Quality of Variant and Invariant Features

Gertjan J. Burghouts

Printing: TNO, The Hague, The Netherlands.

The cover has been designed with the kind and patient help of my grandfather Gerrit Conijn. The depicted images are old postcards of Edam, printed in 1907 (courtesy of respectively Sipkema publishers in Edam and Trenkler publishers in Leipzig). The postcards are drawings that were based on black and white photographs; colors have been added by the publishers manually.
Interestingly, the two postcards both depict the Speeltoren in Edam, yet its appearance is very different in the two depictions. Firstly, the postcards are from a different period (note the change in neighboring houses). Secondly, the viewpoint and the color of the sky is very different. The latter two aspects have a distinctive effect on the highlighted elliptical image regions, yet humans are very well able to identify the two regions as belonging to the Speeltoren, most probably even when shown in isolation. One of the topics of this thesis is to assess for a range of existing or proposed image measurements (referred to as features) how descriptive they are of objects such as the Speeltoren, independent of the conditions at hand such as viewpoint and illumination.

ISBN 978-90-5986-262-3

Quality of Variant and Invariant Features

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus prof. dr D. C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 16 november 2007, te 12:00 uur

door

Gerardus Johannes Burghouts

geboren te Purmerend

Promotiecommissie:

Promotor:	Prof. dr ir A. W. M. Smeulders
Co-promotor:	dr J. M. Geusebroek
Overige leden:	Prof. dr ir L. J. van Vliet Prof. dr ir F. C. A. Groen Prof. dr Th. Gevers dr ir R. P. W. Duin dr F. W. Cornelissen

Faculteit: Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Advanced School for Computing and Imaging

The work described in this thesis has been carried out within the graduate school ASCI, at the Intelligent Systems Lab Amsterdam of the University of Amsterdam. ASCI dissertation series number 157.

UNIVERSITEIT VAN AMSTERDAM

Observations always involve theory.

Edwin Hubble

Contents

Chapter 1

Introduction

1.1 From Photographs to Features

In the past, photographers needed to make many preparations before they could take a good photograph. They had to take into account the time of day, season of the year, the cloudiness, and the type of scene, before putting in the right type of film, and setting the right aperture and exposure time. In other words, photographers were dependent on the material quality, and on the *uncontrollable* environmental conditions. Over the years, cameras have improved, and nowadays even commodity cameras are able to adjust automatically their sensitivity to suit the conditions at hand. To take a good photograph, one has become less and less dependent on the accidental conditions.

Camera hardware has become widely available. Photographs are taken with a wide variety in scene content and in conditions in which the photographs are taken, as illustrated in Figure 1.1. If we look more closely, the objects designated in the scene are recorded under various settings of the illumination and under varying camera viewpoints. Object appearance is heavily affected by such settings. For instance,

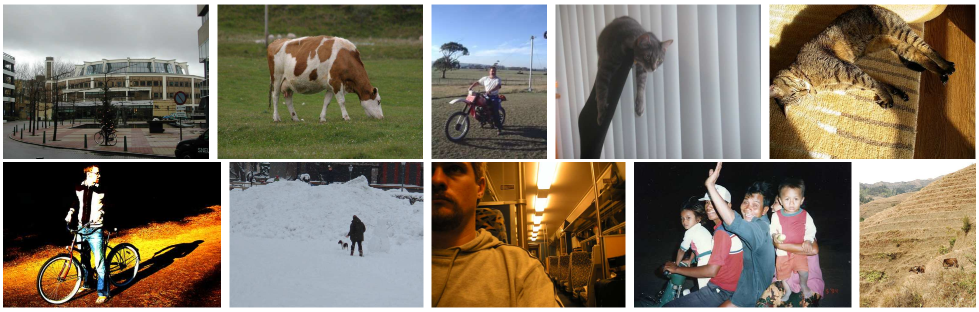


Figure 1.1: Examples of scene photographs.

an object that is in the picture, changes appearance as it is recorded in the afternoon and at dusk. At dusk, the light from the sky turns reddish, which affects the observed color of the object. Even for pictures taken right after each other, the object appearance deviates when the pose of the object has changed. Examples of appearance variation of a single object are depicted in Figure 1.2. The figure illustrates the dominating appearance variations: changes of the illumination direction, of the illumination intensity and color, and a change of the camera viewpoint.

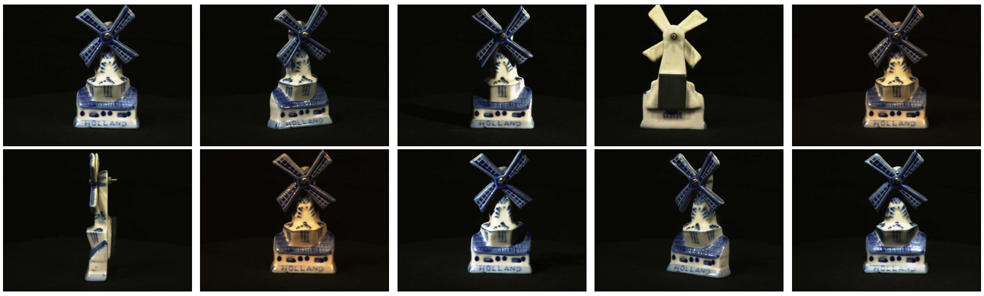


Figure 1.2: Example object with various appearances. Before an object can be recognized from any arbitrary angle and under any arbitrary illumination, these accidental conditions need to be removed without knowing a priori what the parameters of viewpoint and illumination are.

Nowadays, often the photographs are taken by digital cameras, or photographs are digitized afterwards as, for instance, to preserve cultural heritage. Digital photographs, or images for short, have become ubiquitous. This has led to a demand for (semi-) automated analysis of images, for instance to retrieve them from a large collection, or to indicate whether a particular object is present in the given image. Automated analysis carried out by vision systems is expected to give similar results independent of the varying scene conditions under which the objects of interest are recorded. This implies the need for measurements that are robust, in the ideal case even invariant, to appearance changes. Humans are able to distinguish between objects very well in a large variety of circumstances. We are able to observe differences between two appearances of the same object, i.e. we never see the object without environmental determinants. Yet, we are able to classify or identify effortlessly the object. In analogy, for artificial vision systems, the challenge is to abstract from the distorted input such that different appearances of the same object are taken as one and the same.

The image measurements are a first step to determine the similarity between images or objects designated therein. The more invariant a measurement, the more robustness is gained as unimportant details are discarded, but also more information is lost. With a more invariant measurement, objects become more similar. The choice of measurement affects the level of detail retained hence the similarity. Therefore, in order to measure similarity adequately, the level of invariance needs to be balanced with discriminative power. Here, besides the scene variation that is anticipated, also task-specific requirements need to be taken into account. For instance, object shading is an irrelevant property if one aims to order objects according to their color, but it is important if one aims to determine object shape. Thus, to determine similarity, an important choice is to select the measurements that take into account only the properties that matter to the vision system. For various computer vision tasks, the notion of similarity has proven to be very crucial. For example, in image segmentation the objective is to separate homogeneous regions, where the homogeneity may be defined by color independent of shading and shadow effects. Likewise, to recognize an object recorded under various illumination directions, illumination colors and camera viewpoints, similarity should be determined independently of appearance effects. If one knows the accidental conditions under which objects are viewed, see Figure 1.3, then one can design classes of image measurements that are variant or invariant to these accidental conditions. These examples clarify that the choice of image measurements is important to subsequently determine the similarity between objects.

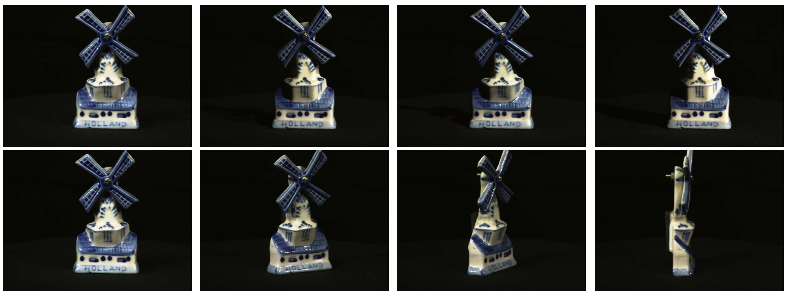


Figure 1.3: When the accidental conditions are ordered with respect to their driving force, then one can design classes of image measurements that are variant or invariant to these accidental conditions. To order variants and invariants and to measure their quality of measuring object properties is the goal of the thesis.

This thesis addresses the quality of image measurements, or features. Quality will be established by determining a feature's invariance and its retained discriminative power for a large dataset of real-world objects and realistic scene variation, and investigate the notion of similarity between objects in that dataset.

In computer vision, a fundamental choice is to take the picture as a starting point, or to take the objects in the real world represented in the picture as a starting point. A large part of image processing methods is intensity specific in that it deals with the intensities as they appear in the image. This is the preferred strategy for general-purpose compression and communication. For other image processing tasks as image segmentation, interpretation, and object-identification the heart of the matter is in the properties of the objects represented by the picture. The objects are not seen as a pictorial depiction of the scene but rather as a pictorial representation of the scene [112]. As laid down above, in object-specific image processing, it is important first to remove the accidental conditions introduced into the image at the moment of recording. For object-specific image processing, invariant features have to be considered. In this thesis, we consider a set of image features each with a tailored degree of invariance. For each feature we establish the quality, such that the feature suited for the computer vision task at hand can be chosen.

1.2 Variants and Invariants for Computer Vision

Both the variant and invariant approaches to computer vision will start with the intensity-specific approach as the image intensities are the only information at the start. A starting point for intensity-specific image measurements are Gaussian-shaped filters for many, well-documented reasons. Among them we note their ability not to introduce new peaks in the image field [67], the capability to process the dimensions separately by subsequent one-dimensional filters, the ability to steer them to a preferred orientation [41], the fact that the Gaussian filters slope towards zero at the tails, the ability of the Gaussian filter and its derivatives to represent a signal completely by means of a Taylor series [40] and their robustness to image noise. These and other reasons assure that the Gaussian filters are commonly used these days [20–22, 46, 72, 74, 77, 103]. Pixel properties can be measured in the one color dimension (sampled by usually three filters), the two spatial dimensions, and/or the one time dimension in case of video, yielding a complete Gaussian-based measurement set including zeroth and higher order derivatives to measure also differential structure in the image [3]. Generally, we denote a Gaussian filter $G(a)$ to measure a variable a, and its i-th order derivative $G_{a^i}(a)$; this notation is used throughout the thesis. With the color dimension parameterized by wavelength λ, the Gaussian color (derivative) filters are denoted $G_{\lambda^i}(\lambda)$, and correspond to Koenderink's Gaussian opponent color model [46]. Spatial and temporal (derivative) filters are denoted $G_{x^i y^j}(x, y)$ [67] and $G_{t^i}(t)$ [2], respectively. A Gaussian in the Fourier domain results in a spatial Gabor filter, hence the Gabor filter can be considered as part of the Gaussian measurement framework. For the measurement of spatial regularity, the Gaussian filter needs to be tuned to a central frequency Ω_0, resulting in a spatial Gabor filter $\tilde{G}^{\Omega_{x0},\Omega_{y0}}(x, y)$ [14]. Likewise, a temporal frequency filter $\tilde{G}^{\Omega_{t0}}(t)$ may be derived. The complete Gaussian measurement set is obtained from combining the separable filters from each dimension. For instance, color edges are measured by subsequent application of a specific color filter and a spatial derivative filter, while for color texture a spatial frequency filter may be applied rather than a derivative filter. Motion information of objects may be measured from subsequently filtering of the result by a temporal derivative or frequency filter. We refer to these measurements as *variant* features.

To measure object-specific properties, the intensity-specific variant features need to be combined in image *invariants* that are independent of the sources of variations in the scene. Throughout this thesis, we consider features that are invariant to the most important scene parameters: varying illumination color, intensity and direction, and varying object position and pose (or, alternatively, varying camera viewpoint). These accidental settings of the scene are illustrated in Figure 1.3. We give preference to photometric invariant features from physical principles, see e.g. [36–38, 42, 50, 123]. We leave the machine learning approach to invariants out of consideration, that is to record the values under all possible conditions from extensive experimenting over representative datasets. Such an approach will yield robust invariant features only if

Table 1.1: Filters, observables, unwanted scene variations, and invariants.

Physical variable	*Filter set*	*Observable*	*Disturbance(s)*	*Invariant(s)*
Wavelength spectrum	$G_{\lambda^i}(\lambda)$ Geusebroek [46]	Object color	Illumination intensity	Geusebroek [46]
			Illumination intensity, shadow, shading, highlights	[46]
			Illumination intensity and spectrum, shadow, shading	[46]
Local geometry	$G_{x^j y^k}(x, y)$ Koenderink [71]	Object shape	Object pose	Florack [39]
			Object distance	Lindeberg [77]
Spatial frequency	$\tilde{G}^{\Omega x_0, \Omega y_0}(x, y)$ Bovik [14]	Shape regularity	Object pose and distance	Jain [61]
Time	$G_{t^i}(t)$ Adelson [2]	Object motion	Object distance	–
Temporal Frequency	$\tilde{G}^{\Omega t_0}(t)$ Burghouts [18]	Object motion periodicity	Object distance	Burghouts [18]

To measure object properties, Gaussian-shaped image and video filters are denoted by G. For measurements of regularity in space or time Gabor functions are used, indicated by $\tilde{G}$. Accidental settings of the scene are categorized into varying lighting conditions (illumination color, intensity and direction), and object pose, rotation of distance (or camera viewpoint), as illustrated in Figure 1.3. To remove scene disturbances, invariants have been proposed in literature.

the dataset is large and representative, but for general conditions, features derived from a precise physical model of the image formation will yield similar results without extensive learning. Further, we consider local features in the image, to achieve robustness to object occlusion and image clutter.

Color measurements are affected by changing illumination color, intensity and direction. To counteract these appearance deviations locally, we adopt the photometric invariants of Geusebroek *et al.* [46], which are based on the Gaussian derivative framework. From the framework, we consider features that are increasingly invariant: features invariant to illumination intensity, additional invariance to shadow and shading, and additional invariance to highlights (specularities). The spatial measurements are disturbed by changing orientation and distance of the viewed object. To obtain rotation invariant measurements, we consider the framework of Florack *et al.* [40], while for scale-invariant application of features we adopt the framework of Lindeberg [77]. Measurements of motion are also affected by a change of object distance. To derive an invariant for motion directly, multiple cues have to be considered to disambiguate the distance and size of the object. An alternative is to measure motion from quasi-periodically moving objects, for which a feature invariant to object distance can be constructed directly as will be proposed by the author. In recapitulation, the local measurement framework is extended with invariant features. The variant features to measure particular observables, and the invariants that can be constructed from them in order to counteract particular disturbing appearance variations, are considered in this thesis are summarized in Table 1.

1.3 Central Questions of the Thesis

The choice of image feature is very important for the design of vision systems, as it determines the input for subsequent analysis. The suitability of a feature for the task at hand depends on a feature's intrinsic properties such as the reliability and accuracy of its values. Furthermore, feature suitability depends on the given dataset: both the amount of objects that are to be discriminated as well as the amount of object appearance variation need to be taken into account. The objective of this thesis is to enable the designer of a vision system to select a suitable feature for the problem at hand. The central problem addressed in this thesis is: **Which image feature should be chosen for the problem at hand?**

Throughout the thesis, local features of object color, shape and motion are considered. We start with establishing the quality of features that can be measured in still images. The problem addressed in Chapter 2 is: **What is the quality of a feature to measure object properties in a single image?** We investigate quality of variant and invariant features by measuring invariance under the various scene disturbances.

In Chapter 3 we consider the following problem: **What is the quality of a feature to distinguish between many real-world objects?** We establish the quality of variant and invariant features for the description of object parts by evaluating the matching accuracy. The matching quality is measured for various imaging conditions and for image transformations that occur frequently in practice: compression and blurring of the image.

As can be concluded from Section 1.2, the Gaussian/Gabor measurement framework is not complete: a temporal frequency measurement is missing. For the offline case, a temporal frequency measurement is performed trivially by a temporal Gabor filter. For the online case, only image frames from the past are available. This requires a different approach, as the filter needs to be reshaped. In Chapter 4 we therefore provide a solution to the question: **In the Gaussian framework, what is the temporal frequency filter?**

In computer vision, many methods are based on grey-value images. A well-known example is the model of Varma and Zisserman [120] to model materials. In Chapter 5, we consider the problem of incorporating color information a posteriori such that the original grey-value models are left untouched: **How can grey-value histogram features be extended a posteriori to include color information?**

Commonly, a fixed feature (set) is used to model a dataset of objects, while objects may have different distinctive properties. In Chapter 6, we consider materials, and we propose a framework to model material-specific properties. The problem addressed is: **How can variant and invariant features be selected or combined for each material specifically?**

One step beyond feature measurement, towards their employment in the context of a specific problem, is to measure goal-oriented features, here referred to as similarities. At the end of this thesis, in Chapter 7, we investigate similarity between objects in a large dataset to facilitate their search. We investigate the problem: **How are similarities between objects distributed across a large dataset?**

Chapter 2

Quality of Variant and Invariant Features for Color Image Processing*

2.1 Introduction

In image and video processing, a fundamental choice is to take the picture as a starting point, or to take the objects in the real world represented in the picture as a starting point. A large part of image processing methods is intensity specific in that it deals with the intensities as they appear in the image. This is the preferred strategy for general-purpose compression and communication. For other image processing tasks as image segmentation, interpretation, and object-identification, the heart of the matter is in the properties of the objects represented by the picture [112]. In object-specific image processing, it is important first to remove the accidental variation introduced into the image at the moment of recording. For one, viewpoint has a distinct effect on the image and should be removed before one would like to deal with object-specific properties. The same holds for the color, direction and intensity of illumination, introducing shadows in the image. Again, this introduces many scene-dependent effects into the image to be removed before object-specific analysis may proceed. Hence, for object-specific image processing image invariants have to be considered.

Any of two approaches will start with the intensity-specific approach as the image intensities are the only information at the start. A starting point for intensity-specific image processing is Gaussian-shaped filters for many, well-documented reasons. Among them we note their ability not to introduce new peaks in the image field [67], the capability to process the dimensions separately by subsequent one-dimensional filters, the ability to steer them to a preferred orientation [41], the fact that the Gaussian filters slope towards zero at the tails, the ability of the Gaussian filter and its derivatives to represent a signal completely by means of a Taylor series [40] and their robustness to image noise. These and other reasons assure that the Gaussian

*Appeared partially in *Proceedings of the European Cognitive Vision Conference*, 2004.

filters are commonly used these days [20–22, 46, 72, 74, 77, 103]. The pixel properties can be measured in the two spatial dimensions, the one color dimension (sampled by usually no more than three values), and/or the one time dimension yielding a complete and exhaustive point-based measurement set. An immediate advantage of the Gaussian filter is evident here, as any combination of a dimension with any of the other dimensions is separable, computationally as well as analytically. In the following, we will use a complete set of Gaussian filters to measure all intensity-specific variant image properties.

To measure object-specific properties, the intensity-specific properties need to be combined in image invariants that are independent of the sources of variations in the scene. Invariant features require a proper balance between constancy of the object measurement regardless of the disturbing scene parameters on the one hand, and retained discriminating power between truly different states of the objects on the other. Hence, both invariance and discriminative power of object-specific measurements should be investigated simultaneously. We provide a complete set of invariant image features to measure object properties independent of the most important scene parameters: varying illumination direction, illumination color and varying viewing directions.

We give preference to invariant features from physical principles. We leave the machine learning approach to invariants out of consideration, that is to record the values under all possible conditions from extensive experimenting over representative datasets. Such an approach will yield robust invariant features only if the dataset is large and representative, but for general conditions, features derived from a physical model of imaging will yield similar results without extensive learning.

Starting from the complete sets of variant (intensity-specific) and invariant (object-specific) feature sets, our contribution is to assess the quality of variants and invariants. We determine the quality of the features according to their robustness, invariance, information content and discriminative power. Finally, we demonstrate the merit of using invariant features rather than variants for real-world imaging conditions.

First, we give a complete set of Gaussian filter families to measure intensity-specific entities in the spatial, color and temporal dimensions (Section 2.5). We combine systematically the intensity-specific measurements such that object-specific properties can be measured independent of disturbing scene parameters (Section 2.5). We list quality measures for variant and invariant features (Section 2.4) and validate the quality of variant and invariant features in Section 2.5. We wrap up with conclusions on the suitability of the features for image processing applications.

2.2 Variants: Measurable Pictorial Properties

Measurements of a picture signal imply integration over a spatial, spectral and temporal region. The visual measurement $\hat{E} : \mathbb{R}^4 \mapsto \mathbb{R}$ of the color video signal $E(x, y, \lambda, t)$ is the linear correlation of E with a Gaussian filter type $G : \mathbb{R}^4 \mapsto \mathbb{R}$:

$$\hat{E}_{x^i y^j \lambda^k t^l}(x, y, \lambda, t) \equiv$$

$$\int\int\int\int E(x,y,\lambda,t)\, G^{x_0,y_0,\lambda_0,t_0;\sigma_x,\sigma_y,\sigma_\lambda,\sigma_t}_{x^i y^j \lambda^k t^l}(x,y,\lambda,t)\, dx\, dy\, d\lambda\, dt, \tag{2.1}$$

with $(\sigma_x, \sigma_y, \sigma_\lambda, \sigma_t) \in \mathbb{R}^4$ the scales of the filters in each dimension and $(x_0, y_0, \lambda_0, t_0) \in \mathbb{R}^4$ its location. The superscript indices in the subscripts denote the order of differentiation. We drop the location parameters. Using the separability of the Gaussian, we list the filters that measure spatial, color and temporal properties in the image independently.

The measurement of local geometry in the image is equivalent to a filter that measures the change in the intensity structure at (x, y) at scale σ_{xy} [71]:

$$G^{\sigma_{xy}}_{x^i y^j}, \tag{2.2}$$

of which the measurements are given by $\hat{E}_{x^i y^j} = G^{\sigma_{xy}}_{x^i y^j} * E$. As an example, image edges are represented by G_x. Up to second order, 6 spatial derivative filters can be constructed.

For the measurement of local regularity, the filter needs to be tuned to a central frequency Ω_0. A Gaussian in the Fourier domain results in a Gabor filter in the spatial domain. In two dimensions, the Gabor filter is given by [14]:

$$\tilde{G}^{\sigma_{xy},\Omega_{x_0},\Omega_{y_0}}(x,y) \equiv G^{\sigma_{xy}}(x,y)\, e^{2\pi i \left(\begin{smallmatrix}\Omega_{x_0}\\ \Omega_{y_0}\end{smallmatrix}\right)\cdot\left(\begin{smallmatrix}x\\ y\end{smallmatrix}\right)}, \quad i^2 = -1, \tag{2.3}$$

where $\sqrt{\Omega_{x_0}^2 + \Omega_{y_0}^2}$ is the radial central frequency and $\tan^{-1}(\frac{\Omega_{y_0}}{\Omega_{x_0}})$ the orientation. The frequency may be zero in one dimension, yielding a combined Gaussian-Gabor filter.

These measurements can also be applied to multi-valued images, where RGB-color values are a sampling of the color dimension yielding three values per pixel. Embedded in the Gaussian framework, color filters have been proposed [46] as an opponent color system which is approximately colorimetric with human vision. The three filters that measure wavelength λ are given by:

$$G^{\sigma_\lambda}(\lambda), G^{\sigma_\lambda}_{\lambda}(\lambda), G^{\sigma_\lambda}_{\lambda\lambda}(\lambda). \tag{2.4}$$

In practice, the Gaussian opponent color values are obtained from linear combination of RGB color values that approximate the sensitivity curves of the filters from Equation 2.4:

$$\begin{bmatrix} \hat{E}(x,y) \\ \hat{E}_\lambda(x,y) \\ \hat{E}_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{bmatrix} E(x,y) * G(\lambda) \\ E(x,y) * G_\lambda(\lambda) \\ E(x,y) * G_{\lambda\lambda}(\lambda) \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{pmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix}. \tag{2.5}$$

The colored spatial differential image structure, for instance, the measurement of color edges, is given by combination of the spatial derivative filters and the color filters [46]:

$$G^{\sigma_\lambda;\sigma_{xy}}_{\lambda^i x^j y^k} = G^{\sigma_\lambda}_{\lambda^i} * G^{\sigma_{xy}}_{x^j y^k}. \tag{2.6}$$

In the same way as in Equation 2.3, regularity can be measured in the color intensity fields by [58, 61]:

$$\tilde{G}^{\sigma_\lambda;\sigma_{xy},\Omega_{x_0},\Omega_{y_0}}_{\lambda^i} = G^{\sigma_\lambda}_{\lambda^i} * \tilde{G}^{\sigma_{xy};\Omega_{x_0},\Omega_{y_0}}. \tag{2.7}$$

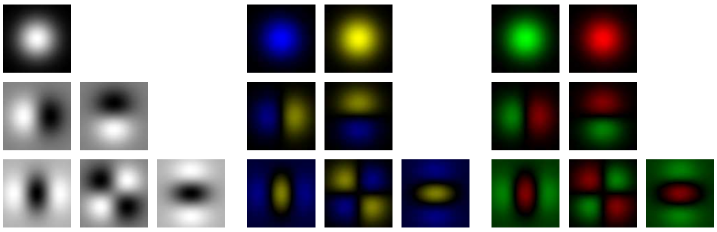


Figure 2.1: The family $G_{x^i y^j \lambda^k}(x, y, \lambda)$ of 18 color differential filters up to second differential order in wavelength. Yellow and red have negative values, blue and green positive.

For temporal analysis in online video only the past is available. A logarithmical reparameterization of the time axis solves this, yielding the measurement to determine temporal change [69]:

$$G^{\sigma_t}_{t^i}(t) = \frac{1}{\sqrt{2\pi}\,\sigma_t}\, e^{-\frac{\log\left(\frac{t'-t'_0}{\sigma_t}\right)^2}{2\,\sigma_t^2}}, \tag{2.8}$$

where response time delay is denoted σ_t. Analogous to the spatial frequency domain, the temporal frequency can be measured by a temporal Gabor filter. As with Equation 2.8, its envelope is logarithmically rescaled for online filtering [18]:

$$\tilde{G}^{\sigma_t, \sigma_{t'}, \Omega_{t_0}}(t) \equiv G^{\sigma_t}(t)\, e^{2\pi i \Omega_{t_0} \frac{t}{\sigma_{t'}}}, \quad i^2 = -1, \tag{2.9}$$

where Ω_{t_0} denotes the temporal frequency, and $G^{\sigma_t}(t)$ the time filter of Equation 2.8.

By the separability of the Gaussian filter any spatial or color filter can be combined with a time filter. For instance, the filter that determines temporal changes in the color differential structure is given by:

$$G^{\sigma_\lambda;\sigma_{xy};\sigma_t}_{\lambda^i x^j y^k t^l} = G^{\sigma_\lambda}_{\lambda^i} * G^{\sigma_{xy}}_{x^j y^k} * G^{\sigma_t}_{t^l}. \tag{2.10}$$

Likewise, to measure regularity rather than differential structure, the spatial and temporal derivative filter can be substituted respectively by a spatial frequency filter (Equation 2.3) and temporal frequency filter (Equation 2.9):

$$\begin{aligned}
G^{\sigma_\lambda;\sigma_{xy},\Omega_{x_0},\Omega_{y_0};\sigma_t}_{\lambda^i t^l} &= G^{\sigma_\lambda}_{\lambda^i} * \tilde{G}^{\sigma_{xy},\Omega_{x_0},\Omega_{y_0}} * G^{\sigma_t}_{t^l} &\quad (2.11)\\
G^{\sigma_\lambda;\sigma_{xy};\sigma_t,\Omega_{t_0}}_{\lambda^i x^j y^k} &= G^{\sigma_\lambda}_{\lambda^i} * G^{\sigma_{xy}}_{x^j y^k} * \tilde{G}^{\sigma_t,\sigma_{t'},\Omega_{t_0}} &\quad (2.12)\\
G^{\sigma_\lambda;\sigma_{xy},\Omega_{x_0},\Omega_{y_0};\sigma_t,\Omega_{t_0}}_{\lambda^i} &= G^{\sigma_\lambda}_{\lambda^i} * \tilde{G}^{\sigma_{xy},\Omega_{x_0},\Omega_{y_0}} * \tilde{G}^{\sigma_t,\sigma_{t'},\Omega_{t_0}} &\quad (2.13)
\end{aligned}$$

In recapitulation, the complete set of all reasonable point-based properties that can be measured in an image is given in Table 2.1. The last column specifies a family of differential filters one to each box. As an example, we depict in Figure 2.1 the color differential filter family defined by Equation 2.6 truncated at second order, yielding 3 color filters × 6 spatial derivative filters, or 18 filters in total. For the 3 color Gabor filters, a filter family is constructed by variation over orientations and center frequencies. A good coverage of the spatial frequency domain is obtained

Table 2.1: Measurable point-based properties in images and corresponding filters.

Measurable image property	*Description*	*Variables*	*Refs*	*Gaussian filter family*	*Eq.*
Differential structure in the intensity field	Local geometry, e.g., edges, curvature	x, y	[71]	$G_{x^i y^j}(x, y)$	(2.2)
∼ color intensity fields	Local color geometry	x, y, λ	[46]	$G_{x^i y^j \lambda^k}(x, y, \lambda)$	(2.6)
∼ and temporal change	Motion of local color geometry	x, y, λ, t	[2]	$G_{x^i y^j \lambda^k t^l}(x, y, \lambda, t)$	(2.10)
∼ and temporal periodicity	Periodic motion of local color geometry	$x, y, \lambda, \Omega(t)$	[18, 74]	$\tilde{G}^{\Omega_t}_{x^i y^j \lambda^k}(x, y, \lambda, t)$	(2.12)
Regularity in the intensity field	Regular texture	$\Omega(x), \Omega(y)$	[14]	$\tilde{G}^{\Omega_x, \Omega_y}(x, y)$	(2.3)
∼ color intensity fields	Regular color texture	$\Omega(x), \Omega(y), \lambda$	[61]	$\tilde{G}^{\Omega_x, \Omega_y}_{\lambda^k}(x, y, \lambda)$	(2.7)
∼ and temporal change	Motion of regular color texture	$\Omega(x), \Omega(y), \lambda, t$	–	$\tilde{G}^{\Omega_x, \Omega_y}_{\lambda^k t^l}(x, y, \lambda, t)$	(2.11)
∼ and temporal periodicity	Periodic motion of regular color texture	$\Omega(x), \Omega(y), \lambda, \Omega(t)$	–	$\tilde{G}^{\Omega_x, \Omega_y, \Omega_t}_{\lambda^k}(x, y, \lambda, t)$	(2.13)

Gaussian-shaped image and video filters are denoted G. For measurements of regularity in space or time Gabor functions are used, indicated by $\tilde{G}$. We have dropped the scale parameters in the table.

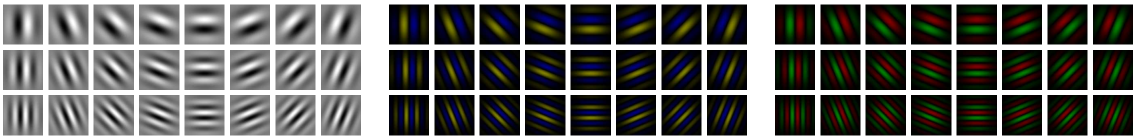


Figure 2.2: The family $\tilde{G}^{\Omega_x, \Omega_y}_{\lambda^k}(x, y, \lambda)$ of 3 colored Gabor filters up to second differential order in wavelength. The filter banks are obtained from varying the orientation and center frequency parameters.

by 24 orientation-frequency selective filters [72], see Figure 2.2 for the Gabor filter family. In as much as human vision is limited to wavelength derivative filters up to second order [57], and considering spatial derivatives up to order four [130], we obtain 3 × 24 color geometry filters for a total of 144 spatial filters. Independent to all of this, temporal differential order is truncated at first order for human vision [31], yielding 2 derivative filters (zeroth and first order) over time which can be replaced by a temporal frequency filter.

2.3 Invariants: Measurable Object Properties

In this section, we regard object-specific measurements. We consider the most important object-specific properties which can be measured at a point such as local shape, texture and color of the object. Object rotation, pose and distance to the observer are scene-dependent and hence not considered here.

To eliminate the effects generated by the scene that are disturbing the analysis of object-specific features, we start from physical models that capture the image formation at different stages. We distinguish the stages in which the light irradiates the object under consideration and the reflected light is projected onto the image plane. In the following, we extend the initial framework presented in [112].

We start with the illumination irradiating the scene. The illumination has a direction relative to objects in the scene as well as a spectral composition and intensity. The light reflected from an object onto the image plane depends on geometry and the object reflectance function. The formation of the color image is modelled by means of the Kubelka-Munk theory [63,73]. The reflected spectrum in the viewing direction is given by:

$$E(\lambda, x, y) = e(\lambda, x, y)(1 - \rho(x, y))^2 R(\lambda, x, y) + e(\lambda, x, y)\rho(x, y), \tag{2.14}$$

where $e(\lambda, x, y)$ denotes the illumination spectrum, $\rho(x, y)$ the Fresnel reflectance and $R(\lambda, x, y)$ the object reflectance function.

When deriving object-specific hence invariant features, the rationale is to form groups by image formation parameters (for instance, the object has a matte surface). The assumption simplifies the reflectance model. From the simplified model expressions can be derived for the invariant features.

With white illumination and intensity variations, the reflectance model for matte surfaces reduces to:
$E(\lambda, x) = i(x)\, R(\lambda, x)$. The expression $\frac{E_x}{E} = \frac{1}{R(\lambda,x)} \frac{\partial R(\lambda,x)}{partial\lambda}$ depends on object reflectance $R(\lambda, x)$ only, hence: it is an object reflectance property independent of shadow and shading [46]. Differentiation of this expression yields a complete set of invariants [95], $\mathcal{C}_{\lambda^m x^n}$:

$$\mathcal{C}_{\lambda^m x^n} = \frac{\partial^n}{\partial x^n} \left\{ \frac{E_{\lambda^n}}{E} \right\}. \tag{2.15}$$

Substitution of $E_{\lambda^m x^n}$ by measurements of the spatiospectral filter family given in Equation 2.6, $\hat{E}^\sigma_{\lambda^m x^n} = E * G^\sigma_{\lambda^m x^n}$, yields the object-specific measurement. The family $\mathcal{C}^\sigma_{\lambda^m x^n}$ contains chrominant geometrical descriptors, robust to changes of shadow and shading.

