD dataset.

Nearest neighbor matches: Figure 4.25 compares the local descriptors based on average number of correct nearest neighbor descriptor matches under various overlap error thresholds. All the local descriptors obtain low number of correct matches at 10% overlap error. For instance, MN-SIFT obtains only 1 correct match out of 7 correspondences (Figure 4.24(a)),

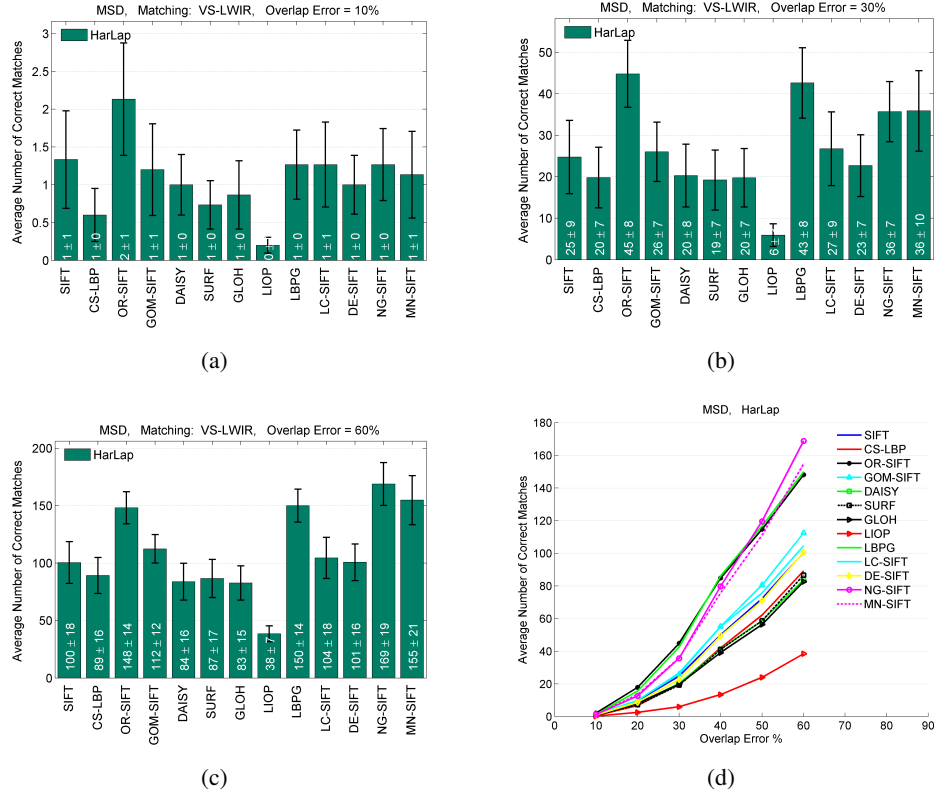


Figure 4.25: Average number of correct nearest neighbor descriptor matches when all scenes of the MSD dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error threshold are shown in Figure 4.22(b).

whereas OR-SIFT demonstrate the best results. OR-SIFT also demonstrates the best performance at 20%, 30% and 40% overlap errors. MN-SIFT and NG-SIFT achieve the second best results. However, at 60% overlap error, NG-SIFT demonstrates the best number of correct matches, i.e., 169, whereas MN-SIFT obtains the second best results and achieves 155 correct matches.

LBP and OR-SIFT demonstrate the third best results at 60% overlap error. GOM-SIFT achieves 112 correct matches and outperforms LC-SIFT and DE-SIFT. LC-SIFT obtains better results than SIFT. DE-SIFT performs comparable to SIFT. GLOH, SURF, CS-LBP, DAISY, and LIOP demonstrate lower performance than SIFT. Compared to the intensity based descriptors of LIOP, all the gradient based descriptors obtain large number of correct matches with increasing overlap error in VS-LWIR image matching.

AUC: Figure 4.26 shows the effects of overlap error thresholds on the average AUC (%) scores of the local descriptors. All the local descriptors demonstrate lower AUC scores at 10% overlap error. However, with certain overlap error thresholds, AUC scores improve. For instance, OR-SIFT outperforms other local descriptors between 10% and 40% overlap errors. Afterwards, OR-SIFT demonstrates lower AUC scores than NG-SIFT, which obtains the best

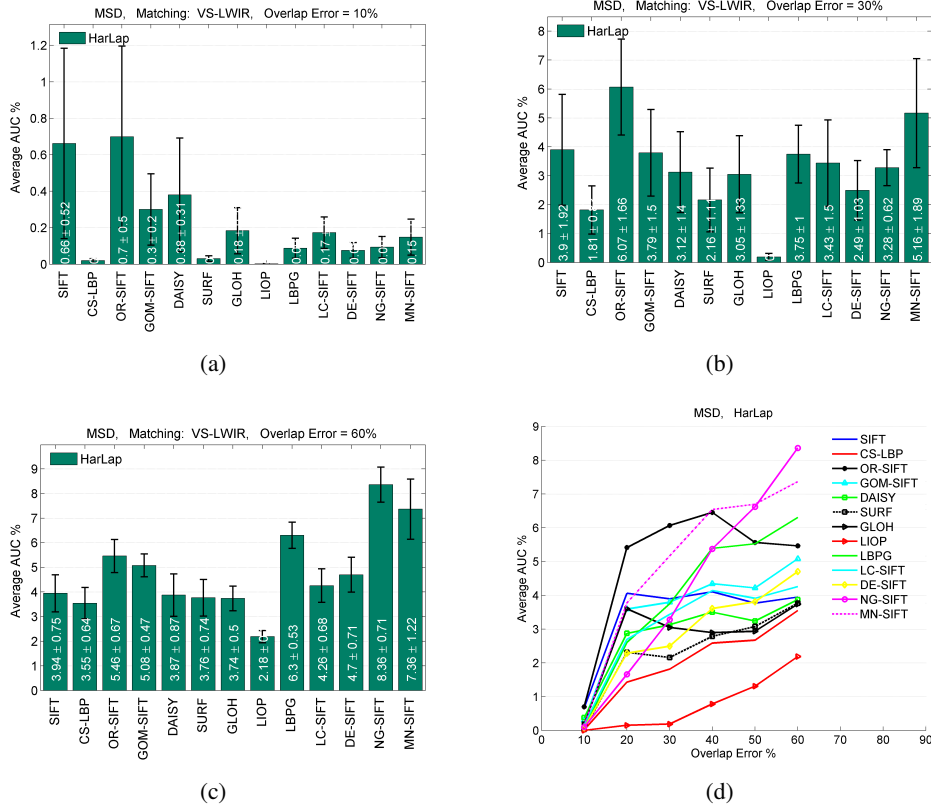


Figure 4.26: Average AUC scores when all scenes of the MSD dataset are considered for image matching under various overlap error thresholds. The results based on 50% overlap error are shown in Figure 4.23(a).

AUC score at 60% overlap error. NG-SIFT achieves an AUC score of 8.4% and demonstrates 4.4% improvement over SIFT. MN-SIFT achieves the second best AUC and outperforms SIFT by 2.9%. LBPG obtains the third best result. SIFT achieves better AUC scores than GLOH, SURF, DAISY, LIOP, and CS-LBP.

OR-SIFT and GOM-SIFT are based on gradient orientation modification. They demonstrate better performance than gradient magnitude modification based descriptors, such as LC-SIFT and DE-SIFT. This is due to high intensity reversal between the VS and LWIR images of the MSD dataset. However, the experimental results show that LC-SIFT and DE-SIFT demonstrate better performance than OR-SIFT and GOM-SIFT on CAVE and RWHI datasets.

4.6 RGB-NIR Scene Dataset

RGB-NIR Scene Dataset⁴ consists of 477 RGB-NIR image pairs in 9 different scene categories [5]. The scene categories are Country, Field, Forest, Indoor, Mountain, Buildings, Street, Urban, and Water. Each scene category depict different scenarios in Visible (RGB) and Near Infra-Red (NIR) bands, as shown in Figure 4.27. To perform image matching on the images of the RGB-NIR scene dataset, a Blue-NIR and a Red-NIR image matching per RGB-NIR pair is performed.

4.6.1 Experiments on RGB-NIR Scene Dataset

This section presents the experiment results for image matching when all the RGB-NIR image pairs of the RGB-NIR scene dataset are considered for Blue-NIR and Red-NIR image matching. The experiments are performed on the pairs of images of the same scene, and then, the average results are reported. To evaluate the performance of the local descriptors, the experimental results for each scene category are separately presented in order to show which local descriptor obtains the best performance in which scene category of the dataset. Additionally, the performance comparison of the local descriptors under various overlap error thresholds are also reported.

Country Category

This section presents the experimental results when all the RGB-NIR image pairs of the Country category are considered for Blue-NIR and Red-NIR image matching.

Nearest neighbor matches: Figure 4.28(a) shows the average number of correct nearest neighbor descriptor matches at 50% overlap error. The highest number of correct matches is demonstrated by MN-SIFT. It obtains 92 and 163 correct matches in Blue-NIR and Red-NIR image matching, respectively. The average of these matches is $(92+163)/2=128$, which is shown in Figure 4.28(b) at 50% overlap error. This figure compares the local descriptors based on average number of correct matches with respect to overlap error thresholds. The figure shows similar results for all the local descriptors, at 10% overlap error, but afterwards, the local descriptors exhibit different performances. They obtain the highest number of correct matches at 60% overlap error, where MN-SIFT and NG-SIFT obtains 150 correct matches and outperform SIFT and other local descriptors. LBPG obtains the second best result. LC-SIFT and DE-SIFT demonstrate better performance than SIFT. GLOH, OR-SIFT, GOM-SIFT, GLOH, DAISY, and CS-LBP perform inferior to SIFT. LIOP obtains the worst performance.

AUC: Figure 4.28(c) compares the local descriptors based on average AUC (%) scores at 50% overlap error. The best AUC score is demonstrated by MN-SIFT. It demonstrates AUC scores of 8.2% and 17% in Blue-NIR and Red-NIR image matching, respectively. The average of these AUC scores is $(8.2+17)/2=12.7\%$, which is shown in Figure 4.28(d) at 50% overlap error. This figure compares the local descriptors based on average AUC scores with respect to overlap error thresholds. This comparison shows wide variations in the AUC scores of the local

⁴http://ivrg.epfl.ch/supplementary_material/cvpr11/

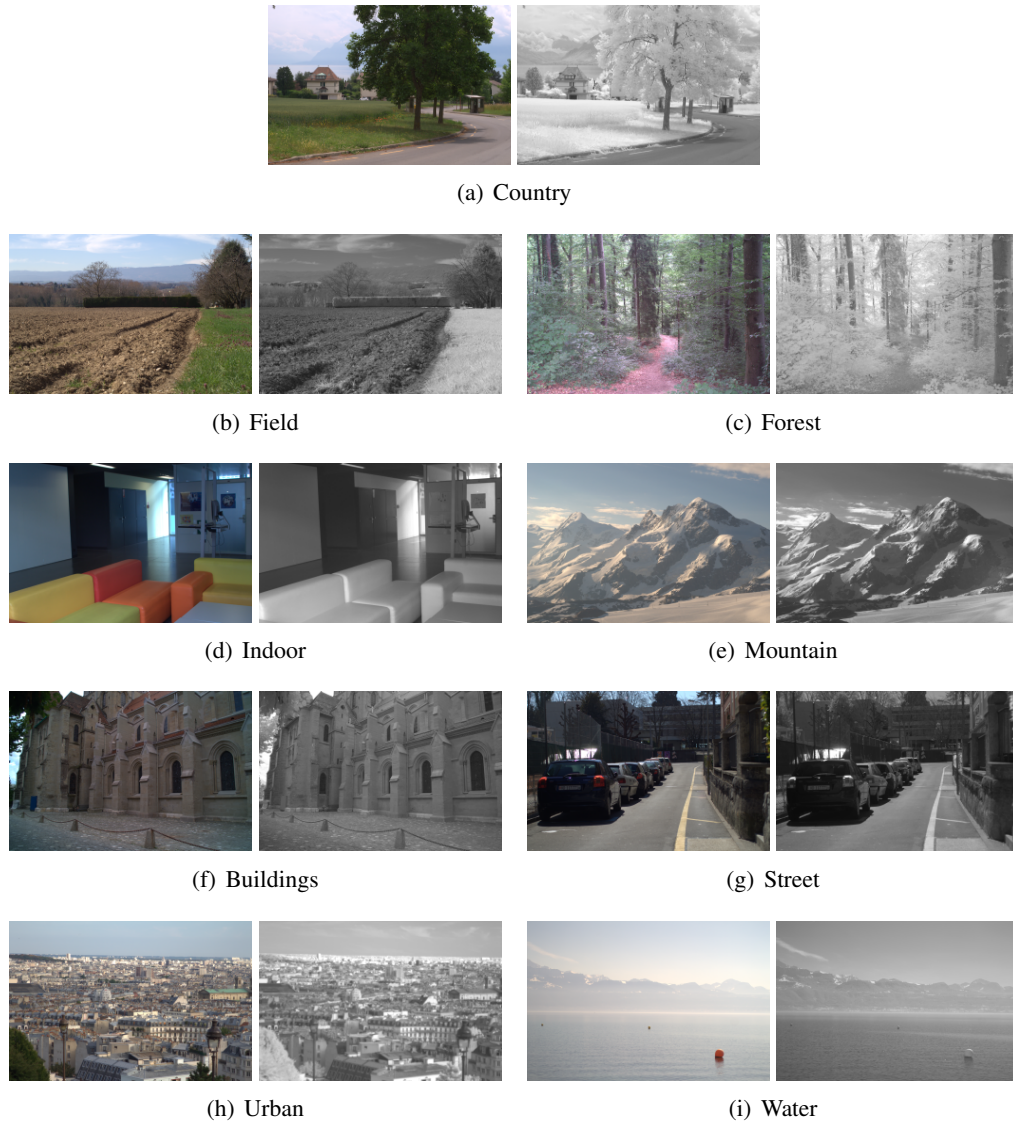


Figure 4.27: Samples from the evaluated RGB-NIR scene dataset [5]. An RGB-NIR image pair from each scene category is shown. Each pair shows the same image contents in RGB (left) and NIR (right) bands.

descriptors. The local descriptors demonstrate the best AUC scores at 20% overlap error. Afterwards, a decline in the AUC scores with respect to overlap error can be seen. MN-SIFT and NG-SIFT show the best AUC scores at 60% overlap error and outperform others. They achieve AUC scores of 10% and 9.5% and demonstrate 2.5% and 2% improvement over SIFT, respectively. LC-SIFT, DE-SIFT, and LBPg achieve similar AUC scores and perform comparable to SIFT.

