

**Illumination invariant interest point
detection for vision based recognition
tasks**

Flore Faille

Ph.D. Thesis

Lehrstuhl für Realzeit-Computersysteme

**Illumination invariant interest point detection for
vision based recognition tasks**

Flore Faille

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. habil. G. Rigoll

Prüfer der Dissertation: 1. Univ.-Prof. Dr.-Ing. G. Färber

2. Univ.-Prof. Dr.-Ing. E. Steinbach

Die Dissertation wurde am 25. September 2006 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 17. Januar 2007 angenommen.

Contents

List of Figures	vi
List of Tables	ix
List of Symbols	x
1 Introduction	1
1.1 Motivation	1
1.2 Main contributions	3
1.3 Outline of the thesis	4
2 State of the art and related work	6
2.1 Image formation model	6
2.2 Interest point detection	10
2.2.1 State of the art	10
2.2.2 The Harris detector	12
2.2.3 Extension of the Harris detector for colour images	14
2.3 Handling illumination variations	16
2.3.1 Grey value image processing	16
2.3.2 Colour image processing	18
2.4 Summary	24
3 Illumination invariant interest point detection for grey value images	25
3.1 Illumination influence on the Harris detector	25
3.2 Local normalisation	26
3.3 Homomorphic Harris detector	31
3.4 Local threshold adaptation	34
3.5 Local clustering	40
3.6 Handling of saturated areas	45
3.7 Comparison framework	47
3.7.1 Quantitative evaluation of the detection stability	47
3.7.2 Compared interest point detectors	49
3.7.3 Image data set	50
3.8 Detector evaluation and comparison	53
3.8.1 Simple illumination changes	53
3.8.2 Complex illumination changes	55
3.8.3 Conclusion	59

3.9	Summary	61
4	Illumination invariant interest point detection for colour images	63
4.1	Colour image acquisition and demosaicing	63
4.1.1	Presentation of the demosaicing algorithms	66
4.1.2	Comparison of the demosaicing algorithms	69
4.1.3	Selection of the most appropriate demosaicing method	73
4.2	Image formation model and Harris detector for colour images	75
4.3	Robust invariant interest point detector	78
4.4	Homomorphic colour interest point detector	81
4.5	M space interest point detector	84
4.6	Preprocessing for the M space detector	91
4.7	Comparison framework	94
4.7.1	Compared interest point detectors	95
4.7.2	Image data set	96
4.8	Detector evaluation and comparison	97
4.8.1	Simple illumination changes	97
4.8.2	Complex illumination changes	100
4.8.3	Conclusion	104
4.9	Summary	107
5	Application to a recognition task	108
5.1	System overview	108
5.2	Interest point characterisation	110
5.2.1	Choice of the descriptor algorithm	110
5.2.2	SIFT descriptors	112
5.3	Stereo reconstruction	115
5.3.1	Principles of stereo vision	115
5.3.2	Finding correspondences	117
5.3.3	3D reconstruction	120
5.4	Matching	122
5.4.1	Descriptor similarity constraint	123
5.4.2	Geometric constraints	126
5.4.3	Match list constraints	128
5.5	Recognition and localisation	129
5.5.1	Choice of the recognition and localisation algorithm	131
5.5.2	Filling the accumulators	131
5.5.3	Taking uncertainties into account	134
5.5.4	Interpreting the accumulators	136
5.6	Evaluation framework	139
5.6.1	Evaluation criteria	139
5.6.2	Compared interest point detectors	141
5.6.3	Object database and test images	142
5.7	Results	145

5.7.1	Recognition and localisation quality	145
5.7.2	Detector suitability for recognition and localisation	151
5.7.3	Conclusion	154
5.8	Summary	156
6	Conclusion	157
6.1	Summary	157
6.2	Further research	159
	Bibliography	161

List of Figures

1.1	Overview of a general object recognition system	2
2.1	The different elements of the image formation model.	7
2.2	Border handling for convolution.	14
2.3	Detection example with the Harris detector.	15
2.4	Detection example with the Harris detector for colour images.	15
3.1	Detection example for the HD on the same scene under two different illuminations	27
3.2	Suppression of the illumination influence on the derivatives with energy normalisation.	29
3.3	Detection example for the N-HD on the same scene under two different illuminations	30
3.4	Suppression of the illumination influence on the derivatives with homomorphic processing	32
3.5	Detection example for the H-HD on the same scene under two different illuminations	34
3.6	Behaviour of the ratio CF/\overline{CF} in an image series of a scene under different illuminations.	34
3.7	Grey value image and corresponding local standard deviation of CF	35
3.8	Mean and standard deviation of the noise on $ CF $	36
3.9	Mean and standard deviation of the noise on $\ln(CF)$	37
3.10	Detection of the textured areas with the local standard deviation of $\ln(CF)$	38
3.11	Histogram of the noise standard deviation on $\ln(CF)$	39
3.12	Detection example for the AT-HD on the same scene under two different illuminations	40
3.13	Discretisation for the local thresholding.	41
3.14	Histogram of $\ln(CF)$ in image patches	41
3.15	Results of the ISODATA algorithm on neighbourhoods of $\ln(CF)$	43
3.16	Detection example for the LI-HD on the same scene under two different illuminations	45
3.17	Detection and handling of the saturated areas in the image	46
3.18	Two images of the series with shutter time variations.	51
3.19	Sample images of the series with complex illumination variations.	52
3.20	Evaluation results for the series with shutter time variations.	54
3.21	Evaluation results for the <i>nesquik</i> series	55

3.22	Evaluation results for the <i>nesquik</i> series depending on the complexity measure CM	57
3.23	Evaluation results for the <i>paper</i> and the <i>calendar</i> series	58
3.24	Evaluation results for the <i>rabbit</i> and for the <i>shelves</i> series	60
4.1	Bayer CFA.	63
4.2	Typical demosaicing artifacts.	64
4.3	Influence of demosaicing on the colour gradient.	65
4.4	Kernel for median filtering.	67
4.5	Interpolation of the G value at a sampled R pixel with ACPI. G values at sampled B pixels are interpolated similarly.	68
4.6	WACPI directions and neighbourhood shown in fig. 4.7.	68
4.7	Weight and contribution of the left direction (see fig. 4.6) to the interpolation of the G value at pixel position $R6$	69
4.8	Five images of the Kodak colour image database.	69
4.9	Demosaicing results for a colourful image part	70
4.10	Demosaicing results for a textured image part	71
4.11	Influence of white balancing on demosaicing results	74
4.12	Detection example for the C–HD on the same scene under two different illuminations	77
4.13	Detection example for the RI–HD on the same scene under two different illuminations	81
4.14	Suppression of the illumination influence on colour derivatives with homomorphic processing.	83
4.15	Detection example for the HC–HD on the same scene under two different illuminations	84
4.16	Suppression of the illumination influence on colour derivatives	87
4.17	Influence of the preprocessing on the m space	88
4.18	Detection example for the MS–HD on the same scene under two different illuminations	90
4.19	Neighbourhoods used for the Nagao filter.	93
4.20	Neighbourhoods used for the simplified Nagao filter.	93
4.21	Results of the preprocessing methods on an enlarged image detail.	94
4.22	Influence of the preprocessing methods on the m space gradient.	95
4.23	Two images of the series with shutter time variations.	97
4.24	Sample images of the series with complex illumination variations.	98
4.25	Evaluation results for the series with increasing shutter time.	100
4.26	Evaluation results for the <i>shelves</i> series	101
4.27	Evaluation results for the <i>box</i> series	102
4.28	Evaluation results for the <i>giraffe</i> and <i>box2</i> series	103
4.29	Evaluation results for the <i>rabbit</i> and <i>snoopy</i> series	105
5.1	Degrees of freedom of the recognition system	109
5.2	Overview of the recognition system	110
5.3	SIFT descriptor overview	113

List of Figures

5.4	Perspective camera model.	115
5.5	Stereo vision and epipolar geometry.	116
5.6	Search areas for stereo vision	119
5.7	Correspondences between two stereo images	119
5.8	3D Reconstruction of a scene point.	120
5.9	Histograms of the descriptor distances between different interest points and between interest points showing the same scene point under different illuminations	124
5.10	Selection of the threshold D_{lim}	125
5.11	Threshold selection for the descriptors based on homomorphic grey value descriptors.	125
5.12	Influence of the symmetry constraint on the matching process.	130
5.13	Used coordinate systems and estimated localisation parameters	132
5.14	Example of accumulator filling	137
5.15	Example of accumulator filling for images of two different objects	138
5.16	Images used to create the database.	143
5.17	The five different poses for the test images of object 10.	144
5.18	Different illuminations for the test images of object 10.	145
5.19	Different test images of an object not contained in the database.	145
5.20	Recognition rate and localisation accuracy for the different detectors.	148
5.21	Number of votes for the best pose hypothesis for the different detectors.	149
5.22	Matching consistency and deviation for the different detectors.	150
5.23	Detector suitability for recognition and localisation.	153

List of Tables

3.1	Overview of the evaluated interest point detectors.	49
3.2	Evaluation results for the series with small illumination changes	53
4.1	Demosaicing performance in textured areas, in homogeneous areas and in entire images.	72
4.2	Demosaicing performances in coloured areas	73
4.3	Overview of the evaluated interest point detectors.	96
4.4	Evaluation results for the series with small illumination changes	99
5.1	Thresholds for the different descriptor types.	126
5.2	Recognition and localisation performances for the different detectors. . . .	146
5.3	Detector suitability for recognition and localisation	152

List of Symbols

Detectors:

AT-HD	Locally adaptive thresholding Harris detector
C-HD	Colour Harris detector
HC-HD	Homomorphic Colour Harris detector
HD	Harris detector
H-HD	Homomorphic Harris detector
H-HD+NBP	Homomorphic Harris detector with selection of the N best interest points
LI-HD	Local ISODATA Harris detector
MS-HD	M Space Harris detector
MS-HD+N	M Space Harris detector with Nagao preprocessing
MS-HD+WB	M Space Harris detector with simple preprocessing and with white balancing before demosaicing
2MS-HD+N	2 channel M Space Harris detector with Nagao preprocessing
3MS-HD+N	3 channel M Space Harris detector with Nagao preprocessing
3MS-HD+WB	3 channel M Space Harris with simple preprocessing and with white balancing before demosaicing
N-HD	Energy normalised Harris detector
RI-HD	Robust Invariant Harris detector

Demosaicing:

ACPI	Adaptive Colour Plane Interpolation
CFA	Colour Filter Array
EMBP	Enhanced Median Based Postprocessing
MBP	Median Based Postprocessing
WACPI	Weighted Adaptive Colour Plane Interpolation

Similarity measures:

SAD	Sum of Absolute Differences
SSD	Sum of Squared Differences

Abstract

Vision based recognition systems learn the appearance of given objects using images. These objects can be recognised and localised in other images after camera motion and illumination changes. The goal of this work is to improve the ability of such systems to recognise objects after illumination changes. Recognition systems usually reduce the amount of image data used for recognition by detecting interest points: small characteristic image patches. Most interest point detectors are sensitive to illumination changes. Illumination invariant interest point detection would increase the proportion of points which are redetected after illumination changes and it would decrease the proportion of false points. It would hence improve the performance of recognition systems.

Several new interest point detectors with higher stability under illumination changes are developed in this work. They are based on the Harris detector, which is often used because of its stability under viewpoint changes. Four new detectors are developed for grey value images. They all adapt detection to the local lighting intensity using different principles: local normalisation, homomorphic processing, local threshold adaptation and local clustering. Two new detectors are developed for colour images. The first one adapts detection to local lighting intensity and colour with homomorphic processing. The second detector is based on an invariant colour space. It fully eliminates light intensity influence and it locally compensates light colour. In addition, an appropriate demosaicing method and a preprocessing based on the Nagao filter are presented, in order to reduce the influence of noise and colour artifacts on the colour detectors.

Detection stability is evaluated for all new detectors and for the existing Harris detector on image series acquired under varying illumination. The new detectors are more stable than the Harris detector for scenes with complex 3D geometry, for non-uniform lighting and for complex illumination changes such as light source movement. The best results are obtained with the homomorphic detectors and, if the scene has good colour edges, with the detector based on the invariant colour space. A robust state of the art recognition system is also developed and used to evaluate the detectors in a practical application. Systems using colour information perform better than systems based on grey values. For grey value images, the new homomorphic detector improves recognition performances for complex objects or it increases the system efficiency, depending on the chosen thresholding method. For colour systems, the new detectors achieve higher performance improvements than for grey value systems. The detector based on the invariant colour space achieves the best recognition performances if the object contains good colour edges. The homomorphic colour detector also performs very well and is suitable for all kinds of objects. The developed algorithms improve hence detection stability and recognition performances.

1 Introduction

Illumination has a strong influence on images. As a consequence, images of the same scene under different illuminations can be very dissimilar, especially when the scene is non-uniformly lighted. Illumination changes are compensated naturally by eyes and neurons in humans and animals, but they are not easy to handle with camera and computer for machine vision applications. This sensitivity to illumination changes makes the use of machine vision difficult in normal, everyday environments.

Illumination variations are natural and occur frequently. For example, intensity and wavelength composition of sunlight vary with weather and with time of day. Due to earth rotation, sunlight direction also changes during the day. Rooms generally have more than one light source: several lamps, windows and doors. As a consequence, lighting also varies for indoor applications. Last, changes of the camera parameters like aperture, shutter time and white balance influence images in a similar way to illumination changes, because camera parameters are optimised to counterbalance part of the illumination influence. In short, lighting cannot be controlled in normal environments. Therefore vision applications must deal with illumination changes.

Many machine vision applications are based on detecting interest points in the current image and matching those to a model. The aim of this work is to increase the robustness of such applications under illumination changes. Among the various applications using interest points, recognition tasks are the special focus of this work. To achieve higher robustness to illumination changes, new interest point detectors allowing a better re-detection of the interest points after a variation of the illumination conditions are developed and evaluated. After a more detailed description of the motivation, a summary of the contributions as well as the outline of this thesis are presented.

1.1 Motivation

This work is focused on an important topic in machine vision: recognition tasks. These find applications in different domains such as for example robotics, human-machine interface, augmented and virtual reality. Most current recognition methods are based on detecting and matching interest points. Interest points are also used in many other machine vision applications, for example: tracking, content-based image retrieval, registration, 3D-reconstruction and wide baseline stereo. This work is of interest for these applications as well although illumination invariance generally plays a less essential role for these than for recognition tasks.

1 Introduction

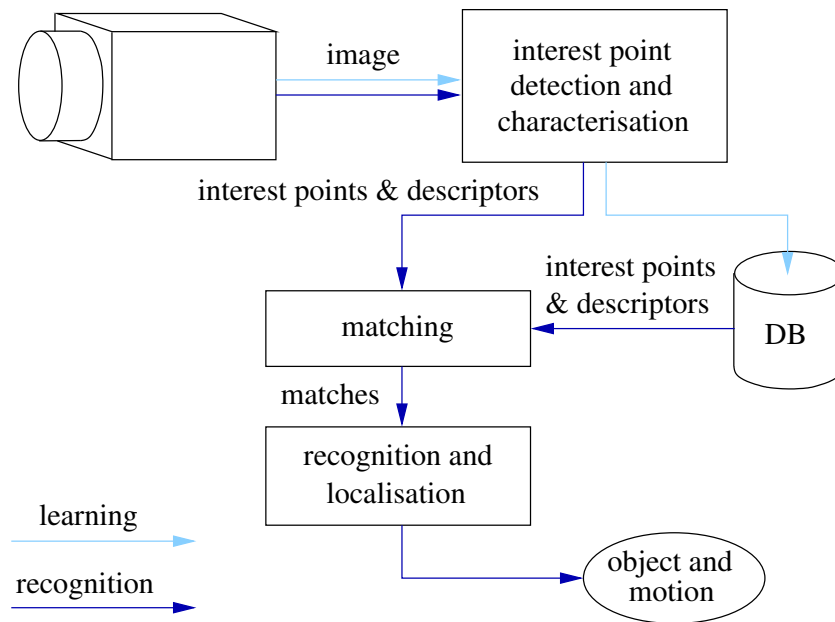


Figure 1.1: Overview of a general object recognition system

Interest points are small characteristic image patches that contain more scene information than other patches. For example, interest points can be corners, dark blobs on a bright background, edge parts with a local curvature maximum. Figure 1.1 shows the overview of a general recognition system using interest points. After detection, the interest points are characterised with descriptors. They are stored together with the descriptors in a database (or model) during the learning phase. During the recognition phase, they are matched to the interest points which are stored in the database by comparing the descriptors. The matches are then used for recognition and localisation. The first three steps (interest point detection, characterisation, and matching) are necessary for all applications based on interest points. Interest point detection reduces the amount of processed data and consequently the computational costs: typically, 50 to 1000 interest points are processed instead of $640 \times 480 = 307200$ pixels. Interest points are local features, i.e. they depend only on a small image part. This has some advantages compared to features computed on large image patches or the whole image such as edges or histograms. For example, changes of the observed scene, occlusions and non-uniform lighting can be handled more easily. They result in the loss of some interest points, which can be easily dealt with by robust recognition algorithms, such as the generalised Hough transform described in [Hou62, Bal81] or RANSAC described in [FB81]. Features depending on large image patches must be recognised with partial information, which is more difficult (see for example [Das02]). Interest points correspond to local scene features, so that scene geometry can be used as additional cue for recognition and localisation. These advantages result in a wide use of interest points in current machine vision applications.

A stable interest point detection is important because interest points and their descriptors are the only image information used for matching and recognition: ideally the same points

with the same descriptors should be detected in both learning and recognition phases. In practise, enough points must be redetected and matched for a reliable application. Robust recognition methods allow to deal with false and missing matches. They are nevertheless more efficient and more accurate with fewer misdetections. Misdetections can be caused among others by: noise, illumination changes, changes of camera parameters, limited changes in viewpoint (due to a movement of the camera or of the object), limited changes in the scene due to occlusions. The stability of existing interest point detectors under viewpoint changes, noise and simple changes of the camera parameters has been investigated and improved for example in [SMB00, MTS⁺05]. This thesis deals with stability under illumination changes. This is important for recognition tasks because learning and recognition phases take place at different time instants. Therefore, illumination variations are very likely to happen, due to a change or a movement of the light source or to a change of the camera parameters.

Many interest point detectors have been developed using various principles, however almost all are sensitive to local image contrast (see subsection 2.2.1). Therefore detection is sensitive to illumination changes. To overcome this sensitivity, as many interest points as possible are usually detected to ensure a minimum number of correct interest points. The use of illumination invariant descriptors helps reducing the number of false matches. When a minimum number of correct matches are found, the right solution can be obtained with robust recognition algorithms. A more stable interest point detection would reduce the number of false interest points, hence reducing the complexity of matching and of recognition. In addition, recognition accuracy would increase because matching and recognition algorithms are faster and more accurate when the number of false interest points and of false matches decreases. There are many different methods to extract illumination invariant image features but surprisingly few illumination invariant interest point detectors. Therefore, in this thesis, principles used to extract illumination invariant features are adapted and applied to interest point detection. The stability of the new interest point detectors is evaluated on images series showing scenes under different illuminations. The new detectors are also evaluated in a practical object recognition and localisation application in chapter 5. The aim and the contributions of this work are summarised in the next section.

1.2 Main contributions

To improve the ability of machine vision applications to deal with illumination changes, several new illumination invariant interest point detectors are developed and evaluated. These new detectors should be as robust to noise and to viewpoint changes as the existing detectors. They should be stable when the type, intensity, position, orientation and number of the light sources change. They should not require any user interaction such as manual white balancing. They should only use a single image for illumination invariant interest point detection: no additional optical filter, no special hardware and no analysis of image sequences should be used.

1 Introduction

To achieve these goals, an existing interest point detector is enhanced: the Harris detector. This algorithm is popular because of its good stability under viewpoint changes. The influence of illumination on the Harris detector is first modelled. This is then used to design the new illumination invariant detectors.

Four new detectors are developed for grey value images. They all adapt detection to the local lighting intensity. Hence, non uniform lighting and unsharp shadow and shading effects can be better handled. All detectors are based on different principles: local normalisation, homomorphic processing, threshold adaptation and threshold selection using clustering. For all detectors, an implementation which is robust to noise is proposed.

Two new detectors are developed for colour images. In comparison to the grey value detectors, they are slower but they can compensate illumination influence more accurately. The first detector compensates local illumination intensity and colour with homomorphic processing. The second detector is based on an invariant colour space. It fully eliminates all intensity, shadow and shading effects and it locally compensates light colour. To reduce the noise influence on the detectors, a robust implementation, special preprocessing and special demosaicing are presented.

The stability of the new detectors and of the current Harris detectors are evaluated and compared to each others on real image sequences acquired under illumination changes. The type, number, position and orientation of the light sources are varied. Several scenes are used, with different 3D geometry properties and different reflectance properties. Last, a state of the art object recognition and localisation application is developed and used to evaluate the influence of interest point detection in a real application. The system performances are evaluated for different objects, different viewpoints and different illuminations. These thorough evaluations show that more interest points are redetected and less false interest points are detected with the new detectors and that this improves recognition and localisation results for scenes with complex 3D geometry and for complex illumination changes.

1.3 Outline of the thesis

Chapter 2 begins with the description of the image formation models for grey value and for colour images. These are used to describe illumination influence in the rest of this work. An overview of the existing detectors is then given and the Harris detector is presented in more detail. Finally, related work is presented: illumination invariant algorithms for interest point detection and for other machine vision tasks.

In chapter 3, the illumination influence on the Harris detector is derived and the instability of this detector under illumination changes is illustrated. All four new detectors for grey value images are then described. A method is presented, which discards false interest points caused by saturation in the images. Finally, the stability of the interest point detectors under illumination changes is evaluated using image sequences acquired under varying illumination.

Chapter 4 deals with interest point detection using colour images. The illumination influence on colour derivatives and on the colour Harris detector are shown. The robust invariant Harris detector is then presented in more details: it is the only existing illumination invariant version of the Harris detector. Next, the two illumination invariant detectors developed in this work for colour images are described. To reduce the influence of colour artifacts and of noise on detection, an appropriate demosaicing method and a preprocessing algorithm are proposed. Finally, the stability of the colour detectors are compared to each others and to the best grey value detector with the same evaluation framework as in chapter 3.

Chapter 5 presents the developed recognition and localisation system. All system blocks are explained in details: stereo reconstruction, characterisation of interest points with descriptors, matching between current interest points and database, and the recognition and localisation algorithm. State of the art and robust methods are used for all system blocks. Last, the influence of interest point detection on the recognition and localisation performances is evaluated. For this, a database of 10 objects and many test images involving illumination and viewpoint changes are used.

Chapter 6 closes the thesis with a summary of the achieved work and suggestions for further research.

2 State of the art and related work

This chapter gives an overview of the related work. First, the image formation model is presented and used to derive the illumination influence on grey value and on colour images in section 2.1. The overview of the existing interest point detectors in subsection 2.2.1 shows that most detectors are sensitive to image contrast, hence motivating this work. The Harris detector and its extension for colour images are then explained in details in subsections 2.2.2 and 2.2.3 as the detectors developed in this thesis are based on them. Finally, an overview of the state of the art methods for dealing with illumination changes in machine vision is given in subsection 2.3.1 for grey value images and in subsection 2.3.2 for colour images. This chapter is summarised in section 2.4.

2.1 Image formation model

The illumination influence on images is derived from the dichromatic image formation model described in [Sha85]. The measured pixel values depend on the spectrum and direction of the incident light, on the spectral and geometrical properties of the scene and on the viewing angle and the spectral sensitivity of the camera, as illustrated in fig. 2.1. This is described by the following formula:

$$C^j = m_b(\mathbf{n}, \mathbf{e}) \int_{\lambda} f^j(\lambda) e(\lambda) c_b(\lambda) d\lambda + m_s(\mathbf{n}, \mathbf{e}, \mathbf{v}) \int_{\lambda} f^j(\lambda) e(\lambda) c_s(\lambda) d\lambda \quad \text{for } j = R, G, B. \quad (2.1)$$

C^R , C^G and C^B are the red, green and blue values of the considered pixel. λ denotes the wavelength. The first term of the sum models body (or Lambertian) reflection, while the second term models surface (or specular) reflection. $m_b(\mathbf{n}, \mathbf{e})$ and $m_s(\mathbf{n}, \mathbf{e}, \mathbf{v})$ express the geometric dependencies of both terms as a function of the light direction \mathbf{e} , of the surface normal \mathbf{n} at the considered scene point and of the viewing direction \mathbf{v} . $f^j(\lambda)$ ($j = R, G, B$) models the spectral sensitivities of the camera channels. $e(\lambda)$ is the spectrum of the incident light. $c_b(\lambda)$ and $c_s(\lambda)$ are the surface albedo and the Fresnel reflectance, which model the spectral scene properties. $c_s(\lambda)$ is usually considered to have a constant value c_s for all wavelengths: this is called the neutral interface reflection (NIR) model in [LBS90]. This means that specular reflections have the same colour as the incident light. For monochrome cameras, the same formula is used with a single intensity channel I . As explained for example in [Tec01], $f^I(\lambda)$ is in general a much broader filter than $f^R(\lambda)$, $f^G(\lambda)$ and $f^B(\lambda)$. In this work, grey value images are acquired using the Y channel of a colour camera. It can be easily shown that the resulting sensitivity $f^I(\lambda) = f^Y(\lambda)$ is a weighted sum of the $f^j(\lambda)$: here $f^I(\lambda) = 0.3f^R(\lambda) + 0.59f^G(\lambda) + 0.11f^B(\lambda)$ (see [Tec01]).

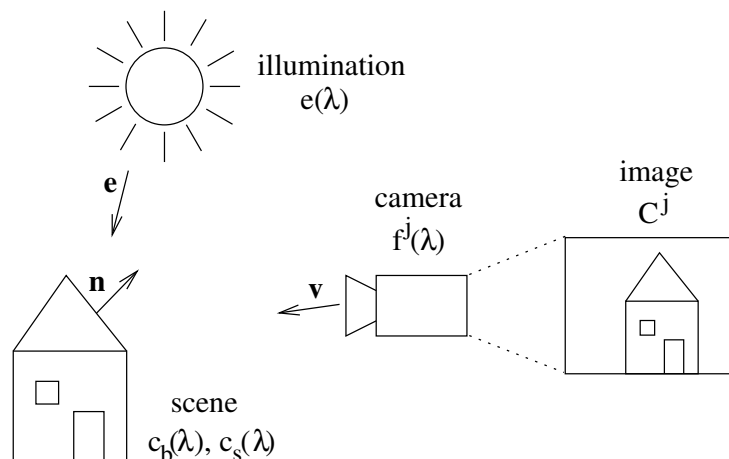


Figure 2.1: The different elements of the image formation model.

Although all elements in eq. (2.1) may vary from one pixel to another, some assumptions can be made concerning their spatial variations. The camera sensitivities $f^j(\lambda)$ can be assumed to remain constant for all pixels. The geometric and the reflection terms m_b , m_s , c_b and c_s may vary abruptly from one pixel to another (high spatial frequencies) as a consequence of the 3D geometry and texture of the scene. On the contrary, the illumination spectrum $e(\lambda)$ is usually assumed to vary slowly, which means it stays constant in small image neighbourhoods. This assumption is necessary to enable the distinction between illumination influence and texture in the scene. It is true if a single light source illuminates the scene and if shadows are considered to change only light intensity. In that case $e(\lambda)$ can be written as $e \bar{e}(\lambda)$ where e is the light intensity and $\bar{e}(\lambda)$ is the normalised spectrum (with norm equal to 1). Shadows are modelled by allowing illumination intensity e to vary abruptly between pixels. The influences of shadows (e) and of shading (m_b) or specularities (m_s) are similar and can be grouped into light intensity factors $i_b = e m_b$ and $i_s = e m_s$. $\bar{e}(\lambda)$ can be assumed to stay constant in the whole image in the case of a single light source. However, this does not model accurately inter-reflections in the scene and scenes lighted by several light sources. More accurate models are presented for example in [Ris01] or in [FHD02]: shadows are lighted by ambient light which results from inter-reflections, hence they may have a different colour from direct light. As explained in [XE01], ambient light has the same spectrum as direct light only when the chromatic average of the scene reflectances is nearly grey (similar for all wavelengths), otherwise shadows are “coloured”. Coloured shadows are not taken into account in this work because they only occur rarely in indoor images. To reduce the approximation error caused by coloured shadows or by multiple light sources, small neighbourhoods should be used during image processing: this allows the model to tolerate more spatial variations of the illumination spectrum.

Changes of the camera parameters can be modelled like illumination changes as they affect images similarly. Aperture and shutter time influence light intensity and white balancing affect light colour. Their influence is identical for all pixels (no spatial variations).

2 State of the art and related work

The dichromatic image formation model is used to model illumination influence on colour and on grey value images. A further approximation is necessary to handle illumination colour changes (changes of the illumination spectrum) more easily. The colour filters $f^j(\lambda)$ are assumed to be narrow band filters, so that they can be modelled by Dirac deltas: $f^j(\lambda) = f^j(\lambda^j)\delta(\lambda - \lambda^j)$ for $j = R, G, B$. If the camera filters are not narrow-band filters, spectral sharpening can be applied as described in [FDF94]: a linear transformation is estimated that results into new colour channels for which the narrow-band filter assumption is fulfilled. Using this approximation and the NIR assumption, eq. (2.1) becomes:

$$C^j = em_b(\mathbf{n}, \mathbf{e})f^j(\lambda^j)\bar{e}(\lambda^j)c_b(\lambda^j) + em_s(\mathbf{n}, \mathbf{e}, \mathbf{v})c_s f^j(\lambda^j)\bar{e}(\lambda^j) \quad \text{for } j = R, G, B. \quad (2.2)$$

The influence of light intensity and spectrum are separated with: $e(\lambda) = e\bar{e}(\lambda)$, where e and $\bar{e}(\lambda)$ are illumination intensity and normalised illumination spectrum. This equation can be reformulated as:

$$C^j = i_b L^j c_b^j + i_s L^j \quad \text{for } j = R, G, B, \quad (2.3)$$

where $i_b = em_b(\mathbf{n}, \mathbf{e})$, $i_s = em_s(\mathbf{n}, \mathbf{e}, \mathbf{v})$, $L^j = f^j(\lambda^j)\bar{e}(\lambda^j)$ and $c_b^j = c_b(\lambda^j)$. i_b models shadows and shading. i_s models shadows and specularities. i_b and i_s also model changes of the shutter time or aperture of the camera. They may vary abruptly between pixels and are identical for all channels. L^j models illumination colour as well as white balancing and is assumed to stay constant in small image neighbourhoods. c_b^j models scene texture and can vary freely. For two images of the same scene under different illuminations, i_b and i_s change for example when the light source moves, and L^j changes if the illumination colour changes. If the scene is assumed Lambertian, the second term of the sum disappears. This yields:

$$C^j = i_b L^j c_b^j \quad \text{for } j = R, G, B. \quad (2.4)$$

i_b , L^j and c_b^j have the same properties as for eq. (2.3). This model is often named the diagonal model, because the illumination influence is modelled by a diagonal matrix if colour values are written as vectors:

$$\begin{pmatrix} C^R \\ C^G \\ C^B \end{pmatrix} = \begin{pmatrix} i_b L^R & 0 & 0 \\ 0 & i_b L^G & 0 \\ 0 & 0 & i_b L^B \end{pmatrix} \begin{pmatrix} c_b^R \\ c_b^G \\ c_b^B \end{pmatrix}. \quad (2.5)$$

If specularities are considered as in eq. (2.3), the model is named diagonal with translation.

When grey value images are used, this model is applied to the grey value channel I, with $L^I = L$ and $c_b^I = c_b$. The distinction between a light intensity change (parameters i_b and i_s) and a light colour change (parameter $L^j = L$) disappears because the image has only a single channel. Equation (2.3) can be simplified to:

$$I = a_b c_b + a_s, \quad (2.6)$$

where $a_b = i_b L$ and $a_s = i_s L$. c_b models the scene texture and can vary freely from one pixel to another. a_b and a_s model the illumination factors and the camera parameters for

body and surface reflections. Both parameters must be assumed to vary slowly from one pixel to another otherwise no distinction between light and texture influence would be possible. This is a coarse approximation of reality which is necessary because little information is available. As a consequence, specular highlights, shadows and shading effects are not modelled accurately (their influence may vary freely in real images). Furthermore, changes of illumination colour cannot be handled correctly, because the intensity channel cannot be modelled properly as narrow-band filter. Therefore, eq. (2.6) is not accurate in neighbourhoods containing several colours. Any illumination change is modelled simply by a variation of both illumination factors a_b and a_s . Like for colour images, the second term of the sum disappears for Lambertian scenes:

$$I = a_b c_b. \quad (2.7)$$

a_b has the same properties as in eq. (2.6).

In the case illumination colour does not change or white balancing is used to counter-balance such a change, a model similar to eq. (2.3) can be derived from the dichromatic reflection model (eq. (2.1)) without using the narrow-band colour filter approximation. The details of this derivation are not given here because illumination colour changes are allowed in this work and no white balancing is applied (see subsection 2.3.2 for explanations). The reader should refer to [GS99] for more details. The model is given as a comparison to the used model (eq. (2.3)):

$$C^j = i_b \alpha^j + i_s \beta \quad \text{for } j = R, G, B, \quad (2.8)$$

with $\alpha^j = \int_{\lambda} f^j(\lambda) \bar{e}(\lambda) c_b(\lambda) d\lambda$ and $\beta = \int_{\lambda} f^j(\lambda) \bar{e}(\lambda) d\lambda$. Only i_b and i_s are allowed to change when the illuminant changes. All elements of the equation vary freely from one pixel to another. The illuminant is assumed to be white (same energy in all wavelengths) as a result of white balancing. In comparison to eq. (2.3), the light colour term (L^j) disappears and the narrow-band filter assumption is not used.

When the colour filters cannot be modelled as narrow-band filters and when light colour varies, a full affine transformation can be used to model illumination influence as explained in [HS97, SH97]. For this, the scene reflectance is approximated with several basis reflectance functions $S_j(\lambda)$. In most cases three functions are enough: $c_b(x, y, \lambda) = \sum_{j=1}^3 \sigma_j(x, y) S_j(\lambda)$. Using eq. (2.1) for a Lambertian surface yields:

$$\begin{pmatrix} C^R \\ C^G \\ C^B \end{pmatrix} = \mathbf{A} \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{pmatrix}, \quad (2.9)$$

with $A_{kj} = \int m_b e(\lambda) f^k(\lambda) S_j(\lambda) d\lambda$. The affine transformation modelled by matrix \mathbf{A} models the whole illumination influence and is assumed to stay constant in small image neighbourhoods. Shadows are therefore not modelled accurately. Shading is only modelled correctly for scenes with smooth underlying 3D surface. In comparison to the used model (see eq. (2.5)), the non-diagonal terms are not zero to model the correlation between the image channels. To consider specularities, a translation can be added as in [MMG02]. The

diagonal model and the full affine model have been compared in [GMD⁺97, MMG02]. In [GMD⁺97], the diagonal model is found to be the best compromise between complexity and modelling accuracy on small neighbourhoods. In [MMG02], the full affine model is found to be necessary for modelling large flat outdoors scenes. For indoor scenes, the diagonal model is found sufficient in [MMG02]. Therefore, the diagonal model of eqs. (2.3) and (2.4) is used in this work.

2.2 Interest point detection

After an overview of the existing interest point detectors in subsection 2.2.1, the Harris detectors for grey value images and for colour images are presented in more details in subsections 2.2.2 and 2.2.3 as this work is based on them.

2.2.1 State of the art

As pointed out in chapter 1, interest points are used in many machine vision applications like object recognition, mobile robot localisation, 3D-reconstruction, tracking, content-based image retrieval... As a consequence, numerous interest point detectors were designed based on various principles and on various definitions. An exhaustive overview of the existing algorithms would go beyond the scope of this section, so a short overview of the different types of interest point detectors is given.

A first detector group is based on the first order image derivatives. In [MS98] and [LE04], an edge map is computed in a first step. Junctions and corners are then found in the edge map using local curvature in [MS98] or a geometric wedge-based junction description in [LE04]. In [LZ03], a fast method for detecting interest points by means of symmetries in the gradient is presented. A general definition of interest point is used in [LZ03], hence not only corners and junctions are detected, but also blobs and more general textured neighbourhoods. In the Harris detector in [HS88], a “corneriness” function is computed using the first order derivatives to select image neighbourhoods with enough gradient in two perpendicular directions. Like in [LZ03], a general definition of interest point is used. The Harris detector is a widely used interest point detector. Its stability under viewpoint changes is improved in [SMB00]. It is also extended to process colour images in [Gou00] and to become invariant to scale changes in [Bau00, Duf01, MS04] or to general viewpoint changes in [MS04]. In [vdW05], the principle is used to design illumination invariant interest point detectors for colour images (see subsection 2.3.2 for more detail).

Another group of algorithms uses second order image derivatives to detect blobs. This interest point definition is well suited to achieve scale invariance. In [Lin98], blobs are detected in scale-space using the Laplacian operator or the determinant of the Hessian matrix. This Hessian detector is extended in [MTS⁺05] to achieve affine invariance. In [Low04], the Laplacian operator is replaced by a difference of Gaussians operator, as it leads to a faster implementation. Another scale-invariant detector based on both first and

second order derivatives is presented in [Lin98] to detect junctions using the curvature of level curves in areas with sufficient gradient.

Another group of methods uses directly the pixel values. In [SB97], [KB01] and [RT01], the detection is based on the local pixel value distribution. In [SB97], the number of pixels having a similar grey value to the centre pixel are counted. The neighbourhoods where this number is minimised are selected as interest points. In [KB01], the local entropy is estimated to select complex, hence characteristic image neighbourhoods. For both methods, a general definition of interest point is used. The method in [KB01] is also scale invariant. In [RT01], colour distributions are used to detect corners and junctions based on a geometric wedge-based description. In [Tuy00] and [MCUP02], image regions with similar grey values and a high contrast to the other pixels (the background) are detected. In [Tuy00], local grey value extrema are chosen to initialise regions. These regions are grown until the contrast between region and background is maximum. In [MCUP02], an algorithm similar to watershed is used to generate all regions resulting from binarisation with all possible thresholds. The maximally stable regions (the regions with the highest contrast to the background) are selected.

Last, several detectors are based on special image processing techniques. In [STL⁺03], an interest point detector using wavelet decomposition is described. A mathematical morphology algorithm for corner detection is presented in [Lag98]. Neuronal modelling of cortical cells is used to design an interest point detector in [WL97]. Interest points can also be detected through saliency mechanisms inspired by the human visual system: several multi-scale cues (for example intensity contrast, edges, cornerness, motion, symmetries, colour contrast) are merged in a saliency map and interest points are detected using centre-surround operators in [LM99, TT04]. In [Kov03], the phase of the different Fourier components is used to detect corners.

The responses of all methods using first or second order derivatives are sensitive to local image contrast and hence to illumination changes (see section 2.1). The only exception is the method in [LZ03] because symmetries can be detected using only gradient orientation. The authors advise however to take gradient magnitude into account too, to reduce noise sensitivity. Some of the detectors, for example in [HS88, Lin98, Low04], are based on local extrema of the detector response. These are stable under illumination changes. Nevertheless, a threshold must be applied to reduce noise sensitivity. This thresholding step is sensitive to illumination changes. For the detectors based directly on the pixel values, the methods in [Tuy00] and [MCUP02] are robust to illumination changes as they detect all local extrema: the region size is used as a detection criterion to reduce noise sensitivity. In [RT01], robustness to contrast changes is achieved by using a non-linear distances in the *CIELAB* colour space. Yet it is not completely invariant to illumination changes and it is also very computation intensive. The last method invariant to illumination changes is presented in [Kov03]: it considers only phase information. Nonetheless, its high computing time limits its usability.

As a conclusion, all existing interest point detectors except the methods in [Tuy00] and [MCUP02] are either sensitive to image contrast or too computation intensive for a wide

use. Both methods in [Tuy00] and [MCUP02] detect free form homogeneous regions with a high contrast to the background. They are sensitive to blur, as shown in [MTS⁺05]. Hence, the goal of this work is to design an illumination invariant algorithm which also detects a different kind of interest points such as corners and junctions. As mentioned in [MTS⁺05], complementary interest point detectors based on different definitions and principles are useful to adapt to the specific application and to the associated image type.

Instead of designing a new illumination invariant detector, an existing detector can be improved by applying it on illumination invariant information like in [vdW05]. The Harris detector is chosen as basis for this work. It is complementary to the detectors in [Tuy00] and [MCUP02] because it detects corners, junctions, blobs and general form interest points using gradient information. Scale and affine invariant versions are presented in [Bau00, Duf01, MS04] and a version for colour images is given in [Gou00]. It is shown in [SMB00, VL01, Lil03, MTS⁺05] to be more stable than many other interest point detectors under viewpoint and illumination changes, blur and camera noise. The detected interest points can be optimally retrieved after a limited camera motion, as proved in [ST94]. They are also shown to possess high information content in [SMB00] and high saliency in [HLS02]. As a consequence, the Harris detector is widely used in various applications such as navigation in [Duf01], mobile robot localisation in [DM02, KSOK00], content-based image retrieval in [SM97], object and face recognition in [CJ02, OI96, WB01], tracking in [ST94], extrinsic camera calibration in [ZDFL95], wide baseline stereo in [Bau00], 3D reconstruction in [VL01], structure from motion in [MB01]... The Harris detector is hence a good basis for developing an illumination invariant interest point detector.

In the next subsections, the Harris detector and its extension to process colour images are explained in details and the implementation used in this work is described.

2.2.2 The Harris detector

The Harris detector¹⁾ is based on a general definition of interest points introduced by Moravec in [Mor79]: interest points are image neighbourhoods in which texture changes significantly in all directions. Texture changes are measured with the sum of squared differences (SSD):

$$SSD(x, y, \delta_x, \delta_y) = \sum_{(u,v) \in W(x,y)} [I(u, v) - I(u + \delta_x, v + \delta_y)]^2. \quad (2.10)$$

(x, y) is the current pixel. (δ_x, δ_y) is the displacement vector. $I(u, v)$ is the grey value of pixel (u, v) . $W(x, y)$ is the neighbourhood centred around (x, y) . Interest points are neighbourhoods for which this measure is high enough for all displacement vectors.

Harris shows in [HS88] how these interest points can be detected with a matrix representing the local statistics of the image derivatives. It is faster because the SSD must not be computed for several displacement vectors. The improvement proposed in [SMB00]

¹⁾ It is also named sometimes the Plessey detector.

is used in this work, as it yields more stability under viewpoint changes. The matrix is defined for each pixel as:

$$\mathbf{M} = G(\sigma_M) \otimes \begin{bmatrix} (I_x)^2 & I_x I_y \\ I_x I_y & (I_y)^2 \end{bmatrix}. \quad (2.11)$$

$G(\sigma_M)$ is a Gaussian with standard deviation σ_M and \otimes is the convolution operator. The first derivatives I_x and I_y are estimated by convolving the grey value image I with the derivatives of a Gaussian with standard deviation σ_{deriv} to reduce noise and aliasing effects: $I_x = G_x(\sigma_{deriv}) \otimes I$ and $I_y = G_y(\sigma_{deriv}) \otimes I$. σ_{deriv} adjusts the amount of noise reduction during derivative estimation. In this work, $\sigma_{deriv} = 1.2$. If \mathbf{M} has two small eigenvalues, texture does not change in any direction: the image neighbourhood is homogeneous. If \mathbf{M} has one small and one high eigenvalue, texture changes only in one direction: the neighbourhood is located near an edge. If \mathbf{M} has two high eigenvalues, texture changes significantly in two perpendicular directions: the neighbourhood is eligible as interest point. σ_M parametrises the neighbourhood size. In this work, σ_M is set to 3.0, which corresponds to a circular neighbourhood with a diameter of approximately 18 pixels.

The ‘‘corneriness’’ function CF allows detection without calculating the eigenvalues:

$$CF = \det(\mathbf{M}) - \alpha \text{trace}^2(\mathbf{M}). \quad (2.12)$$

α is related to the minimum ratio allowed between the two eigenvalues. Harris suggested in [HS88]: $0.04 \leq \alpha \leq 0.06$. Here, α is set to 0.06. CF takes values near 0 in homogeneous regions, negative values near edges and high values near potential interest points. Hence, the interest points are the local maxima of CF above a user-defined threshold T ($T > 0$). Some authors, for example in [VL01, ST94], favour the use of the eigenvalues for detection. In [CJ02], a new normalised detection function taking values between 0 and 1 and based on the eigenvalues is introduced. The corneriness function CF achieves faster processing and as well lower sensitivity to noise and aliasing: less noise-induced interest points are detected near edges when CF is used. It is hence used in this work.

The Harris detector is summarised in the following:

1. Compute the image derivatives I_x and I_y with derivative of Gaussian filters: $I_x = G_x(\sigma_{deriv}) \otimes I$ and $I_y = G_y(\sigma_{deriv}) \otimes I$.
2. Compute the structure matrix \mathbf{M} with eq. (2.11).
3. Compute the corneriness function CF according to eq. (2.12).
4. (x, y) is an interest point:
 - if it is a local maximum of the corneriness function CF
 - and if $CF(x, y) > T$ ($T > 0$).

The calculation of matrix \mathbf{M} is the most computation intensive step of the detector. After the convolution of the image I with the derivatives of a Gaussian, the three different matrix elements are calculated by convolving $(I_x)^2$, $(I_y)^2$ and $I_x I_y$ with a Gaussian. As all involved kernels are separable, convolutions are performed as two sequential 1D convolutions along

2 State of the art and related work

lines and along columns. The derivative of Gaussian filters are implemented straightforwardly since σ_{deriv} is small, which results in a small convolution kernel. For the Gaussian filter (for which $\sigma_M = 3.0$), the recursive implementation proposed by Deriche in [Der93] is used because it reduces the number of performed operations. The filter of order 4 is chosen to obtain good approximation quality. The convolution can be performed with approximately 64 instead of 72 operations per pixel. To reduce border effects during convolution, the values at the image boundaries are extended as shown in fig. 2.2 and the interest points detected within the image border²⁾ are discarded.

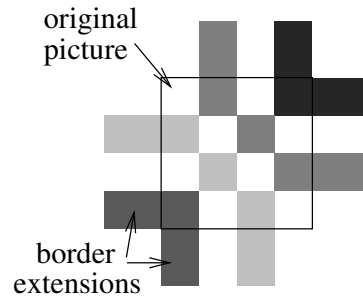


Figure 2.2: Border handling for convolution.

The result of the Harris detector is illustrated in fig. 2.3. The interest points are indicated by circles of radius $3\sigma_M$ to show the considered neighbourhood. It can be seen that the interest points have a general form: corners, blobs, X junctions, T junctions, etc. are detected. The image was acquired with a colour camera. The grey value image is obtained as the Y channel of the YUV colour space: $Y = 0.3C^R + 0.59C^G + 0.11C^B$ (see [Tec01]), where C^R, C^G and C^B are the R, G and B channels of the camera. The threshold is chosen to detect approximately 100 interest points. With this implementation, the Harris detector needs approximately 412ms for an image with 640×480 pixels on a AMD Athlon XP 2200+ computer with 1800MHz and with 256KB cache.

2.2.3 Extension of the Harris detector for colour images

To take into account colour information, the Harris detector must merge the information of several image channels. Several adaptation principles are compared in [Kiß03]: applying the Harris detector to each channel separately and merging the interest points, combining the channel derivatives in the structure matrix \mathbf{M} like in [Gou00], or reducing colour space dimensionality through local thresholding as in [TY96] before applying the Harris detector for grey value images. The adaptation of the Harris detector in [Gou00] achieves the best stability. Using the Di Zenzo colour gradient, the different channel derivatives are combined in the structure matrix according to:

$$\mathbf{M} = G(\sigma_M) \otimes \sum_{j=R,G,B} \begin{bmatrix} (C_x^j)^2 & C_x^j C_y^j \\ C_x^j C_y^j & (C_y^j)^2 \end{bmatrix}, \quad (2.13)$$

where C^R, C^G and C^B are the three image channels. As shown in [Kiß03], it corresponds to using the SSD on colour images with:

$$SSD(x, y, \delta_x, \delta_y) = \sum_{j=R,G,B} \sum_{(u,v) \in W(x,y)} [C^j(u, v) - C^j(u + \delta_x, v + \delta_y)]^2. \quad (2.14)$$

²⁾ The border has a width of $3\sigma_M + 1 = 10$ pixels, corresponding to half the width of a Gaussian kernel.

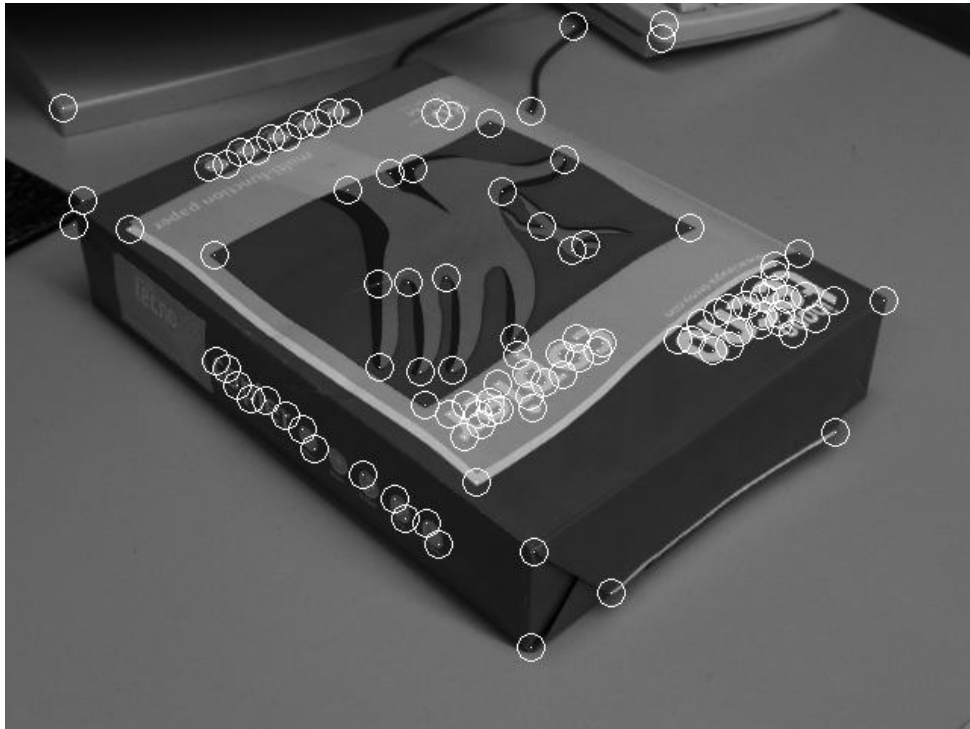


Figure 2.3: Detection example with the Harris detector.

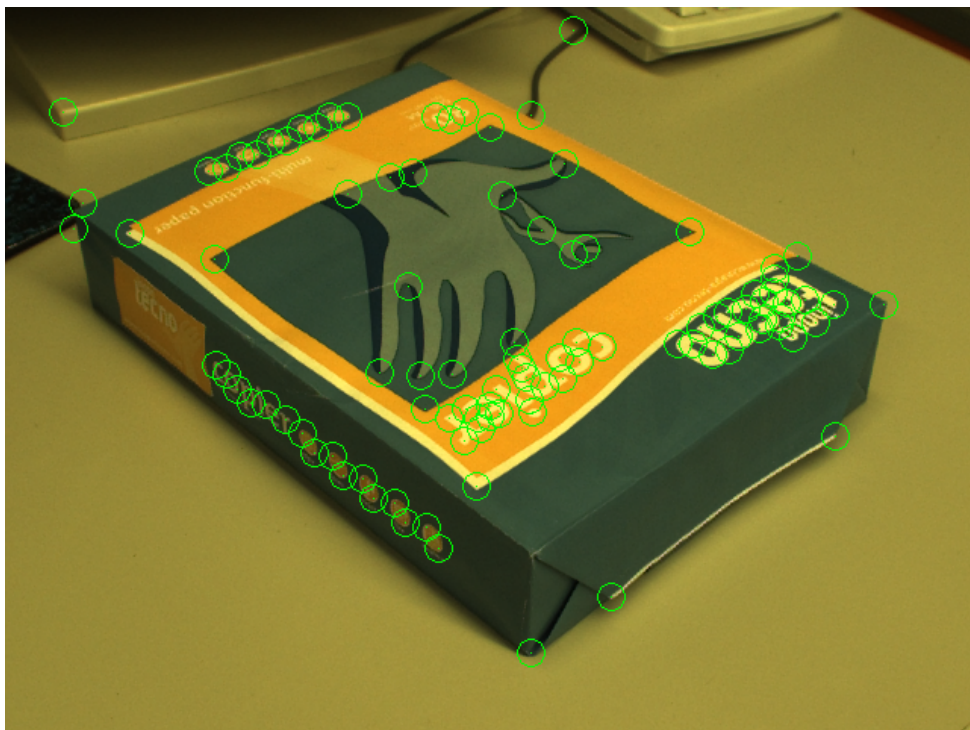


Figure 2.4: Detection example with the Harris detector for colour images.

The cornerness function is then calculated with eq. (2.12) and with $\alpha = 0.06$. The interest points are obtained as before by local maxima detection and thresholding.

The filtering and the border handling are implemented like for the grey value Harris detector. The results are shown in fig. 2.4 on the same scene as in fig. 2.3. The threshold is set to detect 100 interest points like for fig. 2.3. Most interest points are detected by both grey value and colour detectors. However, the colour detector selects more interest points near edges between areas with different colours. The colour Harris detector needs 681ms for an image with 640×480 pixels on the same computer as in subsection 2.2.2. It is 1.66 times the computation time required by the original Harris detector, due to the calculation of six derivative images for the three channels instead of two derivative images for grey values.

2.3 Handling illumination variations

As explained in the introduction, illumination variations are natural and have to be handled by machine vision algorithms. This section gives an overview of the state of the art methods to handle illumination changes for grey value images in subsection 2.3.1 and for colour images in subsection 2.3.2. A complete overview is given in the context of interest point detection. In addition, the different principles which are used in other machine vision applications and which could be adapted to interest point detection are presented.

2.3.1 Grey value image processing

As explained in section 2.1, illumination influence can be modelled by a local affine transformation of the grey values (eq. (2.6)) or by a local multiplicative transformation (eq. (2.7)) if the scene is Lambertian. The first part of this subsection explains how illumination changes are compensated for interest point detection. The second part gives an overview of principles to handle illumination changes in other machine vision tasks.

Interest point detection

As pointed out in subsection 2.2.1, the response of most interest point detectors is sensitive to local contrast and hence to illumination variations. Most of the time, this sensitivity is reduced by detecting local extrema of the detector response because those are not affected by illumination influence (local affine transformations of the grey values). Nonetheless, thresholding is necessary to avoid detecting interest points in homogeneous areas or near straight edges due to noise. Two exceptions are the methods in [Tuy00] and [MCUP02]: all extrema are selected, and the size of the detected region is used as a detection criterion to reduce noise sensitivity. This principle can however not be applied to other detectors. As

shown in [Duf01], the selection of local maxima with a fixed threshold provides enough stability in applications where illumination changes are limited, when a low detection threshold is used.

To increase stability when stronger illumination changes occur, **the detection threshold is adapted to the overall image contrast**. In [SMB00], the used threshold T is proportional to the maximum detector response: $T = 0.01 \max(CF)$. Alternatively, the N local maxima with the highest responses are selected in [Tuy00, MB01]. This threshold adaptation allows to compensate global illumination changes: illumination changes which influence all image pixels equally such as a change of light intensity or of the camera shutter time. It is shown in [Fai03a] that the selection of the N best points allows a more stable detection under illumination changes. The choice of a thresholding method depends however mainly on the application, as explained in [MB01]. With a detection threshold, the number of interest points varies with scene content. As a drawback, the application can become unreliable on simple scenes, for which few interest points are detected. On the other hand, the detection of N interest points results in the detection of noise-induced interest points in simple images. Both thresholding methods only compensate global grey value transformations. Therefore, current grey value interest point detectors cannot compensate illumination influence, as this varies in the image (see section 2.1).

Other machine vision tasks

As explained in chapter 1, recognition systems usually detect many interest points and characterise them with illumination invariant descriptors for the matching. Descriptors are computed in small image neighbourhoods so an affine transformation of the grey values models illumination changes accurately (see section 2.1). This affine transformation is usually compensated by a **normalisation of the pixel values or of the descriptor values**. For this, the mean and the standard deviation of the grey values in the neighbourhood are widely used. This can be performed in an explicit normalisation step like in [MTS⁺05] or implicitly, for example with the normalised crosscorrelation in [KB01, VL01, WB01, ZDFL95, MB01]. In [Sch97], three image characteristics are compared regarding their suitability for local normalisation: mean and standard deviation, minimum and maximum pixel values, as well as energy (defined as the sum of the squared pixel values). The local energy achieves the best results because the descriptors are less noise sensitive. Last, the descriptors can be normalised in a postprocessing step for example with the gradient magnitude in [Low04, MS04, Duf01].

Homomorphic processing is another well-known technique to suppress illumination influence for Lambertian scenes (see for example [GW92]). The local multiplicative model in eq. (2.7) is used. In a first step, the logarithm of the image is taken. Hence, the illumination influence becomes additive. The illumination influence is assumed to be a low frequency signal so it can be suppressed through linear high pass filtering. Finally the exponential of the image is taken if required. Homomorphic processing can be used to enhance non uniformly lighted images as shown in [Smi99]. A current application using

homomorphic processing for illumination invariance is motion detection in [TAM00].

Another machine vision domain in which illumination influence plays an important role is segmentation. Interest point detection can be interpreted as a binarisation (segmentation into two classes) of the detector response. To handle illumination influence in binarisation, the threshold is adapted based on the image histogram or on some other image characteristics such as its information content... An overview and a comparison of such adaptive thresholding methods is given for example in [SS04]. As told in the first part of this subsection, some methods exist to adapt the detection threshold to the overall image contrast for interest point detection. Yet local contrast changes cannot be handled accurately. Hence **local adaptive thresholding** is of particular interest for this work. Two overviews of such methods are given in [TJ95, SS04]. For most algorithms, a threshold is computed for each pixel depending on some characteristics of the pixel values in the neighbourhood. Alternatively, the image can be divided into several non-overlapping windows and the threshold is computed in each of these windows. As less information is available in a local window than in a whole image, only simple methods can be used to adapt or compute the threshold. The threshold is adapted with for example mean (or weighted mean), standard deviation, local minimum and maximum pixel values... Some methods also apply simple clustering methods like the Otsu method in the considered neighbourhoods. For document binarisation, local adaptive thresholding outperforms global thresholding, especially for non-uniformly lighted images as shown in [TJ95].

The last principle presented in this subsection is an interesting and promising technique, which is however not suitable in this work. **A model of the illumination influence is learnt** from several images of the scene lighted from different directions. This allows to learn the shadows and shading effects and hence leads to a more accurate model than the local affine model of eq. (2.6). In this work, illumination invariance should be obtained from a single image, so the method cannot be applied. Furthermore this method is better suited for large image patches or whole objects. It is used for tracking in [HB98] or for object recognition in [Neu01].

2.3.2 Colour image processing

Image formation can be more accurately modelled for colour images than for grey value images, as shown in section 2.1. Illumination also has a higher influence on colour values, especially when light colour changes. Colour is a distinctive information which is important for many applications, especially for visualisation and recognition tasks. As a consequence, many algorithms exist to compute illumination invariant colour values, based on the different image formation models presented in section 2.1. Many algorithms aim at illumination invariant colour values or at modelling human colour perception. Such methods are not presented here as this work aims at illumination invariant filtering of colour images. As in subsection 2.3.1, the first part gives a complete overview of methods to handle illumination changes for interest point detection and the second part

presents related methods to handle illumination changes in other machine vision applications. Last, a short overview of automatic white balancing is given to explain why white balancing is not applied as preprocessing in this work.

Interest point detection

Only few interest point detectors are based on colour information as texture information is almost fully contained in intensity images. As a consequence, grey value detectors are often applied even when colour descriptors are used for subsequent matching, for example in [Tuy00, NG98, Bau00].

In [Gou00], the Harris detector is extended to compute the structure matrix based on colour images (see also subsection 2.2.3). The author does not give any associated thresholding method, but **the same thresholding algorithms as for the grey value Harris detector** can be applied to adapt to the overall image contrast: the detection threshold can be set proportional to the maximum cornerness value or the N local maxima with the highest cornerness value can be selected (see subsection 2.3.1). As for the grey value detector, the resulting method is sensitive to illumination changes causing local changes.

In [vdW05, vdWGG05], **colour invariant and quasi-invariant derivatives** are designed and applied to the colour Harris detector, as well as to the extraction of other features like edges, motion, circles. . . All methods are based on the image formation model of eq. (2.8), therefore white balancing is required before detection. In [vdWGG05], quasi-invariant methods are presented. They allow to detect features caused only by the scene reflectance as they are not influenced by shadow-shading or specular edges. They are computed by projecting the colour derivatives on an invariant plane (for the shadow-shading quasi-invariant) or line (for the shadow-shading-specular quasi-invariant) of the colour space. They are very robust to noise compared to illumination invariant colour features. Nevertheless, the derivatives and cornerness values vary when illumination varies, so they cannot be used in this work. In [vdW05] robust invariant detectors are presented: the edge and cornerness quasi-invariants are normalised to become invariant to illumination changes. This results however also in more noise sensitivity. Robustness is increased through a weighting function designed to counterbalance noise effects. The resulting interest point detectors are only tested on a few images in [vdW05]. The main drawback of these two algorithms is the necessity of white balancing before detection. Existing automatic white balancing methods are indeed not reliable on real images, as shown in [Bar99, FBM98]. The shadow-shading invariant Harris detector is compared to the detectors developed in this work in chapter 4.

Colour-based detectors may also use different principles than the Harris detector. In [TT04], a detector inspired by the human visual attention mechanism is presented. It uses centre-surround operators and several cues including colour to detect salient neighbourhoods. The colour cues are invariant to shadows, to shading and to light colour changes because the RGB values are converted to the **m space** (see [GS99]) before interest point detection. The m space is experimentally shown in [TT04] to provide better

2 State of the art and related work

illumination invariance than the normalised colour space `rgb` or the comprehensive image normalisation presented in [FSC98]. These colour spaces are described more precisely in the second part of this subsection. The resulting interest point detector is however not completely invariant to illumination changes as other cues, for example intensity contrast, are also considered during detection. In [RT01], corners and junctions are detected using the **CIELAB colour space** and a wedge-based interest point description. As it is not only based on chrominance but also on intensity, the detector is not invariant to illumination changes. Robustness is however achieved through the use of non-linear functions.

In [MKK00], invariant colour descriptors are designed. They require at least two distinct colours in the used neighbourhood. Bicolour neighbourhoods are detected based on **colour segmentation in RGB space**. This segmentation also reduces noise sensitivity. No geometry information and a randomised grid are used to generate the analysed neighbourhoods, so the detected bicolour windows have no specific location in the image. As a result of this geometric instability, they can only be used for pure recognition tasks but not for localisation.

Other machine vision tasks

The easiest way to compensate illumination variations is to extend the methods for grey value images by applying them to every channel. This results in the diagonal image formation model with translation of eq. (2.3), with the supplementary assumption that shadow, shading and specular effects are constant in the considered neighbourhood. This method is often applied to compute colour descriptors of interest points for illumination invariant matching. **The neighbourhood is normalised** for each colour channel separately using for example minimum and maximum pixel values, mean and standard deviation or median values in [Gou00]. The normalisation can be performed on the pixel values as preprocessing step as in [OMC03, Bau00, Gou00] or the descriptor values can be normalised as in [MMG99]. If the scene is Lambertian, the local affine transformation becomes a local multiplicative transformation, which can also be corrected by normalisation. In [GMD⁺97], $C^j/\text{mean}(C^j)$ for $j = R, G, B$ is found to provide a good local image normalisation as a preprocessing before interest point detection and matching. In [NB96], a new colour ratio for object recognition is introduced: $(C_1^j - C_2^j)/(C_1^j + C_2^j)$ for $j = R, G, B$. It is an approximation of the invariant ratio C_1^j/C_2^j which reduces noise sensitivity and avoids ill-posed computation for dark pixels. C_1^j and C_2^j ($j = R, G, B$) are the pixel values on the two sides of the considered colour edge. Edges and colour values are obtained through segmentation in RGB space. Local normalisation can also be applied based on the affine model of eq. (2.9). In that case, the colour filters of the sensor are not assumed to be narrow-band. It is used to obtain normalised histogram moments for object recognition in [HS97], normalised mean values in [SH97] and corrected colour values and normalised moments in [LWM02]. The normalisation and the computation of basis reflectance functions for eq. (2.9) are based on eigenvalue decomposition. All these methods based on local normalisation cannot compensate accurately shadows, shading

and specularities because their influence is assumed to stay constant in the considered image neighbourhood. Light colour changes, on the other hand, are correctly handled.