To measure object-specific measurements of regular color patterns, the Gabor filters from Equation 2.3 are convolved with shadow and shading invariant measurements that are again obtained from the variants $\hat{E}^\sigma_{\lambda^m x^n} = E * G_{\lambda^m x^n}$ from Equation 2.6 [58]:

$$\mathcal{T}^{\Omega,\phi}_{\lambda} = \tilde{G}^{\Omega_{x_0},\Omega_{y_0}} * (\frac{\hat{E}_\lambda}{\hat{E}}),\ \mathcal{T}^{\Omega,\phi}_{\lambda\lambda} = \tilde{G}^{\Omega_{x_0},\Omega_{y_0}} * (\frac{\hat{E}\hat{E}_{\lambda\lambda} - \hat{E}^2_\lambda}{\hat{E}^2}). \tag{2.16}$$

In analogy to the shadow and shading invariant family $\mathcal{C}^\sigma_{\lambda^m x^n}$, a family invariant to illumination intensity, $\mathcal{W}^\sigma_{\lambda^m x^n}$, is derived from normalization by the local energy [46], where the object-specific measurements are given by:

$$\mathcal{W}^\sigma_{\lambda^m x^n} = \left\{ \frac{\hat{E}^\sigma_{\lambda^m x^n}}{\hat{E}^\sigma} \right\}_{m \geq 0,\, n \geq 1}. \tag{2.17}$$

Finally, an invariant family robust to shadow, shading and highlight effects, $\mathcal{H}^{\sigma}_{x^n}$, is obtained from the variant spectral measurements $\hat{E}^{\sigma}_{\lambda^m} = E * G^{\sigma}_{\lambda^m}$ from Equation 2.4:

$$\mathcal{H}^{\sigma}_{x^n} = \frac{\partial^n}{\partial x^n} \left\{ \arctan(\frac{\hat{E}^{\sigma}_{\lambda}}{\hat{E}^{\sigma}_{\lambda\lambda}}) \right\}_{n \geq 0} \ . \tag{2.18}$$

We consider geometric variation and deal with translation, rotation and scale effects. When convolving the image with local geometry filters at dense spatial locations, object measurements are translation invariant [67]. By considering local geometry filters in a systematic manner in local gauge coordinates, measurements of object geometry are rotation invariant [39]. Fixing local gauge coordinates (v, w) aligns the original coordinate system (x, y) to the local image structure. For instance, measurements from Equation 2.6 or any invariant from the same differential order can be aligned to the isophote, that is, the gradient: $\vec{w} = (\hat{E}_{\lambda^i x}, \hat{E}_{\lambda^i y})$, where the perpendicular direction is given by $\vec{v} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \vec{w}$. The rotation invariant family of object shape measurements is given by:

$$\mathcal{I}^{\sigma}_{v^m w^n} = \{\hat{F}^{\sigma}_{v^m w^n}\}_{m,n \geq 0, m \neq 1}, \tag{2.19}$$

where the measurement $\hat{F}^{\sigma}_{v^m w^n}$ is obtained from any measurement (variant or invariant) from the same differential orders. For instance, the gradient magnitudes of the photometric invariants $\mathcal{W}_{x^i y^j \lambda^k}$, $\mathcal{N}_{x^i y^j \lambda^k}$ and $\mathcal{H}_{x^i y^j}$ is given by:

$$\mathcal{W}^{\sigma}_{\lambda^i w} = \sqrt{(\mathcal{W}^{\sigma}_{\lambda^i x})^2 + (\mathcal{W}^{\sigma}_{\lambda^i y})^2} \tag{2.20}$$

$$\mathcal{C}^{\sigma}_{\lambda^i w} = \sqrt{(\mathcal{C}^{\sigma}_{\lambda^i x})^2 + (\mathcal{C}^{\sigma}_{\lambda^i y})^2} \tag{2.21}$$

$$\mathcal{H}^{\sigma}_{w} = \sqrt{(\mathcal{H}^{\sigma}_{x})^2 + (\mathcal{H}^{\sigma}_{y})^2}. \tag{2.22}$$

Scale invariant object shape measurements are obtained from maximizing the scale-normalized invariants from Equation 2.19 [77]:

$$\mathcal{I}_{v^m w^n} = \max_{\sigma_{xy}} \mathcal{I}^{\sigma}_{v^m w^n}. \tag{2.23}$$

Object texture measurements are also disturbed by rotational variation. Rotation invariant texture measurements are obtained from integration of the Gabor measurements over the rotation group $W_{\phi} = \ \vec{x}' \ = \ (\ \begin{matrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{matrix} \ ) \ \vec{x}$. The family of rotation invariant object texture measurements $\mathcal{S}^{\Omega}_{\lambda^i}$ is given by [19]:

$$\begin{aligned} \mathcal{S}^{\Omega}_{\lambda^i} &= \tfrac{1}{|W_{\phi}|} \textstyle\int_{W_{\phi}} \hat{S}^{\Omega,\phi}_{\lambda^i} d\phi \\ &\approx \tfrac{1}{n} \textstyle\sum_{i=0}^{n-1} \hat{S}^{\Omega,\phi_i}, \quad \phi_i = \tan^{-1}(\tfrac{\Omega_{y0}}{\Omega_{x_0}}) = \tfrac{i\,\pi}{n}, \end{aligned} \tag{2.24}$$

where color Gabor measurements are denoted $\hat{S}^{\Omega,\phi}_{\lambda^i} = E * \tilde{G}^{\Omega,\phi}_{\lambda^i}$ (Equation 2.7). To combine rotation and photometric invariance, the color Gabor measurements may be replaced by the shadow and shading invariant measurements of object texture from Equation 2.16 to obtain the rotation invariant texture family $\mathcal{T}^{\Omega}_{\lambda^i}$:

$$\mathcal{T}^{\Omega}_{\lambda^i} = \frac{1}{|W_\phi|} \int_{W_\phi} \hat{T}^{\Omega,\phi}_{\lambda^i} d\phi. \tag{2.25}$$

Additionally, the rotation invariant texture measurements may be maximized over various center frequencies Ω [61] to obtain the rotation and scale invariant families $\mathcal{S}_{\lambda^i}$ and $\mathcal{T}_{\lambda^i}$, where the latter is also photometric invariant:

$$\begin{aligned} \mathcal{S}_{\lambda^i} &= \max_{\Omega} \mathcal{S}^{\Omega}_{\lambda^i} && (2.26) \\ \mathcal{T}_{\lambda^i} &= \max_{\Omega} \mathcal{T}^{\Omega}_{\lambda^i}. && (2.27) \end{aligned}$$

In summary, the object-specific measurements that are invariant to irrelevant scene variations when determining object color, shape or texture are given in Table 2.2. The last column specifies a family of differential invariants one to each row. We consider photometric invariants $\mathcal{W}_{v^i w^j \lambda^k}$, $\mathcal{N}_{v^i w^j \lambda^k}$, $\mathcal{H}_{v^i w^j}$ up to to second order. The number of invariants in gauge coordinates (v, w) add up to 5 rotation invariants for each spectral derivative order for each invariant. The family $\mathcal{W}_{v^i w^j \lambda^k}$ contains derivatives up to second spectral order. The family $\mathcal{N}_{v^i w^j \lambda^k}$ is only defined for spectral first and second order derivatives, whereas the hue family $\mathcal{H}_{v^i w^j}$ does not contain spectral derivatives. For the 3 photometric invariant families we obtain 5 rotation invariants times respectively 3, 2 and 1 spectral variations, yielding $15 + 10 + 5$ adding up to 30 invariants up to second order measuring object reflectance and local shape. The invariant families $\mathcal{W}_{v^i w^j \lambda^k}$, $\mathcal{N}_{v^i w^j \lambda^k}$ and $\mathcal{H}_{v^i w^j}$ all contain less measurements than the variant family from which the invariants are obtained: 18 variants compared to respectively 15, 10 and 5 invariants.

2.4 Quality Measures

Quality measures can be distinguished into reproducibility and discriminative power of features. We measure the quality of a feature by considering the mean and standard deviation of its measured value, $\{\hat{q}_i\}$, relative to its expected value, $\hat{m}_i$. Without prior knowledge of the feature values, we normalize the (Euclidean) feature value differences by the reference value,

$$\hat{d}_i = \frac{||\bar{m}_i - \bar{q}_i||}{\bar{m}_i} \quad . \tag{2.28}$$

As a measure of quality, the set of differences are summarized in terms of the mean and standard deviation,

$$Q = \{\mu(\hat{d}_i), \sigma(\hat{d}_i)\} \quad . \tag{2.29}$$

In the following we identify for each of the proposed quality measures the actual measurements $\hat{m}_i$ and $\hat{q}_i$.

Table 2.2: Measurable point-based properties of objects and corresponding invariants.

Measurable object property	*Description*	*Irrelevant scene parameter*	*Refs*	*Invariant family*	*Eqs*
Local geometry	Edge properties	Object pose	[39]	$\mathcal{I}^{\sigma}_{v^i w^j}$	(2.19)
		Object pose and distance	[77]	$\mathcal{I}_{v^i w^j}$	(2.23)
Color local geometry	Object reflectance and shape	Illumination intensity	[46]	$\mathcal{W}^{\sigma}_{x^i y^j \lambda^k}$	(2.17)
	Object reflectance and shape	Illumination intensity and object pose and distance	–	$\mathcal{W}_{v^i w^j \lambda^k}$	(2.20)
	Object reflectance and shape	Illumination intensity, shadow, shading	[46]	$\mathcal{C}^{\sigma}_{x^i y^j \lambda^k}$	(2.15)
	Object reflectance and shape	Illumination intensity, shadow, shading and object pose and distance	–	$\mathcal{C}_{v^i w^j \lambda^k}$	(2.21)
	Object reflectance and shape	Illumination intensity, shadow, shading and highlights	[46]	$\mathcal{H}^{\sigma}_{x^i y^j}$	(2.18)
	Object reflectance and shape	Illumination intensity, shadow, shading, highlights and object pose and distance	–	$\mathcal{H}_{v^i w^j}$	(2.22)
Regularity	Object texture	Object rotation	[19]	$\mathcal{S}^{\Omega}_{\lambda^i}$	(2.24)
	Object texture	Object distance	[61]	$\mathcal{S}_{\lambda^i}$	(2.26)
Color regularity	Object color texture	Shadow and shading	[58]	$\mathcal{T}^{\Omega,\phi}_{\lambda^i}$	(2.16)
	Object color texture	Shadow and shading and object rotation	–	$\mathcal{T}^{\Omega}_{\lambda^i}$	(2.25)
	Object color texture	Shadow and shading and object rotation/distance	–	$\mathcal{T}_{\lambda^i}$	(2.27)

2.4.1 Displacement

Displacement of a feature indicates the reproducibility of locations where the value of the feature is expected to be found. That is, smaller localization errors of a feature indicate better reproducibility. We measure the displacement of the location of the maximum response of a feature, $\hat{q}_i = \arg_{(x,y)} \max \hat{F}(x, y)$ relative to the location of the entity to be measured $m_i = (x, y)$. The omission of the hat sign indicates that it is absolute information, i.e. the ground truth. Then, $\hat{d}_i = ||m_i - \hat{q}_i||$. Larger spatial scales of features introduce more uncertainty in the localization of the measured entity [14]. Hence, for fair comparison between features, $\hat{q}_i$ should be measured at fixed scale.

2.4.2 Stability

Stability is the reproducibility of a measurement when measured over several instances of the same but noisy or transformed data. At a straight color transition T_i, we measure the feature values $\hat{F}_{ij} = \hat{F}(x,y)_{(x,y)\in T_i}$ along the transition. First, as an indication of image noise sensitivity, we consider the differences of the values $\hat{q}_{ij} = \hat{F}_{ij}$ to the mean value $\hat{m}_i = \mu(\hat{F}_{ij})$. We accumulate the sets of differences over multiple transitions. Second, we reduce the intensity of the transition image, T_i', to establish feature sensitivity to low hence noisier pixel values. To that end, we repeat the setup but now with $\hat{q}_{ij} = \hat{F}'_{ij}$ and $\hat{m}_i = \mu(\hat{F}'_{ij})$. As an indication of the feature sensitivity to compression, we consider the JPEG compression of the color transition, T_i''. We consider the difference between the mean values obtained from T_i and from T_i'', $\hat{q}_i = \mu(\hat{F}''_{ij})$ and $\hat{m}_i = \mu(\hat{F}_{ij})$.

2.4.3 Invariance

Invariance is the reproducibility of a measurement when measured over the same image data but recorded under truly different imaging conditions. For a point in the image, (x, y), we measure the feature value $\hat{m}_i = \hat{F}_i$. Fixing the camera, we consider the recording of the same scene, but with a different imaging condition, i'. Measuring the feature at the same position (x, y), we obtain $\hat{m}_j = \hat{F}'_i$.

2.4.4 Discriminative Power

Discriminative power is the contrast of a measurement to other measurements of truly different data. For a point in the image, (x, y), we measure the feature value $\hat{m}_i = \hat{F}_i$. For other points, we measure the feature values, and keep the one that is most similar to $\hat{m}_i$: $\hat{q}_i$. We count the number of points that can be discriminated when requiring that $\hat{d}_i = \frac{||\bar{m}_i - \bar{q}_i||}{\bar{m}_i}$ is equal or higher to a predefined contrast value δ: $Q = \#(\hat{d}_i \geq \delta)$.

2.5 Experiments

In this section, we evaluate the quality of the Gaussian features overviewed in Sections and . We do not report the quality of Gabor features, as their quality proved to be very similar to Gaussian features with similar degrees of invariance (data not shown). In the experiments, we consider two datasets to establish the quality measures:

- Displacement, stability and discriminative power. To establish displacement, stability and discriminative power of the variant and invariant features, we consider images taken from PANTONE patches [96]. The images contain on the left side one patch, and on the right side an other. Hence, we have a ground truth for each image of the line that represents the color transition. Image features are computed along this line, which we repeat for 1,000 random patch combinations.

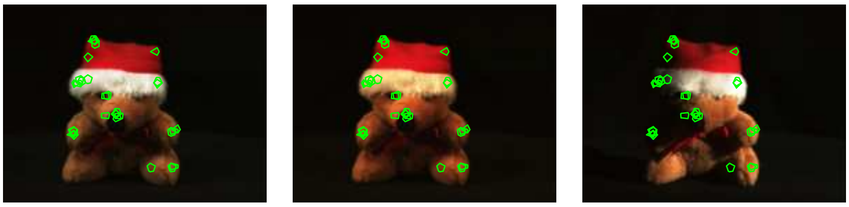


Figure 2.3: The reference image is on the left, with consecutive images illustrating reddish illumination and illumination from the side of the example object.

- Invariance. To establish invariance, we consider the object images from the ALOI database [47]. For 1,000 objects, a large number of images are recorded, each under a different imaging condition. We consider various imaging conditions, of which the camera viewpoint is kept fixed such that we have a ground truth. Image features are computed from Harris interest points [54], which have been determined for the reference image (white and hemispherical illumination) and have been copied to the other conditions (reddish illumination, and white illumination from right side only). Figure 2.3 illustrates Harris points for images of an example object.

The integral testsuite including technical details is publicly available from our website[†]. All variants and invariants are computed from Gaussian filters with $\sigma = 1$ pixels. For all quality measures, we reduce the feature values obtained from multiple color channels to a single value, by considering the total gradient: $\bar{F}_w = \sqrt{\sum_i \hat{F}^2_{\lambda^i w}}$. See the previous section for details on the normalization of the image features.

2.5.1 Displacement

For the variant feature $\bar{E}_w$ and the invariant features $\bar{\mathcal{W}}_w$ and $\bar{\mathcal{C}}_w$ the average displacement is low: 0.9 ± 0.3 pixels. For the invariant $\bar{\mathcal{H}}_w$ the displacement is somewhat higher: 1.0±0.5 pixels. Measuring at a scale of $\sigma = 2$ pixels, the displacements become larger. For increasingly invariant features $\bar{E}_w$, $\bar{\mathcal{W}}_w$, $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$, the displacement is increasingly large: 0.8 ± 0.4, 1.1 ± 0.5, 1.5 ± 0.7 and 1.4 ± 1.0 pixels. The invariants $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$ involve more nonlinear combinations of spatially smoothed images, hence their displacement is higher than of $\bar{E}_w$ and $\bar{\mathcal{W}}_w$. For all features, the displacement is on average smaller than the spatial scale.

2.5.2 Stability

The instability of the variant feature $\bar{E}_w$ to image noise is marginal: on average $1\% \pm 1\%$ change of the feature value, while for the invariants $\bar{\mathcal{W}}_w$, $\bar{\mathcal{C}}_w$, and $\bar{\mathcal{H}}_w$ the noise sensitivity is larger: respectively $2\% \pm 2\%$, $2\% \pm 2\%$, and $2\% \pm 3\%$. When

[†] http://www.science.uva.nl/~burghout/isis; will be put there as soon as the manuscript becomes publicly available.

decreasing the intensity to 10% of the original image intensity, the noise sensitivity of the feature values does not increase significantly. For $\bar{E}_w$ and $\bar{\mathcal{W}}_w$ the noise sensitivity has not increased: respectively $1\% \pm 1\%$ and $2\% \pm 2\%$. For the invariant $\bar{\mathcal{C}}_w$ the noise sensitivity has increased marginally: $3\% \pm 3\%$. For the invariant $\bar{\mathcal{H}}_w$, the sensitivity under low intensity is unacceptable: $8\% \pm 13\%$.

The invariant $\bar{\mathcal{H}}_w$ is based on a color ratio, hence it is also very sensitive to JPEG compression: $4\% \pm 11\%$. The features $\bar{E}_w$, $\bar{\mathcal{W}}_w$, and $\bar{\mathcal{C}}_w$ are much more stable: $2\% \pm 1\%$, $2\% \pm 1\%$, and $2\% \pm 2\%$. For all features except $\bar{\mathcal{H}}_w$, we conclude that the stability to image noise, low image intensities and compression is reasonable.

2.5.3 Invariance

The dataset contains reference images, i.e. objects recorded under white and hemispherical illumination, and experimental images, i.e. reddish illumination, and white illumination from right side only. With reddish illumination, the features $\bar{E}_w$ and $\bar{\mathcal{W}}_w$ perform well, the deviations are on average: $7\% \pm 3\%$ and $3\% \pm 2\%$. $\bar{\mathcal{W}}_w$ is somewhat more stable than $\bar{E}_w$, due to the non-equal intensities in the image to which it is invariant. The features $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$ perform significantly less with colored illumination, respectively $14\% \pm 7\%$ and $20\% \pm 11\%$. The invariants $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$ are computed from combinations of multiple color derivatives, increasing the effect of color deviations.

For white illumination from the side, the increasingly invariant features $\bar{E}_w$, $\bar{\mathcal{W}}_w$ and $\bar{\mathcal{C}}_w$ perform increasingly well as deviations become smaller: $19\% \pm 14\%$, $16\% \pm 12\%$ and $14\% \pm 10\%$. The exception here is $\bar{\mathcal{H}}_w$, which is designed to be most invariant but it has the largest deviation: $20\% \pm 14\%$. We conclude that $\bar{\mathcal{H}}_w$ is not a good invariant feature.

2.5.4 Discriminative Power

For the feature $\bar{E}_w$, the most similar feature value to each feature value is at least 2.5% larger or smaller. For $\bar{\mathcal{W}}_w$ and $\bar{\mathcal{C}}_w$ the discriminative power is somewhat less: respectively 98% and 97% have a nearest neighbor that differs more than 2.5%. For $\bar{\mathcal{H}}_w$, the discriminative power is lowest: 90%. Requiring that the feature values differ more than 10% of the original feature value, yields lower discriminative power. For increasingly invariant features $\bar{E}_w$, $\bar{\mathcal{W}}_w$, $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$, the discriminative power decreases more: respectively 95% (-5%), 91% (-6%), 90% (-7%) and 81% (-10%). As expected, the more invariant, the less discriminative power is maintained.

2.6 Conclusions

In this chapter, we have considered image features with various degrees of invariance, i.e. robust to shadowing, shading, and highlights. We have proposed a set of measures to determine their quality. The established quality of variant and invariant image features is summarized in Table 2.3. For increasingly invariant features $\bar{E}_w$, $\bar{\mathcal{W}}_w$, $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$ the displacement increases. But, the displacement of all features is on average less than the filter scale, hence the errors are acceptable. However, if the localization accuracy is an important issue, as it is with the Hough transform for shape detection, $\bar{E}_w$ is the preferred choice to minimize shape misdetections.

Variant feature $\bar{E}_w$ and invariant features $\bar{\mathcal{W}}_w$ and $\bar{\mathcal{C}}_w$ are very stable when the image intensity decreases or even the image is compressed. The invariant $\bar{\mathcal{H}}_w$ is very unstable to both low intensities and compression. The features $\bar{\mathcal{W}}_w$, $\bar{E}_w$, $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$ are in decreasing order robust to a change of illumination color. In fact, none is invariant to a change of illumination color, and $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$ perform worse due to the many color derivatives that they are based on. The increased robustness of $\bar{\mathcal{W}}_w$ compared to $\bar{E}_w$ is due to additional invariance to the intensity level. For varying the illumination direction, increasing the invariance is beneficial: $\bar{\mathcal{C}}_w$ has the lowest deviation. $\bar{\mathcal{H}}_w$ is not a good invariant: although it is designed to be very invariant, its values deviate under changes of the imaging conditions.

The discriminative power of the variant feature $\bar{E}_w$ is best. For increasingly invariant features $\bar{\mathcal{W}}_w$, $\bar{\mathcal{C}}_w$ and $\bar{\mathcal{H}}_w$, the discriminative power decreases. This result indicates that discriminative power is indeed inversely related to invariance. Hence, for any image processing application, both qualities should be investigated simultaneously. The preferred choice of balancing the invariance and discriminative power depends on the imaging conditions. When only changes of the illumination color are to be expected, $\bar{E}_w$ is the preferred feature. With additional changes of illumination directions, $\bar{\mathcal{C}}_w$ is the preferred feature as its loss in discriminative power compared to $\bar{E}_w$ is marginal. If additionally low image intensities may be observed, $\bar{\mathcal{W}}_w$ is the preferred feature as it is more stable than is $\bar{\mathcal{C}}_w$.

Table 2.3: Quality of Variant and Invariant Features

			Instability		Invariance		Discr.
Feature	*Invariance*	Displacement [pixels]	Low energy [%]	Compression [%]	Illumination color [%]	Illumination direction [%]	power [%]
$\bar{E}_w$	not inv.	0.5 ± 0.5	1 ± 1	2 ± 1	7 ± 3	19 ± 14	100
$\bar{\mathcal{W}}_w$	i	1.0 ± 0.5	2 ± 2	2 ± 1	3 ± 2	16 ± 12	98
$\bar{\mathcal{C}}_w$	+s	1.5 ± 0.5	3 ± 3	2 ± 2	14 ± 7	14 ± 10	97
$\bar{\mathcal{H}}_w$	+h	1.5 ± 1.0	8 ± 13	4 ± 11	20 ± 11	20 ± 14	90

Summary of the measured quality over the ALOI [47] dataset (invariance) and the PANTONE [96] dataset (other). Abbreviations of invariance: illumination intensity (i), shadow and shading (s), and highlights (h). For increasing invariance, the displacement and instability increases somewhat, while the highlight invariant is very unstable. The shadow/shading and highlight invariant are sensitive to color effects, while for a change of illumination direction the shadow/shading invariant remains most stable. Increasing order of invariance implies descreasing discriminative power.

Chapter 3

Performance Evaluation of Local Color Invariants

3.1 Introduction

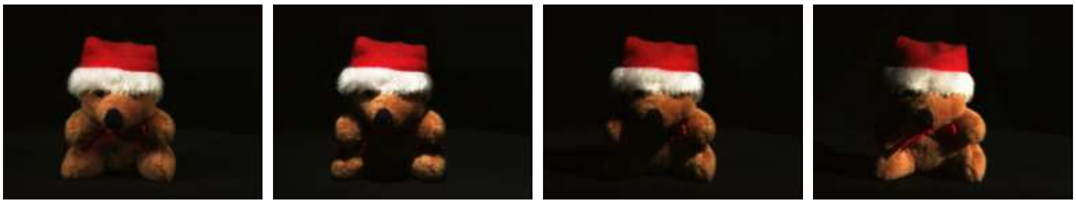
Many computer vision tasks depend heavily on local feature extraction and matching. Object recognition is a typical case where local information is gathered to obtain evidence for recognition of previously learned objects. Recently, much emphasis has been placed on the detection and recognition of locally (weakly) affine invariant regions [79,84,93,101,109]. The rationale here is that planar regions transform according to well known laws. Successful methods rely on fixing a local coordinate system to a salient image region, resulting in an ellipse describing local orientation and scale. After transforming the local region to its canonical form, image descriptors should be well able to capture the invariant region appearance. As pointed out by Mikolajczyk and Schmid [83], the detection of elliptic regions varies covariantly with the image (weak perspective) transformation, while the normalized image pattern they cover and the image descriptors derived from them are typically invariant to the geometric transformation. Recognition performance is further enhanced by designing image descriptors to be photometric invariant, such that local intensity transformations due to shading and variation in illumination have no or limited effect on the region description. State-of-the-art methods in object recognition normalize mean intensity and standard deviation of the intensity image [75, 79, 83]. Moreover, image measurements using a Gaussian filter and its derivatives is becoming increasingly popular as a way of detecting and characterizing image content in a geometric and photometric invariant way. Gaussian filters have interesting properties from an image processing point of view, among others, their robustness to noise [40], their rotational steerability [41], and their applicability in multi-scale settings [77]. Many of the intensity based descriptors proposed in literature are based on Gaussian (derivative) measurements [35, 54, 84, 103, 104], a well engineered exponent being Lowe's SIFT descriptor [79]. Indeed, for grey-value descriptors, the detection of affine regions combined with the SIFT descriptor is demonstrated to be better than many alternatives [84].

In this chapter, we consider the extension to color-based descriptors. Color has high discriminative power; in many cases, objects can well be recognized merely by their color characteristics [20, 42, 50, 87, 113, 117]. However, photometric invariance is less trivial to achieve, as the accidental illumination and recording conditions affect the observed colors in a complicated way. Photometric invariance has been intensively studied for color features [36–38, 42, 50, 123]. Geusebroek *et al.* [46] derived a set of color invariant features based on the Gaussian derivative framework, facilitated by Koenderink's Gaussian color model. The important research question is if color-based descriptors indeed improve upon their grey-based counterparts in practise. The answer depends on the stability of the non-linear combinations of Gaussian derivatives necessary to achieve a similar level of invariance as implemented in grey-value descriptors. For instance, the values of photometric invariants are distorted when the image is JPEG compressed, as the compression distorts the pixel values and spatial layout, and more for the color channels than for the intensity. Therefore, we aim at a comparative study of local color descriptors, in comparison to grey-value descriptors.

To be precise on the scope of the chapter, there is no need to address the issue of (affine) region detection, as many well performing methods exist [56, 65, 78, 81, 83, 119, 122]. Hence, we will concentrate on descriptor performance. Furthermore, to enable a fair comparison between intensity based descriptors and color based descriptors, we demand identical geometric invariance for both intensity based features and color based features. This requirement is conveniently fulfilled by the Gaussian measurement framework.

For the evaluation of local grey-value and color invariants, we adopt the extensive methodology of Mikolajczyk and Schmid [84]. In this chapter, the authors propose the evaluation of descriptor performance by the matching of regions from one image to another image. Correct matches are determined using the homography between the two images. From [84], we adopt the measures to evaluate discriminative power and invariance. Also, we adopt variety in recording conditions, being changes of illumination intensity, of the camera viewpoint, blurring of the image, and JPEG compression. We go beyond [84] by extending this set with images recorded under different illumination colors and illumination directions. These conditions induce a significant variation in the image recording. For an illustration of images recorded under varying illumination directions, see Figure 3.1.

We extend the number of images used in the evaluation framework [84] to 26,000 images, representing 1,000 objects recorded under 26 imaging conditions. Moreover, we further decompose the evaluation framework in [84] to the level of local grey-value invariants on which common region descriptors are based. We measure the performance of photometric invariants for the detection of color transitions only. Hence, we evaluate the performance of the Gaussian grey-value and color invariant derivatives, to indicate the merit of the invariant when plugged into a region descriptor. Finally, we establish performance criteria that are specific to color invariants, indicating the level of invariance with respect to photometric variation, and evaluating the ability to distinguish between various photometric effects.



(a) Example image (b) 90° (c) 60° (d) 30°

Figure 3.1: Example object recorded under semi-hemispherical illumination, and images recorded under an illuminant at decreasing altitude angles. Illuminant azimuth is to the right of the object.

The chapter is structured as follows: In Section 3.2, we shortly overview grey-value and photometric invariants and we discuss previous work on the evaluation of grey-value image invariants, which we relate to the evaluation of photometric invariants as proposed in this chapter. Section 3.3 describes the invariant features used in our comparison. Section 3.4 discusses the performance measures and the datasets, and presents the experimental results. For a realistic application of the invariants, we evaluate the performance on the VOC dataset [134] in Section 3.5. Conclusions are drawn in Section 3.6.

3.2 Previous Work

3.2.1 Grey-value Invariants

Many techniques for the description of images have considered local features. Methods based on local intensity values in the image, see e.g. [62, 131], are successfully applied to image matching. A considerable step forward was the work by Schmid and Mohr [104]. They combined Gaussian derivative measurements in a multi-scale and rotation invariant descriptor. The Gaussian derivatives were computed at Harris corner points [54], achieving general recognition under occlusion and clutter. The choice for the Gaussian filter was fundamental in there method, allowing their descriptor to capture the local differential structure of the image [67] such that scale-invariance was achieved.

To identify an appropriate and consistent scale for Gaussian-based image measurements, Lindeberg [77] determined local maxima over scale. This scheme determines the characteristic scale for the local differential image structure, and has been successfully applied to detect keypoints [79] and multiscale Harris detectors [83]. To achieve invariance to affine planar transformations, Lindeberg and Gårding [78] considered a local affine adaptation. Such an affine adaptation has recently been incorporated in Harris-affine and Hessian-affine detectors [83].

The use of the local Gaussian differential structure has received considerable interest. Gaussian derivative based descriptors have been proven to be very distinctive for matching, see e.g. [5, 7, 102]. Schiele and Crowley [103] modelled differential structure across an image by accumulating image derivatives into histograms, effectively capturing texture information. Belongie *et al.* [9] accumulated image derivatives in a regional grid with multiple bins to model both shape and location information, resulting in the so-called shape-context. Varma and Zisserman [120] modelled texture appearance by accumulation of the Gaussian-based MR8 filterbank. Winn *et al.* [125] are using a Gaussian filterbank for object recognition by a visual dictionary approach.

The most successful local image descriptor so far is Lowe's SIFT descriptor [79]. The SIFT descriptor encodes the distribution of Gaussian gradients within an image region. The SIFT descriptor is a 128-bin histogram that summarizes local oriented gradients over 8 orientations and over 16 locations. This represents the spatial intensity pattern very well, while being robust to small deformations and localization errors. Nowadays, many modifications and improvements exist, among others, PCA-SIFT [66], GLOH [84], Fast approximate SIFT [52], and SURF [8]. These region-based descriptors have achieved a high degree of invariance to overall illumination conditions for planar surfaces. Although designed to retrieve identical object patches, SIFT-like features turn out to be quite successful in bag-of-feature approaches to general scene and object categorization, see e.g. [64].

3.2.2 Photometric Invariants

Color invariants have received extensive theoretical and experimental treatment, due to the additional discriminative power that comes with color information in comparison to grey-value information. Additionally, color information enables one to distinguish between true color variation and photometric distortions, as pointed out by Gershon [45]. Indeed, for color information to be useful, Slater and Healey [110], Finlayson [36], and Gevers and Smeulders [50], have all stressed the importance of achieving invariant color measurements to varying lighting conditions such as a change in illumination color, illumination direction, or camera viewpoint.

Photometric invariants can be derived from the physical laws of light reflection. Methods which normalize mean intensity and standard deviation for grey-value descriptors are assuming Lambert's law of light reflection, $I = \rho l n$. Here, the observed image I is the result of a multiplicative formation process, for which ρ represents the surface albedo, l the light source direction and intensity, and n the surface normal. Normalization of the (local) standard deviation removes the contribution of l in the image descriptor, whereas the mean normalization counteracts the camera sensitivity offset. However, the normalized result still depends on both the surface reflectance ρ and the geometry of the surface represented by it's normal n. Hence, shadow and shading edges are coded by image descriptors. This very effect causes nowadays image descriptors to be effective for planar patches only.

Color images convey more information about the image formation process, and hence may improve on the features which can be discriminated. Inspired by the success of color indexing [117], Funt and Finlayson [36, 42] use the Lambertian assumption to arrive at photometric invariant indexing of images. Although the methodology to achieve photometric invariance is essentially similar to the grey-value case outlined above, they improve in discriminative power by adding the extra color information availably from the image. Furthermore, by exploiting the extra information which comes with color, they discount the effects of shadow and shading on their image descriptor. Gevers and Smeulders [50] elaborate on this work by deriving several sets of invariants. These sets are invariant under the more complicated photometric model proposed by Shafer [106]. In this way, they arrive at features invariant for highlights and for colored illumination. In consecutive work [51], shadows, highlights, and true color boundaries are separated in practise, based on a pixel-wise comparison of invariant values.

Geusebroek *et al.* [46] extended photometric invariance to Gaussian-based derivatives, facilitated by Koenderink's Gaussian framework. Hence, effectively combining photometric color invariance with the highly successful Gaussian geometric invariants. The pixel-based invariants can still be represented by considering the limiting case of the spatial scale for the Gaussian filters small, such that single pixels are covered. However, tuning the filters to a larger scale allows for the more interesting class of geometric *and* photometric invariant features.

Promising recent methods aim at combining color and shape description of the local neighborhood. Mindru *et al.* [86] have considered color moments, which are invariant to illumination color. However, in [84], local moments-based descriptors were found to be relatively unstable. Van de Weijer and Schmid [124] augmented the SIFT descriptor with a histogram of photometric invariant values, effectively combining color and shape information. They have shown that adding color information to the SIFT descriptor improves its discriminative power. Likewise, Geodeme *et al.* [44] have used localized color moments to reduce *a posteriori* the mismatches of SIFT descriptors. Other recent approaches have altered the SIFT descriptor itself. Abdel-Hakim and Farag [1] have based the SIFT descriptor on the hue gradient rather than the intensity gradient. Bosch and Zisserman [13] have computed SIFT from the HSV representation to provide a richer descriptor. Unfortunately, the improvement in performance for such descriptors is unclear, as no well established evaluation method is available for color based descriptors.

3.2.3 Performance Evaluation

For the evaluation of discriminative power of local descriptors, an extensive evaluation framework has been proposed by Mikolayzcyk *et al.* [84, 85]. They aimed at evaluating the different stages of an nowadays object recognition framework, by decomposing the benchmark in the separate evaluation of keypoint detection and local image descriptors. Furthermore, they realized the importance of evaluating robustness against geometric and photometric distortions of the target image. They evaluate the discriminative power and invariance of descriptors over various imaging conditions. Discriminative power for any of the detector-descriptor combinations is evaluated over: illumination intensity, of the camera viewpoint, blurring of the image, and JPEG compression. Invariance is measured by the performance degradation over increasingly hard imaging conditions, e.g. increasing JPEG compression rates. Moreels and Perona [88] elaborated on this framework by considering descriptor evaluation for 3D objects, and included three different lighting directions in their setup.

3.2.4 The Contribution of This Chapter

With the increasing interest in distinctive and robust local features, we propose in this chapter a benchmark for the evaluation of local color invariants. The contribution of this chapter is three-fold:

- We establish a framework for the evaluation of color image descriptors, including a suitable dataset and three measures of performance: discriminative power, constancy under irrelevant image distortions or imaging conditions, and the ability to distinguish true (object) variation from irrelevant (photometric) variation.
- We include color information in SIFT descriptors, and propose three color SIFT methods each having different characteristics with respect to photometric variation.
- We evaluate the performance of these descriptors together with the performance of the Gaussian color invariants on which they are based. We compare with alternative color SIFT implementations from literature.

Regarding the first contribution, we adopt the setup from [84, 85] to evaluate descriptor performance over increasingly hard imaging conditions. We consider the ALOI database [47] to match regions that are computed from 26, 000 images of 1, 000 objects in total. Ground truth is obtained by manual selection of stable Harris-affine regions inside the objects. The dataset contains both image transformations as well as photometric variation in imaging conditions, and is considered more suitable for evaluation of color descriptors than the original database proposed by Mikolayzcyk *et al.* [84, 85] or the one proposed by Moreels and Perona [88]. For example, the database contains six different lighting conditions and, very important for assessing color descriptors, variation in illumination color. Hence, allowing the assessment of color constancy for color image descriptors.

With respect to our second contribution, we will include the Gaussian color invariant gradients proposed in [46] into the SIFT descriptor [79]. We will evaluate their performance with respect their grey-value counterparts, and with respect to color SIFT descriptors from literature [1, 13].

Finally, our third contribution further decomposes the evaluation framework proposed in [84, 85]. Mikolayzcyk *et al.* evaluate discriminative power and invariance of region descriptors. SIFT-based descriptors consist of a set of Gaussian derivative image measurements and a well-designed histogram description thereof. The performance of the Gaussian filter and the non-linear combinations to obtain geometric invariance are well known and taken for granted. However, for photometric invariance, non-linear combinations may significantly alter its performance. Hence, we decompose the benchmark proposed by Mikolayzcyk *et al.* further in order to address this issue separately. We abstract from the descriptors here, and evaluate the underlying, local invariants only. The discriminative power and invariance will be established for local grey-value invariants, and for the Gaussian color invariants of [46]. Furthermore, following [51], we will assess the power of an invariant to distinguish object color variation from photometric variation.

3.3 Invariants

We will evaluate the performance of Gaussian-based invariant features. For completeness, and to introduce notation, we shortly rehearse grey-value differential invariants and color invariants in this section.

3.3.1 Grey-value Invariants

We denote a grey-value image $E(x, y)$, with a scalar value at pixel location (x, y). The filtering of a grey-value image by an (isotropic) Gaussian $G^\sigma(x, y)$ at scale σ is given by (leaving out pixel position parameters): $\hat{E}^\sigma = E * G^\sigma$, where $*$ is the convolution operator. The notational use of the hat symbol $(\hat{\cdot})$ implies dependence on the scale parameter σ, hence we leave the scale parameter out in the following and simply use $\hat{\cdot}$. More generally, we consider the filtering of an image $E(x, y)$ by a Gaussian filter G and its x- and y-derivatives,

$$\hat{E}_j = E * G_j \ , \tag{3.1}$$

where subscript $j \in \{\emptyset, x, y\}$ indicates either smoothing or spatial differentiation.

The gradient is a rotation invariant derivative measurement, given by

$$\hat{E}_w = \sqrt{\hat{E}_x^2 + \hat{E}_y^2} \ . \tag{3.2}$$

Normalizing each gradient value by the local intensity suppresses regional intensity variations [46],

$$\hat{W}_w = \frac{\hat{E}_w}{\hat{E}} \ . \tag{3.3}$$

3.3.2 Color Invariants

We consider the color-based photometric invariants from [46], which are derived from the Gaussian opponent color model. First, we recap this color model. Three opponent colors are obtained per pixel: $E(x, y)$, $E_\lambda(x, y)$ and $E_{\lambda\lambda}(x, y)$, representing respectively the intensity, the yellow-blue channel, and the red-green channel. These color channels are obtained per pixel directly from RGB values according to a linear transformation [46]. The transformation effectuates the decorrelation of RGB values.

Gaussian (derivative) filtering and construction of the gradient for each opponent color channel is similar to the grey-value case. The color-based counterpart of Equation 3.2 becomes

$$\hat{E}_{\lambda^i w} = \sqrt{\hat{E}_{\lambda^i x}^2 + \hat{E}_{\lambda^i y}^2} \ . \tag{3.4}$$

Likewise, the color invariants $\hat{W}_{\lambda w}$ and $\hat{W}_{\lambda\lambda w}$ are a generalization of the grey-value invariant $\hat{W}_w$ from Equation 3.3,

$$\hat{W}_{\lambda^i w} = \frac{\hat{E}_{\lambda^i w}}{\hat{E}} \ . \tag{3.5}$$

Note that for $i = 0$ in λ^i, the results for Equations 3.4 and 3.5 indeed is exactly the grey-value invariants $\hat{E}_w$ and $\hat{W}_w$ (Equations 3.4 and 3.5) by the very construction of the opponent color space: the first channel ($i = 0$) is the intensity channel. The photometric invariants $\hat{W}_w$, $\hat{W}_{\lambda w}$ and $\hat{W}_{\lambda\lambda w}$ are invariant to regional variations of the intensity.