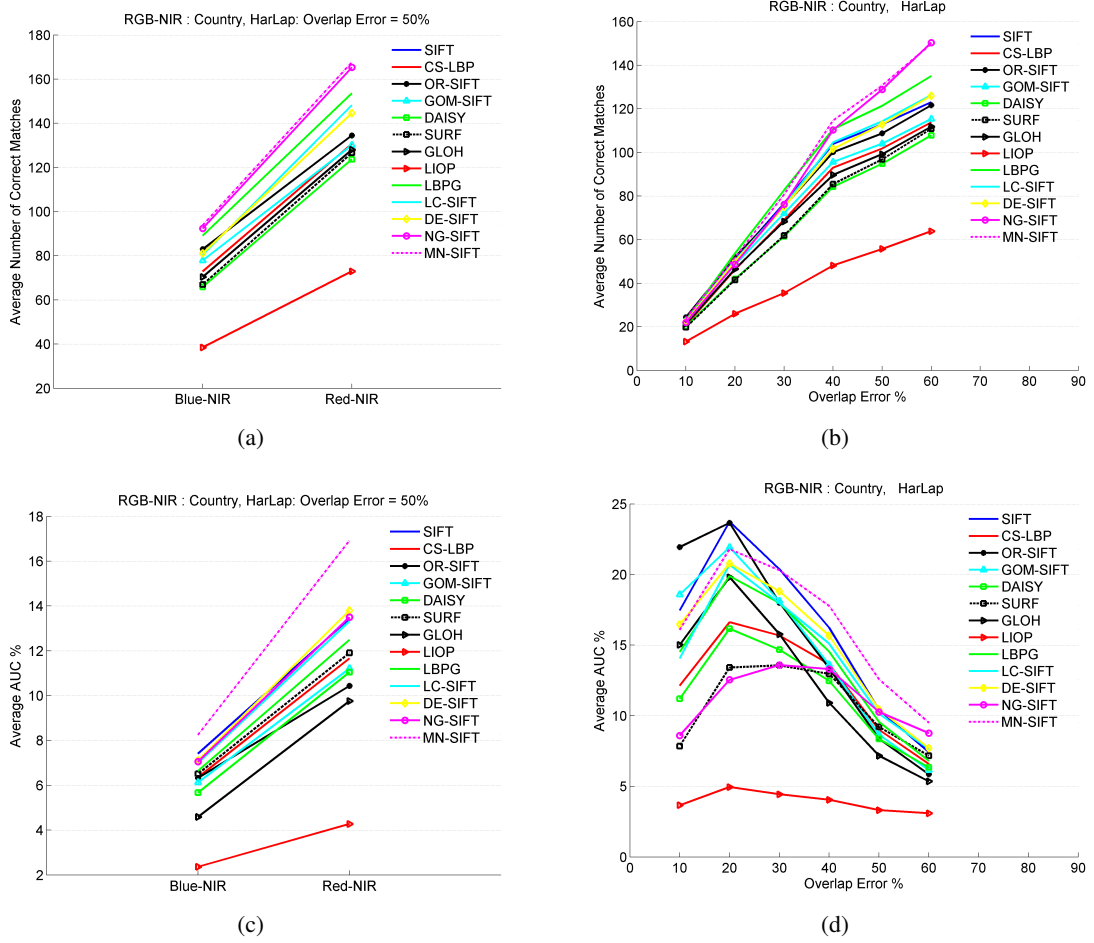


Figure 4.28: (a,b) Average number of correct nearest neighbor matches and (c,d) average AUC (%) scores when all the RGB-NIR image pairs of the Country category are considered for Blue-NIR and Red-NIR image matching under various overlap error thresholds.

Field Category

Figure 4.29 shows the image matching results when all the RGB-NIR image pairs of the Field category are considered for Blue-NIR and Red-NIR image matching. The experimental results show the highest number of correct matches for MN-SIFT under various overlap errors. All the local descriptors demonstrates the best results at 60% overlap error. In the case of AUC scores, all the local descriptors obtain the best results at 20% overlap error, where LBPG and SIFT outperform the other local descriptors. Afterwards, a decline in the AUC scores with respect to overlap error can be seen.

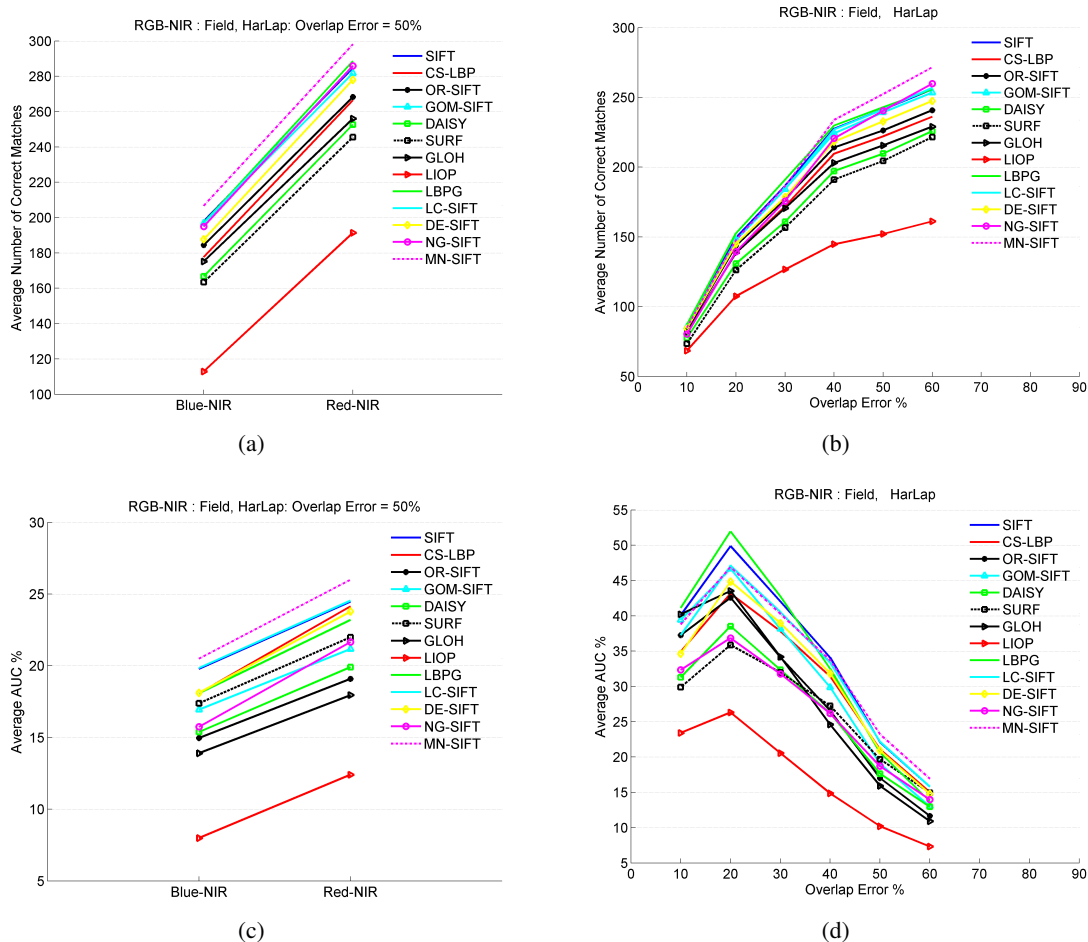


Figure 4.29: Image matching results for Field category.

Forest Category

Figure 4.30 shows the experimental results when all the RGB-NIR image pairs of the Forest category are considered for Blue-NIR and Red-NIR image matching.

Indoor Category

Figure 4.31 shows the experimental results when all the RGB-NIR image pairs of the Indoor category are considered for Blue-NIR and Red-NIR image matching.

Mountain Category

Figure 4.32 shows the experimental results when all the RGB-NIR image pairs of the Mountain category are considered for Blue-NIR and Red-NIR image matching.

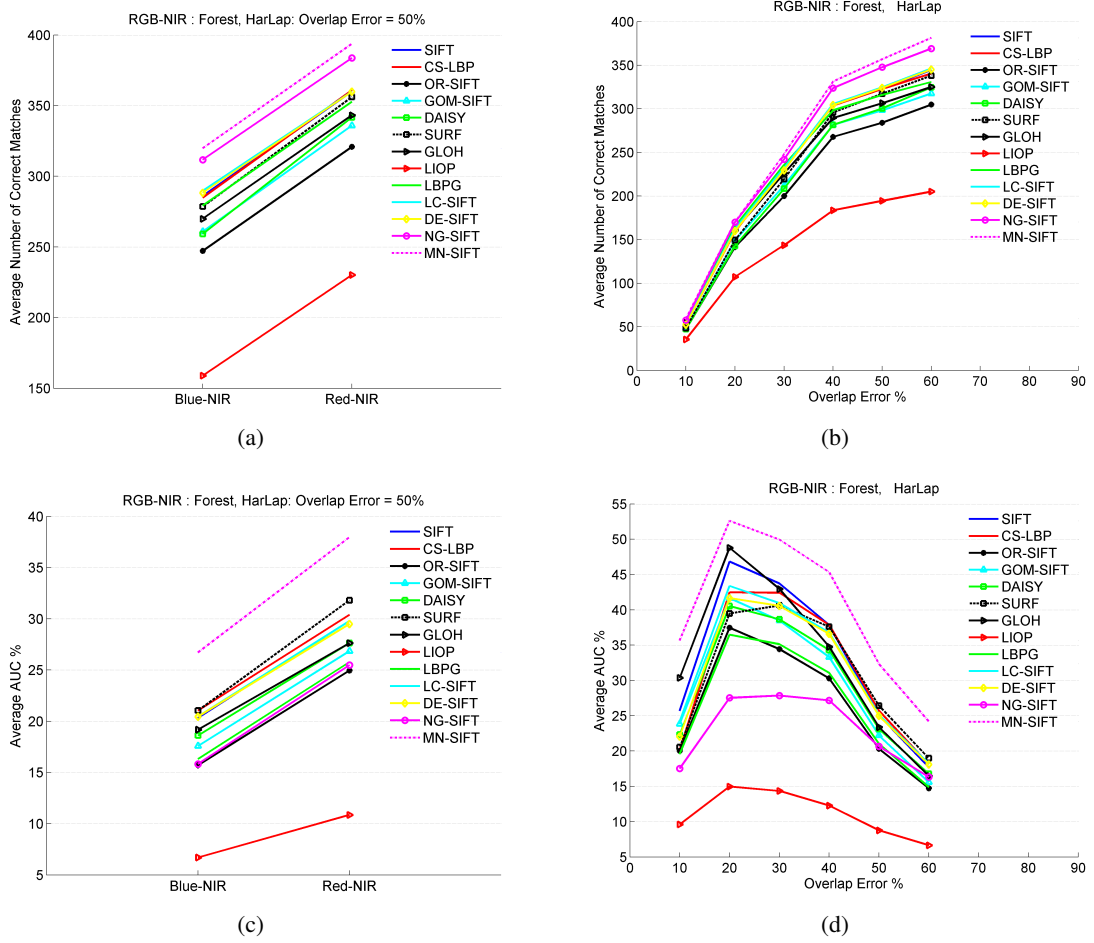


Figure 4.30: Image matching results for Forest category.

Building Category

Figure 4.33 shows experimental results when all the RGB-NIR image pairs of the Buildings category are considered for Blue-NIR and Red-NIR image matching.

Street Category

Figure 4.34 shows the experimental results when all the RGB-NIR image pairs of the Street category are considered for Blue-NIR and Red-NIR image matching.

Urban Category

Figure 4.35 shows the experimental results when all the RGB-NIR image pairs of the Urban category are considered for Blue-NIR and Red-NIR image matching.

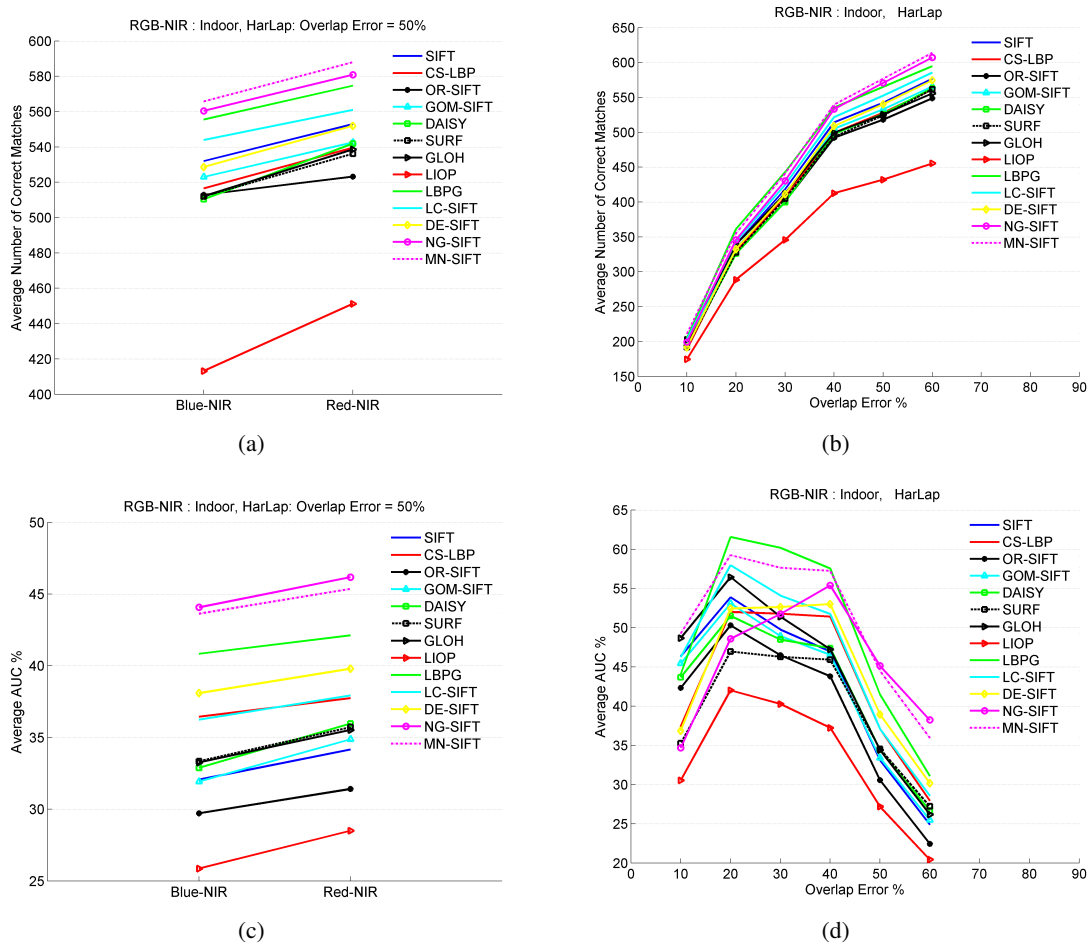


Figure 4.31: Image matching results for Indoor category.

Water Category

Figure 4.36 shows the experimental results when all the RGB-NIR image pairs of the Water category are considered for Blue-NIR and Red-NIR image matching.

4.7 Summary

In this chapter the experimental results for image matching were presented. The performance of the descriptors proposed (LBP, LC-SIFT, DE-SIFT, MN-SIFT and MN-SIFT) were compared with that of state-of-the-art local descriptors (SIFT, OR-SIFT, GOM-SIFT, GLOH, CS-LBP, SURF, DAISY and LIOP) on four different multispectral image datasets (CAVE, RWHI, RGB-NIR and MSD). All the experiments were performed under nonlinear intensity changes by performing image matching between the pairs of different wavelength band images of the

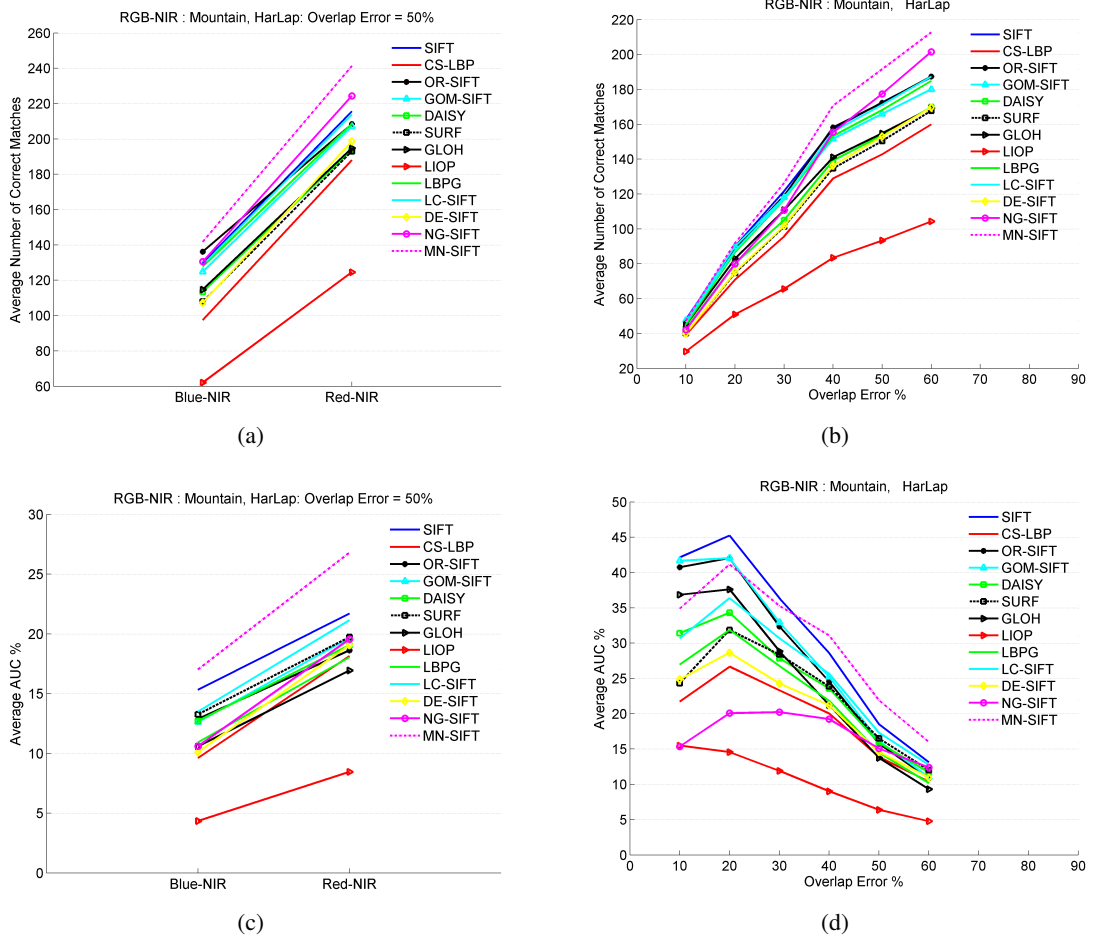


Figure 4.32: Image matching results for Mountain category.

same scene. The performance comparison was based on average number of correct nearest neighbor descriptor matches, average Precision–Recall curves, and average AUC scores. The experimental results showed that SIFT demonstrated on average better performance than other state-of-the-art local descriptors (OR-SIFT, GOM-SIFT, GLOH, CS-LBP, SURF, DAISY and LIOP). Therefore, SIFT was used as a reference descriptor to evaluate the performance of the descriptors proposed.