Homomorphic processing can also be applied to colour images. Like for grey value image processing, it can correct a local multiplicative transformation. As a consequence, homomorphic processing handles the diagonal model of eq. (2.4), with the supplementary assumption that shadow and shading influences are constant in the neighbourhood. Illumination colour changes are compensated accurately. For example, homomorphic processing is applied in [LYHT02] to robustly detect changes in colour image sequences despite simultaneous illumination variations. In [FF95], it is used to obtain illumination invariant features for subsequent histogram-based object recognition.

Ratios between colour channels correct all shadow and shading effects accurately, as all channels are affected identically according to the model of eq. (2.3). Colour spaces are often based on such colour ratios to separate intensity and chrominance information. For example, the rgb colour space ($r = C^R / (C^R + C^G + C^B)$, similar formulae for g and b) compensates shadows and shading for Lambertian scenes. The rgb space is widely used in machine vision, for example in [XE01, FSC98, DWL98, GS99, MKK00]. HSI is another popular colour space. $H = \arctan \frac{\sqrt{3}(C^G - C^B)}{(C^R - C^G) + (C^R - C^B)}$ and $S = 1 - \frac{\min(C^R, C^G, C^B)}{C^R + C^G + C^B}$ are both invariant to shadow and shading effects, as shown in [GS99]. It is also proved in [GS99] that H is invariant to specularities when the illumination colour is white (i.e. equal energy in all wavelengths). Perceptive colour spaces such as CIELAB achieve invariance to shadows and shading as well, but they require more complex transformations. They are therefore only used when human-like colour perception must be achieved. In addition to shadows and shading, **light colour must also be compensated**. It can be performed on the whole image through automatic white balancing like in [DWL98, FSC98, CBS03]. In [DWL98, FSC98], simple white balancing methods similar to the grey-world algorithm are applied on the whole image. In [FSC98], the steps for shadow-shading correction and for light colour correction are iterated to provide a comprehensive image normalisation. In [CBS03], a complex illuminant colour estimation is presented and is shown to provide better results than the grey-world algorithm. Alternatively, the illuminant colour can also be compensated locally as in [GS99, NG98, MKK00, XE01]. In [XE01], light colour is explicitly estimated by tracking illuminant colour in an image sequence. This cannot be applied in this work as the algorithm should work on a single image. In [GS99, NG98, MKK00], the rgb colour space is improved to implicitly correct local light colour. It is named the m space in [GS99]: $m^1 = \frac{C_1^R C_2^G}{C_1^G C_2^R}$, where C_1^j and C_2^j are the colour values of two neighbouring pixels. m^2 and m^3 are defined similarly for C^R/C^B and C^B/C^G . Those colour ratios are invariant to shadows, shading and light colour for Lambertian scenes according to the image formation model of eq. (2.3) (see [GS99]). In [NG98], an approximation for these ratios is introduced to reduce noise sensitivity and ill-posed computation in dark areas: $\phi^{G/R} = \frac{C_1^G/C_1^R}{C_1^G/C_1^R + C_2^G/C_2^R}$ (similar definition for $\phi^{B/R}$). Illumination invariance can be achieved in a preprocessing step as in [FSC98, CBS03] where the pixel values are normalised before image processing. Alternatively, it is performed during feature extraction in [NG98, GS99, MKK00, DWL98, XE01]. These methods are used in various applica-

2 State of the art and related work

tions, for example tracking in [XE01, CBS03], histogram-based or geometry-based object recognition in [GS99, NG98, DWL98], segmentation in [GSS98]. . . A last method to compute invariants to shadows, shading and light colour is based on ratios and on derivatives in both image and wavelength dimensions. It is described in [GvdBSG01]. It is used for example to detect edges in [GDvdB⁺99]. Its main drawback is the requirement of a calibrated camera to define derivative filters in the wavelength dimension.

Invariant colour features based on ratios are **sensitive to noise** especially in dark areas, because the division emphasises noise. In addition, a ratio is not defined when the divisor is zero. For that reason, better behaved approximations are introduced in [NB96, NG98]. For segmentation, clusters with a given form in RGB space (for example a plane or a line) can be searched instead of point-like clusters in invariant spaces as in [Sto00, FCF96]. This principle can however not be applied in this work, because interest point detection is performed even for colour information with a single channel: the cornerness function. In [GS04], noise statistics of the invariants are taken into account during histogram construction for object recognition to counterbalance the higher noise sensitivity in dark areas. Similarly, in [Sto00], noise statistics are considered to set the detection threshold for colour edge detection. In [vdWGG05], quasi-invariant derivatives are defined to robustly detect features depending only on object reflectance. As explained in the first part of this subsection, they cannot be used for detection because the quasi-invariant derivatives vary when illumination changes. Therefore, in [vdW05], a robust method to compute real invariants from these quasi-invariants is presented. It is based on a weighting function designed using the noise statistics to reduce noise sensitivity. In addition to photon and electronic noise, colour artifacts may appear, mainly in the vicinity of colour edges. This is caused by chromatic aberrations of the optics, by demosaicing (if a single-chip colour camera is used) or by misregistration of the sensor chips (if a multi-chip colour camera is used), as explained in [BMCF02b]. For that reason, segmentation is applied as a preprocessing step to reduce noise and to suppress artifacts in [BMCF02b, MKK00].

Finally, in [FHD02], **coloured shadows** are estimated: two light colours are compensated, for direct and for ambient lighting. The colour values are projected on a camera specific line to yield a grey value image which is invariant to coloured shadows, to shading and to light colour changes for Lambertian scenes. This projection can be used to detect invariant features or as in [FHD02] as a preprocessing to remove shadows. It requires a calibrated camera or several images taken in the environment. Coloured shadows are more important for images of outdoor scenes, as explained in [Ris01]. In addition, ambient light are not well approximated by a Planckian light source for indoor scenes, because it results only from inter-reflections in the scene and not from sky and clouds. Therefore, coloured shadows are not compensated in this work.

Light colour estimation and automatic white balancing

Light colour has a strong influence on colour values. Therefore, many algorithms for automatic white balancing exist, mostly for visualisation or colour management purposes.

When colour images are processed, white balancing may be required as a preprocessing step like in [vdW05], or light colour can be compensated implicitly by the invariants like in [GS99]. A short overview over automatic white balancing algorithms is given here.

The simplest white balancing algorithms are based on assumptions inspired by the colour constancy mechanism of the human visual system: **the grey-world and the white-patch algorithms**. The grey-world algorithm assumes the average reflectance in the scene to be grey, so the average colour value in the image should be grey after white balancing. Therefore, all channels are normalised by their average values: $C^j \rightarrow C^j / \text{mean}(C^j)$ for $j = R, G, B$. This algorithm is the basis of the retinex algorithm by Land. The retinex algorithm can be applied on the whole image or locally, using for example a centre-surround operator as in [Lan86]. When the grey-world algorithm is applied locally, homogeneous areas become grey as explained in [GMD⁺97, RJW02]. For visualisation applications, this can be corrected through scale selection and recolouration as proposed in [RJW02]. The white patch algorithm assumes the lightest pixels in the scene to have the same colour as the light source: either the light source is visible in the scene, or it is reflected by a white surface (which, of all surfaces, reflects the most energy), or the lightest pixels are a specular highlight, which according to the NIR assumption has the same colour as the illuminant (see section 2.1). The human visual system uses a mixture of both grey-world and white-patch hypotheses. Such a combination of both algorithms is achieved for example in [RGM03].

More advanced colour constancy algorithms often **use statistical knowledge** on the expected colours. An overview and a comparison of automatic white balancing methods on synthetic and on real images is given in [BMCF02a, BMCF02b]. Automatic white balancing methods are shown to be less reliable on real images. The simple white-patch algorithm provides surprisingly good results on real images, so it provides the best compromise between performance and complexity according to [BMCF02b]. Other recent colour constancy algorithms are **based on physics** to model specularities exactly. They use highlights to estimate light colour like for example in [TNI03]. Statistics based methods like the ones analysed in [BMCF02a] require a scene with many distinct colours, whereas physics based methods like in [TNI03] require surfaces with a single colour. Both principles can be combined, for example in [FS01].

Some machine vision applications assume a perfect correction of the light colour in a preprocessing step, for example in [vdW05]. However, current automatic white balancing methods are shown to work unreliably on real images in [BMCF02b]. In [Bar99, FBM98], the reliability of current colour constancy algorithms is tested in the context of histogram based object recognition. This shows that a preprocessing with automatic white balancing methods improves recognition results. However, the conclusion in [Bar99, FBM98] is that **automatic colour constancy is not good enough yet**. As colour constancy in the human visual system is known to be imperfect, the authors raise the question whether human colour constancy would be good enough for machine vision applications. In [CBS03], it is shown that a normal colour constancy algorithm does not perform well enough for their object tracking application. They design a special method based on Bayesian inference and on learning the object appearance under several illuminants. Their

method cannot be used in this work as it should work on a single image. In [XE01], white balancing is based on tracking the light colour in an image sequence. It is therefore not suitable for this work. In [OMC03], the use of white balancing as preprocessing is questioned. In their application, scenes are matched using interest points and a simple local photometric normalisation. Colour constancy is performed as a postprocessing step, based on the matched neighbourhoods.

As a consequence, the illumination invariant interest point detectors developed in this work for colour images will compensate light colour implicitly like in [OMC03] or [GS99] to avoid any unreliability caused by automatic white balancing. The stability of the invariant detector described in [vdW05] which requires white balancing will be compared to the detectors developed in this work.

2.4 Summary

In the first part of this chapter, the dichromatic image formation model is presented and simplified to obtain the diagonal model with translation (eq. (2.3)), which is the model used in this work. It describes the influence of shadows and shading, of specularities and of light colour on colour images. A similar model is derived for grey value images (eq. (2.6)). The assumptions and the limitations of both models are also presented.

The second part of this chapter gives an overview of the existing interest point detectors. Most detectors are sensitive to the local image contrast and therefore to illumination variations. This motivates this work, in which illumination invariant versions of the Harris detector for grey value and for colour images are developed. The Harris detector is the basis for this work because it is stable under viewpoint changes and because it is used in numerous applications. Its principle and its implementation are described in more details in subsections 2.2.2 and 2.2.3.

Finally, an overview of methods to handle illumination changes for interest point detection and for other machine vision applications is given. For grey value images, the existing interest point detectors adapt the detection threshold to the overall image contrast. This is however not enough to handle illumination changes, as the contrast may change locally due to shadow or shading effects. Such local changes are handled in other machine vision applications by local normalisation, homomorphic processing or local adaptive thresholding. For colour images, interest point detectors invariant to shadows, shading and specularities are presented in [vdW05]. They require however white balanced images. As current automatic white balancing methods are not reliable on real images, colour interest point detectors are developed in this work that do not require any white balancing. Illumination changes can be handled in machine vision applications by using one of the numerous illumination invariant colour spaces. Alternatively local normalisation and homomorphic processing can be applied. Noise sensitivity is also handled in this work because the existing invariants emphasise noise in dark areas and because current cameras produce colour artifacts near edges.

3 Illumination invariant interest point detection for grey value images

In this chapter, the developed illumination invariant detectors for grey value images are presented and their performances are evaluated and compared. After a reminder of the image formation model for grey value images and of the Harris detector, the illumination influence on the Harris detector is derived in section 3.1. Next, the four developed detectors and their implementation are explained: they are based on local normalisation (section 3.2), on homomorphic processing (section 3.3), on local threshold adaptation (section 3.4) and on local clustering (section 3.5). To reduce the influence of specular highlights, interest points detected near saturated image areas are filtered out as described in section 3.6. Evaluation framework and evaluation criteria are described in section 3.7. Finally the comparison results are given in section 3.8. Section 3.9 summarises the chapter.

3.1 Illumination influence on the Harris detector

First, the image formation model for grey value images and the Harris detector are summarised. More details are given in section 2.1 and in subsection 2.2.2. The illumination influence on grey value images is modelled by a local affine transformation:

$$I = a_b c_b + a_s, \tag{3.1}$$

where I is the pixel grey value and c_b is the scene reflectance. a_b and a_s model the illumination influence on body and surface reflections. Both a_b and a_s are assumed to stay constant in small neighbourhoods. As a result of this assumption, the influence of light colour and of sharp shadow, shading and specular patterns is not modelled accurately. These inaccuracies are necessary because the image only has a single channel. Modelling inaccuracies can be reduced by decreasing the size of the neighbourhoods on which a_b and a_s are assumed to be constant. If the scene is assumed to be Lambertian, the image formation model becomes a local multiplicative model:

$$I = a_b c_b, \tag{3.2}$$

where I , a_b and c_b have the same properties as in eq. (3.1).

The Harris detector is based on following structure matrix:

$$\mathbf{M} = G(\sigma_M) \otimes \begin{bmatrix} (I_x)^2 & I_x I_y \\ I_x I_y & (I_y)^2 \end{bmatrix}, \tag{3.3}$$

3 Illumination invariant interest point detection for grey value images

which represents the local statistics of the image derivatives I_x and I_y . $G(\sigma_M)$ is a Gaussian with standard deviation σ_M . \otimes represents convolution. I_x and I_y are obtained by convolving image I with derivatives of Gaussian with standard deviation σ_{deriv} : $I_{x/y} = G_{x/y}(\sigma_{deriv}) \otimes I$. The cornerness function CF is computed from matrix \mathbf{M} with:

$$CF = \det(\mathbf{M}) - \alpha \text{trace}^2(\mathbf{M}). \quad (3.4)$$

The interest points are the local maxima of CF above detection threshold T ($T > 0$). To adapt detection to the overall image contrast, T can be set proportional to the maximum of the cornerness function, or the N interest points with the highest cornerness values can be selected. The processing time of the Harris detector is approximately 412ms for an image with 640×480 pixels (see subsection 2.2.2 for more details).

In the image formation model of eq. (3.1), the illumination influence on grey values is modelled by locally constant factors a_b and a_s . As a consequence, the image derivatives I_x and I_y are influenced by the multiplicative illumination factor a_b :

$$I_x = a_b c_{bx} \text{ and } I_y = a_b c_{by}, \quad (3.5)$$

where c_{bx} and c_{by} are the derivatives of the scene reflectance c_b . The structure matrix \mathbf{M} is hence influenced by factor a_b^2 , because a_b is assumed to be constant on the neighbourhood considered for the convolution with $G(\sigma_M)$. As a result, the illumination influence on cornerness function CF is the local multiplicative factor a_b^4 :

$$CF(I) = a_b^4 CF(c_b). \quad (3.6)$$

a_b^4 stays constant in image neighbourhoods, so the local maxima of the cornerness function are stable under illumination changes. Nevertheless, the detection threshold T should be adapted to the local illumination factor for illumination invariant detection: a_b^4 may vary between distant pixels, especially for non uniformly lighted scenes.

Figure 3.1 shows that the Harris detector cannot compensate complex illumination changes, even if the detection threshold is adapted to the overall image contrast. The left image is illuminated by neon lamps. The right image shows the same scene lighted by sunlight. The detection threshold is set for each image such that approximately N interest points are detected ($N = 100$). Only 46.1% of the interest points in the left image are redetected in the right image. 51.6% of the interest points in the right image do not correspond to any interest point of the left image. Such interest points without any correspondence in the reference image (here the left image) are named false positives. They may produce false matches in applications. Therefore, detection is stable when a high proportion of interest points are redetected and when the proportion of false matches is low. To improve the detection stability under complex illumination changes, new interest point detectors are developed in this thesis, that adapt detection to the local lighting conditions.

3.2 Local normalisation

As explained in subsection 2.3.1, one popular principle to handle locally varying illumination conditions is local normalisation. The Harris detector is composed of several steps,



Figure 3.1: Detection example for the Harris detector with selection of the N best points ($N = 100$). The images show the same scene under two different illuminations. 46.1% of the interest points of the left image are redetected in the right image. 51.6% of the interest points in the right image are false positives: they do not correspond to any interest points of the left image. Both images are gamma corrected for visualisation ($\gamma = 1.4$).

so normalisation could be performed on the derivatives, on the elements of matrix \mathbf{M} or on the cornerness function CF . A normalisation of the cornerness function is equivalent to an adaptation of the detection threshold like the method presented in section 3.4. It is therefore not used here. Derivative normalisation induces less computation than the normalisation of the elements of \mathbf{M} because two derivatives are normalised instead of three distinct elements. In addition, the noise introduced by normalisation is reduced by the convolution with the Gaussian $G(\sigma_M)$ in eq. (3.3). A last alternative consists in a local grey value normalisation in a preprocessing step as proposed for colour images in [GMD⁺97]. Preliminary tests performed in this work showed that local derivative normalisation is better suited for interest point detection than local grey value normalisation. The Harris detector is indeed sensitive to the artifacts introduced near edges by local grey value normalisation (see [GMD⁺97] for an illustration of these artifacts on colour images). As a conclusion, normalisation is performed on the derivatives.

The neighbourhood size used for normalisation must be chosen. To reduce the inaccuracies of the image formation model, the neighbourhood size should be as small as possible. On the other hand, normalisation amplifies image noise more strongly when it is performed on smaller neighbourhoods. Therefore, a compromise is necessary. To reduce the number of parameters, the same pixels are used for both derivative computation and normalisation. This is realised with the normalised convolution proposed in [Sch97].

Next, the characteristics used for normalisation must be chosen. Three local normalisations based on minimum and maximum grey values, on mean and variance and on energy are compared in [Sch97]. The normalisation using local energy shows the best behaviour, because the normalised features are less sensitive to noise, especially for small

3 Illumination invariant interest point detection for grey value images

size neighbourhoods and for images with homogeneous areas. The behaviour of normalisation in homogeneous areas is of particular interest here because the cornerness function is computed for all pixels. If the local normalisation amplifies image noise strongly in homogeneous areas as is the case when local minimum and maximum values or when local standard deviation are used, noise-induced hence unstable interest points are detected in those areas. For the same reason, normalisation based on gradient values as in [Low04, MS04, Duf01] is not used here. As a conclusion, normalisation is performed with local energy here. A drawback is that it cannot compensate the local affine model of eq. (3.1) but only the local multiplicative model of eq. (3.2): this is shown in eq. (3.9).

To summarise, the normalised convolution using local energy is used to compute illumination invariant derivatives for the Harris detector. It is defined in [Sch97] by:

$$output(x, y) = \frac{\sum_{(i,j) \in W} I(x+i, y+j) kernel(i, j)}{\sqrt{\sum_{(i,j) \in W} I(x+i, y+j)^2} \sqrt{\sum_{(i,j) \in W} kernel(i, j)^2}}. \quad (3.7)$$

$I(x, y)$ represents the image grey values. $output(x, y)$ is the result of normalised convolution. $kernel$ is the convolution kernel, here the derivative of Gaussian kernels. W is the window associated with $kernel$. In [Sch97], several kernels with different sizes are combined for object recognition. Here, only the two derivation kernels in x and y directions are used. Both have the same weighting factor $\sqrt{\sum_{(i,j) \in W} kernel(i, j)^2}$ as they are related by a 90° rotation. Therefore, the division by the kernel energy in eq. (3.7) is superfluous and can be suppressed. This results in:

$$output(x, y) = \frac{\sum_{(i,j) \in W} I(x+i, y+j) kernel(i, j)}{\sqrt{\sum_{(i,j) \in W} I(x+i, y+j)^2}}. \quad (3.8)$$

The illumination parameter a_b in eq. (3.2) is assumed constant on the considered neighbourhood W . As a consequence, the result of normalised convolution using local energies is not influenced by local lighting conditions for Lambertian scenes:

$$\begin{aligned} output(x, y) &= \frac{\sum_{(i,j) \in W} a_b c_b(x+i, y+j) kernel(i, j)}{\sqrt{\sum_{(i,j) \in W} a_b^2 c_b(x+i, y+j)^2}} \\ &= \frac{a_b \sum_{(i,j) \in W} c_b(x+i, y+j) kernel(i, j)}{\sqrt{a_b^2 \sum_{(i,j) \in W} c_b(x+i, y+j)^2}} \\ &= \frac{\sum_{(i,j) \in W} c_b(x+i, y+j) kernel(i, j)}{\sqrt{\sum_{(i,j) \in W} c_b(x+i, y+j)^2}}. \end{aligned} \quad (3.9)$$

The local affine model of eq. (3.1) cannot be handled by the energy normalised convolution. Using the invariant derivatives computed by energy normalised convolution in the Harris detector results in an invariant cornerness function. Therefore, thresholding with a user-defined threshold leads to illumination invariant interest point detection. Figure 3.2

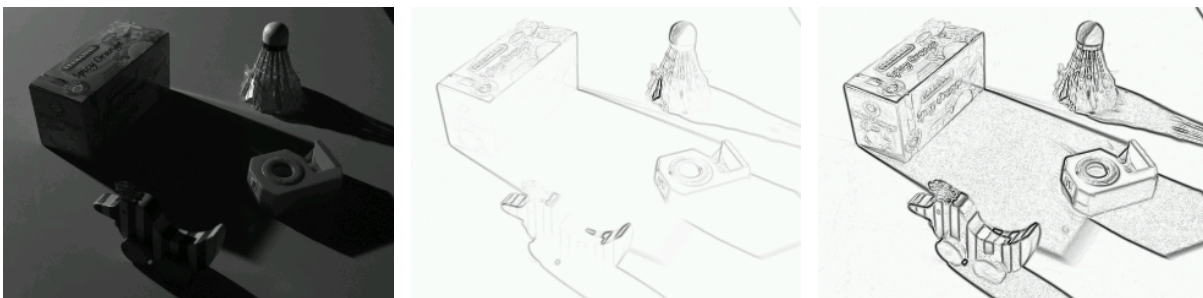


Figure 3.2: Suppression of the illumination influence on the derivatives with energy normalisation. Left: grey value image. Middle: gradient $\sqrt{I_x^2 + I_y^2}$. Right: energy normalised gradient. The gradient images are scaled between 0 and 255 using minimum and maximum gradient values. The grey value image is gamma corrected ($\gamma = 1.3$).

illustrates how the local lighting conditions are compensated using energy normalisation. The energy normalised gradient has similar values in both shadows and directly lighted areas, which is not the case for the normal gradient. A drawback is the noise amplification in dark areas, which is visible in the shadows in fig. 3.2.

The resulting detection algorithm is summarised here:

1. Compute the image derivatives I_x and I_y using the derivatives of Gaussian.
2. Compute the local energies $E(x, y) = \sum_{(i,j) \in W} I(x+i, y+j)^2$.
3. Normalise the derivatives with the local energies: I_x/\sqrt{E} and I_y/\sqrt{E} .
4. Compute the structure matrix \mathbf{M} according to eq. (3.3) with the normalised derivatives.
5. Compute the cornerness function CF according to eq. (3.4).
6. (x, y) is an interest point:
 - if it is a local maximum of the cornerness function CF
 - and if $CF(x, y) > T$ ($T > 0$).

The window for energy summation should have the same size as the derivation kernels (here $7 \approx 6\sigma_{deriv}$). T is a user-defined threshold. It should be set to obtain an appropriate number of interest points. The normalised derivatives can be re-used to compute descriptors: with one step, both interest point detection and characterisation become invariant. This method is called Normalised Harris Detector (N-HD) in the following. The algorithm can be easily extended to other Harris detector versions, in particular to the scale or viewpoint invariant versions in [MS04]. The principle can also be applied to other detectors based on first or second order image derivatives (see subsection 2.2.1).

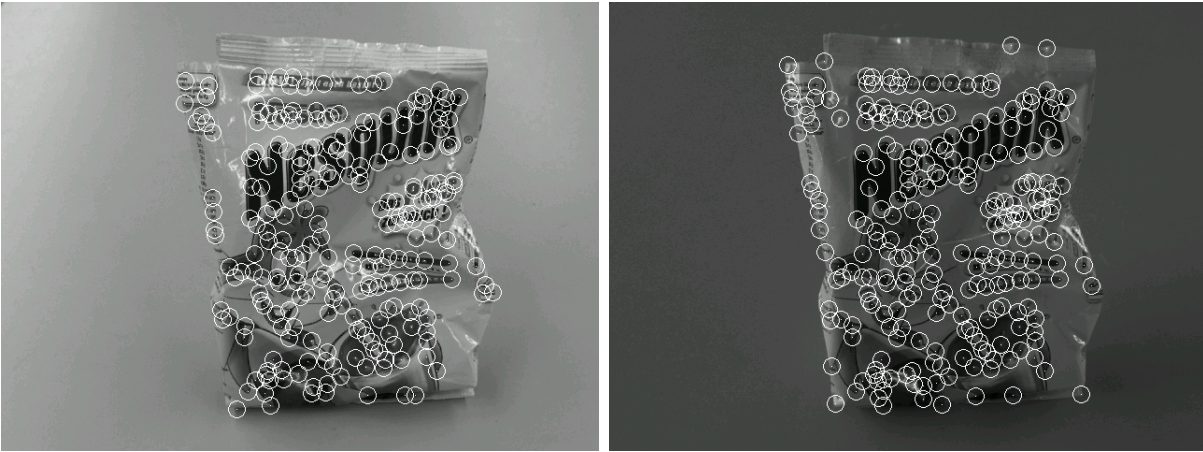


Figure 3.3: Detection example for the energy normalised Harris detector ($T = 0.05$). The same images as in fig. 3.1 are used. 76.1% of the interest points of the left image are redetected in the right image. 26.5% of the interest points in the right image are false positives. Both images are gamma corrected for visualisation ($\gamma = 1.4$).

The convolutions with the derivatives of Gaussian and with the Gaussian are implemented as explained in subsection 2.2.2. To compute the local energies $E(x, y)$, the first step is to compute the square of the grey values. A box filter is then applied on the squared grey values. The box filter should have the same size W as the derivation kernels. It is implemented as two sequential 1D convolutions along lines and along columns because the kernel is separable. In addition, all pixels in the window contribute with the same weight to the local energy. Therefore, a recursive implementation can be used to reduce computational costs. For 1D convolution, the value of a pixel $output(x)$ is computed from the value of its neighbour $output(x - 1)$ using only one addition and one subtraction:

$$output(x) = \sum_{i=-N}^N input(x + i) = output(x - 1) + input(x + N) - input(x - 1 - N),$$

where $2N + 1$ is the size of window W .

The results of the energy normalised Harris detector are illustrated in fig. 3.3. The same images as in fig. 3.1 are used to allow a better comparison to the Harris detector (HD). With N-HD, the interest points are not only detected in the areas with the highest local contrast. N-HD achieves a higher detection stability than the Harris detector on this example. 76.1% of the interest points of the left image are redetected after the illumination change. Only 26.5% of the interest points in the right image are false positives. With the proposed implementation, N-HD requires 480ms processing time for an image with 640×480 pixels (see subsection 2.2.2 for details on the computer): it is 1.17 times the processing time of HD.

3.3 Homomorphic Harris detector

The next developed invariant detector is based on homomorphic processing, which allows to compensate locally varying illumination conditions for Lambertian scenes. The illumination influence on grey values can be modelled with a local multiplicative factor (see eq. (3.2)). The illumination factor a_b is assumed to stay constant in small neighbourhoods. By taking the logarithm of the grey value image (which is positive), homomorphic processing transforms the multiplicative model into an additive one:

$$\ln I = \ln a_b + \ln c_b. \quad (3.10)$$

The illumination influence $\ln a_b$ can be assumed to stay constant in small neighbourhoods, while grey values $\ln I$ and reflectance $\ln c_b$ vary freely. As a consequence, illumination invariant information can be obtained by applying a linear high-pass filter on $\ln I$.

Therefore, homomorphic processing yields illumination invariant image derivatives:

$$\frac{\partial \ln I}{\partial x} = \frac{I_x}{I} = \frac{a_{bx}c_b + a_b c_{bx}}{a_b c_b} \approx \frac{c_{bx}}{c_b} \quad \text{and} \quad \frac{\partial \ln I}{\partial y} = \frac{I_y}{I} \approx \frac{c_{by}}{c_b}. \quad (3.11)$$

This approximation is valid because the illumination factor a_b is constant in small neighbourhoods, so its derivatives are approximately zero: $a_{bx} \approx a_{by} \approx 0$. I_x , I_y and c_{bx} , c_{by} are the derivatives of the image and of the scene reflectance. Like in section 3.2, illumination invariant interest point detection can be achieved by thresholding with a user-defined threshold when the Harris detector is based on these invariant derivatives. As can be seen from eq. (3.11), the invariant derivatives obtained by homomorphic processing can be interpreted as derivatives normalised with the local mean values: I_x/\bar{I} and I_y/\bar{I} , where \bar{I} is the mean value on the neighbourhood used for derivative computation.

If implemented in a straightforward manner, the dark image areas cannot be handled properly with homomorphic processing. The noise influence is indeed strongly amplified, as the divisor in eq. (3.11) takes values near zero. Additionally pixels with a grey value equal to zero cannot be handled, as their logarithm is not defined. The use of $\ln(1+I(x, y))$ instead of $\ln I(x, y)$ proved experimentally to be a good work around for both problems. In bright regions, adding 1 to the grey values has a negligible effect on the derivatives. In dark regions, it actually helps to reduce noise influence. In addition, $\ln(1+I(x, y))$ takes only positive values like a natural grey value image. Depending on the amount of noise introduced by the camera, it may be necessary to further attenuate noise in dark image areas. Otherwise, noise-induced hence unstable interest points may be detected in dark homogeneous image areas. Here, a simple 3×3 box filter is applied in dark areas: all pixels with a grey value smaller than threshold V are replaced by the mean value in their 3×3 neighbourhood. V was chosen experimentally to be 3. This threshold and if necessary the preprocessing filter should be adapted to the camera noise. Fig. 3.4 illustrates on a non-uniformly lighted scene how the illumination influence on the derivatives is suppressed by homomorphic processing and how noise effects in dark regions are reduced by the presented implementation. The standard gradient is very low in shadows. As shown

3 Illumination invariant interest point detection for grey value images

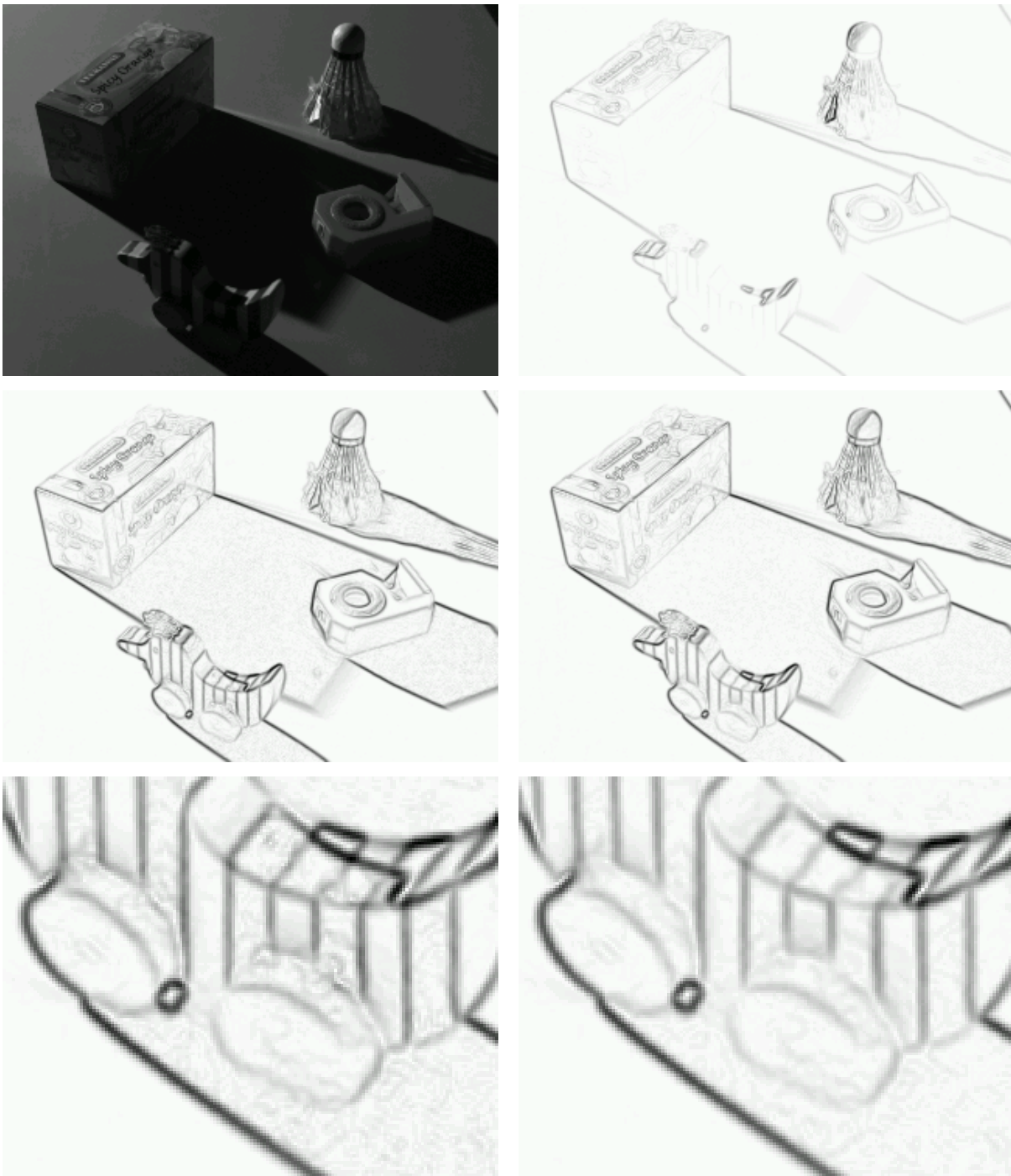


Figure 3.4: Suppression of the illumination influence on the derivatives with homomorphic processing. Top left: grey value image. Top right: gradient $\sqrt{I_x^2 + I_y^2}$. Middle left: gradient of $\ln I$. Middle right: gradient of $\ln(1 + I)$ with the proposed preprocessing. Bottom left: detail of the gradient of $\ln I$. Bottom right: detail of the gradient of $\ln(1 + I)$ with the proposed preprocessing. Gradient images are scaled between 0 and 255 using minimum and maximum gradient values. The grey value image is gamma corrected ($\gamma = 1.3$).

on the middle images, homomorphic processing compensates this effect. The proposed preprocessing reduces efficiently noise in the dark image areas as shown in the bottom images: the gradient is less influenced by noise in the black stripes and wheels of the “Tigerente” figurine.

The Harris detector based on homomorphic processing is summarised here:

1. Preprocess the dark image areas if necessary. For example, preprocess all pixels having a grey value smaller than $V = 3$ with a 3×3 box filter.
2. Take the logarithm of the preprocessed image with: $L = \ln(1 + I)$.
3. Compute the invariant derivatives by convolving the logarithm image with the derivative of Gaussian filters: $L_x = G_x(\sigma_{deriv}) \otimes L$ and $L_y = G_y(\sigma_{deriv}) \otimes L$.
4. Compute the structure matrix \mathbf{M} with eq. (3.3) and with the invariant derivatives L_x and L_y .
5. Compute the cornerness function CF with eq. (3.4).
6. (x, y) is an interest point:
 - if it is a local maximum of the cornerness function CF
 - and if $CF(x, y) > T$ ($T > 0$).

This algorithm is named homomorphic Harris detector (H-HD) in the following. As in section 3.2, the user-defined threshold T should be set to get an appropriate number of interest points. The obtained invariant derivatives can be re-used to compute invariant descriptors. The algorithm can be easily extended to other Harris detector versions, in particular to the scale or viewpoint invariant versions presented in [MS04]. Homomorphic processing can also be used to reduce illumination influence on other detectors if these are based on high-pass filtering. In that case, it should be reminded that the inaccuracies of the image formation model increase with the size of the considered neighbourhood.

The convolution with the derivatives of Gaussian and with the Gaussian is implemented as in subsection 2.2.2. The box filter for the preprocessing of dark areas is implemented straightforwardly as the neighbourhood is small (3×3). In addition, only few dark pixels are processed. Therefore, a complex implementation like recursive or sequential processing as in section 3.2 would not bring any advantage.

The results of the homomorphic Harris detector are illustrated in fig. 3.5 on the same images as in fig. 3.1. Like for N-HD, the interest points are not only detected in the areas with the highest local contrast. The detection stability is increased in comparison to the Harris detector (HD): 74.1% of the interest points of the left image are redetected in the right image. 27.1% of the points in the right image are false positives. The detection stability is similar to the stability of N-HD for this image pair. With the proposed implementation, H-HD requires 457ms for an image with 640×480 pixels (see subsection 2.2.2 for details on the computer): that is 1.11 times the processing time of HD.

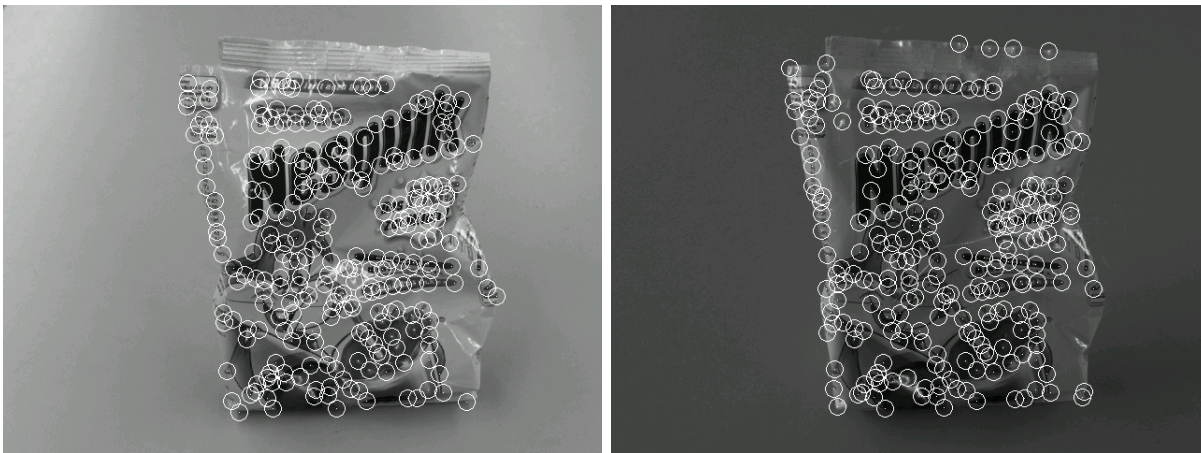


Figure 3.5: Detection example for the homomorphic Harris detector ($T = 10^{-5}$). The same images as in fig. 3.1 are used. 74.1% of the interest points of the left image are redetected in the right image. 27.1% of the interest points in the right image are false positives. Both images are gamma corrected for visualisation ($\gamma = 1.4$).

3.4 Local threshold adaptation

The next detector is based on local adaptive thresholding. It is shown in [TJ95] that even simple local adaptive thresholding methods achieve better binarisation results than global thresholding, in particular for non-uniformly lighted images. This principle is therefore adapted to interest point detection here.

As shown in eq. (3.6), the cornerness function CF is influenced by the local multiplicative illumination factor a_b^4 if the image is influenced according to the local affine model: $I = a_b c_b + a_s$ (eq. (3.1)). a_b^4 is assumed to stay constant in small image neighbourhoods. Therefore its influence can be compensated with local image characteristics like mean, median... The evaluation of local thresholding methods in [TJ95] shows that even a simple threshold adaptation with local mean and standard deviation as in [Nib86] compensates well local contrast changes. The used characteristic is calculated directly from the cornerness function CF to reduce estimation errors on the local contrast factor a_b^4 . The local mean of the cornerness function is a simple characteristic, but it is sufficient to compensate the local multiplicative illumination factor. This is shown in fig. 3.6: the ratio between the cornerness value CF and its local mean \overline{CF}

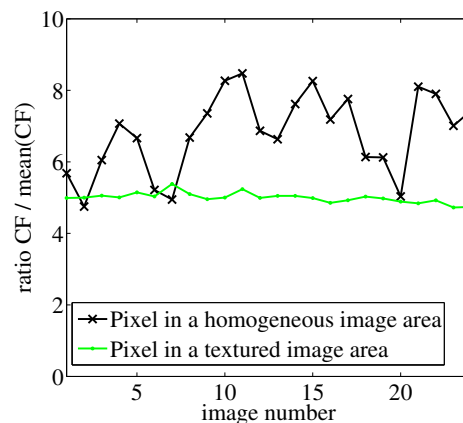


Figure 3.6: Behaviour of the ratio CF/\overline{CF} in an image series of a scene under different illuminations.



Figure 3.7: Grey value image and corresponding local standard deviation of CF estimated on a 15×15 window. The grey value image is gamma corrected ($\gamma = 1.3$). For better visualisation, the logarithm of the local standard deviation of CF is shown after normalisation between 0 and 255.

stays constant under illumination changes for textured areas. This adaptation is however sensitive to noise in homogeneous areas as shown in fig. 3.6: the ratio CF/\overline{CF} is not constant for pixels in homogeneous areas. These variations occur because CF and \overline{CF} both have very small values. Using both local mean and local standard deviation to adapt the threshold as in [Nib86] did not bring any advantages in preliminary experiments compared to only using the local mean.

To avoid detecting noise-induced interest points in homogeneous areas, a step is added to the filter in which textured areas are detected. This allows to switch off interest point detection in homogeneous areas. The spatial variations of CF are higher in textured areas than in homogeneous areas. Therefore, textured areas can be detected by thresholding the local standard deviation of CF : as shown in fig. 3.7, the local standard deviation of CF is higher in textured areas than in homogeneous areas. As the local mean of CF is used to adapt the detection threshold, the local standard deviation of CF can be computed quickly.

Fig. 3.7 shows that the standard deviation of CF in homogeneous areas varies with the grey values: the standard deviation is higher on the white wall than on the grey door or in the shadow areas in the shelf. Hence, the threshold for the detection of textured areas depends on the illumination intensity. This is due to the main noise source in modern cameras: photon noise. Photon noise can be approximated by a multiplicative noise with a standard deviation proportional to the square root of the grey values. As a result, the noise on CF in homogeneous is also approximately multiplicative. This can be verified on image series taken with a constant setup (same camera position and parameters, same scene and same illumination): noise is the only source for pixel changes between two images of the series. Calculating the standard deviation of CF over such an image series gives an estimation of the noise standard deviation on CF . This is calculated for each pixel with: $\sigma(CF) = \sqrt{\sum_{n=1}^N (CF_n - \overline{CF})^2 / (N - 1)}$, where CF_n is the cornerness value of the considered pixel for image n of the series, N is the number of images in the series

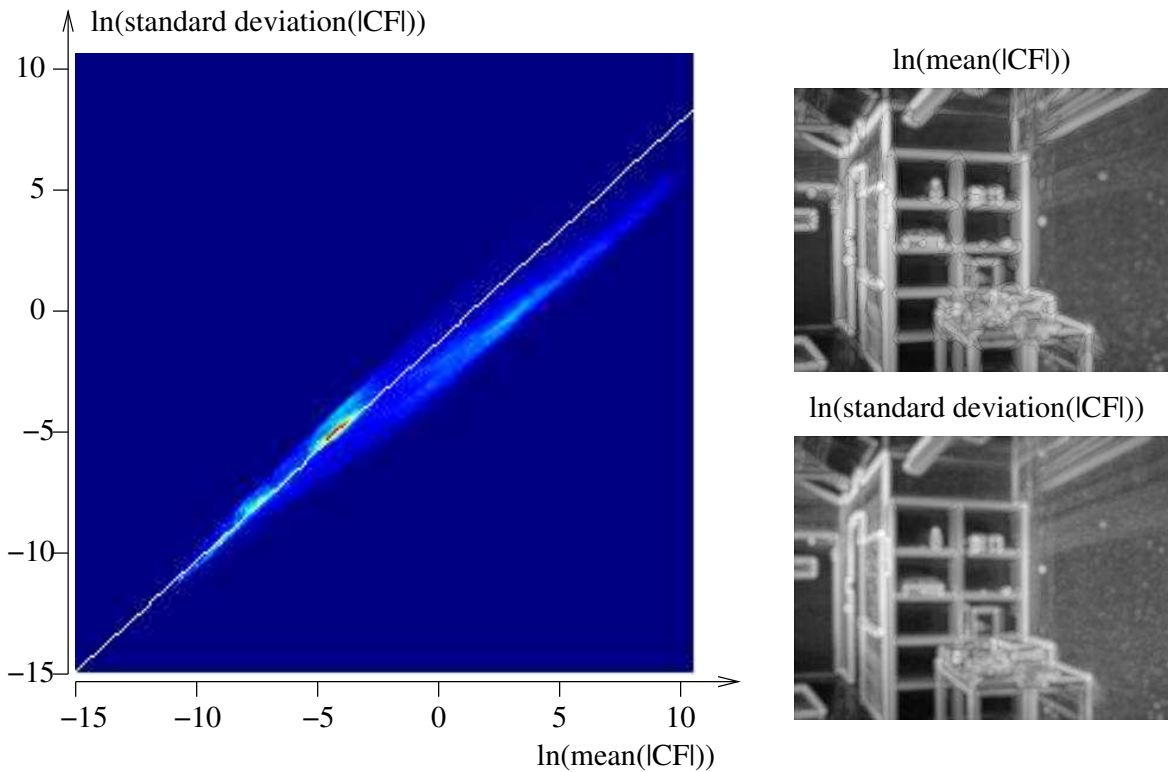


Figure 3.8: Mean and standard deviation of the noise on $|CF|$. The same scene as in fig. 3.7 is used. The 2D histogram is shown on the left. For better visualisation, false colour are used. Dark blue indicates empty histogram cells, yellow to red indicates a high number of samples. A line is fitted for the pixels in homogeneous areas (for which $|CF|$ is low). The corresponding mean and standard deviation images are shown on the right.

and \overline{CF} is the mean of CF over the image series for the considered pixel. Mean and standard deviation are estimated here for $|CF|$, because it avoids negative cornerness values and simplifies the analysis. The result is shown in fig. 3.8. For better visualisation, the logarithm of both mean and standard deviation of $|CF|$ is used. As can be seen, the relation between the logarithm of the mean and the logarithm of the standard deviation is linear. The relation is different for homogeneous and for textured areas (i.e. for low and for high $|CF|$ values). For the homogeneous areas (selected here by manually thresholding $|CF|$), a line with a slope of 0.907 can be fitted to the histogram data. Similar results are obtained for different scenes. Therefore, the standard deviation of $|CF|$ is approximately proportional to its mean. This shows that the noise on CF is approximately multiplicative.

If $|CF|$ is transformed with the logarithm, this multiplicative noise is transformed to an additive noise. The standard deviation of the noise on $\ln(|CF|)$ is approximately constant, as shown in fig. 3.9. Noise has different influences on $\ln(|CF|)$ in homogeneous and in textured areas. The noise influence is the smallest in textured areas: the standard deviation of $\ln(|CF|)$ is almost zero in these areas. In homogeneous areas, the standard deviation of $\ln(|CF|)$ is between 0.2 and 0.6. Finally, the standard deviation of $\ln(|CF|)$ is

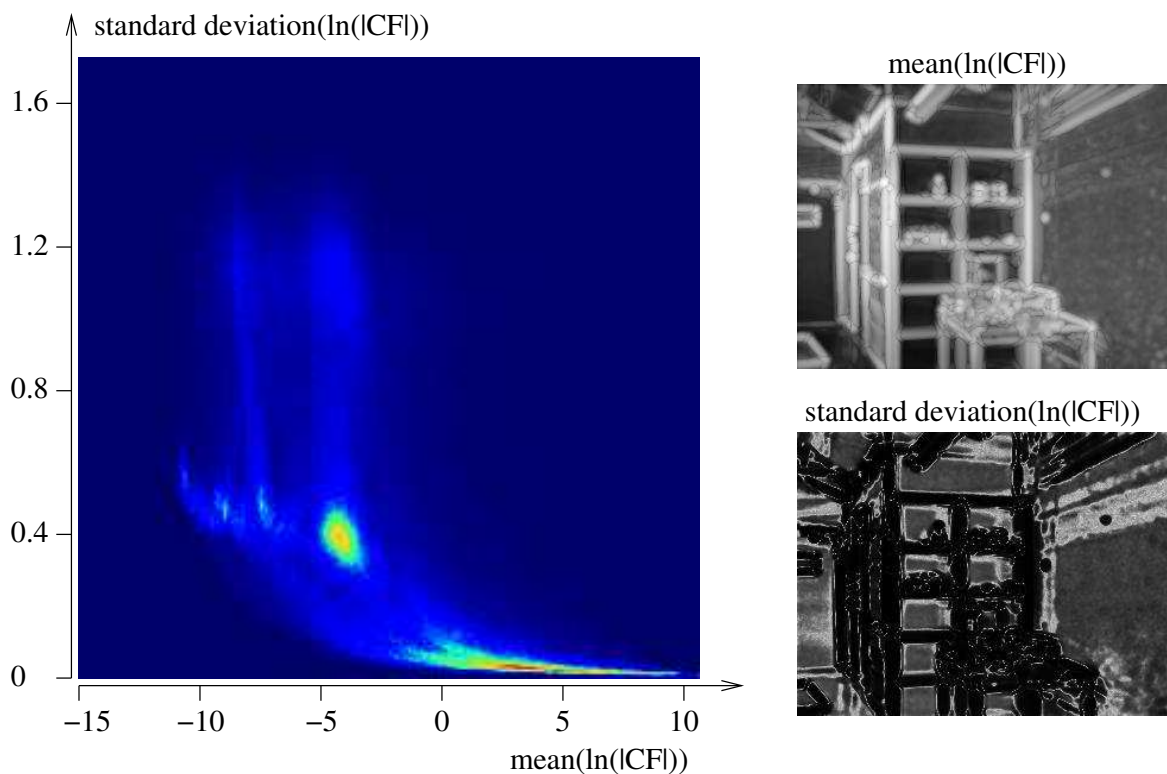


Figure 3.9: Mean and standard deviation of the noise on $\ln(|CF|)$. The same scene as in fig. 3.7 is used. The 2D histogram is shown on the left. For better visualisation, false colour are used. Dark blue indicates empty histogram cells, yellow to red indicates a high number of samples. The corresponding mean and standard deviation images are shown on the right.

the highest (approximately 1.1) in areas with a small grey value gradient, like the unsharp shadow produced by the lamp on the wall. In these areas, the cornerness function is indeed more sensitive to noise due to the existence of a gradient. As a conclusion, the noise standard deviation on $\ln(|CF|)$ does not depend on the illumination conditions. Therefore, textured areas can be detected using a fixed threshold on the standard deviation of $\ln(|CF|)$.

In the detector, the standard deviation of $\ln(|CF|)$ is estimated in image neighbourhoods because a single image is available. The result of such an estimation is shown on the left of fig. 3.10. The spatial standard deviation takes the lowest values in homogeneous areas. Its highest values are at the transition between homogeneous and textured areas. Due to the logarithm transformation, the peaks of $\ln(|CF|)$ near edges and interest points are flatter than for $|CF|$. This leads to moderate values of the spatial standard deviation in textured areas. The detection of textured areas by thresholding the spatial standard deviation of $\ln(|CF|)$ is illustrated in fig. 3.10. The threshold is set to 1.4 based on the histogram of fig. 3.9. The detection of textured areas using $\ln(|CF|)$ is not affected by illumination conditions as the standard deviation of $\ln(|CF|)$ is constant for all homogeneous areas.



Figure 3.10: Detection of the textured areas by thresholding the local standard deviation of $\ln(|CF|)$ estimated on a 21×21 neighbourhood. Left: standard deviation normalised between 0 and 255. Right: detection result. Homogeneous areas are indicated in black. For visualisation, the grey values are gamma corrected ($\gamma = 1.4$).

Textured areas are detected by thresholding the local standard deviation $\sigma(x, y)$ of $\ln(|CF|)$. The local mean $\mu(x, y)$ of $\ln(|CF|)$, which is needed to estimate $\sigma(x, y)$, is used to adapt the detection threshold to the local lighting conditions. As the logarithm is used and the illumination factor on CF is multiplicative, the local threshold adaptation is performed with an addition: $\ln(|CF|) > T + \mu$. This is illumination invariant because both $\ln(|CF|)$ and μ are influenced by the same additive illumination term. According to eq. (3.6), $CF(I) = a_b^4 CF(c_b)$, where a_b is constant in image neighbourhoods. In addition, $a_b > 0$ because it is the illumination factor on the grey values. Therefore, $\ln(|CF(I)|) = 4 \ln(a_b) + \ln(|CF(c_b)|)$. As $\ln(a_b)$ is constant on the neighbourhood used to compute the mean μ , $\mu(I) = 4 \ln(a_b) + \mu(c_b)$.

The proposed interest point detection method is summarised by:

1. Compute CF with eqs. (3.3) and (3.4) and compute $\ln(|CF|)$.
2. Compute the local mean μ and the local standard deviation σ of $\ln(|CF|)$.
3. (x, y) is an interest point:
 - if it is a local maximum of the cornerness function with $CF > 0$,
 - and if $\sigma > T_1$, (step a: detection of textured areas)
 - and if $\ln(|CF|) > \mu + T_2$. (step b: local adaptive thresholding)

This algorithm is referred to as the Adaptive Threshold Harris Detector (AT-HD) in the following. It is equivalent to a local normalisation of CF with an additional detection of textured areas. It can compensate the full affine image formation model of eq. (3.1). It has three parameters: threshold T_1 for the detection of the textured area, threshold T_2 for the illumination invariant selection of the interest points and the size W of the window used for estimating the local mean μ and standard deviation σ .

Threshold T_1 should be adjusted to the noise level on CF . It depends on the camera and on the parameters σ_{deriv} , σ_M and α in eqs. (3.3) and (3.4). To set T_1 , the histogram of the standard deviation of $\ln(|CF|)$ on an image series taken with a constant setup can be used, as it represents the noise influence. The histogram is obtained like for fig. 3.9, except that a 1D histogram of the standard deviation is computed instead of a 2D histogram of mean and standard deviation. This is shown in fig. 3.11. As in fig. 3.9, the histogram contains several peaks: the first one corresponds to textured areas and the second one to homogeneous areas. In addition, higher values of σ occur in areas with a small gradient (areas where the grey values change slowly) in which noise effects are amplified. Here T_1 is set to 1.4. The size W of the window used to compute μ and σ should be chosen such that the local *spatial* standard deviation σ in homogeneous areas matches the noise standard deviation estimated in fig. 3.11. This makes sure that the threshold chosen according to fig. 3.11 is also valid for the spatial standard deviation σ . For this, several window sizes are tested and the results are compared to the histogram in fig. 3.11. It delivers typically a range for the window size: here, windows of sizes 15×15 to 23×23 . As shown in fig. 3.10, the estimated spatial standard deviation σ is the largest at the transition between homogeneous and textured areas and decreases in textured areas because the peaks in $\ln(|CF|)$ are flatter than in CF . Therefore, if the window size for the estimation of σ is too small, only the transition areas between homogeneous and textured areas are selected as texture by thresholding. Therefore, the window size should be chosen big enough to detect the complete textured areas. The thresholding of σ is tested on the window sizes selected in the previous step (here 15×15 to 23×23) with the chosen threshold T_1 . The best compromise between consistency to noise standard deviation and thresholding performance is chosen. In this work, W is set to 21×21 . W depends on the parameters σ_{deriv} and σ_M of eq. (3.3) and on T_1 . The last parameter is threshold T_2 for adaptive thresholding. It must be positive. Like for the detectors N-HD and H-HD presented in sections 3.2 and 3.3, it should be set by the user to get an appropriate number of interest points.

The convolution with the derivatives of Gaussian and with the Gaussian is implemented as described in subsection 2.2.2. To handle areas with very small corneriness values, the logarithm is replaced by the following function:

$$f(|CF|) = \begin{cases} \ln(\epsilon) & \text{if } |CF| \leq \epsilon \\ \ln(|CF|) & \text{if } |CF| > \epsilon \end{cases},$$

where ϵ is set to 10^{-12} . The local mean and standard deviation of $\ln(|CF|)$ are computed using a recursive and sequential implementation of the box filter like for the local energy computation in section 3.2. The window size W for estimating μ and σ only has a minimal influence on the computation time of the detector thanks to the recursive implementation:

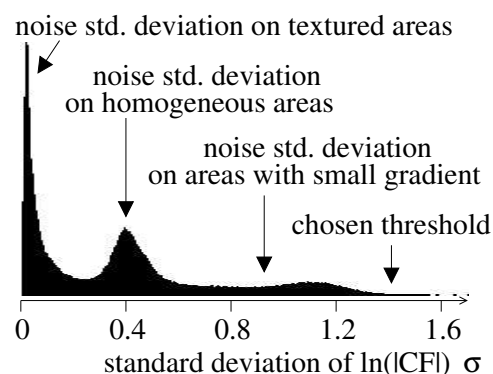


Figure 3.11: Histogram of the noise standard deviation on $\ln(|CF|)$.

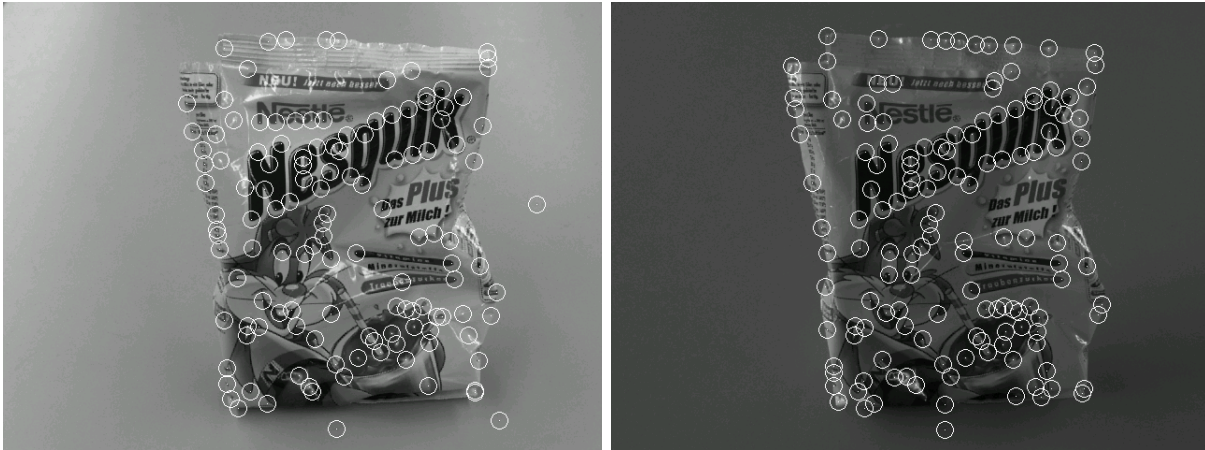


Figure 3.12: Detection example for the Harris detector with local adaptive thresholding ($T_2 = 2$). The same images as in fig. 3.1 are used. 62.8% of the interest points of the left image are redetected in the right image. 39.1% of the interest points in the right image are false positives. Both images are gamma corrected for visualisation ($\gamma = 1.4$).

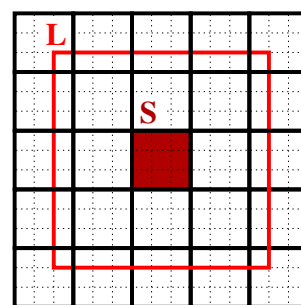
after initialisation, the box filter only requires one addition and one subtraction per pixel instead of W additions.

The results of the AT-HD algorithm are illustrated in fig. 3.12 on the same images as in fig. 3.1. The detection stability is increased in comparison to the Harris detector (HD): 62.8% of the interest points of the left image are redetected in the right image and 39.1% of the interest points in the right image are false positives. The detection stability with AT-HD is lower than with N-HD and H-HD for this image pair. As shown in fig. 3.12, the distance between neighbouring interest points is larger for AT-HD than for the preceding detectors. This is due to the large neighbourhood used for threshold adaptation. The higher sensitivity to noise in nearly homogeneous areas is also visible: some interest points are detected in the background in both images. Those interest points are not only caused by noise as small gradient exists in their neighbourhood. They are however unstable. With the proposed implementation, AT-HD requires 583ms for an image with 640×480 pixels (see subsection 2.2.2 for details on the computer), that is 1.42 times the processing time of HD.

3.5 Local clustering

The last method developed to improve the stability of the Harris detector under illumination changes is also based on local adaptive thresholding. Instead of adapting the threshold to the lighting conditions with a local characteristic like in section 3.4, the threshold is computed based on local clustering of the cornerness function CF . This algorithm is similar to the method in [ETM91], which is one of the best performing binarisation methods in the comparison presented in [TJ95].

The algorithm in [ETM91] computes a local threshold by applying the Otsu clustering method presented in [Ots79] on image neighbourhoods. Clustering results in an automatic division of the image pixels into several classes or clusters. In this work and in [ETM91], the pixels are divided into two classes: object and background. In this work, object corresponds to areas in which interest points can be detected (textured areas) and background corresponds to homogeneous areas. Clustering provides an optimal threshold to separate the two classes. To save execution time, the discretisation shown in fig. 3.13 is adopted: pixels in the large window L are considered for clustering, and the threshold is applied to the small neighbourhood S . This is performed for all small neighbourhoods S in the image (indicated by squares delimited with thick black lines in fig. 3.13).



pixels are delimited with dotted lines

Figure 3.13: Discretisation for the local thresholding.

Almost all clustering methods model each class with a Gaussian distribution. To achieve a better fulfilment of this condition, the logarithm of $|CF|$ is used. The corneriness values CF are too much spread out for being accurately modelled with a Gaussian. Histograms of $\ln(|CF|)$ in 40×40 image windows are shown in fig. 3.14 for a homogeneous and a textured area. Fig. 3.14 shows that the histogram for the textured patch has two distinct clusters. The corneriness values in homogeneous areas and the corneriness values near edges can be modelled as two classes with low and high mean values. The histogram for the homogeneous patch has only one cluster: only one class is visible.

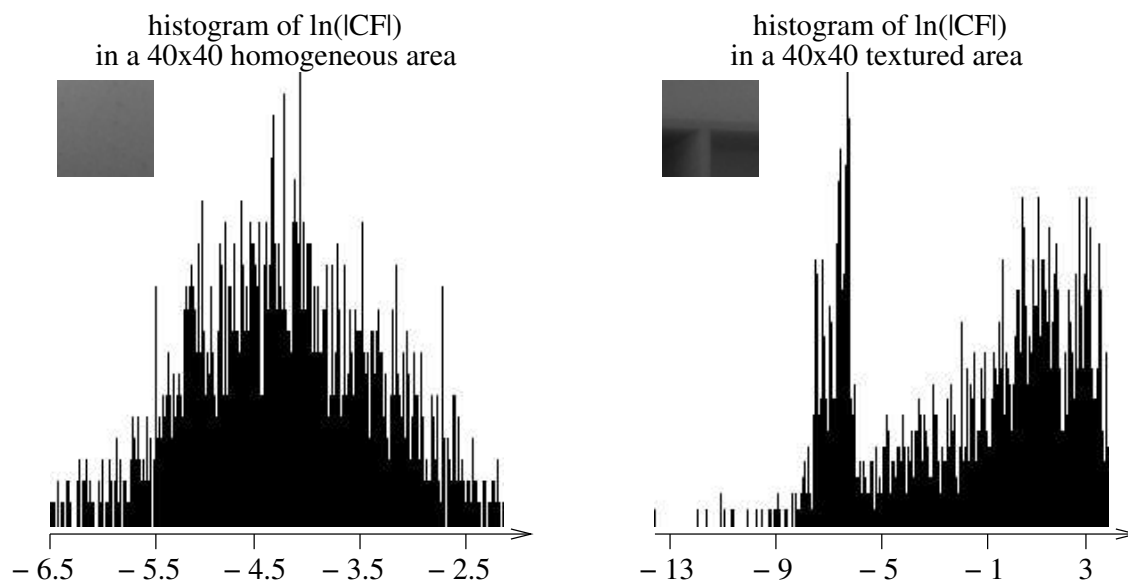


Figure 3.14: Histogram of $\ln(|CF|)$ in homogeneous and textured image patches. The corresponding image patches are shown on the top left corner of the histograms.