Likewise, other photometric invariants can be constructed. The invariants $\hat{W}_{\lambda^i x}$ compute first the gradient and normalize it by the local intensity later. Alternatively, the intensity normalized color values $\frac{\hat{E}_\lambda(x,y)}{\hat{E}(x,y)}$ and $\frac{\hat{E}_{\lambda\lambda}(x,y)}{\hat{E}(x,y)}$ can be differentiated with respect to x or y, which, using the chain rule for differentiation, yields

$$\hat{C}_{\lambda j} = \frac{\hat{E}_{\lambda j}\,\hat{E} - \hat{E}_\lambda\,\hat{E}_j}{\hat{E}^2} \ , \tag{3.6}$$

$$\hat{C}_{\lambda\lambda j} = \frac{\hat{E}_{\lambda\lambda j}\,\hat{E} - \hat{E}_{\lambda\lambda}\,\hat{E}_j}{\hat{E}^2} \ , \tag{3.7}$$

where subscript $j \in \{x, y\}$ indicates spatial differentiation. Under Lambertian reflection, the normalization of color values by the local intensity results in color values independent of the intensity distribution. Hence, $\hat{C}_{\lambda j}$ and $\hat{C}_{\lambda\lambda j}$ and their derivatives are invariant to shadow and shading. The shadow and shading invariant gradients are obtained from: $\hat{C}_{\lambda w} = \sqrt{\hat{C}^2_{\lambda x} + \hat{C}^2_{\lambda y})}$ and $\hat{C}_{\lambda\lambda w} = \sqrt{\hat{C}^2_{\lambda\lambda x} + \hat{C}^2_{\lambda\lambda y}}$.

A next step is to include the Fresnel reflectance, hence additionally modelling highlights. In this case, the local color ratio, $\frac{\hat{E}_\lambda(x,y)}{\hat{E}_{\lambda\lambda}(x,y)}$, is invariant to the intensity distribution and the Fresnel coefficient (see [46] for details). Invariance to the Fresnel coefficient implies invariance to highlights in the image. Again applying the chain rule to obtain spatial derivatives yields

$$\hat{H}_j = \frac{\hat{E}_{\lambda\lambda}\hat{E}_{\lambda j} - \hat{E}_\lambda\hat{E}_{\lambda\lambda j}}{\hat{E}^2_\lambda + \hat{E}^2_{\lambda\lambda}} \ , \tag{3.8}$$

where subscript $j \in \{x, y\}$ indicates spatial differentiation. This yields the gradient $\hat{H}_w = \sqrt{\hat{H}^2_x + \hat{H}^2_y}$, which is invariant to shadow, shading and highlights.

To illustrate the gradient measurements by the photometric invariants, we combine the invariants in each of the sets $\{\hat{W}_w, \hat{W}_{\lambda w}, \hat{W}_{\lambda\lambda w}\}$ and $\{\hat{C}_{\lambda w}, \hat{C}_{\lambda w}\}$ to obtain a single value per pixel ($\hat{H}_w$ already yields a single value per pixel). The combined edge strength is measured by root of the squared sum. For W we compute $\bar{W}_w = \sqrt{\hat{W}^2_w + \hat{W}^2_{\lambda w} + \hat{W}^2_{\lambda\lambda w}}$, whereas for C we have $\bar{C}_w = \sqrt{\hat{C}^2_{\lambda w} + \hat{C}^2_{\lambda\lambda w}}$. Furthermore, we define $\bar{E}_w$ as the non-normalized combined edge strength over all color channels,

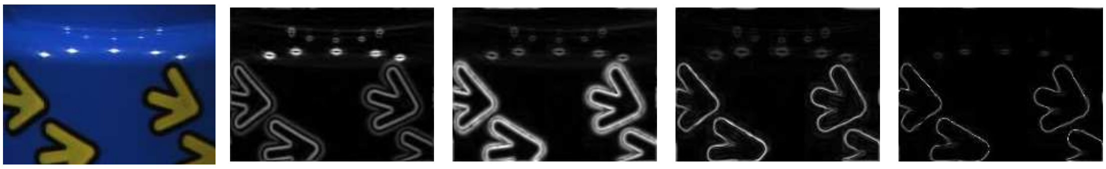

(a) Example image (b) $\bar{E}_w$ (c) $\bar{W}_w$ (d) $\bar{C}_w$ (e) $\bar{H}_w$

Figure 3.2: Photometric invariant gradients. $\bar{E}_w$ is not photometric invariant, $\bar{W}_w$ is invariant to illumination intensity, $\bar{C}_w$ is invariant to shadow and shading, $\bar{H}_w$ is invariant to shadow, shading and highlights.

that is, similar to $\bar{W}_w$ but without the local intensity normalization. The total edge strengths $\bar{E}_w$, $\bar{W}_w$, $\bar{C}_w$, $\bar{H}_w \equiv \hat{H}_w$, each illustrating one set of photometric invariants, are depicted in Figure 3.2. Note that the shading is removed by $\bar{C}_w$ (d), and that the non-saturated highlights are removed by $\bar{H}_w$ (e).

3.4 Performance Evaluation

We compare the local grey-value and color invariants based on three evaluation criteria:

- Discriminative power. We establish the power of each invariant to discriminate between image regions. Discriminative power is measured by the quality of region matching, similar to [84]. The successful matching strategy as proposed by Lowe [79], is based on the rationale that for the recognition of an object, it suffices to correctly match only a few regions of that object. In our experimental framework, we push this to the extreme, and consider the matching of one region of an object against a database of $1,000$ regions: one noisy realization of the same object matched against 999 of other objects. Under noisy conditions we consider image deformations caused by blurring, JPEG compression and out-of-plane object rotation (viewpoint change), and photometric variation induced by changes in illumination direction and illumination color. Precision and recall characteristics reflect the discriminative power of the invariant under evaluation.

- Invariance or robustness. As above, but now we establish the degradation of the number of correct matches as function of an imaging condition or image transformation which increasingly deteriorates, similar to [85]. As with discriminative power, the conditions we test are: blurring, JPEG compression, illumination direction, viewpoint change, illumination color. The degradation in the recall reflects the constancy of the invariant under examination.

- Information content. We establish the power of each invariant to discriminate between true color transitions while remaining constant under non-object related transitions induced by shadow, shading, and highlights. Hence, we assess simultaneously for each invariant its power to discriminate between color transitions, and its invariance to photometric distortions. Note that this is different from the two experiments above, as here we evaluate the property to distinct between the variant and invariant aspects in the photometric condition, in isolation of a possible effect on recognition performance.

3.4.1 Experimental Setup

We consider for $1,000$ objects from the ALOI database [46], the following imaging conditions: JPEG compression, blurring, and changes of the viewpoint, illumination direction and illumination color. Figure 3.3 illustrates the imaging conditions for some of the objects.

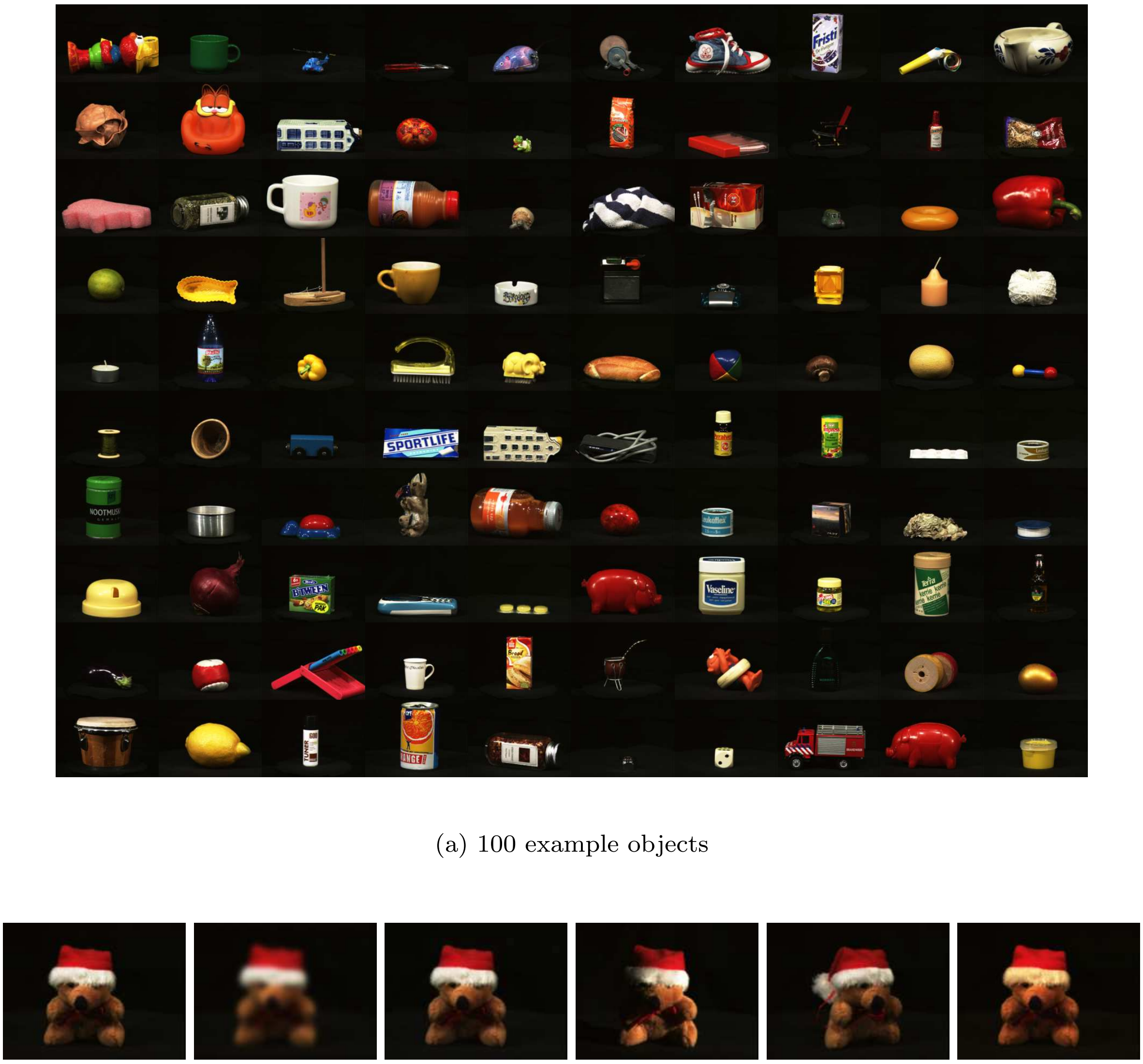
For each object image, we determine its regions. To be consistent with literature, we determine Harris-affine regions [83]. As pointed out in [84], to establish the correct matching of regions, one should either fix the camera viewpoint, or one should consider the homography limiting oneself to more or less flat scenes. For 3D objects, the assertion of a flat scene fails. To overcome this problem, we consider images that have been recorded with fixed camera viewpoint. However, the condition of viewpoint change has to be settled. Therefore, for each object, we manually selected the single region inside the object which is most consistent between the original and the image recorded under a viewpoint change. We copied the region from the original to all remaining imaging conditions, see Figure 3.4 for an example. Note that, as we are dealing with regions inside objects only, the black background does not affect the experiments. Furthermore, trying to find one region from the 1,000 selected regions could be seen as searching the one region in an image of 1,000 cluttered objects, for which all selected regions are visible. Together with the variation in image transformations and imaging conditions, a total of 26,000 regions are available. The regions vary significantly in size and anisotropy, see Figure 3.4a and b, respectively. The ground truth of regions is publicly available on the website of the ALOI database[*].

Next, we compute the invariants from each region. To be consistent with literature, we normalize the regions as in [83]. We consider two experiments:

- Single location computation. In the first experiment, we compute the invariant gradients from one location. We do so by computing them at a fixed scale (i.e. one third of the region size). For each region, we determine the location in which the image gradient $\bar{E}_w$ is maximum. For all copied regions (see for region extraction the description above), this location is identical. From this location, we compute all invariants.

- SIFT-based computation. In the second experiment, we compute the SIFT descriptor from the normalized region identical to Mikolayzcyk's computation

[*] http://www.science.uva.nl/∼aloi; will be put there as soon as the manuscript becomes publicly available.



(a) 100 example objects

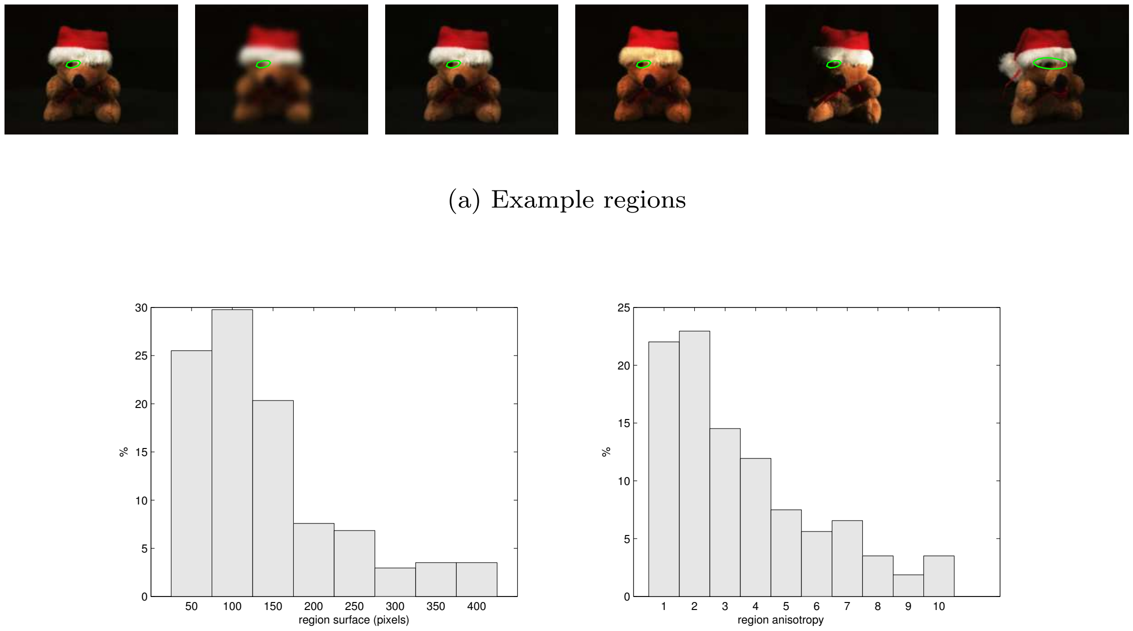
(b) Reference image and testing conditions

Figure 3.3: Randomly selected objects from the ALOI collection are depicted in (a). Imaging conditions are shown in (b), respectively: the reference image, blurring ($\sigma = 2.8$ pixels, image size 192×144), JPEG compression (50%), illumination direction change (to 30° altitude, from the right), viewpoint change (30°), illumination color change ($3075K \rightarrow 2175$K).

[84], but with the grey-value gradient inside the SIFT descriptor replaced by one of the invariant color gradients [†].

For the performance evaluation, we consider the following sets of invariant gra-

[†] software available at: http://www.science.uva.nl/~mark; will be put there as soon as the manuscript becomes publicly available.



(a) Example regions

(b) Surface

(c) Anisotropy

Figure 3.4: (a) Image regions for respectively: the reference image, blurring, JPEG compression, illumination color change, illumination direction change, and viewpoint change. For all imaging conditions except the change of viewpoint, the camera is fixed, so the regions are set identical. For the camera viewpoint change, we have manually selected the most stable region. Histogram of (b) the size of the region surfaces, and (c) of the anisotropy (where anisotropy= 1 indicates isotropy).

dients, see Table 3.4.1. The appendix `-SIFT` implicates SIFT-based computation, otherwise single location Gaussian invariants are considered. Original `SIFT` is also included in the experiments and is equivalent to `W-grey-SIFT`. To ensure results improve in discriminative power with respect to intensity based descriptors –one of our goals in adding color information–, we include the intensity gradient W_w in the H and C color based descriptors. Although this seems contradictory at first sight, the orthogonalization of intensity and intensity-normalized color information proofs effective in matching.

For fair comparison to the original SIFT descriptor, we reduce the dimensionality of all color `SIFT` descriptors to 128 numbers using PCA reduction (the covariances have been determined over 200 example regions computed from the reference images). Furthermore, we will evaluate the hue-based SIFT descriptor of Abdel-Hakim and Farag [1], termed `hue-color-SIFT`, and the HSV-based SIFT descriptor of Bosch and Zisserman [13], termed `hsv-color-SIFT`.

Table 3.1: Grey-value and color invariants

Invariant	Gradients	Property	Eq.	color-SIFT name
`E-grey`	$\{E_w\}$	Not photometric invariant	3.2	–
`E-color`	$\{E_w, E_{\lambda w}, E_{\lambda\lambda w}\}$	Not photometric invariant	3.4	–
`W-grey`	$\{W_w\}$	Invariant to local intensity level	3.3	`(W-color-) SIFT`
`W-color`	$\{W_w, W_{\lambda w}, W_{\lambda\lambda w}\}$	Invariant to local intensity level	3.5	`W-color-SIFT`
`C-color`	$\{W_w, C_{\lambda w}, C_{\lambda\lambda w}\}$	Invariant to local intensity level, plus invariant to shadow and shading	3.6	`C-color-SIFT`
`H-color`	$\{W_w, H_w\}$	Invariant to local intensity level, plus invariant to shadow and shading, and highlights	3.8	`H-color-SIFT`

Grey-value and color invariants used in the experiments.

3.4.2 Discriminative Power

The objective of this experiment is to establish the distinctiveness of the invariants. To that end, we match image regions computed from a distorted image to regions computed from the reference images as in [84]. The discriminative power is measured by determining the recall of the regions that are to be matched, and the precision of the matches:

$$recall = \frac{\#correct\ matches}{\#correspondences}\ , \tag{3.9}$$

$$precision = \frac{\#correct\ matches}{\#correct\ matches + \#false\ matches}\ . \tag{3.10}$$

Here, recall indicates the number of correctly matched regions relative to the ground truth of corresponding regions in the dataset. Precision indicates the relative amount of correct matches in all the returned matches. The definition of recall is specific to the problem of matching based on a ground truth of one-to-one correspondences, hence it deviates from the definition as used in information retrieval. The aim in our experiment is to match correctly all regions (recall of one) with ideally no mismatches (precision of one).

We consider the nearest-neighbor matching as employed in [84]. Distances between values of photometric invariants are computed from the Mahalanobis distance (the covariances have been determined over 200 examples computed from reference images). Over various thresholds, the number of correct and false matches are evaluated to obtain a recall vs. precision curve. A good descriptor would produce a small decay in this curve, reflecting the maintainance of a high precision while matching more image regions.

We randomly draw a test set of regions and use 1,000-fold cross validation to measure performance over our dataset. The number of regions to which a single region is compared is set to 20 for the invariants computed from one location. We consider a successful distinction between 20 image points to be the minimal requirement of a point-based descriptor. For the SIFT-based computation of invariants, we increase this number, as the region-based description is more distinctive. The number of regions to which one region is compared is between 100 or 500, depending on the hardness of the imaging condition. We consider a successful distinction between 100 regions to be the minimal requirement of a region-based descriptor. We consider a successful distinction between 500 regions to be sufficient for realistic computer vision tasks, this is in line with validation in [84, 85].

Experimental Results: Discriminative Power

The results of the region matching for invariant gradients are shown in Figure 3.5. The organization of all figures is as follows, see also the legends. All photometric invariants are plotted using solid lines. All color-based invariants are plotted using red lines, opposed to grey-value invariants which are plotted in black lines.

Overall, the performance of `H-color` is disappointing and apparently lacks discriminative power. Two effects play a role. First, this descriptor misses one color channel of information, and better discriminative power could be achieved when adding a saturation channel. However, in that case one would, at best, expect a performance similar to `W-color`. We will see a comparison later on when establishing performance for the color SIFT descriptors. A second effect is the instabilities caused by the normalization in the denominator of Equation 3.8. The expression becomes unstable for colors which are unsaturated, hence being greyish. Blurring by the Gaussian filter enhances this effect, as color at boundaries –which we are evaluating in this setup– are mixed. Hence, `H-color` seems unsuitable for region descriptors based on Gaussian derivatives.

Furthermore, grey-value derivatives `E-grey` and `W-grey` are outperformed by color based descriptors, except when illumination color is changed (Figure 3.5e). In that case, normalized intensity `W-grey` performs reasonable, but is still outperformed by many color based invariants.

In detail, the effect of blurring, shown in Figure 3.5a, causes the image values to be smoothed. Hence, details are lost, but no photometric variation is introduced. The color gradient with no photometric invariant properties, `E-color`, performs best. Besides the decay in performance due to additional blur, the graph clearly illustrates the gain in discriminative power when using color information.

The compression of images by JPEG, shown in Figure 3.5b, causes the color values to be distorted more than the intensity channel. Still, color information is distinctive, as the color gradient that is invariant to the intensity level, `W-color`, performs best. At the beginning of the recall-precision curves, one clearly sees the advantage of orthogonalizing intensity and color information, as `W-color`, `C-color`, and `H-color` perform significantly better than `E-color`, for which all channels are correlated with intensity. In the latter case, all values of the SIFT descriptor will be severely corrupted by the JPEG compression. For the invariant color descriptors, the intensity channel will be relatively mildly corrupted by the compression, whereas the color channels still add extra discriminative power. Compression effects become more influential at the tail of the recall-precision curves, where one sees `H-color` to drop off quit early due to instability of the descriptor, followed by `C-color`. Although `W-color` had a slower start, it ends up doing quite well due to the more stable calculation of the non-linear derivative combination.

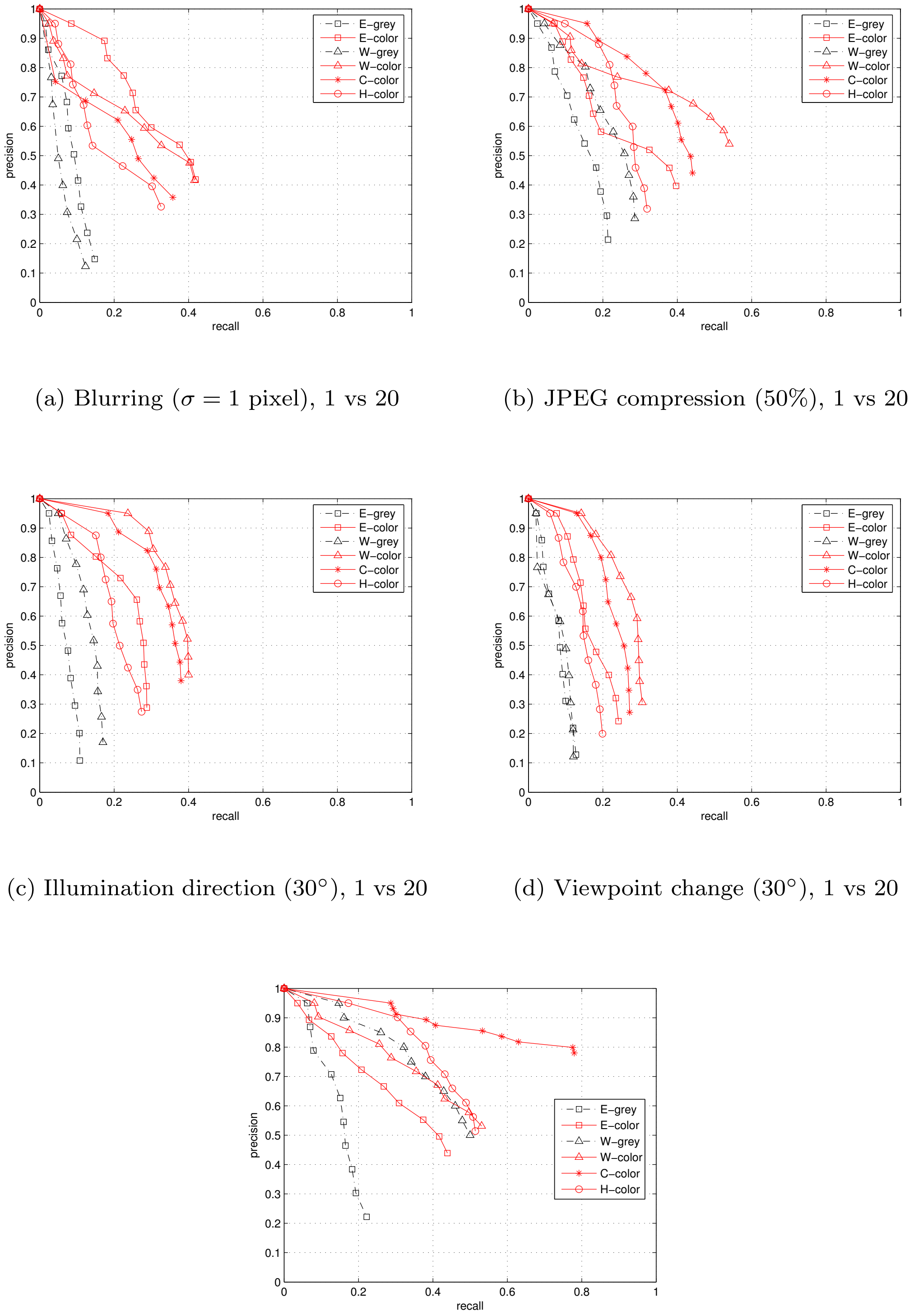
For changes of the illumination direction, Figure 3.5c, the main imaging effects are darker and lighter image patches, and shadow and shading changes. However, for the small scale at which we measure the Gaussian derivative descriptors, we expect intensity changes to dominate over shadow and shading edges. Shadow and shading (geometry) edges are expected to become more important when assessing SIFT based descriptors, which capture information over a much larger region. Hence, both color gradients that are invariant to intensity changes, `W-color` and `C-color`, are performing well. Clearly, the color invariant descriptors outperform grey-value descriptors and non-invariant color descriptors.

The results of a change in viewpoint, Figure 3.5d, clearly demonstrate the advantage of adding color information. The patches, manually indicated to be stable, merely contain a change in information content due to an projective transformation and small errors in the affine region detection. Furthermore, the light field will be distributed somewhat different over the image, causing `W-color` and `C-color` to perform superior over grey-value descriptors, non-invariant color descriptors, and the `H-color` descriptor.

For varying illumination color, Figure 3.5e, obviously the color values become distorted. The color gradient invariant to shadow, `C-color`, shows to be very robust here. Although `C-color` is based on color, its gradients are computed in such a way that can be shown to be reasonably color constant [46]. Furthermore, one would expect the grey-value descriptors not to be affected by illumination color changes. However, a change in overall intensity is also present, making direct use of `E-grey` infeasible. The intensity normalized invariant `W-grey` performs reasonable, but lacks the discriminative power which comes with the use of color.

Experimental Results: Discriminative Power for color-SIFT descriptors

Figure 3.6 shows the discriminative power of the invariants when they are plugged into the SIFT descriptor. The figure has an identical organization as Figure 3.5. The only exception in the experimental setup is that the number of regions, to which a single region is matched, is increased. This number varies over the imaging conditions,



(a) Blurring ($\sigma = 1$ pixel), 1 vs 20

(b) JPEG compression (50%), 1 vs 20

(c) Illumination direction (30°), 1 vs 20

(d) Viewpoint change (30°), 1 vs 20

(e) Illumination color (2100K), 1 vs 20

Figure 3.5: Discriminative power of photometric invariant gradients.

and is either 100 or 500, to obtain suitable resolution in the performance graphs. Furthermore, note that two extra methods from literature have been added, being the `hue-color-SIFT` descriptor [1], and the `hsv-color-SIFT` descriptor [13].

Overall, the relative performance of SIFT-based computation of invariants corresponds largely to relative performance of invariants from single points. Color-based SIFT invariant to shadow and shading effects, `C-color-SIFT`, performs best.

Generally, the SIFT-based computation improves significantly the discriminative power compared to single-point computation. Almost all color and grey-value descriptors perform well under blurring (Figure 3.6a), JPEG compression (Figure 3.6b), and illumination color changes (Figure 3.6e). Note that the `C-color-SIFT` descriptor performs equally well as the intensity based `SIFT` descriptor in the last case, implying a high degree of color constancy for this descriptor.

Discriminative power drops when considering illumination direction or viewpoint changes, see Figure 3.6a,b. These cases are much harder to distinguish using a SIFT descriptor. In these cases, the grey-value based `SIFT` is outperformed by the color-based SIFT descriptors. In particular, the color-based SIFT invariant to shadow and shading effects, `C-color-SIFT`, is very discriminative in these cases. This can be explained by the large spatial area over which the SIFT descriptor captures image structure. Hence, shadow and shading (object geometry) effects are more likely to be captured by the SIFT descriptor, but the effects being cancelled by the C invariant.

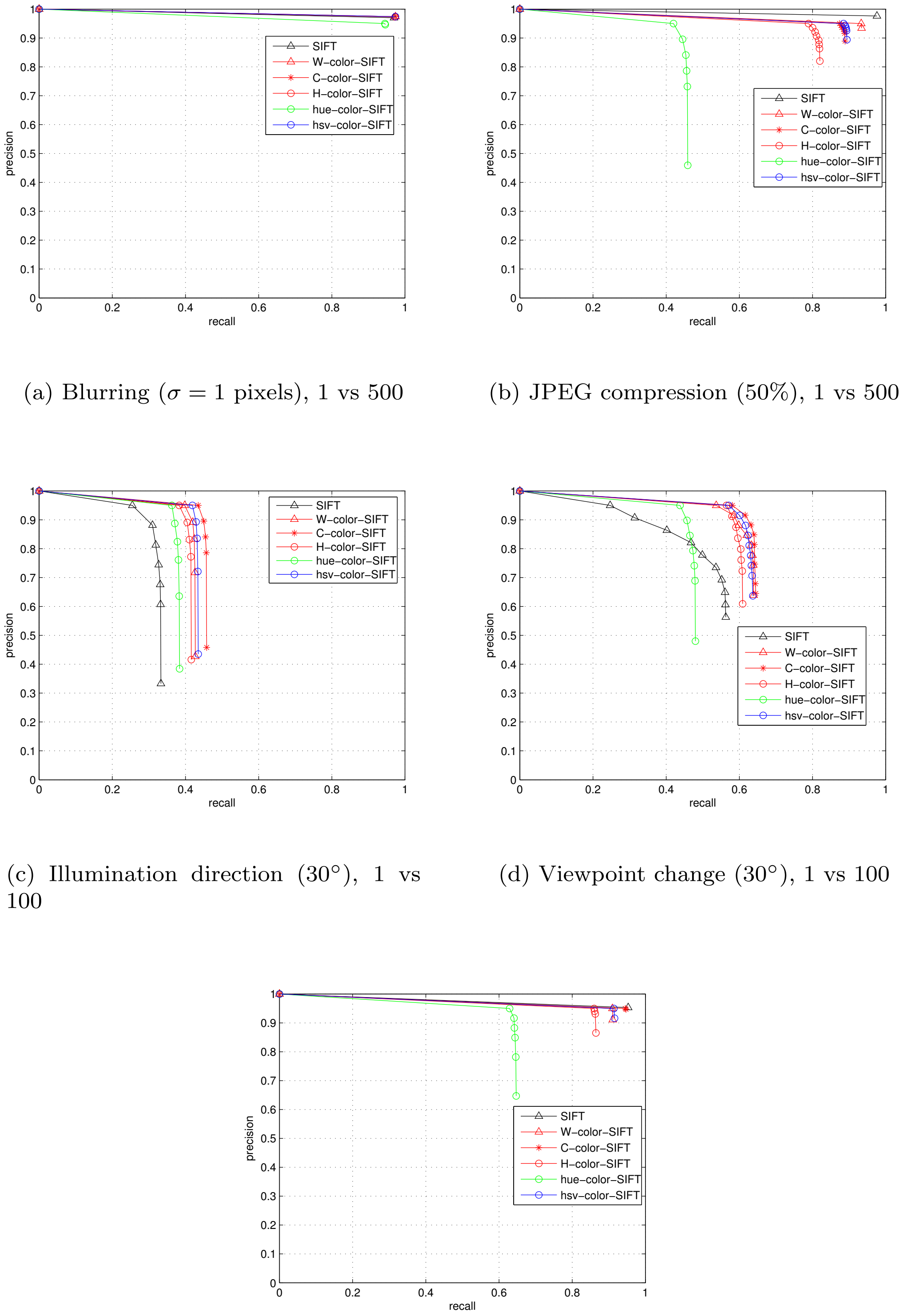
The shadow and highlight invariant `H-color-SIFT` is generally not very distinctive compared to `W-color-SIFT` and `C-color-SIFT`. Lack of discriminative power affects the performance for `hue-color-SIFT`, `H-color-SIFT`, and `SIFT` under blurring. Furthermore, the hue-based descriptors `hue-color-SIFT` and `H-color-SIFT` are affected by JPEG compression, and by illumination color changes. The distinctiveness of `hue-color-SIFT` is generally much less than of `H-color-SIFT`. Hence, using the hue alone is not a distinctive region property. The distinctiveness of `hsv-color-SIFT` is generally somewhat higher than of `H-color-SIFT`. Thus, the saturation s in the hsv color space is a distinctive property. But, the distinctiveness of `hsv-color-SIFT` is generally less than of `W-color-SIFT` and `C-color-SIFT`, due to instability as argued before.

3.4.3 Invariance

The objective of this experiment is to establish the constancy of the invariants against varying imaging conditions. Likewise [85], we measure the degradation of recall (Equation 3.9) over increasingly hard imaging conditions. The experimental setup is identical to the previous experiment. The aim in this experiment is to minimize the degradation over more distorted images.

Experimental Results: Invariance

The results of the region matching over increasingly hard imaging conditions is shown in Figure 3.7. The organization of the figure is identical to Figures 3.5 and 3.6. The present graphs are orthogonal to Figures 3.5 and 3.6, in that now the amount of



(a) Blurring ($\sigma = 1$ pixels), 1 vs 500

(b) JPEG compression (50%), 1 vs 500

(c) Illumination direction (30°), 1 vs 100

(d) Viewpoint change (30°), 1 vs 100

(e) Illumination color (2100K), 1 vs 500

Figure 3.6: Discriminative power of photometric invariant gradients, when plugged into the SIFT descriptor.

degradation is varied, at a fixed recall which corresponds to the end-point of the curves in Figures 3.5 and 3.6. Any decline in performance indicates lack of constancy with respect to the tested condition. Ideally, the decline would be zero (horizontal line), indicating perfect invariance to the set of imaging conditions.

For image blurring, Figure 3.7a, no significant imaging effects are observed. Hence, all descriptors have equal performance with respect to constancy, although initial discriminative power varies from a recall of 0.2 for grey-value derivatives to more than 0.7 for color based derivatives. For JPEG compression, Figure 3.7b, the grey-value invariants `I-grey` and `W-grey` are slightly more constant than the color invariants, as the image intensity is less affected by JPEG compression than the image chromaticity. For changes in the illumination direction, Figure 3.7c, due to the small scale of the derivative descriptors, the main imaging effect is the change of region intensity. Hence, `W-grey`, `W-color`, `C-color` and `H-color` are very stable. For a viewpoint change, Figure 3.7d, only marginal imaging effects are observed. Hence, all measures perform equally well with respect to constancy. For varying illumination color (e), besides the intensity based measures `E-grey` and `W-grey`, `C-color` is very invariant. This measure has theoretically been shown to be reasonably color constant [46].

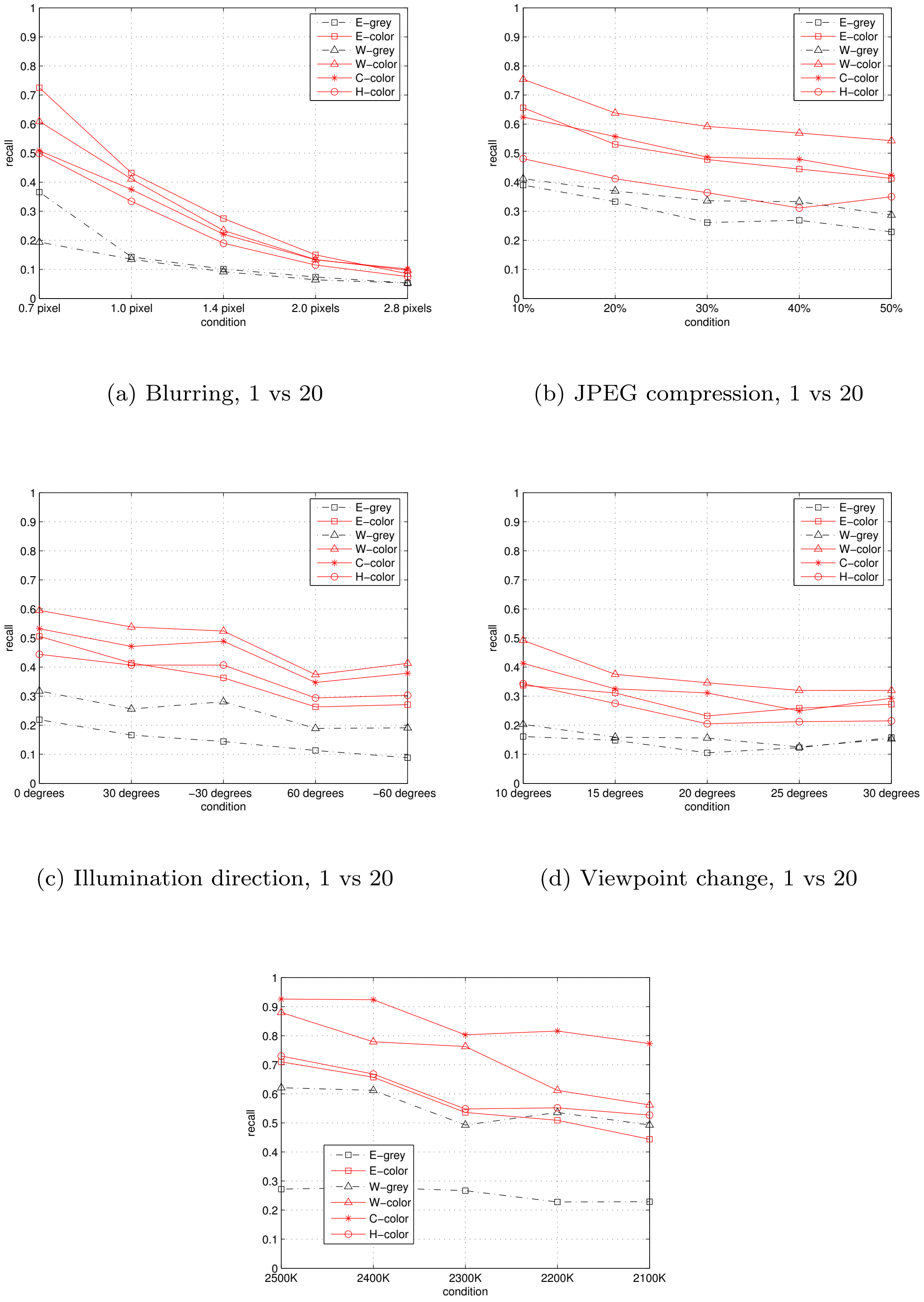
Experimental Results: Invariance for color-SIFT descriptors

We repeat the invariance experiment but now the invariants are plugged into the SIFT descriptor. The results are shown in Figure 3.8.