The experimental results showed that the gradient orientation modification based descriptors, i.e., OR-SIFT and GOM-SIFT, which have been used to overcome the effects of nonlinear intensity changes, demonstrated lower performance than SIFT on CAVE, RHWI and RGB-NIR datasets. But they demonstrated comparable performance to SIFT on MSD dataset. The descriptors proposed, which are based on gradient magnitude modification obtain on average better results than SIFT, OR-SIFT, and GOM-SIFT on all the evaluated datasets. This proves that the modification of the gradient magnitudes, which are proposed for the construction of the de-

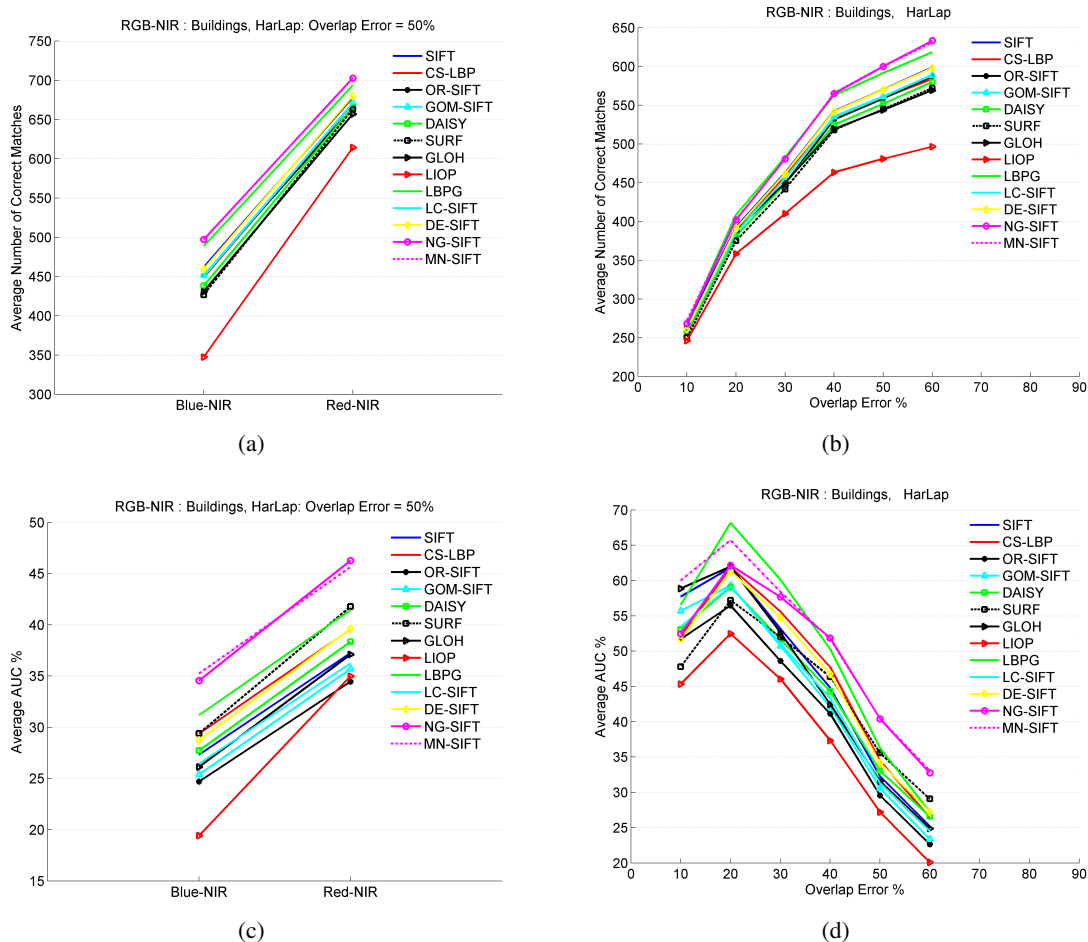


Figure 4.33: Image matching results for Building category.

scriptors proposed, improve the performance of the local descriptors with regard to nonlinear intensity changes.

CS-LBP have been used to overcome the effects of monotonic intensity changes. It also demonstrates good performance under nonlinear intensity changes. The experimental results show that LBPG, which is a gradient extension of CS-LBP, demonstrates even better performance than CS-LBP on all the evaluated datasets. This shows that the modification of image gradients based on the CS-LBP scheme, generates more robust image features, which improve the performance of the descriptors proposed against nonlinear intensity changes than the original/standard image gradients based descriptors.

LIOP demonstrates the worst performance on all the evaluated datasets. This shows that the intensity based descriptors that LIOP computes, are less suitable for image matching under nonlinear intensity changes. GLOH, SURF, and DAISY demonstrate on average a slightly lower performance than SIFT. All these descriptors demonstrate better performance than LIOP, OR-

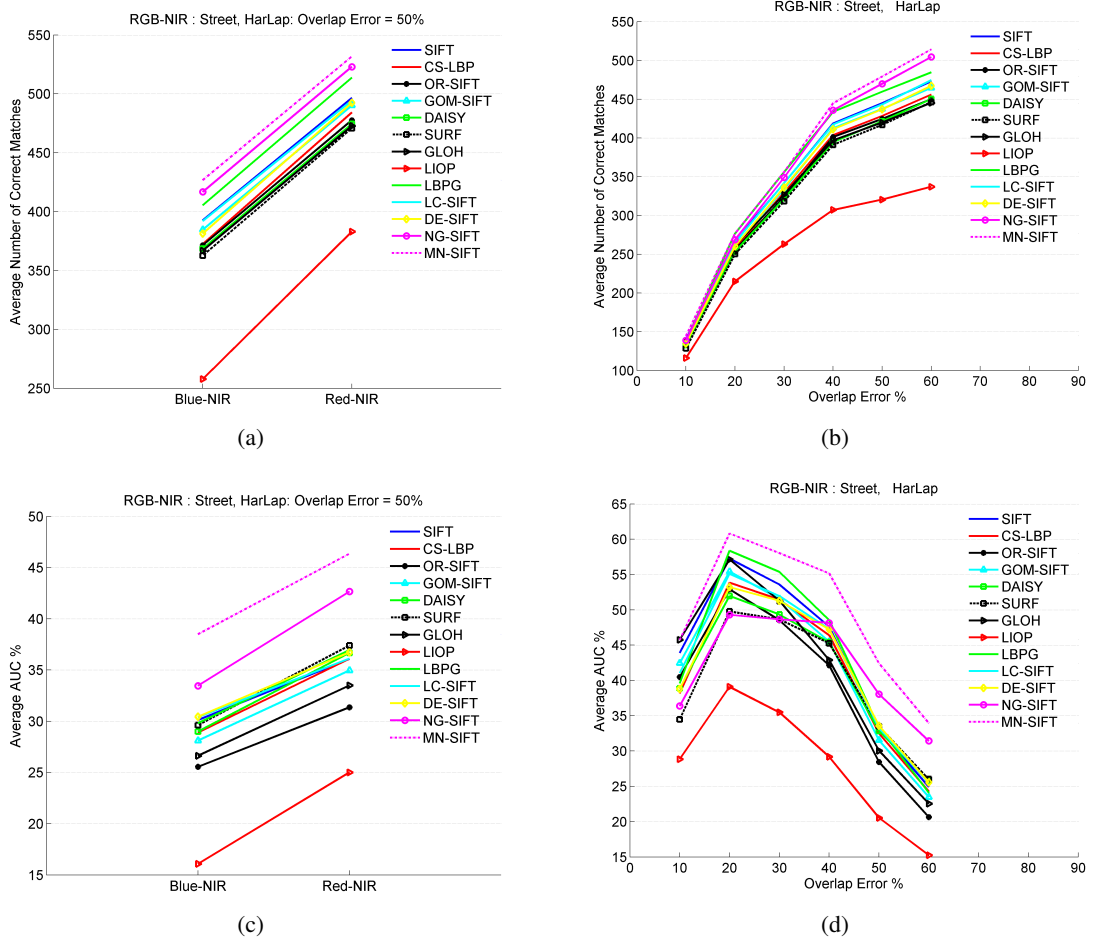
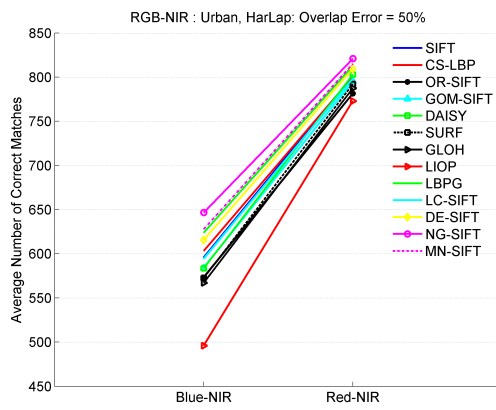


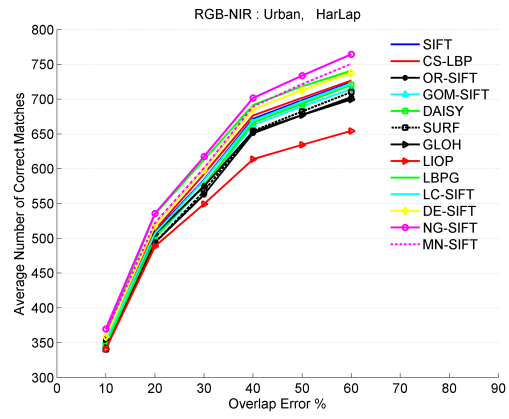
Figure 4.34: Image matching results for Street category.

SIFT, and GOM-SIFT.

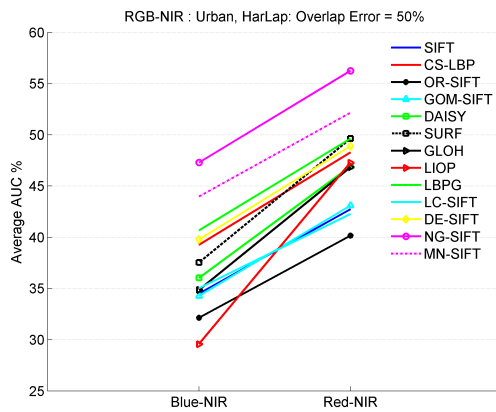
NG-SIFT and MN-SIFT demonstrate the best performance on all the evaluated datasets. They are based on normalized image gradients. The experimental results show that such normalized image gradients are more robust to nonlinear intensity changes than image gradients, CS-LBP features, and Haar wavelet responses, which are used in the construction of SIFT, CS-LBP, and SURF descriptors, respectively. The experimental results show that MN-SIFT outperforms NG-SIFT on the images of textured scene (such as County, Field, Forest, Mountain, Street, and Water), whereas NG-SIFT demonstrates better performance than MN-SIFT on structured scene images (such as Buildings and Urban).



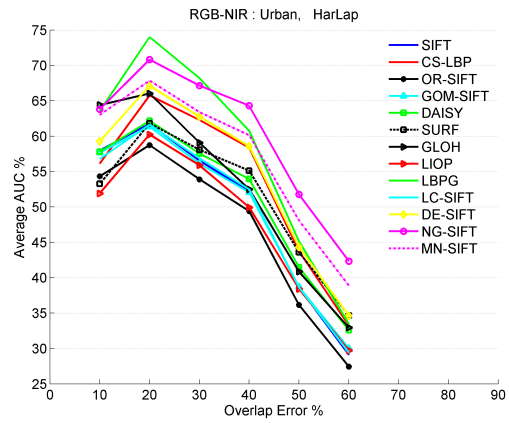
(a)



(b)



(c)



(d)

Figure 4.35: Image matching results for Urban category.

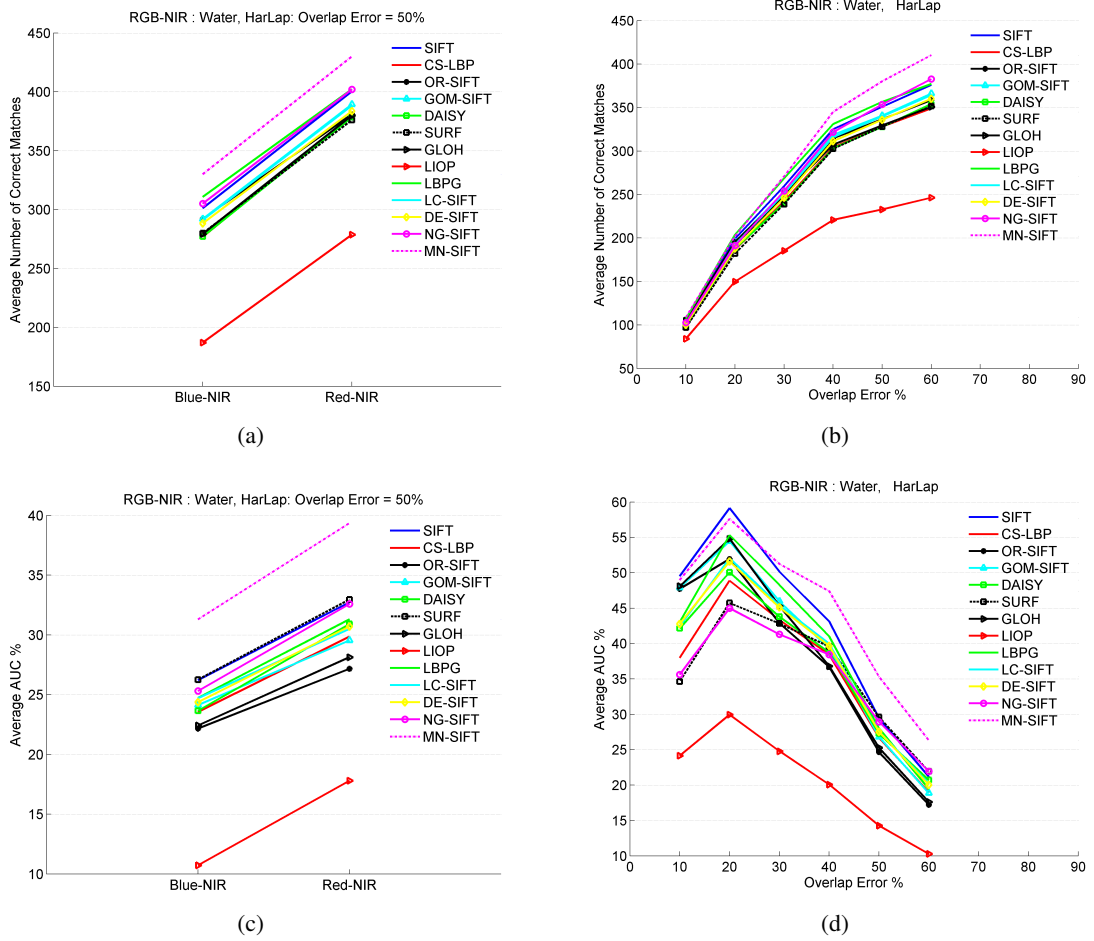


Figure 4.36: Image matching results for Water category.