In this work, a faster method than the Otsu method is used: the k -means algorithm. Both methods optimise the same criterion to find the two classes and the optimal threshold,

but the k-means finds iteratively the best threshold near an initial user-defined threshold. The Otsu method on the contrary performs an exhaustive search, which requires more computing time. The k-means finds only a local optimum if the initial threshold is too far from the real global optimum. The k-means for two classes is named ISODATA. It was presented by Ridler and Calvard in [RC78]. The classes are represented by two Gaussian distributions with different means and with the same standard deviation. As visible in fig. 3.14, the two classes do not have the same standard deviation. Therefore, the threshold obtained with the ISODATA algorithm will not be optimal. Other clustering methods can handle classes with different standard deviations, for example the method by Kittler and Illingworth in [KI86]. This method estimates more parameters for the classes. Therefore, it requires more computation time and is more probable of overfitting the data when applied on a small image window instead of a whole image. To avoid overfitting, the ISODATA algorithm is used in this work, even if it does not work optimally.

The ISODATA algorithm works according to the following scheme:

1. The current threshold is initialised with the initial threshold: $T^{current} = T^{init}$.
2. Using the current threshold $T^{current}$, the pixels are separated into two classes. The means of both classes μ^{low} and μ^{high} are computed ($\mu^{low} < \mu^{high}$).
3. The threshold is updated as the average of those means: $T^{new} = (\mu^{low} + \mu^{high})/2$.
4. Steps 2 and 3 are repeated with $T^{current} = T^{new}$ until all pixels belong to a single class or until the threshold stops changing: $|T^{current} - T^{new}| < \epsilon$, where ϵ is user-defined.

The performance of the ISODATA algorithm depends on the initial threshold. As proposed in [RC78], the mean of the data is used for initialisation. To reduce the computation time, the mean of $\ln(|CF|)$ over the whole image is used to initialise the ISODATA algorithm for all considered neighbourhoods L . Although it represents the global lighting conditions, a consistent convergence of the ISODATA algorithm has been obtained for all tested images. For interest point detection, significant local maxima of CF in textured areas are searched. Therefore, it should be first verified that the area is textured. Histogram of homogeneous areas like in the left part of fig. 3.14 have a single peak. The ISODATA algorithm may deliver two classes with two close mean values, because the sum of two Gaussians is a single peak when the means are close enough. Alternatively, one of the two classes may be empty. As a consequence, interest points are only detected when no class is empty and when the two classes have distant means: $\mu^{high} - \mu^{low} > T_1$. The results of the ISODATA algorithm on $\ln(|CF|)$ are illustrated in fig. 3.15 on the scene of fig. 3.10. The textured areas are reliably detected with the used threshold $T_1 = 2.5$. The coarse discretisation by neighbourhoods S and L is visible. In the detected textured areas, significant local maxima correspond to interest points. The class with the higher mean represents the pixels near edges. Therefore μ^{high} is used to compute the local threshold. Similarly to the AT-HD method described in section 3.4, all local maxima of CF such that $\ln(|CF|) > \mu^{high} + T_2$ are detected.

Like for AT-HD, interest point detection depends only on $\ln(|CF|)$. The illumination



Figure 3.15: Results of the ISODATA algorithm on 25×25 neighbourhoods (L) of $\ln(|CF|)$. The small neighbourhood S is 5×5 wide. Left: low cluster mean μ^{low} . Middle: high cluster mean μ^{high} . Both images are scaled with the same parameters for visualisation. The areas with constant grey value are areas for which the high cluster contains no pixel. Right: detected textured area. Homogeneous areas are indicated in black. T_1 is set to 2.5 for the detection of textured areas with $\mu^{high} - \mu^{low} > T_1$.

influence on $\ln(|CF|)$ is a local additive term according to the used image formation model: $\ln(|CF(I)|) = 4 \ln(a_b) + \ln(|CF(c_b)|)$. $\ln(a_b)$ is assumed to stay constant in image neighbourhoods (see section 3.4 for more details). As $\ln(a_b)$ is constant on neighbourhood L , both cluster mean values are translated by $4 \ln(a_b)$: $\mu^{low}(I) = \mu^{low}(c_b) + 4 \ln(a_b)$ and $\mu^{high}(I) = \mu^{high}(c_b) + 4 \ln(a_b)$. Therefore, the detection of textured areas by thresholding $\mu^{high} - \mu^{low}$ is invariant to illumination changes: $\mu^{high}(I) - \mu^{low}(I) = \mu^{high}(c_b) - \mu^{low}(c_b)$. The selection of the significant interest points with $\ln(|CF(I)|) > \mu^{high}(I) + T_2$ is also invariant to illumination changes because it is equivalent to $\ln(|CF(c_b)|) > \mu^{high}(c_b) + T_2$. Due to the higher computational cost, a coarser discretisation is used than for the other developed detectors. The clustering step makes this method more flexible than the other detectors. In theory, more complex models could be handled than the simple additive model on $\ln(|CF|)$. Therefore areas where the assumptions of the image formation model (eq. (3.1)) are not completely fulfilled can be better handled than with the other detectors developed in this work. On the other hand, clustering requires the use of larger neighbourhoods in comparison to the other detectors: this increases modelling errors.

The detection algorithm is summarised in the following:

1. Compute CF with eqs. (3.3) and (3.4) and compute $\ln(|CF|)$.
2. Apply the ISODATA clustering algorithm to $\ln(|CF|)$ using all pixels in the large neighbourhoods L (see fig. 3.13) to obtain μ^{low} and μ^{high} .
3. In all small neighbourhoods S , (x, y) is an interest point:
 - if it is a local maximum of the cornerness function with $CF > 0$,
 - and if $\mu^{high} - \mu^{low} > T_1$, (step a: detection of textured areas)
 - and if no cluster is empty, (step a: detection of textured areas)

3 Illumination invariant interest point detection for grey value images

- and if $\ln(|CF|) > \mu^{high} + T_2$. (step b: local adaptive thresholding)

This method is referred to as the Local ISODATA Harris Detector (LI-HD) in the following.

The LI-HD algorithm has five parameters. Two parameters have only a small influence on stability: the size of the small neighbourhood S and the threshold ϵ used to detect the convergence of the ISODATA algorithm (see the description of the ISODATA algorithm). Both should be set to obtain a reasonable computing time. Here $\epsilon = 0.1$ and a small neighbourhood S of size 5×5 are used. The size of the large neighbourhood L depends on the Gaussian used to calculate CF (σ_M in eq. (3.3)). If the neighbourhood L is too small, the clustering does not work because not enough data are available. Therefore, L should be chosen experimentally such that each cluster (homogeneous areas for μ^{low} and textured areas for μ^{high}) has enough samples. In this work, L is set to 25×25 . Threshold T_1 depends on the parameters in eq. (3.3) and on the camera noise, because stronger noise and smaller σ_{deriv} or σ_M result in higher μ^{low} . T_1 represents the required signal to noise ratio between cornerness values in homogeneous and in textured areas. To select T_1 in this work, the distance $\mu^{high} - \mu^{low}$ was calculated on several image patches. T_1 is chosen to reduce the detection of interest points in areas having a small gradient as they are more sensitive to noise (see section 3.4). As shown in fig. 3.15, the cluster with the high cornerness values is empty in most homogeneous areas. Therefore if the maximum number of interest points should be detected for the application, T_1 can be set to 0. Its value only has a limited influence on detection stability. In this work, $T_1 = 2.5$ is used. T_2 allows to control the density of detected interest points in textured areas. It should be defined by the user to get an appropriate number of interest points.

The convolution with the derivatives of Gaussian and with the Gaussian kernel is implemented as described in subsection 2.2.2. To compute $\ln(|CF|)$, the same method as in section 3.4 is used. The ISODATA algorithm is implemented straightforwardly as in [Par93]. No optimisation such as recursive implementation is performed: the computing time reduction would be small because the iterations of the ISODATA algorithm use different thresholds even for neighbouring pixels.

The results of the LI-HD are illustrated in fig. 3.16 on the same images as in fig. 3.1. Due to the large neighbourhood used to adapt the threshold, neighbouring interest points are farther away from each other than for HD, N-HD and H-HD. Unstable interest points are detected in areas of the background with weak texture like with AT-HD. 54.2% of the interest points in the left image are redetected after illumination change. 44.8% of the interest points in the right image are false positives. The detection stability of the LI-HD is better than the stability of the Harris detector for this image pair. It is similar to the stability of AT-HD and slightly lower than the stability of N-HD and H-HD. With the proposed implementation, LI-HD requires 664ms for an image with 640×480 pixels (see subsection 2.2.2 for details on the computer): that is 1.61 times the processing time of HD.

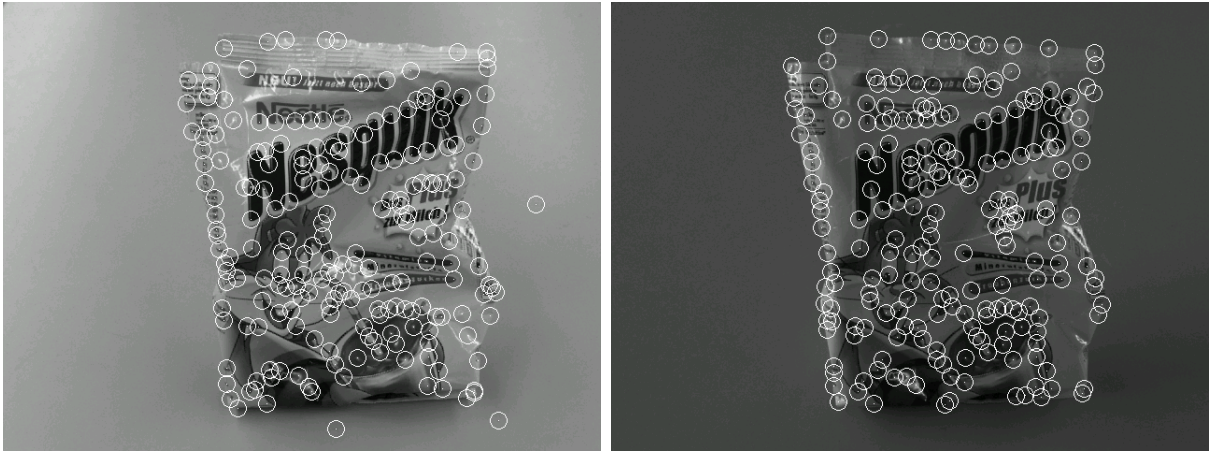


Figure 3.16: Detection example for the local ISODATA Harris detector ($T_2 = 0.5$). The same images as in fig. 3.1 are used. 54.2% of the interest points of the left image are redetected in the right image. 44.8% of the interest points in the right image are false positives. Both images are gamma corrected for visualisation ($\gamma = 1.4$).

3.6 Handling of saturated areas

Images of standard CCD or CMOS cameras contain areas with saturated pixel values, because of the high dynamics of light intensity. This is caused by visible light sources, for example the windows in fig. 3.17, or by specular highlights, for example on the posters in fig. 3.17. In those saturated areas, texture information is partially or completely lost, therefore no reliable detection can be done. Discretisation effects are emphasised at the border between saturated and non-saturated areas. This leads to the detection of interest points caused by aliasing artifacts as shown on the windows in the bottom left image of fig. 3.17. If “true” interest points are detected near these areas, they cannot be matched reliably because part of the image information in the interest point neighbourhood misses. For those reasons, saturated areas are detected and the interest points detected in their vicinity are discarded. This reduces false interest point candidates for matching. It is also important for the comparison between different detectors in section 3.7. Detection instability near or in saturated areas is caused solely by image data. It is therefore independent of the used detector. In practise, the number of interest points affected by saturated areas is however random and could therefore distort the comparison results.

The detection and handling of saturated areas is identical for all detectors. A colour camera is used for image acquisition in this work. Hence, saturated pixels are detected by thresholding the R, G and B channels. They are stored in a saturation map S :

$$S(x, y) = \begin{cases} 1 & \text{if } R(x, y) = 255 \text{ or } G(x, y) = 255 \text{ or } B(x, y) = 255 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all pixels } (x, y).$$

This thresholding is illustrated on the top left image in fig. 3.17. To avoid detection of interest points near saturated areas, the saturation map S is dilated with a simple

3 Illumination invariant interest point detection for grey value images

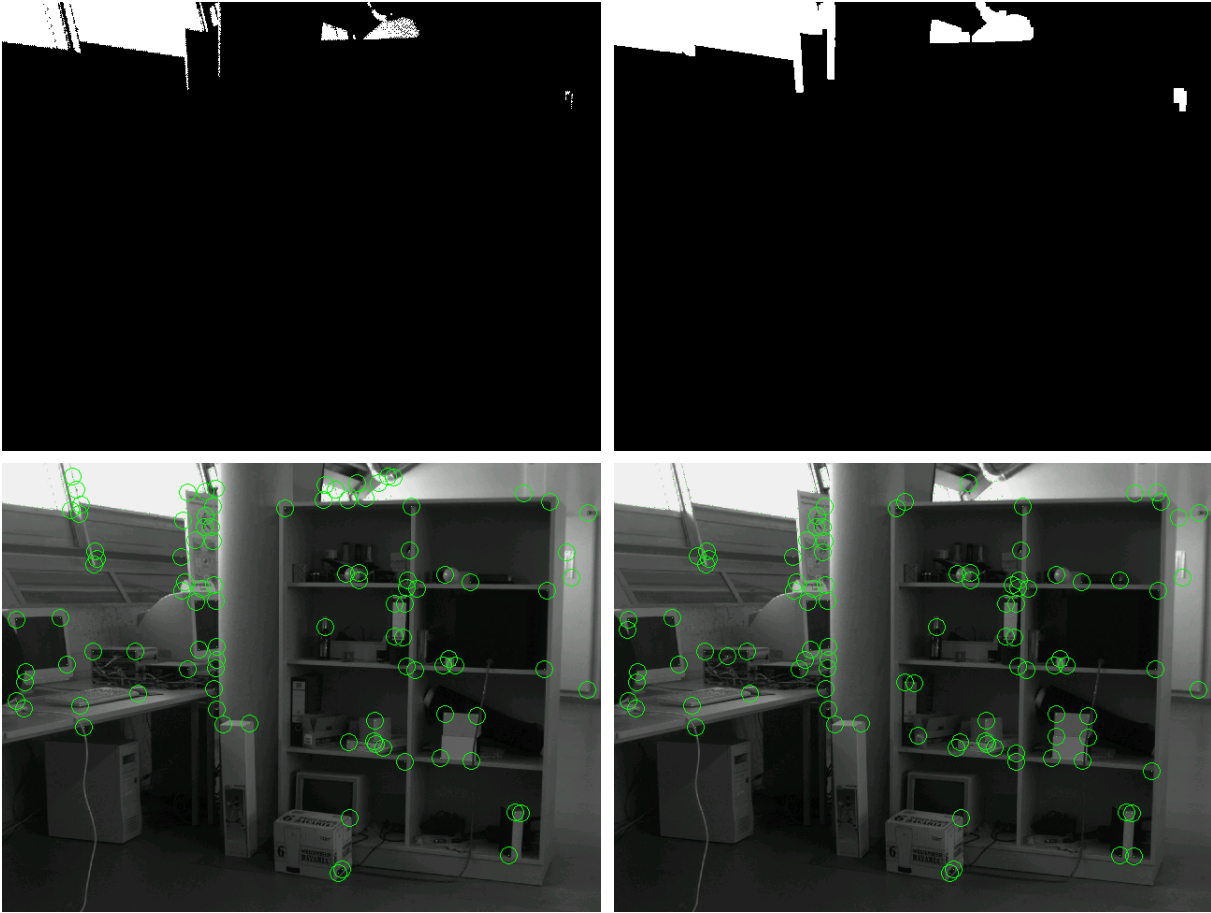


Figure 3.17: Detection and handling of the saturated areas. Top left: detected saturation map S . Top right: Saturation map after dilatation S' . Bottom left: interest points detected with the Harris detector. Bottom right: interest points detected with the Harris detector and handling of saturated areas. For both images, the 100 best interest points are detected.

7×7 square mask. The mask size must be adapted to the size σ_M of the neighbourhood considered for interest point detection (see eq. (3.3)). This dilatation is implemented with a 7×7 box filter followed by thresholding. S is a binary image, so the result of the convolution with the box filter is strictly positive if and only if one or more pixels in the neighbourhood defined by the kernel is equal to one. Similarly, the result of a dilatation on a binary image is one if and only if one or more pixel in the structuring element is equal to one. Hence, the dilatation with a 7×7 square structuring element is implemented by:

1. Filter the binary saturation map S with a 7×7 box filter. It results in the grey value image S_{box} .
2. Binarise the result S_{box} to get the final binary saturation map S' with:

$$S'(x, y) = \begin{cases} 1 & \text{if } S_{box}(x, y) > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all pixels } (x, y).$$

The resulting saturation map S' is illustrated on the top right image of fig. 3.17. All interest points detected in the areas marked by one in the dilated saturation map S' are discarded. The box filter is implemented sequentially and recursively as described in section 3.2. As illustrated in the bottom of fig. 3.17, no interest points are detected in the vicinity of saturated areas when this handling of saturated areas is applied.

3.7 Comparison framework

The stability of the developed detectors under illumination changes is evaluated and the detectors are compared to each others as well as to the original Harris detector. This section presents the comparison framework. The criteria used to measure detection stability are presented in subsection 3.7.1 after an overview of criteria used in similar works. The compared detectors and the detection parameters are given in subsection 3.7.2. Finally the image data are presented in subsection 3.7.3.

3.7.1 Quantitative evaluation of the detection stability

Most authors compare their interest point detector to other detectors with a simple visual comparison on a few test images, for example in [SB97, vdW05, Kov03]. One of the first comparisons based on quantitative evaluation of interest point detection is presented in [SMB00]. The stability of several interest point detectors is estimated using image series showing the same scene under different changes of imaging conditions (noise, camera movement and lighting conditions). In each series, one image is chosen as reference image and the interest points detected in this image are considered to be the “ground truth”. A stable detector would detect interest points corresponding to the same scene points as the reference interest points in all other images. Therefore, detection stability is measured with the repeatability rate, which is defined for image i as:

$$r_i(\epsilon) = \frac{|R_i(\epsilon)|}{\min(n_1, n_i)}. \quad (3.12)$$

$|R_i(\epsilon)|$ is the number of reference interest points redetected in image i . n_1 is the number of reference points visible in the current image (the reference image is assumed to have number one) and n_i is the number of interest points in the current image i visible in the reference image. This visibility test is necessary when the camera is moved, as the scene visible in both images may be partly different due to occlusion. To compute $|R_i(\epsilon)|$, the geometric transformation between both images is estimated and the reference interest points are projected in image i . A reference point is considered redetected if an interest point is detected in image i in a disc of radius ϵ centred on the projection of the reference

3 Illumination invariant interest point detection for grey value images

point. The repeatability rate takes values in $[0, 1]$. High stability is indicated by high repeatability rate.

The repeatability rate is adopted in other works to assess detection stability, for example in [STL⁺03, VL01]. It has also been extended or modified to assess other characteristics of interest point detectors. In [MS04, MTS⁺05], the repeatability rate is used to compare affine invariant interest point detectors. In order to better evaluate the invariance to affine image transformations, a reference interest point is considered redetected only if the neighbourhood estimated by the detectors on image i and the projection of the neighbourhood of the reference interest point have a minimum surface overlap of 40%. In [Gou00], a second measure is added to the repeatability rate to evaluate detection accuracy. Instead of only counting the number of redetected points, the distances between the projections of the redetected reference points and the corresponding points detected in image i are accumulated. Smaller values indicate a more accurate redetection.

The repeatability rate has one main drawback: it only considers the redetected interest points (the true positives), but it does not take into account the number of false positives, which are the interest points detected in the current image which do not correspond to any reference interest point. Therefore, decreasing the detection threshold to obtain more interest points generally increases the repeatability rate, even though the number of noise sensitive and unstable interest points is increased. Therefore other measures are used in [CJ02, MM01] to assess both false positives and true positives. This is realised by considering the variation of the number of detected interest points between images. Ideally, this number should not vary. In addition, the repeatability rate is replaced by the true positive rate (redetection rate) in [CJ02]. This is the ratio of the number of redetected reference interest points to the number of reference interest points visible in the current image: $|R_i(\epsilon)|/n_1$ with the notations of eq. (3.12). In [MM01], the detected interest points are compared to a ground truth based on human judgement: the reference image is replaced by a man-made map of “true” interest points in the scene.

Finally, interest point detection is in general only an intermediate step in an application and it is most of the time followed by matching. Therefore, matching scores can also be used as a criterion to compare interest point detectors, like for example in [MTS⁺05] or in [GMD⁺97]. Such an application-based evaluation is performed in chapter 5 of this thesis.

To take into account both true positives and false positives, redetection and false positive rates are used in this work to evaluate the interest point detectors, like in [CJ02]. The ground truth is the interest points detected in a reference image, as it results in a better estimation of detection stability than a man-made interest point map. As explained in section 3.6, no detection is performed in detected saturated areas. Therefore saturated areas must be taken into account to estimate redetection and false positive rates. This yields a fairer comparison between detectors as discussed in section 3.6. This is performed similarly to the visibility test in the repeatability test. The redetection rate $r'_i(\epsilon)$ and the false positive rate $fp_i(\epsilon)$ for the current image i are defined by:

$$r'_i(\epsilon) = \frac{|R_i(\epsilon)|}{n'_1} \text{ and } fp_i(\epsilon) = \frac{n'_i - |R_i(\epsilon)|}{n'_i}. \quad (3.13)$$

$|R_i(\epsilon)|$ is the number of reference interest points redetected in image i like in eq. (3.12). n'_1 is the number of reference points which do not project in a saturated area in image i . n'_i is the number of detected interest points in image i which do not project in a saturated area in the reference image. This work is focused on illumination changes. Therefore, the images used for comparison present the same scene under different illumination conditions. There is no camera movement, so the projection between reference and current images is the identity transformation: corresponding pixels have the same coordinates in both images. Each reference point is therefore considered redetected if any interest point is detected in the current image i in a disc of radius ϵ centred on the reference point. To take into account noise influence on the position of the interest points, ϵ is set to 1.5: all points detected in the 3×3 neighbourhood of a reference interest point contributes to $|R_i(\epsilon)|$. Redetection and false positive rates sums up to one only when $n'_1 = n'_i$. This is only the case when a constant number of interest points is detected per image and when the images contain no saturated areas.

3.7.2 Compared interest point detectors

The performances of all developed interest point detectors are evaluated. In addition they are compared to an existing interest point detector: the Harris detector. This allows to evaluate the stability improvement yielded by the different developed methods, because all evaluated detectors are built on the same principle and only the adaptation of detection to the lighting conditions is different. No other existing interest point detector is included in the evaluation because it is shown in [MTS⁺05] that interest points based on different principles have different performances on different image types (structured or textured images, ...). This makes the comparison of detectors built on different principles as in [MTS⁺05] difficult.

As explained in subsection 2.3.1, the detection with the Harris detector can be adapted to the global lighting conditions by either selecting the N interest points with the highest cornerness values or by using a detection threshold proportional to the maximum cornerness value in the image. It is shown in [Fai03a] that the selection of the N best interest points yields more stability under illumination changes. Therefore, the Harris detector used for comparison to the developed methods selects the N points with the highest cornerness values. An overview of the five interest point detectors is given in table 3.1.

detector name	abbreviation	description	computing time
Harris detector	HD	section 2.2.2	412 ms
energy normalised Harris detector	N-HD	section 3.2	480 ms
homomorphic Harris detector	H-HD	section 3.3	457 ms
adaptive threshold Harris detector	AT-HD	section 3.4	583 ms
local ISODATA Harris detector	LI-HD	section 3.5	664 ms

Table 3.1: Overview of the evaluated interest point detectors.

3 Illumination invariant interest point detection for grey value images

For all detectors, the handling of saturated areas described in section 3.6 is applied. The following parameters are used for all detectors to compute the cornerness values: $\sigma_{deriv} = 1.2$, $\sigma_M = 3$ and $\alpha = 0.06$. All parameters of the developed methods are set to the values given in the corresponding sections, except the user-defined thresholds which control the number of detected interest points. These thresholds are set to avoid detecting noise-induced interest points and to detect enough interest points on all used scenes. As the used scenes have different complexities, a compromise must be found. The user-defined threshold has however for all developed methods only a limited influence on the detection stability. For HD, $N = 100$ interest points are detected. For N-HD, T is set to 0.05. For H-HD, T is set to 10^{-5} . For AT-HD, T_2 is set to 2. For LI-HD, T_2 is set to 0.5. Those values are used for all experiments presented in section 3.8.

3.7.3 Image data set

Several image series are used to evaluate detection stability under illumination changes. Each series shows one scene under different illumination conditions. The images are acquired with a BASLER A302fc colour CCD camera. No gamma correction and no white balancing is performed. Gain and brightness (multiplicative gain and additive offset applied to the electric signal before A/D conversion in the camera) are set to the values given by the manufacturer as it is the optimal operating point of the electronics: it minimises electronic noise and it maximises linearity between light and pixel values. Only aperture and shutter time of the camera are changed. Those two parameters are set manually to avoid large saturated areas in the images and to obtain the best possible pixel value histogram. Some large dark areas may however appear in some images as a result of the sensor linearity, especially when the illumination is not diffuse as it leads to very dark shadows. The raw camera signal is used and demosaiced with the algorithm in [LT03], as explained in chapter 4. The grey value images are obtained as the Y component of the YUV colour space: $Y = 0.3C^R + 0.59C^G + 0.11C^B$ (see [Tec01]). This results in images equivalent to images acquired directly with a one-channel CCD camera. The advantage is that the detectors based on colour information presented in chapter 4 can be evaluated on the same image data. In each series, one image is chosen as reference image to compute redetection and false positive rates as given in subsection 3.7.1.

The first image series on which the detectors are evaluated are images series with simple illumination changes. The variation is the same for all pixels in the image. As a result, the original Harris detector can compensate these illumination changes. All detectors should therefore reach good detection stability. In the first image series, there is no illumination variation at all. Noise is the only cause for pixel value variations. The second image series is lighted with neon lamps. Therefore pixel variations are caused by neon lamp flickering and by noise. For both series, the choice of the reference image is not important because all images are very similar. The first image is therefore chosen as reference. In the last series, the shutter time of the camera is linearly increased while the aperture stays constant. The reference image is the image in the middle of the series to test the stability for both intensity increase and decrease. These three image series show the same



Figure 3.18: Two images of the series with shutter time variations. For visualisation, the images are gamma corrected ($\gamma = 1.4$).

scene. Two images of the intensity variation series are shown in fig. 3.18. For these three series, the scene geometrical and reflectance properties are not important as they have no influence on the pixel variations under these simple illumination changes.

The other image series present scenes under complex illumination changes. The illuminant type is varied. Realistic illuminants are used: natural light, neon lamps and tungsten halogen lamps. For the tungsten halogen lamps, umbrellas can be added to obtain a more diffuse light. Natural light is approximately white with the fixed white balancing settings of the camera, while neon lamps produce yellow light and tungsten lamps produce red light. Position, orientation and number of light source(s) are also changed in the series. The image with the “best” lighting conditions is selected manually as the reference image in each series: it is the image with the most uniform lighting. For complex illumination changes, scene properties have a big influence on the changes in the images. In scenes with complex 3D geometry, the influence of shadows and shading is higher than in scenes with simple geometry (for example planar scenes). Similarly the reflectance properties of the scene are important as they influence the amount of specular effects in the image. Indeed, specularities cannot be compensated by N-HD and by H-HD because both methods are based on the local multiplicative image formation model of eq. (3.2). The detectors have been tested on scenes with various properties. Two realistic scene types are used. First, typical indoor scenes are used, showing places in our laboratory. Second, typical scenes for object recognition are also used, showing one or a few objects in foreground with an approximately homogeneous background. Detection stability is similar for these two scene types. It depends much more on the object or place properties (such as simple/complex 3D geometry or diffuse/specular surfaces) than on the scene type. Scenes with similar properties result in similar detection stability. Therefore typical image series have been selected out of all acquired series and series with redundant results are not presented in this thesis. For each of the selected series, the reference image and two characteristic images are shown in fig. 3.19. The first three series (*nesquik*, *paper* and *calendar*) present structured scenes with decreasing 3D complexities and specular surfaces. The next two series (*shelves* and *rabbit*) show scenes with complex 3D geometry and diffuse surfaces. Their reflectance is however more complex so that they are textured scenes (they have many interest points with similar cornerness values). The *shelves* series also presents the effect of shadows: most images contain many sharp or soft shadows.

3 Illumination invariant interest point detection for grey value images

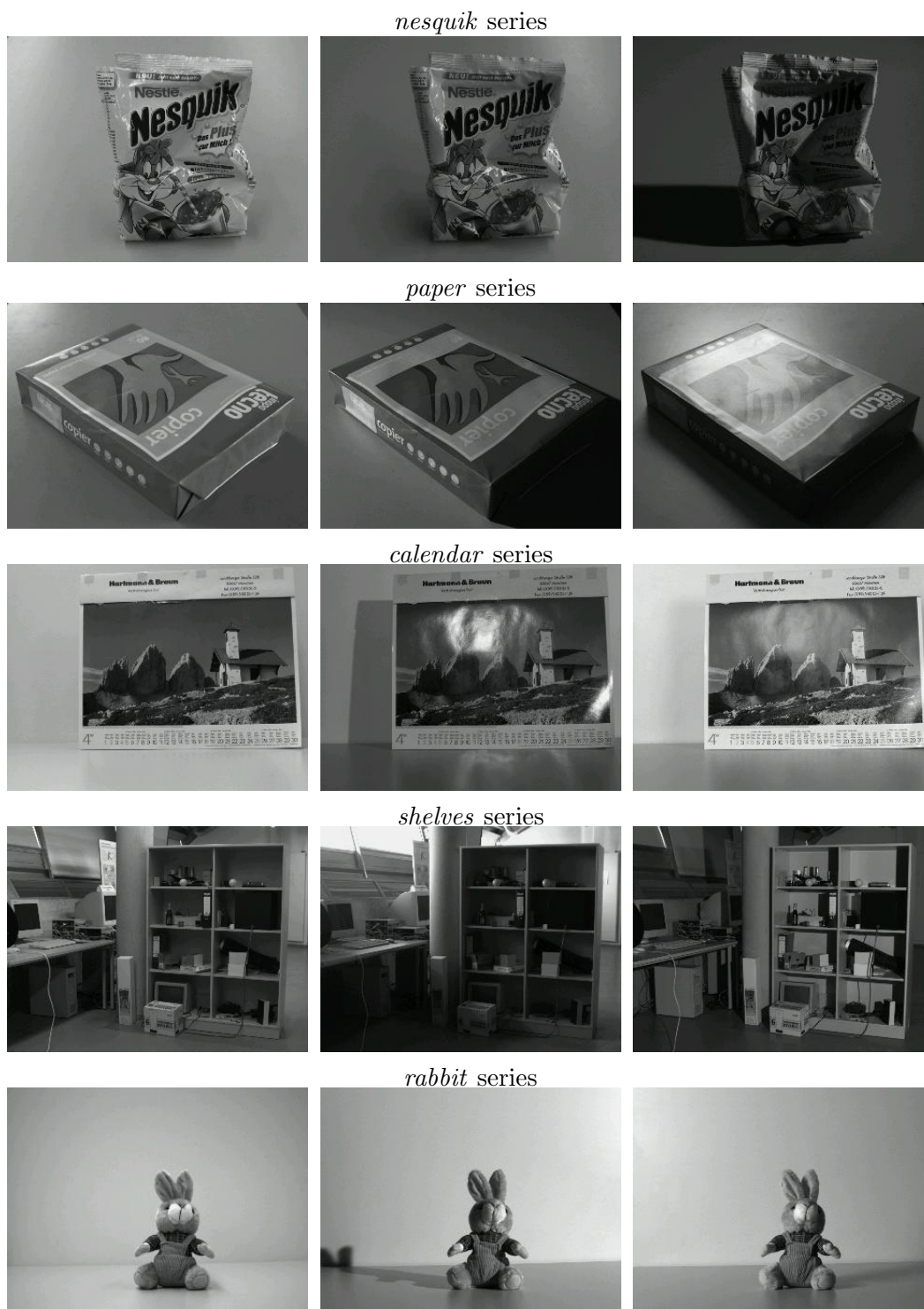


Figure 3.19: Sample images of the series with complex illumination variations. For all image series, the reference image of the series is shown on the left. For visualisation, the images are gamma corrected ($\gamma = 1.4$).

3.8 Detector evaluation and comparison

The results of the detector evaluation are presented in this section. In subsection 3.8.1, the detectors are evaluated on image series with simple illumination changes. In subsection 3.8.2, the results for the image series with complex illumination changes are presented. Finally, a conclusion is given in subsection 3.8.3.

3.8.1 Simple illumination changes

Image series with noise as only source for pixel variations					
	HD	N-HD	H-HD	AT-HD	LI-HD
mean redetection rate	0.983	0.974	0.980	0.933	0.923
mean false positive rate	0.0247	0.0279	0.0388	0.0647	0.0787

Image series with neon flickering as main source for pixel variations					
	HD	N-HD	H-HD	AT-HD	LI-HD
mean redetection rate	0.971	0.977	0.989	0.942	0.936
mean false positive rate	0.0542	0.0133	0.0506	0.0670	0.0554

Table 3.2: Evaluation results for the series with small illumination changes. In the first series, no illumination changes occur so that noise is the only source for pixel variations. In the second series, neon flickering is the main source for pixel variations. The mean rates for the whole series are given (50 images). The scene is shown in fig. 3.18.

The evaluation results for the image series with simple illumination changes are shown in table 3.2 and in fig. 3.20. In the first series, no illumination change occurs. Noise is the only source for pixel variations. The mean redetection and false positive rates for all tested detectors are given in the top of table 3.2. HD, N-HD and H-HD all have similar stability. The small differences of redetection and false positive rates for these three detectors are not significant. AT-HD and LI-HD have a slightly smaller stability for this scene. The detection thresholds have been selected based on many images. The lower stability of AT-HD and LI-HD are caused by a mismatch between detection threshold (T_2) and scene content. Many interest points have a similar corneriness value and the detection threshold selects only part of these similar interest points. As a result, small changes of the corneriness values due for example to noise influence the redetection of many interest points, decreasing hence detection stability. When the AT-HD and the LI-HD detectors are evaluated on similar image series with other scenes, redetection rates are in the range $[0.95, 0.97]$ and false positive rates are in the range $[0.02, 0.06]$, like for the other three detectors.

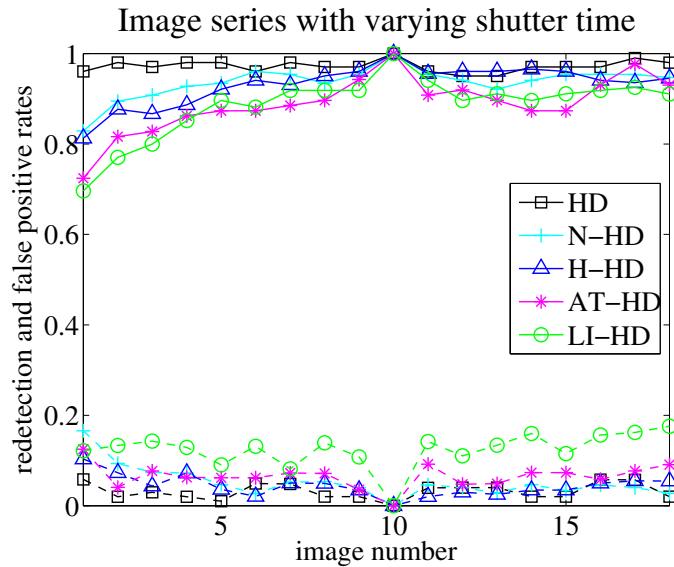


Figure 3.20: Evaluation results for the series with shutter time variations. Redetection rates are indicated with straight lines. False positive rates are indicated with dashed lines. The mean grey value for the images varies from 14.5 on the left for image 1 to 66.8 on the right for image 18. The scene is shown in fig. 3.18.

The second image series shows the same scene illuminated by neon lamps. Neon lamp flickering is the main source for illumination changes. The results are given in the bottom of table 3.2. As before, HD, N-HD and H-HD have similar stability and AT-HP and LI-HD have slightly lower stability due to the used threshold (the same scene is used). The mean redetection rates and the mean repeatability rates are similar for both series shown in table 3.2. This shows that neon flickering is well compensated by all detectors. For comparison, when a fixed threshold is used for the original Harris detector, the mean repeatability rate is 0.933 and the mean false positive rate is 0.0640 (the threshold has been chosen manually to obtain 100 interest points in the reference image).

The last image series with simple illumination changes shows the same scene under varying illumination intensity. This is obtained by varying the camera shutter time. The results are presented in fig. 3.20. The stabilities of all detectors are in the same range. The differences in repeatability and false positive rates are caused mainly by the threshold values. When HD is applied with $N = 200$, the curves for the repeatability and false positive rates are between the curves for N-HD and H-HD. This shows that HD, N-HD and H-HD have similar stability. Like for the first two other series, AT-HD and LI-HD have slightly lower stability due to the chosen detection threshold. Fig. 3.20 shows that all developed detectors (N-HD, H-HD, AT-HD and LI-HD) are less stable in dark images: the redetection rates decrease for images with low mean grey values (small image numbers). Those detectors are all more sensitive to noise in dark images because the values used to adapt the detection to the local illumination conditions (for example the local grey value energy or the local mean cornerness value) are low, hence noise sensitive.

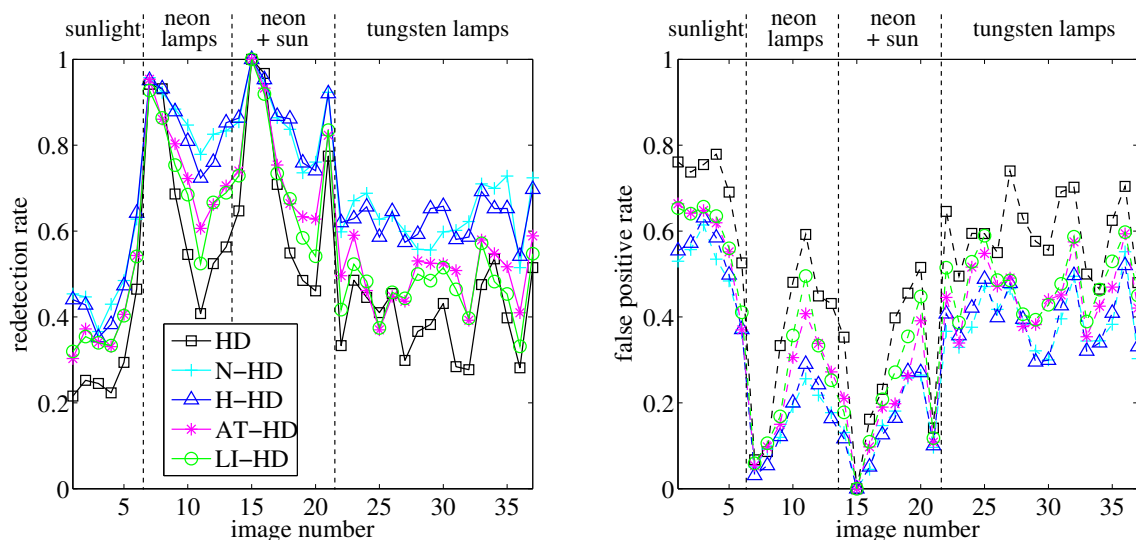


Figure 3.21: Evaluation results for the *nesquik* series (see fig. 3.19). The redetection rate is given in the left diagram and the false positive rate is given in the right diagram. The illumination type for every images is indicated at the top of the diagram.

As a consequence, the adaptation process is more sensitive to noise. Similar results are obtained on other scenes, as shown for example in [Fai03a, Fai03b, Fai04b].

All illumination changes handled in this subsection are simple illumination changes, which have the same influence on all pixels. Therefore, the original Harris detector can compensate these changes when the threshold is adapted using the maximum cornerness value or when the N interest points with the highest cornerness values are selected. It is shown here that the interest point detectors developed in this work achieve similar stability. The noise sensitivity of the new detectors is slightly higher on dark images than for the Harris detector. In addition, the results for AT-HD and LI-HD shows that detection stability is slightly decreased when detection threshold and scene content mismatch.

3.8.2 Complex illumination changes

The evaluation results for all detectors under complex illumination changes are presented in this subsection. Sample images of the series are shown in fig. 3.19. Type, number, position and orientation of the light sources vary from one image to another (see subsection 3.7.3). The reference image is selected manually and is approximately uniformly lighted.

Typical results for a standard scene are presented in fig. 3.21. As explained in subsection 3.7.3, the scene properties have a strong influence on detection stability. This scene has a 3D geometry of middle complexity. The object reflectance is structured and it contains both Lambertian and specular components (see fig. 3.19). As shown in fig. 3.21, HD yields the lowest stability: it has the lowest redetection rate and the highest false positive

rate for most images. The detected interest points are indeed located in the areas with the highest local contrast. The position of these areas in the scene varies due mainly to changes of the light source position and orientation. A local adaptation of the detection to the lighting conditions improves detection stability: all developed detectors achieve better results. N–HD and H–HD provide the best stability. Their performance is similar. AT–HD and LI–HD also have similar performances. Their stability is lower than the stability of N–HD and H–HD. For the new detectors, unstability is mainly caused by specular highlights, which have a high cornerness and change position when the light source or the camera moves. The filtering of saturated areas suppresses some of those interest points, but not all specular highlights are saturated. Fig. 3.21 also shows that a change of the illuminant type has a stronger influence on the image and hence on the detection stability than a movement of the light source or a different number of light sources. This can be explained by two effects. First, a change of the illumination colour cannot be compensated correctly for grey values especially when the neighbourhood contains more than one colour. In addition, the illumination properties (diffuse or directed light...) have a strong influence on shadows: diffuse illumination reduces the contrast between shadows and directly lighted areas, so it prevents strong spatial variations of the light intensity in the scene. The difference between diffuse and point light source illumination is visible in fig. 3.19 for the *shelves* series. The most diffuse illumination is achieved here with neon lamps or sunlight depending on the time of day.

To verify that the developed detectors better compensate local lighting changes (in opposition to global lighting changes which have a similar influence on all pixels), the following complexity measure CM is introduced:

$$CM(I_1, I_2) = CM(I_2, I_1) = \sigma\left(\frac{I_1 - \mu_1}{\sigma_1} - \frac{I_2 - \mu_2}{\sigma_2}\right). \quad (3.14)$$

$CM(I_1, I_2)$ characterises the complexity of the illumination variation between the two grey value images I_1 and I_2 . Both images should show the same scene like in the used image series. μ_i and σ_i are the mean and standard deviation of the grey values of image I_i . $(I_i - \mu_i)/\sigma_i$ is image I_i after zero-normalisation: the mean of the grey values becomes zero and their standard deviation becomes one. This normalisation compensates the global illumination influence: $I = a_b c_b + a_s$ where a_b and a_s are identical for all pixels (see section 3.1). CM is the standard deviation σ of the difference between the two zero-normalised images. If images I_1 and I_2 are related by a global lighting change, the difference between the zero-normalised images is approximately zero and $CM = 0$. Hence, CM indicates if an adaptation to the overall lighting condition is enough for a stable detection.

The stability of HD should decrease when CM increases. For the image series with varying shutter time in fig. 3.20, the complexity measure CM between two images takes values of about 0.05 as a result of noise and saturation. The redetection and the false positive rates for the *nesquik* series (see fig. 3.21) are presented as a function of the complexity measure between reference image and current image in fig. 3.22. CM varies between 0 and 1.5 for this series. This proves the necessity to consider the local lighting conditions for complex lighting changes. Fig. 3.22 shows that the stability of the developed detectors

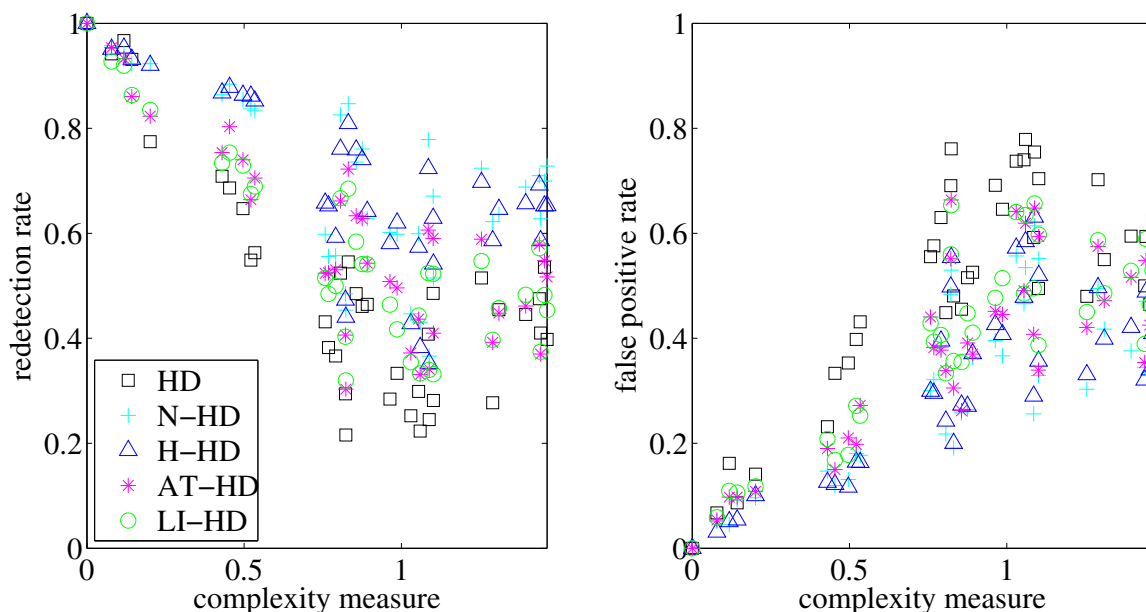


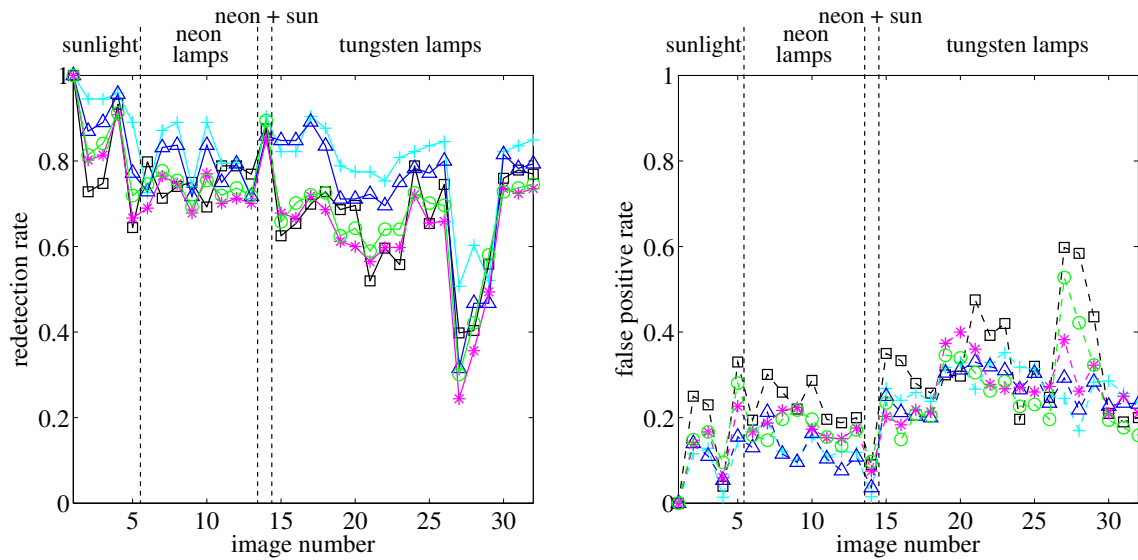
Figure 3.22: Evaluation results for the *nesquik* series (cf. fig. 3.19) depending on the complexity measure CM . The redetection rate is given in the left diagram and the false positive rate is given in the right diagram.

decreases less strongly than the stability of the original Harris detector (HD) when CM increases. Therefore, the developed detectors achieve a better compensation of the local lighting conditions. The developed detectors are hence more useful for complexer lighting changes. As before, H-HD and N-HD achieve the best results.

The detection stability increases for all detectors when the scene 3D geometry is simpler. This is shown in fig. 3.23 which presents the evaluation results for a rectangular object (*paper* series) and for a planar object (*calendar* series). For all detectors, the redetection rate is higher for *paper* than for *nesquik* and for *calendar* than for *paper*. The false positive rate is also smaller for *paper* than for *nesquik* and for *calendar* than for *paper*. The reason for this is that shadows and shading have less influence: the orientation of the surface normals varies less (see eq. (2.1)) and there are less shadows. As a result, the stability increase yielded by the developed detectors in comparison to HD is smaller for simpler 3D geometry, as shown in fig. 3.23. For simple 3D geometry, the illumination influence becomes similar for all pixels. This leads to a better performance of HD. The stability of the developed detectors also increases because they only compensate shadows and shading with slow spatial variations (a_b and a_s in eq. (3.1) are assumed to vary slowly). For scenes with large planar surfaces, large specular areas may appear for some illumination directions as shown for the right image of *paper* in fig. 3.19. These large specular areas may prevent the redetection of interest points, as shown by the drop in the redetection rate for the *paper* series for images 27 to 29 and for the *calendar* series for images 1, 4 and 11 to 14. N-HD and H-HD are particularly sensitive to such large specular areas as they do not compensate specularities: almost no interest points are

3 Illumination invariant interest point detection for grey value images

paper series



calendar series

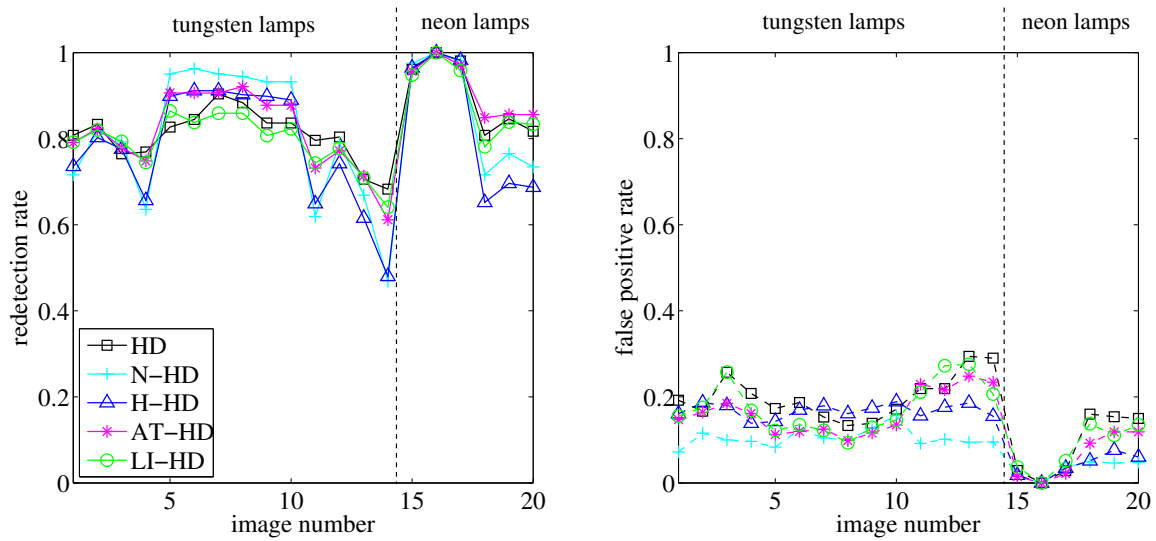


Figure 3.23: Evaluation results for the *paper* series (top) and for the *calendar* series (bottom) (see fig. 3.19 for sample images). The redetection rate is given in the left diagram and the false positive rate is given in the right diagram. The illumination type for every images is indicated at the top of the diagrams. The same legend is used in both diagrams.

detected in specular areas. AT–HD and LI–HD detect a reduced number of interest points, so redetection is less influenced. On the other hand, they detect interest points at the border of specular areas when the spatial transition is abrupt enough and when the grey values in the specular areas are not saturated. As a result, HD, AT–HD and LI–HD have a higher false positive rate than N–HD and H–HD when large specular areas occur. For these two scenes with simple 3D geometry, N–HD and H–HD achieve slightly better stability than HD, AT–HD and LI–HD which perform similarly.

The results of the detectors for textured scenes are presented in fig. 3.24. In both scenes, the reflectance is more complex than for the previous scenes, as shown in fig. 3.19. As a consequence, many interest points have similar cornerness values. This reduces detection stability: for all detectors, the repeatability rate for the *rabbit* and *shelves* series is smaller than for the *nesquik*, *paper* and *calendar* series. The false positive rate is also higher. A higher stability could be reached by adapting the detection threshold to the scene content. The *rabbit* series has a 3D geometry of similar complexity to the *nesquik* series. The scene is also less specular. The main causes for instability are hence self-similarity in the object texture and shadow and shading effects which cannot be completely corrected. Like for the previous series, N–HD and H–HD yield the most stable detection. AT–HD, LI–HD and HD all achieve similar stability. The *shelves* scene has the most complex 3D geometry. As a consequence, shadow and shading effects have higher influence. This is particularly visible when the scene is lighted with tungsten lamps as they provide a less diffuse light. As a result, many sharp shadows appear in the image and some shading edges have high contrast. This is visible on the right image of the *shelves* series in fig. 3.19. This results in lower stability for all detectors because corners are detected near shading or shadow edges, which change with the light source position and orientation. The stability of the developed detectors is better than the stability of HD because at least the slowly varying part of shadows and shading can be compensated. The achieved detection stability is however low. To compensate sharp shadow or shading edges, colour images are required as explained in section 2.1. For the *shelves* series, N–HD and H–HD perform best, followed by AT–HD and LI–HD. HD achieves the worst stability.

3.8.3 Conclusion

All tested detectors (HD, N–HD, H–HD, AT–HD and LI–HD) yield similar high stability when illumination changes affect all pixels identically. This occurs for all kinds of scenes under simple illumination changes and for scenes with planar geometry under all illumination changes. For complex illumination changes and scenes with moderate or complex 3D geometry, the developed detectors (N–HD, H–HD, AT–HD and LI–HD) are more stable than the original Harris detector (HD) because they can better compensate shadow and shading effects. In addition, the stability improvement increases when the scene geometry or the illumination changes become complex. The best stability is obtained by N–HD and H–HD, which both have similar performances. As H–HD is faster than N–HD, H–HD should be preferred. AT–HD and LI–HD both have similar performances and yield the second best stability. AT–HD is less computation intensive than LI–HD and should

3 Illumination invariant interest point detection for grey value images

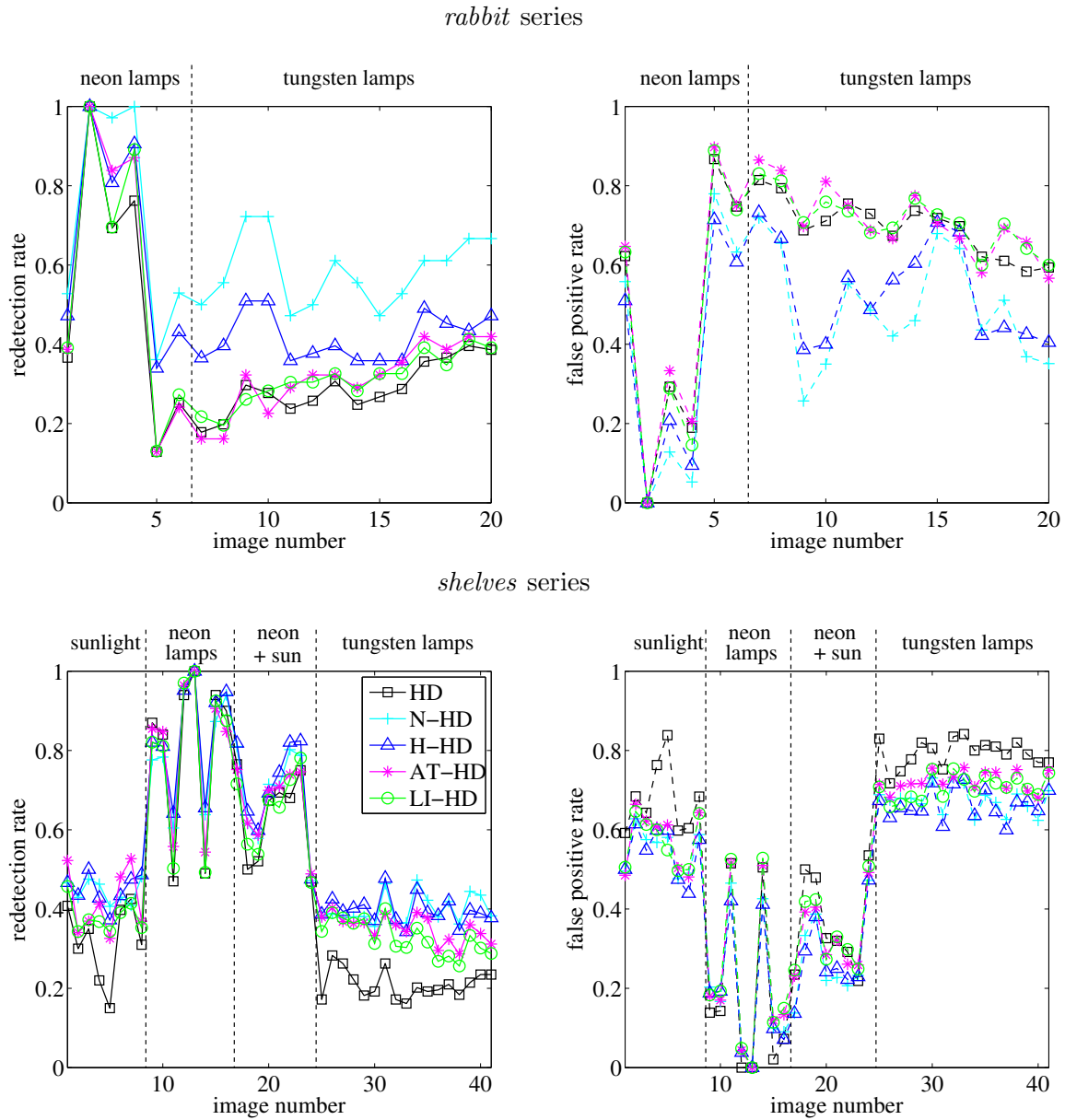


Figure 3.24: Evaluation results for the *rabbit* and for the *shelves* series (cf. fig. 3.19). The redetection rate is given in the left diagram and the false positive rate is given in the right diagram. The illumination type for every images is indicated at the top of the diagram. The same legend is used in both diagrams.

therefore be favoured. The interest points detected with AT–HD are farther away from each other and more uniformly distributed in the image than with H–HD or N–HD. The behaviour of AT–HD and LI–HD in specular areas is also different from the one of H–HD and N–HD. Therefore AT–HD could be interesting for some applications, for example localisation (see chapter 5), even though its processing time is longer and even though it provides less stable interest points than H–HD or N–HD.

The developed detectors remain sensitive to sharp shadow patterns and to highly contrasted shading edges, as only shadow and shading components with low spatial frequencies can be modelled and corrected with a single grey value image (see section 2.1). Therefore, diffuse light yields better stability. High frequency shadow and shading components can be compensated when colour images are used. Such methods are presented in chapter 4. Specularity is a further phenomenon which is not accurately handled by the developed detectors. When specular areas occur due for example to planar surfaces, the redetection rate for N–HD and H–HD decreases because less interest points are detected. AT–HD and LI–HD achieve better redetection rates but detect on the other hand more false positives near specular patterns. Specularities have a smaller influence on stability than sharp shadow or shading edges because they only appear for a narrow range of the angle between camera and light direction. If necessary for the application, specularities can be detected or reduced using polarising filters or using several images as in [LLK⁺02]. Several techniques also exist to correct specular areas, for example in [BSCB00, TLQS03, TI03]. Finally, the user–defined threshold influences stability for all detectors. Many interest points in the image may have similar cornerness values, especially in textured scenes. If the threshold selects only part of those, stability decreases as detection becomes noise sensitive. This could be enhanced by automatically adapting the detection threshold to scene content. These topics should be object of further research.

3.9 Summary

After a reminder of the image formation model and of the Harris detector for grey value images, the illumination influence on interest point detection is derived. Four new detectors are presented which are based on the Harris detector and which achieve a better adaptation of the detection to the local lighting conditions. The first detector, N–HD, is based on a local normalisation of the derivatives using the grey value energy. The second detector, H–HD, uses the principle of homomorphic processing to compute illumination invariant derivatives. The third detector, AT–HD, performs local adaptive thresholding. The threshold is adapted using the local mean of the cornerness values. To reduce noise sensitivity, textured areas are detected using the local standard deviation of the cornerness values and detection is switched off in homogeneous areas. The last detector, LI–HD, also applies local adaptive thresholding. The threshold is computed based on local clustering of the cornerness values with the ISODATA algorithm. The obtained cluster means enable texture detection, hence reducing noise sensitivity. N–HD and H–HD are based on the local multiplicative image formation model, whereas AT–HD and LI–HD are based

3 Illumination invariant interest point detection for grey value images

on the local affine image formation model. The principles of N-HD and of H-HD can be easily adapted to increase the stability of the other interest point detectors based on first or second derivatives. In addition, the invariant derivatives obtained by N-HD and H-HD can be re-used to compute invariant descriptors of the interest points.

The four detectors developed in this work are evaluated and compared to each other and to the original Harris detector (HD) on image series showing a scene under different illuminations. For the original Harris detector, the N interest points with the highest cornerness values are selected to adapt the detection to the overall image contrast. The detection stability is assessed with the redetection rate and the false positive rate. For simple illumination changes which induce the same grey value transformation for all pixels, an adaptation to the overall image contrast is sufficient to compensate the illumination changes. All developed detectors (N-HD, H-HD, AT-HD and LI-HD) achieve similar stability to the original Harris detector (HD). Complex illumination variations, for which position, orientation, type and number of light sources are changed, require on the other hand an adaptation to local lighting conditions. For such image series, all developed detectors achieve higher stability than the original Harris detector. The stability improvement is higher for complexer illumination change (for example when the illuminant type is changed) and for complexer 3D scene geometry. The best results are obtained by H-HD. AT-HD may also be interesting for some applications because it provides a more uniform distribution of the interest points in the image. All developed detectors remain sensitive to sharp shadow or shading edges as these are not modelled accurately by the image formation model. This is handled in the next chapter using colour images. In addition, the stability of the developed detectors is influenced by specularities and by the choice of the detection threshold. This should be object of further research.

4 Illumination invariant interest point detection for colour images

This chapter presents the developed illumination invariant interest point detectors for colour images. As explained in the previous chapters, colour information is advantageous for illumination invariance because shadows, shading and variations of the light colour can be more accurately modelled than with grey value information. Current colour cameras introduce colour artifacts in images, especially near edges. In this work a single chip colour camera is used. Therefore, in section 4.1, an appropriate demosaicing algorithm is selected, that reduces the amount of colour artifacts. This algorithm is applied to acquire all images used in this work. Next, the image formation model and the Harris detector for colour images are recalled and the illumination influence on the detector is discussed in section 4.2. The invariant detector introduced in [vdW05] is explained in more details in section 4.3. The two developed detectors are then presented: the colour homomorphic detector in section 4.4 and the m space detector in section 4.5. A preprocessing method is introduced in section 4.6 to reduce the influence of noise and artifacts on the m space detector. Next, the comparison framework is explained in section 4.7 and the results are presented in section 4.8. Finally, a summary of the chapter is given in section 4.9.