Overall, most descriptors are performing well for blurring (Figure 3.8a), JPEG compression (Figure 3.8b), and illumination color change (Figure 3.8e). Exceptions again are the hue based descriptors `H-color-SIFT` and `hue-color-SIFT`, which lack discriminative power, and are more affected by these conditions. A change in illumination direction or viewpoint is much harder for the SIFT descriptor to deal with, even with color invariance build in. Overall, the `C-color-SIFT` seems the best choice, for which shadow and shading edges are discounted. This descriptor has invariance comparable to the intensity based SIFT descriptor, but gains considerably in discriminative power.

3.4.4 Information Content

The objective of this final experiment is to establish the information content of the photometric invariants. Information content refers to the ability of an invariant to distinguish between color transitions and photometric events such as shadow, shading and highlights. Ideally, the invariant's values covaries with color transitions and it's value is constant to photometric events to which it is designed to be invariant. We illustrate the information content of `W-color` and `C-color`, see Figure 3.9. For the first object, new image edges are introduced by changing the illumination direction in Figure 3.9b and c. Hence, the matching is better with the shadow and shading invariant descriptor `C-color-SIFT`. Figures 3.9e and f show an example where no shadow/shading invariance performs better. Here, no new edges are introduced by the change in illumination direction, and only the local intensity is affected due to



(a) Blurring, 1 vs 20

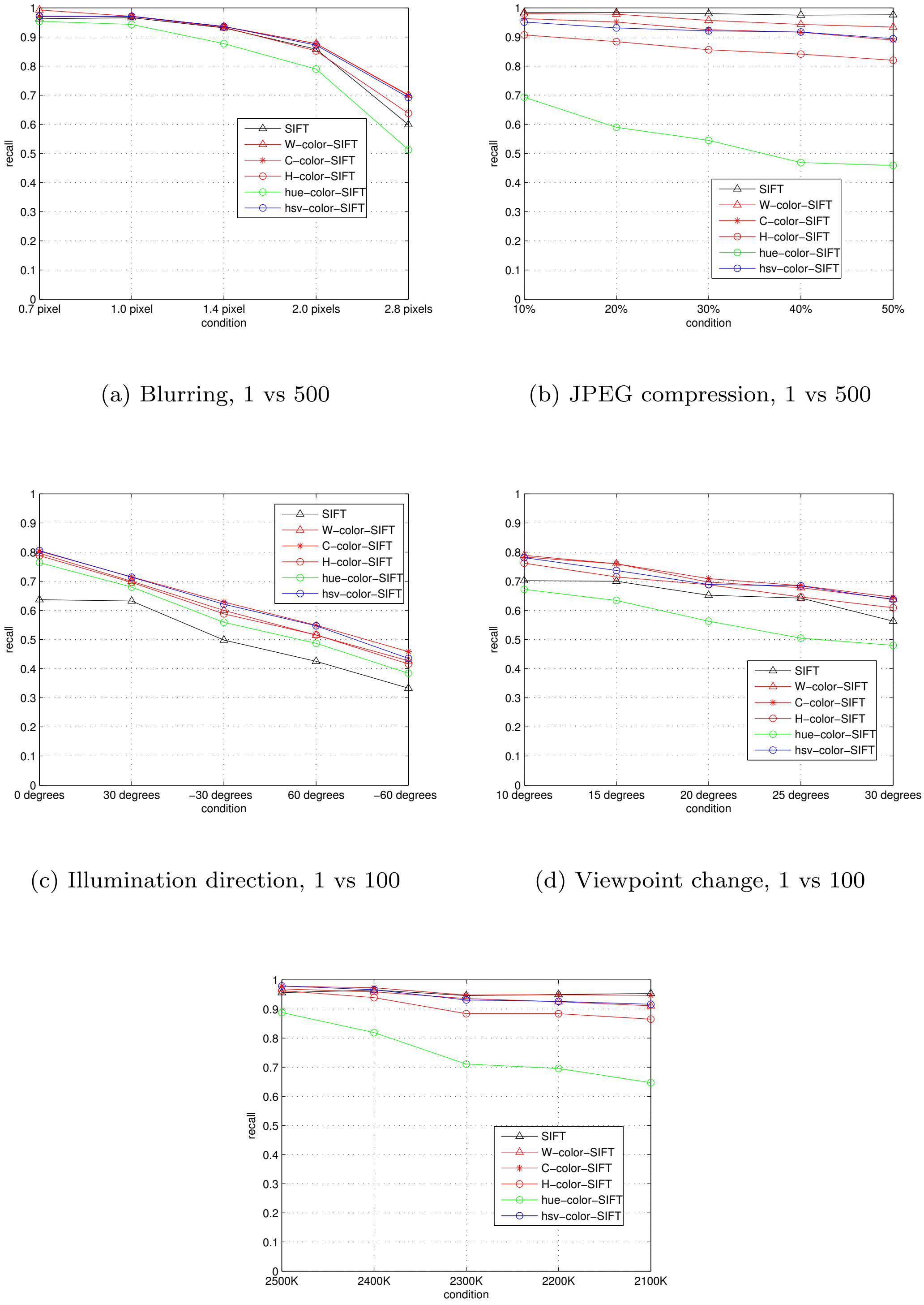
(b) JPEG compression, 1 vs 20

(c) Illumination direction, 1 vs 20

(d) Viewpoint change, 1 vs 20

(e) Illumination color, 1 vs 20

Figure 3.7: Invariance of photometric invariant gradients over increasingly hard imaging conditions.



(a) Blurring, 1 vs 500

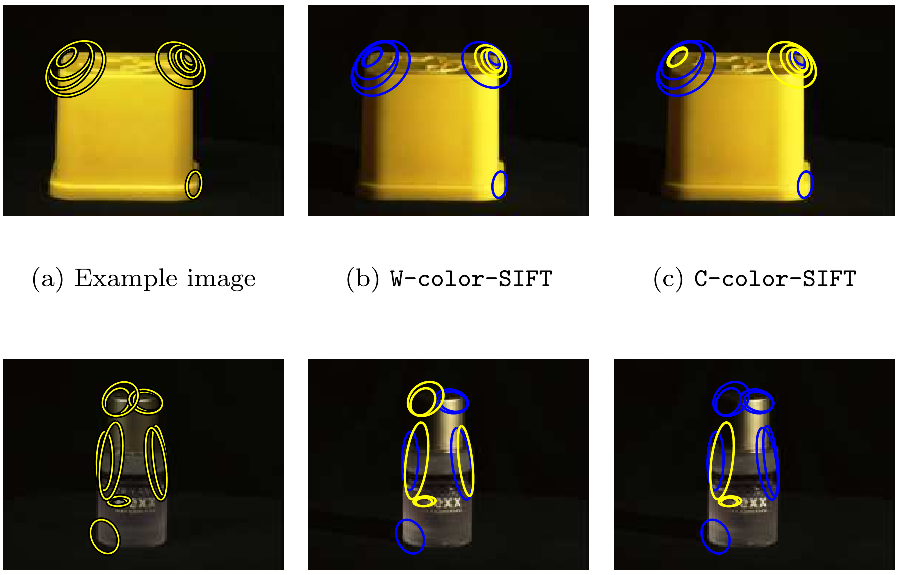
(b) JPEG compression, 1 vs 500

(c) Illumination direction, 1 vs 100

(d) Viewpoint change, 1 vs 100

(e) Illumination color, 1 vs 500

Figure 3.8: Invariance of photometric invariant gradients over increasingly hard imaging conditions, when plugged into the SIFT descriptor.



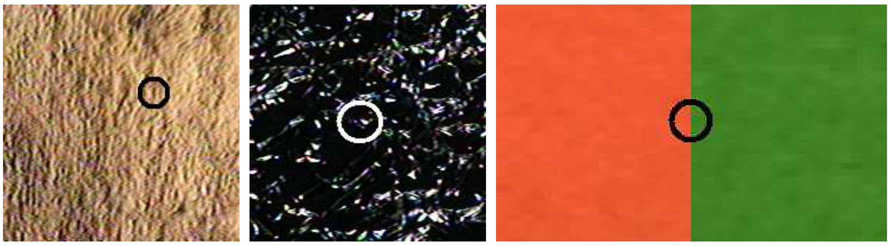
(a) Example image (b) `W-color-SIFT` (c) `C-color-SIFT`

(d) Example image (e) `W-color-SIFT` (f) `C-color-SIFT`

Figure 3.9: Illustration of matching for two objects. One is better matched with `C-color-SIFT`, the other with `W-color-SIFT`, respectively. Correct matches are shown in yellow, false matches are shown in blue.

relatively large-scale shading effects.

To establish the information content, we measure the discriminative power and invariance over individual image regions. Each image region is labelled whether it contains a color transition, or a shadow, shading or highlight transition. In this way, the information content evaluates the invariant's discriminative power and invariance over various photometric events. To that end, we construct a large annotated dataset. This dataset contains tens of images with in the order of hundreds of labelled image points located at the various photometric events. The images are selected from the CURET dataset [30]. The selected texture images contain many edges, where we annotated for each image whether the texture was generated mainly by either shadow/shading (sponge, cracker b, lambswool, quarry tile, wood b, and rabbit fur) or highlight effects (aluminium foil, rug a, and styrofoam). From these images, regions have been detected by applying a Harris corner detector [54]. Figures 3.10a and b illustrate, for two fragments of texture images, shadow/shading and highlight edges, respectively. In addition, we have collected image points located at color transitions. To that end, images have been taken from PANTONE color patches [96], see Figure 3.10c for an illustration. From the PANTONE patch combinations, we have selected the 100 combinations that have the largest hue difference, hence selecting patches which reflect true changes in object color rather than intensity or saturation differences.



(a) Shadow edges (b) Highlight edges (c) Color edges

Figure 3.10: Examples of the photometric events dataset. Detected points are given a label whether the point is located on a (a) shadow/shading edge, (b) highlight edge, or (c) color edge.

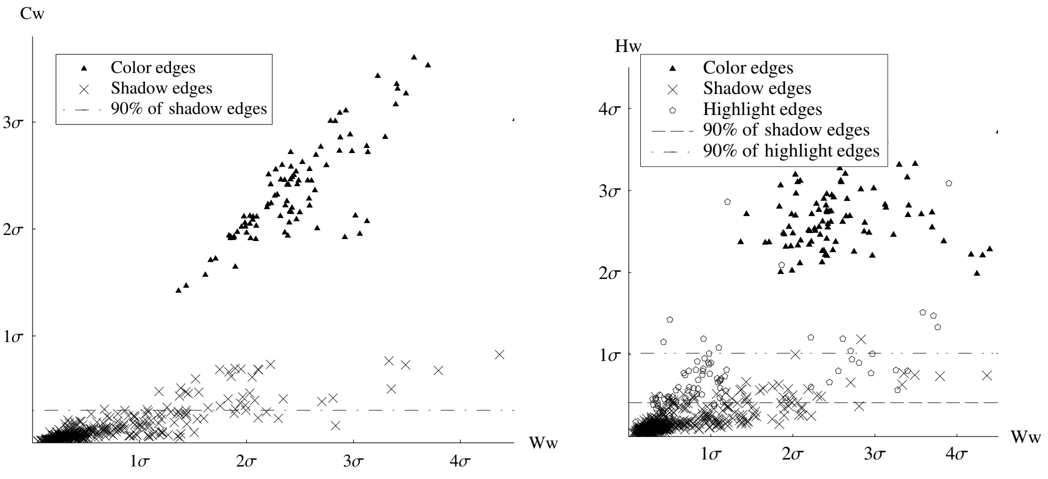
We measure an invariant's power to distinguish between color transitions and disturbing photometric events by the Fisher criterion. From many color transitions, we compute a first cloud of points; from transitions of a particular disturbing photometric event, we compute a second point cloud. The Fisher criterion expresses the separation between the two clouds of points, termed $\{x_1\}$ and $\{x_2\}$ respectively:

$$information = \frac{|\mu(\{x_1\}) - \mu(\{x_2\})|^2}{\sigma^2(\{x_1\}) + \sigma^2(\{x_2\})} \quad . \tag{3.11}$$

Experimental Results: Information Content

The values of photometric invariants to various photometric events are shown in Figure 3.11. The plots show values relative to the total color edge strength $\bar{W}_w$. We do so, to express simultaneously the power of $\bar{W}_w$ and of the shadow and shading invariants $\bar{C}_w$ and $\bar{H}_w$ to distinguish between photometric events and true color edges. As expected, the values of the invariants $\bar{C}_w$ and $\bar{H}_w$ are close to zero for shadow/shading edges (note that values of the reference invariant $\bar{W}_w$ are indeed significant to shadow/shading edges). For shadow/shading disturbances, we obtain $information(\bar{C}_w) = 2.6$, and $information(\bar{H}_w) = 4.9$. Thus, the invariant $\bar{H}_w$ separates shadow/shading from object transitions much better than $\bar{C}_w$. Furthermore, the value of $\bar{H}_w$ is also low for highlights, see Figure 3.11b. However, as expected, not all of the values are close to zero due to pixel saturations at highlights. As a result, the invariance and the information content of $\bar{H}_w$ are somewhat lower for highlight disturbances than for shadow/shading disturbances, $information(\bar{H}_w) = 2.9$.

Overall, the photometric invariant `H-color` is more constant to shadow and shading than `C-color`. Both perform well when separating color transitions from shadow and shading transitions. The separation of color transitions and highlights by `H-color` is harder due to saturated highlights. As a consequence, most of the highlights are separated well, but some highlights are misclassified as color transitions.



(a) (b)

Figure 3.11: Scatter plots of invariant values to photometric events. The figures depict (a) $\bar{C}_w$ vs. $\bar{W}_w$ and (b) $\bar{H}_w$ vs. $\bar{W}_w$. All invariants are sensitive to color edges. $\bar{C}_w$ and $\bar{H}_w$ are invariant to shadow and shading, where $\bar{H}_w$ is additionally invariant to highlights. The horizontal lines describe a 90% interval of the invariant values. This gives an indication of the invariant's ability to distinguish between values to color edges and to disturbing photometric events.

3.5 Experimental Results

In this final experiment, we evaluate the performance of the color-SIFT descriptors on the VOC dataset [134] containing 10 categories of natural and man-made objects in realistic settings. As an experimental framework, we consider the bag-of-feature approach, see e.g. [64]. We outline the approach shortly. Images are encoded by vector quantizing the appearance space by mapping descriptor vectors obtained from the image onto a codebook. The codebook contains descriptor vectors that are representative of the dataset. A common scheme is to construct the codebook by storing the cluster centers obtained from k-means clustering [27, 109]. We create codebook representations according to the method of Perronnin textitet al. [99]. They have proposed a distinctive histogram representation that is tuned to the categories to be classified. The codebook is constructed by clustering 50,000 descriptor vectors into 256 cluster centers.

It is important to notice that we deviate from [99] only in that we do not obtain cluster centers from Gaussian-mixture modelling, but from k-means. We do so for reason of speed, and also to prevent reduction of the dimensionality of the descriptors to 50 as done in [99]. As a consequence of the different clustering, a lower performance is achieved with our implementation than reported in [134]. Even though the performance may be less, our main point here is a relative performance of the grey and color-based SIFT descriptors.

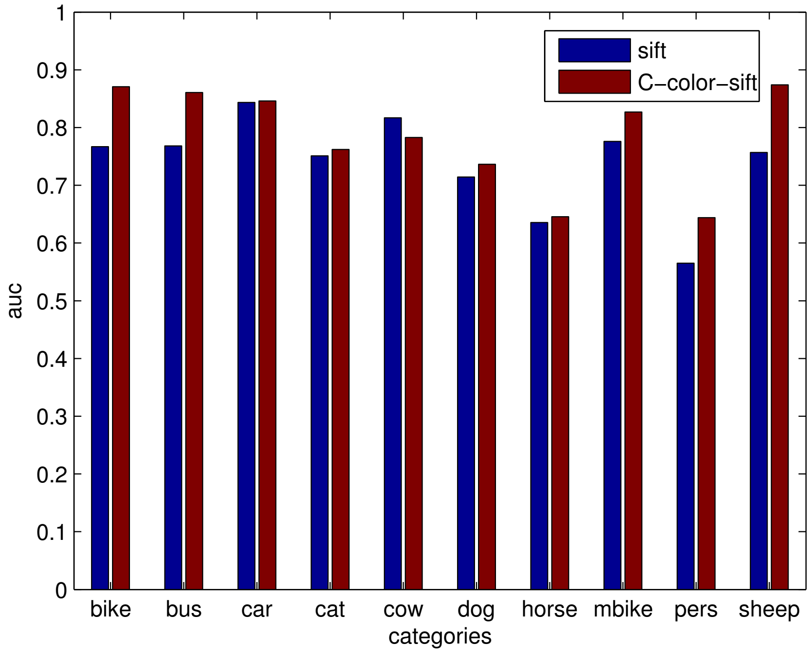


Figure 3.12: VOC classification results obtained with gray (SIFT) and color-SIFT (C-SIFT) descriptors.

The VOC dataset consists of a training, validation and testing set. We prefer the k-nearest neighbor classifier as it performs best (tested among the linear SVM, nearest mean, Fisher and logistic regression classifiers). Optimal k is determined from performance on the validation set. The performance for the SIFT and C-color-SIFT descriptors is determined from the test set. The objective is to compare qualitatively the performance of the SIFT and C-color-SIFT descriptors within a successful bag-of-feature approach.

The performance of the SIFT and C-color-SIFT descriptors for codebook-based classification is depicted in Figure 3.12. As a classification performance measure, we consider the area under the curve (auc). For the cat, car and horse categories, the classification accuracy of SIFT and C-color-SIFT is similar, while for one category (cows) the performance of C-color-SIFT is somewhat less than of SIFT (3%). For the other categories, C-color-SIFT classifies the images significantly better than does C-SIFT, up to approximately 10% improvement for the bike, bus and sheep categories. We conclude that for this realistic categorization task, the C-color-SIFT descriptor is the preferred choice over the traditional SIFT descriptor.

3.6 Conclusions

In this chapter, we have presented an experimental evaluation of local color invariants in the presence of realistic geometric transformations and photometric changes. The goal was to compare local invariants computed on regions from 3D objects. The evaluation was designed to assess performance of local invariants, which can be directly plugged into many of the descriptors that are available from literature. The setup is to evaluate of each invariant its distinctiveness, invariance, and information content. The evaluation protocol, together with test data and ground-truth, is available from the internet, allowing evaluation and comparison of future color descriptors.

We have considered the grey-value based gradient `I-grey`. The grey-value photometric invariant `W-grey` is derived from `I-grey` by locally normalizing it by the image intensity. We have considered their extensions to color, yielding `I-color` and `W-color`. Further, we have taken into account more advanced photometric invariants, being the shadow and shading invariant `C-color`, and the shadow, shading and highlight invariant `H-color`.

Our experimental evaluation showed the most distinctive color invariant to be `C-color`, which is designed to be constant to changes in illumination conditions, and to the geometry of the object. That is, shadow and shading effects are ignored. Furthermore, the invariant is reasonably color constant. Our experiments showed the descriptor to outperform alternatives with respect to discriminative power, while being more constant to illumination direction, viewpoint, and illumination color changes. Hence, the `C-color` based invariant is applicable in many computer vision tasks.

We have plugged the local invariants into the SIFT descriptor. Our experiments showed the `C-color-SIFT` based descriptor to outperform the traditional intensity based SIFT, due to it's significant increase in discriminative power, while being equally constant to the tested conditions as traditional SIFT. Furthermore, `C-color-SIFT` outperforms hue-based SIFT [1] and HSV-based SIFT [13] proposed in literature. The usefulness of `C-color-SIFT` for realistic computer vision applications is illustrated for the classification of object categories from the VOC challenge [134], for which a significant improvement is reported.

Chapter 4

Quasi-periodic Spatio-temporal Filtering*

4.1 Introduction

The temporal frequency of a moving object may be an important property of that object. Real world applications illustrate this, for instance when monitoring the oscillatory beating of a heart. Further, for periodically moving objects, the temporal frequency of the periodic motion directly relates to the velocity of the motion [23]. The velocity of waves propagating through water follows directly from its motion periodicity and its spatial frequency [108]. The velocity of waves is a direct consequence of an harmonic mechanical system, described by the wind force and the depth, width and mass of the water, which is in equilibrium. The measurement of an object's periodic motion hence may enable the estimation of both the object's velocity and environmental properties derived thereof. Estimating velocity from motion periodicity is robust, since periodicity is invariant to the object's distance. On the contrary, estimated motion from optical flow [59] varies with the object's distance. In addition, periodic motion has proven to be an attentional attribute [127], which may facilitate target detection in video (see e.g., visual surveillance in [29]).

To measure the periodicity of object motion, we propose a temporal frequency filter that measures the reoccurrence of an object's surface during a time interval. Note that the class of periodic temporal events is more rigid than the class of stochastically defined dynamic textures [32]. The temporal frequency filter cannot measure both the frequency and the timing of an occurrence of periodic motion with arbitrary precision [14]. The challenge for detecting and identifying temporal frequency is thus to find the right trade-off between timing and frequency analysis. Time-frequency analysis based on the Fourier transform of the video signal [29, 100, 118] ignores temporal discrimination. However, the Fourier transform extracts maximum information

*Appeared in *IEEE Transactions on Image Processing*, 15 (6), 1572-1582, 2006.

about the frequency composition of the signal. Gabor filtering provides the optimal joint resolution in both time and frequency, obtaining equal temporal width at all frequencies [14, 94]. Hence, the Gabor temporal frequency filter measures both the frequency identification ("what") and the frequency detection ("when").

We embed a temporal frequency filter in the Gaussian scale-space paradigm [67] to incorporate the spatial and temporal scale in its measurement. Larger spatial scales incorporate contextual information, hence avoiding pixel matching. A temporal scale allows the periodicity of object motion to be resolved in suitable time windows. For the analysis of temporal frequency, it is natural to measure the temporal signal in the Fourier domain. A Gaussian measurement in the Fourier domain, tuned to a particular frequency, boils down to a Gabor measurement [14] in the temporal domain. For online filtering, only the past is available. We deal with this restriction by a logarithmical mapping of the filter onto the "past" only [68]. However, the sinusoidal sensitivity curve of the temporal Gabor filter becomes logarithmical hence not suitable for frequency measurements. We reparameterize the temporal Gabor filter to optimize it for the local and online measurement of temporal frequency. We introduce color to increase discriminative power when measuring the reoccurrence of a particular surface.

In this chapter, we derive an online temporal frequency filter and demonstrate the filter to respond faster and decay faster than Gabor filters. Additionally, we show the online filter to be more selective to the tuned frequency than Gabor filters (Section 4.2). In color video, the filter detects and identifies the periodicity of natural motion. Further, we determine the velocity of moving gratings in a real world example (Section 4.3). We demonstrate the general applicability of the proposed filter. Consequently, we do not attribute specialized topics that analyze motion of specific kinds in depth, such as motion-based recognition [23, 100, 118]. The experiments include: (a) stable and changing periodic motion of (b) stationary and non-stationary objects with (c) smooth and regularly textured surfaces.

4.2 Temporal Frequency Filter

4.2.1 Derivation

We consider color video to be an energy distribution over space, wavelength spectrum and time. A spatiospectral energy distribution is only measurable at a certain spatial resolution and a certain spectral bandwidth [46,63]. Analogously, the temporal energy distribution is only measurable at a certain temporal resolution. Hence, physical realizable color video measurements inherently imply integration over spectral, spatial and temporal dimensions. Based on linear scale space assumptions [67], we consider Gaussian filters and their derivatives to measure color video. We generally define an i-th order Gaussian derivative filter $G_{a^i}(a)$ probing a variable a at scale σ_a and location a_0:

$$G_{a^i}^{\sigma_a, a_0}(a) = \frac{\partial^i G^{\sigma_a, a_0}(a)}{\partial^i a} = \frac{H_i(\frac{a-a_0}{\sigma_a})}{\sqrt{2\pi}\,\sigma_a} e^{-\frac{(a-a_0)^2}{2\,\sigma_a^2}}, \tag{4.1}$$

where the i-th order Hermite polynomial with respect to $\frac{a-a_0}{\sigma_a}$, $H_i(\frac{a-a_0}{\sigma_a})$, determines the shape of the i-th order Gaussian derivative filter. For orders $i \in \{1, 2, 3\}$ the Hermite polynomials H_i are given by: $\{\frac{2(a-a_0)}{\sigma_a}, 4\left(\frac{a-a_0}{\sigma_a}\right)^2 - 2, 8\left(\frac{a-a_0}{\sigma_a}\right)^3 - \frac{12(a-a_0)}{\sigma_a}\}$. For notational convenience, we omit the scale and location parameters where possible.

An object's surface is defined by its reflectance function $R(x, y, \lambda)$ at a spatial location (x, y), where λ denotes the wavelength [63]. Furthermore, the temporal periodicity of the object is measured in time t [2]. The temporal frequency measurement hence requires a simultaneous measurement of these variables to determine whether an object's surface has reoccurred at a certain spatial location. The periodic reoccurrence of an object's surface at a constant time period p is defined as:

$$\hat{E}(x, y, \lambda, t) = \hat{E}(x + x', y + y', \lambda, t + p), \tag{4.2}$$

with $\hat{E}$ the measurement of the color video signal and (x', y') the translation of the point due to object movement relative to the camera. In the sequel, we consider the temporal frequency measurement at a spatial location (x, y), and correct for the object's translation by tracking the object. The temporal frequency measurement $\hat{E}(x, y, \lambda, t)$ of the color video signal $E(x, y, \lambda, t)$ is performed by a filter $F(x, y, \lambda, t)$, yielding:

$$\hat{E}(x, y, \lambda, t) \equiv E(x, y, \lambda, t) * F(x, y, \lambda, t), \tag{4.3}$$

with $(*)$ the convolution operator as we consider linear measurements.

For convenience, we first concentrate on the measurement of the wavelength distribution. To measure wavelength in color video, we consider the advantage to separate the luminance from the color channels. The opponent color system used in this chapter is formalized by measuring with 3 spectral Gaussian derivative filters [46]: G_{λ^i}. The zeroth order derivative filter measures the energy over all wavelengths (the luminance), whereas the first order derivative filter compares the first half (blue) and second half (yellow) of the spectrum and the second order derivative filter compares the middle (green) and two outer (red) regions of the spectrum. To obtain colorimetry with human vision, the Gaussian filters are to be tuned such that the filters span the same spectral subspace as spanned by the CIE 1964 XYZ sensitivity curves. The location λ_0 and scale σ_λ of the Gaussian spectral filters are optimized such that approximate colorimetry is obtained by setting the parameters to $\sigma_\lambda = 55\text{nm}$, and $\lambda_0 = 540\text{nm}$ [46]:

$$G_{\lambda^i}^{\sigma_\lambda=55\text{nm},\lambda_0=540\text{nm}}(\lambda).i \in \{0, 1, 2\}, \tag{4.4}$$

See Figure 4.1 for the sensitivity curves of the spectral filters.
Spectral derivative filters $G(\lambda)$, $G_\lambda(\lambda)$ and $G_{\lambda\lambda}(\lambda)$ yield respectively the measurements $\hat{E}$, $\hat{E}_\lambda$, and $\hat{E}_{\lambda\lambda}$. In practice, the values are obtained by a linear combination of given RGB sensitivities [46]:

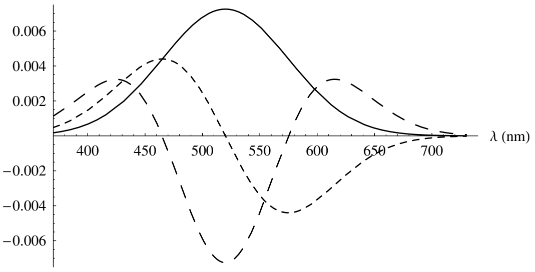


Figure 4.1: Sensitivity curves of the spectral probes $G^{\sigma_\lambda=55\mathrm{nm},\lambda_0=540\mathrm{nm}}(\lambda)$, $G_\lambda^{\sigma_\lambda=55\mathrm{nm},\lambda_0=540\mathrm{nm}}(\lambda)$ and $G_{\lambda\lambda}^{\sigma_\lambda=55\mathrm{nm},\lambda_0=540\mathrm{nm}}(\lambda)$, approximately colorimetric with the CIE 1964 XYZ sensitivity curves hence with RGB sensitivities [46].

$$\begin{bmatrix} \hat{E} \\ \hat{E}_\lambda \\ \hat{E}_{\lambda\lambda} \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} . \tag{4.5}$$

Color can only be measured by integration over a spatial area and a spectral bandwidth. Hence, a color measurement requires a combination of the spectral filter (Equation 4.4) and a spatial filter. For simplicity, we select a zeroth order, isotropic, 2-dimensional spatial filter [67]:

$$G^{\sigma_{xy}}(x,y) = \frac{1}{2\pi\,\sigma_{xy}^2}\, e^{-\frac{x^2+y^2}{2\,\sigma_{xy}^2}} , \tag{4.6}$$

where σ_{xy} indicates the spatial extent of the filter. To measure elongated shapes, we refer to oriented anisotropic spatial filters [49]. Alternatively, spatial Gabor filters [14] can be applied. Combining the spectral filters (Equation 4.4) and the spatial filter (Equation 4.6), we construct the spatiospectral filter [46] to probe the object's reflectance:

$$G_{\lambda^i}^{\sigma_{xy}}(x,y,\lambda) = G_{\lambda^i}(\lambda) * G^{\sigma_{xy}}(x,y). \tag{4.7}$$

We consider the online measurement of temporal frequency, hence we cannot access information about the "future". Consequently, only a half-axis is available: $[-\infty, t_0]$, with t_0 the present moment. Measuring in the domain $t > t_0$ violates causality; a temporal Gaussian filter has infinite extent and, consequently, is only causal over a complete t-axis. A reparameterization $s(t)$ of the time axis t is required, such that the filtering of the $s(t)$ domain with a Gaussian is uniform and homogeneous [68]. The requirement of uniform and homogeneous sampling should be independent of the unit of time. Therefore, sampling in the s-reparameterized time axis should be uniform and homogeneous for both clocks t and at, where a is a constant representing a different time scale. Now consider a periodic generator of events, of which the periodicity is estimated in the two time scales $s(t)$ and $s(at)$. On beforehand, no periodicity is more likely than any other. In other words, the probability density function (pdf) f of periodicities as a function of the reparameterized time $s(t)$ in a finite time window is a constant: $f(s(t)) = c$. Further, we require the pdf in the $s(t)$ domain, $f_1(s(t))$, and the pdf in the $s(at)$ domain, $f_2(s(at))$ to be equal: $f_1(\cdot) = f_2(\cdot)$, or, applying the substitution rule when swapping variables: $f_1(s(t)) = \frac{\partial s(at)}{\partial s(t)}.f_2(s(at))$. From the latter equation it follows that the mapping function $s(t)$ must be logarithmic [68]: $f_1(\log(t)) = \frac{\partial \log(at)}{\partial \log(t)}.f_2(\log(at)) = \frac{\partial \log(a)+\log(t))}{\partial \log(t)}.f_2(\log(a) + \log(t)) = f_2(\log(a) + \log(t))$, thus $f_1(\cdot)$ equals $f_2(\cdot)$ except for a shift $\log(a)$. Requiring $f_1(\cdot) \equiv f_2(\cdot)$ implies that both pdf's are constant. Hence, the reparameterization $s(t) = \log(t)$ satisfies the real-time requirement. Note that we do not single out any position or visible wavelength in the spatiospectral measurements. In analogy, we do not single out any *range* of time.

For temporal frequency analysis, it is natural to turn to the Fourier domain. With the logarithmic rescaling s of the time dimension t, the Fourier transform of a periodic function in t, $f(t)$, becomes in the s domain: $\int f(t(s)).\frac{\partial t(s)}{\partial s}.e^{-2\pi i t(s)\, u} ds$, where $t(s)$ is the inverse of $s(t)$, $t(s) = e^s$. Locally weighing the function f with a kernel g, to obtain a joint representation in time and temporal frequency, results in: $\int f(e^s).g(e^s - e^{s'}).\frac{\partial e^s}{\partial s}.e^{-2\pi i e^s u} ds'$, where g is the logarithmically rescaled Gaussian filter from the t domain. The Fourier transform can thus be rewritten: $\int f(e^s).G(s - s').e^s.e^{-2\pi i e^s\, u} ds'$, with $G(s)$ a Gaussian function. Thus, in the s domain we get a convolution of the e^s transformed periodic signal f with a kernel $G(s).e^s.e^{-2\pi i e^s\, u}$. Translating back to the time domain t, the kernel results in the temporal frequency filter: $G(\log(t)).t.e^{-2\pi i t u}.\frac{\partial s(t)}{\partial t} = G(\log(t)).e^{-2\pi i t u}$. In full form, we get the temporal frequency filter:

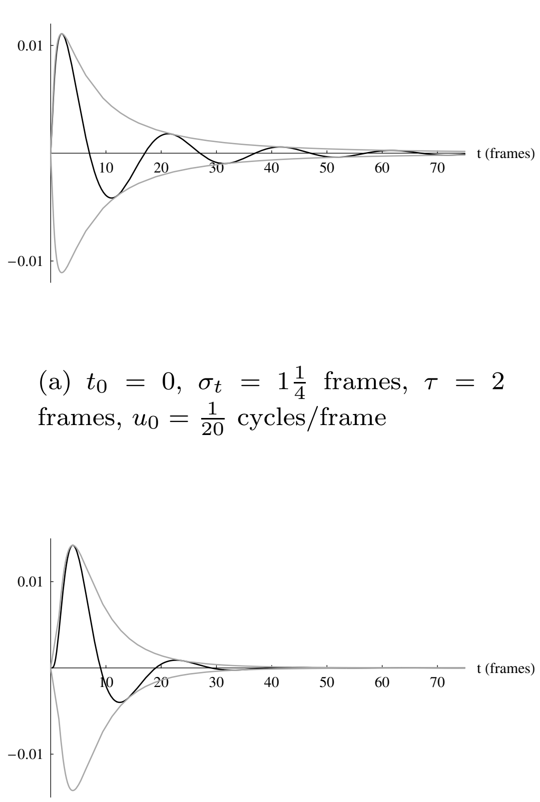
$$\tilde{G}^{u_0;\sigma_t,\tau}(t) \equiv \frac{1}{\sqrt{2\pi}\,\sigma_t}\, e^{-\frac{\log(\frac{t_0 - t}{\tau})^2}{2\,\sigma_t^2}}\, e^{2\pi i u_0 t}, \tag{4.8}$$

with t_0 the present moment and where τ scales the logarithmic reparameterization hence determines the position of the maximum of the temporal frequency filter. The scaling of the filter determines its extent and is given by σ_t. The shape of the obtained filter resembles auditory temporal frequency filters [12, 60, 94]. Figure 4.2 depicts the temporal frequency filter, together with its logarithmically rescaled Gaussian envelope.

The combination of the spatiospectral filters (Equation 4.7) and the online temporal frequency filter (Equation 4.8) yields the online temporal frequency filter for color video:

$$\tilde{G}_{\lambda^i}^{\sigma_{xy};u_0;\sigma_t,\tau}(x, y, \lambda, t) \equiv G_{\lambda^i}^{\sigma_{xy}}(x, y, \lambda) * \tilde{G}^{u_0;\sigma_t,\tau}(t). \tag{4.9}$$

The spatial scale parameter σ_{xy} of the filter determines its spatial extent. Although dependent of the distance between the camera and the object, we will in general consider the object's surface at a coarse scale. The temporal frequency selectivity of the filter depends on the frequency tuning parameter u_0. We do not change the

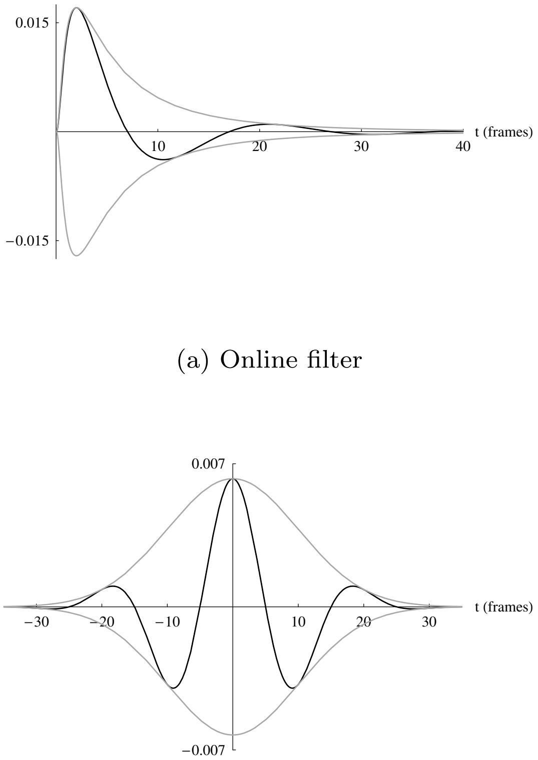


(a) $t_0 = 0$, $\sigma_t = 1\frac{1}{4}$ frames, $\tau = 2$ frames, $u_0 = \frac{1}{20}$ cycles/frame

(b) $t_0 = 0$, $\sigma_t = \frac{3}{4}$ frames, $\tau = 4$ frames, $u_0 = \frac{1}{20}$ cycles/frame

Figure 4.2: Reparameterized temporal Gabor component of the temporal frequency filter. Tuning the parameters σ_t and τ determine the shape of the temporal component; we leave the temporal frequency parameter unchanged. Note that (a) has a smaller delay than (b), but a larger temporal extent.

time unit: for the temporal scale parameter σ_t we simply choose $\sigma_t = 1$ frame. As a consequence, the other temporal scale parameter, τ, can be directly related to the tuned temporal frequency. We select the temporal scale a multiplicative of the inverse temporal frequency: $\tau = c\,\frac{1}{u_0}$ frames, with c a constant. As a result, the temporal shape of the filter does not depend on the tuned temporal frequency. Further, the effective temporal extent of the filter directly relates to the temporal frequency, based on the Nyquist theorem that frequency can only be determined if a period of the signal can be resolved.