Scene Category Recognition

This chapter presents the experimental results for Scene Category Recognition (SCR). It is an image-to-database matching task, which is performed in this thesis under nonlinear intensity changes to evaluate the performance of the descriptors proposed (LBP, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) and to compare their performance measures with that of state-of-the-art local descriptors (SIFT, OR-SIFT, GOM-SIFT, CS-LBP, GLOH, DAISY, and LIOP) for a recognition task. A nearest neighbor based descriptor matching strategy is used for SCR. Brown and Susstrunk [5] use the same descriptor matching strategy and evaluate the performance of SIFT and the color variants of SIFT [63] for SCR by computing the descriptor matching between the images of the same wavelength bands. Unlike the work of Brown and Susstrunk [5], the SCR presented in this chapter is performed by using images of different wavelength bands for training and testing. The idea is to induce nonlinear intensity changes between the images of the training and test sets for the performance evaluation of the local descriptors. In the following sections, the experimental setup and results are described.

5.1 Experimental Setup

The framework [49] used for SCR is shown in Figure 5.1. It consists of an RGB–NIR scene dataset [5], a 10 folds cross validation test/train split, HarLap regions, and a Naive Bayes Nearest Neighbor image classifier [4]. In the following sections, each block of the SCR framework is described in detail.

5.2 Image Dataset

Experiments are performed on an RGB–NIR Scene Dataset [5] (see Figure 4.27). This dataset consists of 477 RGB–NIR image pairs in 9 different scene categories. The scene categories are Country, Field, Forest, Indoor, Mountain, Buildings, Street, Urban, and Water. The framework uses images of different bands for training and testing. For instance, if the images of NIR band

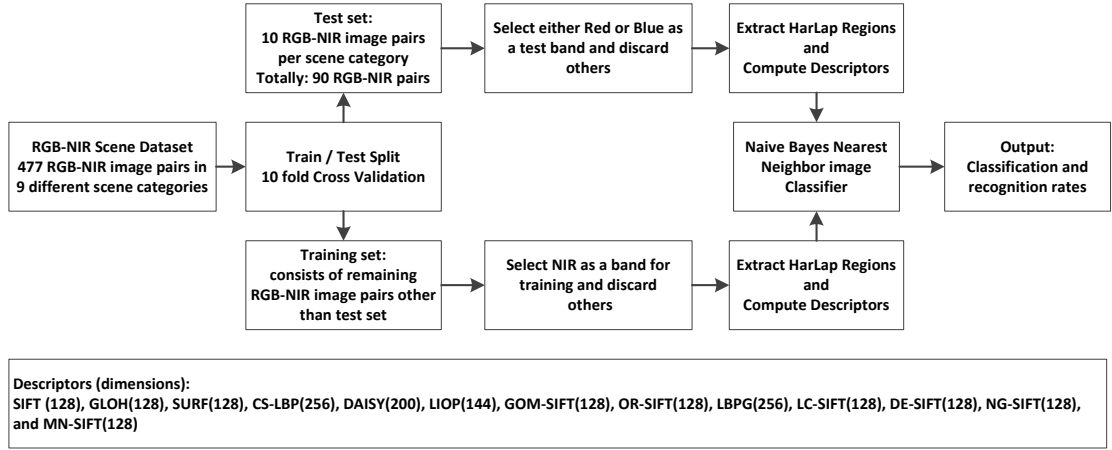


Figure 5.1: Framework for the performance evaluation of the local descriptors for SCR under nonlinear intensity changes.

are chosen for training, then testing is carried out on the images of Red (R) or Blue (B) bands in order to induce nonlinear intensity changes between the images of training and test sets.

5.3 Cross Validation

Experiments are repeated 10 times by randomly performing the train/test split. Then, the average recognition rates obtained by the local descriptors are reported for the performance comparison. For each train/test split, the 477 RGB–NIR image pairs are randomly split into training and test sets: 90 RGB–NIR image pairs (i.e., 10 RGB–NIR image pairs per scene category) are used for testing and training is performed using the rest.

5.4 Testing and Training Images

If the images of Blue band are used for testing, then, the images of other bands in the test set, i.e., Red, Green and NIR are discarded for both testing and training. Similarly, if the images of Red band are selected for testing, then, the images of other bands in the test set, i.e., Blue, Green, and NIR are discarded for both testing and training. The images of NIR band in the training set are only used for training, and the images of other bands (i.e., Red, Green, and Blue) in the training set are discarded for both testing and training.

5.5 Description of Testing and Training Images

To compute the local descriptors, HarLap [38] regions are used, as described in Section 3.2. The local descriptors used for SCR are listed in Figure 5.1 along with their description vector dimensions.

5.6 Naive Bayes Nearest Neighbor Classifier

Naive Bayes Nearest Neighbor (NB–NN) image classifier is used for SCR in this thesis. Brown and Susstrunk [5] use the same image classifier for SCR. Boimam et al. [4] propose this classifier and show that the descriptors z_i in an image I are Independent and Identically Distributed (i.i.d.) Gaussian variables and their class conditional density is approximated with the nearest neighbors as follows [4]:

$$p(I|c) = \prod_i p(z_i|c) \approx \prod_i \mathcal{N}(z_i, NN_c(z_i), \sigma^2) \quad (5.1)$$

where $NN_c(z_i)$ are nearest neighbors of z_i in class c of the training set and σ is the width of the Parzen window, which is used in the estimation of the probability density [4]. Boimam et al. show that if a single nearest neighbor in each class c of the training set is used for z_i , then Equation 5.1 obtains a simple form as follows:

$$p(I|c) = \prod_i p(z_i|c) \propto - \sum_i \|z_i - NN_c(z_i)\| \quad (5.2)$$

where $\|\cdot\|$ is an Euclidean distance between z_i and its nearest neighbor $NN_c(z_i)$ in class c of the training set. Then, by making use of the Naive Bayes classifier with equal class priors a class label for the image I is obtained as follows:

$$\hat{C} = \arg \max_c p(c|I) \approx \arg \min_c \sum_i \|z_i - NN_c(z_i)\| \quad (5.3)$$

5.7 Experimental Results

This section presents the experimental results for SCR. The results are based on average recognition rates (fraction of correct classification) obtained by the local descriptors by randomly performing the train/test split 10 times. In the following subsections, the experimental results for SCR based on the images of Blue–NIR and Red–NIR bands, are described.

5.7.1 SCR based on Blue–NIR band images

In the case of SCR based on Blue–NIR band images, the descriptors of NIR band images in the training set are compared with the descriptors of the Blue band images of the test set by using the NB–NN classifier, as described in Section 5.6. Having compared the descriptors, the scene category labels for the test images are obtained by using Equation 5.3. Then, the recognition rates are computed and the performance of the descriptors proposed is compared with that of the state-of-the-art local descriptors.

To understand the recognition rates reported in this chapter, a confusion table is shown in Figure 5.2. The confusion table shows the average recognition rates obtained by SIFT in nine different scene categories of the RGB–NIR dataset in the SCR based on Blue–NIR band images. In fact, the diagonal elements in the confusion table are the average recognition rates, which are used for the performance evaluation of SIFT with regard to other local descriptors.

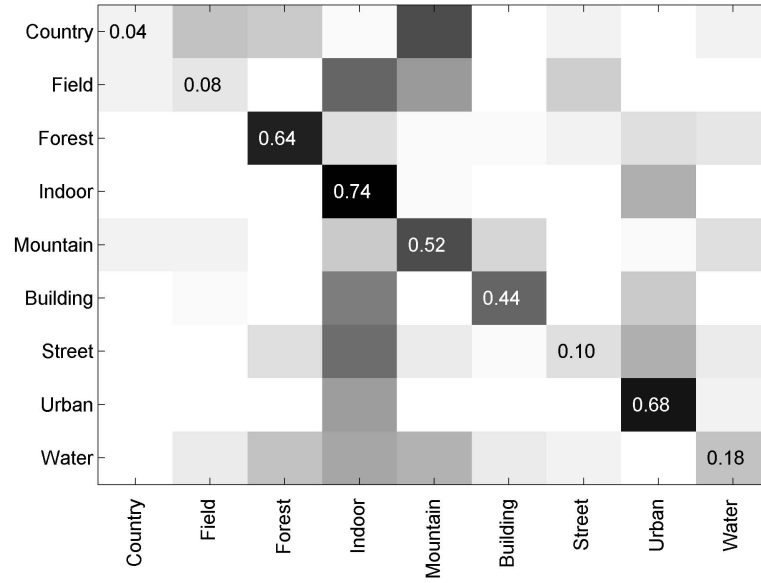


Figure 5.2: Confusion table shows the average recognition rates achieved by SIFT in nine different scene categories of the RGB–NIR scene dataset when testing and training are performed on the images of Blue and NIR bands, respectively.

The recognition rates are shown with different gray shades for better visualization and comparison of the recognition rates. The darkest shade (full black) depicts a recognition rate of 1 and the brightest shade (full white) depicts a recognition rate of 0. The rows of the confusion table correspond to ground truth (actual labels) and the columns correspond to predicted labels. The sum of recognition rates over each row gives a recognition rate of 1, i.e., 100% recognition rate.

The confusion table shows that SIFT achieves the best recognition rate of 0.74 in the indoor category. Therefore, the other columns of the same row show lighter shades. Similarly, the recognition rates in other rows of the confusion table can also be explained. The diagonal elements of the confusion table are also shown in Table 5.1 to compare the performance of SIFT with that of other local descriptors.

Figure 5.3 shows the confusion tables when all the local descriptors are used for SCR based on Blue–NIR images. The diagonal elements of these confusion tables are shown in Table 5.1, where the best result in each column is printed in boldface. The results show wide variations in the average recognition rates of the local descriptors. The best mean recognition rate (i.e., overall average recognition rate) is demonstrated by MN-SIFT. It achieves a mean recognition rate of 0.43 (last column).

GOM-SIFT obtains the second best results. Note that, GOM-SIFT demonstrates a low performance in Blue–NIR image matching, as described in Section 4.6; but in the case of SCR based on Blue–NIR band images, it achieves the second best mean recognition rate, which shows its more robustness to nonlinear intensity changes in the recognition task, i.e., SCR compared to

Table 5.1: Average Recognition rates when testing and training are performed on the images of Blue and NIR bands of the RGB–NIR scene dataset, respectively. Mean represents the average recognition rates over all the scene categories.

Descriptor	Country	Field	Forest	Indoor	Mountain	Buildings	Street	Urban	Water	Mean
SIFT	0.04	0.08	0.64	0.74	0.52	0.44	0.10	0.68	0.18	0.38
CS-LBP	0.00	0.40	0.32	0.84	0.14	0.32	0.16	0.48	0.34	0.33
OR-SIFT	0.10	0.22	0.36	0.54	0.24	0.46	0.10	0.32	0.32	0.30
GOM-SIFT	0.08	0.34	0.74	0.70	0.38	0.50	0.12	0.48	0.34	0.41
DAISY	0.00	0.24	0.76	0.68	0.64	0.44	0.16	0.54	0.14	0.40
SURF	0.06	0.26	0.36	0.78	0.78	0.24	0.20	0.32	0.04	0.34
GLOH	0.06	0.06	0.80	0.68	0.46	0.42	0.12	0.58	0.08	0.36
LIOP	0.04	0.16	0.62	0.48	0.12	0.06	0.18	0.32	0.06	0.23
LBPG	0.08	0.10	0.46	0.34	0.32	0.26	0.30	0.26	0.06	0.24
LC-SIFT	0.08	0.32	0.52	0.60	0.40	0.50	0.22	0.64	0.22	0.39
DE-SIFT	0.04	0.30	0.60	0.60	0.18	0.36	0.36	0.42	0.16	0.34
NG-SIFT	0.26	0.34	0.48	0.54	0.22	0.14	0.14	0.46	0.08	0.30
MN-SIFT	0.10	0.42	0.74	0.74	0.54	0.22	0.16	0.66	0.28	0.43

image matching. Similarly, DAISY demonstrates better mean recognition rates than SIFT. The results show adverse effects of nonlinear intensity changes on the performance of NG-SIFT, which achieves a mean recognition rate of 0.30 and demonstrates the best average recognition rates only in the Country category. LIOP obtains the worst performance, which shows that the intensity based descriptors that LIOP computes, are less suitable for SCR under nonlinear intensity changes.

LC-SIFT obtains a mean recognition rate of 0.39 and demonstrates an improvement of 1% over SIFT. DE-SIFT obtains a lower mean recognition rate than SIFT. LBPG achieves the second lowest mean recognition rates. It performs only 1% better than LIOP. SIFT demonstrates better recognition rate than OR-SIFT, SURF, CS-LBP, GLOH, and LIOP. Figure 5.3 also shows confusion between the scene categories. For instance, it can be seen that the Building category is confused with the Indoor category, and the Urban and Country categories are confused with Field and Forest categories.

5.7.2 SCR based on Red–NIR band images

This section describes the experimental results when Red–NIR band images are used for SCR, i.e., the descriptors of NIR band images of the training set are compared with the descriptors of Red band images of the test set by using the NB–NN, as described in Section 5.6. All the results presented in this section are average recognition rates, which are obtained by performing the train/test split 10 times.

Figure 5.4 shows the confusion tables for SCR based on Red–NIR band images. The diagonal elements (average recognition rates) of these confusion tables are also shown in Table 5.2. The best results in each column are printed in boldface. The results show variations in the average recognition rates of the local descriptors with regard to scene categories. The mean

Table 5.2: Average Recognition rates when testing and training are performed on the images of Red and NIR bands of the RGB–NIR Scene dataset, respectively. Mean represents average recognition rates over all the scene categories.

Descriptor	Country	Field	Forest	Indoor	Mountain	Buildings	Street	Urban	Water	Mean
SIFT	0.14	0.10	0.68	0.68	0.64	0.62	0.10	0.52	0.34	0.42
CS-LBP	0.08	0.30	0.38	0.92	0.26	0.34	0.24	0.54	0.32	0.38
OR-SIFT	0.02	0.32	0.36	0.58	0.26	0.44	0.18	0.36	0.26	0.31
GOM-SIFT	0.00	0.36	0.80	0.72	0.38	0.62	0.12	0.48	0.34	0.42
DAISY	0.20	0.24	0.70	0.60	0.44	0.64	0.12	0.50	0.22	0.41
SURF	0.08	0.18	0.24	0.78	0.64	0.40	0.20	0.34	0.12	0.33
GLOH	0.12	0.10	0.78	0.66	0.38	0.62	0.14	0.52	0.16	0.39
LIOP	0.00	0.16	0.70	0.42	0.16	0.16	0.10	0.20	0.08	0.22
LBPG	0.14	0.12	0.48	0.28	0.32	0.32	0.20	0.46	0.12	0.27
LC-SIFT	0.10	0.26	0.62	0.58	0.38	0.62	0.14	0.38	0.12	0.36
DE-SIFT	0.16	0.24	0.64	0.62	0.34	0.40	0.44	0.56	0.28	0.41
NG-SIFT	0.24	0.42	0.50	0.70	0.24	0.38	0.14	0.42	0.04	0.34
MN-SIFT	0.08	0.34	0.66	0.78	0.56	0.64	0.18	0.56	0.30	0.46

recognition rates in the last column show that MN-SIFT obtains the best results. It achieves a mean recognition rate of 0.46, and outperforms all the other local descriptors. Note that, this recognition rate is 3% higher than 0.43, which MN-SIFT demonstrates in the case of SCR based on Blue–NIR band images (Table 5.1). This performance improvement is due to low nonlinear intensity changes between the images of Red–NIR bands compared to Blue–NIR band images. This shows that low intensity changes less affect the nearest neighbor descriptor matches (Equation 5.2) and result in better recognition rates than SCR based on Blue–NIR band images. Similarly, the improvement in the recognition rates of the other local descriptors can also be seen.

Table 5.2 also shows the effects of image contents on the average recognition rates of the local descriptors. For instance, SIFT obtains the best average recognition rates of 0.64 and 0.34 in Mountain and Water categories, respectively. NG-SIFT obtains the best recognition rates of 0.24 and 0.42 in Country and Field categories, respectively. DE-SIFT shows the best recognition rates in Street and Urban categories. GOM-SIFT achieves the best recognition rates in Forest and Water categories. Both GOM-SIFT and SIFT obtain the best recognition rates in Water category. But, GOM-SIFT outperforms SIFT in Forest category, whereas SIFT outperforms GOM-SIFT in Mountain category. Both DAISY and MN-SIFT obtain the best recognition rates in Building category. Both SURF and SIFT achieve the best recognition rates in Mountain category. Whereas OR-SIFT, LC-SIFT, LBPG, GLOH, and LIOP do not demonstrate the best recognition rates in any of these scene categories of the RGB-NIR dataset.