4.1 Colour image acquisition and demosaicing

As explained in subsection 2.3.2, current colour cameras introduce colour artifacts in images, especially near edges. The main source of colour artifacts is the camera sensor itself. If a multi-chip camera is used, colour artifacts occur when the sensor chips are misregistered. Most digital colour cameras are however based on a single CCD or CMOS sensor chip combined with a colour filter array (CFA): each pixel measures only one of the RGB colours. The most popular CFA is the Bayer CFA presented in [Bay76] and shown in fig. 4.1. Using a single chip with a CFA allows to reduce the camera cost. The spatial resolution of the human eye is lower for colour information than for intensity information. Therefore, the perceived quality is similar for both single chip and multi-chip cameras. This is the reason for the wide use of single chip cameras. In this work as well, a single chip camera is used. The sparsely sampled colour information must be interpolated to obtain a full resolution image with three colour values per pixel. This

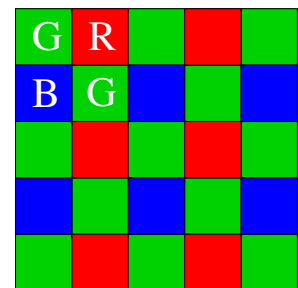


Figure 4.1: Bayer CFA.

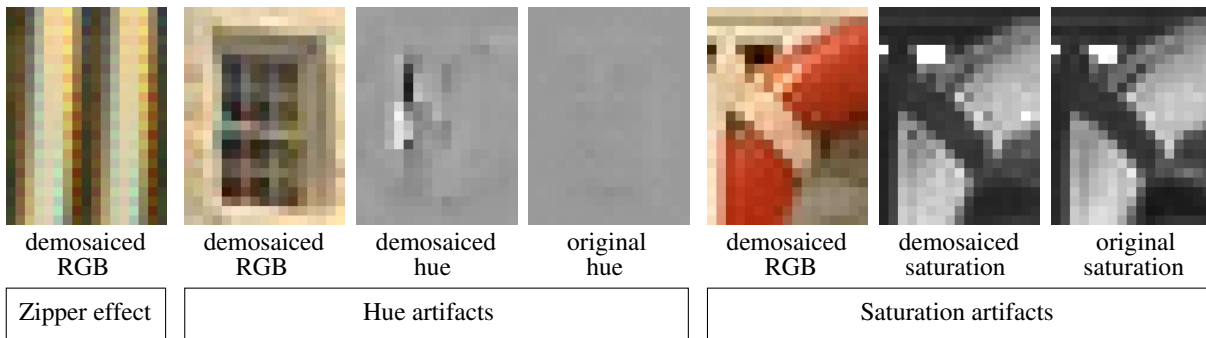


Figure 4.2: Illustration of typical demosaicing artifacts on simulated CFA images. For a better visualisation of hue and saturation artifacts, hue H and saturation S components of the HSI colour space are given for the demosaiced image and for the original three channel image. H and S are scaled between 0 and 255.

is named demosaicing, because CFA are also named mosaic filters. Many demosaicing algorithms exist (see the overview in subsection 4.1.1). Nonetheless, even recent methods are prone to interpolation errors which result in colour artifacts in the images. Typical demosaicing artifacts are shown in fig. 4.2: “zipper” effects, wrong colour hue and wrong colour saturation.

The demosaicing method has a strong influence on the quality of the resulting colour information. Hence, it influences the quality of colour gradient and of colour interest point detectors. This is illustrated in fig. 4.3 for the m space colour gradient. The m space gradient is only sensitive to chrominance (see section 4.5). It visualises hence very well the influence of colour artifacts. Fig. 4.3 shows that the gradient is better estimated after the complex demosaicing algorithm described in [LT03] than after a simple demosaicing algorithm (bilinear interpolation): image (d) contains less false edges and less edges with wrong colours than image (c). Colour gradient is used for interest point detection and also to compute interest point descriptors for the matching (see chapter 5). The reduction of colour artifacts is hence important for the whole application. Several demosaicing algorithms are compared in order to find the algorithm best suited for this work.

The previous comparisons between demosaicing methods, for example in [LT03, RSBS02, Had04], aim at reconstructing visually pleasing images, so their evaluation criteria are, in addition to Mean Squared Error (MSE) in RGB space, visual inspection and measures based on human perception like ΔE_{ab}^* . Here, the quality of colour information for interest point detection is important. The colour Harris detector and the homomorphic colour detector both use the RGB values directly. The robust invariant Harris detector and the m space Harris detector are based on chrominance only. None of the previously used criteria can evaluate chrominance quality alone. For that reason, the comparison results in [LT03, RSBS02, Had04] cannot be used to select the most appropriate demosaicing algorithm in this work. The comparison in subsection 4.1.2 is based on colour spaces HSI and I_{rb} for two reasons: these allow to evaluate chrominance quality and they describe colour information intuitively. Performance differences in coloured, textured and homogeneous areas are emphasised to better characterise the demosaicing results.

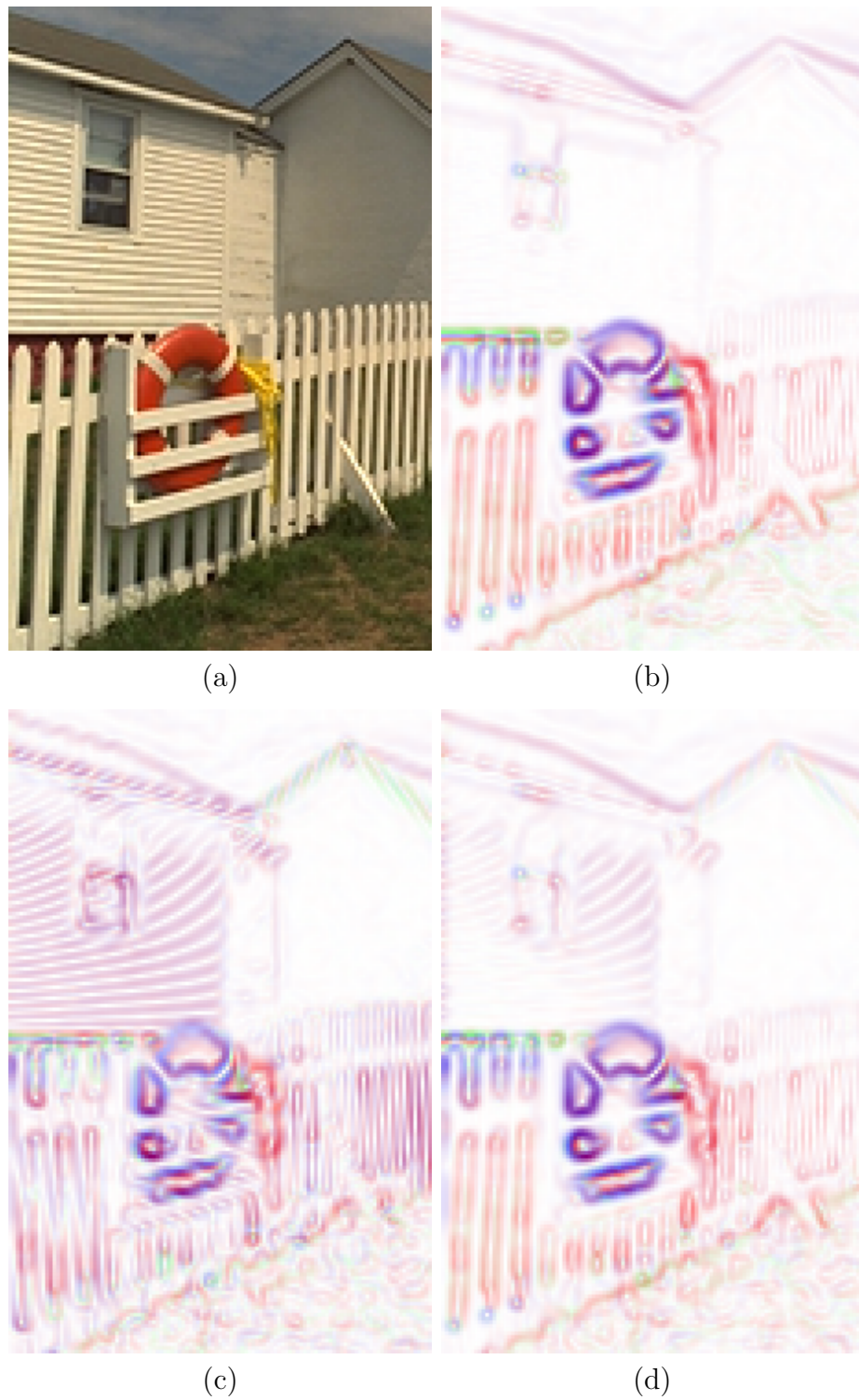


Figure 4.3: Influence of demosaicing on the colour gradient of a simulated CFA image. (a) original three channel image. (b) m space gradient of the original image. (c) m space gradient of the image demosaiced with bilinear interpolation. (d) m space gradient of the image demosaiced with the method in [LT03].

After an overview of the existing demosaicing algorithms, the principles of evaluated methods are introduced in subsection 4.1.1. Subsection 4.1.2 presents the comparison framework and the results. Finally, the demosaicing algorithm is selected in subsection 4.1.3. Comparisons of the same demosaicing methods with a more general background than interest point detection and additional comparison criteria can be found in [Fai04a, Fai05a]. To simplify explanations and formulae, the red, green and blue values of a pixel are represented by R , G and B in this section instead of C^R , C^G and C^B as in the rest of the thesis.

4.1.1 Presentation of the demosaicing algorithms

Overview of the existing demosaicing algorithms

The simplest demosaicing algorithm consists of a separate bilinear interpolation of the three channels. The obtained images are blurry and contain many artifacts as shown in figs. 4.9 (a) and 4.10 (a). Demosaicing quality can be improved by enlarging the considered neighbourhood. In addition, gradient information can be taken into account to adapt the used neighbourhood to image content, as in [HA97]. This reduces artifacts because interpolation is performed along rather than across edges. Finally, many algorithms use the high inter-channel correlation to constrain the interpolation process: either the colour differences $R - G$ and $B - G$ as in [Fre88] or the colour ratios R/G and B/G as in [Cox87] are assumed to remain constant in small neighbourhoods. These principles improve demosaicing at a moderate complexity increase. A good overview and comparison of state of the art demosaicing methods is given in [RSBS02]. In this section, the best two state of the art methods are compared. The median based postprocessing (MBP) in [Fre88] enforces high inter-channel correlation to reduce artifacts after a first demosaicing step. The adaptive colour plane interpolation (ACPI) in [HA97] takes inter-channel correlation into account and uses gradients to select between horizontal and vertical interpolation.

The most interesting recently developed algorithms can be divided into three categories. The algorithms of the first category apply the same principles as ACPI but provide more powerful and flexible methods to adapt the neighbourhood used for interpolation. In the Weighted Adaptive Colour Plane Interpolation (WACPI) in [LT03] and in the first step of [Kim99], each direction contributes to interpolation with a weight which is proportional to the gradient inverse. In [RS03], the weights depend on the similarity to the centre pixel after a first demosaicing step. WACPI also improves the framework to estimate R and B channels. This enhances chrominance quality, so WACPI is evaluated in this work.

The second category of methods locally selects between horizontal and vertical interpolation like in ACPI, based on more complex criteria than gradients. In [HP03], image homogeneity in CIELAB colour space is used. In [OW04], the variation of R/G and B/G colour ratios and the response to the Harris detector are used. These complex measures cannot be estimated directly from the CFA sampled images, so one horizontally interpolated image and one vertically interpolated image are generated first. The local direction

selection is performed subsequently to generate a final image. This results in a high computing time, even when direction selection is only performed in textured areas as in [OW04]. For that reason, these methods are not evaluated here.

The last category of methods works similarly to MBP. After a first demosaicing step, the result is postprocessed to enforce one or more constraints, hence reducing artifacts. In the second step of [Kim99], the locally constant colour ratio assumption is enforced using an adaptive neighbourhood. In [GAM02] a compromise between the following two constraints is reached: locally constant colour differences and fidelity to sampled data. These methods did not yield any significant enhancement over MBP in preliminary tests. They are therefore not evaluated here.

Median Based Postprocessing (MBP and EMBP)

MBP reduces demosaicing artifacts by enforcing high inter-channel correlation. It is presented in details in [Fre88]. After a first demosaicing step, the difference images $\delta_R = R - G$ and $\delta_B = B - G$ are median filtered. The image is then reconstructed using the filtered difference images δ'_R, δ'_B and the CFA sampled data. For example, at a sampled G pixel, $(R', G', B') = (\delta'_R + G, G, \delta'_B + G)$. The algorithm works similarly at sampled R and B pixels. The kernel proposed in [Fre88] is used for median filtering, because it achieves a good compromise between resulting image quality and computing time. The kernel is shown in fig. 4.4: the value of the centre pixel is replaced by the median of the nine pixels indicated in grey.

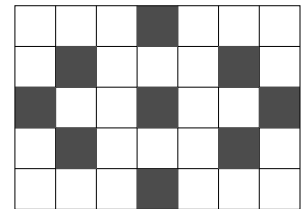


Figure 4.4: Kernel for median filtering.

Demosaicing results are shown in figs. 4.9 (b) and 4.10 (b). Pixels with wrong intensity and wrong saturation appear near colour edges with low inter-channel correlation such as red/white edges.

To avoid these artifacts due to contradictions between the high inter-channel correlation model and the sampled data, the Enhanced Median Based Postprocessing (EMBP) is used in [HP03, LT03]. Sampled values are also changed in EMBP. The reconstruction step becomes the same for all pixels: $(R', G', B') = (\delta'_R + G', (R - \delta'_R + B - \delta'_B)/2, \delta'_B + G')$. The implementation proposed in [LT03] is chosen here. Already processed pixels are used to filter following pixels for a faster diffusion of the postprocessing. In addition, only areas with sufficient gradient are postprocessed, because homogeneous areas are less prone to artifacts. Results are presented in figs. 4.9 (c) and 4.10 (c). For the comparison in subsection 4.1.2, MBP and EMBP are applied after the WACPI method (see [LT03]) to obtain the best possible results.

Adaptive Colour Plane Interpolation (ACPI)

ACPI is a state of the art method using gradient information and inter-channel correlation. It is presented in details in [HA97, RSBS02]. As the Bayer CFA contains twice

as many G pixels as R or B pixels, the G channel is interpolated first and is used to interpolate R and B channels in a second step. Figure 4.5 presents the G channel interpolation: the interpolation is performed in the direction of the minimum gradient. To take the inter-channel correlation into account, gradients and interpolated G values depend on the laplacian of the R (or B) channel. Similarly, interpolated R and B values depend on the laplacian of the interpolated G channel. Gradients are only used to interpolate R(B) values at B(R) pixels, for which four R(B) neighbours exist. The demosaicing results are illustrated in figs. 4.9 (d) and 4.10 (d).

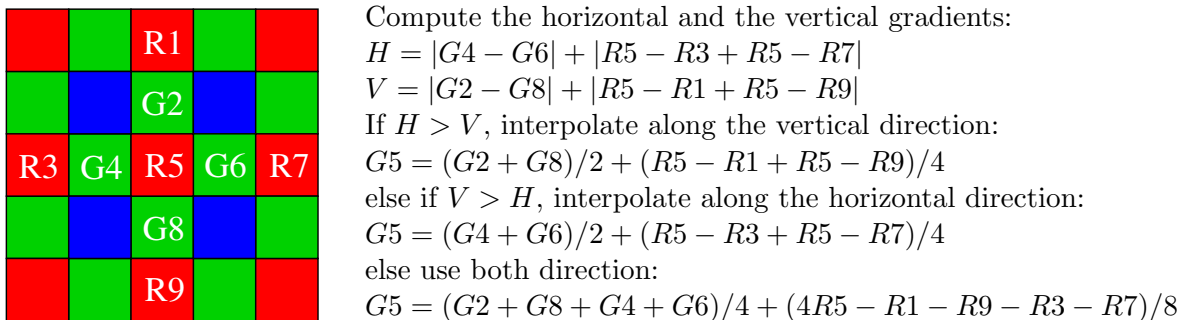


Figure 4.5: Interpolation of the G value at a sampled R pixel with ACPI. G values at sampled B pixels are interpolated similarly.

Weighted Adaptive Colour Plane Interpolation (WACPI)

WACPI is based on the same principles as ACPI. G values are estimated first. Gradients and interpolated values depend on the laplacian of the other colour channels, like in ACPI. WACPI provides a more flexible framework to adapt the interpolation neighbourhood to the image. Four directions (instead of two) are considered as shown in fig. 4.6, which enhances the performance near slanted edges and corners (see figs. 4.9 (e) and 4.10 (e)). The contributions of every direction to the interpolation of G values are weighted by the gradient inverses before they are summed up and normalised:

$$G = \frac{\alpha_{left} \tilde{G}_{left} + \alpha_{right} \tilde{G}_{right} + \alpha_{up} \tilde{G}_{up} + \alpha_{down} \tilde{G}_{down}}{\alpha_{left} + \alpha_{right} + \alpha_{up} + \alpha_{down}}.$$

The formulae for α_{left} and \tilde{G}_{left} are given in fig. 4.7 as an example for all α_i and \tilde{G}_i . As can be seen, the contributions to the interpolation \tilde{G}_i are the same as for ACPI. On the other hand, the gradients are estimated on a larger neighbourhood to compute the weights α_i . The constant additive term in the denominator of α_i avoids division by zero in homogeneous areas. The interpolation of the R and B channels is also improved in

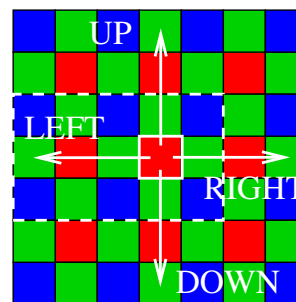


Figure 4.6: WACPI directions and neighbourhood shown in fig. 4.7.

comparison to ACPI: the gradient based neighbourhood adaptation is performed for the interpolation of all R and B values. To achieve this, R (or B) values at sampled B (or R) pixels are interpolated first. After this step, sampled G pixels have two sampled R (or B) neighbours and two interpolated R (or B) neighbours. As a consequence, the neighbourhood adaptation framework can also be applied to those pixels in a second step. The reader should refer to [LT03] for the complete algorithm description. The results are shown in figs. 4.9 (e) and 4.10 (e).

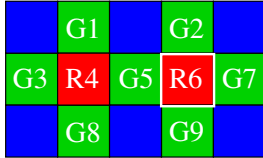
	$\tilde{G}_{left} = G5 + (R6 - R4)/2$ $\alpha_{left} = \frac{1}{1 + G7 - G5 + G5 - G3 + R6 - R4 + \frac{ G2 - G1 + G9 - G8 }{2}}$
---	---

Figure 4.7: Weight and contribution of the left direction (see fig. 4.6) to the interpolation of the G value at pixel position R6.

4.1.2 Comparison of the demosaicing algorithms

The demosaicing algorithms are evaluated on simulated CFA sampled images by comparison with the original three channel images. 24 images of the Kodak colour image database are used. These images depict scenes of various content, like for example landscapes, persons, natural and man-made objects, as can be seen in fig. 4.8.

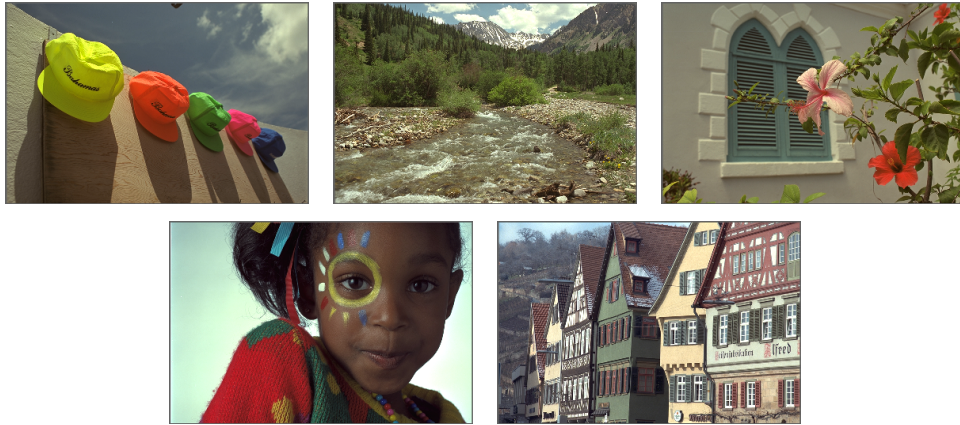


Figure 4.8: Five images of the Kodak colour image database.

The demosaicing results are illustrated in figs. 4.9 and 4.10 to show the different artifacts. The figures show one colourful area and one textured area of the *Small Lighthouse* image, a downsampled version of one of the Kodak database images used in many articles, for example in [Kim99, LT03]. The chosen areas illustrate best the results. For better visual analysis and understanding of the results, hue H, saturation S and intensity I components are also shown, as HSI is an intuitive colour space. The theoretical ranges of hue $[-\pi, \pi]$

4 Illumination invariant interest point detection for colour images

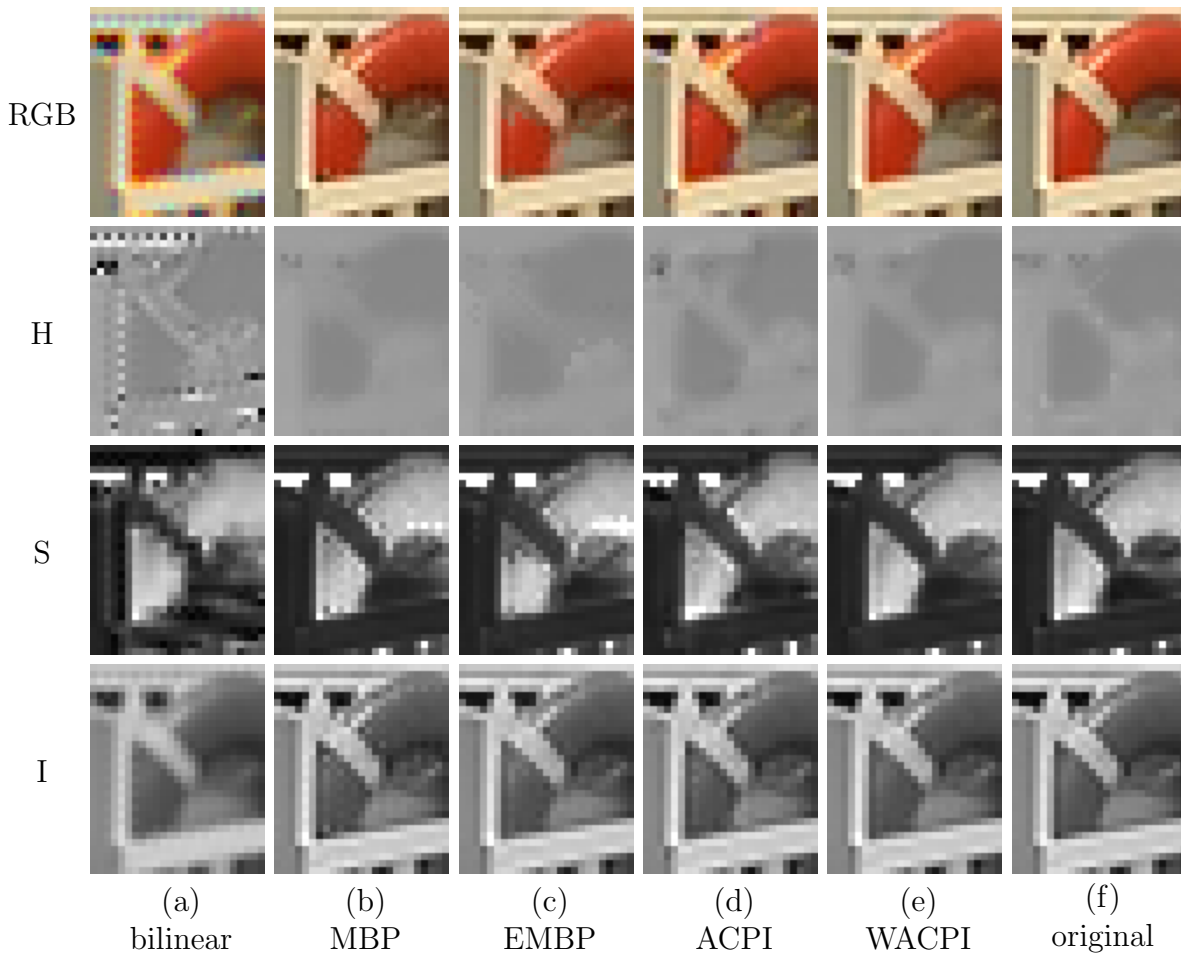


Figure 4.9: Demosaicing results for a colourful detail of the *Small Lighthouse* image showing a buoy. For better visualisation, H , S and I components are displayed.

and of saturation $[0, 1]$ are mapped to $[0, 255]$. Bilinear interpolation produces blurred images and artifacts known as "zipper" effects near edges (see figs. 4.9 (a) and 4.10 (a)). For ACPI and WACPI, colour artifacts appear, especially near slanted edges or corners as shown in fig. 4.10. MBP and EMBP correct these artifacts (see fig. 4.10) but introduce new artifacts near colour edges (see fig. 4.9). MBP generates outliers with wrong intensity and wrong saturation, whereas EMBP generates areas with wrong saturation. In addition, MBP reconstructs hue more accurately than EMBP in colourful areas (see fig 4.9). WACPI achieves the best reconstruction of the colourful detail, whereas MBP and EMBP achieve the best reconstruction of the textured detail. As far as complexity is concerned, ACPI, WACPI, MBP (+WACPI) and EMBP (+WACPI) require on average over all images 2, 8.5, 20 and 14 times as much processing time as bilinear interpolation.

Two detectors presented in this work are based directly on the RGB values and two detectors are based on chrominance information. Therefore, as explained in the introduction in this section, the demosaicing quality is evaluated using three colour spaces: RGB, HSI and Irb. The transformation from the RGB colour space to HSI and Irb is given for

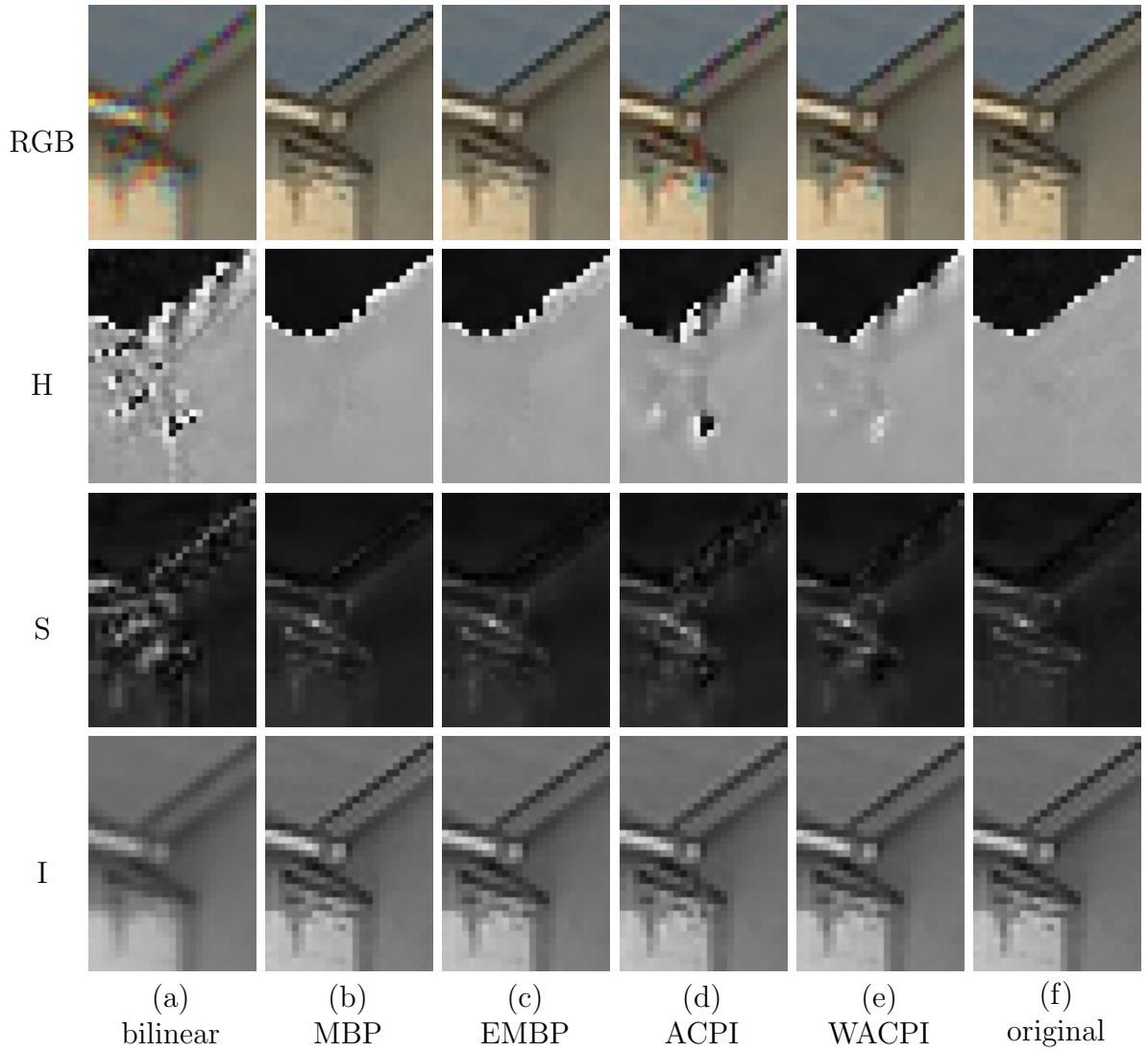


Figure 4.10: Demosaicing results for a textured detail of the *Small Lighthouse* image showing roofs. For better visualisation, H , S and I components are displayed.

example in [GS99]:

$$\begin{pmatrix} H \\ S \\ I \end{pmatrix} = \begin{pmatrix} \arctan \frac{\sqrt{3}(G-B)}{2R-G-B} \\ 1 - \min(R, G, B)/I \\ (R + G + B)/3 \end{pmatrix} \text{ and } \begin{pmatrix} I \\ r \\ b \end{pmatrix} = \begin{pmatrix} (R + G + B)/3 \\ R/I \\ B/I \end{pmatrix}. \quad (4.1)$$

Demosaicing quality is evaluated using the Mean Square Error (MSE) between the original three channel image and the demosaiced image in RGB, HSI and Irb colour spaces. The MSE in RGB space also allows comparison with previous evaluations like in [LT03, RSBS02]. H , S , r and b are scaled by 100 before computing the MSE, so that their order of magnitude is similar to the one of I , R , G , and B . To avoid numerical unstabilities, S , r and b are only computed for sufficiently bright pixels (here detected as $I > 0.9$). Similarly H is only estimated for significantly coloured pixels (here detected as $S > 0.1$).

4 Illumination invariant interest point detection for colour images

Algorithm	Textured areas			Homogeneous areas			Entire images		
	RGB	HSI	rb	RGB	HSI	rb	RGB	HSI	rb
<i>bilinear</i>	215	21.1	1.40	14.1	.900	.0545	104	10.7	.669
	87.1	.966	1.57	5.64	.0549	.0587	42.0	.460	.748
	222	95.4		15.1	6.14		108	45.9	
<i>ACPI</i>	38.1	3.10	.394	4.85	.370	.0245	20.1	1.79	.202
	24.9	.268	.319	2.60	.0223	.0232	13.2	.141	.168
	33.6	15.8		4.52	1.89		18.4	8.46	
<i>WACPI</i>	23.6	1.63	.232	3.44	.273	.0178	12.4	.936	.118
	13.3	.187	.210	1.71	.0173	.0180	7.22	.0959	.108
	21.6	9.44		3.52	1.35		11.7	5.08	
<i>MBP</i>	25.2	1.62	.249	4.01	.269	.0206	12.6	.814	.120
	9.93	.195	.212	1.91	.0204	.0212	5.18	.0976	.106
	18.4	8.13		3.67	1.49		9.72	4.20	
<i>EMBP</i>	29.7	2.30	.543	3.51	.281	.0201	14.4	1.09	.245
	18.1	.264	.283	1.79	.0182	.0188	8.57	.124	.134
	24.0	9.64		3.58	1.36		12.1	5.07	

Table 4.1: Demosaicing performance in textured areas, in homogeneous areas and in entire images. The average MSE over 24 images of the Kodak database is given for RGB, HSI and Irb spaces. For each detector, each row presents a channel (top row for R , H and r , middle row for G , S and b and last row for B and I). As I is the same in HSI and Irb, the MSE for I is only given for HSI. The best performance for each component is indicated in boldface.

To avoid the influence of border effects, a three pixel wide image border is left out during MSE computation. The demosaicing performance depends on image content. Therefore the average MSE over all 24 images of the Kodak database is used. The results are presented in table 4.1. The performance differences in textured and homogeneous areas is emphasised (texture is detected by thresholding the response of a Laplacian filter on the original image).

For all demosaicing algorithms, the G channel is better estimated than R and B channels because there are sampled twice as often. Bilinear interpolation, which does not use gradients and inter-channel correlation, yields by far the worst results. Despite its low complexity, ACPI allows significant enhancement, proving the strength of the used principles. WACPI and MBP perform best. The quality of the colour ratios and of saturation are on average comparable for both algorithms (see MSE for r , b , S). MBP improves texture estimation (see MSE for R , G , B and I) and reduces hue artifacts (see MSE for H). On the other hand, its performance in homogeneous areas is worse. This is mostly due to its higher sensitivity to the inaccuracy of the high inter-channel correlation model (constant colour differences in a neighbourhood) in coloured areas. Even low contrast

ACPI			WACPI			MBP			EMBP		
RGB	HSI	<i>rb</i>	RGB	HSI	<i>rb</i>	RGB	HSI	<i>rb</i>	RGB	HSI	<i>rb</i>
30.8	2.15	.747	19.2	1.12	.435	28.2	.817	.452	27.2	1.06	.806
19.6	.756	.598	10.7	.584	.409	11.3	.524	.407	15.8	.759	.484
27.6	12.8		18.8	7.75		19.2	8.99		21.0	7.95	

Table 4.2: Demosaicing performances in coloured areas (with $S \geq 0.3$). As in table 4.1, the average MSE over all 24 images is given in RGB, HSI and *rb* colour spaces. The results for the bilinear interpolation are omitted, as it yields by far the worst results in table 4.1.

edges in coloured areas may result in pixels with wrong intensity and saturation. As mentioned in the overview in subsection 4.1.1, the recent methods similar to MBP in [Kim99, GAM02] do not significantly reduce this sensitivity to model inaccuracy. The overall performance of EMBP is moderate. It better estimates homogeneous areas than MBP, but achieves poor chrominance quality compared to WACPI and MBP.

Table 4.2 presents the demosaicing performances in coloured image areas. Coloured areas are detected by thresholding the saturation component: $S \geq 0.3$. They are particularly interesting for this work because the *m* space detector and the robust invariant detector are only sensitive to chrominance (see sections 4.3 and 4.5). WACPI and MBP perform best. In comparison to table 4.1, MBP shows the maximal performance drop, especially for the estimation of texture (see MSE for *R*, *G*, *B* and *I*). In coloured areas, WACPI achieves the best texture estimation. MBP’s higher sensitivity to the model inaccuracy is also shown by a higher number of negative (hence invalid) interpolated values: if these are not corrected to 0, the average MSE for all 24 entire images for example for *r* becomes 0.441 for WACPI and 0.965 for MBP.

4.1.3 Selection of the most appropriate demosaicing method

To summarise the evaluation results, WACPI and MBP achieve the best results. WACPI performs better than MBP in coloured and in homogeneous areas. MBP better reconstructs texture and hue than WACPI. As a consequence, MBP is better on images with fine textures (for example landscapes or natural objects), while WACPI is better on images of man-made objects and on close-ups. In this work, a good reconstruction of coloured areas is important because the robust invariant detector and the *m* space detector are only sensitive to colour edges. Therefore, WACPI is chosen for demosaicing. In addition, recognition tasks deal often with images of man-made objects, which are better reconstructed with WACPI than with MBP. Last, WACPI is faster. If a short computing time is important for the application, ACPI is also interesting because it achieves a good compromise between computing time and demosaicing quality.



Figure 4.11: Influence of white balancing on demosaicing results. Left column: no white balancing is applied before demosaicing. Right column: white balancing (white patch algorithm explained in section 4.3) is applied before demosaicing. Top row: demosaicing results. Bottom row: m space gradients computed on the demosaiced images. These gradients allow a better visualisation of colour artifacts. Colour images are gamma corrected for better visualisation ($\gamma = 1.5$). Gradients images are scaled between 0 and 255.

For all methods, demosaicing quality is lower in coloured areas because inter-correlation between channels is lower in these areas. As a consequence, white balancing should be performed before demosaicing: this reduces the size of coloured areas in the image and hence increases demosaicing quality. Fig. 4.11 shows how white balancing reduces the formation of colour artifacts during demosaicing. The left row presents a demosaiced image detail. The right row presents the same image detail, but white balancing is performed before demosaicing. No ground truth exists for the image, therefore demosaicing quality can only be assessed by visual inspection. For better visualisation of the demosaicing quality, the m space gradient images (see section 4.5) are shown. As the m space gradient is only sensitive to chrominance, the colour artifacts are more easily visible. When white balancing is applied before demosaicing, all shadow or shading edges have less colour artifacts (the m space gradient is lower). Colour edges also have a more homogeneous hue (the colour of the m space gradient varies less along a colour edge). Finally the gradient image looks less noisy. All these observations show that less colour artifacts are introduced when white balancing is performed before demosaicing. In this work, automatic white balancing is only applied for the robust invariant detector and for one version of the m space detector. The other detectors do not require white balancing.

4.2 Image formation model and Harris detector for colour images

This section begins with a reminder of the image formation model and of the Harris detector for colour images. More details can be found in sections 2.1 and 2.2.3. The chosen image formation model for colour images is described by the following formula:

$$C^j = i_b L^j c_b^j + i_s L^j \quad \text{for } j = R, G, B. \quad (4.2)$$

C^R , C^G and C^B are the red, green and blue values of the considered pixel. c_b^j represents scene reflectance. i_b models shadows and shading. i_s models shadows and specularities. In contrary to the image formation model for grey value images, i_b and i_s must not be assumed to vary slowly from one pixel to another: they vary freely. L^j represents illuminant colour and is assumed to vary slowly between pixels. In comparison to the image formation model for grey value images, shadows, shading, specularities and illumination colour are more accurately modelled. The image formation model for colour images can also be expressed in vector-matrix form, in which case the model is named diagonal with translation:

$$\begin{pmatrix} C^R \\ C^G \\ C^B \end{pmatrix} = i_b \begin{pmatrix} L^R & 0 & 0 \\ 0 & L^G & 0 \\ 0 & 0 & L^B \end{pmatrix} \begin{pmatrix} c_b^R \\ c_b^G \\ c_b^B \end{pmatrix} + i_s \begin{pmatrix} L^R \\ L^G \\ L^B \end{pmatrix}. \quad (4.3)$$

For Lambertian scenes, the model is simplified to a diagonal model:

$$C^j = i_b L^j c_b^j \quad \text{for } j = R, G, B, \quad \text{or} \quad \begin{pmatrix} C^R \\ C^G \\ C^B \end{pmatrix} = i_b \begin{pmatrix} L^R & 0 & 0 \\ 0 & L^G & 0 \\ 0 & 0 & L^B \end{pmatrix} \begin{pmatrix} c_b^R \\ c_b^G \\ c_b^B \end{pmatrix}. \quad (4.4)$$

4 Illumination invariant interest point detection for colour images

i_b , L^j and c_b^j have the same properties as in eq. (4.2).

The Harris detector is extended to work on colour images in [Gou00]. The information from all colour channels is merged into the structure matrix according to:

$$\mathbf{M} = G(\sigma_M) \otimes \sum_{j=R,G,B} \begin{bmatrix} (C_x^j)^2 & C_x^j C_y^j \\ C_x^j C_y^j & (C_y^j)^2 \end{bmatrix}. \quad (4.5)$$

C_x^j and C_y^j are the derivatives of channel C^j . They are obtained by convolving each channel with a derivative of Gaussian filter with standard deviation σ_{deriv} : $C_{x/y}^j = G_{x/y}(\sigma_{deriv}) \otimes C^j$ for $j = R, G, B$. $G(\sigma_M)$ is a Gaussian with standard deviation σ_M . The cornerness function is then computed with:

$$CF = \det(\mathbf{M}) - \alpha \text{trace}^2(\mathbf{M}). \quad (4.6)$$

The interest points are the local maxima of CF above detection threshold T . Like for grey value images, detection is adapted to the overall image contrast when T is proportional to the maximum cornerness value in the image or when the N interest points with the highest cornerness values are selected. With the described implementation, the colour Harris detector needs 681ms for an image with 640×480 pixels (see subsection 2.2.2 for more details), that is 1.66 times the processing time of the grey value Harris detector (HD).

The illumination influence on the colour detector is complexer than for grey value images because the image formation model is more accurate. With the used model in eq. (4.2), the derivatives of each channel are composed of a sum:

$$C_x^j = i_{bx} L^j c_b^j + i_b L^j c_{bx}^j + i_{sx} L^j \quad \text{for } j = R, G, B. \quad (4.7)$$

f_x is the derivative of signal f in x direction (for $f = C^j, i_b, c_b^j, i_s$). This formula is obtained by considering that L^j varies slowly from one pixel to another. Therefore the derivative of L^j can be approximated by 0. A similar formula can be obtained for the derivative in y direction. Even when i_b and i_s are assumed to vary slowly between neighbouring pixels (hence $i_{bx} \approx i_{sx} \approx 0$), the illumination influence on each channel is still different due to the light colour represented by L^j :

$$C_x^j = i_b L^j c_{bx}^j \quad \text{for } j = R, G, B.$$

As a consequence, no simple formula can describe the illumination influence on the cornerness function as is the case for grey value images. The reason is the more accurate image formation model. It is however clear that an adaptation of the detection threshold to the overall image contrast is insufficient for stable interest point detection under illumination changes. This is shown in fig. 4.12. The same scene is depicted in both images. On the left, it is lighted by neon lamps suspended from the ceiling and directed towards the floor. On the right, it is lighted by tungsten halogen spot lamps directed towards the scene. No automatic white balancing is used. A fixed white balancing is applied

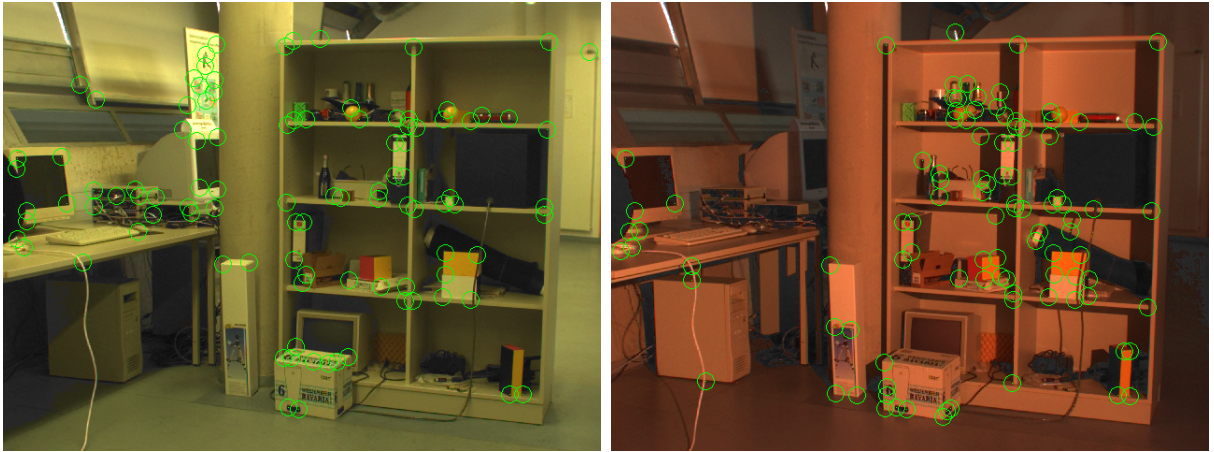


Figure 4.12: Detection example for the colour Harris detector with selection of the N best points ($N = 100$). The images show the same scene under two different illuminations. They are gamma corrected for better visualisation ($\gamma = 1.5$). 28.0% of the interest points of the left image are redetected in the right image. 71.7% of the interest points in the right image are false positives.

for visualisation to reduce the otherwise greenish hue of the images. This fixed white balancing counterbalances the different channel sensitivities and is generally performed in the camera hardware. Here, the red and the blue channels are both scaled with 1.3. These factors transform sunlight to white illumination. Neon and tungsten lamps result in yellow and red images respectively, as seen in fig. 4.12. Like with the original Harris detector for grey value images, interest points are detected mainly in areas with high local contrast. Detection stability is low because these areas move when illumination changes. Only 28.0% of the interest points of the left image are redetected in the right image. 71.7% of the interest points in the right image are false positives. For comparison, the stability of the grey value detectors on this image pair are given at image 26 in the *shelves* series in fig. 3.24.

Like in chapter 3, new interest point detectors are developed in this chapter. They adapt detection to the local lighting conditions. For this, principles used in other machine vision applications are adapted and applied to interest point detection. The overview in subsection 2.3.2 shows that three principles are interesting for interest point detection: local normalisation, homomorphic processing and the use of an illumination invariant colour space. In comparison to the developed detectors for grey value images, no detector based on local threshold adaptation is presented. This has two reasons. First, no simple formula can describe illumination influence on the cornerness function for the colour Harris detector, as shown before. Second, the grey value detectors based on local normalisation N-HD and on homomorphic processing H-HD both achieve a higher stability than the detectors based on local adaptive thresholding (AT-HD and LI-HD) (cf. section 3.8). N-HD and H-HD have similar stability, but H-HD is faster. As a result, the first developed colour interest point detector is based on homomorphic processing. No detector based on local normalisation is presented as it would achieve similar stability. The colour homomorphic

Harris detector is based on the assumption that shadow and shading influence i_b varies slowly between neighbouring pixels (see section 4.4). Therefore, the second developed interest point detector is based on the m space: an illumination invariant colour space which compensates completely shadow and shading influence. Both detectors implicitly correct illumination colour, so no white balancing is required as preprocessing (see subsection 2.3.2 for explanations). Both detectors assume a Lambertian scene. No detector compensating specularities in colour images is presented in this thesis. The main reason is that an accurate estimation of the illumination colour is necessary to compensate the specular reflection term $i_s L^j$. As explained in subsection 2.3.2, no automatic method to estimate the illuminant colour is reliable enough. Before presenting the new detectors, the robust invariant interest point detector introduced in [vdW05] is explained in details, as a comparison.

4.3 Robust invariant interest point detector

Most existing illumination invariant colour features are very sensitive to noise and artifacts, especially in dark and non-coloured areas. In [vdW05], several methods to compute structure matrix \mathbf{M} in eq. (4.5) are introduced, depending on the desired degree of invariance and of robustness to noise. These methods are based on the image formation model presented in eq. (4.2), in which i_b , i_s are assumed to vary freely. White balancing is required as preprocessing. As a consequence, L^j is constant for all pixels and for all images. This is the main difference between the robust invariant interest point detector and the detectors developed in this thesis.

The first method provides a quasi-invariant structure matrix \mathbf{M} . This matrix does not respond to shadow and shading edges or to shadow, shading and specular edges, depending on the chosen degree of quasi-invariance. This means that the matrix elements are only different from zero for material edges (caused by c_b^j). The matrix elements for material edges are however influenced by shadows, shading and specularities. The advantage of the quasi-invariant structure matrix is its high robustness to noise and artifacts. It is however not sufficient for stability under illumination changes because the values in the structure matrix are influenced by illumination. This illumination influence can be suppressed by normalisation. This yields the full invariant structure matrix. Like for the quasi-invariants, the matrix can be invariant to shadows and shading or to shadows, shading and specularities. This invariant matrix provides in theory stable detection under illumination changes, but it is very sensitive to noise. Therefore a last method is proposed, which yields the robust invariant structure matrix. In comparison to the invariant matrix, some weighting is performed to reduce noise influence. As before, the robust invariant matrix can be invariant to shadows and shading or to shadows, shading and specularities. The robust invariant structure matrix is of particular interest for this work because such a compromise between noise sensitivity and invariance to illumination changes could yield higher stability for interest point detection. Several feature detectors based on structure matrix \mathbf{M} are presented in [vdW05], among others the colour Harris

detector. The stability of the resulting applications is however not evaluated in [vdW05]. It is simply illustrated on an example. Therefore the robust invariant Harris detector is evaluated and compared to the developed detectors in section 4.8.

As explained in subsection 2.3.2, the main drawback of this detector is the requirement for white balancing: no existing automatic white balancing method is robust on real images (see [BMCF02b]). Manual white balancing is used in the applications shown in [vdW05]. In this work, on the contrary, no manual interaction should be required. As a consequence, an automatic white balancing method is applied. Based on the comparison between white balancing methods in [BMCF02b], the white-patch algorithm is used as preprocessing for the robust invariant detector. The robust invariant to shadows, shading and specularities relies strongly on the estimated illuminant colour (L^R, L^G, L^B) . To reduce unstability caused by misestimation of the illuminant colour to a minimum, only the robust invariant to shadows and shading is considered in this work.

The chosen automatic white balancing algorithm is the white-patch algorithm. It is applied before demosaicing to reduce the formation of colour artifacts (see section 4.1). The illuminant colour is assumed to be the same for all pixels. It is estimated from the average value of brightest pixels in the image. The brightest pixels in the image are assumed to be produced by either the light source, reflection on a white surface or specular highlights (see also subsection 2.3.2). If the light source is visible in the image and the pixel values are not saturated, the pixel values are clearly equal to the illuminant colour. If the pixels are reflected by a white surface, $c_b^j = 1$ for $j = R, G, B$, hence $C^j = i_b L^j + i_s L^j = (i_b + i_s) L^j$ for $j = R, G, B$. In the case of a specular highlight, i_s is much bigger than i_b , hence $C^j \approx i_s L^j$. Therefore, in all three cases, the RGB values are proportional to the light colour: $C^j = i L^j$ for $j = R, G, B$ and where i is an intensity factor. Therefore, the white patch algorithm works as indicated in the following:

1. Compute the intensity image: $I = C^R + C^G + C^B$
2. Select the brightest non-saturated pixels: all pixels (x, y) such that $C^j(x, y) < 255$ for $j = R, G, B$ and such that $I(x, y) > \alpha \max(I)$ where α is a user-defined threshold ($\alpha = 0.7$ in this work).
3. Compute the average RGB value of the selected pixels: $A^j = \frac{1}{N} \sum_{i=1}^N C^j(x_i, y_i)$ for $j = R, G, B$. (x_i, y_i) for $i = 1 \dots N$ are the coordinates of the selected pixels.
4. Correct the illuminant colour by weighting two image channels, for example red and blue channels with:

$$C_{wb}^R = \frac{A^G}{A^R} C^R, \quad C_{wb}^G = C^G, \quad C_{wb}^B = \frac{A^G}{A^B} C^B,$$

where C_{wb}^R , C_{wb}^G and C_{wb}^B are the channels of the white balanced image.

After this correction, the average of the brightest pixels of the white balanced image is: $(\frac{A^G}{A^R} A^R, A^G, \frac{A^G}{A^B} A^B) = A^G (1, 1, 1)$, which corresponds to white. In the used implementation, the two channels with the smallest A^j values are corrected because the resulting white balanced images look more natural than when two fixed channels are corrected.

4 Illumination invariant interest point detection for colour images

The robust invariant structure matrix is given by:

$$\mathbf{M} = \sum_{j=R,G,B} \begin{bmatrix} \frac{G(\sigma_M) \otimes (S_x^{C^j})^2}{G(\sigma_M) \otimes w} & \frac{G(\sigma_M) \otimes S_x^{C^j} S_y^{C^j}}{G(\sigma_M) \otimes w} \\ \frac{G(\sigma_M) \otimes S_x^{C^j} S_y^{C^j}}{G(\sigma_M) \otimes w} & \frac{G(\sigma_M) \otimes (S_y^{C^j})^2}{G(\sigma_M) \otimes w} \end{bmatrix}, \quad (4.8)$$

where $G(\sigma_M)$ is a Gaussian with standard deviation σ_M like in eq. (4.5). $S_x^{C^j}$ and $S_y^{C^j}$ are the quasi-invariant derivatives of channel j in x and y directions. They are computed with:

$$S_x^{C^j} = C_x^j - \frac{C_x^R \bar{C}^R + C_x^G \bar{C}^G + C_x^B \bar{C}^B}{\bar{C}^R^2 + \bar{C}^G^2 + \bar{C}^B^2} \bar{C}^j \quad \text{for } j = R, G, B. \quad (4.9)$$

C_x^j is the derivative of C^j in x direction and it is obtained by convolving C^j with a derivative of Gaussian filter with standard deviation σ_{deriv} : $C_x^j = G_x(\sigma_{deriv}) \otimes C^j$. \bar{C}^j is obtained by convolving C^j with a Gaussian of standard deviation σ_{deriv} : $\bar{C}^j = G(\sigma_{deriv}) \otimes C^j$. $S_y^{C^j}$ is computed with a similar formula, replacing C_x^j by C_y^j . Finally, w combines normalisation and noise reducing weighting in one step:

$$w = \bar{C}^R^2 + \bar{C}^G^2 + \bar{C}^B^2. \quad (4.10)$$

Those formulae, which define the robust shadow-shading invariant structure matrix, are explained in [vdW05]. The cornerness function is then computed using eq. (4.6) and the interest points are the local maxima of the cornerness function above the user-defined threshold T ($T > 0$). The convolution with the Gaussian and derivatives of Gaussian filters are implemented as indicated in subsection 2.2.2. ¹⁾

The gradient obtained with the robust shadow-shading invariant method is illustrated and compared to the m space gradient in fig. 4.16 (b). The influence of shading is well compensated, for example on the box in the top left image part. The shadow and shading edges are also well suppressed: only edges between differently coloured areas are present. The high noise sensitivity is visible, especially in dark areas. The results of the robust invariant Harris detector (RI-HD) are illustrated in fig. 4.13. The same images as in fig. 4.12 are used. The white patch algorithm is applied before demosaicing. Weaknesses of automatic white balancing are visible in fig. 4.13: corresponding pixels in the two images have similar but not identical colour because some areas are influenced by several light sources. In the right image, the background is lighted by halogen lamps and by some sunlight entering the room despite the closed blinds. As a result, the background appears blue after white balancing. RI-HD is not sensitive to intensity edges. Therefore, less interest points are detected than with the colour Harris detector. The interest points are detected near colour edges. Some false interest points are also detected in very dark areas (on black objects or in shadows) or because of colour artifacts. This is caused by the high sensitivity of RI-HD to noise in dark areas and to colour artifacts. In comparison to the colour Harris detector, a more stable detection is achieved because illumination influence is better compensated. In fig. 4.13, 35.7% of the interest points in the left image

¹⁾ I would like to thank Joost van de Weijer for his help for the implementation of the robust invariant Harris detector (RI-HD).

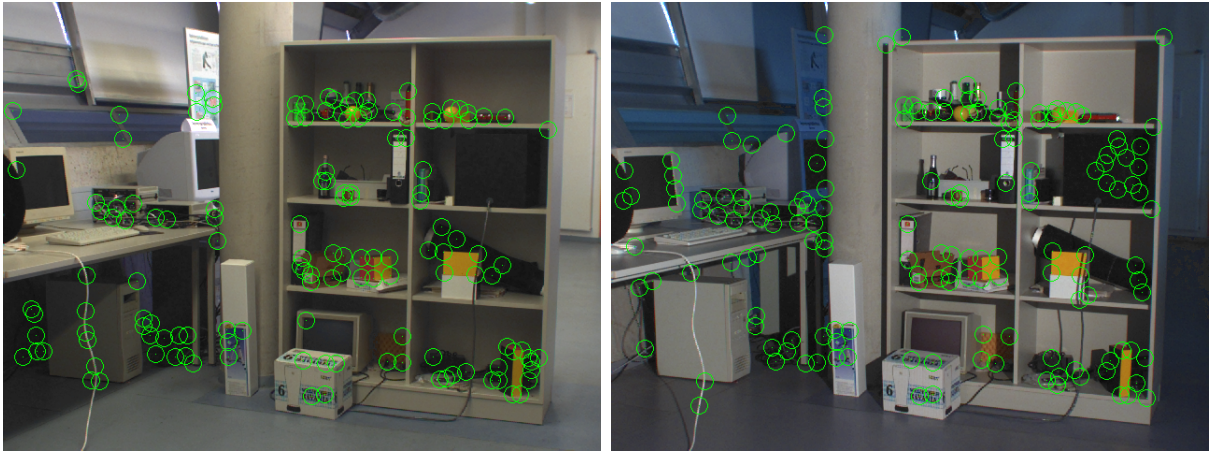


Figure 4.13: Detection example for the robust invariant Harris detector ($T = 10^{-7}$). The same images as in fig. 4.12 are used. White balancing is applied as required by the RI-HD. 35.7% of the interest points of the left image are re-detected in the right image. 68.2% of the interest points in the right image are false positives. Both images are gamma corrected for better visualisation ($\gamma = 1.5$).

are re-detected in the right image. 68.2% of the interest points in the right image are false positives. On this example, the stability increase is not as high as expected because of the detector sensitivity to noise and colour artifacts. With the described implementation, RI-HD requires 1151ms (+ approximately 10ms for white balancing) for an image with 640×480 pixels (see subsection 2.2.2 for details on the computer), this is 2.82 times the processing time of HD and 1.70 times the processing time of C-HD.

4.4 Homomorphic colour interest point detector

Homomorphic processing is often used to compensate illumination influence in colour images for machine vision application, as seen in subsection 2.3.2. It extends a principle used for grey value images to colour images by processing each channel separately.

Homomorphic processing is designed for Lambertian reflection. It is based on the image formation model of eq. (4.4): $C^j = i_b L^j c_b^j$ for $j = R, G, B$. A further restriction is applied: both light colour L^j and intensity factor i_b are assumed to vary slowly between pixels. The first step of homomorphic processing is to take the logarithm of each channel:

$$\ln C^j = \ln i_b + \ln L^j + \ln c_b^j \quad \text{for } j = R, G, B. \quad (4.11)$$

The illumination influence on $\ln C^j$ is additive. $\ln i_b$ and $\ln L^j$ are assumed to vary slowly between pixels: they are constant in small image neighbourhoods. On the contrary, $\ln C^j$ and $\ln c_b^j$ both vary freely. As a consequence, applying a linear high-pass filter to $\ln C^j$ provides illumination invariant information. The derivatives of each channel are therefore

4 Illumination invariant interest point detection for colour images

illumination invariant:

$$\frac{\partial \ln C^j}{\partial x} = \frac{C_x^j}{C^j} = \frac{i_{bx}L^j c_b^j + i_b L_x^j c_b^j + i_b L^j c_{bx}^j}{i_b L^j c_b^j} \approx \frac{c_{bx}^j}{c_b^j} \quad \text{for } j = R, G, B. \quad (4.12)$$

f_x denotes here the derivative of signal f in x direction (for $f = i_b, L^j, c_b^j$). This approximation is based on the fact that light colour L^j and intensity factor i_b are assumed to vary slowly between pixels, hence $i_{bx} \approx 0$ and $L_x^j \approx 0$. The derivative in y direction obeys a similar equation. Eq. (4.12) proves that homomorphic processing on colour images can compensate both light colour influence and intensity factor for channel derivatives. Like for grey value images, the derivatives resulting from homomorphic processing can be interpreted as derivatives which are normalised channel-wise with the local mean values: $C_x^j/\overline{C^j}$ for $j = R, G, B$, where $\overline{C^j}$ is the mean value of channel j on the neighbourhood used for derivative estimation (see eq. (4.12)).

For interest point detection, the invariant derivatives are merged into the structure matrix with eq. (4.5). Last, the cornerness function CF is computed using eq. (4.6). The interest points are the local maxima of CF with a cornerness value above the user-defined detection threshold. As the channel derivatives are invariant to light colour, shadows and shading, the structure matrix \mathbf{M} and the cornerness function CF are also invariant to light colour, shadows and shading. The homomorphic colour Harris detector is therefore stable under illumination changes for Lambertian scenes. Shadow and shading factors are assumed to vary slowly in space like for the grey value detectors. The homomorphic colour Harris detector has an advantage over the homomorphic Harris detector (H-HD): light colour is better corrected, because light colour cannot be accurately modelled for grey value images.

Like for H-HD, the logarithm of the image channels cannot be implemented straightforwardly for two reasons. First, the noise influence on the derivatives would be too strong in dark areas. Second, the logarithm is not defined for pixels with a value equal to zero. The implementation described in section 3.3 for the grey value detector achieves a good compromise between invariance and noise sensitivity. Therefore, this implementation is also used for the colour detector. The logarithm of the channels is computed with $\ln(C^j + 1)$ instead of $\ln C^j$. In addition, dark pixels are preprocessed with a 3×3 box filter applied channel-wise. If C^j is smaller than a user-defined threshold V , its value is replaced by the mean value in their 3×3 neighbourhood. This process is repeated for each channel: $j = R, G, B$. Like for the grey value images, V is set to 3. This preprocessing and the threshold V should be adapted to the camera noise.

Fig. 4.14 shows how the illumination influence on the colour gradient is compensated with homomorphic processing. In comparison to the normal colour gradient, edges in brightly lighted image areas and in shadow areas have similar values. The implicit compensation of the illuminant colour is visible: for example, shadow and shading edges appear grey or black. The noise amplification in dark areas is also visible, for example in the shadows. The homomorphic gradient is compared to the m space gradient and to the robust invariant gradient in fig. 4.16: the homomorphic gradient is less sensitive to noise but it cannot suppress sharp shadow or shading edges.

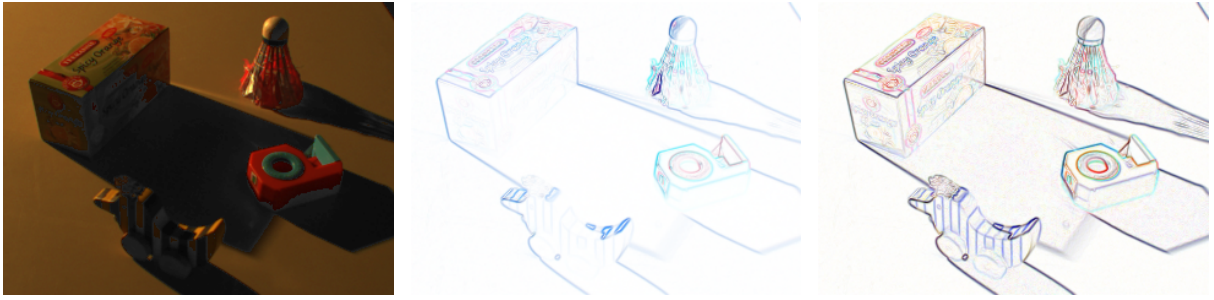


Figure 4.14: Suppression of the illumination influence on colour derivatives with homomorphic processing. Left: colour image. Middle: colour gradient $\sqrt{C_x^{j^2} + C_y^{j^2}}$ for $j = R, G, B$. Right: colour gradient with homomorphic processing. The gradient images are scaled between 0 and 255 (the maximum and minimum gradient values over all pixels and all channels are used). The colour image is gamma corrected ($\gamma = 1.4$).

The homomorphic colour Harris detector is summarised in the following:

1. For each channel $j = R, G, B$:
 - a) Preprocess the dark areas of the channel if necessary. For example, preprocess all pixels with channel value C^j smaller than $V = 3$ with a 3×3 box filter.
 - b) Take the logarithm of the preprocessed channel with: $L^j = \ln(1 + C^j)$.
 - c) Compute the invariant derivatives by convolving the logarithm of the preprocessed channels with the derivatives of Gaussian filters: $L_x^j = G_x(\sigma_{deriv}) \otimes L^j$ and $L_y^j = G_y(\sigma_{deriv}) \otimes L^j$.
2. Combine the invariant derivatives L_x^j and L_y^j for $j = R, G, B$ into the structure matrix \mathbf{M} with eq. (4.5).
3. Compute the cornerness function CF with eq. (4.6).
4. (x, y) is an interest point:
 - if it is a local maximum of the cornerness function CF
 - and if $CF(x, y) > T$ ($T > 0$).