(a) Online filter

(b) Offline filter

Figure 4.3: Temporal frequency filters for $u_0 = \frac{1}{20}$ cycles/frame. The time windows of the online and offline filter differ due to the delay $\tau = 2$ frames of the online filter (a). Note the resemblance in the shapes of the online and offline filter for the past time axis, while the constraint of online filtering is fulfilled. The integral of the filters is normalized to unity, which for the online filter yields a maximum of approximately twice the maximum of the offline filter. Consequently, the online filter will have a faster and higher response than the offline filter.

4.2.2 Properties

For color video that is integrally available, the temporal frequency filter is not restricted to a half time axis. As a consequence, the temporal frequency filter does not have to be reparameterized and periodicity can be measured by a temporal Gabor filter, see Figure 4.3 (b). We consider the properties of the offline and online temporal frequency filters, which have different shapes (see Figure 4.3; the filters have identical parameters).

The temporal frequency measurement is a local correlation of the periodical color video signal and the temporal Gabor filters. The response of the online filter is asymmetric in time with a fast rise and a slow decay. The online filter hence provides

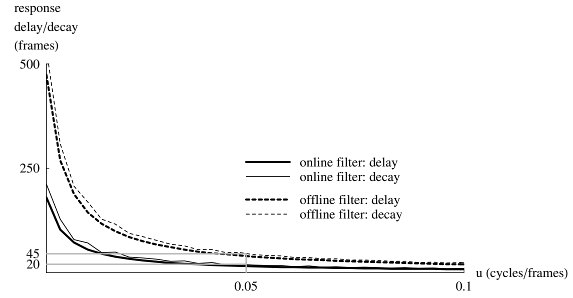


Figure 4.4: Response delays (thick lines) and decays (thin lines) of online (solid lines) and offline (dotted lines) temporal frequency filters tuned to signals with various frequencies. The online filters respond and decay approximately after one period of the signal plus the delay of the filter τ, see indication at $u_0 = 0.05$ cycles/frame $\equiv$ 20 frames/period. In contrast, the offline filters respond and decay approximately after 2.25 periods, being 45 frames at $u_0 = 0.05$ cycles/frame. Hence, the online filter reacts significantly faster than the offline filter.

a better fit to an onset of a periodic event in the video data than the offline filter. Figure 4.4 demonstrates that the online temporal frequency filters respond faster and decay faster than the offline filters. The online filters respond and decay approximately after one period of the signal plus the delay of the filter τ, see indication at $u_0 = 0.05$ cycles/frame $\equiv$ 20 frames/period in Figure 4.4. In contrast, the offline filters respond and decay approximately after 2.25 periods, being 45 frames at $u_0 = 0.05$ cycles/frame. Hence, the online filter reacts significantly faster than the offline filter.

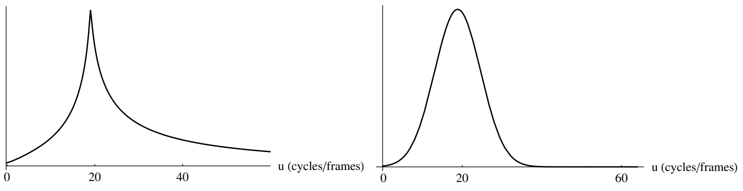
To determine the temporal frequency selectivity of the filters, we turn to the Fourier domain. See Figure 4.5 for the Fourier transforms of the online and offline filter. The online filter is not well localized in the Fourier domain. As a consequence, the online filter yields a low response to higher frequencies than the frequency it is tuned to. However, the Fourier transform of the online filter shows a narrow peak at the tuned frequency. Hence, the online filter is more narrowly tuned to frequencies than the offline filter.

The narrow frequency selectivity of the online filter is demonstrated in Figure 4.6. The online filter bank is tuned to dense temporal frequencies. We relate the discrimination quality of the filter bank to the variance of its responses. The variance of the online temporal frequency filter bank is lower than the variance of offline filter responses.

We conclude that the online temporal frequency filter achieves higher acuity as it 1) responds and decays faster and 2) can be narrowly tuned to a particular frequency.

4.2.3 Algorithm

In the sequel, we define the online temporal frequency measurement for a particular color channel, $\hat{E}_{\lambda^i}^{\sigma_{xy};u_0;\sigma_t,\tau}$, as the magnitude of the complex response of the filter. Filter responses to different color channels are combined by considering their magnitude:



(a) Online filter (b) Offline filter

Figure 4.5: Fourier transforms of the online (a) and offline (b) temporal frequency filters from Figure 4.3. Note the narrow peak and the heavy tail of the online filter, compared to the Gaussian shape of the Fourier transformed offline filter.

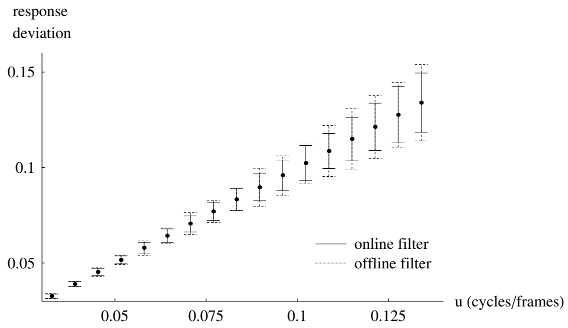


Figure 4.6: Frequency selectivity of the online (solid bars) and offline (dotted bars) filter. Frequency selectivity is derived from the variance of the responses of a bank of online and offline filters tuned to dense frequencies (the bars indicate a magnification of 200 times the variance of responses).

$$\hat{E}^{\sigma_{xy};u_0;\sigma_t,\tau} = \sqrt{\sum_{i=0}^{2}(\hat{E}_{\lambda^i}^{\sigma_{xy};u_0;\sigma_t,\tau})^2}. \tag{4.10}$$

We consider multiple temporal frequency filters, tuned to dense but fixed frequencies, ranging from $\frac{1}{75}$ to $\frac{1}{7}$ cycles per frame. To prune the filter bank, for instance to a range of temporal frequencies that was observed last, a gradient ascent method may be used, taking a filter's response (i.e., its correlation with the signal) as input. Further, the filter is parameterized with a spatial scale. In the experiments, we will preselect a particular spatial scale dependent of the size of the object and the "smoothness" of its motion. Alternatively, the scale of the spatial filter may be derived from scale selection. A common practice is to select the scale according to the maximum of the Laplacian filter [77].

The response of the temporal frequency filter is inherently delayed, depending on the temporal shape of the filter. Responses of filters tuned to lower temporal frequencies are longer delayed. In the experiments, we will both illustrate the delays of different filters and responses where we have aligned filter response delays. The temporal frequency filters primarily respond at half-periods of the periodic motion, with alternating magnitudes. We therefore integrate the filtering result over a past time window of one period of the filter. Further, we normalize this integration for the size of the time window. We assume the reoccurring surface to have a large spatial extent. Therefore, we spatially pool the responses of the temporal frequency filter. The pooled measurements are thresholded to determine periodicity detection. We identify the frequency as the tuned frequency u_0 of the filter that, after spatially pooling by summation, yields the maximum. As a consequence, we constrain ourselves to the periodic motion of one object. Further, we assume that the maximum spatially pooled response is representative of the periodicity of the object under investigation. Segmenting a frame based on spatially localized responses of temporal frequency filters would overcome the problem of measuring motion periodicity of multiple objects. In the experiments, we will constrain ourselves to demonstrating the robustness of the temporal frequency filter for both stationary and nonstationary single objects moving periodically and quasi-periodically.

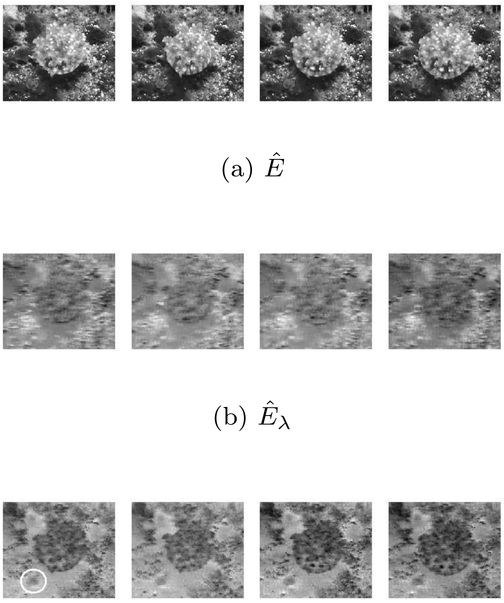
4.3 Application to Color Video

In this section, we apply a bank of temporal frequency filters to color video of natural scenes. We consider: (a) stable and changing periodic motion of (b) stationary and non-stationary objects with (c) smooth surfaces and regularly textured surfaces (gratings). For all experiments, the color video is recorded by a RGB digital video camera (JVC GR-D72) at 768×576 pixels video frame transfer sampled at 25 frames per second.

4.3.1 Periodic Zoological Motion

In this experiment, we *detect* the temporal frequency of the periodic motion of two (stationary) anemones.

Figure 4.7 shows a fragment of color video of the periodic motion of a large anemone and a small anemone. The frames are shown in increasing order, from left to right, indicating quarter-periods of the motion of the large anemone. The frames are represented by the 3 color channels $\hat{E}$ (a), $\hat{E}_\lambda$ (b) and $\hat{E}_{\lambda\lambda}$ (c). The large anemone is located in the center and visible in all color channels (a-c). The small anemone is located in the lower left region, and is only visible in the "green-red" opponent color channel $\hat{E}_{\lambda\lambda}$ (c, indicated with a circle). The large anemone moves periodically at $\frac{1}{18}$ cycles per frame, whereas the small anemone moves periodically at $\frac{1}{8}$ cycles per frame. Note that the area over which the anemones move are marginal, which makes the detection of the anemones' periodic motion non-trivial.



(a) $\hat{E}$

(b) $\hat{E}_\lambda$

(c) $\hat{E}_{\lambda\lambda}$

Figure 4.7: Color video of the periodic motion of a large anemone and a small anemone. The frames are shown in increasing order, from left to right, indicating quarter-periods of the motion of the large anemone. The frames are represented by the 3 opponent color channels $\hat{E}$ (a), $\hat{E}_\lambda$ (b) and $\hat{E}_{\lambda\lambda}$ (c). The large anemone is located in the center and visible in all color channels (a-c). The small anemone is located in the lower left region, and is only visible in the "green-red" opponent color channel $\hat{E}_{\lambda\lambda}$ (c, indicated with a circle). The large anemone moves periodically at $\frac{1}{18}$ cycles per frame, whereas the small anemone moves periodically at $\frac{1}{8}$ cycles per frame. Note that the areas over which the anemones move are marginal, which makes the detection of the anemones' periodic motion non-trivial.

We analyzed the frequencies of the anemones' periodic motion at a spatial scale $\sigma_{xy} = 2$ pixels to moderately smooth the signal.

Figure 4.8 again depicts a fragment of the color video of the periodic motion of the large anemone and small anemone. The frames cover 1 period of the motion of the large anemone. The frames are represented by the color channel $\hat{E}_{\lambda\lambda}$ (a), which has most discriminative power. Responses of temporal filters tuned to various frequencies are shown (b-g). The filter in (d) is tuned to the frequency of the large anemone. Its response is higher than the responses of filters tuned to temporal frequencies that are slightly lower (b and c) and higher (e and f). We emphasize the inherent delays of the filter: a filter tuned to a lower frequency has a longer delay. The response shown in (d) is higher than the threshold set to determine periodicity. This is also the case for the response of the filter tuned to the periodicity of the small anemone (g, indicated). Despite the isoluminance and weak contrast between the small anemone and its background, the proposed filter strategy was able to detect and identify its periodicity. Note that the filter responds to the periodic motion of both the large and small anemone. The ambiguity in the filter's response is caused by the approximate harmonics formed by the frequencies of the motion of the two anemones.

In the description of the algorithm (Section 4.2.3), we mentioned the integration of filter responses over a small time window. As the temporal frequency filter responds maximally at half-periods of a periodic event, integration over a half-period provides a stable response. In the sequel, we consider integrated responses. For convenience of display, we will align the filter delays with the present moment such that the responses can be compared at single time instances. Figure 4.9 (a) shows frames at full-periods of the periodic motion of the large anemone. Figures 4.9 (b) and (c) depict the integrated and aligned responses of the filters tuned to the frequencies of the large and the small anemone, respectively. Integrating the responses of a filter over a half-period of the filter provides stability, as demonstrated by the detection of the periodicity of motion in Figures 4.9 (b) and (c).

The spatial extent over which the two anemones move, pops out from the responses of the filters tuned to their periodic motion. The high responses of the temporal frequency filters evidently reflect the periodicity of the objects under investigation.

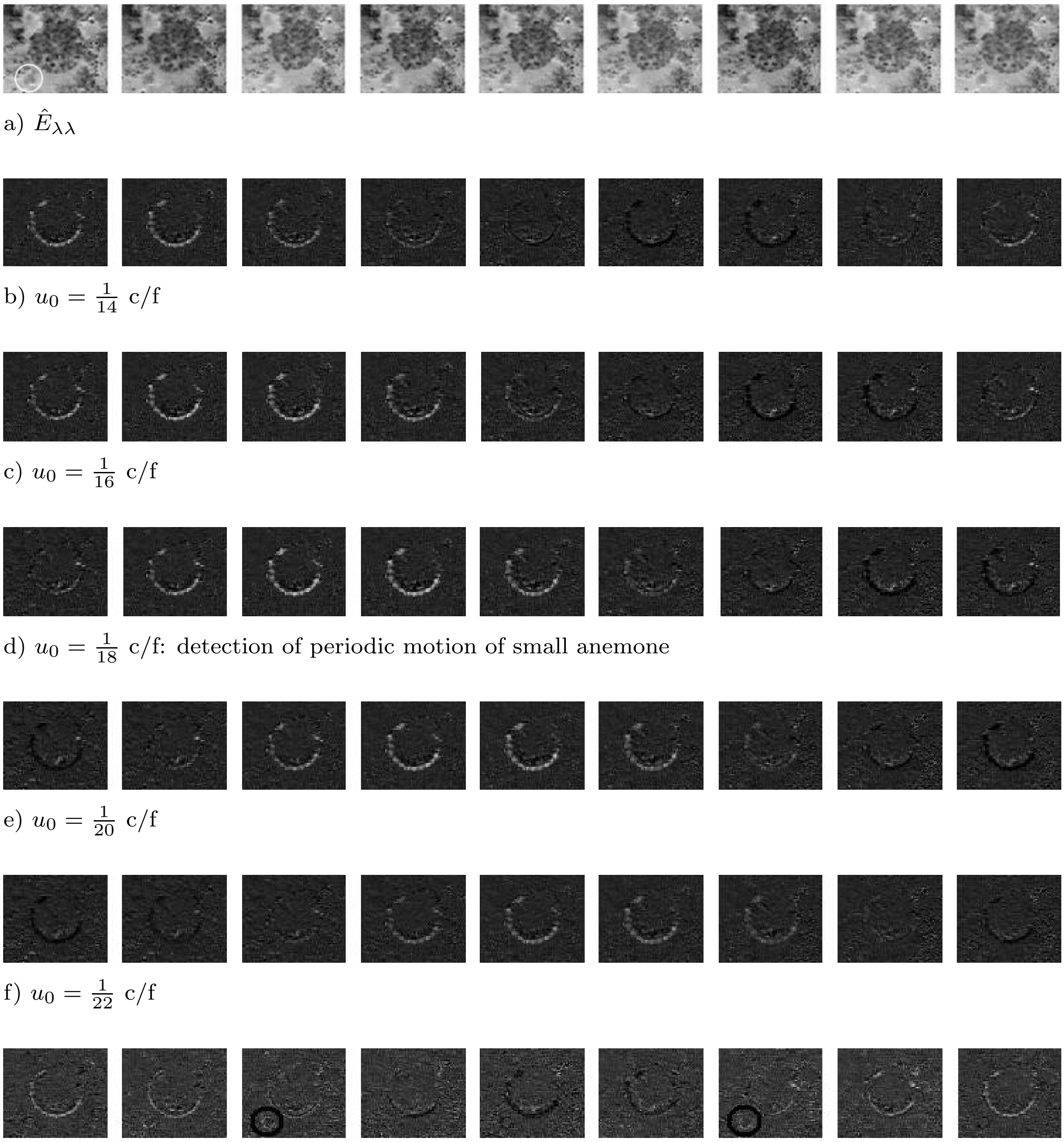
4.3.2 Periodic Animal Motion

In this experiment, we *identify* the temporal frequency of the periodic motion of a flying bird. The frequency of the bird's wings are a measure of its velocity. Figure 4.10 shows a fragment of color video of the periodic motion of a flying bird. The frames cover one period and are represented by the intensity channel $\hat{E}$, which contains most discriminative power. Note the variation in the location of the bird.

The bird's motion inherently causes a translation. To correct for the bird's translation in subsequent frames, we apply kernel-based tracking with scale adaptivity [24]. We thus exploit a prior model of the bird. For approaches that include automatic motion segmentation we refer to [90, 100, 115]. The bird's distance, hence its perceived size, is normalized by a scaling of the tracked kernel regions, before applying the online temporal frequency filter to the obtained regions. We emphasize that the following experiment's robustness to, for instance, clutter and occlusion, heavily depends on the tracking of the object. However, tracking objects is not our primary concern here, and therefore we will not elaborate on this part of the experiment. Figure 4.11 shows the tracking results at half-periods of the bird's motion.

In the frames that differ exactly one period in time (for example, images 1 and 3), the bird has not the same pose. The "misalignment" of the bird's wings is caused by the low sampling rate compared to the high frequency of its moving wings. Due to the misalignment of the bird's wings, the problem of identifying the temporal frequency of the bird's motion is not trivial.

In the sequel, we only depict the zeroth order spectral derivative measurement (i.e., the luminance) for display convenience. Further, due to the relatively large spatial extent of the bird, we analyzed the temporal frequency of its surface at a fairly large spatial scale $\sigma_{xy} = 5$ pixels. The advantage of the spatial extent of the



a) $\hat{E}_{\lambda\lambda}$

b) $u_0 = \frac{1}{14}$ c/f

c) $u_0 = \frac{1}{16}$ c/f

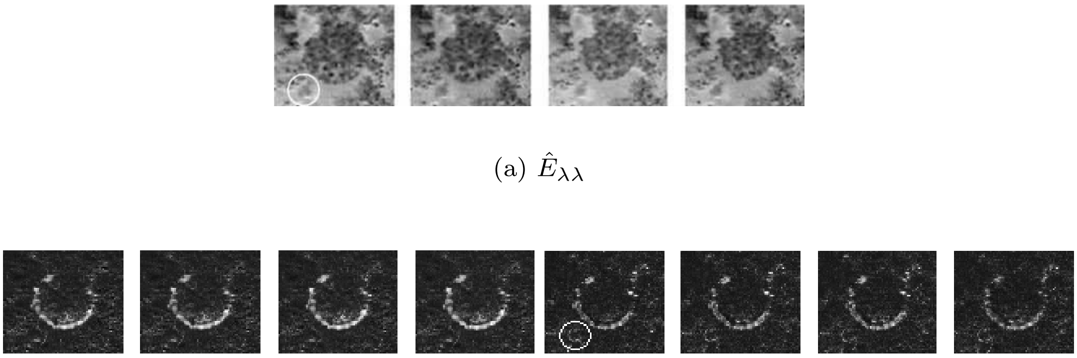
d) $u_0 = \frac{1}{18}$ c/f: detection of periodic motion of small anemone

e) $u_0 = \frac{1}{20}$ c/f

f) $u_0 = \frac{1}{22}$ c/f

g) $u_0 = \frac{1}{8}$ c/f: detection of periodic motion of small anemone

Figure 4.8: Color video of the periodic motion of a large anemone and a small anemone. The frames cover 1 period of the motion of the large anemone. The frames are represented by the color channel $\hat{E}_{\lambda\lambda}$ (a), which has most discriminative power. Responses of temporal filters tuned to various frequencies are shown (b-g), where high responses are indicated white. The filter in (d) is tuned to the frequency of the large anemone. Its response is higher than the responses of filters tuned to temporal frequencies that are slightly lower (b and c) and higher (e and f). The response shown in (d) is higher than the threshold set to determine periodicity. We emphasize the inherent delays of the filter: a filter tuned to a lower frequency has a longer delay. The response of the filter tuned to the periodicity of the small anemone (g, indicated) is higher than the threshold set to determine periodicity. Despite the isoluminance and weak contrast between the small anemone and its background in the color video fragment (a), the proposed filter strategy was able to detect and identify its periodicity.



(a) $\hat{E}_{\lambda\lambda}$

(b) $u_0 = \frac{1}{20}$ c/f

(c) $u_0 = \frac{1}{8}$ c/f

Figure 4.9: Color video of the periodic motion of a large anemone and a small anemone. The frames are randomly selected and represented by the color channel $\hat{E}_{\lambda\lambda}$ (a). Two temporal filters (b and c) are tuned to the respective frequencies of the large and small anemone. For convenience of display, we have aligned the longer delay (b) and smaller delay (c) of the filter responses with the moment at which the frames were presented. Integrating the responses of a filter over a half-period of the filter provides stability over frames within this half-period, as demonstrated by the detection of periodicity at randomly selected frames. The spatial extent over which the two anemones move is detected. See "Supplemental Material" [16] for the original color video plus overlayed responses.

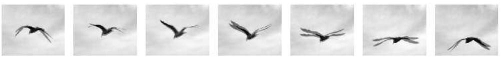


Figure 4.10: Color video of the periodic motion of a flying bird. The frames cover one period and are represented by the intensity channels $\hat{E}$, which contains most discriminative power. Note the variation in the location of the bird.

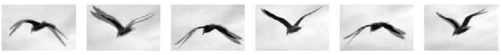
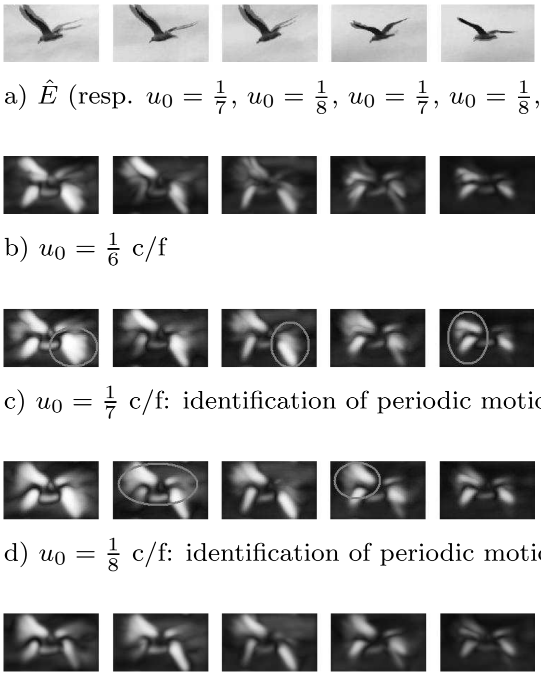


Figure 4.11: Normalized frames as a result of tracking the bird. The frames are taken at half-periods of the periodic motion.

filter is that contextual information is incorporated, making the filter robust to the "misaligned" pose of the bird (see "Supplemental Material" [16]).

Figure 4.12 (a) shows frames at full-periods of the flying bird. The temporal frequency of the bird's periodic motion changes within these samples, i.e. the motion is quasi-periodic. We annotated the temporal frequencies at the full-periods. Note that the frequencies differ only 1 frame per period. At the full-periods, we measured



a) $\hat{E}$ (resp. $u_0 = \frac{1}{7}$, $u_0 = \frac{1}{8}$, $u_0 = \frac{1}{7}$, $u_0 = \frac{1}{8}$, $u_0 = \frac{1}{7}$) c/f

b) $u_0 = \frac{1}{6}$ c/f

c) $u_0 = \frac{1}{7}$ c/f: identification of periodic motion of flying bird

d) $u_0 = \frac{1}{8}$ c/f: identification of periodic motion of flying bird

e) $u_0 = \frac{1}{9}$ c/f

Figure 4.12: Color video of the periodic motion of a flying bird. The frames represent full-periods of the motion. The frames are represented by the intensity channel $\hat{E}$. The frequency of the motion changes throughout the represented fragment. We annotated the frequency in a time window around the frames as shown in (a). Responses of temporal filters tuned to various frequencies are shown (b-e). The filters in (c) and (d) are tuned to two frequencies present in the fragment. At frames 1, 3 and 5, the filter tuned to the annotated frequency of $\frac{1}{7}$ c/f responds maximally (c). Its response is slightly higher than the responses of filters tuned to temporal frequencies that are slightly lower (b) and higher (d and e), see indications. At frames 2 and 4, the filter tuned to the annotated frequency of $\frac{1}{8}$ c/f responds maximally (d). Again, its response is slightly higher than the responses of filters tuned to temporal frequencies that are slightly lower (b and c) and higher (e), see indications. For the identification of temporal frequency of the bird's motion in color video, we refer to "Supplemental Material" [16].

and *identified* the temporal frequency of the periodic motion. The responses depicted in Figures 4.12 (c) and (d) identify alternatively the temporal frequency of the bird's motion, see the indication. Due to the small differences in the actual frequencies apparent in the bird's motion, the responses do not differ much. Nonetheless, the identification resembles the annotation. Lower responses of filters tuned to slightly different temporal frequencies are included in Figures 4.12 (b) and (f).

4.3.3 Velocity of Moving Gratings

In this experiment, we measure the *velocity* of a moving grating. A moving grating may be characterized by its orientation, spatial frequency and temporal frequency.

The grating velocity is determined by its temporal frequency divided by its spatial frequency [108]. We therefore identify both the temporal and spatial frequency.

Analogously to the temporal frequency filter derivation, we analyze spatial frequency in the Fourier domain. When translating back to the spatial domain, we obtain the 2-dimensional spatial Gabor filter [14]:

$$\tilde{G}^{\sigma_{xy},v_0,w_0}(x,y) \equiv G^{\sigma_{xy}}(x,y)\, e^{2\pi i\,(v_0 x + w_0 y)}, \quad i^2 = -1. \tag{4.11}$$

with (v_0, w_0) the frequency in cycles per pixel for 2 dimensions. The radial center spatial frequency $\sqrt{v_0^2 + w_0^2}$ is given in cycles per pixels and $\tan^{-1}(\frac{w_0}{v_0})$ represents the orientation of the filter.

We substitute the spatial component of the spatiospectral filter (Equation 4.7) by the spatial frequency component:

$$G_{\lambda^i}^{\sigma_{xy};v_0,w_0}(x,y,\lambda) = G_{\lambda^i}(\lambda) * \tilde{G}^{\sigma_{xy},v_0,w_0}(x,y). \tag{4.12}$$

For a particular color channel, we obtain the spatial frequency measurement $\hat{E}_{\lambda^i}^{\sigma_{xy};v_0,w_0}$ by considering the magnitude of the complex filter response. Note that the spatial frequency measurement does not incorporate time as we consider it at a particular moment.

The color video contains waves propagating through water. Let us define the velocity v of the waves as the ratio of the measured temporal frequency $\hat{E}^{\sigma_{xy};u_0;\sigma_t,\tau}$ and the measured spatial frequency $\hat{E}^{\sigma_{xy};v_0,w_0}(x,y,\lambda)$. Consequently, for a particular location (x_i, y_i) with reflectance λ_i at time t_i, we obtain the velocity:

$$v = \frac{\hat{E}^{\sigma_{xy};u_0;\sigma_t,\tau}(x_i, y_i, \lambda_i, t_i)}{\hat{E}^{\sigma_{xy};v_0}(x_i, y_i, \lambda_i)}. \tag{4.13}$$

The propagation of the waves has an orientation of approximately 284°, see the fragment of color video in Figure 4.13 (a). Therefore, we tuned the spatial frequency filter to an orientation of $\tan^{-1}(\frac{w_0}{v_0}) \equiv 284°$, such that the frequency parameters v_0 and w_0 yield a radial center frequency of $\sqrt{v_0^2 + w_0^2}$ cycles per pixel. Further, we selected a spatial scale of $\sigma_{xy} = 8$ pixels, to cover a sufficient area to robustly measure the occurring frequencies, which are approximately in the range of $\langle \frac{1}{14}, \ldots, \frac{1}{8} \rangle$ cycles per pixel. Responses of the oriented spatial frequency filter responses to this grating are shown in Figure 4.13 (b-g). In analogy to temporal frequency identification, the identified frequency corresponds to the frequency of the filter with a maximum spatially pooled response, see indications. For instance, the filter response in Figure 4.13 (f) to the second frame is higher than filters tuned to slightly different frequencies in Figures 4.13 (e) and (g). Assigning a maximum response to the first frame is more ambiguous: filter responses in Figures 4.13 (b), (c) and (d) seem very similar. For the first frame, the algorithm appointed Figure 4.13 (c) as the maximum response, whereas for the third frame (e) gives the maximum response.

Combining the identified spatial and temporal frequency of these moving gratings, we obtain the velocity of the grating, see Table 4.1 for randomly selected frames.

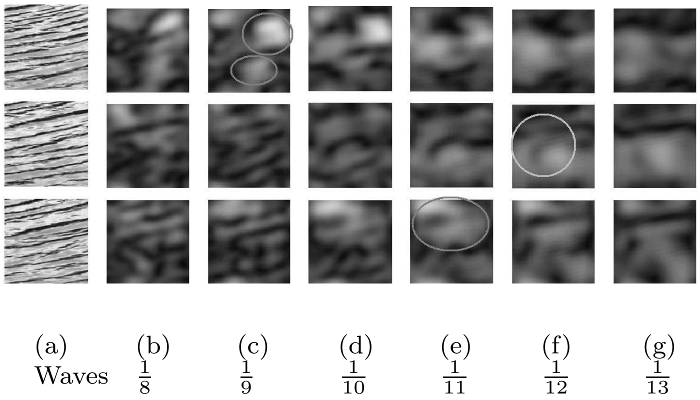


Figure 4.13: Color video of waves propagating through water. The frames are randomly selected and represented by the intensity channel $\hat{E}$ (a). The spatial frequency of surface of the water changes throughout the represented fragment. Responses of spatial filters tuned to various frequencies (in cycles per pixel) are shown (b-g). In analogy to temporal frequency identification, the identified frequency corresponds to the frequency of the filter with a maximum spatially pooled response, see indications. For instance, the filter response in (f) to the second frame is higher than filters tuned to slightly different frequencies in (e) and (g). Assigning a maximum response to the first frame is more ambiguous: filter responses in (b), (c) and (d) seem very similar. For the first frame, the algorithm appointed (c) as the maximum response, whereas for the third frame (e) gives the maximum response.

The velocity measurements confirm that the velocity of the grating changes gradually. The spatial frequency of the water, and the temporal frequency of its speed change throughout the video. However, the velocity measurements in random frames reflect that the velocity is more or less stable throughout the video fragment (mean $\mu = 2.0$ and standard deviation $\sigma = 0.8$ pixels per frame, whereas for the whole video fragment $\mu = 1.8$ and $\sigma = 1.0$ pixels per frame).

4.3.4 Temporal Frequency as an Attentional Attribute

This final experiment illustrates the periodicity of an object's motion to be an *attentional attribute*. Motion periodicity, like flicker, is a probable attribute to guide visual attention [127]. Debate exists whether only luminance polarity or both luminance and color polarity draws the attention towards the object [127]. Therefore, we only consider temporal regularity apparent in the luminance channel. Recall that the zeroth order spectral derivative filter measures the luminance.

We analyzed a color video fragment showing both stochastically moving leaves in the wind and one periodically moving leaf. The temporal regularity in the latter leaf guides the attention towards that leaf (see "Supplemental Material" [16]). In Figure 4.14, frames $\{0, 5, \ldots, 40\}$ depict half-periods of the periodically moving leaf. The initial amplitude of the leaf is indicated by dashed vertical lines. Frames 0-40

Table 4.1: Velocities of the moving gratings.

Subsequent frames (random starting point)	*150*	*151*	*152*	*153*	*154*	*155*	*156*	*157*	*158*	*159*
Spatial frequency (cycles per pixel)	8	6	7	13	10	13	13	10	13	13
Temporal frequency (cycles per frame)	13	6	6	7	7	7	7	7	7	7
Velocity (pixels per frame)	0.6	1.0	1.2	1.9	1.4	1.9	1.9	1.4	1.9	1.9

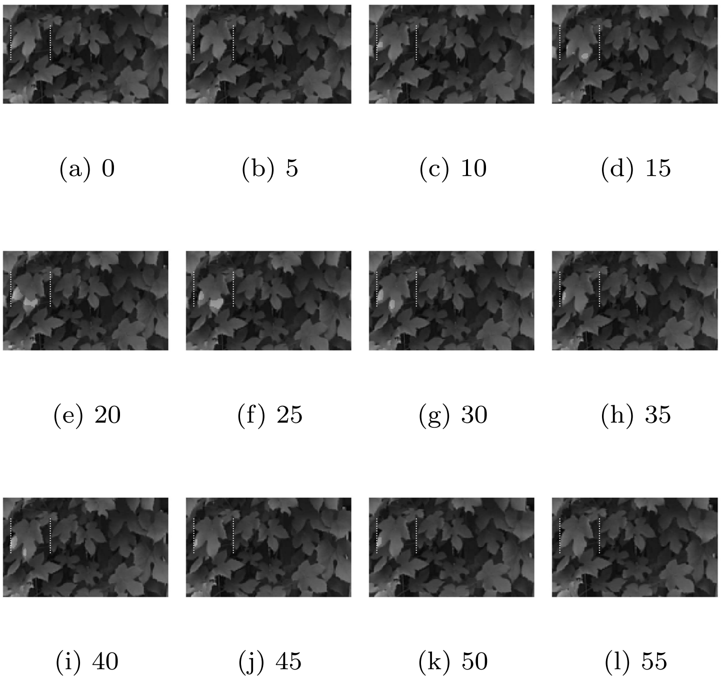
Random frames	*33*	*36*	*78*	*115*	*132*	*147*	*158*	*162*	*169*	*183*
Spatial frequency (cycles per pixel)	10	13	6	8	13	8	13	13	10	10
Temporal frequency (cycles per frame)	4	4	5	8	8	3	7	7	3	8
Velocity (pixels per frame)	2.5	3.3	1.2	1.0	1.6	2.6	1.9	1.9	3.3	1.3

Velocity of moving gratings. The moving gratings are taken from color video of waves propagating through water (samples are depicted in Figure 4.13). Hence, the gratings exhibit a temporal frequency. The ratio of the measured spatial and temporal frequency determines the velocity of the grating [108].

show 4.5 periods of the moving leaf, thus the leaf moves approximately at a frequency of $\frac{1}{10}$ cycles per frame. We overlayed the response of the temporal frequency filter tuned to $u_0 = \frac{1}{10}$ cycles per frame. The highlighted regions in Figure 4.14 indicate the detection of periodicity. Note that the temporal frequency filter responds well after one period, that is, after frames 0-5 have occurred. Frames 45-55 are included to illustrate the immediate decay in the filter's response after the leaf starts to move slower and with less amplitude.

Since the leaves themselves only differ in their motion, and not in luminance, color, shape, size, texture, or velocity [127], and the object of attention is not in the center of the color video, we conclude that the attention is only due to the object's temporal regularity. Hence, when temporal regularity is considered in isolation, temporal frequency detection draws the focus of attention.

Real-Time Performance



(a) 0 (b) 5 (c) 10 (d) 15

(e) 20 (f) 25 (g) 30 (h) 35

(i) 40 (j) 45 (k) 50 (l) 55

Figure 4.14: Color video of the periodic motion of one leaf, in the midst of stochastically moving leaves. The frames represent half-periods of the motion and are represented by the intensity channel $\hat{E}$. The subscripts denote the frame offset from the first frame that is displayed. When temporal regularity is considered in isolation, temporal frequency detection draws the focus of attention [127]. The maximum response among different temporal frequency filters detects the temporal regularity (filter frequency: $u_0 = \frac{1}{10}$ c/f) and is overlayed as a highlight area. Hence, the periodically moving leaf hence guides the attention towards itself. See "Supplemental Material" [16] for the original color video plus overlayed responses.

The filter was applied to color video using a Pentium XEON processor at 2.4 GHz. The computation time of the recursive spatial convolutions described in [128, 129] is independent of the scale σ_{xy} and relates only to the recursion order of the filter and the dimensions of the video. We set the recursion order to 3. For the European PAL and American NTSC MPEG video standard the dimensions are: 720×578 pixels at 25 Hz, and 720×480 pixels at 30 Hz, respectively. Spatial convolutions for PAL (NTSC) consume 21 (17) ms/frame. Computing 3 periods spanning 3 seconds in total, the temporal convolution consumes an additional 18 (15) ms. Total computation thus takes 40 (33) ms, achieving real-time performance. However, these real-time results were obtained for one filter with a large temporal extent. To meet the real-time requirement for multiple and simultaneous filters as described in Section 4.2.3, parallel computation [105] is required.