The last column in Table 5.2 shows the best mean recognition rate of 0.46 for MN-SIFT, whereas both SIFT and GOM-SIFT obtain the second best mean recognition rates of 0.42. LIOP obtains the worst mean recognition rates. Although, LIOP achieves some recognition rates, but compared to that of gradient based descriptors, such recognition rates are much lower, which show that gradient based descriptors are more robust to nonlinear intensity changes. It can also

be seen that SIFT, DE-SIFT, GOM-SIFT, MN-SIFT, and DAISY obtain comparable recognition rates for SCR based on Red–NIR bands, whereas other local descriptors demonstrates lower recognition rates.

5.8 Summary

In this chapter, the experimental results for SCR were presented, in order to compare the performance of descriptors proposed (LBPG, LC-SIFT, DE-SIFT, NG-SIFT and MN-SIFT) with that of state-of-the-art local descriptors (SIFT, GLOH, SURE, CS-LBP, DAISY, LIOP, GOM-SIFT, and OR-SIFT) for a recognition task. SCR was performed on nine different scene categories of the RGB-NIR scene dataset. All experiments were performed under nonlinear intensity changes, i.e., by using the images of different wavelength bands for training and testing.

The experimental results show that the performance of LC-SIFT and MN-SIFT is comparable to SIFT, GOM-SIFT, DAISY, and GLOH when testing and training are performed on the images of Blue and NIR bands, respectively. Whereas, NG-SIFT, DE-SIFT, and LBPG obtain lower scene recognition rates.

The experimental results show that the performance of DE-SIFT and MN-SIFT is comparable to SIFT, GOM-SIFT, and DAISY when testing and training are performed on the images of Red and NIR bands, respectively. Whereas, NG-SIFT, LC-SIFT and LBPG demonstrate lower scene recognition rates.

LIOP obtains the worst results for SCR, which shows that intensity based descriptors are less suitable for the recognition task under nonlinear intensity changes. All the gradient based descriptors demonstrate better performance than LIOP.

The best recognition rates in both Red–NIR and Blue–NIR bands based SCR are demonstrated by MN-SIFT. The experimental results show that all the local descriptors demonstrate good performance for SCR when the images of Red–NIR bands are used due to low nonlinear intensity changes compared to ones between the images of Blue–NIR bands.

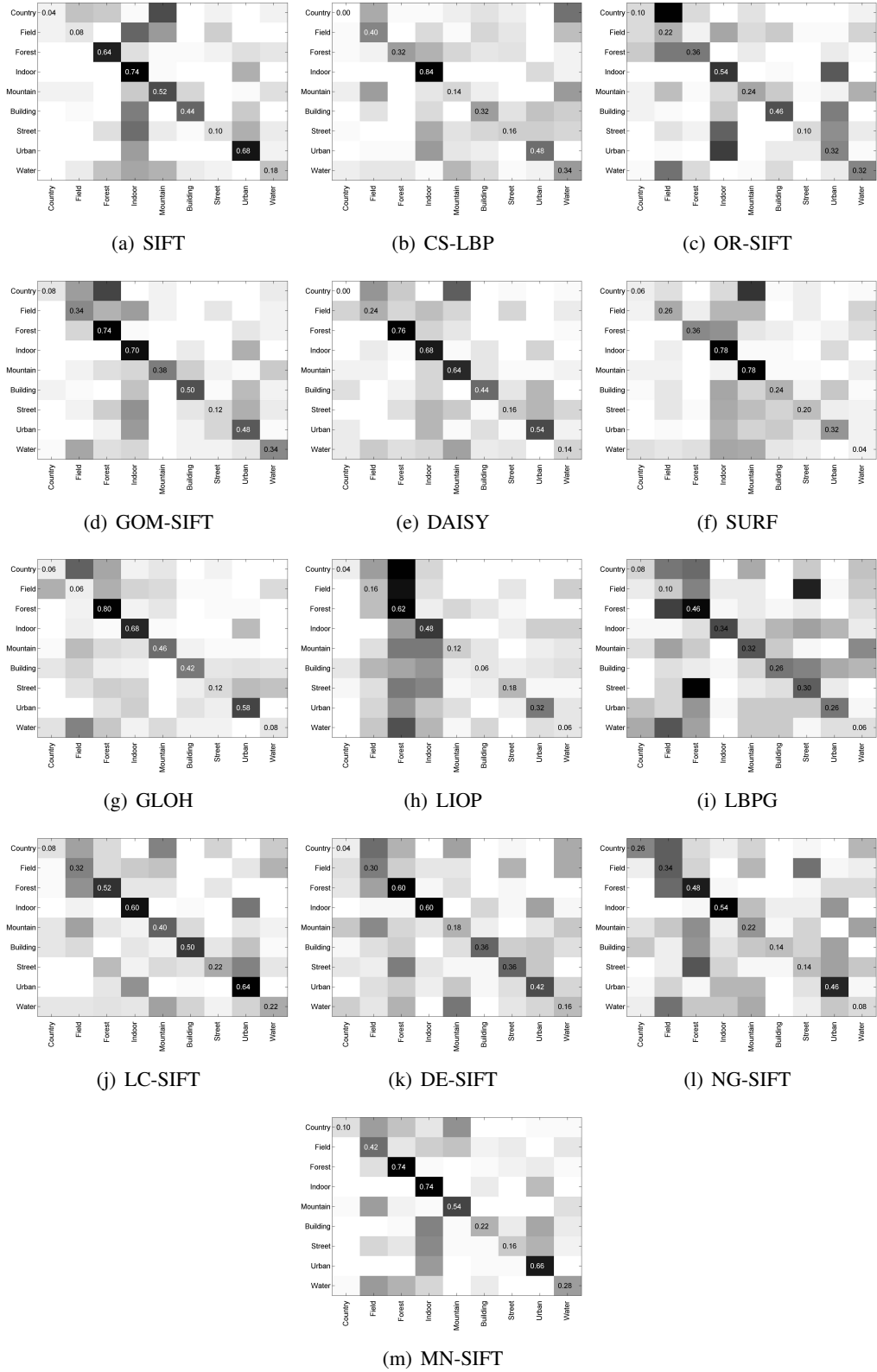


Figure 5.3: Confusion table shows the average recognition rates achieved by the local descriptors in nine different scene categories of the RGB–NIR scene dataset when testing and training are performed on the images of Blue and NIR bands, respectively. The average recognition rates (diagonal elements) are also shown in Table 5.1.

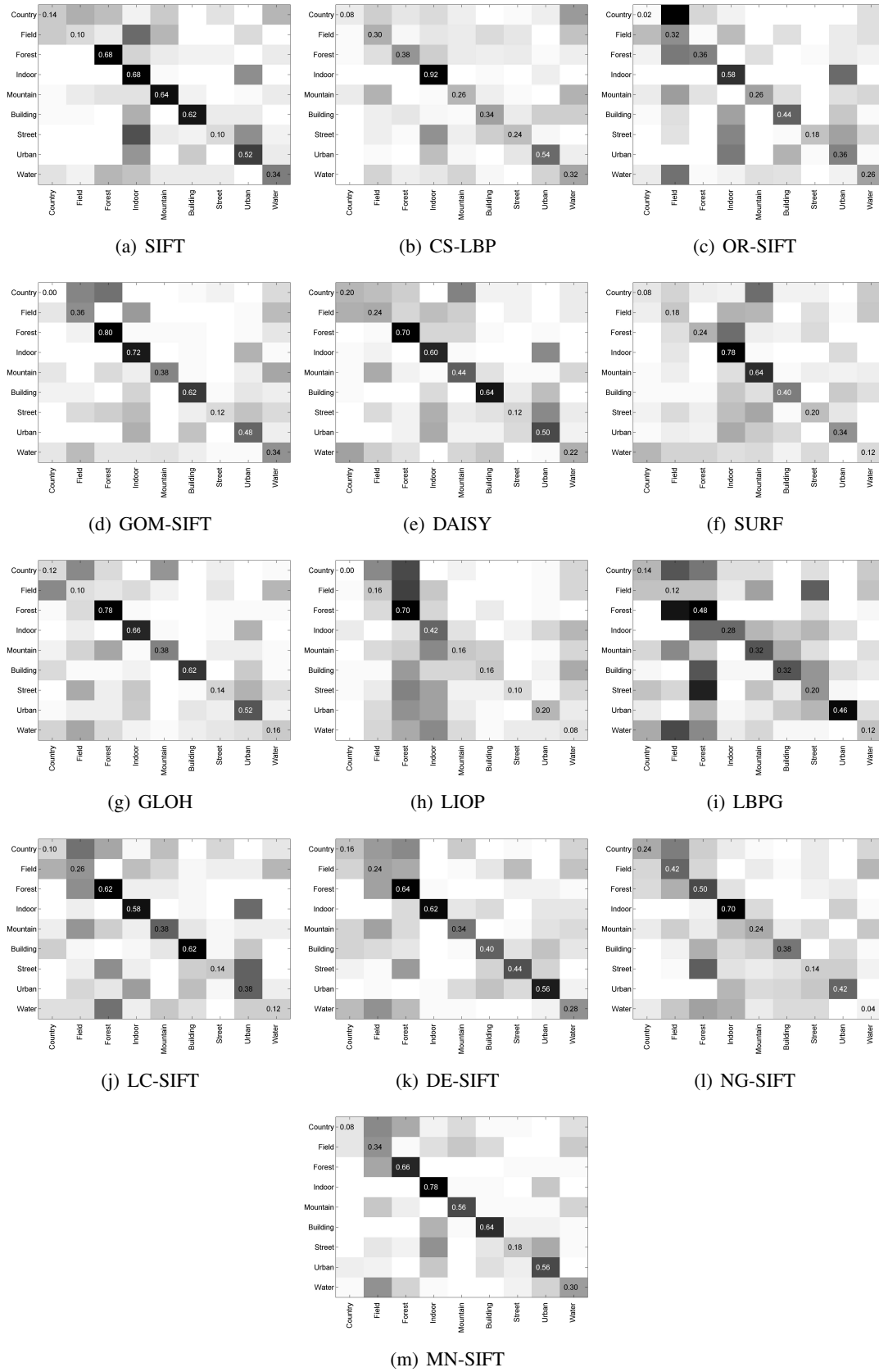


Figure 5.4: Confusion table shows average recognition rates achieved by the local descriptors in nine different scene categories of the RB–NIR dataset when testing and training are performed on the images of the Red and NIR bands, respectively. The average recognition rates (diagonal elements) are also shown in Table 5.2.

Optical Character Recognition

This chapter presents a new method for the recognition of Glagolitic characters. The method proposed is designed for historical documents, named *Missale Sinaiticum* [36], where the Glagolitic character recognition based on the state of the art methods is hard to achieve due to high degradation and background noise [10, 15]. These degradations are due to the parchment aging, environmental effects, fading ink, bleed through, folding marks, and stains [36]. The method proposed deals with such degradations by making use of Dense SIFT algorithm [63] and the nearest neighbor distance maps [50]. The maps encode the Euclidean distances between the Dense SIFT descriptors of the test documents and the SIFT descriptors of the training set. This creates local minima in the distance maps, which help in the localization and recognition of the Glagolitic characters. To evaluate the performance of the method proposed, experiments on three datasets are performed and the performance is compared with that of the SIFT based method proposed by Diem et al. [10] for the similar documents. Additionally, the descriptors proposed (LBPG, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) are used instead of Dense SIFT in the method proposed to compare their Glagolitic character recognition rates with that of Dense SIFT, CS-LBP, OR-SIFT, GOM-SIFT, and SURF descriptors. Note that, Glagolitic Character recognition is an image to database matching task, which is performed in this chapter to evaluate the performance of the descriptors proposed.

6.1 Challenges in Optical Character Recognition

Optical Character Recognition (OCR) is a long standing problem in document image analysis [43, 66]. It deals with the conversion of document images into machine readable format by recognizing the identity of each and every document image character [43]. The conversion process generally takes place in a sequence of steps, which includes text line detection, character segmentation, feature calculation, and classification [30]. The complexity and accuracy of each step depends on the document type [43], for instance, on handwritten documents the complexity of each step is normally high and the accuracy is low than machine-printed documents [16, 66].

In the case of historical documents, such as shown in Figure 6.1(a), the complexity of each step grows and the accuracy degrades, mainly due to background noise, degradation, stains, and fading ink [10].

This chapter concerns with the recognition of Glagolitic characters in Missale Sinaiticum (Cod. Sin. Slav. 5N) manuscripts, which were written in the 11th century and discovered in 1975 at St. Catherine's Monastery. An image portion belonging to Missale Sinaiticum manuscripts, is shown in Figure 6.1(a), where background noise and degradations can be seen. Such degradations affect the localization (i.e., detection and segmentation) and recognition of the Glagolitic characters. To deal with these problems, a new and robust method for the recognition of Glagolitic character is proposed.

Several methods for handwritten character recognition [9, 10, 62] have been proposed, which can be used for the Glagolitic character recognition. However, these methods differ in implementation and also depend on the document type, amount of background noise, and the nature of degradation in the documents. Some methods use text line detection in order to localize the characters and then, isolate the characters of the text line to compute the features on isolated characters for the character recognition [62]. These methods perform well if document images possess homogenous background and are free from background noise [10]. Some methods use local image features to localize and then, recognize the characters. They do not use text line detection as a preprocessing step [10].

In the case of Missale Sinaiticum manuscripts, text line detection and character isolation are hard tasks, which make the local image features based methods more suitable. Another reason, which makes the local image features based methods more suitable, is image binarization used in the text line detection [62]. In the case of Missale Sinaiticum, image binarization is also a challenging problem, as shown in Figure 6.1. If Otsu's image binarization [45] is used, it results in a binarization, which distorts the shape of the majority of the characters, as shown in Figure 6.1(b), mainly due to background noise and fading ink. Such a binarization further complicates the text line detection and makes the isolation of the text line characters difficult [10, 26]. The state-of-the-art methods for image binarization can be used, such as [56] and [57], whose results are shown in Figure 6.1(c) and Figure 6.1(d), respectively. But their results also suggest that achieving a perfect image binarization of the Missale Sinaiticum manuscripts is difficult. If we achieve a perfect image binarization and text line detection, then the isolation of the text line character still remains a challenging task. Here, isolated characters are required for the feature computation, so that such features can be matched with that of isolated training set characters to recognize the identity of the manuscripts' characters. But the image portion in Figure 6.1(a) shows that Glagolitic characters in the manuscripts are overlapping and there is no fixed spacing between them, which make the isolation of the Glagolitic characters difficult.

Diem et al. [10] show that by using local image features a better character recognition can be achieved on Missale Sinaiticum manuscripts. To this end, they use SIFT algorithm [33] and cluster the SIFT keypoints of the manuscripts to localize the characters. Then, by using a SVM classifier on the descriptors of the clustered SIFT keypoints; they recognize the characters. They obtain good results with regard to background noise and fading ink. Similarly, Uchida et al. use SURF algorithm [2] and recognize the handwritten characters [61]. However, they use a different approach than [10], i.e., by decomposing the characters into SURF patterns and then

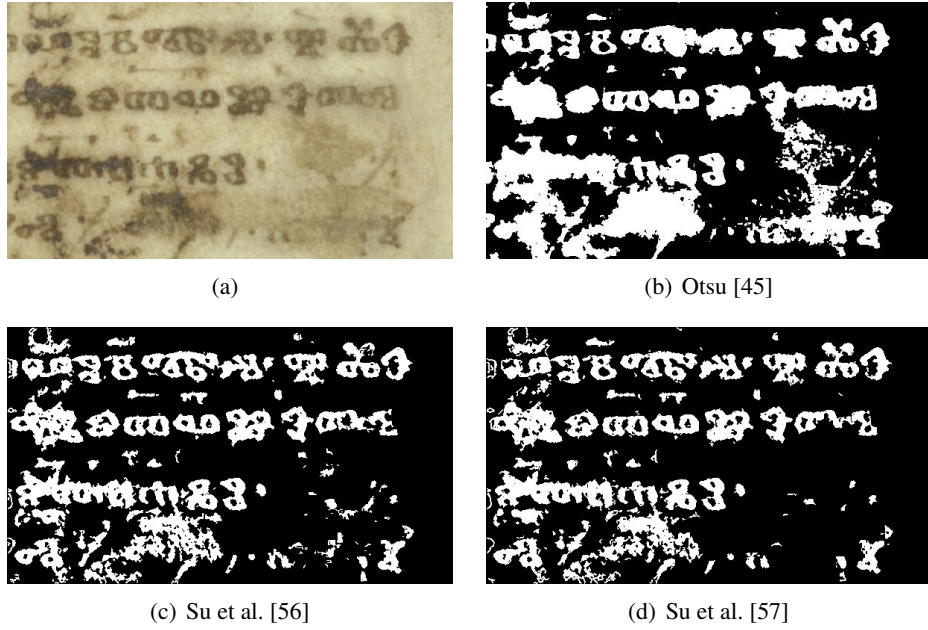


Figure 6.1: (a) A sample belonging to Missale Sinaiticum [36] manuscript is shown. The other images are the binarized versions of (a), which are obtained by applying (b) Otsu's image binarization [45] (c) local maximum and minimum approach [56] and (d) the combined approach for the text binarization [57].

recognizing the patterns with a nearest neighbor classifier [61].