This algorithm is named the homomorphic colour Harris detector (HC–HD). Like for the grey value detector, the preprocessing should be adapted to the amount of noise in the image. The user–defined threshold T should be set to detect an appropriate number of interest points. The invariant channel derivatives can be re–used during the computation of illumination invariant descriptors. The algorithm could be easily extended to other versions of the Harris detector, for example to the scale or affine invariant detectors, and to other interest point detectors based on high–pass filtering. It should be however kept in mind that the approximation of shadow and shading factors by a signal with low spatial frequencies becomes worse when larger neighbourhoods are considered.

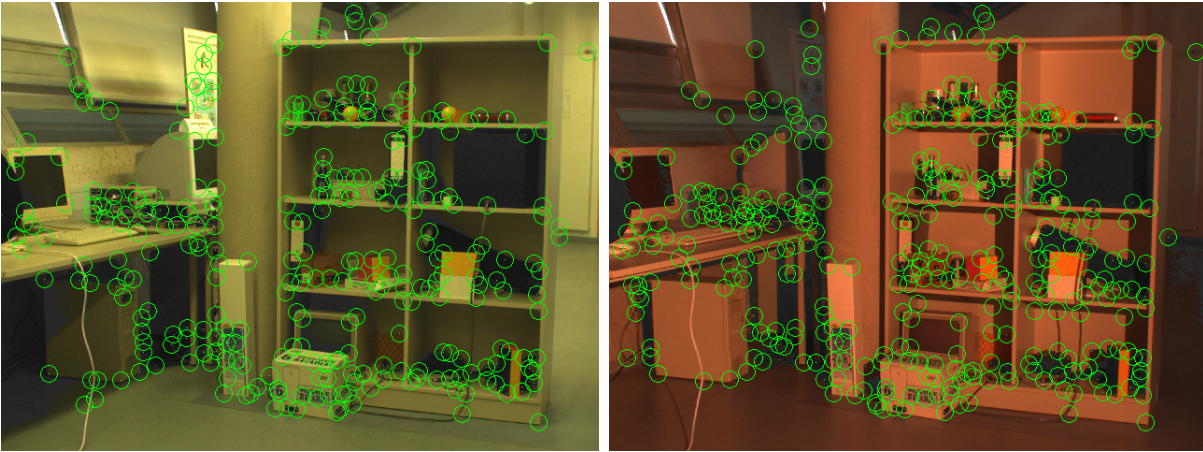


Figure 4.15: Detection example for the homomorphic colour Harris detector ($T = 10^{-4}$). The same images as in fig. 4.12 are used. 41.0% of the interest points of the left image are redetected in the right image. 65.0% of the interest points in the right image are false positives. Both images are gamma corrected for better visualisation ($\gamma = 1.5$).

The convolutions with the Gaussian and the derivatives of Gaussian filters are implemented as indicated in subsection 2.2.2. The preprocessing with the box filter is implemented like for the grey value detector H-HD. This is described in section 3.3.

The stability of the homomorphic colour Harris detector is illustrated in fig. 4.15 on the same images as in fig. 4.12. No automatic white balancing is applied. HC-HD is sensitive to intensity and colour edges, therefore more interest points are detected than with RI-HD. In comparison to C-HD, interest points are detected in all image areas, independently of the local contrast. This results in a higher detection stability: 41% of the interest points in the left image are redetected in the right image. 65% of the interest points in the right image are false positives. Fig. 4.15 also shows that HC-HD is less sensitive to noise than RI-HD because both intensity and chrominance informations are used for detection. With the described implementation, HC-HD requires 826ms for an image with 640×480 pixels (see subsection 2.2.2 for details on the computer), this is 2.01 times the processing time of HD and 1.21 times the processing time of C-HD.

4.5 M space interest point detector

Sharp shadow and shading edges are one source of instability for the interest point detectors presented in chapter 3, because illumination influence is assumed to vary slowly between pixels for all these detectors. The homomorphic colour Harris detector (HC-HD) uses the same assumption. Sharp shadow and shading edges can be compensated for colour images as explained in subsection 2.3.2. For this, the image formation model of eq. (4.2) or eq. (4.4) is used without any restricting assumption on the intensity factors

i_b and i_s . Specularities are not compensated here. This corresponds to the model of eq. (4.4): $C^j = i_b L^j c_b^j$ for $j = R, G, B$. The intensity factor i_b affects all colour channels identically. It can be suppressed by building ratios between colour. This principle is used by many invariant colour spaces (cf. subsection 2.3.2). In addition to compensating all shadow and shading edges, the interest point detector should also be invariant to illumination colour without any prior manual white balancing. Among all invariant colour features that provide these properties, the m space is particularly interesting because it is based on a local implicit compensation of light colour. The other invariant features use more or less elaborate automatic white balancing methods and correct illumination colour globally: the same correction is applied to all pixels. The m space is therefore more flexible and can better handle scenes lighted by more than one light source.

The m space is used in [GS99, NG98, MKK00, TT04]. Its invariance properties are presented in details in [GS99]. It is shown to provide good illumination invariance in the context of interest point detection in [TT04]. Shadow and shading effects are suppressed by building ratios between colour channels. Hence, also sharp shadow or shading edges disappear. Light colour is assumed to vary slowly and it is therefore compensated using neighbouring pixel values. The m space components are defined by:

$$m_1 = \frac{C^R(x_1, y_1) C^G(x_2, y_2)}{C^G(x_1, y_1) C^R(x_2, y_2)}, m_2 = \frac{C^R(x_1, y_1) C^B(x_2, y_2)}{C^B(x_1, y_1) C^R(x_2, y_2)}, m_3 = \frac{C^B(x_1, y_1) C^G(x_2, y_2)}{C^G(x_1, y_1) C^B(x_2, y_2)}. \quad (4.13)$$

(x_1, y_1) and (x_2, y_2) are two neighbouring pixels. These components are invariant to light colour L^j and intensity factor i_s for Lambertian scenes:

$$m_1 = \frac{i_b(x_1, y_1) L^R(x_1, y_1) c_b^R(x_1, y_1) i_b(x_2, y_2) L^G(x_2, y_2) c_b^G(x_2, y_2)}{i_b(x_1, y_1) L^G(x_1, y_1) c_b^G(x_1, y_1) i_b(x_2, y_2) L^R(x_2, y_2) c_b^R(x_2, y_2)} = \frac{c_b^R(x_1, y_1) c_b^G(x_2, y_2)}{c_b^G(x_1, y_1) c_b^R(x_2, y_2)}. \quad (4.14)$$

Eq. (4.14) is true because light colour is assumed to vary slowly between pixels, so that $L^j(x_1, y_1) = L^j(x_2, y_2)$ for $j = R, G, B$. The invariance of m_2 and m_3 is derived similarly. Eq. (4.14) shows that the m space components are equal to one except near edges between areas with different colours because pixels (x_1, y_1) and (x_2, y_2) must have different colours. All intensity edges, including shadow and shading effects, are suppressed.

When the logarithmic m space is used, multiplications and divisions are transformed to additions and subtractions. This results in faster computation. In addition, the robust implementation developed for the homomorphic detectors (H-HD and HC-HD) can be used to reduce noise sensitivity. The transformation from the RGB space is given by:

$$\ln m_1 = (\ln C^R(x_1, y_1) - \ln C^G(x_1, y_1)) - (\ln C^R(x_2, y_2) - \ln C^G(x_2, y_2)). \quad (4.15)$$

$\ln m_2$ and $\ln m_3$ are obtained similarly. The components of the logarithmic m space are equal to zero except when (x_1, y_1) and (x_2, y_2) have different colours. Eq. (4.14) shows that the invariance properties of the m space are true as long as the two pixels (x_1, y_1) and (x_2, y_2) are in the neighbourhood of each other, because the restricting factor for the invariance properties is the spatial variation of light colour (L^R, L^G, L^B). Therefore any high-pass linear filtering of $\ln C^j - \ln C^k$ with $j \neq k$ yields illumination invariant

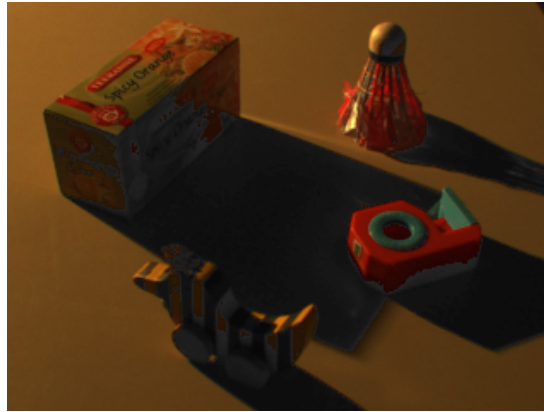
4 Illumination invariant interest point detection for colour images

information. As a result, the derivatives of $\ln C^R - \ln C^G$, $\ln C^R - \ln C^B$ and $\ln C^B - \ln C^G$ are invariant to shadows, to shading and to light colour for matte surfaces. This is used here for illumination invariant interest point detection.

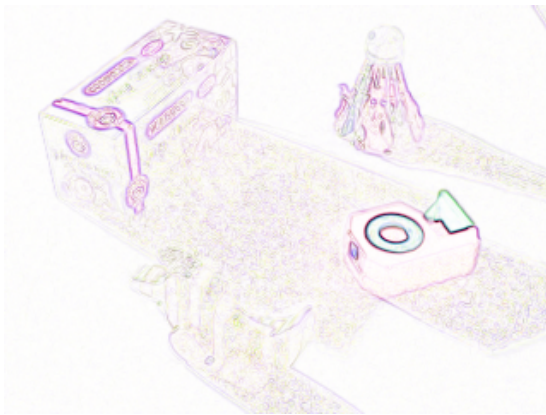
Like for the homomorphic detectors, noise sensitivity is an important topic, especially in the dark image areas. To reduce noise influence and to handle pixels with one of the channel values equal to zero, $\ln(C^j + 1)$ is used instead of $\ln C^j$ to compute the channel logarithm. As explained in subsection 2.3.2, colour features based on channel ratios are very sensitive to noise and to colour artifacts. As colour artifacts occur especially near edges, they influence detection stability strongly because the Harris detector is very sensitive to noise near edges. Hence, interest point detector based on chrominance (m space detector and RI-HD) are more sensitive to colour artifacts than detectors based on the full colour information (C-HD and HC-HD). In this thesis, the WACPI demosaicing algorithm is applied to reduce the formation of colour artifacts (see section 4.1). In addition, special preprocessing is introduced in section 4.6 to further reduce the remaining colour artifacts. Two versions of the m space detector are compared in this thesis. The first one uses the simple preprocessing applied for HC-HD (box filtering for dark pixels, see section 4.4). To reduce colour artifacts, automatic white balancing is applied before demosaicing (see section 4.1). The second version uses the special preprocessing presented in section 4.6 (no white balancing is applied).

Fig. 4.16 illustrates how the illumination influence on the colour gradient is compensated by the invariant algorithms presented in this chapter. The two versions of the m space are used, to show the influence of the preprocessing on the results. For all algorithms, the gradient has similar values in both bright and dark image areas. This means that the low frequency intensity influence is well compensated. The homomorphic algorithm cannot suppress high frequency intensity influence: the sharp shadow and shading edges appear grey, like intensity edges. The m space and the robust invariant algorithms attenuate well shadow and shading edges. Those are however not completely suppressed, due to unmodelled effects like specularities, colour artifacts and coloured shadows. The m space and the robust invariant algorithms are both more noise sensitive than the homomorphic algorithm in dark areas. The edges do not have the same colour in m space and robust invariant images because both algorithms rely on different principles to compensate light colour. For both images (b) and (d), automatic white balancing is applied before demosaicing to reduce the formation of colour artifacts. As a result, the shadow edges are less visible in images (b) and (d) than in image (e). On the other hand, the preprocessing applied in image (e) reduces best the noise and colour artifact influence in dark areas. The robust invariant detector (image (b)) is more influenced by noise and artifacts in dark areas than the developed m space gradient (images (d) and (e)).

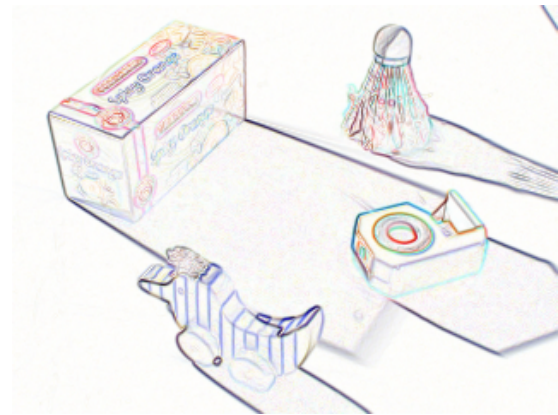
For better visualisation of the preprocessing influence, fig. 4.17 shows an enlarged part of the m space gradient images of fig. 4.16. The shadow edge in the top part of the image is more visible when no white balancing is applied. On the other hand, the preprocessing method of section 4.6 reduces well the influence of noise and colour artifacts in dark areas (especially on the black stripes of the object). Without accurate demosaicing and without preprocessing, the m space is not stable enough for any application, as shown in



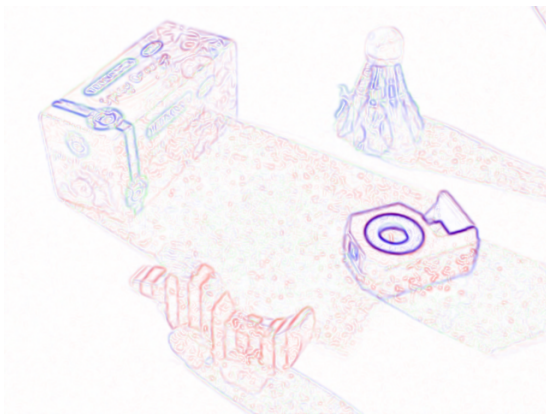
(a)



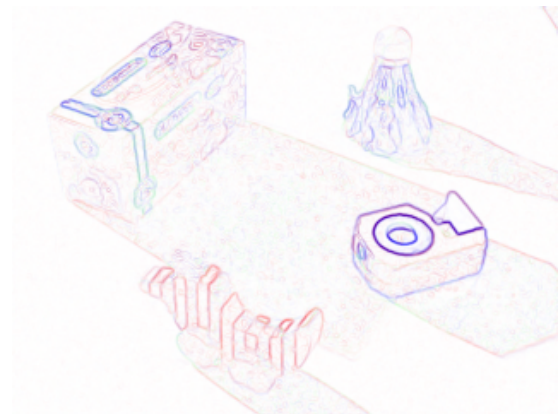
(b)



(c)



(d)



(e)

Figure 4.16: Suppression of the illumination influence on colour derivatives. (a) colour image (gamma corrected). (b) robust invariant gradient (see section 4.3). (c) homomorphic colour gradient. (d) m space gradient with preprocessing as for HC-HD and with white balancing before demosaicing. (e) m space gradient with the preprocessing designed in section 4.6). Gradient images are scaled between 0 and 255 (the maximum and minimum values over all pixels and all channels are used).



Figure 4.17: Influence of the preprocessing on the m space for a detail of the gradient images in fig. 4.16. Left: m space gradient with preprocessing as for HC-HD and with white balancing before demosaicing. Right: m space gradient with the preprocessing designed in section 4.6.

preliminary work described in [Mar02].

Noise influence can be further reduced by taking into account only the two least noisy components of the logarithmic m space in the structure matrix \mathbf{M} . The same interest points are detected as with three components because of the high correlation between colour channels. This also reduces computation time, as four derivatives instead of six are computed. In current cameras, the green channel has typically the lowest noise and the blue channel has typically the highest noise. This is caused by colour filter width and by sensor chip sensitivity. As a result, the pixel values in the green channel are generally the highest hence the most reliable, and the pixel values in the blue channel are the lowest hence the least reliable. In addition, most current single chip cameras use the Bayer pattern, in which the green channel is sampled twice as often as the red and the blue channels. Therefore green values are less affected by demosaicing artifacts (see section 4.1). To summarise, the green channel is the least noisy channel in current colour images. The two least noisy channels in the logarithmic m space are therefore $\ln m_1$ and $\ln m_3$. In experimental evaluations, the two channel m space detector achieved similar stability than the three channel m space detector, independently of the colours present in the scene. As a result, only the two channel m space detector is evaluated in section 4.8.

The m space Harris detector (MS-HD) is the only detector developed in this thesis which is invariant to all shadow or shading edges. It has the same invariance properties as the

robust invariant Harris detector (RI–HD) (see section 4.3), but no prior white balancing is required. Light colour is locally corrected with the m space which is a further advantage because scenes lighted by several light sources with different colours (for example sunlight and lamps) are better handled. The m space Harris detector stays however sensitive to edges caused by coloured shadows. The MS–HD algorithm is described by:

1. Apply preprocessing to reduce noise (same method as for HC–HD or method presented in section 4.6).
2. Perform the logarithmic transformation with: $l^j = \ln(1 + C^j)$ for $j = R, G, B$.
3. Convolve $(l^R - l^G)$ and $(l^B - l^G)$ with the derivative of Gaussian filters: $(l^i - l^G)_x = G_x(\sigma_{deriv}) \otimes (l^i - l^G)$ and $(l^i - l^G)_y = G_y(\sigma_{deriv}) \otimes (l^i - l^G)$ for $i = R, B$.
4. Compute the illumination invariant structure matrix with:

$$\mathbf{M} = G(\sigma_M) \otimes \sum_{i=R,B} \begin{bmatrix} (l^i - l^G)_x^2 & (l^i - l^G)_x (l^i - l^G)_y \\ (l^i - l^G)_x (l^i - l^G)_y & (l^i - l^G)_y^2 \end{bmatrix}. \quad (4.16)$$

5. Compute the cornerness function CF with eq. (4.6).
6. (x, y) is an interest point:
 - if it is a local maximum of the cornerness function CF
 - and if $CF(x, y) > T$ ($T > 0$).

The detection threshold T should be set to detect an appropriate number of interest point. If the three m space channels are used, one term is added to the sum in eq. (4.16) for the chrominance channel $l^R - l^G$. Like RI–HD, MS–HD is only sensitive to chrominance. Therefore, interest points are only detected near boundaries between areas of different colours. As a consequence, MS–HD is also inappropriate when the images do not contain enough colour information. The invariant derivatives computed in step 3 of the MS–HD can be used to compute illumination invariant descriptors. The principle can be easily extended to other versions of the Harris detector, like the scale or affine invariant detectors, and also to other interest point detectors based on first or second derivatives.

The preprocessing method and its implementation are discussed in sections 4.4 and 4.6. The convolutions with the Gaussian and with the derivative of Gaussian filters are implemented as described in subsection 2.2.2.

The results of the m space Harris detectors are illustrated in fig. 4.18 on the same images as in fig. 4.12. The top images show the results for the MS–HD with the same preprocessing as the HC–HD and with white balancing before demosaicing. The bottom images show the results for the MS–HD with the preprocessing presented in section 4.6. This preprocessing attenuates colour artifacts, therefore no white balancing is performed before demosaicing. MS–HD is only sensitive to chrominance. As a result, less interest points are detected than with C–HD or HC–HD and these interest points are located near colour edges. Some false positives are detected in dark areas or near shadow or shading edges due to noise and

4 Illumination invariant interest point detection for colour images

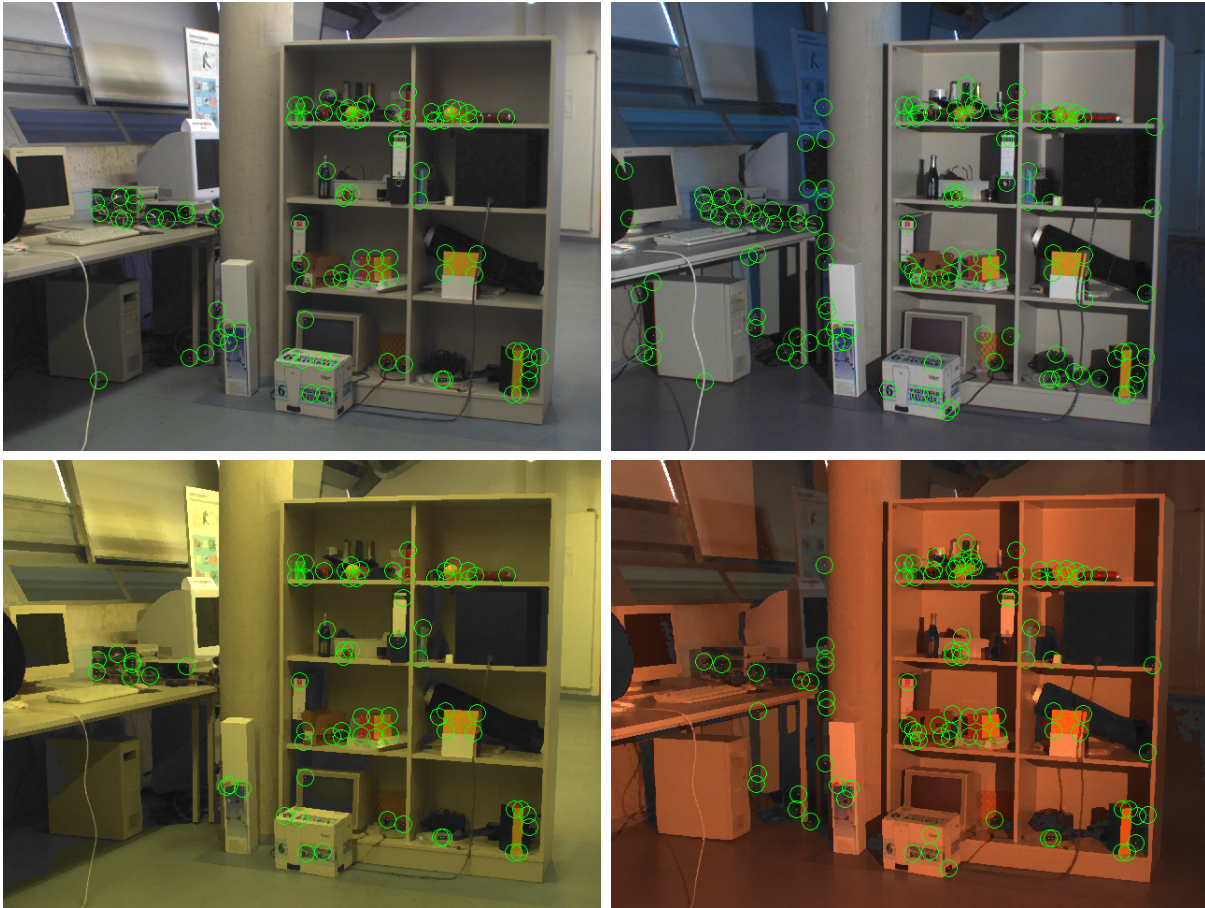


Figure 4.18: Detection example for the m space Harris detector ($T = 10^{-5}$) on the same images as in fig. 4.12. Top: detection results with white balancing before demosaicing and with the same preprocessing as HC-HD. 52.8% of the interest points of the left image are redetected in the right image. 63.6% of the interest points in the right image are false positives. Bottom: detection results with the preprocessing of section 4.6 (no white balancing). 57.8% of the interest points of the left image are redetected in the right image. 57.7% of the interest points in the right image are false positives. All images are gamma corrected for better visualisation ($\gamma = 1.5$).

colour artifacts. The preprocessing of section 4.6 reduces the sensitivity of the MS–HD to noise and colour artifacts: less false positives are detected. Both versions of the MS–HD are less noise sensitive than the RI–HD. With white balancing and preprocessing like for the HC–HD, 52.8% of the interest points in the left image are redetected in the right image and 63.6% of the interest points in the right image are false positives. With the preprocessing of section 4.6, 57.8% of the interest points in the left image are redetected in the right image and 57.7% of the interest points in the right image are false positives. Of all presented detectors, the MS–HD with the preprocessing of section 4.6 achieves the best stability on this image pair. Nevertheless, no perfect compensation of the local illumination conditions is achieved. With the described implementation, MS–HD requires 858ms for an image with 640×480 pixels if three chrominance channels are used, this is 2.08 times the processing time of HD and 1.26 times the processing time of C–HD. If only two chrominance channels are used, MS–HD requires 706ms, this is 1.71 times the processing time of HD and 1.04 times the processing time of C–HD (see subsection 2.2.2 for details on the computer).

4.6 Preprocessing for the M space detector

Even if the formation of colour artifacts can be reduced with accurate demosaicing, colour artifacts are not completely suppressed. Artifacts also occur when images are acquired with multi-chip cameras, due to misregistration of the chips (see [BMCF02b]). Last, colour artifacts may be caused by chromatic aberration of the camera optics. This means that the lens have different refracting indexes for different wavelengths. The focal length is hence slightly different for each colour channel, which results in colour artifacts near edges. Preprocessing should thus reduce both noise and colour artifacts. Colour artifacts are colour outliers: isolated pixels with a different colour from the neighbouring pixels. Edges should also be preserved to obtain stable interest point detection. Preprocessing has hence three goals: noise reduction, colour outlier elimination and edge preservation. The preprocessing for the homomorphic colour Harris detector does not fulfil these conditions: it only reduces noise in dark areas. Therefore several algorithms are presented and compared here to find the most appropriate preprocessing for the m space detector.

For grey value images, **median filtering** is often used for edge preserving smoothing. The extension of median filtering to colour images is not straightforward. If median filtering is applied to each channel separately (named marginal processing), colour artifacts may appear near edges because edges are not constrained to occur at the same pixels for all channels during filtering. Therefore, vector processing is better: pixel values are considered as colour vectors $\mathbf{C} = (C^R, C^G, C^B)$. Vector median filtering requires first the definition of a distance between colour vectors. The median value is the colour vector with the minimum cumulated distance to all other pixels in the filter kernel:

$$\mathbf{C}_{median} = \min_{j=1\dots N} \left(\sum_{i=1}^N \|\mathbf{C}_j - \mathbf{C}_i\| \right).$$

4 Illumination invariant interest point detection for colour images

$\mathbf{C}_i = (C_i^R, C_i^G, C_i^B)$ is the colour value of pixel i and $\{\mathbf{C}_1, \dots, \mathbf{C}_N\}$ represents all pixels in the filter kernel. $\|\mathbf{C}_j - \mathbf{C}_i\|$ is the chosen distance between \mathbf{C}_j and \mathbf{C}_i . Several colour median filters are compared for example in [KA01]. Colour median filters have the drawback of being computation intensive because the distance between every pixels in the filter kernel must be computed. The vector median filter based on the L_1 norm is evaluated here. The L_1 norm is defined as: $\|\mathbf{C}_j - \mathbf{C}_i\| = |C_j^R - C_i^R| + |C_j^G - C_i^G| + |C_j^B - C_i^B|$. A 3×3 kernel is used keep computing time small. The vector median filter requires 202ms for an image with 640×480 pixels on the computer described in subsection 2.2.2.

Edge preserving smoothing can also be achieved with **bilateral filtering** (see [TM98]). Bilateral filtering only uses pixels which are in the neighbourhood of each other and which have similar grey or colour values. As a result, the image is smoothed while edges are preserved. Bilateral filtering is applied in [TA02] to reduce noise influence on interest point detection. The Gaussian bilateral filter is evaluated here. Like for the vector median filter, kernel pixels are represented with subscript $i \in [1, N]$. The currently filtered pixel has subscript i_0 . The standard Gaussian filter is defined by following weights:

$$G_i = \frac{1}{k} e^{-d(i,i_0)^2/2\sigma^2} \quad \text{with} \quad i \in [1, N] \quad \text{and} \quad k = \sum_{i=1}^N G_i.$$

$d(i, i_0)$ is the Euclidian distance between kernel pixel i and the currently filtered pixel i_0 . σ is the filter standard deviation and k is the normalisation factor. With the same notations, the bilateral Gaussian filter is defined by following weights:

$$G_i(\mathbf{C}) = \frac{1}{k(\mathbf{C})} e^{-d(i,i_0)^2/2\sigma^2} e^{-s(\mathbf{C}_i, \mathbf{C}_{i_0})^2/2\sigma_s^2} \quad \text{with} \quad k(\mathbf{C}) = \sum_{i=1}^N G_i(\mathbf{C}).$$

$s(\mathbf{C}_i, \mathbf{C}_{i_0})$ is the similarity between the value \mathbf{C}_i of kernel pixel i and the value \mathbf{C}_{i_0} of the currently filtered pixel i_0 . σ_s is the standard deviation for pixel similarity. In contrary to the standard Gaussian filter, the weights of the bilateral Gaussian filter depends on the pixel values \mathbf{C} . Hence, the kernel weights are calculated for each pixel separately. This results in high computing time. The similarity measure s must be chosen adequately to attenuate noise and colour artifacts. The distance between the two C^G values is used: $s(\mathbf{C}_{i_0}, \mathbf{C}_i) = C_{i_0}^G - C_i^G$, because texture is more accurately reconstructed in the G channel (see section 4.1). In order to keep computing time small, a small spatial standard deviation is chosen: $\sigma = 1$, and the kernel weights are computed with a look-up table. A high similarity standard deviation is chosen to filter colour outliers: $\sigma_s = 30$. The bilateral Gaussian filter requires 853ms for an image with 640×480 pixels on the computer described in subsection 2.2.2.

Last, segmentation of the RGB values can also be applied in order to reduce noise and colour artifacts as in [BMCF02b, MKK00]. Pixel values are replaced by the mean values of their class. This reduces the number of possible pixel values, hence decreasing noise and artifacts while preserving edges. The results depend however strongly on the chosen number of classes. The **Nagao filter** in [NM79] is an edge preserving smoothing filter for grey value images. It is based on a principle similar to segmentation and does not require

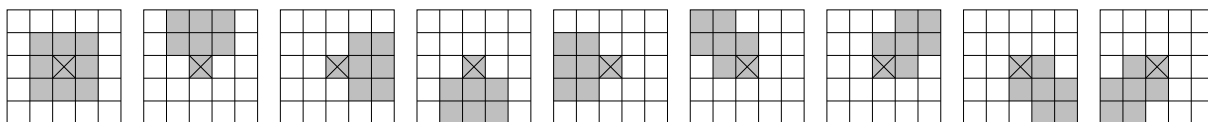


Figure 4.19: The nine neighbourhoods used in the original Nagao filter are shown in grey. The currently processed pixel is indicated with a cross.

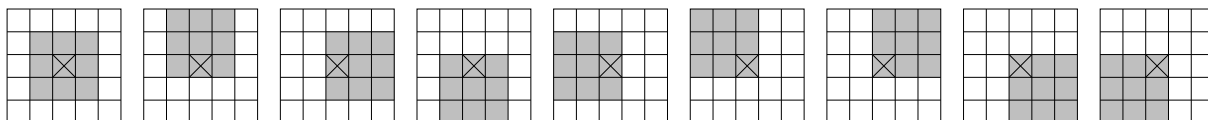


Figure 4.20: The nine neighbourhoods used in the simplified Nagao filter are shown in grey. The currently processed pixel is indicated with a cross.

any parameter. Nine neighbourhoods containing the currently filtered pixel are considered, as shown in fig. 4.19. The neighbourhood with the smallest variance is selected. All pixels within this neighbourhood are assumed to belong to the same class because of the small variance. The value of the currently filtered pixel is replaced by the mean value in this selected neighbourhood. The Nagao filter performs therefore a spatially constrained local segmentation. To adapt this filter to colour images, the cumulated variance for all channels is computed in the neighbourhoods: $\sigma^2 = \sigma_R^2 + \sigma_G^2 + \sigma_B^2$. In addition, the simplified neighbourhoods shown in fig. 4.20 are used. All neighbourhoods in fig. 4.20 have the same form so means and variances can be computed in a preliminary step with a 3×3 box filter. This results in a very fast preprocessing because the box filter can be implemented as two sequential recursive one dimensional filters (see section 3.2). This simplification results in a slightly worse texture preservation, but the resulting image is good enough for this work. The simplified Nagao filter requires 430ms for an image with 640×480 pixels on the computer described in subsection 2.2.2.

The results of the vector median filter, of the bilateral Gaussian filter and of the simplified Nagao filter are shown in fig. 4.21 on an enlarged image detail. The hue component of the image detail is also shown in fig. 4.21, because it allows a good visualisation of colour artifacts. All preprocessing methods attenuate colour outliers. The bilateral Gaussian filter produces a blurred image due to the high similarity standard deviation σ_s . The similarity standard deviation cannot be decreased because colour outliers are not attenuated otherwise. The bilateral filter is indeed optimal for noise reduction with texture preservation, but not for outlier attenuation. Due to its very high computing time and to this blurring, the bilateral filtering will not be used as preprocessing. Both vector median and simplified Nagao filters suppress completely outliers. They also sharpen edges. The simplified Nagao filter does not preserve texture as well as the median filter: small details are attenuated. On the other hand, the median filter does not smooth as well as the simplified Nagao filter. These properties of both filters are also shown on the m space gradient in fig. 4.22: the simplified Nagao filter smooths more in homogeneous areas and attenuates more outliers, while the vector median filter preserves better image texture. In comparison to the m space gradient obtained without preprocessing, the colour of the gra-

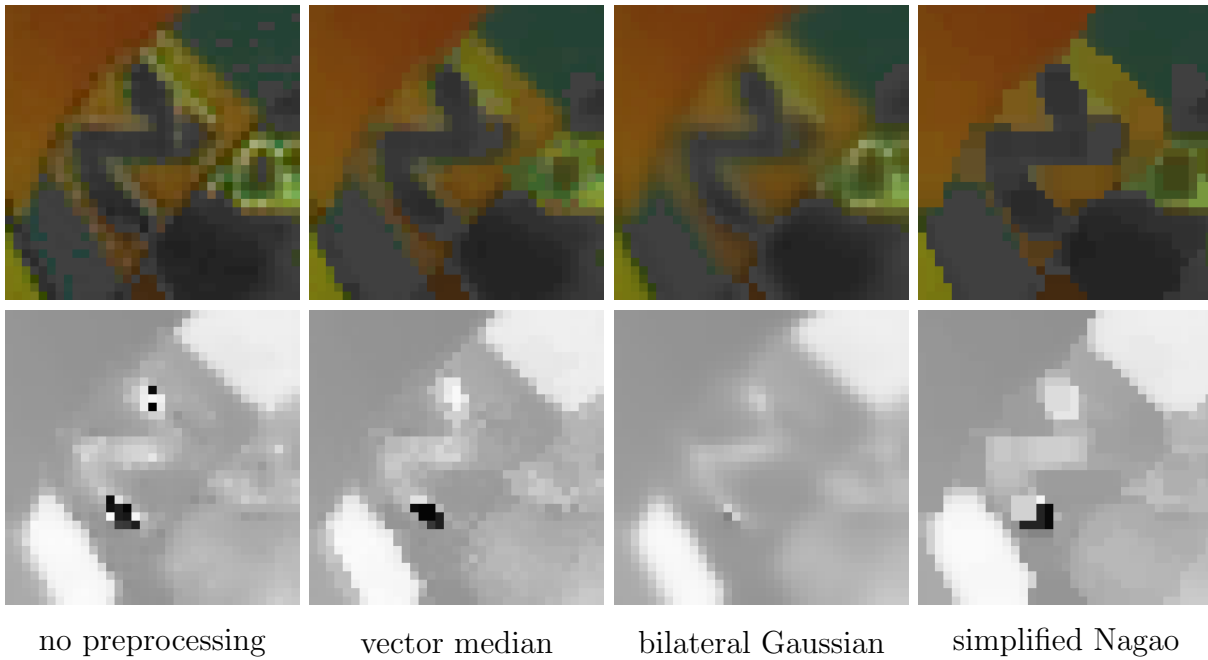


Figure 4.21: Results of the preprocessing methods on an enlarged image detail. Top: gamma corrected image detail ($\gamma = 2$). Bottom: normalised hue component. The theoretical range of hue $[-\pi, \pi]$ is mapped to $[0, 255]$. Therefore, 0 and 255 encode similar hues.

dient along edges is more homogeneous (see the long green edge in fig. 4.22) and the noise in nearly homogeneous areas is reduced. The simplified Nagao filter attenuates better noise and colour outliers than the vector median filter. The obtained gradient is accurate, even though texture is less preserved than with median filtering. As a consequence, the simplified Nagao filter is chosen as preprocessing method in this work.

4.7 Comparison framework

The developed interest point detectors are evaluated and compared to each other and to state of the art detectors. The comparison framework is similar to the one in section 3.7. The detection stability is evaluated on image series showing scenes under different illumination conditions. A reference image is chosen in each series. The interest points detected in reference and in current images are compared to evaluate detection stability with redetection rate and false positive rate (defined in subsection 3.7.1). A reference interest point is considered redetected if an interest point is detected in its 3×3 neighbourhood in the current image. The saturated areas are handled as described in section 3.6. An overview of the compared detectors and their parameters is given in subsection 4.7.1. The used image series are presented in subsection 4.7.2.

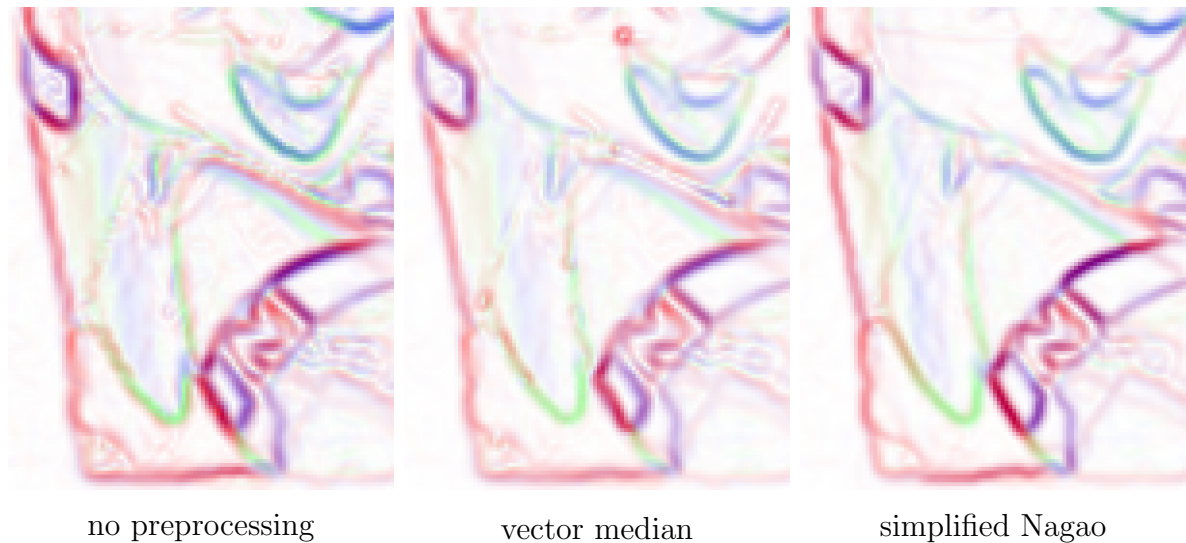


Figure 4.22: Influence of the two best preprocessing methods on the m space gradient. The gradient images are scaled between 0 and 255 (the maximum and minimum values over all pixels and all channels are used). The image in 4.21 is a part of this image (the letter N).

4.7.1 Compared interest point detectors

The performance of the new detectors developed in this chapter are evaluated: the colour homomorphic Harris detector (HC–HD) and the two proposed versions of the m space Harris detector. In the first version, automatic white balancing is applied before demosaicing to reduce the formation of colour artifacts and the simple preprocessing of HC–HD is applied in step 1 of the MS–HD. It is named here m space Harris detector with white balancing (MS–HD+WB). In the second version, the Nagao preprocessing chosen in section 4.6 is applied in step 1 of the MS–HD and no automatic white balancing is applied. It is named m space Harris detector with Nagao preprocessing (MS–HD+N). For both versions, only 2 chrominance channels are used. In addition, they are compared to two existing interest point detectors for colour images: the colour Harris detector (C–HD) in [Gou00] and the robust invariant Harris detector (RI–HD) in [vdW05]. C–HD adapts detection to the overall image contrast by selecting the N interest points with the highest cornerness values. By including C–HD in the comparison, the stability increase yielded by the new detectors can be assessed. RI–HD has similar invariance properties to those of the m space Harris detector (MS–HD). A comparison between RI–HD and MS–HD evaluates noise sensitivity and white balancing influence on detection stability. Last, the best algorithm developed for grey value images (cf. chapter 3) is included in the comparison to assess the utility of colour information for interest point detection. It is the homomorphic Harris detector (H–HD). An overview of the 6 compared detectors is given in table 4.3.

For all methods, saturated areas are handled as described in section 3.6. The following parameters are used to compute the cornerness function for all detectors: $\sigma_{deriv} = 1.2$, $\sigma_M = 3$ and $\alpha = 0.06$. As in chapter 3, each detector applies the same user-defined

detector name	abbreviation	description	computing time
colour Harris detector	C-HD	section 2.2.3	681 ms
robust invariant Harris detector	RI-HD	section 4.3	1161 ms
homomorphic Harris detector	H-HD	section 3.3	457 ms
homomorphic colour Harris detector	HC-HD	section 4.4	826 ms
m space Harris detector with white balancing	MS-HD+WB	section 4.5	716 ms
m space Harris detector with Nagao preprocessing	MS-HD+N	section 4.5	1136 ms

Table 4.3: Overview of the evaluated interest point detectors.

detection threshold for all images. This detection threshold is chosen manually as a compromise for all scenes. For C-HD, $N = 100$ interest points are detected. For RI-HD, T is set to 10^{-7} . For H-HD, T is set to 10^{-5} like in chapter 3. For HC-HD, T is set to 10^{-4} . For MS-HD+WB and MS-HD+N, T is set to 10^{-5} . All other parameters values are given in the respective algorithm descriptions.

4.7.2 Image data set

Several image series are used to evaluate the detector stability under illumination changes. Each series shows one scene under different illumination conditions. The images are acquired as described in more details in subsection 3.7.3. The BASLER A302fc colour CCD camera is used, without gamma correction or white balancing. Gain and brightness are set to the values given by the manufacturer. Only aperture and shutter time are set manually for each image. The raw camera signal is demosaiced with WACPI (see [LT03]) for all images, as explained in section 4.1. For RI-HD and MS-HD+WB, white balancing with the white patch algorithm is applied before demosaicing.

Like in chapter 3, the detectors are first compared on image series with simple illumination changes. In these series, all pixels are influenced identically by the illumination change. As a result, a simple adaptation to the overall lighting conditions as performed by the C-HD is sufficient for stable detection. These series allow to assess the noise sensitivity of the developed detectors. The same series as in chapter 3 are used and the same reference image is chosen. For the first series, no illumination changes occur. Noise is the only source for pixel variations. In the second image series, pixel values are influenced by neon lamp flickering. Finally, the third image series is obtained by varying the camera shutter time. All three series show the same scene, which is illustrated in fig. 4.23.

The detectors are also evaluated on image series showing scenes under complex illumination changes. Type, number, position and orientation of the light source(s) are changed. Realistic illuminants and scenes are used (see subsection 3.7.3 for more details). Like in



Figure 4.23: Two images of the series with shutter time variations. For visualisation, the images are gamma corrected ($\gamma = 1.4$).

chapter 3, the reference image is chosen manually: it is the image with the most uniform illumination. The evaluation has been performed on many series showing different kind of scenes (same image data as in chapter 3). As in chapter 3, the stability of the detectors is similar for scenes with similar properties (simple or complex 3D geometry, textured or structured reflectance, diffuse or specular surface, saliency of the colour information). Therefore, the results are given in section 4.8 for typical image series which have been selected out of all acquired image series. Series with redundant results are not presented. The reference image and two sample images of each selected series are shown in fig. 4.24. The first image series is the *shelves* series, which is also used in chapter 3. It has a complex 3D geometry and hence is strongly influenced by shadows and shading. Its reflectance is rather complex too. The next image series is the *box* series. It shows detection stability on a scene with simple 3D geometry and structured reflectance. In addition, the object is very specular. The next two objects show how the presence of colour and intensity edges in the object reflectance influences detection. The *giraffe* series contains an object with only colour edges, whereas the scene in the *box2* series contains very salient intensity edges. The last two series *rabbit* and *snoopy* show detection stability on textured objects (objects for which many interest points are similar).

4.8 Detector evaluation and comparison

The evaluation and comparison results are presented in this section. In subsection 4.8.1, the results for the image series with simple illumination changes are presented. Subsection 4.8.2 gives the results for the image series with complex illumination changes are presented. Finally, a conclusion closes the evaluation in subsection 4.8.3.

4.8.1 Simple illumination changes

The evaluation results for image series with simple illumination changes are given in table 4.4 and in fig. 4.25. In the first image series (top of table 4.4), noise is the only source for pixel variations. Hence the detector sensitivity to noise can be assessed. C-HD, H-HD and

4 Illumination invariant interest point detection for colour images

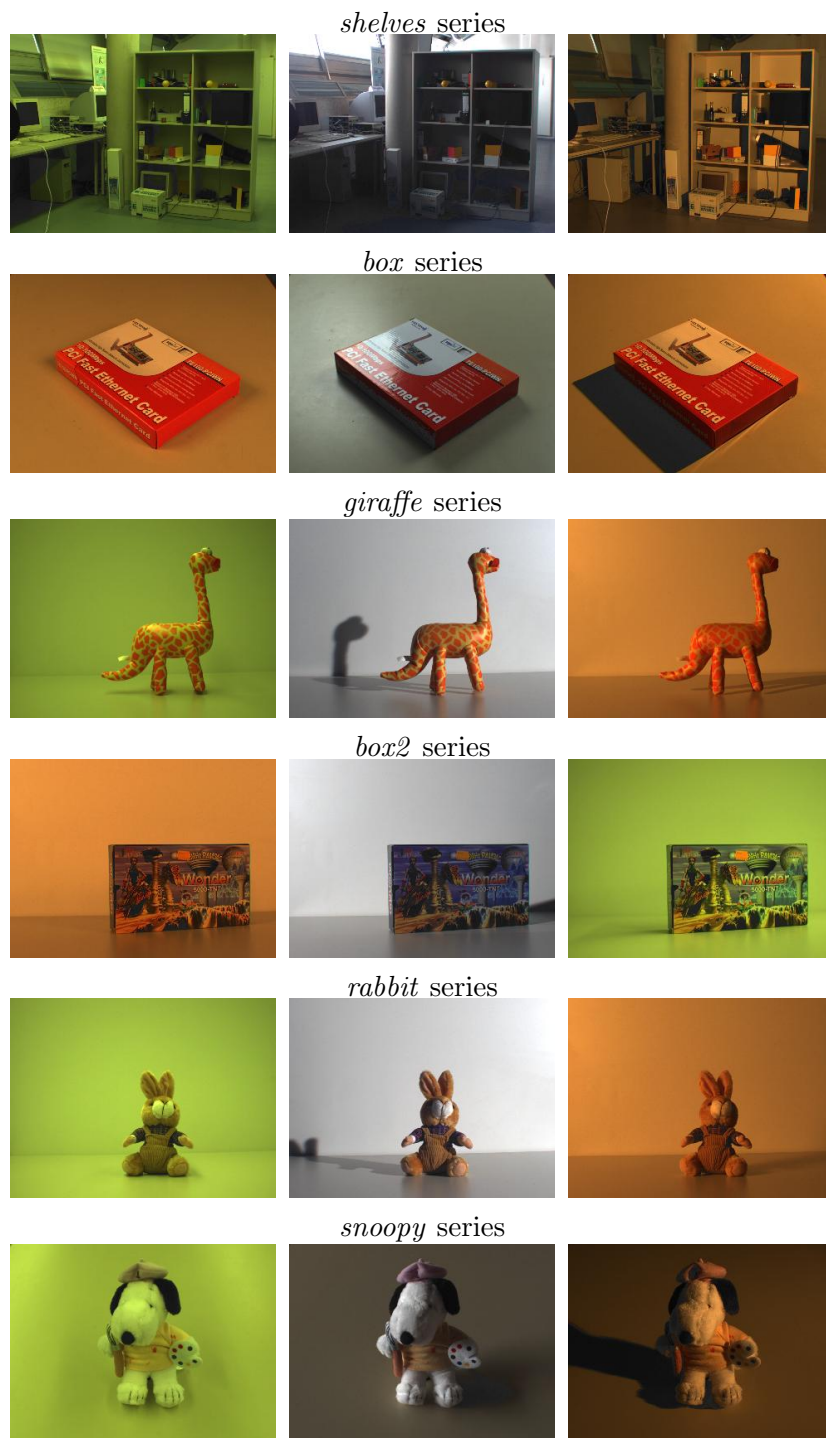


Figure 4.24: Sample images of the series with complex illumination variations. For all image series, the reference image of the series is shown on the left. For visualisation, the images are gamma corrected ($\gamma = 1.4$). For better visualisation of the colours, manual white balancing has been applied to all middle images.

Image series with noise as only source for pixel variations						
	C-HD	H-HD	HC-HD	RI-HD	MS-HD+WB	MS-HD+N
mean redetection rate	0.974	0.980	0.964	0.836	0.870	0.933
mean false positive rate	0.0005	0.0388	0.0320	0.165	0.154	0.0982
Image series with neon flickering as main source for pixel variations						
	C-HD	H-HD	HC-HD	RI-HD	MS-HD+WB	MS-HD+N
mean redetection rate	0.950	0.989	0.970	0.920	0.887	0.932
mean false positive rate	0.0283	0.0506	0.0465	0.118	0.104	0.108

Table 4.4: Evaluation results for the series with small illumination changes. In the first series, no illumination changes occur: noise is the only source for pixel variations. In the second series, neon flickering is the main source for pixel variations. The mean rates for the whole series are given (50 images).

HC-HD all have similar stability. The detectors based only on chrominance information (RI-HD, MS-HD+WB and MS-HD+N) are less stable: their redetection rates are lower and their false positive rates are higher. This is caused by the higher sensitivity of chrominance to noise and colour artifacts. The Nagao preprocessing increases stability: MS-HD+N is more stable than RI-HD and MS-HD+WB. RI-HD achieves the lowest stability because it is very noise sensitive in dark image areas such as here the shadows and the black binders in the shelf (see fig. 4.23). This noise sensitivity is also visible in fig. 4.13: many false positives are detected by RI-HD in dark image areas.

The results of the second series are presented at the bottom of table 4.4. Pixel variations are caused by neon lamp flickering and noise. All detectors can compensate accurately this illumination change. Only the stability of C-HD decreases slightly. Neon lamps produce a more uniform lighting in this series than sunlight in the previous series. As a result, shadows and dark objects appear less dark in the images. This leads to a higher stability for RI-HD and MS-HD+WB. As before, the detectors using intensity (C-HD, H-HD and HC-HD) are more stable than the detectors based only on chrominance (RI-HD, MS-HD+WB and MS-HD+N), because they are less noise sensitive.

Fig. 4.25 shows the result for the series with varying shutter time. Illumination intensity varies identically for all pixels. Like in the previous series, the highest stability is obtained by the detectors using intensity (C-HD, H-HD and HC-HD). The results of H-HD and HC-HD are similar. Both are slightly less stable than C-HD for the darkest images (smallest image numbers) because the local adaptation to lighting conditions increases noise sensitivity in dark areas. RI-HD, MS-HD+WB and MS-HD+N are all less stable than C-HD, H-HD and HC-HD, especially for darker images: redetection rate is smaller and false positive rate is higher. Like in the first image series, RI-HD achieves the lowest stability due to its high noise sensitivity in dark image areas. The implementation of MS-HD+WB and MS-HD+N, using $\ln(C^i + 1)$ and preprocessing to reduce noise influence,

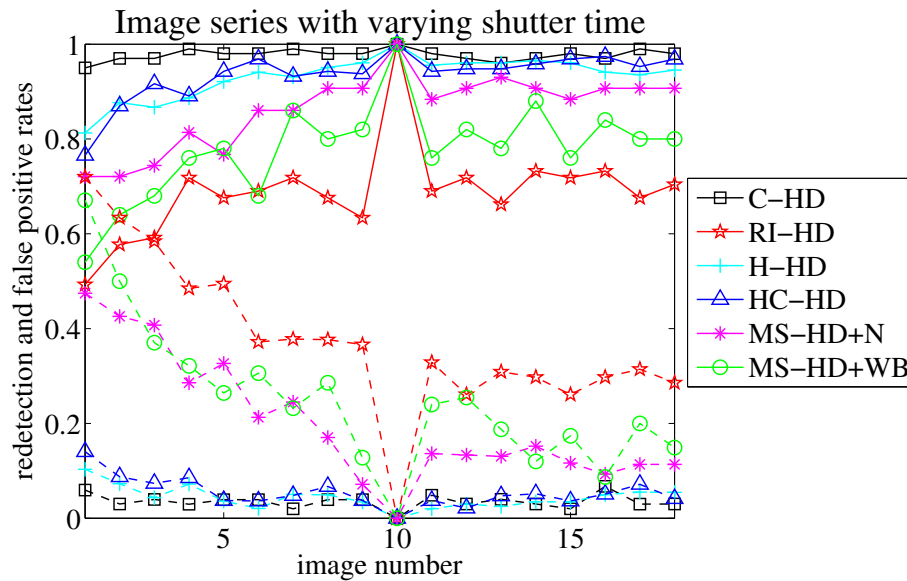


Figure 4.25: Evaluation results for the series with increasing shutter time. Darker images have smaller image numbers. Redetection rates are indicated with straight lines. False positive rates are indicated with dashed lines.

makes both detectors more robust to noise than RI-HD. The Nagao preprocessing reduces very well noise sensitivity: MS-HD+N is more stable than MS-HD+WB and RI-HD.

For every images series in this subsection, all image pixels are influenced identically by the illumination variations. Therefore, all detectors, even the simple C-HD, can compensate the illumination changes: all detectors achieve high stability on all series. The results show that the detectors based only on chrominance (RI-HD, MS-HD+WB and MS-HD+N) are less stable than the detectors using also intensity (C-HD, H-HD and HC-HD), especially when the image contains dark areas. This is caused by the noise sensitivity of chrominance for dark pixel values. RI-HD is the least stable detector in the presence of dark areas. Stability can be increased with a robust implementation like in MS-HD+WB. The Nagao preprocessing is particularly efficient to increase stability because it reduces noise and colour artifacts: MS-HD+N reaches higher redetection rates and lower false positive rates than MS-HD+WB and RI-HD. H-HD and HC-HD both have similar stability to C-HD. They are slightly less stable than C-HD for very dark images due to the local adaptation to the lighting conditions.

4.8.2 Complex illumination changes

This subsection presents the detector stability under complex illumination changes. Type, number, position and orientation of the light source(s) are varied. The used series are described in subsection 4.7.2. Further results are given in [Fai05b]. The main motivation for interest point detection with colour information is the higher accuracy of the image formation model for colour images. Light colour changes, shadow and shading effects can

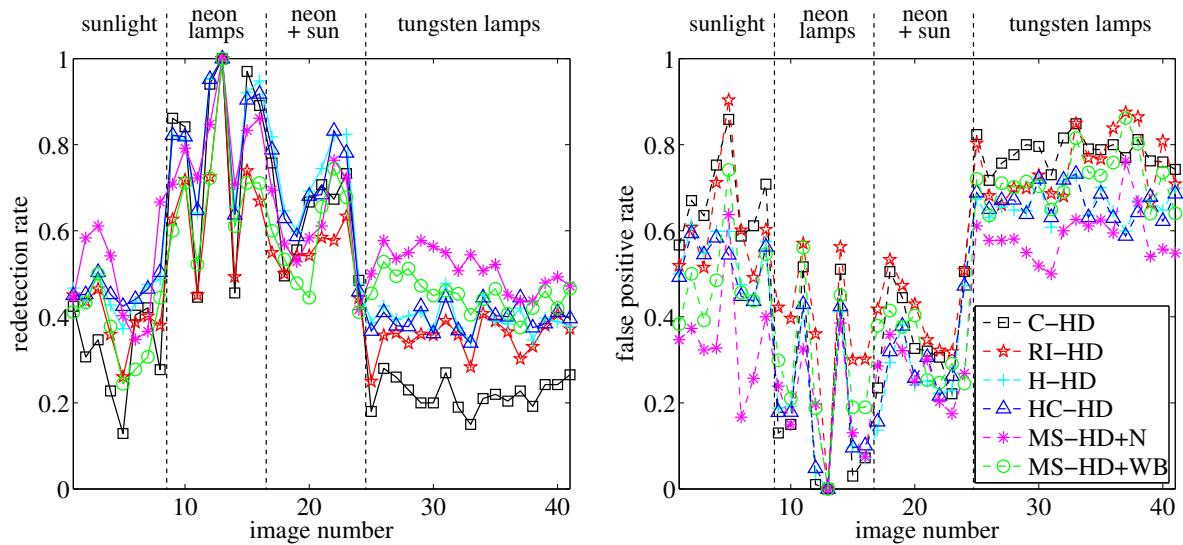


Figure 4.26: Evaluation results for the *shelves* series (Fig. 4.24 shows images 13, 7 and 28). The redetection rate is given in the left diagram and the false positive rate is given in the right diagram. The illumination type is indicated at the top of the diagram.

hence be better compensated. Therefore, the stability under tungsten lamps or after a change of illuminant type are of special interest here: this results in the most complex influences in images.

The scene of the *shelves* series has a complex 3D geometry. Shadow and shading effects have hence a strong influence on the images. The evaluation results are presented in fig. 4.26 (see fig. 3.24 for the stability of the grey value detectors). MS-HD+N achieves the best overall stability. All detectors achieve similar redetection and false positive rates when the illuminant type stays similar (here neon lamps and neon + sun). MS-HD+N is more stable than the other detectors when the illuminant type changes (sunlight or tungsten lamps): it has either a higher redetection rate or a lower false positive rate or both. MS-HD+WB achieves the second best stability. It is less stable than MS-HD+N, especially under tungsten lamps because it is more sensitive to noise in dark areas. RI-HD achieves the second worst stability because many false positives are detected in dark areas like shadows (see also fig. 4.13). To conclude, invariance to sharp shadow and shading edges increases detection stability on this series, when the implementation is robust to noise and colour artifacts. H-HD and HC-HD achieve very similar stability. This shows that light colour does not influence much detection stability (the main difference between HC-HD and H-HD is a better compensation of light colour). C-HD is the least stable because it only adapts detection to the overall illumination intensity. Interest points are hence detected in the areas with the highest local contrast. These areas change position under complex lighting changes.

The detection stability for a scene with simple 3D geometry is shown in fig. 4.27. The scene contains a rectangular box and is lighted by a single illuminant type (the detector

4 Illumination invariant interest point detection for colour images

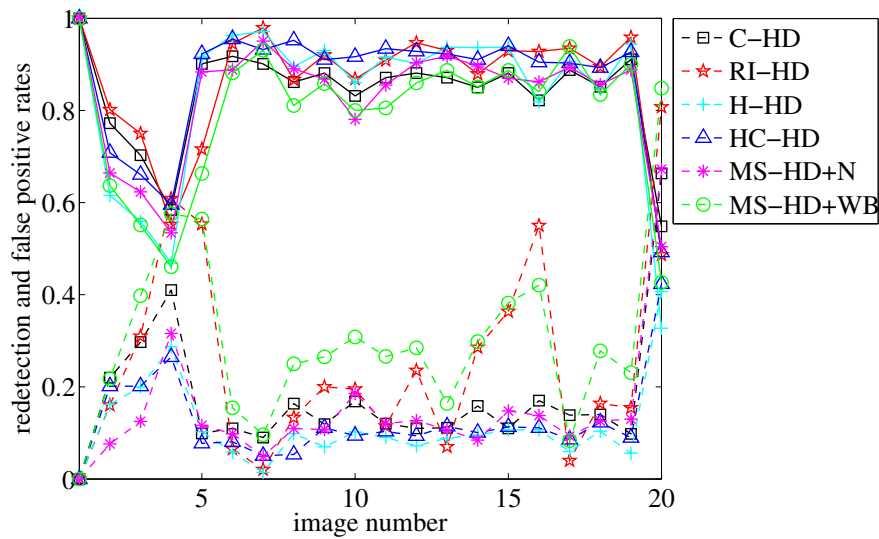


Figure 4.27: Evaluation results for the *box* series (fig. 4.24 shows images 1, 3 and 11). The redetection and false positive rates are indicated by continuous and dotted curve. The scene is lighted by tungsten lamps for all images.

stabilities stay similar when the illuminant type is changed). All detectors achieve similar and good stability because illumination influence is smaller than for scenes with complex geometry. For all detectors, the stability for the *box* series is higher than for scenes with complex 3D geometry like the *shelves* series. The higher noise sensitivity of RI-HD and MS-HD+WB in the presence of very dark shadows is visible for images 8 to 16: RI-HD and MS-HD+WB have higher false positive rates than the other detectors. This noise sensitivity is further increased because the automatic white balancing algorithm fails on some of those images, leading to a higher amount of colour artifacts. As a conclusion, like in chapter 3, simple detectors (here C-HD or grey value detectors) are sufficient for stable detection on such scenes with simple 3D geometry, because all detectors achieve similar stability. In addition, the effect of specularities is visible in images 2 to 5 and in image 20, which all contain large specular areas: all detectors have lower redetection rate and higher false positive rates for these images, because none can compensate specularities.

Fig. 4.28 shows how the presence of chrominance and intensity edges in the scene influences detection stability. As a result, the most appropriate detector depends on scene content. The *giraffe* series (top of fig. 4.28) displays an object with only coloured reflectance edges. In addition, the intensity difference between the two main colours (yellow and orange) is small. Hence, the most prominent intensity edges are due to specularities, shadow and shading effects. As a result, H-HD achieves the worst results because it uses only intensity information. C-HD achieves the second worst stability because detection is only adapted to the overall lighting intensity. HC-HD achieves a good stability increase in comparison to H-HD and C-HD because it uses colour information and adapts detection to the local lighting conditions. The best stability is reached by MS-HD+N, MS-HD+WB and RI-HD. All three achieve similar redetection and false positive rates. MS-HD+WB and

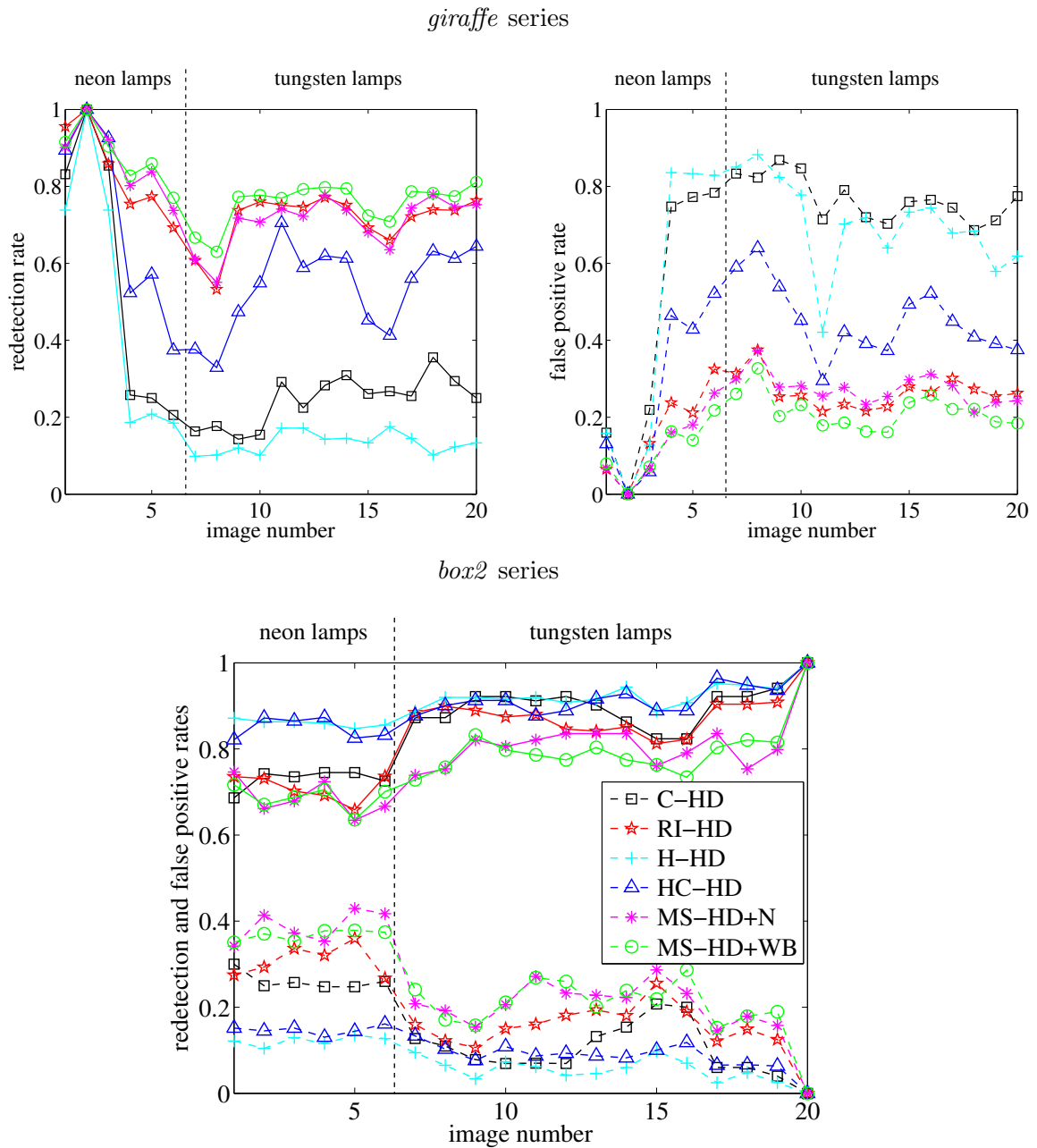


Figure 4.28: Evaluation results for the *giraffe* and *box2* series (fig. 4.24 shows images 2, 8, 16 for *giraffe* and images 20, 15, 6 for *box2*). The redetection and false positive rates are indicated by continuous and dotted curve (left and right diagrams for *giraffe*). The illumination type is indicated at the top of the diagram. The same legend as in fig. 4.25 is used.