4.4 Conclusions and Discussion

In this chapter, we have derived an online, real-time temporal frequency filter. The filter measures in space and wavelength spectrum to estimate the object's surface reflectance. The filter measures temporal frequency to determine the periodicity of the reoccurrence of the surface. Embedded in the scale-space paradigm, the measurement boils down to a 4-dimensional filter, representing a Gaussian filter in the spatiospectral domain and a Gabor filter in the temporal domain. When measuring online, only the past information can be accessed. We therefore have applied a reparameterization of the temporal filter to deal with this constraint. We have introduced color to increase discriminative power to determine the reoccurring surface. Additionally, we introduced spatial extent thereby incorporating local information. We have demonstrated that with moving objects that do not periodically exhibit exactly the same pose, spatial contextual information makes the filter more robust. The constructed online temporal frequency filter measures both frequency identification ("what") and frequency detection ("when").

For simplicity, we have assumed that the spectrum that is reflected from the object does not change under object movement. In general, this assumption does not hold for a moving object. A translation of the object relative to the light source causes primarily shadow and shading deviations. Under the Lambertian reflection model [63], the color video signal $E(x, y, \lambda, t)$ may be decomposed into an intensity component $i(x, y, t)$ and the spectral distribution $e(x, y, \lambda, t)$ representing the color at each location: $E(x, y, \lambda, t) = i(x, y, t)\, e(x, y, \lambda, t)$. A local normalization of the simultaneous measurement of color and temporal frequency, $\hat{E}_{\lambda}^{\sigma_{xy};u_0;\sigma_t,\tau}$ and $\hat{E}_{\lambda\lambda}^{\sigma_{xy};u_0;\sigma_t,\tau}$, by the intensity measurement $\hat{E}$, are robust against shadow and shading [46, 58].

In our experiments, we have restricted ourselves to the measurement of temporal frequency of periodic events and the velocity of periodic motion. We demonstrated the general applicability of the proposed filter. Further, we have demonstrated that the online temporal frequency filter is more selective for frequency measurements than the offline filter, as it responds and decays faster.

We have left specialized topics that analyze motion of specific kinds in depth out of consideration. We consequently have not attributed motion-based recognition and gait analysis. The experiments incorporate both the detection and identification of temporal frequency of stationary and nonstationary objects moving periodically and quasi-periodically. In color video, the proposed filter has proven to robustly measure the periodicity of natural motion of objects isoluminant with their background hence only visible in color. The filter has shown to segment the periodically moving object from its background. Although dynamic texture algorithms [32] do not extract explicit frequency information, these algorithms are very efficient in detecting temporal regularity. Hence, dynamic texture segmentation [33] may be useful to determine initially a region of interest to initialize the spatial parameters of the temporal frequency filter. Further, we demonstrated the filter, in combination with a spatial frequency filter, to estimate the velocity of moving gratings well. The estimation of velocity from the periodicity of an object's motion is robust due to its invariance to the object's distance. Although with varying distance the spatial scale of the filter has to be updated by either scale selection or tracking kernel normalization, the frequency of the object does not change. On the contrary, motion estimation from optical flow varies with an object's distance. Further, we illustrated the attentional attribute of periodic motion. Determining the focus of attention is important as it may detect targets for surveillance video. Finally, we provided examples where periodical events are direct consequences of harmonic mechanical systems in equilibrium. The measurement of an object's periodic motion hence may enable a vision system to estimate parameters of the harmonic mechanism under investigation.

Chapter 5

Color Textons for Texture Recognition*

5.1 Introduction

The appearance of rough 3D textures is heavily influenced by the imaging conditions under which the texture is viewed [30, 116]. The texture appearance deviates as a consequence of a change of the recording setting. Among others, the imaging conditions have an influence on the texture shading, self-shadowing and interreflections [30], contrast and highlights. Texture recognition [15,116,120] and categorization [55] algorithms have been proposed to learn or model the appearance variation in order to deal with varying imaging conditions. In this chapter, we consider the challenge of recognizing textures from few examples [120], for which discriminative models are needed to distinguish between textures, but also invariance of models is required to generalize over texture appearances. Note that texture recognition differs from texture categorization [55], where also generalization is needed over various textures belonging to one category.

Texture recognition methods based on texture primitives, i.e. textons, have successfully learned the appearance variation from grayscale images [120]. Although color is a discriminative property of texture, the color texture appearance model of [116] was tested on the Curet dataset [30] and has been outperformed by the grayvalue-based texton model [120]. This can be explained partly from the use of color image features that are not specific for texture, e.g. raw color values [116], and partly from using color features that are not robust to the photometric effects that dominate the appearance variation of textures. In this chapter, we aim at describing robustly both spatial structure and color of textures to improve the discriminative power for learning textures from few examples. Due to their high discriminative power, we extend the texton models of Varma and Zisserman (VZ) [120] to incorporate robustly color texture information.

Textons are typical representatives of filter bank responses. The MR8-filterbank [120],

*Appeared in *Proceedings of the British Machine Vision Conference*, 2006.

on which VZ is based, is designed such that it describes accurately the spatial structure of texture appearances [120]. A straightforward extension of the grayvalue-based MR8 filterbank would be to apply it to each channel of a multivalued image to describe the spatial structure of color images. However, the true color variations and the appearance deviations due to e.g. shading, interreflections and highlights are manifold. Hence, incorporating color information directly in the filter bank requires many examples to learn the color textons well. Moreover, color textons that are learned directly from the data may not be representative for all appearance deviations in the dataset, with the consequence that the representation of each texture will become less compact. Color invariants (e.g. [42]) have provided means to capture only object-specific color information which simplifies the learning and representation of appearances. However, this leaves one with a suitable choice of color invariant features. This is a nontrivial problem as most color invariants aim to disregard intensity information [42], which is a very discriminative property of textures [116, 120]. A change of the local intensity level is a common effect when textures are viewed under changing settings of the illumination and camera viewpoint [116]. Our first contribution is to propose color texture invariants that are largely insensitive to the local intensity level, while maintaining local contrast variations.

The learning of representatives of the spatial structure *and* colors of textures may be hampered by the wide variety of apparent structure-color combinations. An alternative approach to incorporate color directly, would be to incorporate color information in a post-processing step, leaving VZ intact. We propose a color-based weighting scheme for the coloring of grayvalue-based textons. The weighting scheme is based on the characteristics of color invariant edges, based on non-linear combinations of Gaussian derivative filters [46]. The Gaussian filter provides robustness to image noise. The quality of the color images may be poor, hence uncertainties are introduced in the extraction of color edge information. We characterize locally the color edges by their magnitude and direction, where we propagate magnitude and direction uncertainties to obtain a robust color edge model. We exploit this model to provide an efficient color-weighting scheme to extend VZ to incorporate color information.

We consider the recognition of textures from few examples, for which challenging datasets, containing a wide variety of appearances of textures, have been recorded. The Curet dataset [30] contains color images of textures under varying illumination and viewing direction. Recognition rates of 77% have been reported when textures are learned from two images only [120]. In this chapter, we improve VZ's discriminative power to increase recognition performance when learning Curet textures from few images.

The chapter is organized as follows. In Section 5.2 we shortly overview the original texton algorithm by Varma and Zisserman [120]. To incorporate color, we consider two alternative modifications of the VZ algorithm, as introduced above. In Section 5.3, we repeat the experiments of [120] to investigate the discriminative power of (a) the grayvalue-based textons, (b) the grayvalue-based textons plus weighting, and (c) color invariant textons.

5.2 Combining Textons and Color Texture Information

5.2.1 VZ

Before we propose two alternative modifications of the original grayvalue-based texture recognition algorithm of Varma and Zisserman (VZ) [120], we briefly overview it.

The VZ algorithm normalizes all grayvalue images to zero mean and unit variance. The MR8-filterbank is convolved with a train set of grayvalue images. The filters are $L2$-normalized and their outputs are rescaled according to Weber's law. See [120] for details.

From the filterbank-responses, textons are learned by performing k-means clustering (Euclidean distance), yielding a texton dictionary. The texton dictionary is found to be universal: a different learn set achieves similar results.

Next, each image is represented as a texton model. To that end, each image is filtered with the MR8 filter bank and at each pixel the texton that is closest in feature space is identified. The texton model of an image is a histogram, where each bin represents a texton and its value indicates the number of occurrences of the texton as it occurs in the image.

5.2.2 VZ-color

As a first attempt to extend the VZ algorithm to use color information, we incorporate color directly at the filterbank level [125]. Here, we extend the original MR8-filterbank [120] to filter color channels directly, where we manipulate the color channels to obtain invariance to intensity changes. First, each image is transformed to opponent color space. Using opponent color space, we benefit from the advantage that the color channels are decorrelated. As a consequence, the intensity channel is separated from the color chromaticity values. The transformation from RGB-values to the Gaussian opponent color model is given by [46]:

$$\begin{bmatrix} \hat{E}(x,y) \\ \hat{E}_\lambda(x,y) \\ \hat{E}_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{pmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix}, \tag{5.1}$$

where $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$ denote the intensity, blue-yellow and green-red channel.

The VZ-algorithm has shown to deal with intensity information in a very robust manner, by normalizing the grayvalue image first to zero mean and unit variance, thereby obtaining a large degree of invariance to changes of the viewing or illumination settings. We normalize the intensity channel in the same way.

We propose a physics-based normalization of the color values, such that the color values are invariant to local intensity changes, we term this scheme VZ-color. Here, the color values are rescaled by the intensity variation, but not normalized to zero mean to avoid further loss of chromaticity information. We start with the model: for direct and even illumination, the observed energy E in the image may be modelled by:

$$E(x,\lambda) = i(x)e(\lambda)R(x,\lambda), \tag{5.2}$$

where $i(x)$ denotes the intensity which varies over location x, effectively modelling local intensity including shadow and shading. Further, $e(\lambda)$ denotes the illumination spectrum, and $R(x, \lambda)$ denotes object reflectance depending on location x and spectral distribution which is parameterized by λ. Depending on which parts of the wavelength spectrum are measured, $E(x, \lambda)$ represents the reflected intensity, and $E_\lambda(x, \lambda)$ compares the left and right part of the spectrum, hence may be considered the energy in the "yellow-blue" channel. Likewise, $E_{\lambda\lambda}(x, \lambda)$ may be considered the energy in the "red-green" channel. The actual opponent color measurements of $E(x, \lambda)$, $E_\lambda(x, \lambda)$ and $E_{\lambda\lambda}(x, \lambda)$ are obtained from RGB-values by Equation 5.1.

A change of the region's intensity level is a common effect when textures are viewed under changing settings of the illumination and camera viewpoint. We consider manipulations of $E(x, \lambda)$, $E_\lambda(x, \lambda)$ and $E_{\lambda\lambda}(x, \lambda)$ to obtain some invariance to such appearance changes. With the physical model from Equation 5.2, the measured intensity $\hat{E}$ can be approximated by:

$$\hat{E}(x, \lambda) \approx i(x)e(\lambda)R(x, \lambda). \tag{5.3}$$

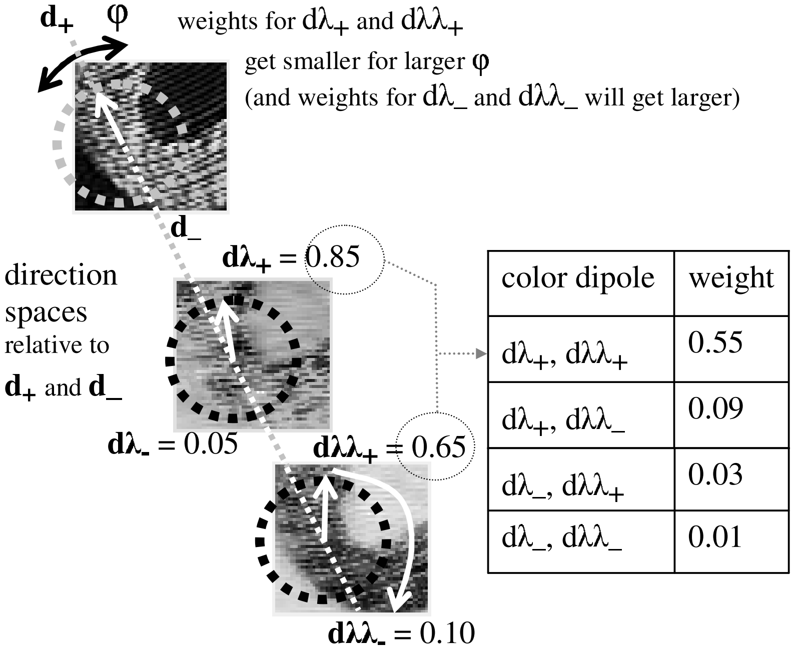
For the spectral derivatives, we obtain the approximations:

$$\begin{aligned} \hat{E}_\lambda(x, \lambda) &\approx \frac{d}{d\lambda} i(x)e(\lambda)R(x, \lambda) = i(x)\frac{d}{d\lambda}e(\lambda)R(x, \lambda), & (5.4) \\ \hat{E}_{\lambda\lambda}(x, \lambda) &\approx i(x)\frac{d}{d\lambda\lambda}e(\lambda)R(x, \lambda). & (5.5) \end{aligned}$$

We obtain the color measurements $\hat{E}_\lambda(x)$ and $\hat{E}_{\lambda\lambda}(x)$ directly from RGB-values according to Equation 5.1. The global variation in these color measurements due to variations of illumination intensity, shadow and shading is approximated by $i(x)$. The intensity measurement $\hat{E}(x)$, also directly obtained from Equation 5.1, is a direct indication of the intensity fluctuation. Therefore, the standard deviation of $\hat{E}$ over all pixels is used to normalize globally each of the color measurements $\hat{E}_\lambda(x)$ and $\hat{E}_{\lambda\lambda}(x)$, thus dividing by $\sigma(\hat{E})$, to obtain better estimates of the actual color variation. We do a global normalization here, and not per pixel, as local intensity variation in the color channels is considered important texture information. Finally, the MR8-filterbank is applied to these 3 color invariant signals.

5.2.3 VZ-dipoles

As an alternative to incorporating color at the level of the filterbank, the VZ algorithm is extended by a post-processing step where textons are weighted according to the color edge at the location of a particular texton. Color edges are measured by color gradients, of which the magnitudes and directions are used to characterize texture edges (subsection 5.2.3). The directions of color gradients are taken relative to the direction of the intensity gradient. Weights are computed to determine to which degree the color gradient direction corresponds to the intensity direction. The weights are combined to obtain an indication of the color transition at the location of a particular texton (subsection 5.2.3), with which the texton is weighted when adding it to the texton histogram (subsection 5.2.3). This process is outlined in Figure 5.1.



color dipole	weight
dλ₊, dλλ₊	0.55
dλ₊, dλλ_	0.09
dλ_, dλλ₊	0.03
dλ_, dλλ_	0.01

Figure 5.1: The color dipole framework. The three images denote the color representation of a color texture. For each color channel, the gradient is computed, depicted by the arrows. The direction of the opponent color gradients (d_λ and $d_{\lambda\lambda}$) are taken relative to the direction of the intensity gradient (d). For both the same ($+$) and opposite ($-$) direction, weights are determined. The smaller the weight gets for one direction, the larger it gets for the opposite direction. To obtain a combined weight for each combination of directions, the weights are multiplied.

Color Invariant Gradients

To exploit color information in a robust fashion, we base ourselves on noise-robust Gaussian image measurements. From these measurements, we extract color invariant gradients that are robust to changes of the intensity level [46], to achieve the same level of invariance as in the previous subsection. We overview shortly the derivation of color invariant gradients. First, we consider the transformation of RGB-values to opponent color space, yielding opponent color values E, E_λ and $E_{\lambda\lambda}$ to represent the intensity, blue-yellow and green-red channel, respectively, likewise the previous subsection (Equation 5.1) . From the opponent color values, spatial derivatives in the x-direction are computed by convolution with Gaussian derivative filters $G_x^\sigma(x,y)$ with scale σ:

$$\begin{aligned}\hat{E}_x^\sigma(x,y) &= E(x,y) * G_x^\sigma(x,y), &(5.6)\\ \hat{E}_{\lambda x}^\sigma(x,y) &= E_\lambda(x,y) * G_x^\sigma(x,y), &(5.7)\\ \hat{E}_{\lambda\lambda x}^\sigma(x,y) &= E_{\lambda\lambda}(x,y) * G_x^\sigma(x,y), &(5.8)\end{aligned}$$

where $(*)$ denotes convolution. The spatial derivatives of opponent color values, $\hat{E}_x^\sigma$, $\hat{E}_{\lambda x}^\sigma$ and $\hat{E}_{\lambda\lambda x}^\sigma$, are transformed respectively into color invariants $\hat{\mathcal{W}}_x^\sigma$, $\hat{\mathcal{W}}_{\lambda x}^\sigma$ and $\hat{\mathcal{W}}_{\lambda\lambda x}^\sigma$

providing robustness to changes of the intensity level by normalizing by the local intensity $\hat{E}^{\sigma}$:

$$\mathcal{W}^{\sigma}_{x}(x,y)=\frac{\hat{E}^{\sigma}_{x}(x,y)}{\hat{E}^{\sigma}(x,y)},\ \mathcal{W}^{\sigma}_{\lambda x}(x,y)=\frac{\hat{E}^{\sigma}_{\lambda x}(x,y)}{\hat{E}^{\sigma}(x,y)},\ \mathcal{W}^{\sigma}_{\lambda\lambda x}(x,y)=\frac{\hat{E}^{\sigma}_{\lambda\lambda x}(x,y)}{\hat{E}^{\sigma}(x,y)}. \tag{5.9}$$

This normalization may become unstable for low pixel values, but with the local smoothing some robustness to noise is obtained.

The color invariant features are computed at multiple scales to obtain scale invariance. We compute each scale-normalized invariant at 3 scales ($\sigma \in \{1,2,4\}$ pixels) and select the scale of the invariant that maximizes the response. Next, the color invariant gradients are computed. The gradient magnitude is determined from: $\hat{\mathcal{W}}_{\lambda^i w}=\sqrt{\hat{\mathcal{W}}_{\lambda^i x}(x,y)^2+\hat{\mathcal{W}}_{\lambda^i y}(x,y)^2}$, whereas its direction is determined from: $\arctan(\frac{\hat{\mathcal{W}}_{\lambda^i y}(x,y)}{\hat{\mathcal{W}}_{\lambda^i x}(x,y)})$. We obtain per pixel the color and scale invariant gradients $\hat{\mathcal{W}}_{w}$, $\hat{\mathcal{W}}_{\lambda w}$ and $\hat{\mathcal{W}}_{\lambda\lambda w}$. After application of the color invariants to the image set that is used for training (see Experiments), we learn their standard deviation. We normalize each invariant by its standard deviation, which effectively boosts color information.

Color Dipoles

An edge in a color image may be characterized by measuring for each color channel the energy gradient, as outlined in [132]. In order to exploit the a-priori structure in texture images, we investigate the correlation between intensity edges and color edges. Therefore, we determine at each pixel the orientation of intensity and color gradients, and measure the correlation between the orientations of intensity and color gradients over all pixels in the Curet dataset [30]. To measure the correlation between orientations at edge locations only, we determine the weighted correlation, where weights are provided by the total gradient magnitude at a particular pixel, measured by $\sqrt{\mathcal{W}_w(x,y)^2+\mathcal{W}_{\lambda w}(x,y)^2+\mathcal{W}_{\lambda\lambda w}(x,y)^2}$. The orientations of intensity and color gradients are strongly correlated: $r(\mathcal{W}_w,\mathcal{W}_{\lambda w})=0.77$, $r(\mathcal{W}_w,\mathcal{W}_{\lambda\lambda w})=0.81$ and $r(\mathcal{W}_{\lambda w},\mathcal{W}_{\lambda\lambda w})=0.82$.

We have observed that edges are largely characterized by color gradient magnitudes, and whether these gradients are directed in the same or opposite direction as the intensity gradient. The characterization of a color edge by a dichotomic framework is termed a color dipole. An example of a color dipole is displayed in Figure 5.1. The figure also displays poor image quality, indicating that robust modelling of color information is required.

We start with the alignment of the color dipole framework to the direction of the intensity gradient $\mathcal{W}_w$. The direction of $\mathcal{W}_{\lambda w}$ is compared to the direction of $\mathcal{W}_w$. Two Gaussian kernels in direction-space measure the certainty that the direction is the same or opposite to the intensity gradient direction, see Figure 5.2. The choice of the size of the kernels has no significant effect on texture recognition results (data not shown). The kernels in direction-space yield 2 direction weights, one for the same direction as the intensity gradient and one for the opposite direction. The more

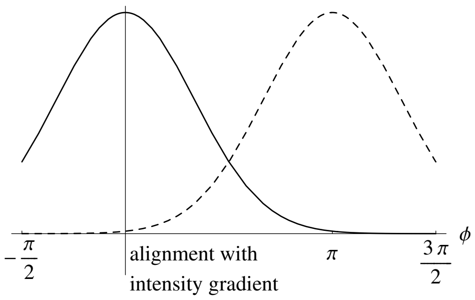


Figure 5.2: Two Gaussian kernels in direction-space measure the certainty that the direction is the same or opposite to the intensity gradient direction.

the direction and opposite direction differ with respect to the direction and opposite direction of the intensity gradient, the lower the weight. Also, the smaller the weight gets for one direction, the larger it gets for the opposite direction. Analogously, two direction weights are determined for the gradient $\mathcal{W}_{\lambda\lambda w}$.

In total, we obtain for each feature two direction weights per pixel, which for two features yields $2 \times 2 = 4$ combinations. For each of the four combinations, we obtain a single weight by multiplying the two corresponding feature direction probabilities, see Figure 5.1. For each of the four dipole possibilities, we have obtained a single weight representing the probability that the edge under investigation is characterized by it.

To ensure that the feature directions are stable, we weight the dipole framework per pixel by the total edge strength $\sqrt{\mathcal{W}_w(x,y)^2 + \mathcal{W}_{\lambda w}(x,y)^2 + \mathcal{W}_{\lambda\lambda w}(x,y)^2}$ at that pixel. We normalize the sum of all weights over the image to unity. The dipole framework provides robustly the probability for each of the four color dipoles per pixel.

Color-weighted Textons

The color-weighting scheme only extends the VZ-algorithm in the way in which the occurrences of grayvalue-based textons contribute to the histogram bins of the texton model. Rather than accumulating a unity weight for each occurrence of a particular texton, we add weights according to the dipole measured at the location of interest. Since we have four weights, each of the original histogram bins of VZ are split into four, such that each of the four weights per texton can be added to the four bins that correspond to the particular texton. Like VZ, the histograms are normalized to unity, and compared using the χ^2-statistic.

In recapitulation, the VZ-dipole algorithm affects only the cardinality of the texton model. Hence, VZ-dipole is a low-cost strategy to obtain colored textons, while avoiding the introduction of essentially different textons for the learning and representation of textons in the image. The color invariant textons, VZ-color-norm, affects the cardinality of the filterbank. In addition, the learning of textons from the color-based VZ-color-norm filterbank requires a learn set that is both representative of the texture shape primitives in the dataset as well as their colors. Table 2 summarizes the proposed modification of and extension to the original grayvalue-based texture recognition VZ algorithm [120].

Table 5.1: Characteristics of the VZ algorithm and proposed modifications of VZ.

	Texton learn set representative of	Size of filterbank	Size of representation
VZ	texture shape primitives	8	# textons
VZ-color	texture shape primitives *and* colors	24	# textons
VZ-dipoles	texture shape primitives	8	4 × # textons

5.3 Texture Recognition Experiment

In this section, we demonstrate the discriminative power of colored textons for texture recognition. We follow the experimental setup of Varma and Zisserman [120] to classify the 61 textures of the Curet dataset [30]. Textons are learned from the same 20 textures as used in [120] and [28]. For each texture, 13 random images are convolved with the MR8-filterbank [120], from which all responses are collected and 10 cluster means are learned to obtain 10 textons. Hence, using 20 textures to learn textons from, 200 textons are learned; this is the texton dictionary. For each of the 61 textures in the Curet dataset, 92 images have been selected by [120] to obtain a total of 5612 images. Each image is represented by a histogram of grayvalue-based texton frequencies [120].

For the recognition of textures, we also follow [120]. To classify textures, each texture is represented by 46 models obtained from alternating images in the total of 92 images per texture. These 46 models are the learning set; the remaining 46 images are test images.

5.3.1 Baseline Performance

We consider the recognition of only 20 textures as used in [120] and [28], based on all 46 models. The VZ algorithm, termed VZ, based on grayvalue-textons achieves a recognition performance of 97.8%. With the physics-based normalization of opponent color values, exploited in VZ-color, the results are better: 98.4%. With the dipole-weighted textons, VZ-dipoles, the highest performance is achieved: 98.7% of the 20 textures is classified correctly.

Due to their improvement in recognition performance over grayvalue-based textons, we consider VZ-color and VZ-dipoles in comparison to VZ for the recognition of all 61 textures from the Curet dataset. As a baseline, with `VZ`, the accuracy of classifying all 61 textures is: 96.4%.

With `VZ-color`, a recognition accuracy of 97.1% is achieved. This is a good result, but we want to know the effect of the choice of the image set to learn textons from. To that end, we have selected randomly alternative sets of images to learn the textons from. We have conducted 10 trials, for each trial random images are taken from the textures used in [120] and [28]. For `VZ-color`, the texture recognition results depend significantly on the texton learn set: recognition accuracy varies from 92.7% to 97.1%, while for the grayvalue-based textons the results vary mildly from 96.0% to 96.4%. We conclude that the learning of discriminative color textons is more sensitive to the choice of the learn set.

Because for grayvalue textons the choice of the learn set is of much less importance, the results obtained with `VZ-dipoles` are stable under the choice of the learn set: recognition accuracy varies from 96.1% to 96.5%. It should be noted that 200 textons are used, identical to the textons used in VZ, but with 4 weights attached to each texton. Increasing the number of textons to 800 increases only very marginally the performance of VZ [120].

5.3.2 Reducing the Learn Set

To test the recognition accuracy when fewer models are incorporated in the learn set, we start to decrease the number of learn models, likewise [120]. The learn set is reduced by discarding models that contribute least to the recognition performance. Models were discarded in each iteration step based on a greedy reduced nearest-neighbor algorithm.

We emphasize that, by reducing the number of models, first the noisy models are discarded, improving the texture recognition performance. Here we consider the performance of the algorithms `VZ` and `VZ-dipoles`, which have demonstrated most stable under the choice of the texton learn set (see above). Experiments over all 61 textures, where models are removed from the learn set by means of the reduced nearest neighbor rule, the best recognition accuracy obtained with the color textons of `VZ-dipoles` is 98.3%. Thus, the best results obtained with `VZ-dipoles` are somewhat lower than achieved by Broadhurst: 99.2% [15]. Broadhurst modelled filterbank responses directly, i.e. without the abstraction step of modelling textons, by a 26-dimensional Gaussian, which was subsequently used in a Bayes recognition framework. It is interesting that the compact texton models achieve a performance that is almost similar to the performance of the elegant models proposed by Broadhurst.

It is interesting how the recognition accuracy decreases when using only few learning models. When 2 models are used, the texture recognition performance increases from 77.1% (`VZ`) to 85.6% (`VZ-dipoles`) when including color information. We conclude that exploiting color information facilitates the learning of texture appearances.

The results can be summarized as follows. `VZ-dipoles` outperforms consistently both the original `VZ` textons as well as the color textons obtained from color invariant filterbank responses, see Table 3. Interestingly, the results obtained with the color-weighted textons (VZ-dipoles) are most stable over: (a) image sets to learn textons from, and, more importantly, (b) image sets to learn textures from.

Table 5.2: Performance of the VZ algorithm and proposed modifications of VZ.

Textures:	20	61		61
Textons:	20	20		20
Models:	46	46		2
Algorithm		best	worse	
VZ	97.8%	96.4%	96.0%	77.1%
VZ-color	98.4%	97.1%	92.7%	–
VZ-dipoles	98.7%	96.5%	96.1%	85.6%

5.4 Conclusion

In this chapter, we have proposed methods to incorporate robustly color information in VZ textons [120] to model the appearance of textures. The textons are learned from filterbank responses. First, we have incorporated color directly at the level of the filterbank. We have shown that the learning of discriminative color textons that are representative of both the textures' shape primitives and colors is not trivial and the recognition accuracy is very dependent on the set of images to learn the color textons from.

As an alternative to incorporate color directly at the filterbank, we have proposed a color weighting scheme to weight grayvalue-based textons by the color edges that generate the texture. This framework captures robustly essential color texture information, is efficient to compute, and provides a simple extension to the original texton model.

In the experiments, we have modelled color texture images from the Curet dataset by the traditional textons and the color-weighted textons. With color-weighted textons, the texture recognition performance is increased significantly, up to ten percent when only two texton models per texture are used. Incorporating color in a robust manner by means of the proposed dipole model adds discriminative power for texture recognition, which facilitates the learning of color textures.

Chapter 6

Material-specific Adaptation of Color Invariant Features

6.1 Introduction

The appearance of materials change significantly under different imaging settings, depending on the settings themselves [30] and also on the physical properties of a material [70]. Hence, materials-specific image representations may improve on the recognition performance, as they capture properties that are distinctive to the material and are balanced with the variation of imaging settings. For instance, for one material the intensity variation is a distinctive property, while the other is distinguished best from other materials based on its color properties. Figure 6.1 depicts some materials from the ALOT dataset*. The first and second material are distinguished best when comparing their colors, more specifically, the red channel. For the third and fourth material, the most discriminative feature is the amount of intensity edges, while the fifth image in the first row and the second image in the third row are distinguished best when comparing the information in the green channel. These examples illustrate the advantage of material-specific representations. The objective in this chapter is to learn material-specific representations for more than 250 materials.

For material recognition [76, 120] and classification [55], but also for object and scene classification [133], the mapping of image features onto a codebook of feature representatives [64,92] has received extensive treatment. Commonly used features are the class of SIFT-based features [79,84], see e.g. [75]. Alternatively, filterbank outputs are in use as features. Promising methods that use filterbanks to model object and scenes, have been proposed by Winn *et al.* [125] and by Shotton *et al.* [107].

In previous work by the author [17], image edges were filtered by a filterbank and subsequently annotated by their color improving the discriminative power of

*http://www.science.uva.nl/∼aloi/public_alot

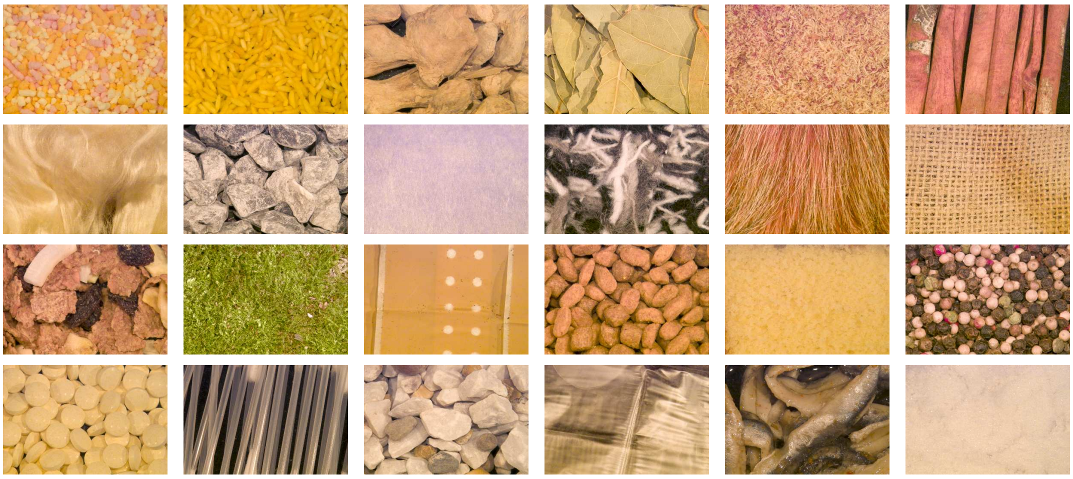


Figure 6.1: Example materials from the ALOT dataset.

filterbanks further. The objective of [17] was to extend a method that was originally proposed for grey-value images to include color information. The extension works well for images with many edges, and we do not expect it to work for more general images. Furthermore, the purpose of this chapter is broader: we will integrate various ways of measuring color by filterbanks. We consider filterbanks for reason of their discriminative power, simplicity and generality.

To adapt the representation to a particular material, we consider various ways to represent an image. To that end, consider various intensity and color filterbanks. They are adapted from the MR8-filterbank which performs well in a recent evaluation [120]. Each of the filterbanks measures different color channels, and each achieves a different degree of photometric invariance. We adopt techniques from the literature on invariant feature design, see e.g. [37, 46, 50]. The general scheme to construct a representation of a filtered image, typically a histogram, is to first establish representatives of the filter outputs, or textons [76]. A standard solution that aims to minimize the average reconstruction error is the k-means algorithm, employed originally by Sivic and Zisserman [109] and Csurka *et al.* [27]. Alternatively, Winn *et al.* [125] employed an information-maximization approach. For any of these approaches to establish textons, the problem is how to arrive at a representation that is specific to the material at hand.

A recent method proposed by Perronnin *et al.* [99] establishes class-specific textons for each of N classes. As the authors point out, the straightforward accumulation of all textons into one large codebook is not feasible, as the learning of materials will be hampered as a result of the large histograms representing the images (curse of dimensionality). To avoid this problem, they suggest to use the N sets of class-specific textons to create respectively N codebooks. Each image is subsequently represented by the N codebooks resulting in N histograms. Elegantly, for each image the N histograms are fed to N class-specific classifiers. Classification of the image is based on the N thus obtained posterior probabilities. In [99], high performance is reported for the classification of $7 - 10$ categories. However, for the classification of more than 250 materials, the method in [99] will be hampered by the creation of more than 250 histograms for each image. With 24 images per class, over $24 \cdot 250 \cdot 250$ histograms need to be constructed, which is not feasible in practice. Rather, we will propose a scalable alternative to construct material-specific representations, by representing the image by $M << N$ histograms. The M histograms are obtained from M color invariant codebooks, each learned from one filterbank with specific color and invariant properties. As a result, the class-specificity of codebooks is not in the learned textons, but in their color and invariance properties.

The chapter is organized as follows. In Section 6.2, the MR8 filterbank and its color invariant versions are introduced. We propose the framework to learn material-specific color information and invariance in Section 6.3. In Section 6.4, we evaluate first the performance of the intensity and color filterbanks on the CURET [30] and ALOT datasets based on discriminative power, invariance to image settings and clutter. Second, we evaluate the framework to adapt the use of filterbanks to the material. Conclusions are drawn in Section 6.5.

6.2 Color Invariant Filterbanks

In this section, we introduce the MR8 filterbank and several extensions to color. The MR8 filterbank is shown in Figure 6.2a. Typically, before the image is convolved with the MR8 filterbank, the image is normalized to zero mean and unit variance to achieve to a large extent invariance to imaging conditions, see e.g. [120]. In the following subsections, we extend the MR8 filterbank to incorporate color information, and we consider various transformations to achieve color invariance from literature.

6.2.1 MR8-NC

In a first modification of the MR8 filterbank to extend it to use color information, we apply the filterbank to the image's color channels directly. This is a straightforward extension that is also employed by Winn *et al.* [125], who have applied the MR8 filterbank to *Lab* color values. We largely follow [125] here. However, we restrain to a linear subspace of RGB, and apply the filterbank to the three opponent color channels of the image. Opponent colors have the advantage that the color channels are largely decorrelated. Here, we consider the Gaussian opponent color model, which is computed from RGB values directly by [46]:

$$\begin{bmatrix} \hat{E}(x,y) \\ \hat{E}_\lambda(x,y) \\ \hat{E}_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{pmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix}, \tag{6.1}$$

where $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$ denote the intensity, blue-yellow and green-red channel.

Likewise the usage of the MR8 filterbank in the VZ algorithm [120], we normalize each of the color channels $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$, to zero mean and unit variance. Next, each of the normalized color channels is convolved with the MR8 filterbank, yielding

24 filter outputs per pixel. This first extension of the MR8 filterbank is termed MR8 with normalized colors, or `MR8-NC`, which is formalized as:

$$\mathrm{MR8-NC} \;=\; \{\mathrm{MR8}(\frac{\hat{E}-\mu_{\hat{E}}}{\sigma_{\hat{E}}}),\;\mathrm{MR8}(\frac{\hat{E}_{\lambda}-\mu_{\hat{E}_{\lambda}}}{\sigma_{\hat{E}_{\lambda}}}),\;\mathrm{MR8}(\frac{\hat{E}_{\lambda\lambda}-\mu_{\hat{E}_{\lambda\lambda}}}{\sigma_{\hat{E}_{\lambda\lambda}}})\},$$

where $\mu_{\hat{E}_{\lambda^i}}$ denotes the mean of the i-th color channel, and $\sigma_{\hat{E}_{\lambda^i}}$ the standard deviation.