The method proposed is also a local image feature based method. It is a text line detection free method. It does not require image binarization as a pre-processing step. It uses Dense SIFT (DSIFT) algorithm [63] to overcome the effects of degradations in Missale Sinaiticum manuscripts, which affect the detection and repeatability of the SIFT keypoints [33], and result in low Glagolitic character recognition rates [10]. DSIFT efficiently overcomes such degradations and obtains better recognition rates than [10]. In the next sections, the method proposed is described in detail.

6.2 Proposed Method

The method proposed is illustrated in Figure 6.2. First, it computes DSIFT descriptors on a test document image and then compares such descriptors with the SIFT descriptors of the training set. It uses a nearest neighbor based descriptor matching strategy and compares the descriptors with the Euclidean distance. This results in Nearest Neighbor Distance Maps (NNDM), which identify probable locations of the Glagolitic characters in the test document image and help in the localization as well as the recognition of the Glagolitic characters. In the following subsections each block of the flowchart is described in detail.

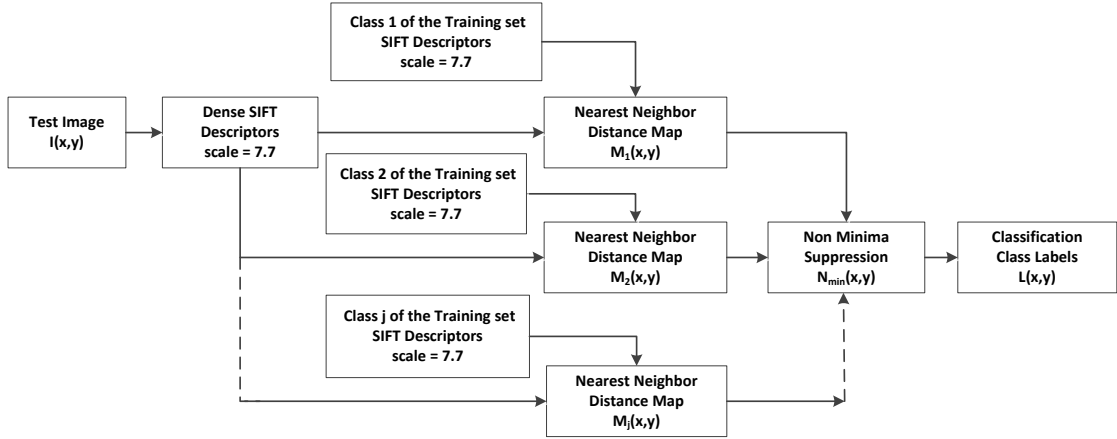


Figure 6.2: Illustration of the method proposed for the recognition of Glagolitic characters in degraded document images.

6.2.1 Test Image

A test document image $I(x, y)$ is shown in Figure 6.1(a), where (x, y) represents a pixel location. This image is used in the following subsections to explain the building blocks of the flowchart.

6.2.2 Dense SIFT Descriptors

To compute DSIFT descriptors, the SIFT descriptor algorithm [63] is applied at a fixed scale value of $\sigma = 7.7$ at each pixel of the test image $I(x, y)$. This scale value is based on the average size of the single Glagolitic character images in the training set. The average width and height of the training set character images are 65 and 51 pixels, respectively. For the estimation of scale, the average width is used and the scale is estimated as $\sigma = 65 / (2 \times \sqrt{2} \times 3)$.

6.2.3 Training SIFT Descriptors

To compute training SIFT descriptors, the SIFT algorithm [63] is applied at the same scale value of $\sigma = 7.7$ but only at the center pixel of each single Glagolitic character image of the training set. To further elaborate on this, three samples from classes 1, 5, and 20 of the training set are shown in Figure 6.3, where a '+' sign on each sample depicts the center pixel. Training set consists of manually segmented single Glagolitic character images. The width and height of such segmented character images vary within Glagolitic character classes and also between the classes. Therefore, the average width is used for the construction of both training and testing SIFT descriptors to cope with the size changes.

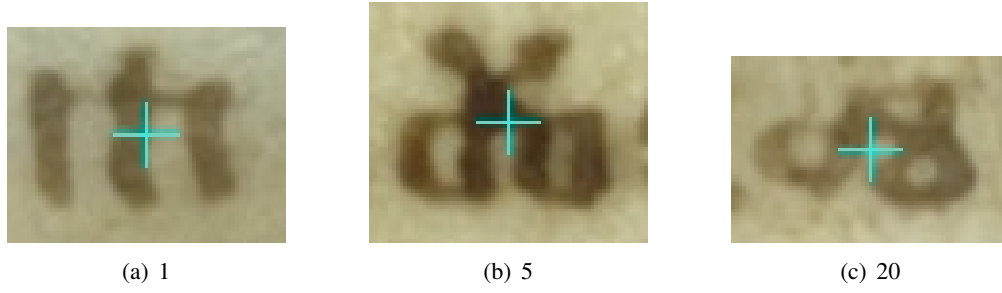


Figure 6.3: Samples from Glagolitic character classes (a) 1, (b) 5, and (c) 20 of the training set. The center pixel of each sample is depicted with a '+' sign, which is used in the method proposed to compute a SIFT descriptor on the sample at a fixed scale value of $\sigma = 7.7$

6.2.4 Nearest Neighbor Distance Map

The number of evaluated Glagolitic character classes is 25 in this thesis. To localize and recognize the characters of these classes, Nearest Neighbor Distance Maps (NNDM) $M_j(x, y)$ are computed as follows:

$$M_j(x, y) = \|d_{x,y} - d_{nn}^j\|; \quad j \in \{1, 2, \dots, 25\} \quad (6.1)$$

where j represents the j^{th} Glagolitic character class of the training set, $d_{x,y}$ is a DSIFT descriptor of $I(x, y)$ at a pixel location (x, y) , d_{nn}^j is the nearest neighbor of $d_{x,y}$ in the j^{th} class of the training set, and $\|\cdot\|$ represents the Euclidean distance between $d_{x,y}$ and d_{nn}^j .

For each pixel of $M_j(x, y)$, a value is computed (see Equation 6.1), which is in fact, an Euclidean distance between a DSIFT descriptor $d_{x,y}$ and its nearest neighbor in the j^{th} class of the training set. Such distances help in the localization as well as the recognition of the characters. To further elaborate on this, distance maps $M_1(x, y)$, $M_5(x, y)$ and $M_{20}(x, y)$ are shown in Figure 6.4. A Glagolitic character image is also shown on the top left corner of the distance map to interpret the map. Dark spots in each distance map can be seen, which are local minima and suggest probable locations of the Glagolitic characters in $I(x, y)$. For instance, a Glagolitic character of class 5 (Figure 6.3(b)) occurs near the top right corner of $I(x, y)$; therefore, a dark spot can be seen at the same location in the $M_5(x, y)$ map, which is shown inside a bounding box in Figure 6.4(c). Similarly, other dark spots in $M_1(x, y)$, $M_5(x, y)$ and $M_{20}(x, y)$ maps show that Glagolitic characters similar to classes 1, 5, and 20 are also present in $I(x, y)$, respectively.

6.2.5 Non-Minima Suppression

To localize the Glagolitic characters in the test image, an image $N(x, y)$ is created. This image is based on NNDM and created as follows by computing a minimum value for each pixel of $N(x, y)$:

$$N(x, y) = M_{j'}(x, y); \quad j' = \arg \min_{j \in \{1, 2, \dots, 25\}} M_j(x, y) \quad (6.2)$$

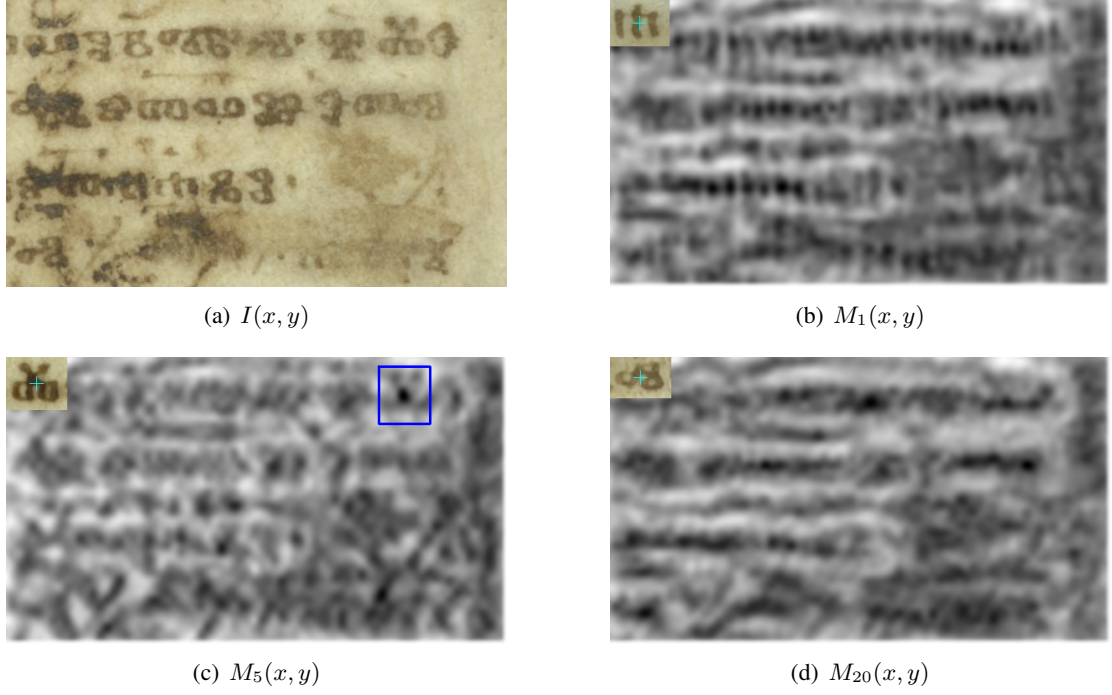


Figure 6.4: Nearest Neighbor Distance Maps $M_1(x, y)$, $M_5(x, y)$ and $M_{20}(x, y)$, which show the probable locations of Glagolitic characters in the test image similar to Glagolitic characters in classes 1, 5, and 20 of the training set, respectively.

Afterwards, a Non-Minima Suppression is applied on $N(x, y)$ and an image $N_{min}(x, y)$ is obtained as:

$$N_{min}(x, y) = \begin{cases} 1 & N(x, y) = \min(\{N(x_w) : x_w \in \mathbb{W}\}) \\ 0 & \text{otherwise} \end{cases} \quad (6.3)$$

where $x_w = (x_w, y_w)$ is a pixel location that lies inside a circular window $\mathbb{W}$ of radius $\lfloor 65/2 + 0.5 \rfloor$ pixels centered at (x, y) . This window is slid over all the pixel locations of $N(x, y)$ and an image $N_{min}(x, y)$ is obtained. For each slide, a local minimum is computed and the pixel (x, y) of the image N_{min} is set to 1 if the window center, i.e., (x, y) matches with the pixel location of the local minimum, otherwise zero is assigned. Figure 6.5 shows the results for non-minima suppression, where the pixel locations correspond to $N_{min}(x, y) = 1$ are depicted with blue dots. It can be seen that, for each Glagolitic character at least one local minimum is obtained. This local minimum is used in the subsequent step to recognize the Glagolitic character.

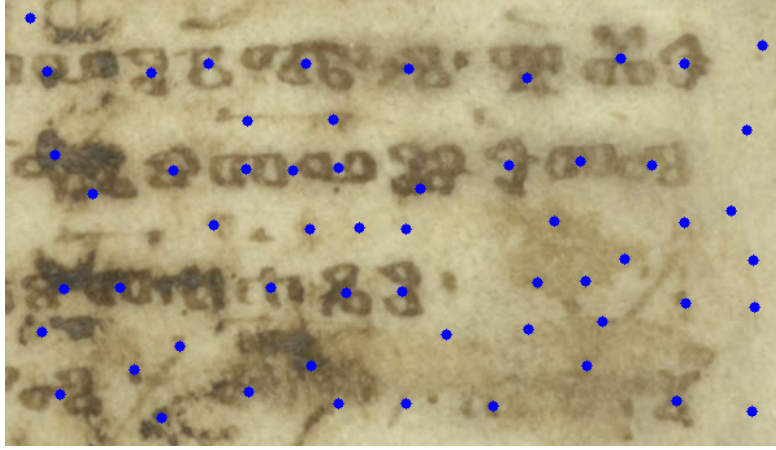


Figure 6.5: Blue dots depict the pixel locations of the local minima. They are obtained by applying a non-minima suppression on the distance maps of the test image.

6.2.6 Classification

Having localized the characters (i.e., local minima), image $N_{min}(x, y)$ is processed as follows to obtain a class labeled image $L(x, y)$:

$$L(x, y) = \begin{cases} 0 & N_{min}(x, y) = 0 \\ j' & j' = \arg \min_{j \in \{1, 2, \dots, 25\}} M_j(x, y) \end{cases} \quad (6.4)$$

For the assignment of class labels, each pixel corresponds to $N_{min}(x, y) = 0$ is assigned zero or a background class label, whereas a class label j is obtained for $N_{min}(x, y) = 1$ by using the NNDM, as described in Equation 6.4.

Figure 6.6(a) shows a result for Glagolitic character recognition based on method proposed. The result shows correctly classified characters in green. The red colored characters with red markers are misclassified characters, and the red colored characters without markers are not retrieved characters. White markers are not eligible; they represent either background or characters of unevaluated Glagolitic classes (refer to Section 6.2.4). It can be seen that method proposed obtains a better character recognition than [10], whose result is shown in Figure 6.6(b).

6.3 Experimental Results

This section presents the experimental setup and results for Glagolitic character recognition. Experiments on three different image datasets (of Missale Sinaiticum manuscripts) [10] are performed. In the following subsections, the experimental results are described in detail.

6.3.1 Experiments on Single Character Images

This section presents the experimental results for the recognition of single Glagolitic character images. Two datasets are used, which are called Dataset-1 and Dataset-2.

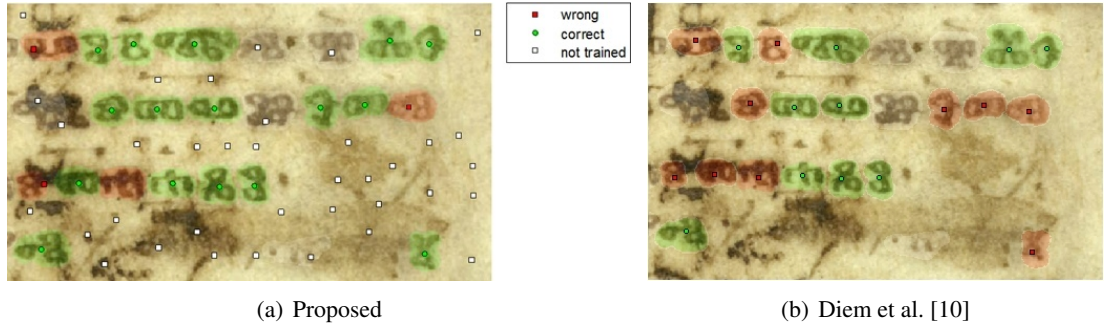


Figure 6.6: Glagolitic character recognition based on (a) method proposed and (b) the method suggested by Diem et al. [10]. The red colored characters without markers are not retrieved characters, whereas characters with green and red markers are correctly and misclassified characters, respectively. White markers are not eligible; they represent either background or the characters of the unevaluated Glagolitic classes.