RI-HD perform as well as MS-HD+N because the image does not contain very dark areas in which they are noise sensitive. On the other hand, in the *box2* series, intensity edges are more prominent than colour edges: most object boundaries in the scene are enhanced by strong intensity edges, like in a drawing. As a result, detectors based only on chrominance (MS-HD+N, MS-HD+WB and RI-HD) are less stable than detectors using intensity (H-HD, C-HD and HC-HD). This is emphasised by the simple 3D scene geometry: shadow and shading effects have only a small influence on the images. As a result, H-HD and HC-HD achieve both the highest stability, while MS-HD+N, MS-HD+WB and RI-HD have the lowest stability. To summarise, chrominance based detectors (MS-HD+N, MS-HD+WB and RI-HD) require good colour edges in the reflectance for detection stability, while intensity based detectors (H-HD) require good intensity edges in the reflectance.

Fig. 4.29 presents detection stability for textured scenes. The *rabbit* series (top of fig. 4.29) shows an object with 3D geometry of middle complexity and coloured textured reflectance. RI-HD and MS-HD+WB achieve the best stability. This has two reasons. First, the images do not contain any dark areas in which RI-HD and MS-HD+WB are noise sensitive. Second, the Nagao preprocessing distorts the fine texture on the object. As a result, MS-HD+N achieves lower stability than RI-HD and MS-HD+WB. For such a scene with fine texture, another preprocessing which better preserves texture would be preferable. The stability of H-HD and HC-HD is similar and only slightly lower than the stability of RI-HD and MS-HD+WB. For this scene, detectors based solely on chrominance are only slightly more stable, because of the object reflectance: the object contains many colour interest points with similar cornerness values, which results in lower detection stability for all detectors. The *snoopy* series shows another textured object. Main parts of the object do not contain any colour edges and the intensity information is more textured than the colour information. The detectors based only on chrominance (MS-HD+N, MS-HD+WB and RI-HD) are in most images more stable than the detectors using intensity (H-HD and HC-HD). This shows that many colour edges are not necessary for good stability of chrominance based detectors. Few stable interest points can be more useful than many unstable interest points. Due to the presence of dark areas, RI-HD achieves clearly lower stability than MS-HD+N and MS-HD+WB. Under tungsten lamps, MS-HD+N is more stable than MS-HD+WB because the dark areas lead to high false positive rates for MS-HD+WB. On the other hand, under the more uniform lighting of neon lamps or sunlight, MS-HD+WB performs better than MS-HD+N because of the texture distortion by the Nagao preprocessing. H-HD and HC-HD achieve lower stability because intensity is more textured than colour information in this scene (more similar interest points).

4.8.3 Conclusion

All detectors have similar and best stability for structured scenes with simple 3D geometry or for simple illumination changes. C-HD is however not stable enough under complex lighting changes. Adapting detection to the local lighting conditions with H-HD or HC-HD increases stability for complex lighting changes. When the scene reflectance contains colour edges with low intensity contrast (here in the *giraffe* series), HC-HD performs

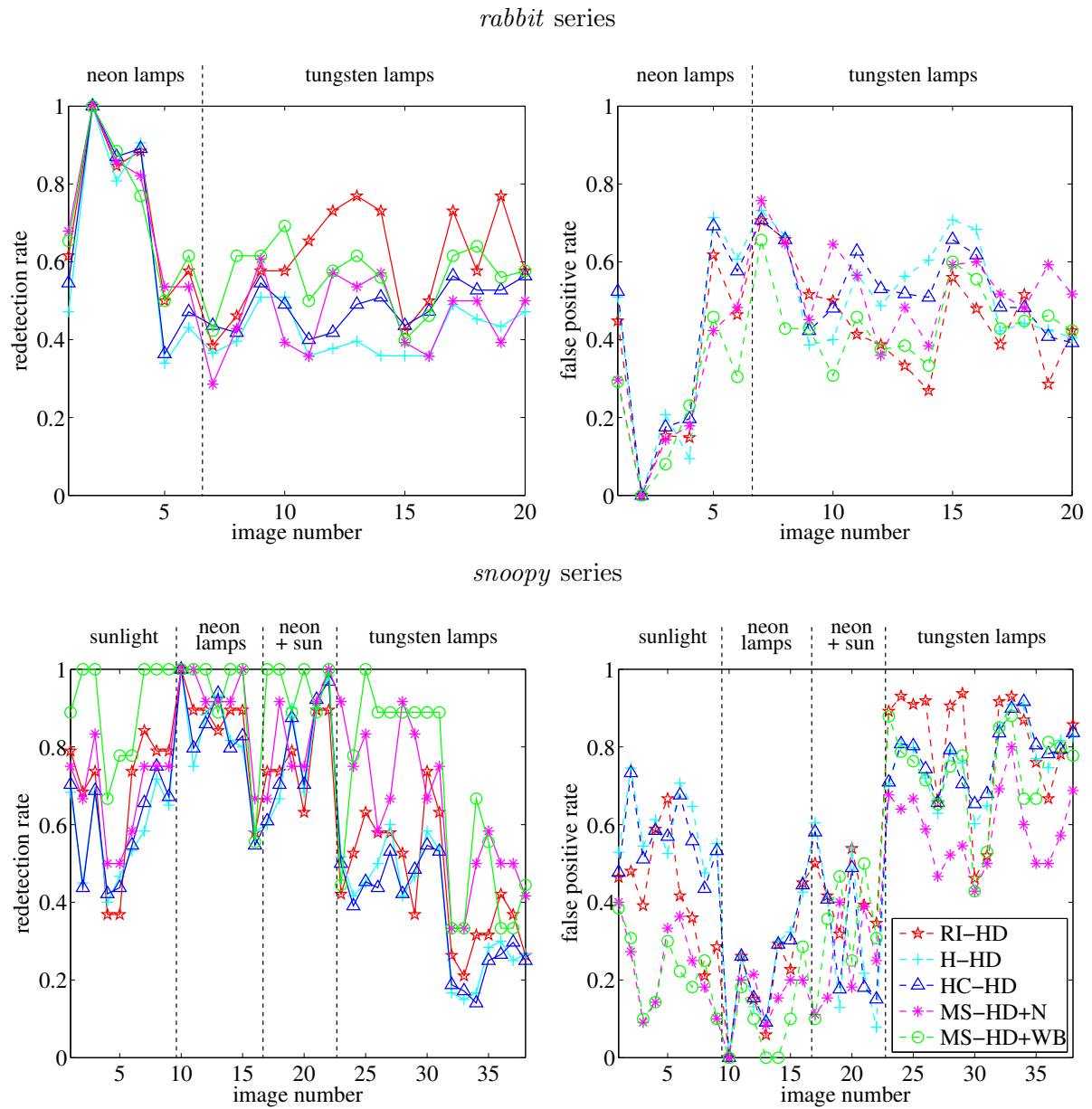


Figure 4.29: Evaluation results for the *rabbit* and *snoopy* series (fig. 4.24 shows images 2, 7, 16 for *rabbit* and images 10, 6, 28 for *snoopy*). The redetection and false positive rates are given in the left and right diagrams. The illumination type is indicated at the top of the diagram. For better legibility, the results of C-HD are not given (it achieves the worst stability).

better than H-HD. In general however, H-HD and HC-HD achieve the same stability, so H-HD should be preferred because it is faster. The other detectors (RI-HD, MS-HD+WB and MS-HD+N) are based only on chrominance and are invariant to all shadow and shading effects. As a result, they are in general more stable than C-HD, H-HD and HC-HD under complex illumination changes, especially for complex 3D scene geometry and when the illuminant type varies. Chrominance is less robust to noise than intensity. RI-HD, MS-HD+WB and MS-HD+N are therefore more noise sensitive than the other detectors, particularly in dark image areas. RI-HD is the most noise sensitive: it detects many false positives in dark areas. Thanks to its robust implementation, MS-HD+WB is less noise sensitive. MS-HD+N is the least noise sensitive of all three chrominance based detectors because the Nagao preprocessing attenuates noise and colour artifacts. The Nagao preprocessing distorts however textures, so that MS-HD+WB is more stable than MS-HD+N for scenes with fine texture. RI-HD, MS-HD+WB and MS-HD+N require stable colour edges in the scene for stable detection. Scene content should therefore influence the choice of the interest point detector. A grey value detector, like H-HD, is the best choice when intensity is the most prominent image information and when scene geometry and changes are of medium complexity. MS-HD+N is most appropriate for structured scenes with stable colour edges, complex geometry and for complex lighting changes. MS-HD+WB is the best choice for textured scenes with stable colour edges, complex geometry and complex lighting changes.

All detectors, especially the chrominance based detectors, are sensitive to colour artifacts. This could be enhanced by better demosaicing methods. Alternatively, a more powerful preprocessing method than the Nagao filter could be applied to achieve better texture preservation while still attenuating colour artifacts and noise. Finally, better camera hardware could increase chrominance quality. The X3 technology by Foveon Inc. (see [Fov06]) is particularly promising for the reduction of colour artifacts because it samples all three colour channels for every pixel with a single CMOS sensor. Chrominance is inherently noise sensitive in dark areas. More elaborate preprocessing of dark areas could enhance detection stability of the m space Harris detector. Alternatively, dark areas could for example be postprocessed like the saturated areas. Like for the grey value detectors, the user-defined detection threshold influences detection stability, especially for textured scenes (scenes with many similar interest points). Automatic threshold adaptation to scene content would therefore enhance stability. Finally, all developed detectors are sensitive to specular effects. The shadow-shading-specular robust invariant detector presented in [vdW05] compensates specularities in addition to shadows and shading. It is however like RI-HD sensitive to noise and to the accuracy of the automatic white balancing method. If necessary, specular effects could be detected or reduced with polarising filters or using several images as shown in [LLK⁺02]. Specular highlights can also be removed using image inpainting as in [BSCB00] or based on the dichromatic image formation model as in [TLQS03, TI03].

4.9 Summary

This chapter deals with illumination invariant interest point detection using colour images. First, several demosaicing methods are compared to find an algorithm that induces only few colour artifacts in the colour images: the algorithm in [LT03]. The image formation model and the Harris detector for colour images are then reviewed and the illumination influence on the detector is derived. Next, the only existing illumination invariant detector is described: the robust invariant Harris detector (RI-HD). RI-HD is invariant to shadow and shading and it requires white balancing. The two developed detectors are then presented. The first detector is the homomorphic colour Harris detector (HC-HD). It adapts detection to the local lighting conditions similarly to the homomorphic Harris detector for grey value images (H-HD). Slowly varying components of shadows, shading and of illuminant colour are compensated. The second detector is based on the m space, an invariant colour space which is based on chrominance and which is fully invariant to shadows and shading. Light colour is locally compensated so that no white balancing is necessary. The m space Harris detector (MS-HD) detects less interest points than detectors using intensity because it only responds to colour edges. Chrominance information is sensitive to noise and to colour artifacts induced by image acquisition. Therefore, a special preprocessing is introduced for MS-HD, which is based on the Nagao filter. It reduces well noise and artifacts. The resulting detector is named MS-HD+Nagao (MS-HD+N). The preprocessing distorts however texture. Therefore a second version of the MS-HD is evaluated, in which the formation of colour artifacts is reduced by applying automatic white balancing before demosaicing. This is the MS-HD+white balancing (MS-HD+WB).

All colour detectors and the best grey value detector (H-HD) are evaluated and compared on realistic image series showing a scene under different illuminations. The colour Harris detector (C-HD) is the least stable under complex illumination changes. Stability is increased with H-HD and HC-HD. They achieve similar stability on most scenes. H-HD is thus better because it is faster. HC-HD performs better when the scene reflectance has only colour edges with little intensity contrast. The detectors based on chrominance (RI-HD, MS-HD+WB and MS-HD+N) are more stable than the other detectors for scenes with complex 3D geometry and for complex lighting changes, especially when the illuminant type varies. They require stable colour edges for stable detection. As a result, the most appropriate detector depends on scene content. RI-HD, MS-HD+WB and MS-HD+N are sensitive to noise and colour artifacts in dark image areas. RI-HD is the most sensitive: it detects many false positives in dark areas. The best stability among RI-HD, MS-HD+WB and MS-HD+N is obtained by MS-HD+N for structured scenes and by MS-HD+WB for textured scenes. All tested detectors are sensitive to specularities. The stability of all colour detectors could be further increased through automatic thresholding, better demosaicing, better preprocessing or postprocessing or better camera hardware. This should be subject to further research.

5 Application to a recognition task

In this chapter, an object recognition and localisation system is developed. It is used to illustrate and evaluate the influence of interest point detectors in applications. Illumination invariance is of special interest for recognition tasks because learning and recognition phases take place at distant time instants: illumination changes are therefore very probable. After the system overview in section 5.1, each section presents one of the system blocks. In section 5.2, the descriptors for interest point characterisation are described. Section 5.3 presents how the interest point 3D positions are estimated with stereo vision. The matching strategy is described in section 5.4. Finally, the recognition and localisation algorithm is explained in section 5.5. The influence of the interest point detector on the recognition system is then evaluated in an application. The evaluation framework is described in section 5.6. Recognition results are presented in section 5.7. Section 5.8 summarises the chapter.

5.1 System overview

The goal of the system is to recognise learnt objects and to estimate the camera pose¹⁾ for the current image. The application is kept small because the aim is to evaluate the influence of interest point detection on a recognition system. The database is composed of 10 objects. The objects should be recognised independently of illumination conditions. Limited camera motion should also be handled: objects should be recognised from similar but distinct viewpoints. The system has three degrees of freedom: translation in a horizontal plane and rotation about the vertical axis. The image plane is approximately vertical. This is illustrated in fig. 5.1. The camera moves hence with a constant height from the floor and its optical axis is parallel to the floor. This is realised by fixing the camera on a tripod with a pan rotation unit. During the learning phase, the database is built with one image of each object. Many test images taken from different viewpoints and under different illuminations are then used to evaluate the influence of interest point detection on recognition. Image data are presented in more details in section 5.6. The 3D position of the interest points is estimated in order to simplify localisation. This is performed with stereo vision because this requires only cameras and is sufficiently accurate. Two cameras are hence mounted next to each other on the tripod.

An overview of the whole recognition system is given in fig. 5.2. One of the interest point detectors presented in chapters 3 and 4 is applied on the image of the left camera. The

¹⁾ Camera pose denotes camera position and orientation.

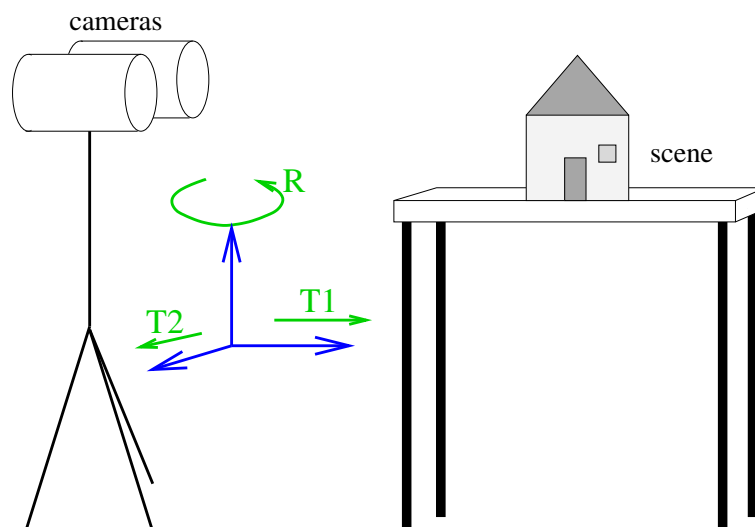


Figure 5.1: Degrees of freedom of the recognition system. The two translation directions and the rotation are indicated in green.

pixel values in the interest point neighbourhood are then used to compute a descriptor for each point. These descriptors characterise the interest points and are used for matching. They are described in details in section 5.2. Next, the 3D positions of the interest points are reconstructed with stereo vision: for each interest point, a corresponding point is found in the right image. The displacement between the two corresponding points is used to estimate the interest point 3D position. Stereo processing is explained in section 5.3. During the learning phase, the interest points are stored together with their descriptors and 3D positions in the database. This is performed for each object. The database is implemented straightforwardly because it is small. During the recognition phase, the interest points in the current image²⁾ are matched to the interest points stored in the database. For this, 3D positions and descriptors are used: two interest points are only matched if their descriptors are similar and if a displacement between their two positions is possible. The matching strategy is described in section 5.4. The obtained matches are used by the recognition and localisation algorithm, which merges the information of all matches to estimate the object viewed in the current image and the camera pose. Localisation is based on the 3D positions of the interest points and assumes rigid body motion. The uncertainty of the interest point positions is taken into account to enhance recognition and localisation results. The recognition and localisation algorithm is described in section 5.5. The different system blocks all apply state of the art algorithms, except the interest point detectors which are new (cf. chapters 3 and 4) because the topic of this thesis is illumination invariant interest point detection.

²⁾ In fact, an image pair is required because of stereo vision. The right image is only used for 3D reconstruction (see fig. 5.2). For simplicity, image denotes in this chapter the left image of the image pair together with the 3D positions of the interest points, except in the context of stereo vision.

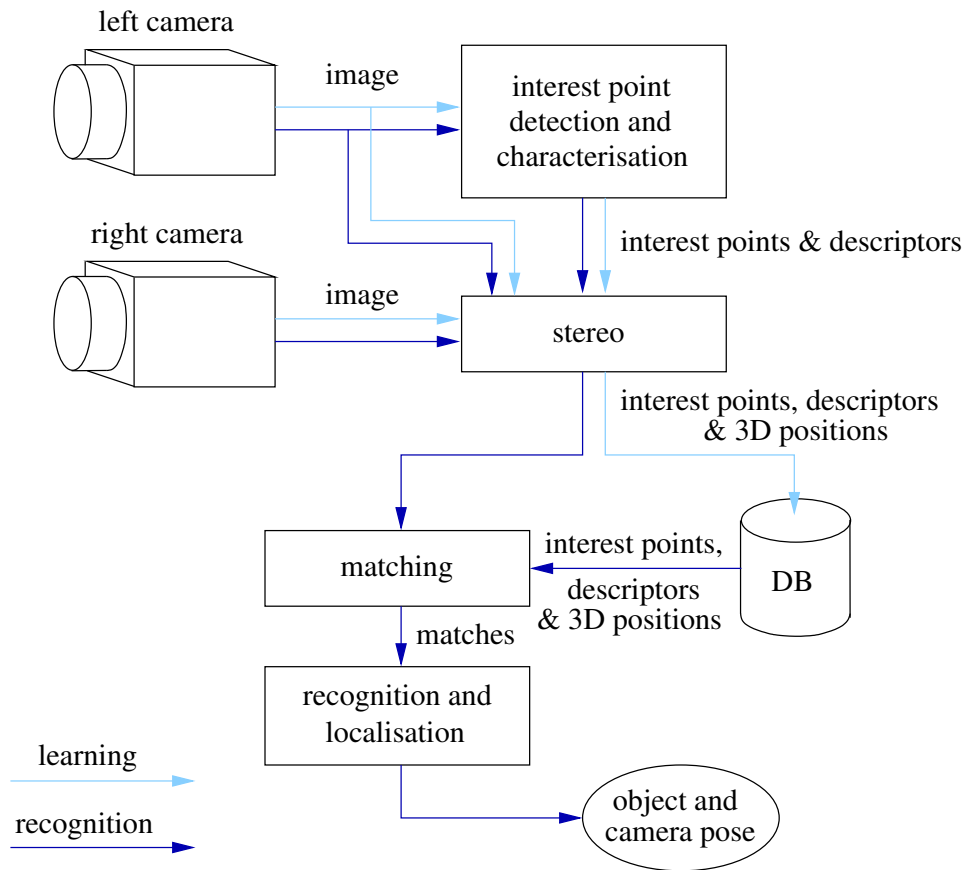


Figure 5.2: Overview of the recognition system

5.2 Interest point characterisation

In subsection 5.2.1, after a short overview of the existing descriptors, the most appropriate algorithm for the developed recognition system is chosen: the SIFT descriptor described in [Low04]. This algorithm and its implementation are described in more details in subsection 5.2.2.

5.2.1 Choice of the descriptor algorithm

After detection with one of the algorithms described in chapters 3 and 4, each interest point is characterised by a descriptor. This descriptor encodes texture information in the neighbourhood defined by the interest point. It is used to constrain matching: only similar interest points are matched. The similarity of two interest points is obtained by comparing their descriptors. As a consequence, descriptors should be discriminative, i.e. contain characteristic texture information, in order to reduce the number of false matches. As stated in section 5.1, the objects should be recognised under illumination changes and limited viewpoint changes. The influence of such changes on the descriptors should hence

be small. Illumination influence is described in chapter 2. Limited viewpoint changes can be approximated by local affine transformations in the interest point neighbourhoods, as explained in [Tuy00]. A last requirement for the descriptors is compactness. This is however not essential here because the application handles a small database.

The simplest method to build a descriptor is to store all pixel values in the interest point neighbourhood. In that case, interest points are compared with a correlation measure, such as normalised crosscorrelation, sum of squared differences or sum of absolute differences. Descriptor dimensionality can be reduced by applying principal component analysis on the descriptor set like in [SD99, LPF04]. Illumination influence can be suppressed by normalising the pixel values in the neighbourhood with one of the methods presented in subsection 2.3.1. Similarly, viewpoint changes can only be compensated if the local affine transformation is corrected beforehand. Such a viewpoint correction is described for example in [MS04, Tuy00, MCUP02]. Another possibility to handle viewpoint changes is to store several descriptors for each interest point, each corresponding to a different viewpoint, like in [LPF04]. This descriptor form is however most of the time used in applications where the viewpoint does not change much, for example stereo vision in [VL01] or tracking in [DM02, SD99].

Another popular descriptor form is based on moments. Moments encode the spatial distribution of the pixel values. They represent texture very compactly. For example, in [Tuy00], the whole neighbourhood is described by 18 or 9 moment invariants, depending on the chosen degree of invariance. Invariance to illumination changes is obtained for example by moment normalisation in [MMG99, Tuy00, MCUP02]. Invariance to rotation in the image plane can be achieved by selecting only some of the moments (cf. [Tuy00]). If invariance to full affine transformation is necessary, viewpoint changes must be compensated beforehand, for example like in [Tuy00, MCUP02].

Decomposition of the pixel values using several filters, for example derivative filters, is another popular method to compute descriptors. The filters can be applied to a single neighbourhood point as in [SM97, MS04] or to several neighbourhood points as in [CJ02]. Filter based descriptors are compact: descriptors in [MGDP00, SM97, MS04] are composed of 8 to 12 invariants. Like for moments, invariance to illumination changes can be achieved by normalising the filter responses as in [SM97]. Alternatively, the image neighbourhoods can be normalised before filtering as in [MGDP00]. Invariance to rotation in the image plane can be obtained by steering the filters in the direction of the highest gradient as in [FA91, CJ02, MS04]. Another solution consists in computing rotational invariants by combining the response to several filters like in [SM97, MGDP00]. For further invariance to viewpoint changes, multi-scale descriptors can be used like in [SM97] or viewpoint changes can be compensated before descriptor computation as in [MS04].

Last, texture can be represented with histograms. The most popular histogram based descriptor is the SIFT descriptor: a multi-dimensional histogram of gradient values (see [Low04]). The first two histogram dimensions are the image directions x and y . This encodes the spatial distribution of the texture. The third dimension is gradient direction. The compactness of this descriptor depends on the number of bins in each direction. In

5 Application to a recognition task

[Low04, MS05], $4 \times 4 \times 8$ descriptors are used. Illumination invariance is obtained by normalising the descriptors. Rotation in the image plane are compensated by steering the descriptor in the direction of the highest gradient. Scale invariance is achieved by adapting neighbourhood size to image content. Further mechanisms (weighting and interpolation) make the descriptors robust to the remaining geometric transformation. More elaborate versions of the SIFT descriptors are presented in [MS05].

In [MS05], existing descriptors are compared to each other. The descriptors are evaluated on several image series with viewpoint changes, simple illumination changes, image blur and JPEG compression artifacts. Comparison criterion is matching quality. The SIFT descriptor reaches the best performance. This proves the robustness of this detector to viewpoint, noise and simple illumination changes. Its higher dimensionality makes it also more discriminative than smaller descriptors based on moments or filters. Smaller SIFT descriptors ($2 \times 2 \times 8$) also achieve higher robustness and better matching quality than filter based descriptors, as shown in [Kra05]. Last, SIFT descriptors simplify matching because the similarity between interest points is simply the Euclidian distance between descriptors. Other descriptors such as the filter based descriptors requires weighted distances like the Mahalanobis distance (see [SM97]). Therefore SIFT descriptors are used in this work to characterise the detected interest points. They are computed from gradients. They can therefore re-use the invariant derivatives computed by some of the developed interest point detectors (N-HD, H-HD, RI-HD, HC-HD and MS-HD). The algorithm and its implementation are described in more detail in subsection 5.2.2.

5.2.2 SIFT descriptors

As explained in the overview of subsection 5.2.1, SIFT descriptors represent texture with a three dimensional gradient histogram. This process is illustrated in fig. 5.3. The original algorithm described in [Low04] is used here. The three histogram dimensions are image directions x and y , and gradient orientation θ . Gradient magnitude is accumulated in the histogram bins: each bin (x_i, y_j, θ_k) (represented by a red arrow in the right part of fig. 5.3) is the sum of the gradient magnitudes of all pixels contained in the image area represented by (x_i, y_j) with gradient orientation corresponding to θ_k . In fig. 5.3, the image area for each bin contains 4×4 pixels and is indicated with bold lines. There are 4 image areas, each with 4 different gradient directions (delimited with dashed lines in fig. 5.3), which corresponds to a descriptor with $2 \times 2 \times 4 = 16$ bins. The number of bins per direction should be chosen as a compromise between discriminative power and compactness. The database in this application is small, therefore small descriptors are discriminative enough: descriptors of size $2 \times 2 \times 4$ are used. This discretisation is shown in fig. 5.3: the spatial bins represent the upper left, upper right, lower left and lower right neighbourhood parts and the gradient orientations are left, right, up and down.

To achieve invariance to scale changes and rotation in the image plan, neighbourhood size and orientation are adapted to image content in [Low04]. In this application, only limited viewpoint changes occur and there is no rotation in the image plane. Therefore, the

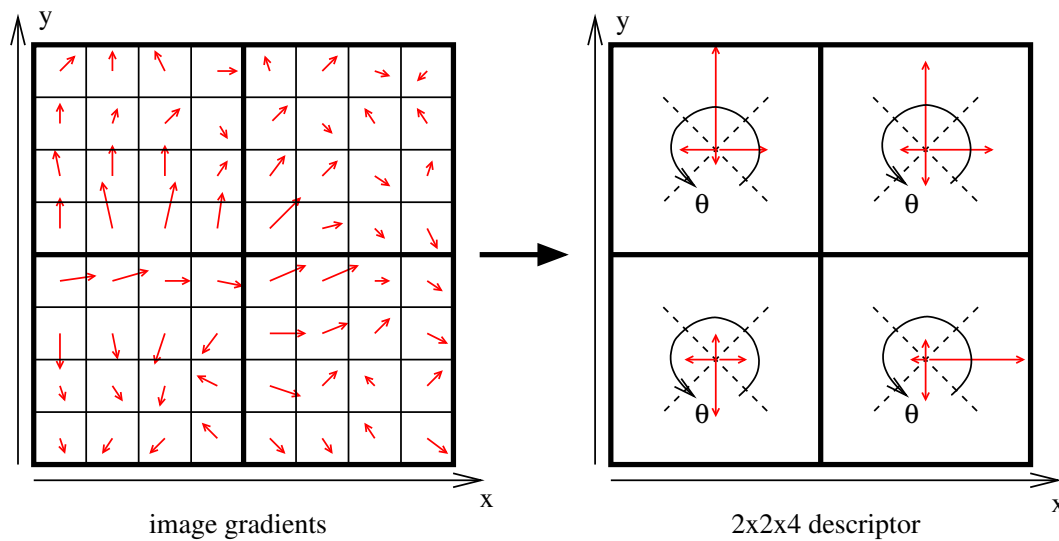


Figure 5.3: SIFT descriptor overview. Right: the image gradient for each pixel in the considered neighbourhood is represented by a red arrow. The arrow indicates gradient magnitude and direction. Left: each red arrow visualises one descriptor bin. In this descriptor, each image direction (x and y) is represented with two intervals and gradient orientation θ is represented by four intervals, leading to 16 bins. Each bin contains the sum of the gradient magnitudes of all pixels in the corresponding image area (delimited with bold lines) with corresponding gradient orientation (delimited with dashed lines).

neighbourhood is not adapted to image content here: the inherent robustness of the SIFT descriptors to viewpoint changes is sufficient. The neighbourhood size for characterisation should be related to the neighbourhood size for interest point detection. A Gaussian filtering with standard deviation $\sigma_M = 3$ is used for interest point detection. Hence, a 15×15 square neighbourhood is used for descriptor computation: it contains most pixels that were taken into account for interest point detection. The SIFT descriptor is designed to be robust to limited viewpoint changes. This principle is inspired from neural mechanisms in the human brain (see [Low04]). To achieve robustness, histogram bins should be chosen big enough. In addition, the histogram is filled using trilinear interpolation: each pixel influences 2 bins per direction. This reduces viewpoint influence because viewpoint changes smoothly modify the descriptors. Finally, gradient magnitudes are weighted with a Gaussian to decrease the influence of pixels at the neighbourhood border and to reduce the influence of inaccurate interest point detection. Like in [Low04], the standard deviation of the weighting Gaussian σ_w is half the width of the considered neighbourhood. σ_w is therefore set to 7.5.

Robustness to illumination changes is achieved by normalising the vectors representing the descriptors to unit length. Bins are proportional to sums of gradients. If illumination influence is assumed constant in the considered neighbourhoods as in eq. (2.6), this normalisation completely compensates illumination changes. Some of the developed interest point detectors are based on invariant gradients (N-HD, H-HD, RI-HD, HC-HD and

5 Application to a recognition task

MS-HD). When those detectors are applied, the derivatives are re-used to compute the descriptors. In that case, no normalisation is performed. To achieve further robustness to specularities and saturation, a second step is applied in [Low04]: the normalised descriptors are thresholded so that no bin value is bigger than 0.2, and they are subsequently renormalised. This reduces the influence of strong gradients which are most of the time caused by specularities or saturation. This second step is not performed here: the handling of saturated areas during interest point detection already prevents the detection of interest points near saturated areas and near most specular highlights. The descriptors described in [Low04] use grey value images. For colour images, one SIFT descriptor is computed for each channel using the channel gradients. In that case, the normalisation compensates not only the influence of slowly varying shadows and shading but also the influence of the illuminant colour.

SIFT descriptors are computed for each interest point according to the following:

1. For each pixel (x, y) in the 15×15 neighbourhood around interest point (x_{IP}, y_{IP}) and for each channel:
 - a) Compute gradient magnitude g and gradient orientation θ from the image derivatives I_x and I_y .
 - b) Multiply gradient magnitude with the Gaussian: $g' = g e^{-\frac{(x-x_{IP})^2+(y-y_{IP})^2}{2\sigma_w^2}}$ where $\sigma_w = 7.5$.
 - c) Compute the two gradient orientation bins to increment θ_1 and θ_2 and the interpolation weights $w(\theta_1)$ and $w(\theta_2)$: $w(\theta_i) = 1 - \frac{|\theta - \theta_i|}{|\theta_2 - \theta_1|}$
 - d) Compute the two horizontal direction bins to increment x_1 and x_2 and the interpolation weights $w(x_1)$ and $w(x_2)$: $w(x_i) = 1 - \frac{|x - x_i|}{|x_2 - x_1|}$. If (x, y) is in the leftmost or rightmost border, only one bin x_1 and its weight $w(x_1)$ are computed.
 - e) Compute the two vertical direction bins to increment y_1 and y_2 and the interpolation weights $w(y_1)$ and $w(y_2)$: $w(y_i) = 1 - \frac{|y - y_i|}{|y_2 - y_1|}$. If (x, y) is on the upper or lower border, only one bin y_1 and its weight $w(y_1)$ are computed.
 - f) Increment all selected bins with: $g' w(\theta_i) w(x_j) w(y_k)$ where $i, j, k = 1$ or 2 .
2. If required, normalise the descriptor channelwise to unit length.

Here, the descriptor has only 4 possible gradient orientations: θ_i is either 0° , 90° , 180° or 270° . Similarly, x_j and y_k are either -3.75 or $+3.75$ because there are 2 cells per image direction and the neighbourhood is 15 pixels wide. The Gaussian weighting and the interpolation weighting for x and y directions are implemented as look-up tables because image coordinates x and y are discrete. Gradient orientation θ can take any value. The weights $w(\theta_i)$ are therefore computed online. The descriptor computation requires approximately 0.06ms per interest point for grey value systems and 0.18ms per interest point for colour systems on the computer described in subsection 2.2.2.

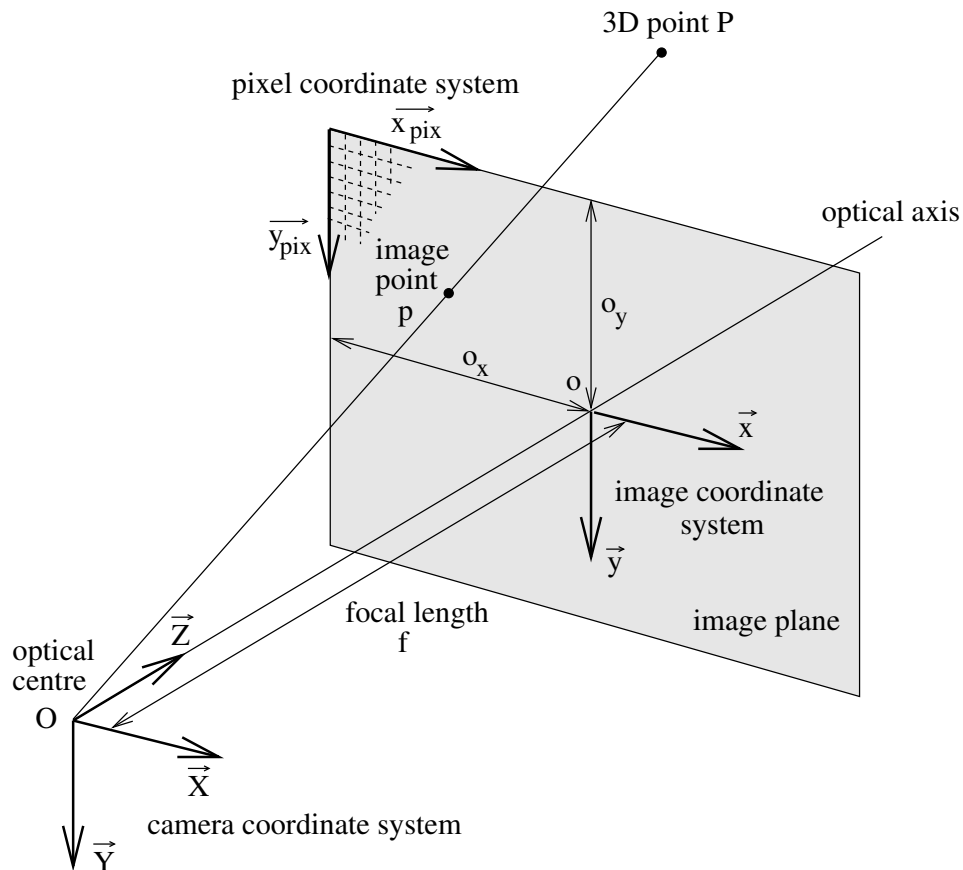


Figure 5.4: Perspective camera model.

5.3 Stereo reconstruction

The interest point 3D positions are reconstructed using stereo vision. This 3D information is used for matching and for recognition and localisation. Stereo vision is chosen because it is a reliable and cheap method to reconstruct 3D information for image points. Only two cameras are needed: no additional sensor such as a laser range finder and no additional light source (for 3D reconstruction using structured light) is necessary. The principles of stereo vision are explained in subsection 5.3.1. The two steps of stereo reconstruction are then described in more detail: subsection 5.3.2 explains how reliable correspondences between the two images are found and subsection 5.3.3 describes how this is used to estimate the 3D position of the interest points.

5.3.1 Principles of stereo vision

To reconstruct 3D scene geometry, a mathematical model describing the geometry of image formation is necessary. Here, the perspective camera model³⁾ is used (see for

³⁾ This model is also named pinhole camera model.

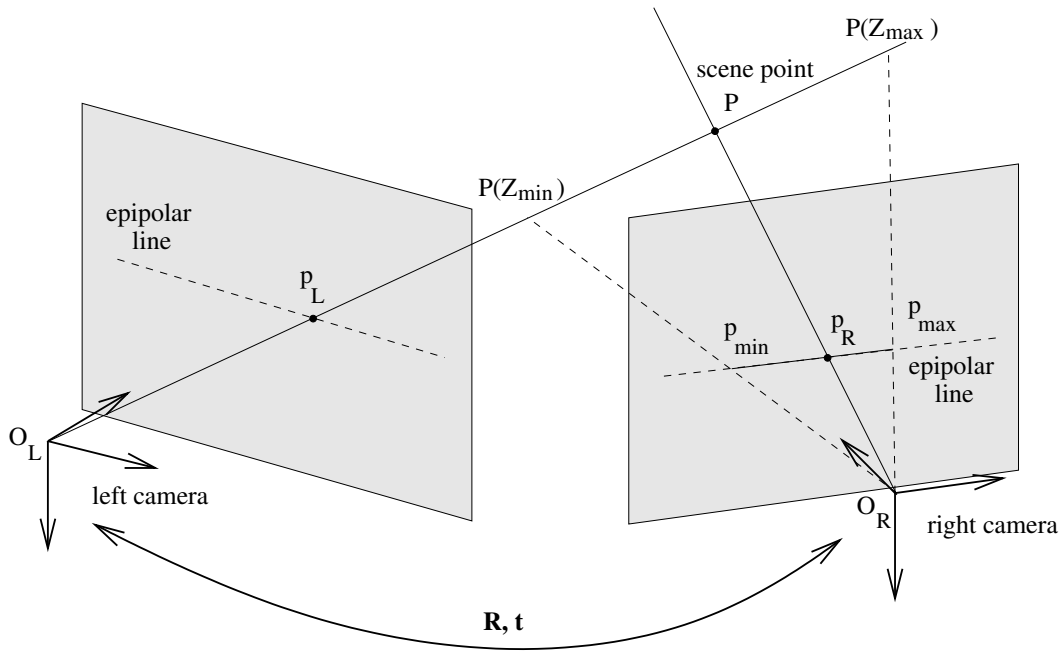


Figure 5.5: Stereo vision and epipolar geometry.

example [TV98]). This model is shown in fig. 5.4. The perspective camera model describes mathematically how a 3D point P is projected into a corresponding image point p . From a geometric point of view, the image point p is the intersection of line (OP) with the image plane. This can be described with linear algebra. If the coordinates of point P are $(X, Y, Z)^T$ in the camera coordinate system $(O \vec{X} \vec{Y} \vec{Z})$, the coordinates of image point p in the image coordinate system $(o \vec{x} \vec{y})$ are given by:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X \\ Y \end{pmatrix}. \quad (5.1)$$

This projection is followed by a discretisation step, in which the continuous image coordinates $(x, y)^T$ are transformed into discrete pixel coordinates $(x_{pix}, y_{pix})^T$ with:

$$x_{pix} = \frac{x}{s_x} + o_x \quad \text{and} \quad y_{pix} = \frac{y}{s_y} + o_y. \quad (5.2)$$

s_x and s_y are the pixel sizes in horizontal and vertical directions. o_x and o_y are the distances between image borders and the projection o of the optical centre in the image plane. The lens might cause radial distortion which can be corrected beforehand. This is handled in more detail in subsection 5.3.3.

For stereo vision, two cameras are used. 3D geometry can be reconstructed for scene points visible by both cameras. Therefore both cameras are directed in similar directions. The geometry of stereo vision is named epipolar geometry. It is shown in fig. 5.5. One coordinate system is associated to each camera: here, $(O_L \vec{X}_L \vec{Y}_L \vec{Z}_L)$ for the left camera and $(O_R \vec{X}_R \vec{Y}_R \vec{Z}_R)$ for the right camera. The two coordinate systems are related by

a rotation and a translation. One camera is chosen as reference and the 3D positions are estimated in its coordinate system. Here, the left camera is the reference. The transformation between both coordinate systems is described by:

$$\begin{pmatrix} X_R \\ Y_R \\ Z_R \end{pmatrix} = \mathbf{R} \left(\begin{pmatrix} X_L \\ Y_L \\ Z_L \end{pmatrix} - \mathbf{t} \right). \quad (5.3)$$

$(X_L, Y_L, Z_L)^T$ and $(X_R, Y_R, Z_R)^T$ are the coordinates of scene point P in the left and the right camera coordinate systems. Rotation matrix \mathbf{R} and translation vector \mathbf{t} represent the transformation between both coordinate systems. Scene point P is projected in both image planes according to the perspective camera model. The intersection of line $(O_L P)$ with the left camera image plane defines p_L . The intersection of line $(O_R P)$ with the right camera image plane defines p_R . The plane $(O_L O_R P)$ is named epipolar plane. It intersects both image planes in two lines named epipolar lines. As shown in fig. 5.5, the projection of line $(O_L P)$ in the right image is the epipolar line. The position of p_R on the epipolar line depends on the distance between O_L and P . Stereo vision uses this property to estimate 3D information. Here, the stereo system is calibrated with a commercial software: the small vision system (see [SRI02, SRI03]). All system parameters are obtained: the intrinsic parameters f, s_x, s_y, o_x, o_y for both cameras and the extrinsic parameters \mathbf{R}, \mathbf{t} . A simple geometric stereo vision algorithm can therefore be applied, which is based only on eqs. (5.1), (5.2) and (5.3).

To reconstruct the 3D position of image point p_L in the reference image (here the left image), line $(O_L p_L)$ is projected in the right image to obtain the epipolar line. A corresponding point p_R is searched in the right image along this epipolar line. The intersection of lines $(O_L p_L)$ and $(O_R p_R)$ gives scene point P , i.e. the 3D position of image point p_L . Stereo vision can therefore be divided into two steps. First, corresponding points must be found in the images. The search is constrained using epipolar geometry. Minimum and maximum Z_L coordinates are used to further reduce the search area in the right image, as shown in fig. 5.5. This is explained in subsection 5.3.2. Once the two corresponding points are obtained, the intersection of lines $(O_L p_L)$ and $(O_R p_R)$ is estimated. A simple geometric method is used. It is described in subsection 5.3.3.

5.3.2 Finding correspondences

To reconstruct interest point positions, correspondences must be found in the right image for all interest points detected in the left image. The same algorithm is performed for each interest point independently. In a first step, a search area is determined in the right image based on epipolar geometry. Next, the most similar point in this search area is determined using the method presented in [Hub04]. Sum of Absolute Differences (SAD) in a 16×16 neighbourhood centred on the interest point is applied to measure similarity. Finally, correspondence uniqueness is tested with a criterion developed in [Stö01]. 3D geometry is only reconstructed for unique correspondences to obtain only reliable 3D information.

5 Application to a recognition task

The first step determines the search area. As shown in [Hub04], constraining the search area with epipolar geometry increases correspondence reliability. The method described in subsection 5.3.1 and illustrated in fig. 5.5 is applied. The search area is defined by the epipolar line and by a valid range $[Z_{min}, Z_{max}]$ for the Z_L coordinate. This defines points p_{min} and p_{max} in fig. 5.5. The valid range should be adapted to the application. In the object recognition system, the objects are placed near the camera. The valid range is set to [30cm, 2.5m] here. The following procedure is used to compute p_{min} and p_{max} :

1. Compute the coordinates of the interest point p_L in the image coordinate system of the left camera using eq. (5.2).
2. Compute the 3D positions of points $P(Z_{min})$ and $P(Z_{max})$ in the left camera coordinate system using eq. (5.1) with $Z = Z_{min}$ and Z_{max} .
3. Transform the coordinates in the right camera coordinate system using eq. (5.3).
4. Project the scene points $P(Z_{min})$ and $P(Z_{max})$ in the image plane of the right camera using eq. (5.1). This gives image points p_{min} and p_{max} .
5. Compute the coordinates of p_{min} and p_{max} in the pixel coordinate system of the right camera using eq. (5.2).

This is illustrated in fig. 5.6. The epipolar lines are almost horizontal because the used cameras are approximately in a parallel configuration: both image planes are approximately coplanar (see [TV98]). To account for calibration errors and for radial distortion, the line segment $[p_{min} p_{max}]$ is extended vertically to define the search area. The search area is defined as all pixels (x, y) such that $|y - y_{epi}| \leq d$ where (x, y_{epi}) is a point of the epipolar segment $[p_{min} p_{max}]$. d is a distance threshold (in pixels). It should be adapted to the calibration accuracy and to the lens radial distortion. Here, d is set to 5 pixels.

Once the search area is defined, the similarity between the interest point and all points in the search area is computed. The point with the highest similarity is the corresponding point. The two cameras used for stereo are close to each other and the images are taken at the same time instant. Therefore, viewpoint and illumination do not change much between left and right images. As a result, a simple correlation measure is sufficient to estimate the similarity of neighbourhoods. Similarity must be computed efficiently because it is performed for all interest points and all points in the corresponding search areas. Here, the method presented in [Hub04] is used: it implements in software the real-time correspondence framework developed in [Stö01] and it extends the framework for stereo vision. Similarity is computed with the Sum of Absolute Differences (SAD) between the 16×16 grey value neighbourhoods around two points. SAD is chosen because of its speed. It is implemented efficiently using MMX assembler commands. The correspondences found between the images of fig. 5.6 are shown in fig. 5.7. On this small example, all correspondences are correct.

The search area may contain several points similar to the interest point when the interest point texture is not specific enough. In that case, a wrong correspondence may be chosen because of noise or viewpoint and illumination influence. To prevent such

Left image showing the detected interest points



Right image showing the corresponding epipolar lines



Figure 5.6: Search areas for stereo vision. The epipolar segments $[p_{min} p_{max}]$ are indicated in the right image for all interest points of the left image.

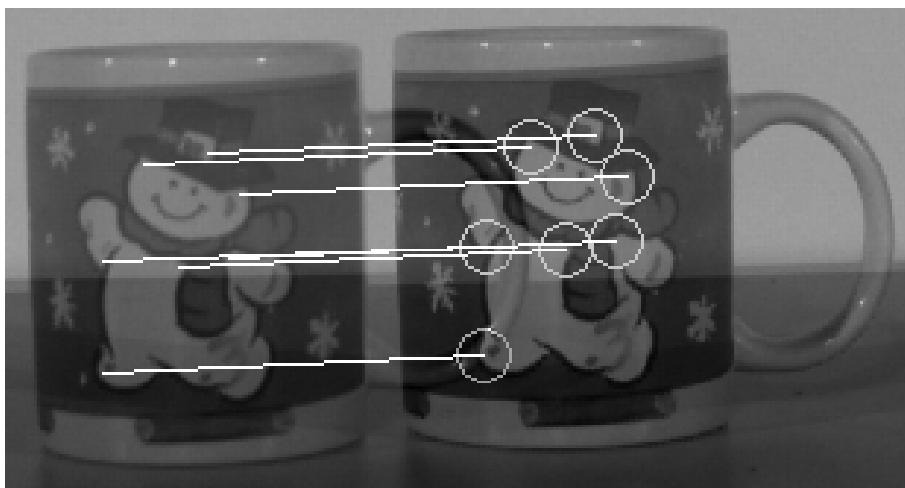


Figure 5.7: Correspondences between two stereo images. The correspondences are indicated by line segments between corresponding points. Both images are superposed for visualisation. Only the relevant image part is shown.

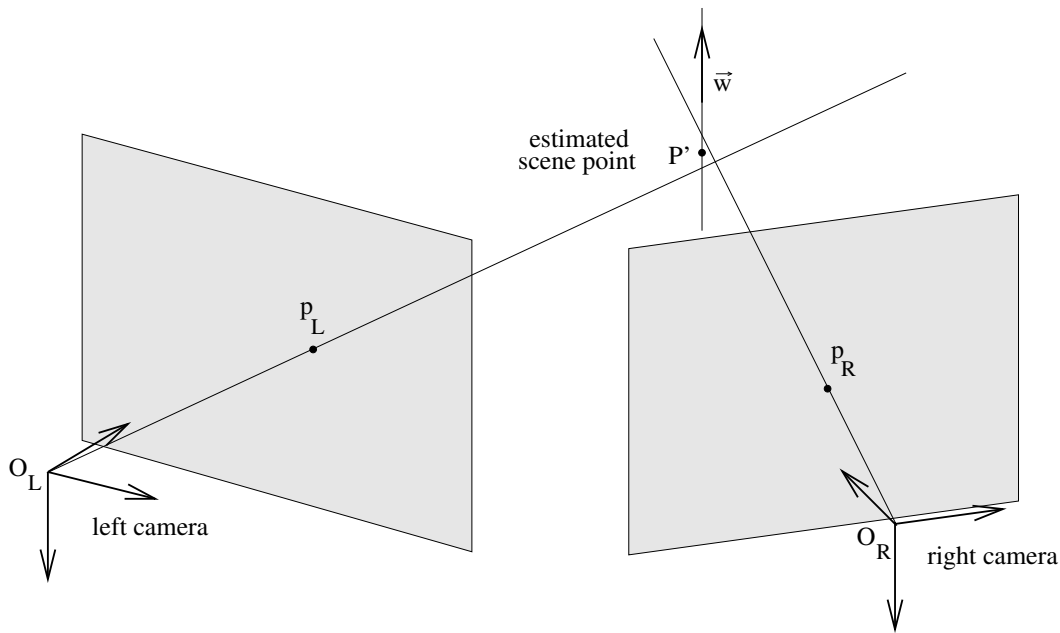


Figure 5.8: 3D Reconstruction of a scene point.

errors, correspondence uniqueness is tested. 3D information is only reconstructed for unique correspondences. The interest points failing the uniqueness test are discarded. For this, the SAD values of the maximum and of the second maximum in the search area are compared to each other: if the difference between both SAD values is high enough ($\text{SAD}(\text{max}) - \text{SAD}(\text{second max}) \geq t$, t is a user-defined threshold), the correspondence is considered unique. This simple uniqueness measure is compared to other uniqueness measures in [Stö01]. It is shown to be effective and more robust to the choice of the user-defined threshold t than the other measures. t is set to 20 here. All reliable correspondences are given to the next module: the 3D reconstruction.

5.3.3 3D reconstruction

As explained in subsection 5.3.1, the 3D position of scene point P is reconstructed with a simple geometric method, given its projections in both images p_L and p_R . The scene point is the intersection of the two lines $(O_L p_L)$ and $(O_R p_R)$ (see fig. 5.5). In practise, these two lines do not intersect because the estimated positions of p_L and p_R and the calibration are not accurate enough. Therefore the estimated 3D point P' is the point with minimum distance to these two lines. For this, the vector \vec{w} perpendicular to both lines is used. The line segment parallel to \vec{w} with one endpoint on each line is determined. The scene point P' is the midpoint of this segment (see [TV98]). This is illustrated in fig. 5.8.

Computations are performed in the left camera coordinate system. The equation of line $(O_L p_L)$ is $\mathbf{x} = a \mathbf{p}_L$ where $\mathbf{x} = (x, y, z)^T$ represents a scene point in the left camera

coordinate system. $\mathbf{p}_L = (x_L, y_L, f_L)^T$ is the image point p_L in the left camera coordinate system. x_L and y_L are obtained from pixel coordinates using eq. (5.2). Similarly, the equation of line $(O_R p_R)$ in the left camera coordinate system is $\mathbf{x} = b \mathbf{R}^T \mathbf{p}_R + \mathbf{t}$. Rotation matrix \mathbf{R} and translation vector \mathbf{t} represent the transformation between both camera coordinate systems (see eq. (5.3)). $\mathbf{p}_R = (x_R, y_R, f_R)^T$ represents image point p_R in the right camera coordinate system. Vector \vec{w} is perpendicular to both lines. Its coordinates in the left camera coordinate system are given by:

$$\mathbf{w} = \mathbf{p}_L \times \mathbf{R}^T \mathbf{p}_R, \quad (5.4)$$

where \times represents vector product. The line segment parallel to \vec{w} and with endpoints on lines $(O_L p_L)$ and $(O_R p_R)$ is therefore obtained by solving following equation:

$$a \mathbf{p}_L + c \mathbf{w} = b \mathbf{R}^T \mathbf{p}_R + \mathbf{t}. \quad (5.5)$$

Once a, b and c are obtained, the coordinates of scene point P' are given by:

$$\mathbf{P}' = a \mathbf{p}_L + \frac{c}{2} \mathbf{w}. \quad (5.6)$$

\mathbf{P}' contains the 3D coordinates of the current interest point in the left camera coordinate system. More details can be found in [TV98]. As explained before, the two lines $(O_L p_L)$ and $(O_R p_R)$ do not intersect because the coordinates of p_L and p_R and the system parameters are only known inaccurately. The 3D reconstruction precision can therefore be increased with better estimates of p_L and p_R or better calibration. Here, two problems are handled. First, the 2D position of p_R is improved using subpixel correspondence. Second, the radial distortion caused by the lens is corrected to model the cameras more accurately. These two steps are described in the following paragraphs.

As explained in subsection 5.3.2, the pixel with the maximum similarity to the interest point p_L is the corresponding point p_R used for reconstruction. Pixel positions are discrete, hence inaccurate. This can be improved with subpixel correspondence. The simplest method for that is to use the similarity values of the pixels around the determined point p_R . The similarity measure reaches a maximum for p_R . If the similarity measure is well behaved, the similarity measure can be modelled as a paraboloid in the neighbourhood of p_R . In practise, it works for most pixels. A paraboloid is fitted to the SAD values of p_R and of its eight neighbours using least square estimation. The peak position of the estimated paraboloid is obtained by setting the paraboloid derivatives in x and y directions to 0. The whole process is described in details in [Gon03]. It yields coordinates for p_R with subpixel accuracy. On some neighbourhoods, paraboloid estimation fails. In such a case, subpixel coordinates are computed for x and y directions separately. For this, a 1D parabola is fitted to the SAD values of p_R and of its two neighbours in the considered direction x or y . The subpixel coordinates are computed like before by setting the derivative of the obtained parabola to 0. This 1D method is less accurate than the 2D method and is therefore only used in the case the 2D method fails. The 1D method always converges. As a result, subpixel correspondences can always be obtained.

5 Application to a recognition task

Camera lenses may induce geometric distortions. These can be modelled as radial distortions with following equations:

$$x = x_d(1 + \kappa_1 r^2) \quad \text{and} \quad y = y_d(1 + \kappa_1 r^2). \quad (5.7)$$

x and y are the undistorted coordinates in the image coordinate system. x_d and y_d are the distorted coordinates in the image coordinate system and $r^2 = x_d^2 + y_d^2$. κ_1 is the distortion parameter. Radial distortions are caused by the lens. They occur hence between the projection in the image plane described by eq. (5.1) and the image discretisation in eq. (5.2). As a result, the undistorted coordinates (x, y) verify eq. (5.1). They are distorted by the lens according to eq. (5.7). Subsequently the distorted coordinates (x_d, y_d) are discretised according to eq. (5.2). Further details can be found in [TV98]. The used commercial calibration software estimates the κ_1 value (see [SRI02, SRI03]). The influence of κ_1 is hence corrected before 3D reconstruction. This proved experimentally to increase reconstruction accuracy.

The 3D reconstruction is summarised in the following:

1. Compute subpixel correspondences for the correspondences obtained with the method of subsection 5.3.2.
2. Transform the pixel coordinates of p_L and p_R in image coordinates with eq. (5.2).
3. Correct radial distortion with eq. (5.7).
4. Compute \mathbf{w} with eq. (5.4) and solve the system of eq. (5.5).
5. Compute the 3D position of the interest point with eq. (5.6).

The whole process (correspondence finding and 3D reconstruction) provides reliable estimates of the interest point 3D positions. It requires approximately 0.5ms per interest point on the computer described in subsection 2.2.2. The accuracy decreases however when interest points are further away from the cameras, like for all stereo reconstruction algorithms. The uncertainty of the 3D information is therefore modelled for matching and for localisation, as explained in sections 5.4 and 5.5.

5.4 Matching

Once the interest points are characterised with descriptors and 3D positions, the whole information extraction process is finished. The interest points are stored in the database with descriptors and 3D positions during the learning phase. During the recognition phase, the current interest points are matched to the database interest points. The resulting list of matches between current image points and database points is then used for recognition and localisation. As the database is small here, matching is implemented straightforwardly. The database is a list of interest points. Each current interest point is compared to all database interest points to verify if a match should be created. Such an exhaustive search is impossible for bigger databases. In that case, the number of

comparisons must be reduced, for example by indexing. Such techniques can be found for example in [Low04, Neu01]. To increase the efficiency and accuracy of recognition and localisation, the resulting match list should contain as many true matches as possible and as few false matches as possible. This is shown for the RANSAC algorithm⁴⁾ in [VL01]. Several matching strategies are presented in [VL01] to eliminate false matches while keeping true matches in the case of uncalibrated stereo. Three of these strategies are applied in this work. First, image information is used: only interest points with similar descriptors are matched, as explained in subsection 5.4.1. Second, geometric information is considered. The recognition system has only three degrees of freedom (cf. section 5.1). This constraint is used to verify if the displacement between the two interest point 3D positions is possible. This is described in subsection 5.4.2. Finally, subsection 5.4.3 presents match list constraints: only a given number of matches is allowed per interest point and matching must be symmetric (i.e. the same matches should be obtained if the roles of current image and database image are interchanged).

5.4.1 Descriptor similarity constraint

The first constraint regards image information: current interest point and database interest point in a match must have similar neighbourhoods. As explained in section 5.2, the similarity between two interest points can be estimated by the euclidian distance between their two descriptors. Therefore, if the current interest point p and the database interest point p_{DB} are characterised by descriptors \mathbf{D} and \mathbf{D}_{DB} , a match between p and p_{DB} is valid if: $\|\mathbf{D} - \mathbf{D}_{DB}\| \leq D_{lim}$. $\|\mathbf{D} - \mathbf{D}_{DB}\|$ is the euclidian distance between vectors \mathbf{D} and \mathbf{D}_{DB} . D_{lim} is the similarity threshold. Thresholding the similarity is a simple solution to verify match validity. It is sufficient for the small recognition system developed in this chapter. For bigger systems, more complex criteria may be necessary (see for example [Low04]). D_{lim} should be small enough to prevent false matches. On the other hand, it should be high enough to match interest points corresponding to the same scene point seen from a different viewpoint and under different illumination. Hence, a compromise between both conditions must be found.

The similarity threshold D_{lim} is selected before the recognition phase using two histograms which correspond to the two conditions. This is illustrated in fig. 5.9. The first histogram shows the distribution of the distances between descriptors representing different scene points (hence false matches). This is obtained by comparing the descriptors of the different interest points detected in one image. The second histogram shows the distribution of the distances between descriptors representing the same scene point under different viewing conditions (hence true matches). Ideally different viewpoints and different illuminations should be used. True matches are necessary to build this histogram. This is difficult to obtain when viewpoint changes. This is however easily obtained for illumination changes. Image series like the ones presented in sections 3.7 and 4.7 can be used: they show the same scene under different illuminations. Corresponding interest points have identical

⁴⁾ RANSAC is a robust estimation method which is often applied for recognition and localisation.

5 Application to a recognition task

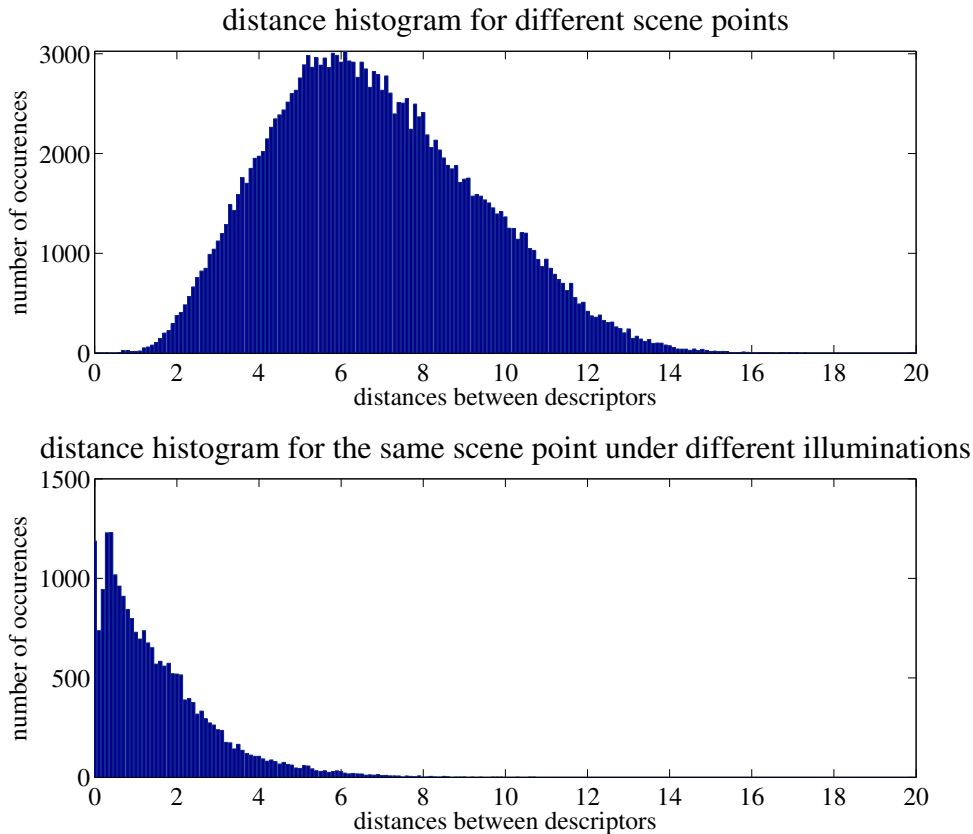


Figure 5.9: Histograms of the descriptor distances for different scene points (top) and for the same scene point under different illuminations (bottom). H–HD is used for interest point detection. The descriptors are computed as described in section 5.2 using the homomorphic derivatives. The histograms are built using several images of the database objects.

pixel coordinates. For simplicity, the second histogram is built only with interest points viewed under different illuminations here. As can be seen in fig. 5.9, distances are significantly higher for different scene points than for the same scene point under different illuminations. The threshold D_{lim} should be chosen between the two histogram peaks to reach a compromise between both conditions. A sufficient number of interest points must be used to build representative histograms. The histograms and the obtained threshold do not vary much with scene content when enough interest points are used. In this work, the histograms are built from a representative subset, which contains 200 images showing each database object under all used illuminants.