6.2.2 `MR8-INC`

In a second modification, we normalize the color channels such that they maintain more color information than is the case with `MR8-NC`. With `MR8-NC`, the means of the yellow-blue and red-green channels are normalized to zero, effectively discarding the actual chromaticity in the image, and only considering the variation. The color channels will be affected mainly by the lighting direction relative to the object and to the camera [116], which are mostly characterized by intensity fluctuations. Hence, we propose to normalize the three opponent color channels only by the standard deviation of the intensity. Here, the intensity variation over the pixels in the image is measured directly from the first opponent color channel $\hat{E}(x, y)$. Normalizing the intensity channel by the standard deviation of intensity, $\sigma(\hat{E})$, sets the variance of this channel to unity. Normalizing the yellow-blue and red-green channels by $\sigma(\hat{E})$ yields a more stable responses when the intensity variation fluctuates as a consequence of lighting or viewpoint changes. At the same time, it maintains information about the chromaticity in the image. Likewise `MR8-NC`, each of the normalized color channels is convolved with the MR8 filterbank, yielding 24 filter outputs per pixel. We refer to this filterbank as MR8 with intensity-normalized colors, or `MR8-INC`:

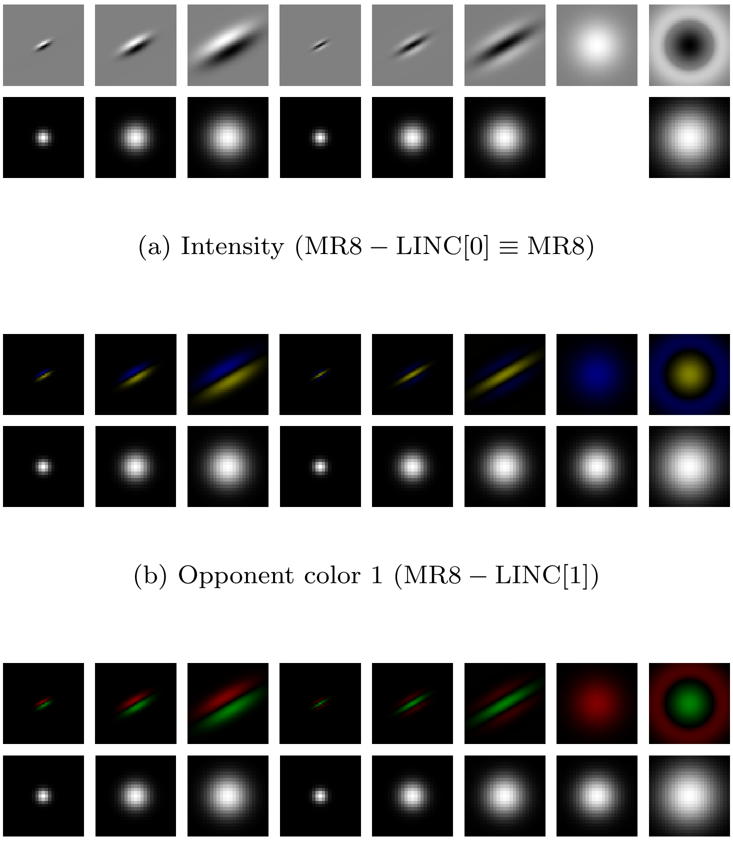
$$\mathrm{MR8-INC} \;=\; \{\mathrm{MR8}(\frac{\hat{E}-\mu_{\hat{E}}}{\sigma_{\hat{E}}}),\;\mathrm{MR8}(\frac{\hat{E}_{\lambda}}{\sigma_{\hat{E}}}),\;\mathrm{MR8}(\frac{\hat{E}_{\lambda\lambda}}{\sigma_{\hat{E}}})\},$$

with $\mu_{\hat{E}}$ and $\sigma_{\hat{E}}$ the mean and standard deviation of the intensity channel.

6.2.3 `MR8-LINC`

In a third modification, we modify the MR8-filterbank to achieve invariance to local intensity changes by a local color normalization rather than a global one. We follow closely the invariant Gaussian features developed in [46]. In [46], each of the local image measurements is normalized by the intensity in a small neighborhood. This achieves invariance to the local intensity level.

We propose to filter for each pixel the non-normalized opponent color values using the MR8-filterbank, to obtain 24 filter outputs per pixel. Also, for each pixel, we measure the local intensity with a Gaussian kernel. Per pixel, we normalize each output of the MR8 filterbank by the local intensity which is measured by a Gaussian at the same scale, see Figure 6.2. Obviously, the zeroth order Gaussian filter from the MR8-filterbank is not normalized by the local intensity, otherwise its output would be constant. We refer to this final color filterbank as MR8 with local intensity-normalized colors, or `MR8-LINC`.



(a) Intensity (MR8 − LINC[0] ≡ MR8)

(b) Opponent color 1 (MR8 − LINC[1])

(c) Opponent color 2 (MR8 − LINC[2])

Figure 6.2: `MR8-LINC`: a color invariant filterbank. The original MR8-filterbank (a – top row) is convolved with each of the image's opponent colors channels (a-c – upper rows), to yield 24 responses per pixel. Each of the 24 filter outputs is normalized by the local intensity as is measured by a Gaussian kernel of the same size of the MR8 filter (a-c – lower rows). The only MR8 filter that is not normalized is the Gaussian kernel that measures intensity (otherwise it would yield a constant output). The normalization achieves invariance to local intensity changes.

Formally, for the `MR8-LINC` filterbank, each of the filter outputs is normalized by the measured intensity $\hat{E}$. Let $\hat{F}_{\lambda^i x^j}$ denote the filter output, with $i \in \{0, 1, 2\}$ the opponent color channel (Equation 6.1), and $j \in \{0, 1, 2\}$ indicating smoothing or spatial differentiation up to first or second order. The scheme of `MR8-LINC` is formalized as:

$$\mathrm{MR8} - \mathrm{LINC} \quad = \{\tfrac{\mathrm{MR8}(\hat{E})}{\hat{E}}, \ \tfrac{\mathrm{MR8}(\hat{E}_\lambda)}{\hat{E}}, \ \tfrac{\mathrm{MR8}(\hat{E}_{\lambda\lambda})}{\hat{E}}\}.$$

6.2.4 MR8-SLINC

Finally, we construct a shadow and shading invariant filterbank, termed MR8-SLINC. Similar to MR8-LINC, the invariance is achieved locally. With MR8-LINC, first the filterbank outputs are computed before normalization by the local intensity. Alternatively, the color values $\hat{E}_\lambda(x, y)$ and $\hat{E}_{\lambda\lambda}(x, y)$ can be normalized locally first, $\frac{\hat{E}_\lambda(x,y)}{\hat{E}(x,y)}$ and $\frac{\hat{E}_{\lambda\lambda}(x,y)}{\hat{E}(x,y)}$, before filtering the thus obtained images. Under Lambertian reflection, the normalization of color values by the local intensity results in color values independent of the intensity distribution. Hence, the filterbank outputs of MR8-LINC are invariant to shadow and shading:

$$\mathrm{MR8-SLINC} \quad = \quad \{\frac{\mathrm{MR8}(\hat{E})}{\hat{E}},\ \mathrm{MR8}(\frac{\hat{E}_\lambda}{\hat{E}}),\ \mathrm{MR8}(\frac{\hat{E}_{\lambda\lambda}}{\hat{E}})\}.$$

6.2.5 Filterbank Properties

Similar to MR8, the color-based filterbanks MR8-NC and MR8-INC involve a global color normalization. In other words, the normalization is dependent on the contents of the image. Hence, clutter will affect the normalization. This makes the output of MR8-NC and MR8-INC *scene-dependent.* In contrast, the local normalizations that are employed in MR8-LINC and MR8-SLINC are not scene-dependent, but only *locally dependent* on the actual color values.

Further, the filterbanks can be ordered by their degree of invariance. MR8-SLINC is most invariant as its color channels aim to discard intensity variation. MR8 and MR8-NC retain respectively the intensity and color variation, but they discard their mean and variance. MR8-LINC retains more of the intensity and color variations, as it discards locally the variance due to intensity fluctuations. Finally, MR8-INC is less invariant than MR8-LINC, as it discards only the global variance due to intensity fluctuations.

6.3 Color Invariant Codebooks and Material-specific Adaptation

In this section, we consider the construction of color invariant codebooks from the several filterbanks, and the methodology to apply the codebooks in a material-specific setting. First, we formalize the color invariant filterbanks as follows: MR8-X = { MR8-X[0], MR8-X[1], MR8-X[2]}, where X ∈ {NC, INC, LINC, SLINC}. To avoid the joint learning of color channels, we learn one codebook for each color channel MR8-X[i], with $i \in \{0, 1, 2\}$. For codebook construction, we follow the common scheme of learning textons by k-means clustering of filterbank outputs [27, 76, 109, 120]. We consider a single set of 20 images randomly drawn from the learning set of material images. Each is filtered by one of the filterbanks MR8-X[i], and from each filtered image we store 10 cluster centers. As a result, for each filterbank MR8-X[i], we obtain a codebook of 200 textons. For the filterbank MR8-X, we have obtained 3 codebooks of length 200. For fair comparison with the single-channel MR8 filterbank, the length of the MR8 codebook is increased to 600 by storing 30 instead of 10 cluster centers per learning image.

To represent an image in terms of codebooks, it is filtered by each of the color channel filterbanks MR8-X[i] first, before mapping the filter outputs onto the corresponding codebook and counting the most similar occurrences. For each MR8-X[i], a histogram of length 200 is obtained; hence for MR8-X three histograms are obtained. After concatenation of the histograms per color channel, a histogram of length 600 is obtained that corresponds to the filterbank MR8-X. The codebook representation is outlined in Figure 6.3.

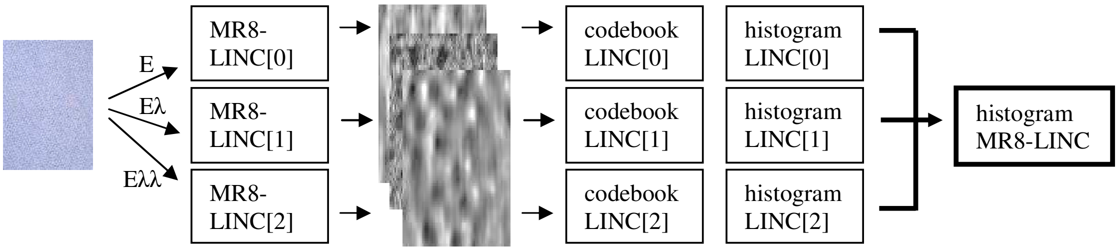
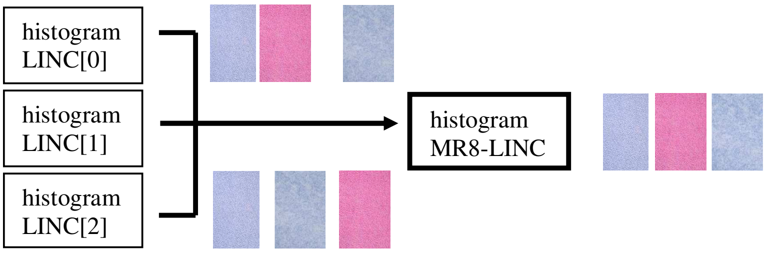


Figure 6.3: Color codebook approach where the three color channels are separately filtered and represented by a histogram. Subsequently, the histograms are combined into one.

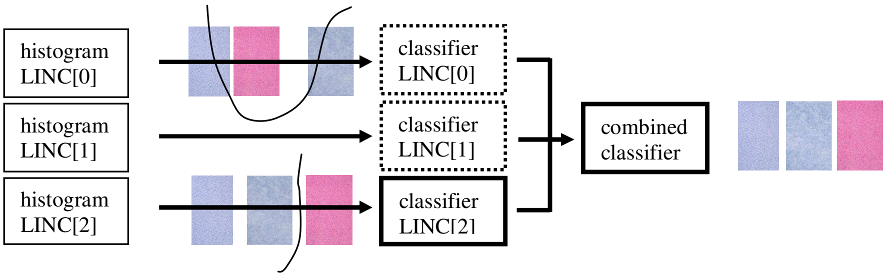
6.3.1 Material-specific Adaptation

The limitation of the color codebook representation as proposed above, is that the discriminative power of the color channels is averaged by using a single histogram comparison measure. For instance, the intensity information may be less distinctive for a given material than is the color information. The averaging of the information in the color channels may lead to incorrect classification of materials. The misclassification of an image of the blueish material, mistakenly considered to be more similar to the pink material, is illustrated in Figure 6.4a. To overcome the limited resolving power of the direct combination of the three color channels, we start with classification of a material at the level of individual color channels and to give preference to a distinctive combination thereof. Figure 6.4b illustrates that the blueish material is well separated from the pink material using the information in the third color channel.

We propose to train one classifier per color channel per filterbank to discriminate one material from all other materials. Hence, with I filterbanks, $F_{1...I}$, and J color channels, $c_{1...J}$, we obtain $I \times J$ classifiers. With N materials, each classifier outputs N posterior probabilities. With this procedure, $I \times J \times N$ values are produced by the first classifier stage. This procedure is illustrated for material 1 in Figure 6.5, where each of the $I \times J$ axes represents a classifier trained to distinguish material 1 from materials $2-6$ given a filterbank and an (invariant) color representation. The



(a) Fixed combination scheme; resulting ordering on the right



(b) Adaptive combination scheme; resulting ordering on the right

Figure 6.4: Separation of two images of the same material from one image of an other material. The fixed representation in (a) is not able to distinguish correctly between the two, while the material-specific representation is able to distinguish between the two (third color channel).

values plotted in this material-specific feature space represent the $N = 6$ posterior probabilities assigned to the materials by the individual classifiers.

In the combination stage, one classifier is trained using the $I \times J \times N$ values obtained for each material image. This one versus all classifier learns per material the discriminant function from the posterior probabilities assigned to each material by the individual classifiers. In Figure 6.5, this is indicated by the dashed line. As a result, the combined classifier learns the filterbank and color channel that is most distinctive for the specific material. In the example of Figure 6.5, the most discriminative combination of filterbank and color representation is (F_i, c_j) to distinguish material M_1 from materials $M_{2\ldots6}$ is represented on the x-axis.

To infer from the material-specific discriminant function provides information which filterbank and color representation combination is most distinctive for a given material, we determine for each material which of the individual classifier's outputs approximates the normal to the discriminant function of the combining classifier best. This measure indicates the importance of a particular filterbank for the classification of the given material.

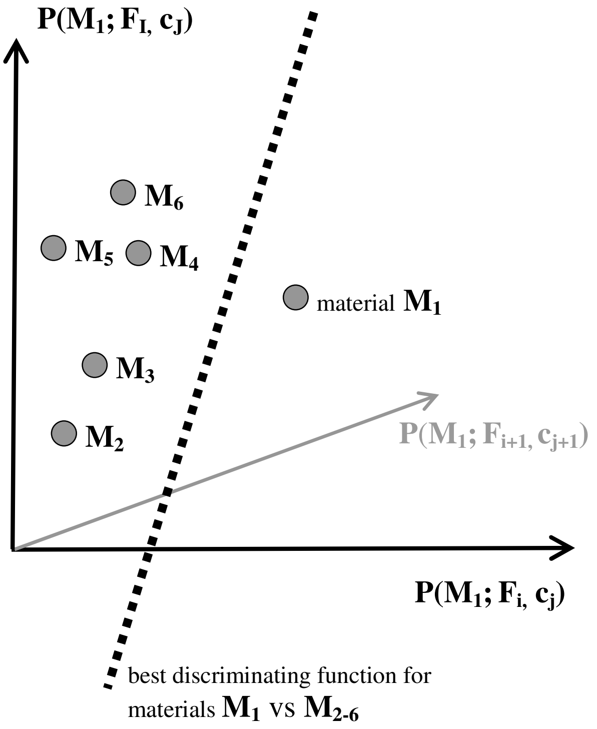


Figure 6.5: Material-specific feature space. The axes represent invididual classifiers trained to distinguish material M_1 from materials $M_{2...6}$ given a filterbank F_i and an (invariant) color representation c_j. The values $P(M_1; F_i, c_j)$ plotted in this new feature space represent the posterior probabilities assigned to the materials $M_{1...6}$ by the individual classifiers. The dotted line indicates the best discriminating function to distinguish material M_1 from materials $M_{2...6}$. The combination of filterbank and color representation (F_i, c_j), on the x-axis, is most in alignment with the discrimination function hence (F_i, c_j) is most discriminative in this example.

6.4 Experiments

In the experiments, we evaluate the color filterbanks and their combination. We take two datasets into account to cover a wide range of real-world materials and imaging conditions under which they can be viewed. First, we consider the well-known CURET dataset [30]. This dataset enables one to test the robustness under varying imaging conditions, i.e. changes of the illumination direction and of the camera viewpoint. For color-based methods, a critical issue is whether the method is robust to color transformations in the image as a consequence of varying illumination color. Second, we consider the ALOT dataset [48] to also include variations of the illumination color. Additionally, this dataset contains more color and 3D variation. Some of the materials that are included in the ALOT dataset are illustrated in Figure 6.1, while some test images are shown in Figure 6.6. In total, we evaluate the filterbanks on 61 textures of the CURET dataset and on 200 textures of the ALOT dataset. In total, in the experiments we use in total $5,612$ CURET images and $7,200$ ALOT images, respectively. For CURET, the train, test and texton learn sets are mentioned in [120];

Figure 6.6: Test images for an example ALOT material.

for ALOT the sets are publicly available on the website of the ALOT database. In the experiments, the number of textons is always set to 200 (as shown in [120] this parameter does not affect the results significantly). For the individual and combined classifiers, we prefer respectively the nearest mean classifier (Euclidean distance) and the linear Bayes-normal classifier [34], as these are performing best.

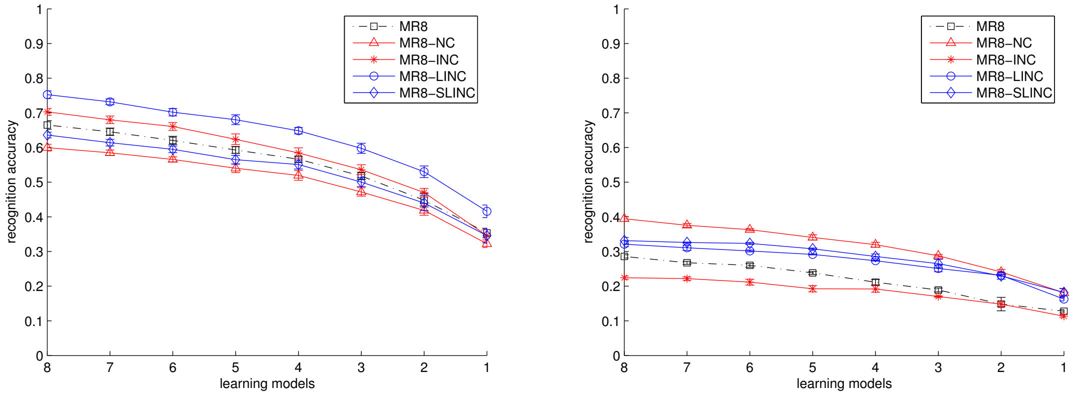
6.4.1 Color Invariant Codebooks

Random Images

We start the performance evaluation by establishing the classification accuracy when selecting randomly the learning images. This experiment gives an indication of the discriminative power and robustness of each of the color filterbanks. We include the original MR8 as a baseline comparison. We consider the mean and standard deviation of classification accuracy over 1,000 repetitions (random selections).

Figures 6.7 (a) and (b) show the recognition results for the CURET and ALOT datasets, respectively. First, we discuss the results for the CURET dataset. The filterbanks with most invariant properties, MR8, MR8-NC and MR8-SLINC filterbanks perform less than the less invariant MR8-INC and MR8-LINC filterbanks. MR8 performs somewhat better than MR8-NC and MR8-SLINC, as it's nearest mean classifier puts all emphasis on the intensity information. With MR8-NC and MR8-SLINC, emphasis of the nearest mean classifier is also put on the color channels, of which almost all information is lost due to the normalization of the mean and variance. The MR8-LINC filterbank performs better than does MR8-INC, as it provides a better approximation of the changing intensity effects by doing so locally.

As expected, for ALOT the performance of the filterbanks is different, as this dataset contains more color and 3D variation. The severe 3D variations causes the intensity to change in such a way that it cannot be approximated well globally. This explains the low performance of the MR8-INC filterbank. At the same time, with much more colorful materials, the global normalization of image colors makes sense: local color variations in the image are now kept albeit relative to each other. Also, the severe 3D variations across materials causes their appearance to change significantly with different illumination. Keeping color variations while being very invariant, explains the good performance of the MR8-NC filterbank. The MR8-INC and MR8-LINC filterbanks are less invariant, hence they perform somewhat less than MR8-NC. The distinctive color information maintained by MR8-INC and MR8-LINC explains their better performance compared to the MR8 filterbank.



(a) CURET

(b) ALOT

Figure 6.7: Accuracy of material recognition for various filterbanks with randomly selected images of (a) the CURET dataset and (b) the ALOT dataset. The vertical bars indicate standard deviation over 1,000 repetitions.

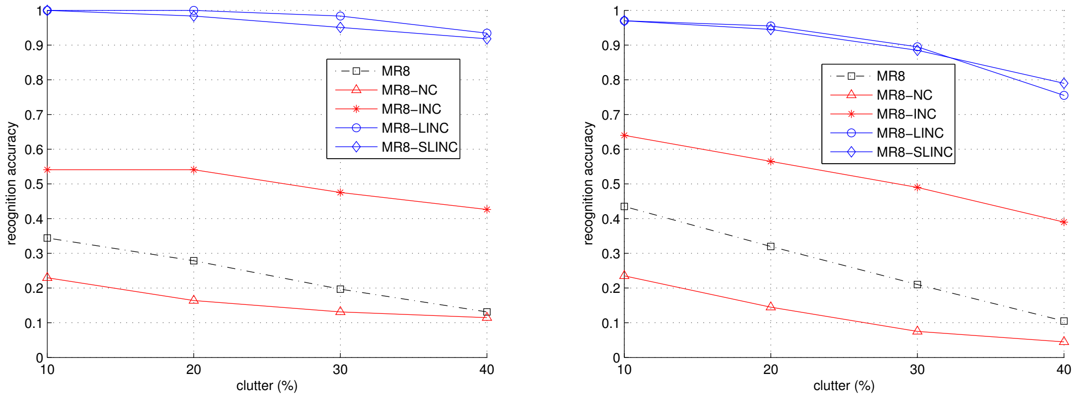
Cluttered Images

Robustness to clutter is of importance for image modelling where the image frame is not fixed, or/and where no image segmentation is available. The setup of the previous experiments involves images that contain no clutter, as the image frames are fixed and each image captures one material only. In this experiment, we evaluate the sensitivity of the color-based filterbanks `MR8`, `MR8-NC`, `MR8-INC` and `MR8-LINC` to clutter.

First, we select randomly one learning image for each texture. Second, we simulate clutter by concatenating the learning image with a randomly selected image of an other texture. For the first cluttered test image, the percentage of original vs. clutter is 90% vs. 10%. To simulate various degrees of clutter, we increase the clutter percentage, up to 40% (note: with 50%, the classification would become chance). The cluttered images are publicly available on the website of the ALOT database[†]. Obviously, for generalization purposes, we use the texton dictionary from the previous experiment (i.e. we do not learn new textons from cluttered images).

Figures 6.8 (a) and (b) show the results for increasingly cluttered images of the CURET and ALOT datasets, respectively. The `MR8-LINC` filterbank performs significantly better than the other filterbanks, `MR8`, `MR8-NC`, and `MR8-INC`, over various degrees of clutter. The low performance of `MR8`, `MR8-NC`, and `MR8-INC` is due to the global normalization schemes that they employ. A global normalization is distorted by clutter, so the filterbank input is different when dealing with variations of clutter. The local normalization employed in `MR8-LINC` is not distorted by clutter. The small performance drop here is due to ambiguity in the images themselves as a result of the cluttering. However, even with 40% clutter, the `MR8-LINC` filterbank achieves a clas-

[†] http://www.science.uva.nl/∼aloi/public_alot



(a) CURET

(b) ALOT

Figure 6.8: Accuracy of material recognition for various filterbanks with increasingly cluttered images of (a) the CURET dataset and (b) the ALOT dataset.

sification accuracy of 75.5% on the ALOT dataset, while the runner-up (MR8-LINC) has an accuracy of 39.0% only.

The results of individual filterbanks are summarized as follows. From the previous two experiments, we conclude that the locally-invariant MR8-LINC and MR8-SLINC filterbanks are very robust to clutter, and that they perform well on different datasets. The MR8-LINC is performing best on the CURET dataset (limited 3D variation), whereas MR8-SLINC performs second-best on the ALOT dataset (severe 3D variation).

6.4.2 Adaptive Color Invariant Codebooks

Since MR8-LINC and MR8-SLINC perform well but on different datasets, and given that the datasets contain very different types of materials, we establish in this experiment whether the tuning of each of the filterbanks to a particular material is beneficial.

As expected, Figures 6.9a and c indicate that the classification accuracy is increased by combining the MR8-LINC and MR8-SLINC filterbanks. While the classification accuracy of MR8-LINC is almost saturated for the CURET dataset, 0.96, the combination achieves a marginal improvement, 2%. For the ALOT dataset, the performance is increased from 0.35 to 0.42 achieving an improvement of 19.8%.

Indeed, as laid down in Figures 6.9b and d, the most distinctive filterbank per material varies significantly across the datasets, and also across the individual materials. The CURET dataset contains many materials of which the structure is similar. Hence, the intensity variation, although very discriminative (see previous experiments), is not *most* discriminative. Rather, color information is most discriminative, as the color channels of the filterbanks are often most distinctive. The information in the filterbanks that are not invariant to shadow and shading, MR8-LINC, is in 56% most distinctive. Most CURET materials are uni-colored, hence the color information is distinctive. With uni-colored materials, too much information is lost when discarding shadow and shading variation. Hence, the shadow and shading invariant filterbank `MR8-SLINC` is in less cases, 27%, most distinctive.

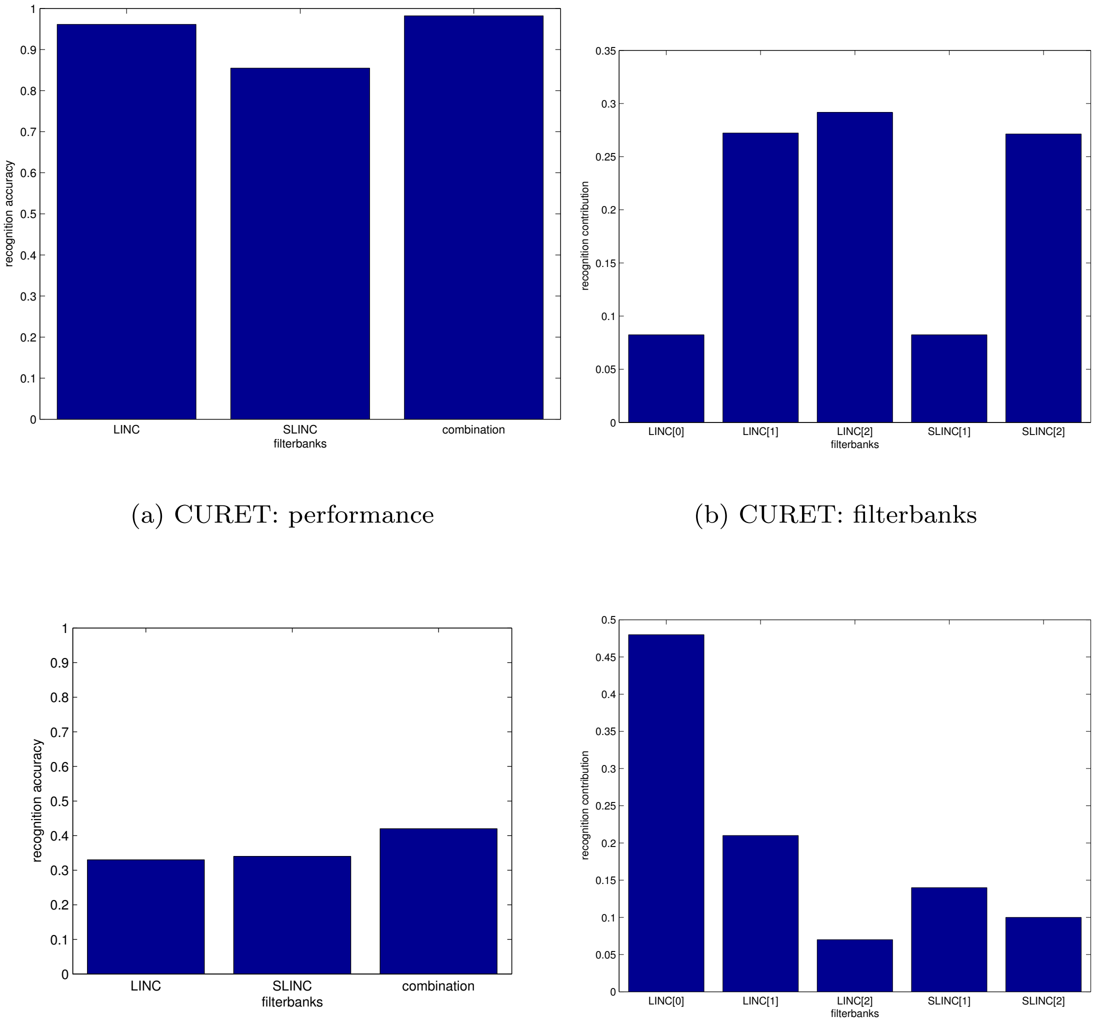
For the ALOT dataset, the performance improvement due to filterbank tuning is significant. As this dataset contains more variation of the material properties, and because more materials are included, the results generalize better. For ALOT the most distinctive filterbanks corresponds to intensity information. This can be explained from the fact that intensity variation rather than color variation is the dominating factor in material appearance [70]. The information in the filterbanks that are not invariant to shadow and shading, `MR8-LINC`, is in 28% most distinctive. The shadow and shading invariant filterbank `MR8-SLINC` is in 25% most distinctive. We conclude that `MR8-LINC` and `MR8-SLINC` are discriminative for large but different sets of materials, respectively.

Finally, we stress that the recognition of materials from the ALOT dataset is obviously a far from solved problem. Here, we have demonstrated the merit of automatically tuning filterbanks with different invariant properties to individual materials with different physical properties.

6.5 Conclusion

In this chapter, we have proposed a framework to learn for each material specifically the most distinctive filterbank from a set of intensity and color invariant filterbanks. The considered filterbanks are adopted from the distinctive `MR8` filterbank of Varma and Zisserman, from which color invariant filterbanks are constructed using techniques from literature. First we have established the distinctiveness, and the robustness to image settings and clutter, for individual filterbanks for the classification of more than 250 materials from the CURET and ALOT datasets, recorded under various illumination directions, viewpoints, and illumination colors. `MR8-NC` is the straightforward extension of `MR8` to color, and likewise it normalizes the mean and variance per color channel. We have shown that this proves to be a good strategy if multiple colors are apparent. `MR8-INC` normalizes each color channel by the variation of the intensity channel. This is a good strategy if the 3D variation of materials is limited. Two color filterbanks normalize locally the filterbank outputs. `MR8-LINC` normalizes locally by the intensity level to counteract intensity fluctuations, whereas `MR8-SLINC` aims at shadow and shading invariant filterbank output. The locally-invariant filterbanks perform on average best, where `MR8-LINC` (`MR8-SLINC`) distinguishes better between materials with limited (significant) 3D variation. Additionally, we have demonstrated that the locally-invariant filterbanks are significantly more robust to image clutter than are filterbanks that involve global normalizations.

Second, we have considered the performance of adapting filterbank combinations to each material specifically. This allows to tune for each material the color channel(s) and invariant properties that discriminates it best from other materials. We have proposed a scheme to do so by learning automatically the best discriminant function



(a) CURET: performance

(b) CURET: filterbanks

(c) ALOT: performance

(d) ALOT: filterbanks

Figure 6.9: Accuracy of material recognition for the best performing filterbanks and their combination for the CURET dataset (a) and the ALOT dataset (c). Percentages indicate how often a particular filterbank is most distinctive (b,d).

in joint filterbank space. Indeed, we have shown that the most distinctive filterbank differs across the CURET and ALOT datasets and across their individual materials. We have demonstrated that this automated tuning of color information and invariance to individual materials results in performance improvements of up to 20%. This result illustrates the merit of tuning a set of invariants to instances that have different physical properties.

Chapter 7

The Distribution Family of Similarity Distances*

7.1 Introduction

In this chapter we derive theoretically and validate experimentally the probability density function family to which similarity distance measures on feature vectors adhere. Knowing the distribution of similarities across between database instances facilitates their search. Databases of documents, images, sounds and other types of data become increasingly large [126]. For images, the commercially available COREL collection [26] of 40,000 images is a well-known example. Also, benchmarks to evaluate content-based retrieval techniques are becoming larger, for instance, the TREC-Video dataset of 184 hours of video [111]. With such large amounts of data, a fundamental issue for retrieval tasks is to index the database contents accurately yet efficiently. To compare database instances, a similarity measure needs to be defined between the descriptors or feature vectors of the instances. The distribution of the similarities from one to other feature vectors is of great practical importance when indexing the dataset [121]. It enables one to confine the search for nearest neighbors of a given feature vector within a given tolerance [4]. This example makes clear that it is fruitful to have a reliable estimation of the range and distribution of similarity values to a feature vector. In this chapter, we derive theoretically under specific but rather general assumptions the distribution family that describes similarity distances from one to other feature vectors. In the experiments, we will limit the scope to feature vectors computed from image and image region descriptors, and we establish whether their distances adhere to the Weibull distribution indeed. We consider SIFT-based features [84], computed from various region types [85]. Furthermore, we consider a global image feature [43] as is used in the TREC-Video benchmark of [114].

This chapter is structured as follows. In Section 2, we overview literature on

*To appear in *Advances in Neural Information Processing Systems*, 2007.

similarity distances and distance distributions. In Section 3, we discuss the theory of distributions of similarity distances from one to other feature vectors. In Section 4, we validate the resulting distribution experimentally for image feature vectors, and in Section 5 experiments are conducted to illustrate consequences of the resulting distribution. The conclusions are given in Section 6.

7.2 Related Work

7.2.1 Similarity Distance Measures

To measure the similarity between two feature vectors, many distance measures have been proposed [82]. A common metric class of measures is the L_p-norm [6]. The distance from one reference feature vector s to one other feature vector t can be formalized as:

$$d_p(s,t) = (\sum_{i=1}^{I} |s_i - t_i|^p)^{1/p}, \tag{7.1}$$

where n and i are the dimensionality and indices of the vectors. Let the random variable D_p represent distances $d_p(s,t)$ where t is drawn from the random variable T representing feature vectors. Independent of the reference feature vector s, the probability density function of L_p-distances will be denoted by $f(D_p = d)$.

7.2.2 Distance Distributions

Ferencz *et al.* [35] have considered the Gamma distribution to model the L_2-distances from image regions to one reference region: $f(D_2 = d) = \frac{1}{\beta^\gamma\,\Gamma(\gamma)}\, d^{\gamma-1}\, e^{-d/\beta}$, where γ is the shape parameter, and β the scale parameter; $\Gamma(\cdot)$ denotes the Gamma function. In [35], the distance function was fitted efficiently from few examples of image regions. Although the distribution fits were shown to represent the region distances to some extent, the method lacks a theoretical motivation.

Based on the central limit theorem, Pekalska and Duin [98] assumed that L_p-distances between feature vectors are normally distributed: $f(D_p = d) = \frac{1}{\sqrt{2\pi}\,\beta}\, e^{-(d^2/\beta^2)/2}$. As the authors argue, the use of the central limit theorem is theoretically justified if the feature values are independent, identically distributed, and have limited variance. Although feature values generally have limited variance, unfortunately, they cannot be assumed to be independent and/or identically distributed as we will show below. Hence, an alternative derivation of the distance distribution function has to be followed.

7.2.3 Contribution of This Chapter

The existing models will be shown to describe the similarity distances not very accurately. Our contribution is to derive a parameterized distribution for L_p-norm distances between feature vectors. And, as a consequence we arrive at a different family of distributions.

7.3 Theory

In this section, we derive the distribution function family of L_p-distances from a reference feature vector to other feature vectors. We consider the notation as used in the previous section, with t a feature vector drawn from the random variable T. For each vector t, we consider the value at index i, t_i, resulting in a random variable T_i. The value of the reference vector at index i, s_i, can be interpreted as a sample of the random variable T_i. The computation of distances from one to other vectors involves manipulations of the random variable T_i resulting in a new random variable: $X_i = |s_i - T_i|^p$. Furthermore, the computation of the distances D requires the summation of random variables, and a reparameterization: $D = (\sum_{i=1}^{I} X_i)^{1/p}$. In order to derive the distribution of D, we start with the statistics of the summation of random variables, before turning to the properties of X_i.

7.3.1 Statistics of Sums

As a starting point to derive the L_p-distance distribution function, we consider a lemma from statistics about the sum of random variables.

Lemma 7.3.1 *For non-identical and correlated random variables X_i, the sum $\sum_{i=1}^{N} X_i$, with finite N, is distributed according to the generalized extreme value distribution, i.e. the Gumbel, Frechet or Weibull distribution.*

For a proof, see [10, 11]. Note that the lemma is an extension of the central limit theorem to non-identically distributed random variables. And, indeed, the proof follows the path of the central limit theorem. Hence, the resulting distribution of sums is different from a normal distribution, and rather one of the Gumbel, Frechet or Weibull distributions instead. This lemma is important for our purposes, as later the feature values will turn out to be non-identical and correlated indeed. To confine the distribution function further, we also need the following lemma.

Lemma 7.3.2 *For Y an upper-bounded random variable, the generalized extreme value distribution is the Weibull distribution:*

$$f(Y = y) = \frac{\gamma}{\beta}(\frac{y}{\beta})^{\gamma-1}\, e^{-(\frac{y}{\beta})^{\gamma}} \quad , \tag{7.2}$$

with γ the shape parameter and β the scale parameter.

For a proof, see [53]. Figure 1 illustrates the Weibull distribution for various shape parameters γ. This lemma is equally important to our purpose, as later the feature values will turn out to be upper-bounded indeed.

The combination of Lemmas 1 and 2 yields the distribution of sums of non-identical, correlated and upper-bounded random variables, summarized in the following theorem.