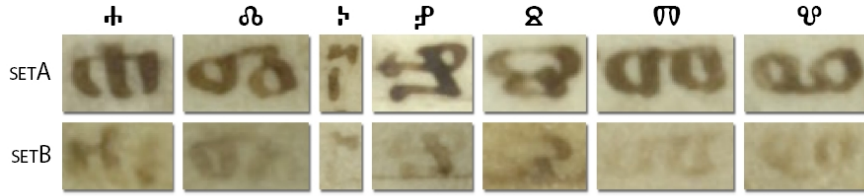


Figure 6.7: Samples from evaluated Dataset-1. Images from [10].

Dataset-1

Dataset-1 consists of two sets; SETA and SETB. A subset of samples from SETA and SETB is shown in Figure 6.7, and the description of SETA and SETB is given in Table 6.1. SETA consists of 10 Glagolitic character classes with 10–12 single Glagolitic character images per class (in total 107 characters). SETB also consists of the same 10 Glagolitic character classes, but there are 9 single Glagolitic character images per class, which are either degraded or partially visible than the ones in SETA, as shown in Figure 6.7. SETB is used as a test set to evaluate the performance of method proposed on degraded and faded characters by using the SIFT descriptors of SETA.

On each training sample of SETA, only one SIFT descriptor is computed at a fixed scale of $\sigma = 7.7$, as described in Section 6.2.3, whereas on each test sample of SETB, DSIFT descriptors are computed at the same scale value of $\sigma = 7.7$. SIFT descriptors of the SETA are then matched with the DSIFT descriptors of each test sample of the SETB to create NNDM for the test sample. The number of Glagolitic character classes in SETA is 10, therefore, $j = \{1, 2, \dots, 10\}$ is used in Equation 6.1. Then, non-minima suppression is performed and a class label for the test sample is obtained, as described in Equation 6.4.

To compare the performance of method proposed with that of the method proposed by Diem

Table 6.1: Dataset-1 and Dataset-2: type, number of samples and number of classes.

		Type	number of samples	number of classes
Dataset 1	SETA	training	107	10
	SETB	test	90	10
Dataset 2	SETC	training	500	25
	SETD	test	198	25

Table 6.2: Precision scores achieved on Dataset-1 and Dataset-2.

	Diem et al. [10]	Proposed (DSIFT)
Dataset-1	0.789	0.822
Dataset-2	0.712	0.751

et al. [10], precision score is computed, which is defined as the ratio of correctly classified characters to total number of characters. Table 6.2 shows the precision scores. It can be seen that the method proposed demonstrates the best performance. It obtains a precision score of 0.822 and outperforms [10] by 3.3%.

Dataset-2

Dataset-2 also consists of two sets: SETC and SETD. The description of each set is given in Table 6.1. Dataset-2 is, in fact, an extended version of Dataset-1. This extension is obtained by increasing the number of Glagolitic character classes from 10 to 25, with 500 and 198 samples in SETC and SETD, respectively. SETC consists of single Glagolitic character images similar to the ones in SETA, whereas SETD consists of degraded and faded characters similar to the ones in SETB. To compute the training SIFT descriptors, SETC is used and the class labels for the test samples in SETD is obtained. Table 6.2 shows the precision scores achieved on Dataset-2 (SETC–SETD). A performance decrease can be seen for each method, compared to the precision scores achieved by them on Dataset-1, inspite of that, method proposed obtains a precision score of 0.751, and outperforms [10] by 3.9%.

6.3.2 Experiments on Document Images

To perform experiments on document images, Dataset-3 is used. This dataset consists of 15 different image portions. An image portion belonging to Dataset-3 is shown in Figure 6.1(a). The number of Glagolitic characters in these image portions is 1055, which are divided into two sets for the performance evaluation: (i) normal, and (ii) degraded. Normal set consists of 913 single Glagolitic character images similar to the ones in SETA and SETC, whereas the degraded set consists of 142 character images similar to the ones in SETB, SETD, and Figure 6.1(a). To recognize the Glagolitic characters in the image portions of Dataset-3, SETC is used as a training set and for performance comparison Recall, Precision, and F-measure are computed, which are

Table 6.3: Results achieved on Dataset-3 (15 image portions: totally 913 normal and 142 degraded Glagolitic characters).

Method	Recall		Precision		F-measure	
	Normal	Degraded	Normal	Degraded	Normal	Degraded
Diem et al. [10]	0.732	0.296	0.862	0.539	0.792	0.382
Proposed (DSIFT)	0.811	0.478	0.962	0.837	0.879	0.609

defined as follows [10]:

$$Recall = \frac{TP}{TP + FN} \quad (6.5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6.6)$$

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (6.7)$$

where TP represents True Positives (class labels that correspond with the ground truth), FP stands for False Positives (correctly located values with false class labels), and FN denotes False Negatives (ground truth values that are not detected by the system). The precision measure represents the fraction of correctly classified characters in the retrieved characters, whereas recall represents the ratio of retrieved characters to those present in the image portion. F-measure represents a weighted average of precision and recall.

Table 6.3 shows the results achieved on Dataset-3. The experimental results show that method proposed obtains recall of 0.811 and 0.478 on normal and degraded sets, which are 7.9% and 18.2% higher than [10], respectively. Similarly, the precision scores of the method proposed show 10.0% and 29.8% improvement over [10] on normal and degraded sets, respectively. These results show that the method proposed is more robust to degradation and background noise in the Missale Sinaiticum manuscripts than the SIFT based method of Diem et al. [10].

6.4 OCR Using Descriptors Proposed

This section presents the experimental results when descriptors proposed are used instead of SIFT in the method proposed for the recognition of Glagolitic character. As discussed, the Missale Sinaiticum manuscripts possess background noise, stain, and fading ink, which induce intensity changes between the images of the training and test sets (Figure 6.7). Such intensity changes affect the SIFT descriptor, as shown in Figure 6.8, where two images of the same Glagolitic character are shown, but their SIFT descriptors look quite different, due to fading ink and background noise. The elements of the SIFT descriptors are shown as line segments inside 4×4 location bins [65]. To overcome the effects of such intensity changes, descriptors proposed (LBP, LC-SIFT, DE-SIFT, NG-SIFT and MN-SIFT) are used instead of SIFT in the method proposed to achieve a better recognition of Glagolitic characters.

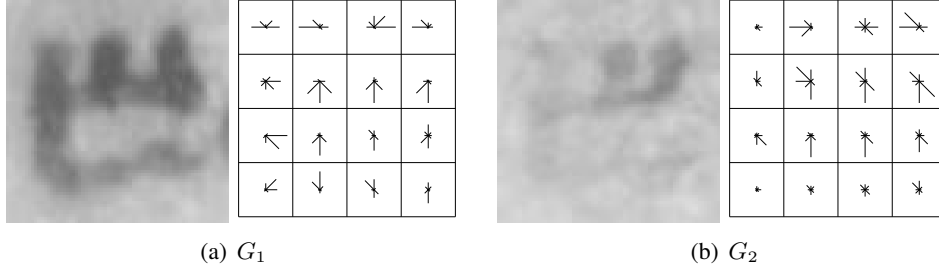


Figure 6.8: Two images of the same Glagolitic character are shown, where intensity changes between the images can be seen, which affect the SIFT descriptor. The elements of the SIFT descriptors are shown as line segments inside 4×4 location bins [65].

6.4.1 Experiments on Single Character Images

To evaluate the performance of descriptors proposed, first, experiments on single Glagolitic character images are performed. Two datasets are used, which are: Dataset-1 and Dataset-2 (refer to Table 6.1). DSIFT (i.e., the proposed method with SIFT descriptor) is used as a reference to evaluate the performance of the descriptors proposed. The precision scores achieved on these datasets are shown in Table 6.4. Note that, dense descriptor construction, as described in Section 6.2.2, is used for the construction local descriptors including the descriptors proposed to recognize the Glagolitic characters of Dataset-1 and Dataset-2.

The experimental results show that MN-SIFT demonstrates the best precision scores. It obtains a precision score of 0.889 on Dataset-1 and outperforms all the other local descriptors. It obtains 6.7% better precision scores than DSIFT. This shows its more robustness to intensity changes than DSIFT. Similarly, NG-SIFT, LC-SIFT, and DE-SIFT also achieves 2.20%, 1.10%, and 2.20% better precision scores than DSIFT, respectively. Only LBPG demonstrates 18% lower precision score than DSIFT. This is due to the CS-LBP scheme, which is sensitive to image noise and is used twice in the LBPG method [51]. Other local descriptors, like CS-LBP, OR-SIFT, and GOM-SIFT, demonstrate lower precision scores than DSIFT.

Table 6.4 shows the precision scores achieved on Dataset-2. The results show lower precision scores for all the local descriptors compared to Dataset-1. Only MN-SIFT obtains the best precision score of 0.801 on Dataset-2 and outperforms DSIFT by 5%. All other local descriptors demonstrate lower precision scores than DSIFT.

6.4.2 Experiments on Document Images

To evaluate the performance of descriptors proposed on document images, Dataset-3 is used (see Section 6.3.2) and the performance of the descriptors proposed is compared with that of DSIFT. Dense descriptor construction, as described in Section 6.2.2, is used for all the local descriptors to recognize the Glagolitic characters of Dataset-3. Table 6.5 shows the results. The best performance by MN-SIFT can be seen; it obtains the best results on normal as well as degraded characters. LC-SIFT also demonstrate better performance than DSIFT. NG-SIFT and DE-SIFT demonstrate better recall and f-measure scores than DSIFT only on degraded char-

Table 6.4: Precision scores achieved on Dataset-1 and Dataset-2 when descriptors proposed are used.

Method	Dataset-1	Dataset-2
Diem et. al. [10] (SIFT)	0.789	0.712
DSIFT	0.822	0.751
Dense CS-LBP	0.789	0.731
Dense OR-SIFT	0.722	0.617
Dense GOM-SIFT	0.789	0.647
Dense SURF	0.744	0.647
Dense GLOH	0.567	0.493
Dense LBPG	0.633	0.483
Dense LC-SIFT	0.833	0.726
Dense DE-SIFT	0.844	0.741
Dense NG-SIFT	0.844	0.711
Dense MN-SIFT	0.889	0.801

Table 6.5: Results achieved on Dataset-3 (15 image portions: totally 913 normal and 142 degraded Glagolitic characters) when descriptors proposed are used.

Method	Recall		Precision		F-measure	
	Normal	Degraded	Normal	Degraded	Normal	Degraded
Diem et al. [10] (SIFT)	0.732	0.296	0.862	0.539	0.792	0.382
DSIFT	0.811	0.478	0.962	0.837	0.879	0.609
Dense CS-LBP	0.760	0.466	0.946	0.806	0.842	0.591
Dense OR-SIFT	0.833	0.432	0.954	0.805	0.890	0.562
Dense GOM-SIFT	0.829	0.429	0.942	0.719	0.882	0.537
Dense SURF	0.423	0.224	0.808	0.610	0.555	0.327
Dense GLOH	0.440	0.348	0.672	0.571	0.531	0.432
Dense LC-SIFT	0.879	0.634	0.962	0.857	0.919	0.729
Dense DE-SIFT	0.789	0.509	0.926	0.788	0.852	0.619
Dense LBPG	0.653	0.275	0.937	0.772	0.769	0.406
Dense NG-SIFT	0.804	0.503	0.949	0.810	0.870	0.621
Dense MN-SIFT	0.893	0.652	0.972	0.890	0.931	0.753

acters. However, LBPG gives lower results than DSIFT. The gradient orientation modification based descriptors, i.e., GOM-SIFT [71] and OR-SIFT [68] also obtain better results than DSIFT.

6.5 Summary

In this chapter a new method for the recognition of Glagolitic characters were presented. It is designed for highly degraded Missale Sinaiticum manuscripts. It is based on Dense SIFT algorithm and uses nearest neighbor descriptor matching strategy to localize and recognize the

Glagolitic characters. It is a simple and scalable method. It does not require feature learning as a pre-processing step for the recognition of characters. It is also a binarization free method and does not require the text line detection as pre-processing step to localize the characters in the document images.

The experimental results show better performance of method proposed on single Glagolitic character images than a SIFT based method, which has been used for Missale Sinaiticum. The results show better performance on normal as well as degraded and faded out characters. The results also show that the method proposed efficiently localizes and recognizes the Glagolitic characters under background noise, stains, and fading ink.

The chapter also presented the results for Glagolitic character recognition when descriptors proposed are used instead of Dense SIFT in the method proposed. The idea is to overcome the effects of intensity changes, which occur between the images of the Missale Sinaiticum manuscripts due to degradation. The experimental results show that LC-SIFT, DE-SIFT, NG-SIFT and MN-SIFT obtain better performance, whereas LBPB demonstrate lower performance than Dense SIFT.

Conclusion

In this thesis new local descriptors are proposed to overcome the effects of nonlinear intensity changes in image-to-image and image-to-database matching tasks. The new descriptors are based on novel image features, which are obtained by combining the strengths of image gradients, DE, LC, and LBP. These features are used in the SIFT like feature histogram schemes and five new local descriptors are computed, which are: LBP-G, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT.

In the case of image-to-image matching, experiments on pairs of images are performed. Homographies between the pairs of images are used as ground truth data to count the number of correct descriptor matches in order to compare the performance of new descriptors with that of state-of-the-art local descriptors (SIFT, GLOH, CS-LBP, SURF, GOM-SIFT, OR-SIFT, DAISY, and LIOP). In the case of image-to-database matching, a nearest neighbor based descriptor matching strategy is used to obtain the recognition rates for two different recognition tasks, which are: SCR and OCR. The experimental results show that the new descriptors obtain a better performance than the state-of-the-art local descriptors in image-to-image matching, SCR, and OCR.

In the case of image-to-image matching, four different multispectral image datasets (CAVE, RWHI, RGB-NIR, and MSD) are used. All the experiments are performed under nonlinear intensity changes, i.e., image matching between the pairs of different band images of the same scene. The experimental results show that OR-SIFT and GOM-SIFT, which have been used against nonlinear intensity changes, demonstrate lower performance than SIFT on CAVE, RWHI, and RGB-NIR datasets. Their performances are comparable to SIFT only on MSD dataset. Both OR-SIFT and GOM-SIFT are extended versions of SIFT and are based on gradient orientation modifications to overcome the effects of nonlinear intensity changes. The new descriptors, which are based on gradient magnitude modification, demonstrate a better performance than OR-SIFT, GOM-SIFT, and SIFT, on all the evaluated datasets.

This shows that the modifications of gradient magnitudes, which are proposed for the construction of new local descriptors, are more robust to nonlinear intensity changes than gradient orientation modifications. The experimental results show that CS-LBP, which has been used

to overcome the effects of monotonic intensity changes, also demonstrates good performance under nonlinear intensity changes. However LBPG, which is a gradient extension of CS-LBP, demonstrates even better performance than CS-LBP on all the evaluated datasets.

LIOP demonstrates the worst performance in image-to-image matching. LIOP have been used against monotonic intensity changes, but the experimental results show that it performs worst on all the evaluated datasets due to nonlinear intensity changes. This shows that the intensity based description vectors that LIOP computes are less suitable and less robust against nonlinear intensity changes. GLOH, SURF, and DAISY in image-to-image matching perform comparable to SIFT. But compared to the new descriptors, they demonstrate inferior performance. NG-SIFT and MN-SIFT obtain the best results in image-to-image matching. They are based on normalized image gradients, which shows that such gradients are more robust towards nonlinear intensity changes than image gradients, CS-LBP features, and Haar wavelet responses, which are used in the construction of SIFT, CS-LBP, and SURF descriptors, respectively.