Fig. 5.10 shows how the two histograms can be used to select the threshold. If two interest points have a distance smaller than the threshold, the match is valid. Otherwise, the match is rejected. The threshold divides each histogram into two surfaces. For the distances between different points (top histogram), the histogram part to the left of the

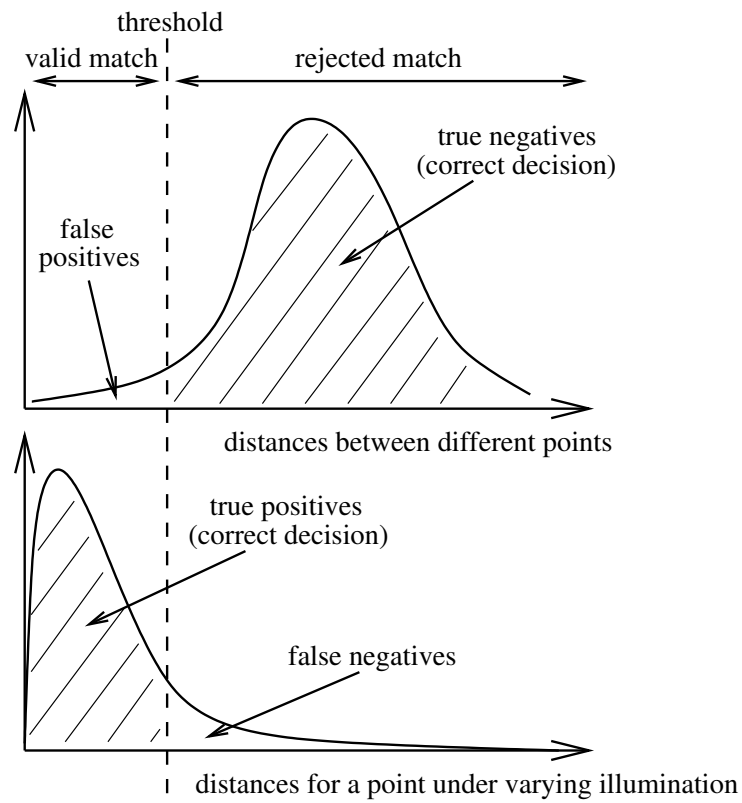


Figure 5.10: Selection of the threshold D_{lim} .

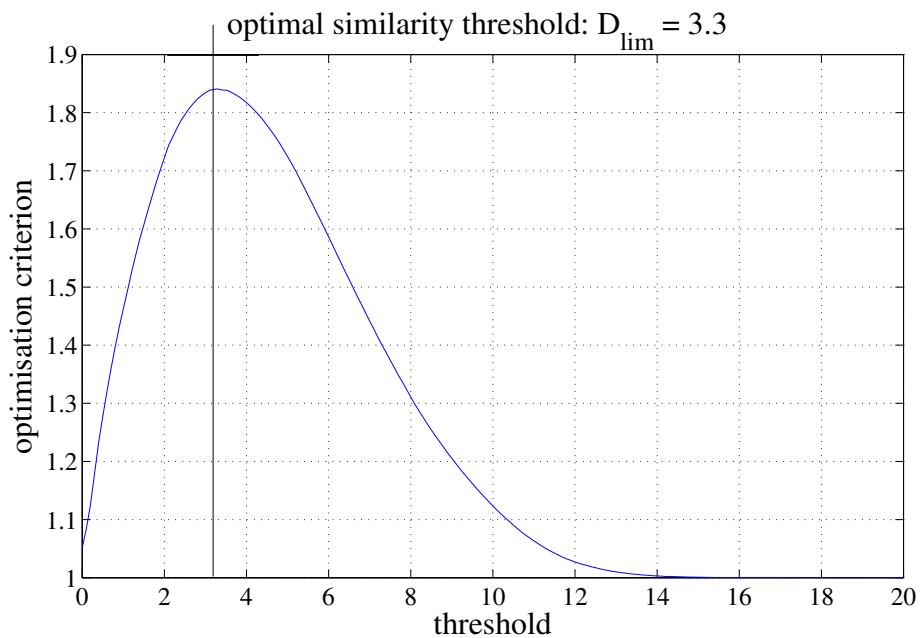


Figure 5.11: Threshold selection for the descriptors based on homomorphic grey value descriptors.

5 Application to a recognition task

threshold represents the percentage of false matches (false decision). The part to the right represents the percentage of true negatives (correct decision). For the distances for the same scene point, the histogram part to the left of the threshold represents the percentage of true matches (correct decision), while the part to the right represents the percentage of false negatives (false decision). A good threshold achieves a high percentage of correct decisions in both histograms. Therefore, the criterion for threshold selection is the sum of the two hatched areas in fig. 5.10. Maximising this criterion maximises both percentages of correct decisions. Before computing the criterion, the histograms are normalised so that the sum of their bins is 1 because both histograms are built with a different number of distances. After this normalisation, the sum of both hatched surfaces takes values between 1 and 2. The similarity threshold is the value for which the criterion reaches its maximum. The curve obtained for the histograms of fig. 5.9 is shown in fig. 5.11. The optimal threshold is $D_{lim} = 3.3$. To obtain a good threshold, a representative image subset must be used to build the histograms (with images of different objects). The thresholds obtained for the different descriptors used in this thesis are given in table 5.1.

derivatives used by the descriptor	optimal threshold
standard grey value derivatives	0.51
energy normalised grey value derivatives	20.5
homomorphic grey value derivatives	3.3
standard colour derivatives	0.94
robust invariant derivatives	1.05
homomorphic colour derivatives	5.3
3 channel m space derivatives + white balancing	3.5
3 channel m space derivatives + Nagao	3.6
2 channel m space derivatives + white balancing	1.9
2 channel m space derivatives + Nagao	2.1

Table 5.1: Thresholds for the different descriptor types.

5.4.2 Geometric constraints

The developed recognition system has only three degrees of freedom as described in section 5.1 and in fig. 5.1: the camera (or equivalently the object) is only translated horizontally and rotated about the vertical axis. Both translation and rotation components displace the scene points in a horizontal plane. Therefore, corresponding interest points have the same height from the ground. As shown in fig. 5.4, height is represented by the Y coordinate in the camera coordinate system. As a result, database and current interest points in a match p_{DB} and p should have the same vertical 3D coordinate Y . Notice that this constraint is different from having the same vertical image coordinate y because the vertical image coordinate y depends on the distance between camera and object (see eq. (5.1)). In preliminary work described in [Web05], this constraint was implemented

with: $|Y_{DB} - Y| \leq Y_{lim}$. Y_{lim} is a user-defined threshold necessary to take into account 3D reconstruction inaccuracy. Nonetheless, the inaccuracy of stereo reconstruction increases when scene points are further away from the camera. Therefore, the geometric constraint in this work takes into account the stereo reconstruction uncertainty, using a statistical model of uncertainties and the theory of uncertainty propagation presented in [Siv96].

The cameras used here for stereo reconstruction have almost a parallel configuration (see section 5.3). Stereo reconstruction is much simpler for exact parallel configuration, as shown in [TV98]. Therefore, the camera system is approximated by an ideal system with parallel configuration to compute reconstruction uncertainties. This means that the rotation matrix \mathbf{R} between both cameras in eq. (5.3) becomes the identity matrix \mathbf{Id} . The translation becomes here a horizontal translation along the \vec{X} axis: $\mathbf{t} = (t_X, 0, 0)^T$. With this ideal system, the 3D coordinates $(X, Y, Z)^T$ of a scene point are reconstructed with:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \frac{t_X}{x_L - x_R} \begin{pmatrix} x_L \\ y_L \\ f_L \end{pmatrix} = \frac{t_X}{d} \begin{pmatrix} x_L \\ y_L \\ f_L \end{pmatrix}. \quad (5.8)$$

(x_L, y_L) and (x_R, y_R) are the coordinates of the two corresponding points in the left and right image coordinate systems. Detailed explanations can be found in [TV98]. $d = x_L - x_R$ is named disparity. Stereo reconstruction depends hence on four parameters: translation t_X , disparity d , interest point position (x_L, y_L) and focal length f_L . Calibration errors are not considered here. Therefore, the two uncertainty sources are disparity d and interest point 2D position (x_L, y_L) . Disparity uncertainty is caused by inaccurate stereo correspondences. Interest point position uncertainty occurs because the image positions of current interest point and database interest point may not correspond exactly.

Measurement errors are approximated here with Gaussians. A measured value u is represented by: $\bar{u} \pm \sigma_u$, where \bar{u} is the estimated value of u and σ_u is the standard deviation of the Gaussian modelling the errors in the measurement of u . When several measurements u, \dots, w are combined with function $f(u, \dots, w)$ to obtain $x = f(u, \dots, w)$, the uncertainty on x can be estimated by:

$$\sigma_x^2 = \left(\frac{\partial f}{\partial u} \sigma_u \right)^2 + \dots + \left(\frac{\partial f}{\partial w} \sigma_w \right)^2, \quad (5.9)$$

if the uncertainties in u, \dots, w are random, independent of each others and relatively small (see [Siv96]). $\partial f / \partial u, \dots, \partial f / \partial w$ are the partial derivatives of f with respect to u, \dots, w . Here, the measured values are disparity d and interest point positions x_L and y_L . The errors on these measurements are assumed small, random and independent of each others. For matching, the uncertainty on the 3D coordinate $Y = t_X y_L / d$ is evaluated depending on the uncertainty on disparity d and interest point position (x_L, y_L) . This yields:

$$\sigma_Y^2 = \frac{t_X^2}{d^4} y_L^2 \sigma_d^2 + \frac{t_X^2}{d^2} \sigma_{y_L}^2 = \frac{Y^2}{d^2} \sigma_d^2 + \frac{t_X^2}{d^2} \sigma_{y_L}^2. \quad (5.10)$$

The uncertainty on x_L is not needed to estimate σ_Y . For matching, the current interest point p and the database interest point p_{DB} are constrained to have the same Y coordinate.

5 Application to a recognition task

For this, the difference $\delta_Y = Y_{DB} - Y$ is considered. The match is only valid if δ_Y is within the bounds of reconstruction uncertainty. Using eq. (5.9), the uncertainty of δ_Y is given by:

$$\sigma_{\delta_Y}^2 = \sigma_{Y_{DB}}^2 + \sigma_Y^2, \quad (5.11)$$

where $\sigma_{Y_{DB}}^2$ and σ_Y^2 are obtained with eq. (5.10). A match is valid if $-3\sigma_{\delta_Y} \leq \delta_Y = Y_{DB} - Y \leq 3\sigma_{\delta_Y}$. $3\sigma_{\delta_Y}$ is used because it corresponds to 99.7% of the surface of the Gaussian with standard deviation σ_{δ_Y} .

A summary of the geometric constraint is given in the following:

1. Estimate the disparity d for both interest points p and p_{DB} and for an ideal parallel camera configuration using eqs. (5.3) and (5.1) with $\mathbf{R} = \mathbf{Id}$ and $\mathbf{t} = (t_X, 0, 0)^T$.
2. Compute the uncertainties on the Y coordinates σ_Y and $\sigma_{Y_{DB}}$ with eq. (5.10).
3. Compute $\delta_Y = Y_{DB} - Y$ and σ_{δ_Y} with eq. (5.11).
4. The match between p and p_{DB} is valid if $-3\sigma_{\delta_Y} \leq \delta_Y \leq 3\sigma_{\delta_Y}$.

The uncertainty for disparity σ_d and interest point position σ_{y_L} are parameters of the matching algorithm. They are set here to: $\sigma_d = s_x$ and $\sigma_{y_L} = 2s_y$, which corresponds to uncertainties of 1 and 2 pixels. The interest point position uncertainty is bigger than the disparity uncertainty because the Harris detector is known to be geometrically inaccurate.

5.4.3 Match list constraints

The first two matching constraints deal with image similarity and 3D geometry. They consider only the two interest points in a match. The last two constraints deal on the contrary with the match list. They are related to the matching algorithm itself.

As explained in section 5.1, the database contains in this simple application one image per object. The current image is matched to each database image, as this is equivalent to matching the current image to each object. The similarity between current image and each database object is subsequently computed by the recognition and localisation algorithm, based on the matches. To keep the whole computation cost small, at most one match is created for each current interest point and for each object. This limits the number of matches that are processed during recognition and localisation. Matching quality is shown in [VL01] to increase if only the best match of each interest point is used: some true matches may be discarded but the number of suppressed false matches is much higher, resulting on the whole in higher match reliability. The best match is here the match with the highest descriptor similarity. The geometric constraint ($Y \approx Y_{DB}$) is not considered to select the best match as the accuracy of stereo reconstruction is low. To summarise, each current interest point is compared to all interest points of the considered database image. The match validity is verified using both descriptor similarity and geometric constraints. The valid match with the best descriptor similarity is added to the match list. Alternatively, the N best matches can be added in the match list. This

is interesting when the image contains repetitive patterns because many interest points are similar. In this application, it does not change matching results significantly. It is therefore not used. A last alternative is to allow one best match per interest point for the whole database. This is not used here as it may lead to an insufficient number of matches for objects with little texture and hence to a recognition failure.

The second constraint applied to the match list is symmetry. This constraint is shown in [VL01] to improve matching quality. It means that the same match list should be obtained when the roles of current image and database image are interchanged. For each database image, only the best match is created for each current interest point. Symmetric matching implies hence that the current interest point is also the best match of the database interest point. To implement the symmetry constraint, the current image is first matched to the considered database image. The interest points of the database image are then matched to the current image. Matches which are not contained in both lists are discarded. This process eliminates mainly false matches occurring when new image information appears after camera movement. The interest points detected in those new areas have no true correspondence in the database image, but they may have valid matches. As a result, several matches involve the same database interest point. The influence of the symmetry constraint is shown in fig. 5.12.

The matching process for each database image is summarised in the following:

1. For each current interest point:
 - a) For each interest point in the database image:
 - i. If both descriptor similarity and geometric constraints (see subsections 5.4.1 and 5.4.2) are satisfied:
 - A. Check if this is the best match for the current interest point using descriptor similarity.
 - b) Add the best match for the current interest point to the match list if it exists.
2. Check the match list symmetry by interchanging the roles of database and current images. Discard asymmetric matches.

This process is repeated for each database image. Matching results are shown in fig. 5.12. The processing time of the matching algorithm varies a lot depending on the data. To match 100 current interest points to 100 database interest points, 0.4ms to 1.4ms are required on the computer described in subsection 2.2.2.

5.5 Recognition and localisation

Matching delivers a match list for each database object. The recognition and localisation algorithm merges the information of these matches to verify if the image contains a database object and to determine the camera pose of the current image. Some matches

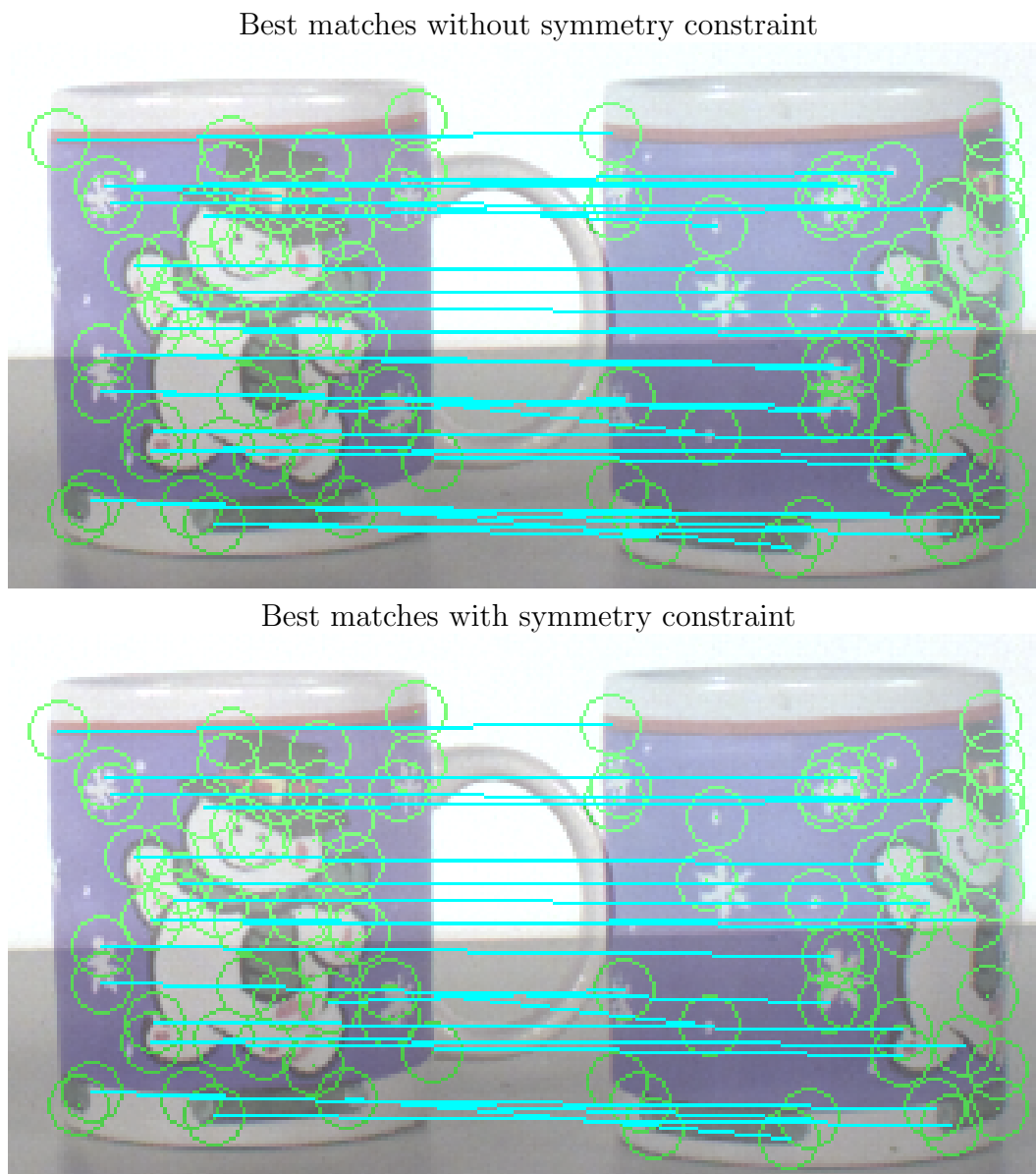


Figure 5.12: Influence of the symmetry constraint on the matching process. Descriptor similarity constraint and geometric constraint are used in both examples. White balancing is applied for visualisation. The interest points are obtained with the HC-HD. The cup on the left side is the database image. The symmetry constraint suppresses matches which involved the same database points several times, for example on snow flakes and the lower cup border.

may be false and some matches may miss. Hence, a robust recognition and localisation algorithm is necessary. The two most popular algorithms for this are the generalised Hough transform described in [Hou62, Bal81] and RANSAC described in [FB81].

5.5.1 Choice of the recognition and localisation algorithm

Generalised Hough transform and RANSAC both test many object and pose hypotheses obtained from the matches and select the best pose hypothesis. The Hough transform clusters all obtained hypotheses in an accumulator representing the object and pose space. The maximum in the accumulator is the best hypothesis: the object and pose with the highest number of consistent matches. The Hough transform is used for example in [Web05, SLL01a] for mobile robot localisation and in [Ols97, Ols01, Low04] for object recognition. The RANSAC algorithm tests a random hypothesis subset. The subset size depends on the false match probability. The quality of each object and pose hypothesis is evaluated: for example the number of consistent matches can be used. If the quality is high enough, the algorithm stops successfully. If no hypothesis in the subset is good enough, recognition fails. RANSAC is used for example in [SLL02] for mobile robot localisation and in [VL01, Tuy00] to estimate epipolar geometry. RANSAC is less computation and memory intensive than the Hough transform when there are few false matches. On the contrary, the Hough transform is more robust in the presence of many false matches. Several variants of the Hough transform have been designed to reduce its computing time or its memory requirement. An overview is given in [Ols97]. In [Ols01], a hybrid algorithm is presented, which combines advantages of both Hough transform and RANSAC.

This chapter evaluates the influence of interest point stability on recognition and localisation. The Hough transform is interesting for this evaluation because analysing the accumulators gives information about the percentage of false matches. As a consequence, the generalised Hough transform is used for recognition and localisation here. One accumulator is associated to each object in the database. All accumulators are filled using the pose hypotheses generated from the match list. This is presented in more detail in subsection 5.5.2. This accumulator filling algorithm is enhanced in subsection 5.5.3 to take geometric uncertainties into account. Finally, the content of the accumulator is interpreted to verify if one database object is contained in the current image. This is explained in subsection 5.5.4.

5.5.2 Filling the accumulators

One accumulator is created for each database object. Each object is considered independently of the others. The accumulator represents the pose space, i.e. all theoretically possible camera poses. It is discretised into bins, similarly to a histogram. To fill the accumulator, the matches are used to generate all possible pose hypotheses. The bins corresponding to the generated pose hypotheses are incremented. The accumulator represents hence the distribution of the pose hypotheses in the pose space. The maximum

5 Application to a recognition task

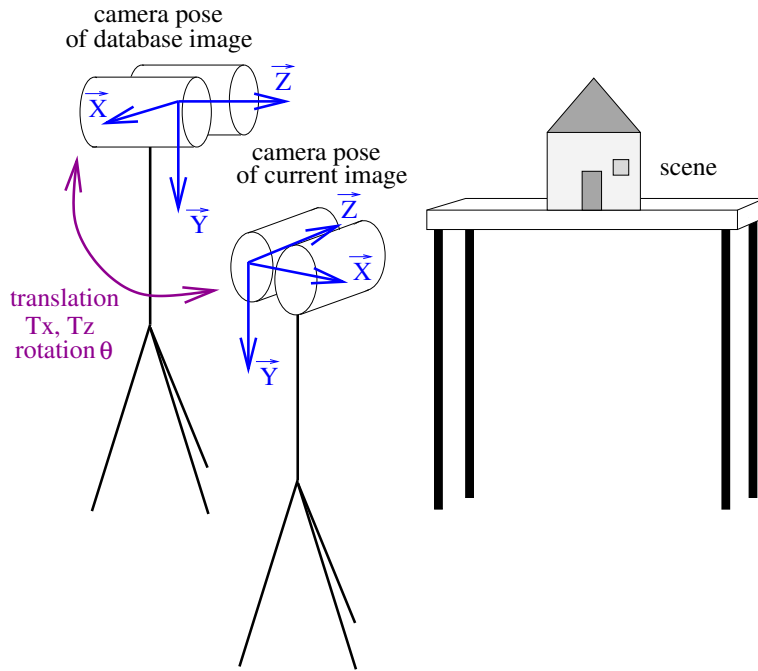


Figure 5.13: Used coordinate systems and estimated localisation parameters.

in the accumulator is the best pose hypothesis as it is supported by the highest number of matches. The Hough transform is hence similar to voting.

In this application, three degrees of freedom are allowed. Therefore, three parameters describe camera pose: θ for the rotation about the vertical axis and T_X, T_Z for the horizontal translation. This is illustrated in fig. 5.13. The 3D positions of the database interest point $(X^{DB}, Y^{DB}, Z^{DB})^T$ and of the current interest point $(X, Y, Z)^T$ are related by:

$$\begin{pmatrix} X^{DB} \\ Y^{DB} \\ Z^{DB} \end{pmatrix} = \begin{pmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} T_X \\ 0 \\ T_Z \end{pmatrix}. \quad (5.12)$$

The rotation angle θ is defined here about the vertical axis pointing up (i.e. opposite to \vec{Y}). Each match between database and current image provides only two equations, because $Y^{DB} = Y$ (see eq. (5.12))⁵⁾. The system has three degrees of freedom, so a match pair is necessary to obtain a pose hypothesis. The accumulator is filled here using match pairs, because this simplifies uncertainty consideration in subsection 5.5.3. Another solution would consist in filling the accumulators with curves corresponding to single matches.

A match pair provides an overdetermined nonlinear equation system:

$$\begin{cases} X_a^{DB} = \cos \theta X_a - \sin \theta Z_a + T_X \\ Z_a^{DB} = \sin \theta X_a + \cos \theta Z_a + T_Z \\ X_b^{DB} = \cos \theta X_b - \sin \theta Z_b + T_X \\ Z_b^{DB} = \sin \theta X_b + \cos \theta Z_b + T_Z \end{cases} \quad (5.13)$$

⁵⁾ $Y^{DB} = Y$ is used to constrain matching (see subsection 5.4.2).

where $((X_a^{DB}, Y_a^{DB}, Z_a^{DB})^T, (X_a, Y_a, Z_a)^T)$ and $((X_b^{DB}, Y_b^{DB}, Z_b^{DB})^T, (X_b, Y_b, Z_b)^T)$ represent the 3D positions of the points in the two matches. Using eq. (5.13), $\cos \theta$ and $\sin \theta$ can be computed. This yields θ and a condition on the match pair (for $\cos^2 \theta + \sin^2 \theta = 1$):

$$\begin{cases} C = \cos \theta = \frac{(X_a^{DB} - X_b^{DB})(X_a - X_b) + (Z_a^{DB} - Z_b^{DB})(Z_a - Z_b)}{(X_a - X_b)^2 + (Z_a - Z_b)^2} \\ S = \sin \theta = \frac{(Z_a^{DB} - Z_b^{DB})(X_a - X_b) - (X_a^{DB} - X_b^{DB})(Z_a - Z_b)}{(X_a - X_b)^2 + (Z_a - Z_b)^2} \\ \theta = \arctan S/C \\ (X_a^{DB} - X_b^{DB})^2 + (Z_a^{DB} - Z_b^{DB})^2 = (X_a - X_b)^2 + (Z_a - Z_b)^2 \end{cases} \quad (5.14)$$

From a geometrical point of view, θ is the angle between segment $[P_a^{DB} P_b^{DB}]$ from the database image and segment $[P_a P_b]$ from the current image. P_i and P_i^{DB} are the 3D points represented by $(X_i, Y_i, Z_i)^T$ and $(X_i^{DB}, Y_i^{DB}, Z_i^{DB})^T$ for $i = a$ or b . The condition on the match pair constrains the length of segments $[P_a^{DB} P_b^{DB}]$ and $[P_a P_b]$ to be equal. In practise, this constraint is implemented with:

$$-\Delta_{lim} \leq \Delta = ((X_a^{DB} - X_b^{DB})^2 + (Z_a^{DB} - Z_b^{DB})^2) - ((X_a - X_b)^2 + (Z_a - Z_b)^2) \leq \Delta_{lim}, \quad (5.15)$$

where Δ_{lim} is a distance threshold. Once θ is known, T_X and T_Z are easily obtained with:

$$\begin{cases} T_X = X_i^{DB} - \cos \theta X_i + \sin \theta Z_i = X_i^{DB} - C X_i + S Z_i \\ T_Z = Z_i^{DB} - \sin \theta X_i - \cos \theta Z_i = Z_i^{DB} - S X_i - C Z_i \end{cases}, \text{ where } i = a \text{ or } b. \quad (5.16)$$

From a geometrical point of view, translation (T_X, T_Z) is obtained as the vector between P_i^{DB} and P_i after correction of the rotation ($i = a$ or b). In practise, a more accurate solution is obtained with the mean over both matches a and b :

$$\begin{cases} T_X = (X_a^{DB} + X_b^{DB})/2 - C(X_a + X_b)/2 + S(Z_a + Z_b)/2 \\ T_Z = (Z_a^{DB} + Z_b^{DB})/2 - S(X_a + X_b)/2 - C(Z_a + Z_b)/2 \end{cases}. \quad (5.17)$$

To conclude, an overview of a simple accumulator filling algorithm is given in the following:

1. For all matches $a = 1$ to N_{match}
 - For all matches $b = a$ to N_{match}
 - a. Compute Δ with eq. (5.15).
 - b. If eq. (5.15) is verified (distance condition on the match pair):
 - i. Compute $C = \cos \theta$ and $S = \sin \theta$ with eq. (5.14).
 - ii. Compute θ with eq. (5.14).
 - iii. If $-90^\circ < \theta < 90^\circ$:
 - A. Compute T_X, T_Z with eq. (5.17).
 - B. Increment the accumulator bin corresponding to θ, T_X, T_Z by 1.

5 Application to a recognition task

In step iii, pose hypotheses for which the estimated angle is bigger than 90° or smaller than -90° are discarded, because such rotation is impossible: a completely different part of the object would be visible. This simple algorithm is enhanced in subsection 5.5.3 to take uncertainties into account. In order to reduce the required memory space, two accumulators are used. One 1D accumulator represents rotation parameter θ . The second accumulator is a 2D accumulator representing translation T_X, T_Z . This does not decrease much the localisation performance because rotation and translation parameters are well correlated: different angles θ correspond to different translation parameters T_X, T_Z . This reduces well memory requirements: only $B_{T_X} \times B_{T_Z} + B_\theta$ bins are required instead of $B_{T_X} \times B_{T_Z} \times B_\theta$, where B_i is the number of bins for dimension $i = T_X, T_Z$ or θ . This modifies only slightly the algorithm: step B is replaced by “Increment the accumulator bin corresponding to θ and the accumulator bin corresponding to T_X, T_Z by 1”.

5.5.3 Taking uncertainties into account

Only geometric information is directly taken into account for accumulator filling. Image information is used indirectly by means of the matches. Stereo reconstruction is less accurate for points situated far away from the camera. In addition, the Harris detector is known to be geometrically inaccurate. Therefore the simple algorithm filling algorithm in subsection 5.5.2 is enhanced here to take these geometric uncertainties into account. This improves both recognition and localisation performances.

Uncertainty is modelled and propagated like in subsection 5.4.2 for matching. The stereo system is approximated by a parallel configuration. For each 3D point, two uncertainty factors are considered: uncertainty of the interest point position in the image (x, y) and uncertainty of the disparity d for stereo reconstruction. These uncertainty factors are modelled by Gaussians with mean equal to 0 and with small variances. They are assumed to be independent of each other. Therefore, the uncertainty propagation framework of subsection 5.4.2 can be used. If several measurements u, \dots, w are combined with function $f(u, \dots, w)$ to obtain $x = f(u, \dots, w)$, the uncertainty on x is estimated with:

$$\sigma_x^2 = \left(\frac{\partial f}{\partial u} \sigma_u \right)^2 + \dots + \left(\frac{\partial f}{\partial w} \sigma_w \right)^2. \quad (5.18)$$

This framework is applied to T_X, T_Z and θ . These depend on the X and Z coordinates of four points P_a^{DB}, P_b^{DB}, P_a and P_b . According to eq. (5.8), X depends on horizontal image position x and on disparity d and Z depends only on disparity d . The uncertainty is modelled with the same variances σ_x and σ_d for all four points. This yields:

$$\begin{aligned} \sigma_{T_X}^2 &= \left(\frac{\partial T_X}{\partial x_a^{DB}}{}^2 + \frac{\partial T_X}{\partial x_b^{DB}}{}^2 + \frac{\partial T_X}{\partial x_a}{}^2 + \frac{\partial T_X}{\partial x_b}{}^2 \right) \sigma_x^2 + \left(\frac{\partial T_X}{\partial d_a^{DB}}{}^2 + \frac{\partial T_X}{\partial d_b^{DB}}{}^2 + \frac{\partial T_X}{\partial d_a}{}^2 + \frac{\partial T_X}{\partial d_b}{}^2 \right) \sigma_d^2 \\ \sigma_{T_Z}^2 &= \left(\frac{\partial T_Z}{\partial x_a^{DB}}{}^2 + \frac{\partial T_Z}{\partial x_b^{DB}}{}^2 + \frac{\partial T_Z}{\partial x_a}{}^2 + \frac{\partial T_Z}{\partial x_b}{}^2 \right) \sigma_x^2 + \left(\frac{\partial T_Z}{\partial d_a^{DB}}{}^2 + \frac{\partial T_Z}{\partial d_b^{DB}}{}^2 + \frac{\partial T_Z}{\partial d_a}{}^2 + \frac{\partial T_Z}{\partial d_b}{}^2 \right) \sigma_d^2 \\ \sigma_\theta^2 &= \left(\frac{\partial \theta}{\partial x_a^{DB}}{}^2 + \frac{\partial \theta}{\partial x_b^{DB}}{}^2 + \frac{\partial \theta}{\partial x_a}{}^2 + \frac{\partial \theta}{\partial x_b}{}^2 \right) \sigma_x^2 + \left(\frac{\partial \theta}{\partial d_a^{DB}}{}^2 + \frac{\partial \theta}{\partial d_b^{DB}}{}^2 + \frac{\partial \theta}{\partial d_a}{}^2 + \frac{\partial \theta}{\partial d_b}{}^2 \right) \sigma_d^2 \end{aligned} \quad (5.19)$$

The complete formulae can be easily derived from eqs. (5.14) and (5.17). To take uncertainties into account, the accumulator filling is performed using a Gaussian with variance computed with eq. (5.19) instead of incrementing one single bin for θ or for T_X, T_Z . The Gaussian is not normalised: for example $G(x) = \exp(-(x - \theta)^2/\sigma_\theta^2)$ for angle hypothesis θ . This ensures that the bin values in the accumulator corresponds better to the number of match pairs consistent with the pose hypothesis. This improves recognition and localisation in practise because the influence of false hypotheses with low uncertainty is decreased. Each accumulator bin also stores the number of votes, i.e. the number of times it has been updated. This number represents exactly the number of match pairs consistent with this pose, and is used for recognition.

If the uncertainty on θ is high, the uncertainty on T_X, T_Z is also high because T_X and T_Z depend on $\cos \theta$ and $\sin \theta$ (see eq. (5.17)). This is used as an additional filter: only the pose hypotheses with an angle uncertainty σ_θ smaller than σ_{lim} are used to update the accumulators. The threshold σ_{lim} is set here to 18° . This discards match pairs for which the size of segments $[P_a^{DB} P_b^{DB}]$ and $[P_a P_b]$ is too small to reliably estimate θ (see subsection 5.5.2). Last, the uncertainties are taken into account for the match pair constraint in eq. (5.15). The threshold Δ_{lim} is set to $\Delta_{lim} = 3\sigma_\Delta$, similarly to the geometric constraint for matching in subsection 5.4.2. σ_Δ is obtained like the uncertainties on rotation and translation with:

$$\sigma_\Delta^2 = \left(\frac{\partial \Delta^2}{\partial x_a^{DB}} + \frac{\partial \Delta^2}{\partial x_b^{DB}} + \frac{\partial \Delta^2}{\partial x_a} + \frac{\partial \Delta^2}{\partial x_b} \right) \sigma_x^2 + \left(\frac{\partial \Delta^2}{\partial d_a^{DB}} + \frac{\partial \Delta^2}{\partial d_b^{DB}} + \frac{\partial \Delta^2}{\partial d_a} + \frac{\partial \Delta^2}{\partial d_b} \right) \sigma_d^2. \quad (5.20)$$

The resulting accumulator filling algorithm is recapitulated in the following:

1. For all matches $a = 1$ to N_{match}
 - For all matches $b = a$ to N_{match}
 - a.** Compute distance threshold $\Delta_{lim} = 3\sigma_\Delta$, where σ_Δ is given by eq. (5.20).
 - b.** If eq. (5.15) is verified (distance condition on the match pair):
 - i.** Compute $C = \cos \theta$ and $S = \sin \theta$ with eq. (5.14).
 - ii.** Compute θ with eq. (5.14) and σ_θ with eq. (5.19).
 - iii.** If $-90^\circ < \theta < 90^\circ$ and if $\sigma_\theta \leq \sigma_{lim} = 18^\circ$
 - A.** Compute T_X, T_Z with eq. (5.17) and $\sigma_{T_X}, \sigma_{T_Z}$ with eq. (5.19).
 - B.** Update the angle accumulator using the Gaussian centred on θ with standard deviation σ_θ , and the translation accumulator using the Gaussian centred on T_X, T_Z with standard deviations $\sigma_{T_X}, \sigma_{T_Z}$.

All changed steps of the simple accumulator filling algorithm are indicated in boldface. The uncertainty for disparity σ_d and interest point position σ_{y_L} are set here to: $\sigma_d = s_x/2$ and $\sigma_{x_L} = s_x$, which corresponds to uncertainties of half a pixel and one pixel. The interest

point position uncertainty is bigger than the disparity uncertainty because of the geometric inaccuracy of the Harris detector. The accumulator size and discretisation must be adapted to the application. Only limited viewpoints are handled. In addition, the distance between camera and scene is small. Hence, small camera motions induce significant viewpoint changes. The translation accumulators cover the interval $[-1\text{m}, 1\text{m}] \times [-1\text{m}, 1\text{m}]$ with bins of size $1\text{cm} \times 1\text{cm}$. The rotation accumulators cover the interval $[-180^\circ, 180^\circ]$ with bins of size 1° . Accumulator filling is illustrated in fig. 5.14. The effect of false matches is visible on the accumulators: two peaks are obtained. The estimated pose parameters correspond to the manually estimated parameters. As shown in [Web05], this algorithm can also be used for mobile robot localisation. The processing time for accumulator filling varies strongly. Less than 1ms to a few ms are required when the two images show different objects. Several hundreds of ms are required when the two images show the same object. When approximately 100 interest points are detected for both objects, accumulator filling requires about 200ms on the computer presented in subsection 2.2.2.

5.5.4 Interpreting the accumulators

Once all accumulators are filled, these are interpreted to verify if a database object is contained in the current image and to estimate the pose parameters. This is performed in two steps: first the best pose hypothesis is obtained for each object, then the quality of the object and pose hypotheses is evaluated. Recognition is based on this quality.

The bin with the maximum value in each accumulator provides the best pose hypothesis for each object. This is performed for both rotation and translation accumulators. To reduce the influence of pose space discretisation, the peak position is interpolated. Similarly to the subpixel stereo correspondences, a paraboloid is fitted to the bin values around the maximum in the translation accumulator and a parabola is fitted to the bin values around the maximum in the rotation accumulator. The interpolated peak position is obtained by setting the derivatives of the fitted curves to zero. More detail is given in subsection 5.3.3.

Once a best pose is obtained for each object, the quality of these hypotheses is evaluated. When two images of the same object are matched, many matches are obtained and the accumulators have one distinct peak like in fig. 5.14. On the contrary, when two images of different objects are matched, few matches are obtained and the accumulators are empty or contain few hypotheses with only few votes. This is illustrated in fig. 5.15. In comparison to fig. 5.14, a similar number of matches is obtained although a much higher number of interest points are detected. The peak has only 4 votes in fig. 5.15, in comparison to 38 votes in fig. 5.14. Therefore, two criteria based on the number of votes are used to evaluate the quality of a pose hypothesis. A threshold is defined for each criterion. An object is only recognised if both thresholds are exceeded.

The first criterion is simply the number of votes that contributed to the pose hypothesis. The number of votes is influenced by the number of interest points detected in the current image. Therefore, hypothesis quality is evaluated by comparing all hypotheses to each

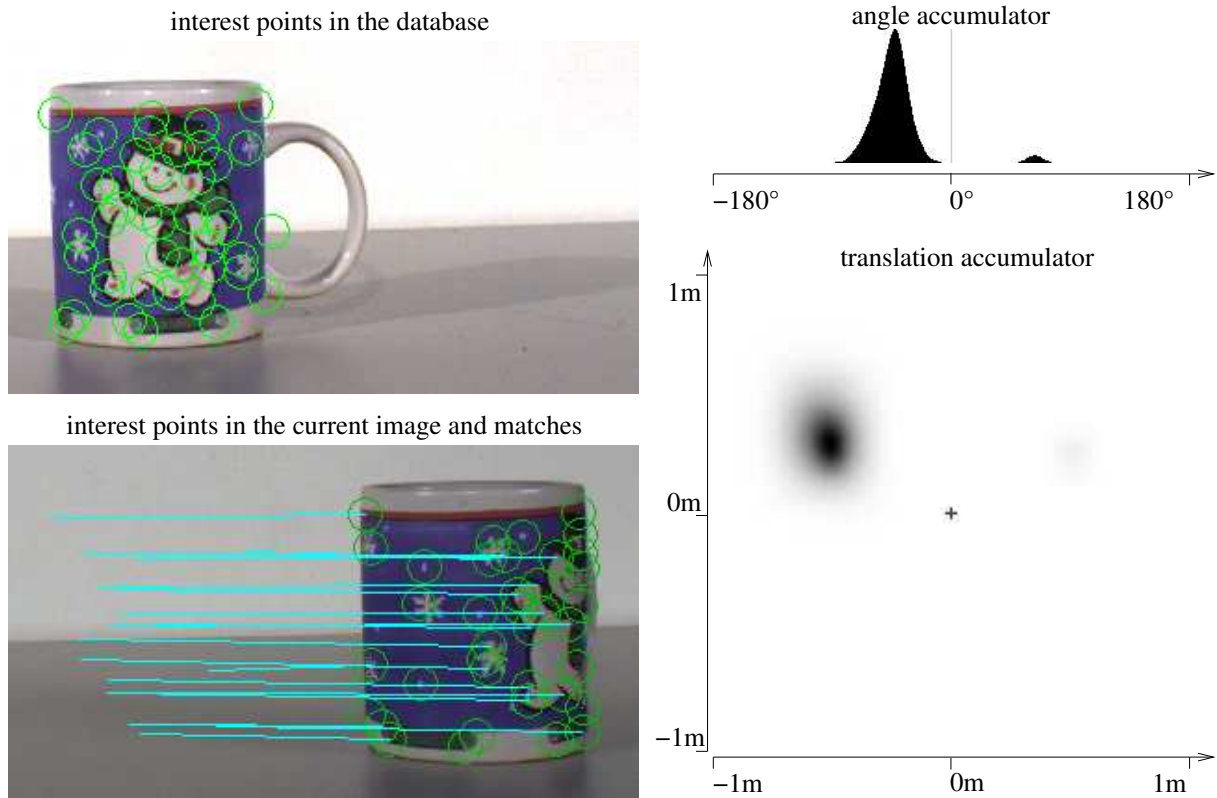


Figure 5.14: Example of accumulator filling (uncertainties are considered). In the translation accumulator, white represents 0 and black represents the maximum accumulator value. White balancing is used for visualisation only. The interest point detector is HC-HD. The database image is described by 39 interest points. The current image is described with 28 points. 18 matches are obtained. The peaks in both accumulators have 38 votes. The estimated parameters are $\theta = -42.7^\circ$, $T_X = -0.505\text{m}$ and $T_Z = 0.299\text{m}$. A manual localisation leads to $\theta = -51^\circ$, $T_X = -0.546\text{m}$ and $T_Z = 0.371\text{m}$.

others. This comparison is based on the assumption that a single database object is visible in the image. The object and pose hypotheses with the highest number of votes and with the second highest number of votes are determined. If a database object is contained in the current image, the number of votes of the best hypothesis should be notably higher than the number of votes of the second best hypothesis. The quality threshold therefore depends on the number of votes of the second best hypothesis. In addition to this condition, a minimum number of votes is necessary to obtain a good object and pose hypothesis. As a result, the quality of the best pose hypothesis is verified with: $N_{max} > N_{lim} = a(N_{2ndmax} + 1)$, where N_{max} and N_{2ndmax} is the number of votes of the best and of the second best hypotheses. $N_{2ndmax} + 1$ is used to handle the case when $N_{2ndmax} = 0$. a has been chosen experimentally here: $a = 2$ when grey value images or two channel images (for the 2 channel M space Harris detector) are used and $a = 3$ when three channel images are used.

5 Application to a recognition task

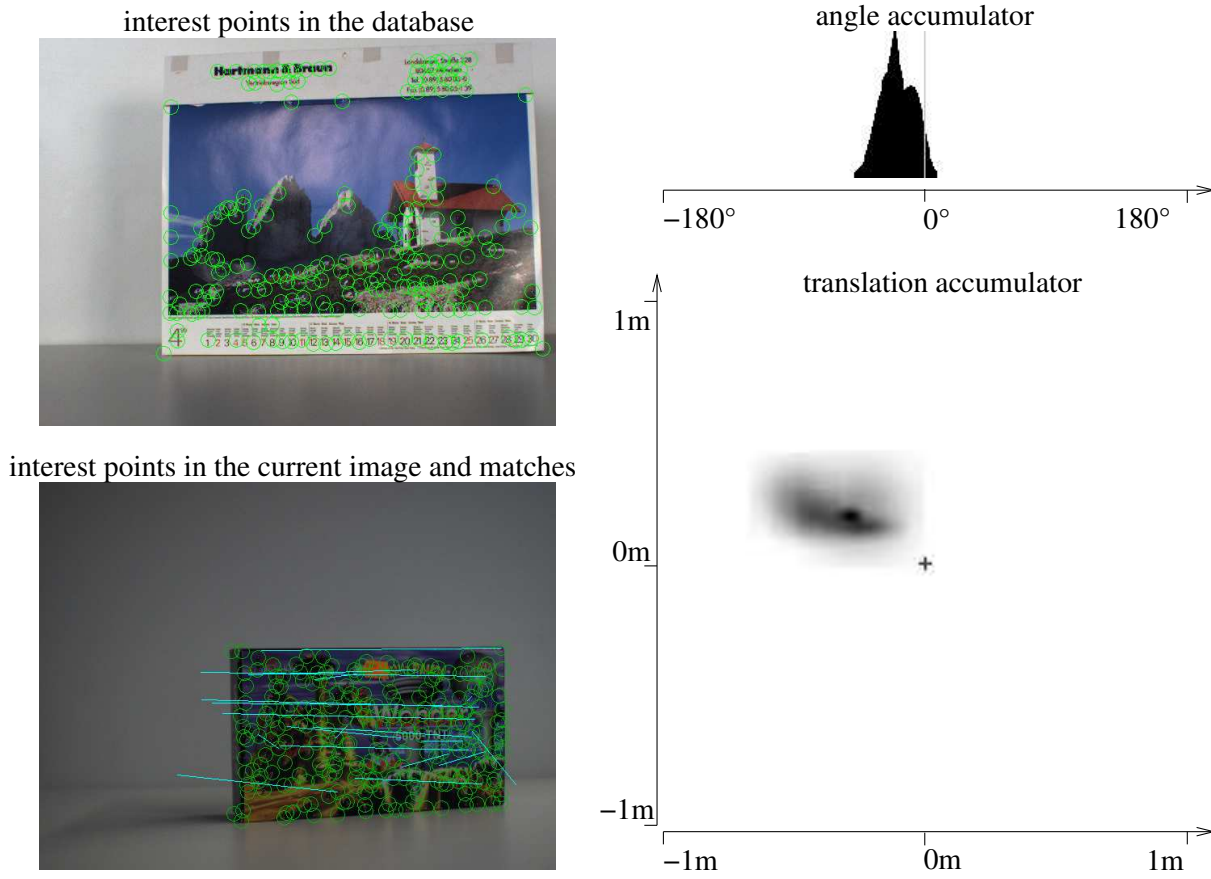


Figure 5.15: Example of accumulator filling for images of two different objects. White balancing is used for visualisation only. The interest point detector is the HC–HD. The database image is characterised by 188 interest points. The current image is characterised by 230 points. 19 matches are obtained. The peaks in both accumulators have 4 votes.

The second criterion is the percentage of match pairs that contributed to the pose hypothesis. It is named here match consistency. Unlike the first criterion, quality is not evaluated with a comparison between different object hypotheses. Match consistency is computed for each object separately. It is the percentage of votes that contributed to the best hypothesis relatively to the number of match pairs, which is the maximum possible number of votes. If N_{match} matches are obtained, $N_{match}(N_{match} - 1)/2$ distinct match pairs exist. Therefore, match consistency is defined as:

$$mc = \frac{2 N_{votes}}{N_{match}(N_{match} - 1)}, \quad (5.21)$$

where N_{votes} is the number of votes that contributed to the best hypothesis. Match consistency depends on the percentage of false matches which result in false pose hypotheses. It is also influenced by the percentage of invalid match pairs (match pairs for which $|\Delta| > \Delta_{lim}$ or $\sigma_\theta > \sigma_{lim}$, see subsection 5.5.3). Here, the threshold mc_{lim} has been set

experimentally to 0.25: 25% of the distinct match pairs should contribute to the peak in the accumulator.

To conclude, the accumulator interpretation is summarised in the following:

1. For each database object:
 - a) Determine the bins with the maximum value in rotation and translation accumulators and their corresponding number of votes.
 - b) Interpolate the peak position to obtain the best pose hypothesis.
2. Determine the object and pose hypotheses with the highest and second highest number of votes N_{max} and $N_{2nd\ max}$.
3. If $N_{max} > N_{lim} = a(N_{2nd\ max} + 1)$:
 - a) Compute the match consistency mc of the best object and pose hypothesis with eq. (5.21).
 - b) If $mc > mc_{lim}$
 - i. The result of the algorithm is the best object and pose hypothesis.
4. In all other cases, no object is recognised.

The interpretation is based on the assumption that a single database object is contained in the image. The algorithm should hence be modified if several database objects can be present. For example, the number of votes of each best pose hypothesis could be compared to the number of votes of second best pose hypothesis for the same object.

5.6 Evaluation framework

The goal of this chapter is to evaluate the influence of interest point detectors in a recognition application. Several detectors are compared to each others using the presented recognition and localisation system. A database of 10 objects is created. The recognition and localisation performances are evaluated with test images of those objects under different illuminations and viewed from different viewpoints. The evaluation criteria are introduced in subsection 5.6.1. An overview of the evaluated detectors is given in subsection 5.6.2. Finally, object database and test images are presented in subsection 5.6.3.

5.6.1 Evaluation criteria

The first two evaluation criteria are recognition rate and localisation accuracy. They characterise the quality of recognition and localisation. To compute both criteria, the results of the system are compared to a manual solution. The manual pose is obtained from manual point matches between the two left images of current and database stereo

5 Application to a recognition task

images. The 3D position of these manual points is reconstructed as described in section 5.3. Accumulators are filled using the 3D positions and the manual matches as explained in subsection 5.5.3. The manual pose is the accumulator maximum (see subsection 5.5.4). The manual poses are therefore obtained with the developed recognition and localisation system, except that interest point detection and matching are performed manually. The recognition rate is the ratio between the number of correctly recognised objects and the number of test images. Localisation accuracy is computed with:

$$\text{localisation accuracy} = \sqrt{\frac{(\theta - \theta^m)^2}{s_\theta^2} + \frac{(T_X - T_X^m)^2}{s_{T_X}^2} + \frac{(T_Z - T_Z^m)^2}{s_{T_Z}^2}} \quad (5.22)$$

where (θ, T_X, T_Z) and (θ^m, T_X^m, T_Z^m) are the system and the manual poses. s_θ , s_{T_X} and s_{T_Z} are the sizes of the accumulator bins in the three dimensions θ , T_X and T_Z (here, 1° and 1cm). Localisation accuracy is only computed if the object is correctly recognised.

The third evaluation criterion is match consistency mc defined in eq. (5.21) of subsection 5.5.4. It characterises the proportion of pose hypotheses consistent with the best hypothesis in the accumulators. It is therefore influenced by the percentage of false matches. In addition, it also depends on the percentage of invalid or uncertain match pairs (see eq. (5.15) and $\sigma_\theta \leq \sigma_{lim}$ in subsection 5.5.3). The next criterion is the deviation from the highest peak in the accumulators. The accumulators are similar to histograms. Therefore, deviation from the peak is computed with:

$$\text{deviation} = \sqrt{\frac{\sum_{b=1}^B \text{acc}(b) \text{dist}(b, b^{max})^2}{\sum_{b=1}^B \text{acc}(b)}} \quad (5.23)$$

where $\text{acc}(b)$ represents the value of accumulator bin b . B is the number of accumulator bins. $\text{dist}(b, b^{max})$ is the distance between the current accumulator bin b and the bin with the maximum value b^{max} . The pose space is represented by two accumulators: one for rotation and one for translation. The pose deviation is therefore computed with:

$$\text{deviation} = \sqrt{\frac{\sum_{b=1}^{B_\theta} \text{acc}_\theta(b) \text{dist}_\theta(b, b_\theta^{max})^2}{\sum_{b=1}^{B_\theta} \text{acc}_\theta(b)} + \frac{\sum_{b=1}^{B_T} \text{acc}_T(b) \text{dist}_T(b, b_T^{max})^2}{\sum_{b=1}^{B_T} \text{acc}_T(b)}} \quad (5.24)$$

$$\text{with } \begin{cases} \text{dist}_\theta(b, b_\theta^{max})^2 = (\theta^b - \theta^{max})^2 / s_\theta^2 \\ \text{dist}_T(b, b_T^{max})^2 = (T_X^b - T_X^{max})^2 / s_{T_X}^2 + (T_Z^b - T_Z^{max})^2 / s_{T_Z}^2. \end{cases}$$

acc_θ and acc_T represent rotation and translation accumulators. (θ^b, T_X^b, T_Z^b) is the pose corresponding with bin b . $(\theta^{max}, T_X^{max}, T_Z^{max})$ is the best pose hypothesis. s_θ , s_{T_X} and s_{T_Z} are the sizes of the accumulator bins in the three dimensions θ , T_X and T_Z . The pose deviation is similar to match consistency because it is influenced by false pose hypotheses. It depends additionally on the distances between false pose hypotheses and the best pose hypothesis as well as on geometric uncertainties. Both match consistency and pose deviation are only computed when the object is correctly recognised.

These four criteria all evaluate the performance of the whole system. Three further criteria are added, which are focused on the suitability of interest point detection for recognition

and localisation. The first criterion is the percentage of characterised points, i.e. points for which a descriptor and 3D position exist. As explained in section 5.3, 3D position is only estimated if the stereo correspondence is unique. Therefore the percentage of characterised interest points depends on the uniqueness of the interest point textures in their neighbourhoods, i.e. on their information content. It is computed with:

$$\text{percentage of characterised points} = \frac{\text{number of characterised interest points}}{\text{number of detected interest points}} \quad (5.25)$$

The next criterion is the percentage of matched interest points. It is related to detection stability between two images of an object. It is also influenced by the stability and the discriminative power of the descriptors. The percentage of matched interest points is:

$$\text{percentage of matched points} = \frac{\text{number of matches}}{\text{number of detected interest points}} \quad (5.26)$$

It is only computed when two images of the same object are matched. Finally, the last criterion is the percentage of consistent points. It is computed similarly to match consistency (see eq. (5.21)), but uses the number of interest points detected in the current image. Only one match is allowed per interest point, so the maximum number of possible matches between two images of the same object is the number of interest points. A match pair is necessary to compute a pose hypothesis. Therefore the percentage of consistent points is approximated as the square root of point consistency:

$$\text{percentage of consistent points} = \sqrt{\frac{2 N_{votes}}{N_{points} (N_{points} - 1)}} \quad (5.27)$$

where N_{votes} is the number of votes for the best pose hypothesis and N_{points} is the number of interest points detected in the current image pair. This criterion is only computed when two images of the same object are matched.

5.6.2 Compared interest point detectors

The evaluation cannot be performed for all developed interest point detectors. A few detectors for grey value images and for colour images are selected. This is explained in this subsection.

Four detectors for grey value images are selected. The homomorphic Harris detector (H-HD) is chosen because it achieves the highest stability in chapter 3. In addition, the locally adaptive thresholding Harris detector (AT-HD) is evaluated. It achieves lower stability than H-HD, but it yields a more homogeneous distribution of the interest points in the image and the distances between interest points are bigger. This is interesting for localisation. The standard Harris detector (HD) is used as reference for state of the art detectors. Like in chapter 3, for HD, the N interest points with the highest cornerness values are selected. This adapts detection to the global lighting conditions. The

5 Application to a recognition task

detection of a fixed number of interest points leads to good recognition and localisation performances. Therefore, a variant of the homomorphic Harris detector is introduced here, for which the N best interest points are selected (H–HD+NBP). For all detectors, the same detection thresholds as in chapter 3 are used for all images. For H–HD, T is set to 10^{-5} . For AT–HD, T_2 is set to 2. For HD and H–HD+NBP, $N = 100$ interest points are detected. The other parameters are set as given in the algorithm descriptions.

Five colour detectors are selected. The standard colour Harris detector (C–HD) is used as reference for the existing colour interest point detectors. Like for the grey value variant (HD), the N best points are selected. The homomorphic colour Harris detector (HC–HD) is also evaluated. Last, three detectors based on chrominance are selected. Preliminary tests showed that using three channels enhances recognition and localisation results. Therefore, the three channel versions of the M Space Harris detector are evaluated here for the two preprocessing methods: with white balancing before demosaicing (3MS–HD+WB) and with Nagao preprocessing (3MS–HD+N). The two channel M Space Harris detector with Nagao preprocessing (2MS–HD+N) is also evaluated to show the performance gain with the third chrominance channel. The results of the robust invariant Harris detector (RI–HD) are very similar to the 3MS–HD+WB, but RI–HD requires a longer processing time. Its results are therefore not included here. The same detection thresholds as in chapter 4 are used. For C–HD, $N = 100$ interest points are detected. For HC–HD, T is set to 10^{-4} . For 2MS–HD+N, 3MS–HD+N and 3MS–HD+WB, T is set to 10^{-5} .

Following detectors are based on invariant derivatives: H–HD, H–HD+NBP, HC–HD, 2MS–HD+N, 3MS–HD+N and 3MS–HD+WB. The invariant derivatives are re-used to compute the descriptors. For HD, AT–HD and C–HD, the descriptors are computed from standard derivatives and they are normalised as indicated in section 5.2 to compensate illumination influence. All parameters for stereo reconstruction, matching and recognition are set as given in sections 5.3, 5.4 and 5.5.

5.6.3 Object database and test images

The image acquisition framework (camera type, camera parameters, demosaicing algorithm and computation of grey value images) described in subsection 3.7.3 is used. The database is created with one image for each of the 10 different objects. The objects have different reflection properties (Lambertian or specular reflection), different reflectance properties (structured or textured reflectance) and different 3D geometry (simple or complex 3D geometry). The images used for database creation are shown in fig. 5.16. Objects 4, 5 and 6 have more specular reflections than the others. Objects 3, 6, 8 and 9 have textured reflectances (i.e. many similar interest points), whereas the others have structured reflectances. Objects 4, 5, 7, 8, and 10 have simpler 3D geometry than the other objects. All database images have an approximately homogeneous illumination.

The test images show these 10 objects from different viewpoints and under different illuminations. For each object, 5 different viewpoints and 20 different illuminations are used during image acquisition, leading to 100 images per object: 1 database image and

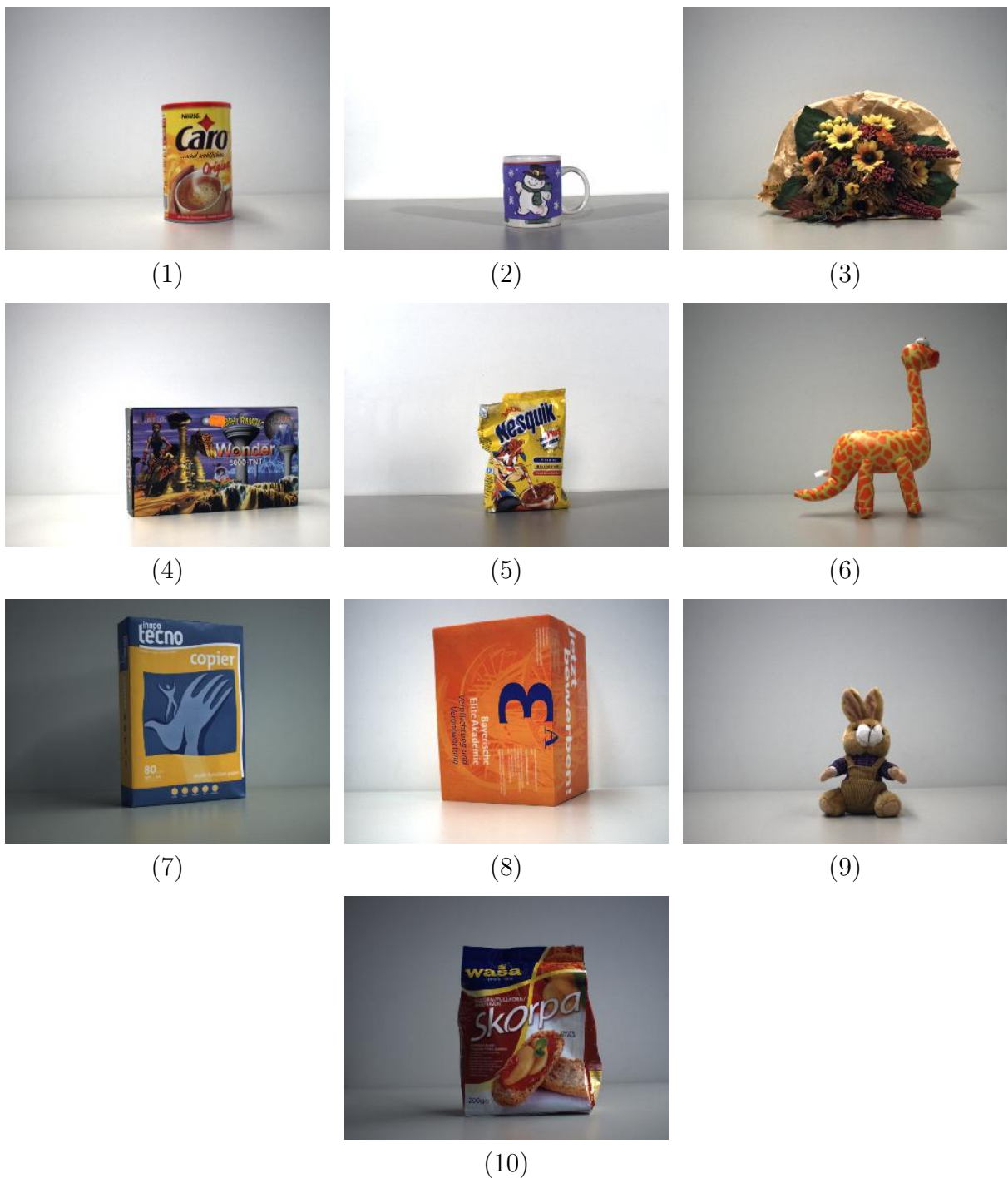


Figure 5.16: Images used to create the database. Manual white balancing has been applied for visualisation. The object numbers are given under the images.

5 Application to a recognition task



Figure 5.17: The five different poses for the test images of object 10. Manual white balancing has been applied for visualisation.

99 test images. The camera system is placed on a tripod with pan rotation unit as described in section 5.1 to limit camera motion to the three degrees of freedom of the system. The camera system is rotated around the object or translated in front of the object. The used viewpoints differ from one object to another. All manually estimated rotation angles are in the interval $[-51^\circ, 70^\circ]$. All manually estimated translations are in the interval $[-0.6\text{m}, 0.8\text{m}] \times [-0.1\text{m}, 0.4\text{m}]$. This is illustrated for database object 10 in fig. 5.17. For illumination, three neon lamps fixed to the ceiling and directed towards the floor and three tungsten halogen lamps placed on tripods and directed towards the object are used. For the tungsten halogen lamps, umbrellas can be added to obtain a more diffuse light. The different light sources are all placed at different positions in the room. To generate the test images, the number and type of turned on light sources is varied, with the constraint that only one illuminant type is used per image (i.e. no image is lighted simultaneously by neon and tungsten halogen lamps). This is illustrated for database object 10 in fig. 5.18. All test images are simple images showing one object in front of a non-cluttered background.

In addition, 100 test images showing an object not contained in the database are also used, in order to test the system response to a “false” object. The 100 images are obtained by varying illumination and viewpoints. Some of them are shown in fig. 5.19. This object has a simple geometry, which increases the chance that it is mistaken for another object or part of another object. These test images have been used to set thresholds a and mc_{lim} for accumulator interpretation in subsection 5.5.4. Both thresholds are set such that no object is recognised on “false” object images. As a result, for all test images, either the object is correctly recognised, or no object is recognised. This allows a better



Figure 5.18: Three different illuminations for the test images of object 10. No white balancing is applied.



Figure 5.19: Three different test images showing an object not contained in the database. No white balancing is applied.

comparability of the systems based on the different interest points.

5.7 Results

In this section, the evaluation results are presented for the recognition and localisation system combined with the interest point detectors selected in subsection 5.6.2. In subsection 5.7.1, the recognition and localisation quality is evaluated. Subsection 5.7.2 presents the suitability of the detectors for recognition and localisation. Finally, a conclusion is given in subsection 5.7.3.

5.7.1 Recognition and localisation quality

Recognition and localisation quality is evaluated with the first four criteria presented in subsection 5.6.1: recognition rate, pose accuracy, match consistency and deviation. The recognition rate gives the percentage of correctly recognised objects. Pose accuracy compares the pose computed by the system to a manual solution. The last two criteria evaluate the quality of the recognition and localisation process by estimating the proportion of false pose hypotheses and their deviation from the determined pose. First, the

5 Application to a recognition task

mean results for the whole database are presented. This is followed by a more detailed analysis where the results are given depending on the database objects.

Mean recognition and localisation performances for the whole database

The mean values of the four criteria over all objects and all test images are given in table 5.2. On the whole, the performance differences between the different systems are rather small, because robust algorithms are used for stereo reconstruction, matching, recognition and localisation and because the descriptors are invariant to illumination and viewpoint changes. Therefore, even if interest point detection is influenced by illumination or viewpoint changes, the obtained object and pose hypothesis is stable. All tested systems achieve recognition rates of about 90% and localisation accuracies of about 3 (this corresponds to an accuracy of a few degrees and a few centimetres in this application).

detector type	recognition rate	localisation accuracy	match consistency	deviation
HD	0.961	2.799	0.664	33.3
H-HD	0.919	3.06	0.658	33.1
AT-HD	0.859	4.99	0.709	29.6
H-HD+NBP	0.966	2.77	0.644	35.1
C-HD	0.936	2.76	0.687	32.7
HC-HD	0.966	2.59	0.680	32.0
2MS-HD+N	0.890	3.35	0.676	32.0
3MS-HD+N	0.955	2.34	0.689	32.1
3MS-HD+WB	0.977	2.82	0.684	32.0

Table 5.2: Recognition and localisation performances for the different detectors. The top of the table presents the results for the grey value detectors and the bottom presents the results for colour detectors. The mean value for all objects and for all test images is given. The best performance for each criterion is indicated in boldface.

Of all systems based on grey value detectors, H-HD+NBP and HD achieves the best overall performance. Their recognition rate and localisation accuracy are higher than H-HD and AT-HD, while match consistency and deviation are only slightly worse. The detection of a constant number of interest points has a positive influence on the robust recognition system, because it ensures enough matches and votes even on difficult test images. This effect is better visible in the analysis of recognition and localisation quality depending on the database objects. These higher recognition and localisation performances are obtained at the cost of more false hypotheses: both match consistency and deviation are worse for H-HD+NBP than for H-HD and both H-HD and AT-HD achieve better deviation than HD and H-HD+NBP. H-HD+NBP performs better than HD, which shows that adapting

interest point detection to the local lighting conditions improves recognition and localisation. The distances between interest points are bigger for AT–HD than for the other detectors (see section 3.4). As a result, AT–HD detects the smallest number of interest points in an image and it achieves the worst recognition and localisation performances. However, the higher distance between interest points improves the quality of the matches: AT–HD achieves the best match consistency and the lowest deviation of all systems. Less match pairs are discarded and the hypotheses are more accurate (see section 5.5).

Systems based on colour detectors achieve on the whole better performances than systems based on grey value detectors. They have similar or slightly better recognition rates and localisation accuracy, better match consistency and better deviation. This shows that the use of colour information reduces the proportion of false hypotheses. The 2MS–HD+N achieves the worst performance of all colour systems. Only two chrominance channels are used for detection and for descriptors, hence fewer interest points are detected and matched than with other detectors. This results, like for AT–HD and H–HD, into low recognition and localisation performances. Using all three chrominance channels improve well the results: both 3MS–HD+N and 3MS–HD+WB achieve the best overall performance of all tested systems. While 3MS–HD+N achieves the best localisation accuracy, 3MS–HD+WB obtains the best recognition rate. HC–HD also achieves very good performances. Although both H–HD and HC–HD have similar stability (see chapter 4), HC–HD reaches better recognition and localisation quality: it has better results than H–HD for all four criteria. On the contrary, the use of colour information for the standard Harris detector does not enhance much the results. C–HD even achieves a lower recognition rate than HD. This shows that the compensation of illumination influence is more important for colour images than for grey value images.

Recognition and localisation performances depending on the database objects

The recognition rate and the mean localisation accuracy are given for each database object in fig. 5.20. The performances may vary strongly from one object to another. Objects 4 and 7 are the easiest objects to recognise and localise because their 3D geometry is simple and their reflectance is structured (see fig. 5.16). All algorithms achieve very good recognition and localisation for these objects. The performance differences between the systems occur for “difficult” objects. Objects are difficult to recognise when many interest points have similar texture, such as objects 6, 8 and 9, and because of their complex 3D geometry (objects 1, 2, 6, 9). Object 6 is also difficult to handle with grey value detectors as the coloured texture has low intensity edges. For these “difficult” objects, HD and H–HD+NBP achieve the best performances of all grey value systems. Colour information helps handling such objects: 3MS–HD+WB, 3MS–HD+N and HC–HD achieve good performances. C–HD also performs well, except for object 6. Localisation accuracy is influenced by the 3D geometry of the object. All algorithms localise accurately objects with approximately planar surfaces, like objects 4, 5, 7, 8 and 10. Round or complex objects like objects 1, 2, 6 and 9 are more difficult to localise because perspective distortion and geometrically inaccurate interest point detection have a stronger influence on the pose

5 Application to a recognition task

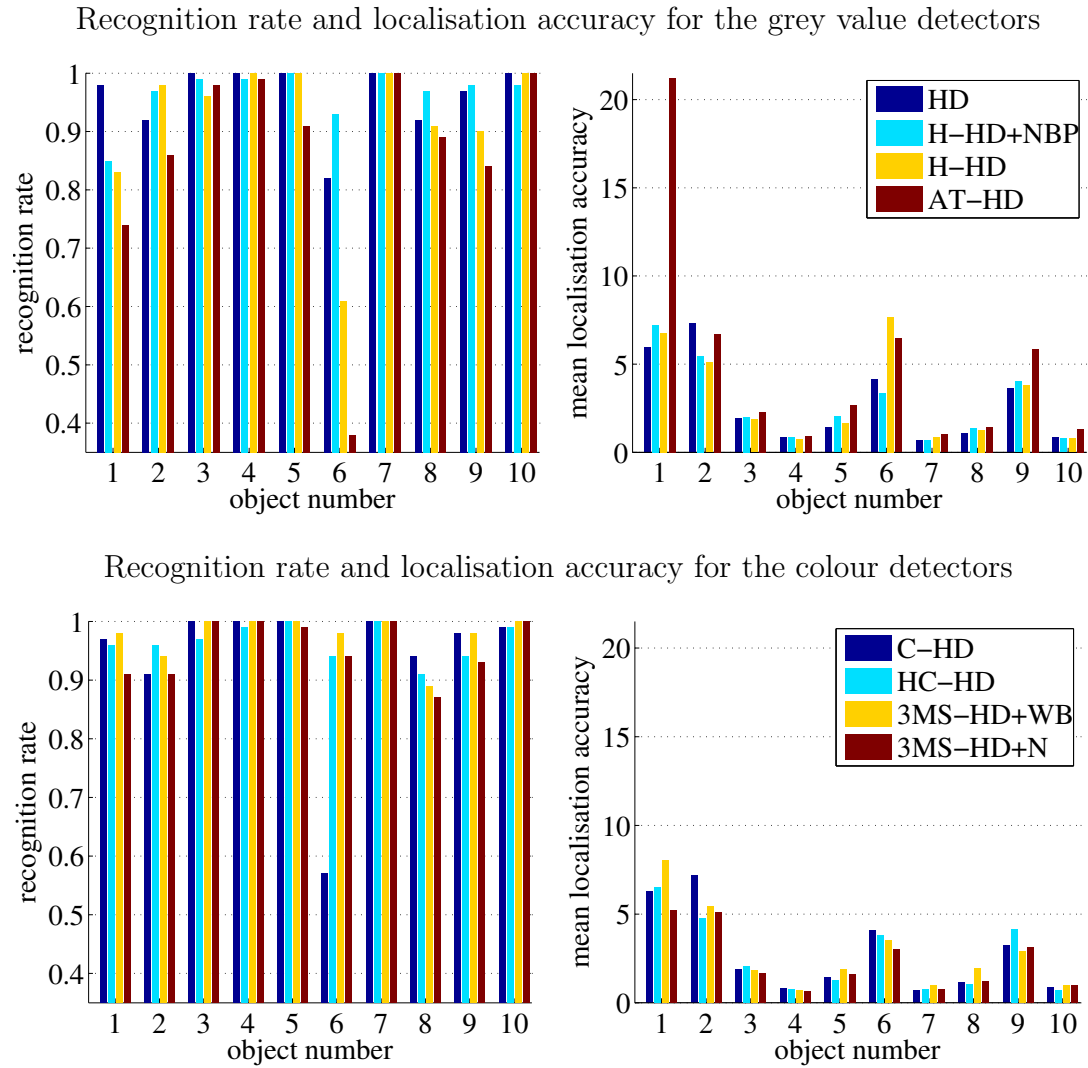


Figure 5.20: Recognition rate and localisation accuracy for the different detectors. For better legibility, the results of 2MS-HD+N are not included, as it achieves worse results than the other colour detectors. For each object, the mean value over all test images is shown. The objects are presented in fig. 5.16.

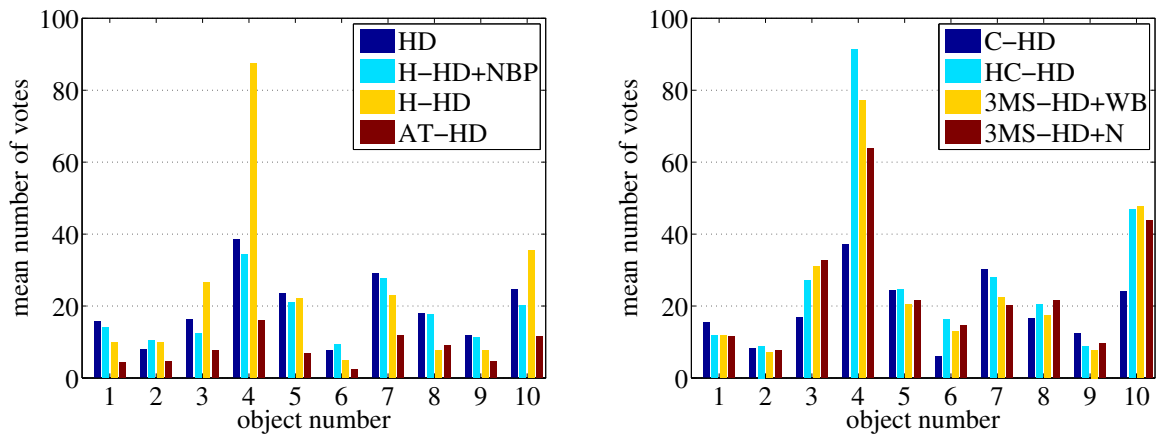


Figure 5.21: Number of votes for the best pose hypothesis for the different detectors. As in fig. 5.20, the results of the 2MS-HD+N are not included for better legibility. For each object, the mean value over all test images is shown. The objects are presented in fig. 5.16.

hypotheses. 3MS-HD+N and HC-HD achieve the best accuracy for these objects.

The correlation between the number of votes for the best pose hypothesis and recognition and localisation performance is illustrated in fig. 5.21. The objects for which the number of votes are low correspond to the objects for which recognition and localisation is difficult: objects 1, 2, 6, 8 and 9. Systems using colour detectors achieve a higher number of votes and also better recognition and localisation results than grey value systems. AT-HD has the smallest number of votes and the lowest recognition and localisation performances. This shows that a higher number of votes enhances the accuracy of the object and pose hypothesis. Only a sufficient number of votes is necessary: for object 4, H-HD and HC-HD have a much higher number of votes than the other algorithms but all algorithms achieve the same high performances. This explains why the detection of a fixed number of interest points like in HD and H-HD+NBP enhances recognition and localisation quality: it ensures a sufficient number of votes for all images.

Mean matching consistency and mean deviation are shown for each database object in fig. 5.22. Like in fig. 5.20, the performances vary a lot between objects. The two easiest objects are objects 4 and 7: the simple geometry and the structured reflectance result in few false pose hypotheses for all systems (high matching consistency and low deviation). Objects 3 and 6 are difficult objects for both matching consistency and deviation. Many false hypotheses are generated because geometry is relatively complex and reflectances are textured. Object 3 shows that good recognition and localisation results can be achieved when enough hypotheses are generated (see fig. 5.20). For objects 1 and 2, a high proportion of false hypotheses is generated (i.e. match consistency is low). Nonetheless, deviation is relatively good. This occurs because detection and matching are relatively stable but localisation is imprecise. On the contrary, objects 8 and 9 have good match consistency but relatively bad deviation. This is caused by a high proportion of false

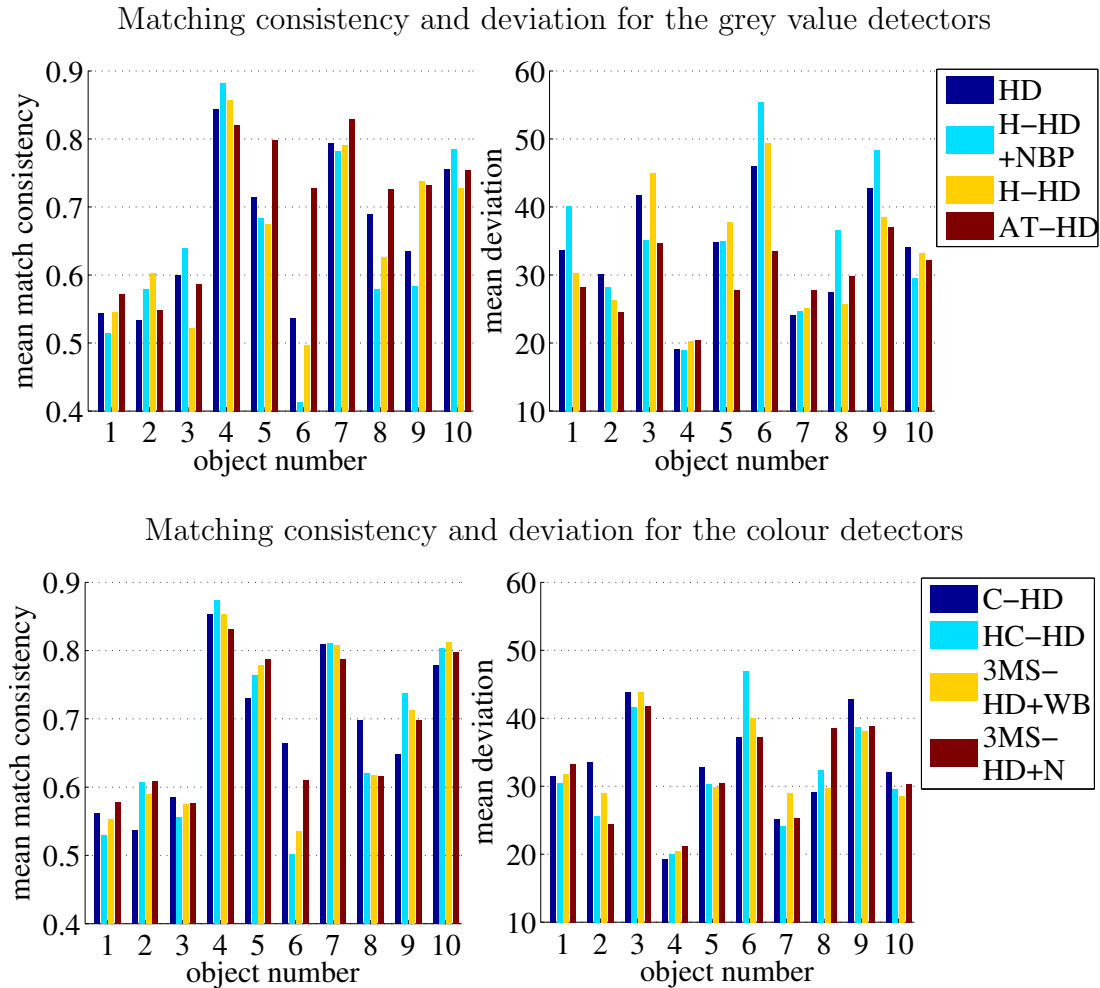


Figure 5.22: Matching consistency and deviation for the different detectors. As in fig. 5.20, the results of the 2MS-HD+N are not included for better legibility. For each object, the mean value over all test images is shown. The objects are presented in fig. 5.16.

matches. AT-HD and 3MS-HD+N achieve the best overall performances in fig. 5.22. Hence, they would be the most appropriate detectors for less robust recognition and localisation methods. Finally, H-HD performs better than H-HD+NBP for all difficult objects: the good recognition and localisation performances of H-HD+NBP are achieved at the cost of more false hypotheses which are well handled by the robust recognition and localisation algorithm.

5.7.2 Detector suitability for recognition and localisation

The suitability of the interest point detectors for recognition and localisation are evaluated using the last three criteria presented in subsection 5.6.1: percentage of characterised points, percentage of matched points and percentage of consistent points. The first criterion is related to the uniqueness of the interest point in its neighbourhood, hence to its information content. The second criterion is related to detection stability and to the discriminative power and the stability of the descriptors. Finally the last criterion estimates the percentage of interest points that contributed to the best pose hypothesis. The mean results for the whole database are presented in a first part. After that, the detector suitability is shown depending on the database objects.

Mean detector suitability for the whole database

The mean values of the three criteria over all test images and all database objects are given in table 5.3. The percentage of points that are passed from one system block to the next decreases continuously. Matching eliminates the highest proportion of points: only approximately 50% of the characterised points are matched. Due to the successive eliminations, the differences between the systems become smaller for the percentage of consistent points.

H-HD achieves the best performances of all tested systems. It has the highest percentages of characterised, matched and consistent points. Therefore, it is the most efficient detector for recognition and localisation: most of the detected, characterised and matched points contribute to recognition and localisation. The detection of a fixed number of interest points, on the contrary, results in a high proportion of “false” points (see the results of HD and H-HD+NBP). This occurs especially when the number of detected interest points is too high for the image content. More details about this are given in the analysis of detector suitability depending on the database objects.

Colour interest point detectors have on the whole a smaller percentage of characterised points. This is particularly visible for 3MS-HD+N. This is due to the stereo reconstruction system which uses only grey values. As a result, some colour interest points are not reconstructed because the intensity contrast in their neighbourhood is too small. This could be improved by considering the full colour signal during stereo reconstruction. The percentage of consistent points is similar for systems based on colour detectors and on grey

5 Application to a recognition task

detector type	percentage of characterised points	percentage of matched points	percentage of consistent points
HD	0.757	0.377	0.192
H-HD	0.837	0.449	0.218
AT-HD	0.806	0.355	0.183
H-HD+NBP	0.749	0.363	0.176
C-HD	0.756	0.366	0.190
HC-HD	0.802	0.410	0.210
2MS-HD+N	0.806	0.385	0.186
3MS-HD+N	0.768	0.385	0.195
3MS-HD+WB	0.822	0.397	0.203

Table 5.3: Detector suitability for recognition and localisation. The top of the table presents the results for the grey detectors and the bottom presents the results for colour detectors. The mean value for all objects and for all test images is given. For information, the mean number of detected interest points for all objects and all test images is approximately 100 for HD, H-HD, H-HD+NBP, C-HD and 2MS-HD+N, approximately 130 for HC-HD and 3MS-HD+N and 40 for AT-HD. The best performance for each criterion is indicated in boldface.

value detectors, even though less interest points are characterised (see the results of H-HD and HC-HD). This shows that colour information enhances matching, recognition and localisation (see also subsection 5.7.1). The use of a third chrominance channel in 3MS-HD+N also reduces the proportion of false matches in comparison to 2MS-HD+N, because the percentage of consistent points is higher for 3MS-HD+N while both systems have the same percentage of matched points. 3MS-HD+WB achieves a slightly better suitability than 3MS-HD+N because only few images contain very dark image areas and many images contain fine texture. HC-HD is the most suitable colour interest point detector for the current recognition and localisation system. The performance of chrominance based detectors (3MS-HD+N and 3MS-HD+WB) could however improve much if stereo reconstruction considered the full colour signal.

Detector suitability depending on the database objects

The mean values for all three criteria are shown in fig. 5.23 for each database object. The performances vary strongly between objects. The performance differences between the different systems are the smallest for the percentage of consistent points and the biggest for the percentage of characterised points. This shows that the robust recognition and localisation system discards false interest points and false matches successfully.

The percentage of characterised points is related to the information content of the detected interest points. The three detectors HD, H-HD+NBP and C-HD achieve all very similar

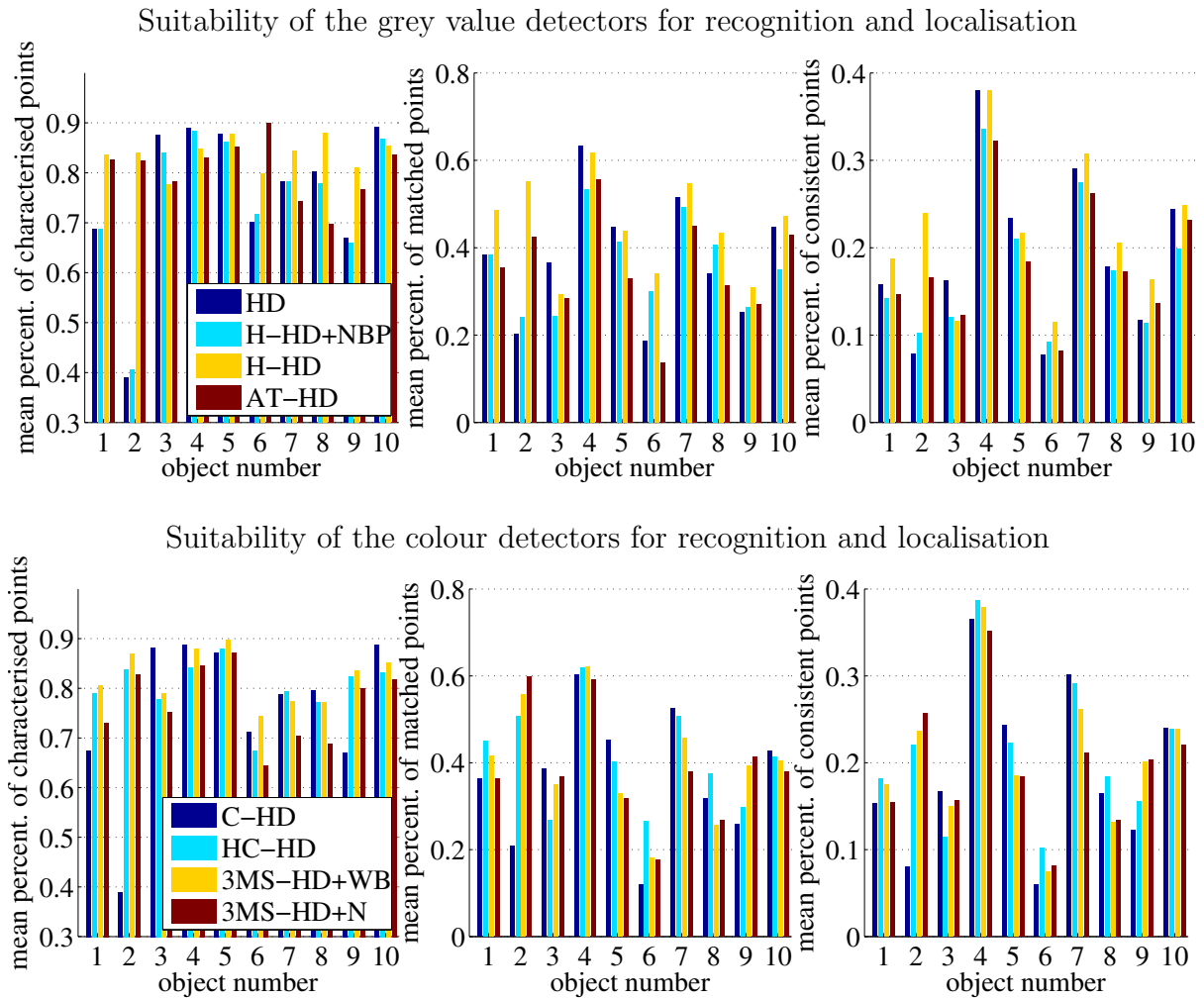


Figure 5.23: Detector suitability for recognition and localisation. As in fig. 5.20, the results of the 2MS-HD+N are not included for better legibility. For each object, the mean value over all test images is shown. The objects are presented in fig. 5.16.

performances which vary strongly between objects, because they detect a fixed number of interest points per image. As a result, many interest points with little information content are detected in images with little texture, for example for object 2. The performances of the other detectors vary less between objects, as the number of interest points is better adapted to image content. Objects 6, 7 and 8 contains many edges with high chrominance contrast but low intensity contrast (see fig. 5.16). This results in a low percentage of characterised points for the detectors based on colour, especially for 3MS-HD+N.

The performance differences between the systems are smaller for the percentage of matched points than for the percentage of characterised points. This shows that the interest points with little information content are successfully discarded by stereo reconstruction. H-HD achieves the best performance for almost all objects. All algorithms reach good results for objects 4 and 7, because these have simple 3D geometry and structured reflectance. Objects 3, 6, 8 and 9 are on the contrary more difficult to handle for all algorithms because of textured reflectance, complex 3D geometry or texture with low intensity contrast (see fig. 5.16). HD, H-HD+NBP and C-HD achieve low performance for object 2 because the number of detected interest points is too high for the image content.

The mean percentages of consistent points vary between objects like the mean percentages of matched points. This shows that only few false matches are obtained. As before, objects with simple 3D geometry and structured reflectances are easier to handle. H-HD, HC-HD, 3MS-HD+N and 3MS-HD+WB achieve the best performances for most objects. The compensation of illumination influence during interest point detection and the use of a fixed detection threshold makes therefore recognition and localisation more efficient. A high discrepancy between image content and the chosen number of detected interest point reduces strongly the suitability of HD, H-HD+NBP and C-HD for image 2. AT-HD achieves higher match consistency than the other systems for all objects but its percentage of consistent points is similar to the percentages of the other systems. Therefore, its good performances in match consistency and deviation are not due to higher stability of the interest points but only to the higher distances between these, which improves the quality of the generated pose hypotheses.

5.7.3 Conclusion

The developed recognition and localisation system is composed of several blocks, such as stereo reconstruction or matching, that use robust methods to discard unreliable and false interest points. The performance differences between all tested systems are small, which shows that false points are successfully eliminated. Performances may vary strongly from one object to another. Objects with simple 3D geometry and structured reflectance are well handled by all systems. Performance differences between systems occur for objects with complex 3D geometry or with textured reflectance (i.e. which contain many similar interest points). The used stereo algorithm is based on grey values. This has the drawback that valid colour interest points with low intensity contrast are discarded. In spite of that, systems based on colour detectors achieve better recognition and localisa-

tion quality. In particular, less false hypotheses are generated. The results show that a sufficient number of votes for the best object and pose hypothesis is necessary for reliable recognition and accurate localisation. As a consequence, the detection of a fixed number of interest points like in HD, H-HD+NBP and C-HD improves recognition and localisation at the cost of more false points and more false hypotheses, because it ensures a sufficient number of votes even on difficult images. False points and false hypotheses reduce the efficiency of the robust system but the quality of the results stays good. The compensation of local lighting conditions during interest point detection improves recognition and localisation performances. It also reduces the proportion of false object and pose hypotheses. Last, it increases the system efficiency: less false points are detected. The compensation of illumination influence is particularly important for systems using colour images. H-HD+NBP achieve the best recognition and localisation performances of all systems based on grey value detectors. The grey value detectors using a detection threshold, H-HD and AT-HD, do not detect enough interest points on difficult test images. This reduces their recognition and localisation performances. On the other hand, H-HD achieves the best efficiency (less false points are detected). AT-HD has the lowest proportion of false pose hypotheses, because the high distance between the interest points improves hypothesis accuracy. 3MS-HD+N, 3MS-HD+WB and HC-HD achieve the best overall performances of all systems based on colour detectors. 3MS-HD+N achieves the best accuracy and 3MS-HD+WB achieves the best recognition rate. Both generate less false hypotheses than HC-HD, but they require enough chrominance edges in the scene. HC-HD achieves very good overall results and is suitable for all scenes.

To increase the proportion of characterised colour interest points, the stereo algorithm should consider the full colour signal. This would improve the recognition and localisation performance of HC-HD, 3MS-HD+WB and 3MS-HD+N. A sufficient number of votes for the best object and pose hypothesis is necessary for reliable recognition and accurate localisation. A better adaptation of the interest point detection threshold to scene content would therefore enhance both system efficiency and recognition and localisation quality. Alternatively, the best interest points could be taken into account one after the other until an object and pose hypothesis of sufficient quality is obtained or until recognition fails. Weighting the votes depending on interest point quality could also enhance recognition and localisation quality. Interest point quality could be given for example by their corner-ness value. To reduce the number of false matches caused by texture with similar points, the uniqueness of interest points in the image could be estimated. Interest points with low uniqueness value could be discarded or the votes could be weighted by the uniqueness values. Finally the accuracy of interest point detection could be improved for example using subpixel interest point detection. This would improve localisation accuracy, especially for objects with complex 3D geometry.

5.8 Summary

In this chapter, a state of the art recognition system is developed and used to evaluate the influence of interest point detection on recognition. An object database is created with one image per object. The system can recognise these objects after limited viewpoint changes and after illumination changes. It also estimates the camera motion. After interest point detection, the 3D positions of the points are reconstructed with stereo reconstruction. Interest points for which stereo correspondence is not unique are discarded so that only reliable information is passed to the next system block. The SIFT descriptors are used to characterise the texture in the interest point neighbourhoods. They are invariant to local illumination changes and robust to perspective transformations. Hence they measure the similarity of two interest points even if viewpoint and illumination changes. The SIFT descriptors and the 3D positions are used to match the interest points in the current image to the interest points in the database. To reduce the number of false matches, the match list contains only one best match per interest point and database object. It is also constrained to be symmetric, i.e. the same list is obtained if the roles of current and database images are interchanged. Finally, the generalised Hough transform merges the information of the matches for recognition and localisation. It takes geometric uncertainties into account. One best pose hypothesis is determined for each object. If the quality of the best object and pose hypothesis is sufficient, the corresponding object is recognised. Otherwise, no object is recognised.

The performances of the recognition system is evaluated for different interest point detectors. Many test images are used. They show the database objects under varying illumination and from different viewpoints. Evaluation criteria measure recognition and localisation performance, the proportion of false pose hypotheses and the proportion of false interest points. All tested systems perform similarly good on objects with simple 3D geometry and structured reflectance. The performance differences occur for objects with complex 3D geometry or with textured reflectance. Systems based on colour information achieve better recognition and localisation results and generate a smaller proportion of false hypotheses. The compensation of illumination influence during interest point detection increases recognition and localisation quality and reduces the proportion of false interest points. This effect is particularly strong for colour images. A sufficient number of votes for the best object and pose hypothesis is necessary for reliable recognition and accurate localisation. As a result, the detection of a fixed number of interest points improves recognition and localisation with the robust system, at the cost of more false points and more false hypotheses. This is especially useful for grey value systems, as colour systems detect and match in general more interest points. Hence, the H-HD system with selection of the N best points performs best of all grey value systems. For colour systems, HC-HD 3MS-HD+WB and 3MS-HD+N perform best. Recognition and localisation could be improved with a better adaptation of the number of interest points to image content. The recognition system could be further enhanced with a stereo algorithm using colour information, with weighting of the votes by interest point quality and with subpixel interest point detection.

6 Conclusion

A summary of the work and of the results is given in section 6.1. Suggestions for further research are given in section 6.2.

6.1 Summary

This work aims at improving the ability of machine vision systems to deal with illumination changes. It is focused on recognition tasks, for which invariance to illumination changes is particularly important. The first step of most state of the art recognition systems is to reduce the amount of processed data by detecting interest points: small characteristic image patches. Most interest point detectors are sensitive to illumination changes. As a result, only some of the interest points are redetected when the scene is viewed from a similar viewpoint under different illumination. To handle this problem, the matching between the current interest points and the interest points in the system database is based on information which is invariant to viewpoint and illumination changes. In addition, robust recognition methods are used to reduce the influence of false matches on the results. A more stable interest point detection would improve recognition performances, even for robust systems. Therefore, in this work, several interest point detectors with increased stability under illumination changes are developed and evaluated in a realistic application. These new detectors do not require any manual white balancing or additional information to handle illumination changes.

The developed algorithms all enhance a very popular interest point detector: the Harris detector (HD). This algorithm is based on the grey value derivatives. It is particularly stable under viewpoint changes. The first step of this work is to model the influence of illumination on the Harris detector. This shows that the stability of the Harris detector under illumination changes can be enhanced by locally adapting detection to the light intensity. Four detectors are developed, that uses different principles to perform this local adaptation. The first detector is the energy normalised Harris detector (N-HD). It normalises the image derivatives with the local grey value energies before detection. The second detector is the homomorphic Harris detector (H-HD). It uses homomorphic processing to eliminate the influence of local lighting conditions on the derivatives. The third detector is based on locally adaptive thresholding (AT-HD): the detection threshold is adapted for each pixel using the local mean of the detector response. For AT-HD, interest point detection is only performed in the detected textured image areas, in order to reduce noise influence. The last detector locally adapts the detection threshold based

6 Conclusion

on a local clustering of the detector response with the ISODATA algorithm (LI-HD). The stability of these four detectors and of the Harris detector are evaluated on image series acquired under varying illumination. The four new detectors achieve better stability for scenes with complex 3D geometry and for complex illumination changes (i.e. when illuminant type or position varies). The best results are obtained by H-HD and AT-HD. H-HD is fast and the most stable, but AT-HD achieves a more uniform interest point distribution in the image.

Some elements of the illumination influence, such as light colour, shadow or shading, can be modelled more accurately with colour images than with grey value images. The colour Harris detector (C-HD) [Gou00] extends the principle of the Harris detector to colour images. This makes the use of colour information very interesting for illumination invariant interest point detection. An illumination invariant colour Harris detector is presented in [vdW05]: the robust invariant Harris detector (RI-HD). This detector is invariant to shadow and shading effects but illuminant colour must be corrected beforehand. In this work, two invariant colour interest point detectors are developed which automatically correct illuminant colour. The homomorphic colour Harris detector (HC-HD) extends the principle of H-HD to colour images: this locally eliminates the influence of light colour and light intensity. The second detector is the m space Harris detector (MS-HD). It is based on chrominances, so it can fully eliminate shadow and shading influence. In addition, illuminant colour is locally compensated. Chrominance is sensitive to noise and to colour artifacts introduced by image acquisition. Therefore, a preprocessing method based on the Nagao filter is introduced to reduce both noise and artifact influence on the MS-HD. In addition, a demosaicing method which reduces colour artifact formation is applied for image acquisition for all detectors. The stability of all colour detectors under illumination changes is evaluated like for the grey-value detectors. MS-HD is the most stable detector for scenes with clear chrominance edges, especially when 3D geometry is complex. It can be combined with the Nagao preprocessing to reduce noise sensitivity in dark image areas. It is however not suitable for scenes with few or no chrominance information. In that case, HC-HD and H-HD are the most stable.

Finally, a state of the art recognition system is developed in order to estimate the influence of interest point detection in an application. A database of 10 objects is created. The system recognises these objects after limited camera motion and after illumination changes. It also estimates the camera motion. After interest point detection, the 3D positions of the interest points are computed with stereo reconstruction. Only interest points with unique stereo correspondences are kept. This filters out interest points which are not characteristic enough. The remaining interest points are characterised with SIFT descriptors [Low04]. These invariant descriptors are used to match current interest points to the most similar database interest points. The number of false matches is further reduced by verifying that the 3D positions of the two points in a match are compatible with the system degrees of freedom. In addition, the match list is constrained to be symmetric: the same list is obtained if current image and database image are interchanged. Last, the generalised Hough transform is used to robustly estimate an object and camera motion hypothesis. The algorithm is enhanced by taking geometric uncertainties into

account. The final recognition result is based on the quality of the object and camera motion hypothesis. The new detectors improve recognition and localisation results for objects with complex 3D geometry or with textured reflectance (i.e. with many similar interest points). They reduce the number of false interest points, which increases the system efficiency. The best results are obtained with systems using colour information. HC–HD, MS–HD+WB and MS–HD+N perform best: MS–HD+N is the most accurate, MS–HD+WB has the highest recognition rate, HC–HD is very good and suitable for all kind of scenes. For grey value systems, the best recognition results are obtained by H–HD combined with the selection of the N best interest points and the smallest proportion of false interest points is reached by H–HD combined with a fixed detection threshold.

6.2 Further research

The developed interest point detectors improve detection stability, recognition and localisation. Illumination influence is however not completely eliminated. This work also showed further problems which would be interesting to solve. These problems are summarised in the following and solutions are suggested.

Shadow and shading effects with sharp edges cannot be compensated using a single grey value image. As a result, all developed grey value interest point detectors only handle slowly varying shadow and shading effects. Shadow and shading effects with sharp edges can be compensated with colour images, for example with MS–HD. This work shows that this full elimination of shadow and shading effects enhances detection stability, recognition and localisation but it is only suitable for scenes with good chrominance edges. It would therefore be interesting to compensate all shadow and shading effects for grey value detectors. Shading effects could be handled by taking into account the 3D geometry of the interest point neighbourhood. To handle both shadow and shading effects, several images of the scene under different illuminants could be used to learn an illumination model of the scene or to filter interest points caused by shadow or shading effects. Such methods are used for example in [HB98, Neu01].

None of the developed interest point detectors compensates specularities. The handling of saturated image areas (see section 3.6) reduces the influence of specular highlights but it does not solve the problem of specular reflections which do not result in saturated pixel values. Specular effects can be reduced or detected with optical polarising filters (see [LLK⁺02]), using two or more images of the scene as in [LLK⁺02] or using a single white balanced colour image as in [TI03]. If an accurate estimate of the light colour is available, colour features which are invariant to specularities can be used for interest point detection (see [GS99, vdW05]). Last, specular highlights can be corrected in a preprocessing step with image inpainting as in [BSCB00, TLQS03, TI03].

The interest point detectors based on chrominance information are sensitive to noise in dark image areas and to colour artifacts introduced by image acquisition. Colour artifacts could be reduced with better camera hardware. The X3 technology by Foveon Inc. (see

6 Conclusion

[Fov06]) is particularly interesting because three colour channels are acquired simultaneously for all pixels with a single CMOS chip. Alternatively, demosaicing algorithms could be improved. Constraining the edges of all colour channels to occur at the same image positions are very promising to reduce colour artifacts. Better preprocessing methods than the Nagao filter proposed in chapter 4 could be designed. In particular, better texture preservation would be desirable. Last, noise influence in dark areas could be reduced by discarding chrominance interest points detected near dark areas, like for saturated areas (see section 3.6).

Chrominance based interest point detection, for example with MS-HD, achieves the highest stability but it requires good chrominance edges in the scene. In addition, chrominance is more noisy than intensity information. The combination of chrominance and intensity information for interest point detection could therefore achieve good stability on all kind of scenes. MS-HD+N and H-HD achieved the best stability in this work. These detectors could be combined for example by adding their responses before interest point detection. Alternatively the best interest points of both detectors could be used for recognition.

Stability evaluations in chapters 3 and 4 and evaluation of recognition results in chapter 5 all showed the importance of adapting the number of interest points to scene content. A too low detection threshold decreases detector stability, system efficiency and localisation accuracy. If the detection threshold is too high, not enough interest points are detected for reliable recognition and localisation. Adapting the detection threshold to the scene is hence a promising further research topic. This could be realised by automatically selecting the threshold using for example the histogram of the detector response. Alternatively, the recognition algorithm could take the best interest points sequentially into account until a reliable object and pose hypothesis is obtained or until recognition fails. Last, the influence of the matches on the recognition system could be weighted by the interest point quality, based for example on the detector response or on their uniqueness. This would reduce the influence of false matches on localisation accuracy.

Finally, the influence of interest point detection could be evaluated in different or more complex applications. First, it would be interesting to know if a stereo algorithm using the full colour information enhances the performances of systems based on colour interest point detectors. The influence of the different detectors could also be tested in more complex recognition applications: for example for a bigger database, for images with cluttered background or with several database objects, and for objects with more complex 3D geometry or for more complex lighting changes. Finally, interest points are not only used for recognition, but also for further machine vision applications: for example tracking, 3D reconstruction, wide baseline stereo, registration and content-based image retrieval. It would therefore be interesting to evaluate if the new detectors improve the performances of other machine vision tasks.

Bibliography

- [Bal81] D. H. Ballard. Generalizing the Hough transform to detect arbitrary patterns. *Pattern Recognition*, 13(2):111–122, April 1981. 2, 131
- [Bar99] Kobus Barnard. *Practical colour constancy*. Ph.d. thesis, Simon Fraser University, School of Computing Science, 1999. 19, 23
- [Bau00] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000. 10, 12, 19, 20
- [Bay76] B. E. Bayer. Color imaging array. United States Patent 3,971,065, 1976. 63
- [BMCF02a] K. Barnard, L. Martin, A. Coath, and B. Funt. A comparison of computational color constancy algorithms – part I: Methodology and experiments with synthesized sata. *IEEE Trans. on Image Processing*, 11(9):972–983, September 2002. 23
- [BMCF02b] K. Barnard, L. Martin, A. Coath, and B. Funt. A comparison of computational color constancy algorithms – part II: Experiments with image data. *IEEE Trans. on Image Processing*, 11(9):985–996, September 2002. 22, 23, 79, 91, 92
- [BSCB00] Marcelo Bertalmío, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. Image inpainting. In *Proc. SIGGRAPH*, pages 417–424, 2000. 61, 106, 159
- [CBS03] G. De Cubber, S.A. Berrabah, and H. Sahli. A bayesian approach for color constancy based visual servoing. In *Proc. of the 11th Conference on Advanced Robotics (ICAR 2003)*, pages 983–990, 2003. 21, 22, 23
- [CJ02] Gustavo Carneiro and Allan D. Jepson. Local phase-based features. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 282–296, Copenhagen, Denmark, 2002. 12, 13, 48, 111
- [Cox87] D. R. Cox. Signal processing method and apparatus for producing interpolated chrominance values in a sampled colour image signal. United States Patent 4,774,565, 1987. 66
- [Der93] R. Deriche. Recursively implementing the gaussian and its derivatives. Technical Report RR-1893, INRIA, 1993. 14

Bibliography

- [DM02] A. J. Davison and D. W. Murray. Simultaneous localization and map-building using active vision. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 24(7):865–880, 2002. 12, 111
- [Duf01] Yves Dufournaud. *Navigation aérienne et guidage terminal à partir de données bidimensionnelles*. PhD thesis, Institut National Polytechnique de Grenoble, Jun 2001. 10, 12, 17, 28
- [DWL98] Mark S. Drew, Jie Wei, and Ze-Nian Li. Illumination-invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 533–540, 1998. 21, 22
- [ETM91] L. Eikvil, T. Taxt, and K. Moen. A fast adaptive method for binarization of document images. In *Proc. First International Conference on Document Analysis and Recognition*, pages 435–443, 1991. 40, 41
- [FA91] 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, September 1991. 111
- [FB81] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24:381–395, 1981. 2, 131
- [FBM98] Brian Funt, Kobus Barnard, and Lindsay Martin. Is machine colour constancy good enough? In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 445–459, 1998. 19, 23
- [FCF96] Graham D. Finlayson, Subho S. Chatterjee, and Brian V. Funt. Color angular indexing. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 16–27, 1996. 22
- [FDF94] G. D. Finlayson, M. S. Drew, and B. V. Funt. Spectral sharpening: Sensor transformation for improved color constancy. *Journal of the Optical Society of America A*, 11(5):1553–1563, May 1994. 8
- [FF95] Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995. 21
- [FHD02] Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images. In *Proc. 7th European Conference on Computer Vision (ECCV)*, pages 823–836, 2002. 7, 22
- [Fov06] Foveon Inc., 2006. www.foveon.com. 106, 160
- [Fre88] W. T. Freeman. Method and apparatus for reconstructing missing color samples. United States Patent 4,774,565, 1988. 66, 67

- [FS01] Graham D. Finlayson and Gerald Schaefer. Solving for colour constancy using a constrained dichromatic reflection model. *International Journal of Computer Vision*, 42(3):127–144, 2001. 23
- [FSC98] Graham D. Finlayson, Bernt Schiele, and James L. Crowley. Comprehensive colour image normalization. In *Proc. of the European Conference on Computer Vision (ECCV)*, 1998. 20, 21
- [GAM02] B. K. Gunturk, Y. Altunbasak, and R. M. Mersereau. Color plane interpolation using alternating projections. *IEEE Trans. on Image Processing*, 11(9):997–1013, 2002. 67, 73
- [GDvdB⁺99] Jan-Mark Geusebroek, Anuj Dev, Rein van den Boomgaard, Arnold W. M. Smeulders, Frans Cornelissen, and Hugo Geerts. Color invariant edge detection. In *Scale-Space Theories in Computer Vision*, pages 459–464, 1999. 22
- [GMD⁺97] P. Gros, G. Mclean, R. Delon, R. Mohr, C. Schmid, and G. Mistler. Utilisation de la couleur pour l'appariement et l'indexation d'images. Technical Report 3269, INRIA, September 1997. 10, 20, 23, 27, 48
- [Gou00] Valérie Gouet. *Mise en correspondance d'images en couleur - Application à la synthèse de vues intermédiaires*. PhD thesis, Université de Montpellier II, France, 2000. 10, 12, 14, 19, 20, 48, 76, 95, 158
- [GS99] T. Gevers and A. W. M. Smeulders. Color-based object recognition. *Pattern Recognition*, 32(3):453–464, March 1999. 9, 19, 21, 22, 23, 24, 71, 85, 159
- [GS04] Theo Gevers and Harro M. G. Stokman. Robust histogram construction from color invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):113–117, 2004. 22
- [GSS98] T. Gevers, A. W. M. Smeulders, and H. Stokman. Photometric invariant region detection. In *Proc. of the British Machine Vision Conference (BMVC)*, 1998. 22
- [GvdBSG01] Jan-Mark Geusebroek, Rein van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001. 22
- [GW92] Rafael C. Gonzales and Richard E. Woods. *Digital Image Processing*. Addison-Wesley, 1992. 17
- [HA97] J. Hamilton and J. Adams. Adaptive color plane interpolation in single sensor color electronic camera. US Patent 5,629,734, 1997. 66, 67
- [HB98] Gregory D. Hager and Peter N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998. 18, 159

Bibliography

- [HLS02] Daniela Hall, Bastian Leibe, and Bernt Schiele. Saliency of interest points under scale changes. In *Proc. of the British Machine Vision Conference (BMVC)*, Cardiff, Wales, 2002. 12
- [Hou62] P. V. C. Hough. Method and means for recognizing complex patterns. United States Patent 3,069,654, 1962. 2, 131
- [HP03] K. Hirakawa and T. W. Parks. Adaptive homogeneity-directed demosaicing algorithm. In *ICIP03*, pages III: 669–672, 2003. 66, 67
- [HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the 4th Alvey Vision Conference*, 1988. 10, 11, 12, 13
- [HS97] G. Healey and D. Slater. Computing illumination-invariant descriptors of spatially filtered color image regions. *IEEE Trans. on Image Processing*, 6(7):1002–1013, July 1997. 9, 20
- [Hub04] Tobias Huber. Korrespondenzverfolgung in Stereobildfolgen zur schritt haltenden Eigenbewegungs- und Szenenrekonstruktion. TU München, Lehrstuhl für Realzeit-Computersysteme, Diplomarbeit, January 2004. 117, 118
- [KA01] Andreas Koschan and Mongi Abidi. A comparison of median filter techniques for noise removal in color images. In *Proc. 7th German Workshop on Color Image Processing*, pages 69–79, 2001. 92
- [KB01] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001. 11, 17
- [KI86] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986. 42
- [Kim99] Ron Kimmel. Demosaicing: Image reconstruction from color CCD samples. *IEEE Trans. on Image Processing*, 8(9):1221–1228, 1999. 66, 67, 69, 73
- [Kov03] P. Kovesi. Phase congruency detects corners and edges. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 309–318, 2003. 11, 47
- [KSOK00] Markus Knappek, Ricardo Swain-Oropeza, and David Kriegman. Selecting promising landmarks. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, San Francisco, USA, 2000. 12
- [Lag98] R. Laganière. Morphological corner detection. In *Proc. International Conference in Computer Vision*, pages 280–285, 1998. 11
- [Lan86] E. Land. Recent advances in retinex theory. *Vision Research*, 26(1):7–21, 1986. 23
- [LBS90] H. C. Lee, E. J. Breneman, and C. P. Schulte. Modeling light reflection

- for computer vision color vision. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 12:402–409, 1990. 6
- [LE04] R. Laganière and R. Elias. The detection of junction features in images. In *Proc. International Conference on Acoustic, Speech and Signal Processing*, pages 573–576, 2004. 10
- [Lin98] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998. 10, 11
- [LLK⁺02] Stephen Lin, Yuanzhen Li, Sing Bing Kang, Xin Tong, and Heung-Yeung Shum. Diffuse-specular separation and depth recovery from image sequences. In *Proc. of the 7th European Conference on Computer Vision (ECCV)*, pages 210–224, 2002. 61, 106, 159
- [LM99] Salvatore Livatino and Claus B. Madsen. Autonomous robot navigation with automatic learning of visual landmarks. In *Proc. of the 7th International Symposium on Intelligent Robotic System (SIRS'99)*, pages 501–506, 1999. 11
- [Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. 10, 11, 17, 28, 110, 111, 112, 113, 114, 123, 131, 158
- [LPF04] Vincent Lepetit, Julien Pilet, and Pascal Fua. Point matching as a classification problem for fast and robust object recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '04)*, 2004. 111
- [LT03] W. Lu and Y. P. Tan. Color filter array demosaicking: New method and performance measures. *IEEE Trans. on Image Processing*, 12(10):1194–1210, October 2003. 50, 64, 65, 66, 67, 69, 71, 96, 107
- [LWM02] Zhenyong Lin, Junxian Wang, and Kai-Kuang Ma. Using eigencolor normalization for illumination-invariant color object recognition. *Pattern Recognition*, 35(11):2629–2642, 2002. 20
- [LYHT02] Jianguang Lou, Hao Yang, Weiming Hu, and Tieniu Tan. Illumination invariant change detection for visual surveillance. In *Proc. of the Asian Conference on Computer Vision (ACCV)*, 2002. 21
- [LZ03] G. Loy and A. Zelinsky. Fast radial symmetry for detecting points of interest. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 25(8):959–973, 2003. 10, 11
- [MB01] Nicholas Molton and Michael Brady. Practical structure and motion from stereo when motion is unconstrained. *International Journal of Computer Vision*, 39(1):5–23, 2001. 12, 17
- [MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo

Bibliography

- from maximally stable extremal regions. In *Proc. of the British Machine Vision Conference (BMVC)*, pages 384–393, 2002. 11, 12, 16, 111
- [MGDP00] P. Montesinos, V. Gouet, R. Deriche, and D. Pelé. Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, 18(9):659–672, June 2000. 111
- [MKK00] J. Matas, D. Koubaroulis, and J. Kittler. Colour image retrieval and object recognition using the multimodal neighbourhood signature. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 48–64, June 2000. 20, 21, 22, 85, 92
- [MM01] K. Mohanna and F. Mokhtarian. Performance evaluation of corner detection algorithms under similarity and affine transforms. In *Proc. of the British Machine Vision Conference (BMVC)*, pages 353–362, 2001. 48
- [MMG99] Florica Mindru, Theo Moons, and Luc J. Van Gool. Recognizing color patterns irrespective of viewpoint and illumination. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pages 1368–1373, 1999. 20, 111
- [MMG02] Florica Mindru, Theo Moons, and Luc J. Van Gool. Comparing intensity transformations and their invariants in the context of color pattern recognition. In *Proc. of the 7th European Conference on Computer Vision*, pages 448–460, 2002. 9, 10
- [Mor79] H. P. Moravec. Visual mapping by a robot rover. In *Proc. of the 6th International Joint Conference on Artificial Intelligence*, 1979. 12
- [MS98] Farzin Mokhtarian and Riku Suomela. Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1376–1381, 1998. 10
- [MS04] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 10, 12, 17, 28, 29, 33, 48, 111
- [MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, October 2005. 112
- [MTS⁺05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schafalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, November 2005. 3, 10, 12, 17, 48, 49
- [NB96] Shree K. Nayar and Ruud M. Bolle. Reflectance based object recognition. *International Journal of Computer Vision*, 17(3):219–240, 1996. 20, 22
- [NG98] Kenji Nagao and W. Eric L. Grimson. Using photometric invariants for 3D

- object recognition. *Computer Vision and Image Understanding*, 71(1):74–93, July 1998. 19, 21, 22, 85
- [Nib86] W. Niblack. *An Introduction to Digital Image Processing*, pages 115–116. Prentice Hall, 1986. 34, 35
- [NM79] M. Nagao and T. Matsuyama. Edge preserving smoothing. *Computer Graphics and Image Processing*, 9(4):394–407, April 1979. 92
- [OI96] K. Ohba and K. Ikeuchi. Recognition of the multi specularity objects using the eigen-window. In *Proc. of the 13th IEEE International Conference on Pattern Recognition*, pages I: 692–696, 1996. 12
- [Ols97] C. F. Olson. Efficient pose clustering using a randomized algorithm. *International Journal of Computer Vision*, 23(2):131–147, 1997. 131
- [Ols01] C. F. Olson. A general method for geometric feature matching and model extraction. *International Journal of Computer Vision*, 45(1):39–55, 2001. 131
- [OMC03] S. Obdrzalek, J. Matas, and O. Chum. On the interaction between object recognition and colour constancy. In *Proc. of the IEEE International Workshop on Color and Photometric Methods in Computer Vision (CPMCV'03)*, 2003. 20, 24
- [Ots79] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979. 41
- [OW04] I. Omer and M. Werman. Using natural image properties as demosaicing hints. In *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pages 1665–1670, 2004. 66, 67
- [Par93] J. R. Parker. *Practical Computer Vision Using C*. John Wiley & Sons, 1993. 44
- [RC78] T. W. Ridler and S. Calvard. Picture thresholding using an iterative selection method. *IEEE Trans. on Systems, Man and Cybernetics*, 8(8):630–632, 1978. 42
- [RGM03] Alessandro Rizzi, Carlo Gatta, and Daniele Marini. A new algorithm for unsupervised global and local color correction. *Pattern Recognition Letters*, 24(11):1663–1677, 2003. 23
- [Ris01] Valéry Risson. *Application de la Morphologie Mathématique à l'Analyse des Conditions d'Eclairage des Images Couleurs*. PhD thesis, Ecole des Mines de Paris, 2001. 7, 22
- [RJW02] Zia-ur Rahman, Daniel J. Jobson, and Glenn A. Woodell. Retinex processing for automatic image enhancement. In *Human Vision and Electronic Imaging, SPIE Symposium on Electronic Imaging, Proc. SPIE 4662*, 2002. 23

Bibliography

- [RS03] R. Ramanath and W. E. Snyder. Adaptive demosaicking. *Journal of Electronic Imaging*, 12(4):633–642, 2003. 66
- [RSBS02] R. Ramanath, W. E. Snyder, G. L. Bilbro, and W. A. Sander. Demosaicking methods for bayer color arrays. *Journal of Electronic Imaging*, 11(3):306–315, 2002. 64, 66, 67, 71
- [RT01] M. A. Ruzon and C. Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 23(11):1281–1295, 2001. 11, 20
- [SB97] S. M. Smith and J. M. Brady. SUSAN – a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, May 1997. 11, 47
- [Sch97] Bernt Schiele. *Object Recognition Using Multidimensional Receptive Field Histograms*. PhD thesis, Institut National Polytechnique de Grenoble, 1997. 17, 27, 28
- [SD99] Robert Sim and Gregory Dudek. Learning visual landmarks for pose estimation. In *International Conference on Robotics and Automation*, Detroit, MI, May 1999. 111
- [SH97] D. Slater and G. Healey. The illumination-invariant matching of deterministic local structure in color images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(10):1146–1151, October 1997. 9, 20
- [Sha85] S. A. Shafer. Using color to separate reflection components. *Color Research Applications*, 10(4):210–218, 1985. 6
- [Siv96] D. S. Sivia. *Data Analysis: A Bayesian Tutorial*. Clarendon Press, Oxford, UK, 1996. 127
- [SLL01a] S. Se, D. Lowe, and J. Little. Local and global localization for mobile robots using visual landmarks. In *Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 414–420, Maui, Hawaii, oct 2001. 131
- [SLL02] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 226–231, Lausanne, Switzerland, 2002. 131
- [SM97] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997. 12, 111, 112
- [SMB00] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, June 2000. 3, 10, 12, 17, 47

- [Smi99] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing – Second Edition*. California Technical Publishing, 1999. 17
- [SRI02] SRI International. *SRI Small Vision System – User's Manual*, 2002. 117, 122
- [SRI03] SRI International. *SRI Small Vision System – Calibration Supplement to the User's Manual*, 2003. 117, 122
- [SS04] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004. 18
- [ST94] Jianbo Shi and Carlo Tomasi. Good features to track. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994. 12, 13
- [STL⁺03] N. Sebe, Q. Tian, E. Louprias, M .S. Lew, and T. S. Huang. Evaluation of salient point techniques. *Image and Vision Computing*, 21(13-14):1087–1095, 2003. 11, 48
- [Sto00] Harro Stokman. *Robust Photometric Invariance in Machine Color Vision*. PhD thesis, University of Amsterdam, 2000. 22
- [Stö01] Norbert Stöfler. *Realzeitfähige Bestimmung und Interpretation des optischen Flusses zur Navigation mit einem mobilen Roboter*. PhD thesis, Technische Universität München, 2001. 117, 118, 120
- [TA02] B. Telle and M.-J. Aldon. Interest points detection in color images. In *Proc. of the IAPR Workshop on Machine Vision Application (MVA)*, pages 550–555, 2002. 92
- [TAM00] D. Toth, T. Aach, and V. Metzler. Illumination-invariant change detection. In *Proc. of the 4th IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 3–7, 2000. 18
- [Tec01] Basler Vision Technologies. *BASLER A302f Camera Manual*, 2001. 6, 14, 50
- [TI03] Robby T. Tan and Katsushi Ikeuchi. Separating reflection components of textured surfaces using a single image. In *Proc. of the 9th IEEE International Conference on Computer Vision (ICCV)*, pages 870–877, 2003. 61, 106, 159
- [TJ95] Øivind D. Trier and Anil K. Jain. Goal-directed evaluation of binarization methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1191–1201, 1995. 18, 34, 40
- [TLQS03] Ping Tan, Stephen Lin, Long Quan, and Heung-Yeung Shum. Highlight removal by illumination-constrained inpainting. In *Proc. of the 9th IEEE In-*

Bibliography

- ternational Conference on Computer Vision (ICCV)*, pages 164–169, 2003. 61, 106, 159
- [TM98] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. of the 1998 IEE International Conference on Computer Vision*, 1998. 92
- [TNI03] Robby T. Tan, Ko Nishino, and Katsushi Ikeuchi. Illumination chromaticity estimation using inverse-intensity chromaticity space. In *Proc. of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, pages 673–682, 2003. 23
- [TT04] Eduardo Todt and Carme Torras. Detecting salient cues through illumination-invariant color ratios. *Robotics and Autonomous Systems*, 48(2-3):111–130, 2004. 11, 19, 85
- [Tuy00] T. Tuytelaars. *Local, Invariant Features for Registration and Recognition*. PhD thesis, Katholieke Universiteit Leuven, 2000. 11, 12, 16, 17, 19, 111, 131
- [TV98] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998. 116, 118, 120, 121, 122, 127
- [TY96] Wen-Hsiang Tsai and Chen-Kuei Yang. Reduction of color space dimensionality by moment-preserving thresholding and its application for edge detection in color images. *Pattern Recognition Letters*, 17(5):481–490, 1996. 14
- [vdW05] J. van de Weijer. *Color Features and Local Structure in Images*. PhD thesis, University of Amsterdam, March 2005. 10, 12, 19, 22, 23, 24, 47, 63, 78, 79, 80, 95, 106, 158, 159
- [vdWGG05] Joost van de Weijer, Theo Gevers, and Jan-Mark Geusebroek. Edge and corner detection by photometric quasi-invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):625–630, April 2005. 19, 22
- [VL01] Etienne Vincent and Robert Laganière. Matching feature points in stereo pairs: A comparative study of some matching strategies. *Machine Graphics & Vision*, 10(3):237–259, 2001. 12, 13, 17, 48, 111, 123, 128, 129, 131
- [WB01] C. Wallraven and H. Bülthoff. View-based recognition under illumination changes using local features. In *IEEE Conference on Computer Vision and Pattern Recognition - Workshop on Identifying Objects Across Variations in Lighting: Psychophysics and Computation*, 2001. 12, 17
- [WL97] R. P. Würtz and T. Lourens. Corner detection in color images by multiscale combination of end-stopped cortical cells. In *Artificial Neural Networks – ICANN’97*, pages 901–906, 1997. 11

- [XE01] Ming Xu and Tim Ellis. Illumination-invariant motion detection using colour mixture models. In *Proc. of the British Machine Vision Conference (BMVC)*, pages 163–172, September 2001. 7, 21, 22, 24
- [ZDFL95] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78:87–119, October 1995. 12, 17

Own publications and supervised theses

- [Das02] Ingo Dasch. Histogrammbasierte Lokalisierung in der mobilen Robotik. TU München, Lehrstuhl für Realzeit-Computersysteme, Diplomarbeit, October 2002. 2
- [Fai03a] Flore Faille. Adapting interest point detection to illumination conditions. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 499–508, 2003. 17, 49, 55
- [Fai03b] Flore Faille. Detektion visueller Merkmale für die Lokalisierung eines mobilen Roboters: Anpassung an die Beleuchtung. In *Autonome Mobile Systeme, Informatik Aktuell*, pages 53–63, 2003. 55
- [Fai04a] Flore Faille. Comparison of demosaicking methods for color information extraction. In *Proc. of the Int. Conf. on Computer Vision and Graphics (ICCVG)*, September 2004. 66
- [Fai04b] Flore Faille. A fast method to improve the stability of interest point detection under illumination changes. In *Proc. of the International Conference on Image Processing (ICIP)*, 2004. 55
- [Fai05a] Flore Faille. *Comparison of Demosaicking Methods for Colour Information Extraction*, chapter II.3, pages 105–114. Pro Literatur Verlag, July 2005. In Cutting Edge Robotics. 66
- [Fai05b] Flore Faille. Stable interest point detection under illumination changes using colour invariants. In *Proc. of the British Machine Vision Conference (BMVC 2005)*, September 2005. 100
- [Gon03] Zhiqiang Gong. Stereo vision based 3D-reconstruction of visible landmarks. TU München, Lehrstuhl für Realzeit-Computersysteme, Master’s thesis, August 2003. 121
- [Had04] Sami Haddadin. Farbinterpolation in der Bildverarbeitung. TU München, Lehrstuhl für Realzeit-Computersysteme, Bachelorarbeit, April 2004. 64

Bibliography

- [Kiß03] Christian Kißling. Entwicklung und Auswertung eines Farb–Harris Corner Detectors. TU München, Lehrstuhl für Realzeit-Computersysteme, Bachelorarbeit, July 2003. 14
- [Kra05] Wolfgang Krank. Evaluierung zweier Matching–Verfahren hinsichtlich Beleuchtungs– und Bewegungs–Invarianz. TU München, Lehrstuhl für Realzeit-Computersysteme, Diplomarbeit, February 2005. 112
- [Lil03] Hu Lili. Repeatability and accuracy of Harris corner detector. TU München, Lehrstuhl für Realzeit-Computersysteme, Master’s thesis, August 2003. 12
- [Mar02] Florian Martin. Illumination independent detection and recognition of points of interest. TU München, Lehrstuhl für Realzeit-Computersysteme, Diplomarbeit, November 2002. 88
- [Neu01] Stephan Neumaier. Realisierung eines appearance–based Objekterkennungsmoduls für den autonomen mobilen Roboter MARVIN. TU München, Lehrstuhl für Realzeit-Computersysteme, Diplomarbeit, May 2001. 18, 123, 159
- [Web05] Thomas Weber. Stereo–basierte Lokalisierung eines mobilen Roboters. TU München, Lehrstuhl für Realzeit-Computersysteme, Interdisziplinäres Projekt, March 2005. 126, 131, 136