Experiments in the case of SCR, are performed on a RGB–NIR scene dataset. All the experiments are performed under nonlinear intensity changes by using images of different bands for training and testing. The scene recognition rates are computed and the performance of the new descriptors are compared with that of the state-of-the-art local descriptors. The experimental results show that LC-SIFT and MN-SIFT perform comparable to SIFT, GOM-SIFT, DAISY, and GLOH, when Blue and NIR band images are used for testing and training, respectively. However, NG-SIFT, DE-SIFT, and LBPG demonstrate lower scene recognition rates. The experimental results show that the performance of DE-SIFT and MN-SIFT is comparable to SIFT, GOM-SIFT, and DAISY when testing and training are performed on Red and NIR band images, respectively, whereas NG-SIFT, LC-SIFT, and LBPG demonstrate lower scene recognition rates.

MN-SIFT demonstrates the best SCR performance, whereas LIOP obtains the worst performance. Compared to LIOP, all the gradient based descriptors, i.e., LC-SIFT, MN-SIFT, DE-SIFT, SIFT, GLOH, GOM-SIFT, OR-SIFT, and DAISY obtain better performances. This shows that the intensity based description vectors that LIOP computes, are less suitable for the scene recognition task under nonlinear intensity changes. Additionally, the experimental results show that all the local descriptors obtain better scene recognition rates when Red and NIR band images are used. This is mainly due to low nonlinear intensity changes between the images of Red and NIR bands than Blue and NIR band images.

In the case of OCR, a gray scale image dataset of Missale Sinaiticum manuscripts is used. This dataset consists of historical documents written in the 11th century in Glagolitic. But due to parchment aging, fading ink, and environmental effects the documents are in degraded conditions, which aggravates the recognition of Glagolitic characters. To deal with such problems, a new method is proposed, which is based on Dense SIFT algorithm and uses a nearest neighbor based descriptor matching strategy to localize and recognize the Glagolitic characters in the documents. The method proposed is a simple and scalable method. It does not require feature learning for character recognition. It is also a binarization free method and does not require text line detection as a pre-processing step to localize the characters.

The experimental results show that the method proposed obtains a better Glagolitic character recognition than the method proposed by Diem et al. [10], which has been used for solving the similar problem. Consistent improvement in the performance is achieved on normal as well as

degraded characters. The localization of the characters is shown under background noise, stains, and fading ink. Even better performance is achieved when the new descriptors are used instead of Dense SIFT in the method proposed. This is due to intensity changes (degradations) between the images of the Missale Sinaiticum manuscripts, which affect the performance of Dense SIFT. The experimental results show that LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT demonstrate better performance, whereas LBPG obtains a lower performance than Dense SIFT.

7.1 Summary of Contributions

The contributions of this thesis include,

- A performance evaluation of eight different state-of-the-art local descriptors (SIFT, CS-LBP, SURF, GOM-SIFT, OR-SIFT, DAISY, and LIOP) under nonlinear intensity changes.
- Five new local descriptors (LBPG, LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT) are proposed in order to improve the performance of the local descriptors under nonlinear intensity changes.
- LBPG features are proposed, which are based on image gradients and CS-LBP features in order to construct descriptors robust to nonlinear intensity changes.
- Four different modifications to gradient magnitudes are proposed, i.e., LC, DE, NG, and MN. These modifications boost the contribution of local edges in the SIFT method, and result in new descriptors i.e., LC-SIFT, DE-SIFT, NG-SIFT, and MN-SIFT. These new descriptors demonstrate better matching performance than gradient magnitudes, CS-LBP features, and Haar wavelet responses, which are used in the construction of SIFT, CS-LBP, and SURF descriptors, respectively.
- A simple SIFT like feature histogram scheme is proposed, which is used for the construction of NG-SIFT and MN-SIFT descriptors in order to achieve better performance against nonlinear intensity changes.
- A new method for the recognition of Glagolitic characters is proposed.

7.2 Future Research

Three potential research topics that can be built upon the work presented in this thesis are described in the following sections.

7.2.1 Feature Point Detection

The state-of-the-art feature point detectors, such as Harris Laplace and Hessian Laplace use Harris and Hessian functions, respectively, for the detection of feature points. These functions are based on image gradients, which are used in order to identify the image locations that have large image gradients in all directions for the detection of feature points. This thesis shows that

due to nonlinear intensity changes, the image gradients vary. It is necessary to investigate if the proposed gradient modifications will make the Harris and Hessian functions invariant to such nonlinear intensity changes. This will increase the repeatability of the feature points and also result in better correspondences between the feature points under nonlinear intensity changes.

7.2.2 Image Registration

Multispectral imaging provides higher spectral resolution than typical RGB imaging. It captures images in discrete wavelength bands. But due to scene motion, such imaging induces misalignment between different band images of the same scene. To deal with this problem, image registration has been used. The performance of the methods used for image registration depends on local descriptors, which are used to establish the feature point correspondences, in order to estimate the unknown homographies between the pairs of misaligned images. This thesis shows that if wavelength differences between the wavelength band images are small, low nonlinear intensity changes occur and the new descriptors perform comparable to the state-of-the-art local descriptors. But under high wavelength differences, the new descriptors demonstrate better performance than the state-of-the-art local descriptors. Further investigation is required for multi-modal images, which possess even higher nonlinear intensity changes and make the image matching and image registration more challenging and difficult based on local descriptors.

7.2.3 Document Image Analysis

This thesis presents a new method for the recognition of isolated ancient characters. It is an image binarization free method, does not require text line detection and bounding box fitting around the characters for recognition. It shows promising results under intensity changes, background noise, fading ink, stain, and folding marks. This method needs to be tested for the recognition of Latin, Greek, Arabic, and Bangla characters in printed as well as handwritten documents. A performance investigation of this method for word spotting and the localization of text in videos, are also required.

List of Symbols

F_h	Image gradient in horizontal direction
F_v	Image gradient in vertical direction
Ω	Gradient Magnitude
β	Gradient Orientation
$\hat{\Omega}$	Normalized Gradient Magnitude
$\hat{\beta}$	Normalized Gradient Orientation
$\bar{\Omega}$	Modified Normalized Gradient Magnitude
ϕ	Restricted Gradient Orientation
$\top$	Vector transpose
dx	Haar wavelet response in horizontal direction
dy	Haar wavelet response in vertical direction
$\sum$	Vector sum
Σ	Standard deviation of a 2D Gaussian kernel
ϵ	Overlap error
v	Infinitely small positive number
(x, y)	Pixel index (cartesian coordinate)
N_a	Number of correspondences
N_c	Number of correct matches
N_f	Number of false matches
$\%$	Percentage

Definition of Terms

SIFT	Scale Invariant Feature Transform
GLOH	Gradient Location Orientation Histogram
SURF	Speeded Up Robust Features
LBP	Local Binary Patterns
CS-LBP	Center Symmetric-Local Binary Patterns
DAISY	Dense local image descriptor
LIOP	Local Intensity Order Pattern
GOM-SIFT	Gradient Orientation Modification - SIFT
OR-SIFT	Orientation Restricted - SIFT
LBPG	Local Binary Pattern of Gradients
LC-SIFT	Local Contrast - SIFT
DE-SIFT	Differential Excitation - SIFT
NG-SIFT	Normalized Gradient-SIFT
MN-SIFT	Modified Normalized gradient-SIFT
DSIFT	Dense SIFT
CAVE	Multispectral Image Dataset
RHWI	Real-World Hyperspectral Image Dataset
MSD	Multi-modal Stereo Dataset
RGB-NIR	Red-Green-Blue and Near Infra-Red bands
VS-LWIR	Visible Spectrum and Long Wave Infra-Red bands
AUC	Area under a Precision–Recall Curve
HarLap	Harris Laplace
PCA	Principle Component Analysis
nm	Nanometer (unit of wavelength)
SCR	Scene Category Recognition
OCR	Optical Character Recognition
NNDM	Nearest Neighbor Distance Map

Bibliography

- [1] Fernando Barrera, Felipe Lumbreras, and Angel D Sappa. Multispectral piecewise planar stereo using Manhattan-world assumption. *Pattern Recognition Letters*, 34(1):52–61, 2013.
- [2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*, pages 404–417, 2006.
- [3] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.
- [4] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [5] M Brown and S Susstrunk. Multi-spectral SIFT for scene category recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 177–184, 2011.
- [6] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European Conference on Computer Vision*, pages 778–792. 2010.
- [7] A. Chakrabarti and T. Zickler. Statistics of real-world hyperspectral images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–200, 2011.
- [8] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikainen, Xilin Chen, and Wen Gao. WLD: a robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, 2010.
- [9] Markus Diem and Robert Sablatnig. Recognition of degraded handwritten characters using local features. In *International Conference on Document Analysis and Recognition*, pages 221–225, 2009.
- [10] Markus Diem and Robert Sablatnig. Are characters objects? In *International Conference on Frontiers in Handwriting Recognition*, pages 565–570, 2010.

- [11] Shimon Edelman, Nathan Intrator, and Tomaso Poggio. Unpublished: Complex cells and object recognition: <http://cogprints.org/561/>. 1997.
- [12] Bin Fan, Fuchao Wu, and Zhanyi Hu. Aggregating gradient distributions into intensity orders: A novel local image descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2384, 2011.
- [13] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:524–531, 2005.
- [14] William T. Freeman and Edward H Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991.
- [15] Angelika Garz and Robert Sablatnig. Multi-scale texture-based text recognition in ancient manuscripts. *International Conference on Virtual Systems and Multimedia*, pages 336–339, 2010.
- [16] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.
- [17] Chris Harris. Geometry from visual motion. In *In Active Vision*, A. Blake and A. Yuille (Eds.), MIT Press, pages 263–284, 1992.
- [18] Chris Harris and Mike Stephens. A combined corner and edge detector. In *In Fourth Alvey Vision Conference, Manchester, UK*, pages 147–151, 1988.
- [19] M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436, 2009.
- [20] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with center-symmetric local binary patterns. In *Computer Vision, Graphics and Image Processing*, pages 58–69. 2006.
- [21] Dana E Ilea and Paul F Whelan. Image segmentation based on the integration of colour texture descriptors - A review. *Pattern Recognition*, 44(10):2479–2501, 2011.
- [22] Luo Juan and Oubong Gwun. A comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing*, 3(4):143–152, 2009.
- [23] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513.
- [24] Ron Kimmel, Cuiping Zhang, Alexander M Bronstein, and Michael M Bronstein. Are MSER features really interesting? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2316–2320, 2011.

- [25] Christoph H Lampert, Matthew B Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [26] Victor Lavrenko, Toni M Rath, and R Manmatha. Holistic word recognition for handwritten historical documents. *International Workshop on Document Image Analysis for Libraries*, pages 278–287, 2004.
- [27] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:2169–2178, 2006.
- [28] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using affine-invariant regions. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–319, 2003.
- [29] Shengcai Liao, Anil K Jain, and Stan Z Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1193–1205, 2013.
- [30] Laurence Likforman-Sulem, Jérôme Darbon, and Elisa H Barney Smith. Enhancement of historical printed document images by combining total variation regularization and non-local means filtering. *Image and vision computing*, 29(5):351–363, 2011.
- [31] Tony Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, 1993.
- [32] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(2):225–270, 1994.
- [33] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [34] David G Lowe. Object recognition from local scale-invariant features. In *international conference on Computer vision*, volume 2, pages 1150–1157, 1999.
- [35] Jiri Matas, Ondrej Chum, Martin Urban, and Tomáš Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.
- [36] H. Miklas, M. Gau, F. Kleber, M. Diem, M. Lettner, M. Vill, R. Sablatnig, M. Schreiner, M. Melcher, and G. Hammerschmid. St. Catherine’s Monastery on Mount Sinai and the Balkan-Slavic Manuscript-Tradition. In *Slovo: Towards a Digital Library of South Slavic Manuscripts*, Boyan Penev Publishing Center, pages 13–36, 2008.
- [37] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. *IEEE International Conference on Computer Vision*, pages 525–531, 2001.

- [38] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
- [39] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.
- [40] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1-2):43–72, 2005.
- [41] H Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. *Tech Report CMU-RI-TR-3, Carnegie-Mellon University, Robotics Institute*, September 1980.
- [42] Ana Cris Murillo, José Jesús Guerrero, and C Sagues. SURF features for efficient robot localization with omnidirectional images. In *IEEE International Conference on Robotics and Automation*, pages 3901–3907, 2007.
- [43] George Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, 2000.
- [44] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.
- [45] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
- [46] Antonio Plaza, Jon Atli Benediktsson, Joseph W Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, et al. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009.
- [47] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, pages 430–443. 2006.
- [48] Sajid Saleem, Abdul Bais, and Robert Sablatnig. A performance evaluation of SIFT and SURF for multispectral image matching. In *Image Analysis and Recognition*, pages 166–173. 2012.
- [49] Sajid Saleem, Abdul Bias, Fabian Hollaus, and Robert Sablatnig. Evaluating local descriptors against nonlinear intensity changes. *under review for: IET Image Processing Journal*, 2014.
- [50] Sajid Saleem, Fabian Hollaus, and Robert Sablatnig. Recognition of degraded ancient characters based on Dense SIFT. *International Conference on Digital Access to textural Cultural Heritage*, pages 15–20, 2014.

- [51] Sajid Saleem and Robert Sablatnig. Interest region description using local binary pattern of gradients. *Scandinavian Conferences on Image Analysis*, pages 468–477, 2013.
- [52] Sajid Saleem and Robert Sablatnig. A modified SIFT descriptor for image matching under spectral variations. *International Conference on Image Analysis and Processing*, pages 652–661, 2013.
- [53] Sajid Saleem and Robert Sablatnig. A robust SIFT descriptor for multispectral images. *IEEE Signal Processing Letters*, 21(4):400–403, 2014.
- [54] Cordelia Schmid, Roger Mohr, et al. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997.
- [55] Patrice Y Simard, Léon Bottou, Patrick Haffner, and Yann LeCun. Boxlets: A fast convolution algorithm for signal processing and neural networks. *Advances in Neural Information Processing Systems*, pages 571–577, 1999.
- [56] Bolan Su, Shijian Lu, and Chew Lim Tan. Binarization of historical document images using the local maximum and minimum. *IAPR International Workshop on Document Analysis Systems*, pages 159–166, 2010.
- [57] Bolan Su, Shijian Lu, and C.L. Tan. Combination of document image binarization techniques. *International Conference on Document Analysis and Recognition*, pages 22–26, 2011.
- [58] Engin Tola, Vincent Lepetit, and Pascal Fua. A fast local descriptor for dense matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [59] Philip HS Torr and David W Murray. Outlier detection and motion segmentation. In *Optical Tools for Manufacturing and Advanced Automation*, pages 432–443. International Society for Optics and Photonics, 1993.
- [60] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.
- [61] Seiichi Uchida and Marcus Liwicki. Part-based recognition of handwritten characters. *International Conference on Frontiers in Handwriting Recognition*, pages 545–550, 2010.
- [62] Georgios Vamvakas, Basilios Gatos, Nikolaos Stamatopoulos, and Stavros J Perantonis. A complete optical character recognition methodology for historical documents. In *IAPR International Workshop on Document Analysis Systems*, pages 525–532, 2008.
- [63] Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
- [64] Luc Van Gool, Theo Moons, and Dorin Ungureanu. Affine/photometric invariants for planar intensity patterns. *European Conference on Computer Vision*, pages 642–651, 1996.

- [65] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/>, 2008.
- [66] Alessandro Vinciarelli. A survey on off-line cursive word recognition. *Pattern recognition*, 35(7):1433–1446, 2002.
- [67] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.
- [68] M.F. Vural, Y. Yardimci, and A. Temizel. Registration of multispectral satellite images with orientation-restricted SIFT. *IEEE International Geoscience and Remote Sensing Symposium*, 3:243–246, 2009.
- [69] Zhenhua Wang, Bin Fan, and Fuchao Wu. Local intensity order pattern for feature description. In *IEEE International Conference on Computer Vision*, pages 603–610, 2011.
- [70] F. Yasuma, T. Mitsunaga, D. Iso, and S.K. Nayar. Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum. *Technical Report, Department of Computer Science, Columbia University CUCS-061-08*, pages 1–30, 2008.
- [71] Z. Yi, C. Zhiguo, and X. Yang. Multi-spectral remote image registration based on SIFT. *Electronics Letters*, 44(2):107–108, 2008.
- [72] Guoshen Yu and J-M Morel. A fully affine invariant image comparison method. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1597–1600, 2009.
- [73] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 809–816, 2011.
- [74] Zhengyou Zhang, Rachid Deriche, Olivier Faugeras, and Quang-Tuan Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial intelligence*, 78(1):87–119, 1995.
- [75] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.