Theorem 7.3.3 *For non-identical, correlated and upper-bounded random variables X_i, the random variable $Y = \sum_{i=1}^{N} X_i$, with finite N, adheres to the Weibull distribution.*

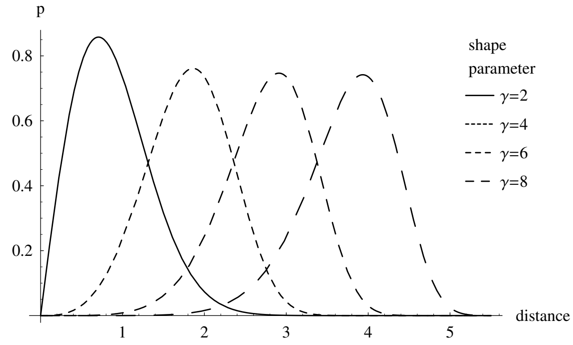


Figure 7.1: Examples of the Weibull distribution for various shape parameters γ.

The proof follows trivially from combining the different findings of statistics as laid down in Lemmas 1 and 2. Theorem 1 is the starting point to derive the distribution of L_p-norms from one reference vector to other feature vectors.

7.3.2 L_p-distances from One to Other Feature Vectors

Theorem 1 states that Y is Weibull-distributed, given that $\{X_i = |s_i - T_i|^p\}_{i \in [1,\dots,I]}$ are non-identical, correlated and upper-bounded random variables. We transform Y such that it represents L_p-distances, achieved by the transformation $(\cdot)^{1/p}$:

$$Y^{1/p} = (\sum_{i=1}^{N} |s_i - T_i|^p)^{1/p}. \tag{7.3}$$

The consequence of the substitution $Z = Y^{1/p}$ for the distribution of Y is a change of variables $z = y^{1/p}$ in Equation 7.2 [97]: $g(Z = z) = \frac{f(z^p)}{(1/p-1)z^{(1-p)}}$. This transformation yields a different distribution still of the Weibull type:

$$g(Z = z) = \frac{1}{(1/p-1)} \frac{\gamma}{\beta^{1/p}} (\frac{z}{\beta^{1/p}})^{p\gamma - 1} e^{-(\frac{z}{\beta^{1/p}})^{p\gamma}}, \tag{7.4}$$

where $\gamma' = p\gamma$ is the new shape parameter and $\beta' = \beta^{1/p}$ is the new scale parameter, respectively. Thus, also $Y^{1/p}$ and hence L_p-distances are Weibull-distributed under the assumed case.

We argue that the random variables $X_i = |s_i - T_i|^p$ and X_j $(i \neq j)$ are indeed non-identical, correlated and upper-bounded random variables when considering a set of values extracted from feature vectors at indices i and j:

- X_i and X_j are upper-bounded. Features are usually an abstraction of a particular type of finite measurements, resulting in a finite feature. Hence, for general feature vectors, the values at index i, T_i, are finite. And, with finite p, it follows trivially that X_i is finite.

- X_i and X_j are correlated. The experimental verification of this assumption is postponed to Section 7.4.1.

- X_i and X_j are non-identically distributed. The experimental verification of this assumption is postponed to Section 7.4.1.

We have obtained the following result.

Corollary 7.3.4 *For finite-length feature vectors with non-identical, correlated and upper-bounded values, L_p distances, for limited p, from one reference feature vector to other feature vectors adhere to the Weibull distribution.*

7.3.3 Extending the Class of Features

We extend the class of features for which the distances are Weibull-distributed. From now on, we allow the possibility that the vectors are preprocessed by a PCA transformation. We denote the PCA transform $g(\cdot)$ applied to a single feature vector as $s' = g(s)$. For the random variable T_i, we obtain T'_i. We are still dealing with upper-bounded variables $X'_i = |s'_i - T'_i|^p$ as PCA is a finite transform. The experimental verification of the assumption that PCA-transformed feature values T'_i and T'_j, $i \neq j$ are non-identically distributed is postponed to Section 7.4.1. Our point here, is that we have assumed originally correlating feature values, but after the decorrelating PCA transform we are no longer dealing with correlated feature values T'_i and T'_j. In Section 7.4.1, we will verify experimentally whether X'_i and X'_j correlate. The following observation is hypothesized. PCA translates the data to the origin, before applying an affine transformation that yields data distributed along orthogonal axes. The tuples (X'_i, X'_j) will be in the first quadrant due to the absolute value transformation. Obviously, variances $\sigma(X'_i)$ and $\sigma(X'_j)$ are limited and means $\mu(X'_i) > 0$ and $\mu(X'_j) > 0$. For data constrained to the first quadrant and distributed along orthogonal axes, a negative covariance is expected to be observed. Under the assumed case, we have obtained the following result.

Corollary 7.3.5 *For finite-length feature vectors with non-identical, correlated and upper-bounded values, and for PCA-transformations thereof, L_p distances, for limited p, from one to other features adhere to the Weibull distribution.*

7.3.4 Heterogeneous Feature Vector Data

We extend the corollary to hold also for composite datasets of feature vectors. Consider the composite dataset modelled by random variables $\{T_t\}$, where each random variable T_t represents non-identical and correlated feature values. Hence, from Corollary 7.3.5 it follows that feature vectors from each of the T_t can be fitted by a Weibull function $f^{\beta,\gamma}(d)$. However, the distances to each of the T_t may have a different range and modus, as we will verify by experimentation in Section 7.4.1. For heterogeneous distance data $\{T_t\}$, we obtain a mixture of Weibull functions [80].

Corollary 7.3.6 (Distance distribution) *For feature vectors that are drawn from a mixture of datasets, of which each results in non-identical and correlated feature values, finite-length feature vectors with non-identical, correlated and upper-bounded values, and for PCA-transformations thereof, L_p distances, for limited p, from one reference feature vector to other feature vectors adhere to the Weibull mixture distribution: $f(D = d) = \sum_{i=1}^{c} \rho_i \cdot f_i^{\beta_i,\gamma_i}(d)$, where f_i are the Weibull functions and ρ_i are their respective weights such that $\sum_{i=1}^{c} \rho_i = 1$.*

7.4 Experiments

For experimentation, we consider image features. First, we validate assumptions and Weibull goodness-of-fit for the region-based SIFT, GLOH, SPIN, and PCA-SIFT features. We include these features for two reasons as: a) they are performing well for realistic computer vision tasks and b) they provide different mechanisms to describe an image region [84]. The region features are computed from regions detected by the Harris- and Hessian-affine regions, maximally stable regions (MSER), and intensity extrema-based regions (IBR) [85]. Also, we consider PCA-transformed versions for each of the detector/feature combinations. Later, we consider also a global image feature. For reason of its extensive use, the experimentation is based on the L_2-distance. We consider distances from 1 randomly drawn reference vector to 100 other randomly drawn feature vectors, which we repeat 1,000 times for generalization. In all experiments, the features are taken from multiple images, except for the illustration in Section 7.4.1 to show typical distributions of distances between features taken from single images.

7.4.1 Validation of the Corollary Assumptions for Image Features

Intrinsic Feature Assumptions

Corollary 7.3.5 rests on a few explicit assumptions. Here we will verify whether the assumptions occur in practice.

- Differences between feature values are correlated. We consider a set of feature vectors T_j and the differences at index i to a reference vector s: $X_i = |s_i - T_{ji}|^p$. We determine the significance of Pearson's correlation [25] between the difference values X_i and X_j, $i \neq j$. We establish the percentage of significantly correlating differences at a confidence level of 0.05. We report for each feature the average percentage of difference values that correlate significantly with difference values at an other feature vector index.

  As expected, the feature value differences correlate. For SIFT, 99% of the difference values are significantly correlated. For SPIN and GLOH, we obtain 98% and 96%, respectively. Also PCA-SIFT contains significantly correlating difference values: 95%. Although the feature's name hints at uncorrelated values, it does not achieve a decorrelation of the values in practice. For each of the features, a low standard deviation $< 5\%$ is found. This expresses the low variation of correlations across the random samplings and across the various region types.

We repeat the experiment for PCA-transformed feature values. Although the resulting values are uncorrelated by construction, their differences are significantly correlated. For SIFT, SPIN, GLOH, and PCA-SIFT, the percentages of significantly correlating difference values are: 94%, 86%, 95%, and 75%, respectively.

- Differences between feature values are non-identically distributed. We repeat the same procedure as above, but instead of measuring the significance of correlation, we establish the percentage of significantly differently distributed difference values X_i by the Wilcoxon rank sum test [25] at a confidence level of 0.05. For SIFT, SPIN, GLOH, and PCA-SIFT, the percentages of significantly differently distributed difference values are: 99%, 98%, 92%, and 87%. For the PCA-transformed versions of SIFT, SPIN, GLOH, and PCA-SIFT, we find: 62%, 40%, 64%, and 51%, respectively. Note that in all cases, correlation is sufficient to fulfill the assumptions of Corollary 7.3.5.

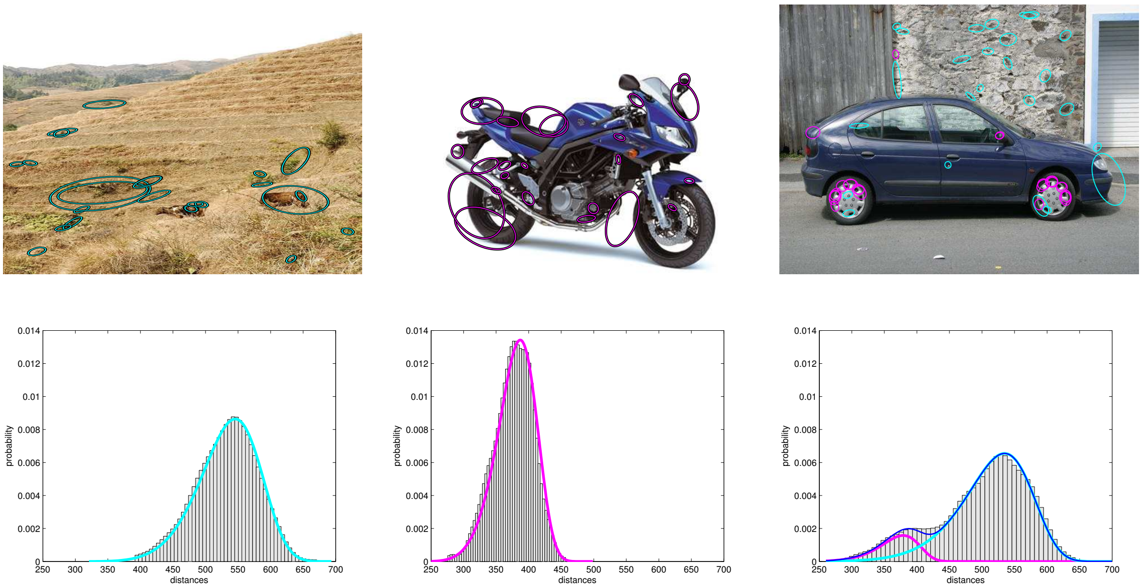
We have illustrated that feature value differences are significantly correlated and significantly non-identically distributed. We conclude that the assumptions of Corollary 7.3.5 about properties of feature vectors are realistic in practice, and that Weibull functions are expected to fit distance distributions well.

Inter-Feature Assumptions

In Corollary 7.3.6, we have assumed that distances from one to other feature vectors are described well by a mixture of Weibulls, if the features are taken from different clusters in the data. Here, we illustrate that clusters of feature vectors, and clusters of distances, occur in practice. Figure 7.2a shows Harris-affine regions from a natural scene which are described by the SIFT feature. The distances are described well by a single Weibull distribution. The same hold for distances from one to other regions computed from a man-made object, see Figure 7.2b. In Figure 7.2c, we illustrate the distances of one to other regions computed from a composite image containing two types of regions. This results in two modalitites of feature vectors hence of similarity distances. The distance distribution is therefore bimodal, illustrating the general case of multimodality to be expected in realistic, heterogeneous image data. We conclude that the assumptions of Corollary 7.3.6 are realistic in practice, and that the Weibull function, or a mixture, fits distance distributions well.

7.4.2 Validation of Weibull-shaped Distance Distributions

In this experiment, we validate the fitting of Weibull distributions of distances from one reference feature vector to other vectors. We consider the same data as before. Over 1,000 repetitions we consider the goodness-of-fit of L_2-distances by the Weibull distribution. The parameters of the Weibull distribution function are obtained by



(a) (b) (c)

Figure 7.2: Distance distributions from one randomly selected image region to other regions, each described by the SIFT feature. The distance distribution is described by a single Weibull function for a natural scene (a) and a man-made object (b). For a composite image, the distance distribution is bimodal (c). Samples from each of the distributions are shown in the upper images.

maximum likelihood estimation. The established fit is assessed by the Anderson-Darling test at a confidence level of $\alpha = 0.05$ [91]. The Anderson-Darling test has also proven to be suited to measure the goodness-of-fit of mixture distributions [89].

Results for Image Region Features

Table 1 indicates that for most of the feature types computed from various regions, more than 90% of the distance distributions is fit by a single Weibull function. As expected, distances between each of the SPIN, SIFT, PCA-SIFT and GLOH features, are fitted well by Weibull distributions. The exception here is the low number of fits for the SIFT and SPIN features computed from Hessian-affine regions. The distributions of distances between these two region/feature combinations tend to have multiple modes. Likewise, there is a low percentage of fits of L_2-distance distributions of the SPIN feature computed from IBR regions. Again, multiple modes in the distributions are observed. For these distributions, a mixture of two Weibull functions provides a good fit ($\geq 97\%$).

Global Image Feature

We consider the dataset collected by Snoek *et al.* [114], describing 101 concepts from video data. For each concept, 30,000 images have been collected. From these images,

Table 7.1: Accepted Weibull fits for COREL data [26].

	Harris-affine		Hessian-affine		MSER		IBR	
	$c=1$	$c \leq 2$	$c=1$	$c \leq 2$	$c=1$	$c \leq 2$	$c=1$	$c \leq 2$
SIFT	95%	100%	60%	99%	98%	100%	92%	100%
SIFT (g =PCA)	95%	99%	60%	98%	98%	100%	92%	99%
PCA-SIFT	89%	100%	96%	100%	94%	100%	95%	100%
PCA-SIFT (g =PCA)	89%	100%	96%	100%	94%	100%	95%	100%
SPIN	71%	99%	12%	99%	77%	99%	45%	98%
SPIN (g =PCA)	71%	100%	12%	97%	77%	99%	45%	98%
GLOH	87%	100%	91%	100%	82%	99%	86%	100%
GLOH (g =PCA)	87%	100%	91%	99%	82%	99%	86%	100%

Percentages of L_2-distance distributions fitted by a Weibull function ($c = 1$) and a mixture of two Weibull functions ($c \leq 2$) are given.

Table 7.2: Accepted Weibull fits for TRECVID data of the benchmark in [114]

	Positive examples		Negative examples	
	$c=1$	$c \leq 2$	$c=1$	$c \leq 2$
outdoor	93%	98%	89%	98%
maps	59%	97%	92%	99%
crowd	89%	99%	92%	100%
building	92%	99%	90%	99%

Percentages of Weibull fits for distributions of L_2-distances from one to other images. Distributions are fit with a Weibull function ($c = 1$) and a mixture-of-Weibull functions ($c \leq 2$).

$1-10\%$ are annotated as positive examples of the concept. The remaining images are annotated as negative examples. Each image is described by a vector of 120 statistical image features, see [43]. We consider two natural concepts, outdoor and crowd, and two man-made concepts, maps and buildings.

Results. For the negative examples, on average $90\% \pm 2\%$ of the distributions of distances fit well to the Weibull distribution. For positive examples, the percentage of fitted distributions is: $83 \pm 16\%$. For maps, the fit percentage is low: 59%. The mixture of two to four Weibulls provides generally a good fit for $> 97\%$ of the cases, as indicated in Table 2.

7.5 Consequence

We consider the estimation of quantiles in the distance distribution. Quantiles are important to determine, amongst others, they are used in the determination of: a) the distance to be set to establish N nearest neighbors, b) the median of the distances required for balanced construction of indexing trees, and c) critical values for the detection of outliers. Our point here is that for quantiles, it is important to know the skewness. For non-skewed distributions, occurring in 53% of the cases found by experimentation in the previous section (including the components of mixtures), the normal distribution suffices [98]. For $>$ 90% of the cases the Weibull distribution provides a better alternative. The Weibull distribution is preferred here, as the distributions are in practice often skewed significantly: skewed positively ($\gamma \leq 2.5$ represents 15% of the cases) and skewed negatively ($\gamma \geq 6$ representing 32%). Together with ones that are marginally skewed ($2.5 < \gamma < 6$ representing 53%), we divide the distributions in three categories, represented by respectively $\gamma = 2$, 4 and 8. For each category we consider 1,000 distributions, fitted by both the Weibull and the normal distribution. Subsequently, we determine from their parameters the distance threshold to be set to retrieve the objective quantile of the features. We measure the error in the percentage of retrieved features compared to the objective quantile.

Without loss of generalization, we limit ourselves to the single-component distributions here. Note that multi-component distributions are linear combinations of single-components.

Results. Over 1,000 repetitions, we have considered the minimum and maximum of retrieved features. For both values, we indicate in Table 3 the error relative to the objective quantile of features to be retrieved. First, we discuss the median of similar instances to be retrieved. Obviously, the Weibull distribution provides generally a good estimation of the quantiles (error $\leq$ 0.4%). For positively skewed distributions, $\gamma = 2$, the normal distribution produces an overestimation of the median value: $\geq$ 8.4%. For marginally skewed distributions, $\gamma = 4$, the median is systematically underestimated by the normal distribution: bias $\approx -2.0\%$. For negatively skewed distributions, the median is systematically underestimated: bias $\approx -7.5\%$.

Second, we discuss the estimation of the tail properties, that is the retrieval of the 1% most similar features to a randomly selected feature. Estimating tail properties is noisier then estimating the median value: the relative error of the Weibull estimation increases but is at maximum 10%. This means that for a database of $1,000,000$ features, the objective is to retrieve $10,000$ most similar features, while the estimation retrieves $10,000 \pm 1,000$ features. Next, we turn to the results obtained for the normal distribution in more detail. For $\gamma = 4$, the retrieval of most similar features is comparable to that of the Weibull distribution. For negatively skewed distributions, by example of $\gamma = 8$, systematically too many instances are retrieved when requesting the 1% most similar ones: at least twice the number of requested features is retrieved (error of 100%). For the example above, this means that $10,000$ features too many are retrieved. Moreover, for positively skewed distributions, by example of $\gamma = 2$, no instances are retrieved at all when requesting the 1% most similar ones. In recapitulation, the Weibull distribution retrieves the most similar features accurately up to only

Table 7.3: Accuracy of Quantile Estimation

Shape parameter of distribution	Representative for % of the cases	Fitting function	[min, max] relative error at median	at 1% most similar
$\gamma = 2$	15%	Weibull	$[-0.4\%, 0.4\%]$	$[-10\%, 10\%]$
		normal	$[8.4\%, 9.4\%]$	[none*,none*]
$\gamma = 4$	53%	Weibull	$[-0.2\%, 0.4\%]$	$[-10\%, 10\%]$
		normal	$[-1.8\%, -2.2\%]$	$[-10\%, 10\%]$
$\gamma = 8$	32%	Weibull	$[-0.4\%, 0.4\%]$	$[-10\%, 10\%]$
		normal	$[-8.2\%, -7.2\%]$	$[100\%, 120\%]$

For distance distributions of various shapes (γ), we establish quantiles from the parameters of the fitted function (Weibull and normal distribution). The second column indicates the percentages of cases for the SIFT, PCA-SIFT, SPIN and GLOH features computed from the COREL dataset. Over 1,000 repetitions, minimum and maximum relative errors are given for the percentage of features that are retrieved. (∗): none of the features are retrieved in this case.

a 10% deviation, while for the normal distribution the errors range from no retrieved features at all up to twice the amount of features to be retrieved. We conclude that the Weibull distribution is the preferred choice for determining the distance bounds in which a given number of nearest neighbors to a given feature are to be expected.

7.6 Conclusion

In this chapter, we have derived that similarity distances between one and other image features in databases are Weibull distributed. Indeed, for various types of features, i.e. the SPIN, SIFT, GLOH and PCA-SIFT features, and a global feature of an image, and for two datasets, i.e. the COREL image collection and keyframes from TRECVID video data, we have demonstrated that the similarity distances from one to other features, computed from L_p norms, are Weibull-distributed. These results are established by the experiments presented in Table 1. Also, between PCA-transformed feature vectors, the distances are Weibull-distributed. Furthermore, when the dataset is a composition, a mixture of few (typically two) Weibull functions suffices, as established by the experiments presented in Tables 1 and 2.

The resulting Weibull distributions are distinctively different from the distributions suggested in literature, as they are positively or negatively skewed while the Gamma [35] and normal [98] distributions are positively and non-skewed, respectively. We have shown by experiments presented in Table 3 that the approximation of distance distributions by a normal distribution yields systematic and unacceptable errors in almost 50% of all cases. We demonstrate that the Weibull distribution is the preferred choice for estimating properties of sets of similarity distances. In particular, the Weibull distribution estimates accurately the distance, below which to expect the most similar instances to an image feature.

Chapter 8

Summary and Conclusions

8.1 Summary

In this thesis, we have explored the quality of image features, each with a specific degree of invariance to the conditions that dominate the appearance of objects in images. The results obtained in the thesis are discussed per chapter in the following paragraphs:

Chapter 2: Quality of Variant and Invariant Features for Color Image Processing. In this chapter, we evaluate color invariants in terms of properties that are important to image processing. We start with an overview of various invariants with different invariant properties. They are based on Gaussian filters and are a measure of color and local shape. We shortly overview how invariants can be derived from the Gaussian variant measurements. Our contribution is to assess the quality of variants and invariants.

In spite of their nonlinear nature, we demonstrate the invariant features to perform nearly as well in terms of localization and stability to low image intensity and JPEG compression as the well-known Gaussian filters. Furthermore, we show that the responses of invariants – with the exception of the hue-based invariant – are low under irrelevant scene variations, while they are shown to covary with relevant image variation.

Chapter 3: Performance Evaluation of Local Color Invariants. In this chapter, we compare local color descriptors to grey-value descriptors. We adopt the evaluation framework of Mikolayzcyk and Schmid [84, 85]. We modify the framework in several ways. We decompose the evaluation framework to the level of local grey-value invariants on which common region descriptors are based. We compare the discriminative power and invariance of grey-value invariants to that of color invariants. In addition, we evaluate the invariance of color descriptors to photometric events such as shadow and highlights. We measure the performance over an extended range of common recording conditions including significant photometric variation.

We demonstrate the intensity-normalized color invariants and the shadow invariants to be highly distinctive, while the shadow invariants are more robust to both changes of the illumination color, and to changes of the shading and shadows. Overall, the shadow invariants perform best: they are most robust to various imaging conditions while maintaining discriminative power. When plugged into the SIFT descriptor, they show to outperform other methods that have combined color information and SIFT. The usefulness of `C-color-SIFT` for realistic computer vision applications is illustrated for the classification of object categories from the VOC challenge [134], for which a significant improvement is reported.

Chapter 4: Quasi-periodic Spatio-temporal Filtering This chapter presents the online estimation of temporal frequency to simultaneously detect and identify the quasi-periodic motion of an object. We introduce color to increase discriminative power of a reoccurring object and to provide robustness to appearance changes due to illumination changes. Spatial contextual information is incorporated by considering the object motion at different scales. We combined spatiospectral Gaussian filters and a temporal reparameterized Gabor filter to construct the online temporal frequency filter.

We demonstrate the online filter to respond faster and decay faster than offline Gabor filters. Further, we show the online filter to be more selective to the tuned frequency than Gabor filters. We contribute to temporal frequency analysis in that we both identify (“what”) and detect (“when”) the frequency. In color video, we demonstrate the filter to detect and identify the periodicity of natural motion. The velocity of moving gratings is determined in a real world example. We consider periodic and quasi-periodic motion of both stationary and non-stationary objects.

Chapter 5: Color Textons for Texture Recognition Texton models have proven to be very discriminative for the recognition of grayvalue images taken from rough textures. To further improve the discriminative power of the distinctive texton models of Varma and Zisserman (`VZ` model) (IJCV, vol. 62(1), pp. 61-81, 2005), we propose two schemes to exploit color information. First, we incorporate color information directly at the texton level, and apply color invariants to deal with straightforward illumination effects as local intensity, shading and shadow. But, the learning of representatives of the spatial structure *and* colors of textures may be hampered by the wide variety of apparent structure-color combinations. Therefore, our second contribution is an alternative approach, where we weight grayvalue-based textons with color information in a post-processing step, leaving the original `VZ` algorithm intact.

We demonstrate that the color-weighted textons outperform the `VZ` textons as well as the color invariant textons. The color-weighted textons are specifically more discriminative than grayvalue-based textons when the size of the example image set is reduced. When using 2 example images only, recognition performance is 85.6%, which is an improvement over grayvalue-based textons of 10%. Hence, incorporating color in textons facilitates the learning of textons.

Chapter 6: Material-specific Adaptation of Color Invariant Features. For the modelling of materials, the mapping of image features onto a codebook of feature representatives receives extensive treatment. For reason of their generality and simplicity, filterbank outputs are in use as features. The MR8 filterbank of

Varma and Zisserman is performing well in a recent evaluation [120]. In this chapter, we construct color invariant filter sets from the original MR8 filterbank. We evaluate several color invariant alternatives over more than 250 real-world materials recorded under a variety of imaging conditions including clutter. Our contribution is a material recognition framework that learns automatically for each material specifically the most discriminative filterbank combination and corresponding degree of color invariance. For a large set of materials each with different physical properties, we demonstrate the material-specific filterbank models to be preferred over models with fixed filterbanks.

Chapter 7: The Distribution Family of Similarity Distances. We report that the L_p-norms – a class of commonly applied distance metrics – from one feature vector to other vectors are Weibull-distributed if the feature values are correlated and non-identical. Although these properties are common for different types of multimedia features in practice, we illustrate them to hold for the domain of image features. We consider image features that are commonly used in the realm of computer vision tasks.

In the experiments, we show that the values of image features are correlated and non-identical, and demonstrate that the Weibull distribution describes the distances between them very well. The Weibull distribution estimates accurately the distance below which to expect the most similar instances to an image feature.

8.2 Conclusions

In this thesis, we have investigated the quality of variant and invariant features. Various reasonable degrees of invariance that can be investigated on a local scale have been considered. As a consequence, the thesis makes a contribution to low-level image analysis: it provides solutions to the quest which local features to choose for more or less specific problems.

First, we conclude on which features perform well generally and which do not. From Chapters 2 we conclude that the hue-based invariant is not stable, hence we exclude it from further general conclusions. As expected, we have found in Chapter 2 that for any invariant the more invariant a feature is, the less stable it becomes, but also the more robustness to variations of the accidental conditions is gained. We conclude that an intermediate level of invariance is suited for many tasks. We have found in Chapters 2, 3 and 5 the intensity-normalized image derivatives to be very descriptive of image edges and yet robust to imaging conditions. Similarly, in Chapter 6 the intensity-normalized filterbank has proven more distinctive for texture modelling than are other filterbanks with different variant or invariant properties.

The success of the measurement of image edges by invariants based on intensity-normalization may be explained by the dominating appearance effects. Image edges are highly localized, and their appearance is dominated by surface orientation relative to the illumination direction and camera viewpoint. For changes of these parameters, but also for a change of camera gain or illumination intensity, the main appearance effect is a change of local intensity. We conclude that these effects are suppressed well by a straightforward normalization of intensity.

Second, we conclude that for region-based descriptors, rather than point-based features, other observation effects start to play a role. Regions descriptors are still localized but have a larger support area than have image edges. As a consequence, they are more probable to be perturbed by shadow and shading effects. For region-based descriptors, we have found in Chapter 3 that indeed shadow and shading invariants are the preferred features. This is an interesting result, as many state-of-the-art applications in computer vision are built upon region-based descriptors.

Third, we have contributed to the understanding of statistics of similarity. In Chapter 7, we have derived the distribution family of similarity distances between one and other descriptors in a dataset. Comparison to and retrieval of images or image regions is a fundamental issue to be resolved in many computer vision and image retrieval applications. We conclude from our results that knowing the statistics of similarities improves significantly the retrieval of similar descriptors.

Fourth, we have addressed in Chapter 6 the problem of selecting invariants not only for the task at hand but also for the object at hand. The proposed framework is a first attempt to maximize the information about an object in terms of extracted features. The improvements in performance are promising, which shows that object-specific representation is an fruitful direction for future work.

Bibliography

Bibliography

[1] A. E. Abdel-Hakim and A. A. Farag. Csift: A sift descriptor with color invariant characteristics. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 1978–1983, 2006.

[2] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2):284–299, 1985.

[3] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In M. Landy and J. A. Movshon, editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press, Cambridge, 1991.

[4] A. S. Arantes, M. R. Vieira, A. J. M. Traina, and C. Traina. The fractal dimension making similarity queries more efficient. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 12–17, 2003.

[5] R. Basri and D. W. Jacobs. Recognition using region correspondences. *International Journal of Computer Vision*, 25(2):145–166, 1997.

[6] B. G. Batchelor. *Pattern Recognition: Ideas in Practice*. Plenum Press, New York, 1995.

[7] A. Baumberg. Reliable feature matching ocross widely separated views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.

[8] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision*, 2006.

[9] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:509–522, 2002.

[10] E. Bertin. Global fluctuations and gumbel statistics. *Physical Review Letters*, 95(170601):1–4, 2005.

[11] E. Bertin and M. Clusel. Generalised extreme value statistics and sum of correlated variables. *Journal of Physics A*, 39:7607, 2006.

[12] E. de Boer and H. R. de Jong. On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America*, 63:115–135, 1978.

[13] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, 2006.

[14] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:55–73, 1990.

[15] R. E. Broadhurst. Statistical estimation of histogram variation for texture classification. In *Proceedings of Texture 2005*, pages 25–30, 2005.

[16] G. J. Burghouts.

[17] G. J. Burghouts and J. M. Geusebroek. Color textons for texture recognition. In *Proceedings of Brittish Machine Vision Conference*, volume 3, pages 1099–1108, 2006.

[18] G. J. Burghouts and J. M. Geusebroek. Quasi-periodic spatio-temporal filtering. *IEEE Transactions on Image Processing*, 15(6):1572–1582, 2006.

[19] H. Burkhardt and S. Siggelkow. Invariant features in pattern recognition - fundamentals and applications. In C. Kotropoulos and I. Pitas, editors, *Nonlinear Model-Based Image/Video Processing and Analysis*, pages 269–307. John Wiley and Sons, 2001.

[20] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[21] D. Charalampidis and T. Kasparis. Wavelet-based rotational invariant roughness features for texture classification and segmentation. *IEEE Transactions on Image Processing*, 11(8):825–837, 2002.

[22] J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz. Adaptive perceptual color-texture image segmentation. *IEEE Transactions on Image Processing*, 14(10):1524–1536, 2005.

[23] F. Cheng, W. J. Christmas, and J. Kittler. Recognising human running behavior in sports video sequences. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 1017–1020, 2002.

[24] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, 2003.

[25] W. J. Conover. *Practical nonparametric statistics.* Wiley, New York, 1971.

[26] Corel Gallery. www.corel.com.

[27] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proceedings of the European Conference on Computer Vision*, 2004.

[28] O. G. Cula and K. J. Dana. 3d texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*, 59(1):33–60, 2004.

[29] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000.

[30] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real world surfaces. *ACM Transactions on Graphics*, 18(1):1–34, 1999.

[31] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences*, 18:451–458, 1995.

[32] G. Doretto, A. Chiuso, S. Soatto, and Y. N. Wu. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.

[33] G. Doretto, D. Cremers, P. Favaro, and S. Soatto. Dynamic texture segmentation. In *Proceedings of the International Conference Computer Vision*, pages 1236–1242. IEEE Computer Society, 2003.

[34] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley, New York, 2000.

[35] A. Ferencz, E.G. Learned-Miller, and J. Malik. Building a classification cascade for visual identification from one example. In *Proceedings of the International Conference Computer Vision*, pages 286–293. IEEE Computer Society, 2003.

[36] G. D. Finlayson. Color in perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):1034–1038, 1996.

[37] G. D. Finlayson, M. S. Drew, and B. Funt. Color constancy: generalized diagonal transforms suffice. *Journal of the Optical Society of America A*, 11(11):3011–3019, 1994.

[38] G. D. Finlayson, S. D. Hordley, and P. M. Hubel. Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1209–1221, 2001.

[39] L. M. J. Florack. *Image Structure*, volume 10 of *Computational Imaging and Vision Series.* Kluwer Academic Publishers, Dordrecht, 1997.

[40] L. M. J. Florack, B. ter Haar Romeny, M. Viergever, and J. J. Koenderink. The gaussian scale-space paradigm and the multiscale local jet. *International Journal of Computer Vision*, 18:61–75, 1996.

[41] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[42] B. V. Funt and G. D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

[43] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A.W.M. Smeulders. Robust scene categorization by learning image statistics in context. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2006.

[44] T. Geodeme, T. Tuytelaars, G. Vanacker, M. Nuttin, and L. Van Gool. Omnidirectional sparse visual path following with occlusion-robust feature tracking. In *Proceedings of the International Conference on Computer Vision*, 2005.

[45] R. Gershon, D. Jepson, and J. K. Tsotsos. Ambient illumination and the determination of material changes. *Journal of the Optical Society of America A*, 3:1700–1707, 1986.

[46] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[47] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

[48] J. M. Geusebroek and A. W. M. Smeulders. The Amsterdam library of object textures. submitted.

[49] J. M. Geusebroek and A. W. M. Smeulders. Fragmentation in the vision of scenes. In *Proceedings of the International Conference Computer Vision*, pages 130–135. IEEE Computer Society, 2003.

[50] Th. Gevers and A. W. M. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.

[51] Th. Gevers and H. M. G. Stokman. Classification of color edges in video into shadow-geometry, highlight, or material transitions. *IEEE Transactions on Multimedia*, 5(2):237–243, 2003.

[52] M. Grabner, H. Grabner, and H. Bischof. Fast approximated sift. In *Proceedings of the Asian Conference Computer Vision*, number 1, pages 918–927, 2006.

[53] E. J. Gumbel. *Statistics of Extremes.* Columbia University Press, New York, 1958.

[54] C. Harris and M. Stephans. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 189–192, Manchester, 1988.

[55] E. Hayman, B. Caputo, M. Fritz, and J. O. Eklundh. On the significance of real-world conditions for material classification. In *Proceedings of the European Conference Computer Vision*, number 3, pages 253–266. Springer Verlag, 2004.

[56] G. Heidemann. Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):817–830, 2004.

[57] E. Hering. *Outlines of a Theory of the Light Sense.* Harvard University Press, Cambridge, 1964.

[58] M. A. Hoang, J. M. Geusebroek, and A. W. M. Smeulders. Color texture measurement and segmentation. *Signal Processing*, 85(2):265–275, 2005.

[59] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 16:185–203, 1981.

[60] T. Irino and R. D. Patterson. A time-domain, level-dependent auditory filter: the gammachirp. *Journal of the Acoustical Society of America*, 101:412–419, 1997.

[61] A. Jain and G. Healey. A multiscale representation including opponent color features for texture recognition. *IEEE Transactions on Image Processing*, 7(1):124–128, 1998.

[62] A. Johnson and M. Hebert. Object recognition by matching oriented points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 684–689, 1997.

[63] D. Judd and G. Wyszecki. *Color in Business, Science, and Industry.* John Wiley and Sons, 1975.

[64] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of International Conference on Computer Vision*, 2005.

[65] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal on Computer Vision*, 2(45):83–105, 2001.

[66] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2005.

[67] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[68] J. J. Koenderink. Operational significance of receptive field assemblies. *Biological Cybernetics*, 58:163–171, 1988.

[69] J. J. Koenderink. Scale-time. *Biological Cybernetics*, 58:159–162, 1988.

[70] J. J. Koenderink, A. J. van Doorn, K. J. Dana, and S. Nayar. Bidirectional reflection distribution function of thoroughly pitted surfaces. *International Journal of Computer Vision*, 31:129–144, 1999.

[71] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 63:291–297, 1987.

[72] P. Kruizinga and N. Petkov. Nonlinear operator for oriented texture. *IEEE Transactions on Image Processing*, 8(10):1395–1407, 1999.

[73] P. Kubelka. New contributions to the optics of intensely light-scattering materials. *Journal of the Optical Society of America*, 38:448–457, 1948.

[74] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005.

[75] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(8):1265–1278, 2005.

[76] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[77] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–154, 1998.

[78] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[79] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[80] J. M. Marin, M. T. Rodriquez-Bernal, and M. P. Wiper. Using weibull mixture distributions to model heterogeneous survival data. *Communications in statistics*, 34(3):673–684, 2005.

[81] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceeding of the British Machine Vision Conference*, pages 384–393, 2002.

[82] R. S. Michalski, R. E. Stepp, and E. Diday. A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts. In L. N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, pages 33–56. North-Holland Publishing Co., Amsterdam, 1981.

[83] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[84] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[85] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[86] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94:3–27, 2004.

[87] P. Montesinos, V. Gouet, R. Deriche, and D. Pel. Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, 18(9):659–671, 2000.

[88] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. In *Proceedings of the International Conference on Computer Vision*, volume 1, pages 800–807, 2005.

[89] K. Mosler. Mixture models in econometric duration analysis. *Applied Stochastic Models in Business and Industry*, 19(2):91–104, 2003.

[90] H. T. Nguyen and A. W. M. Smeulders. Fast occluded object tracking by a robust appearance filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):1099–1104, 2004.

[91] NIST/SEMATECH. *e-Handbook of Statistical Methods.* NIST, http://www.itl.nist.gov/div898/handbook/, 2006.

[92] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of the European Conference on Computer Vision.* Springer Verlag, 2006.

[93] S. Odbrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proceedings of the Brittish Machine Vision Conference*, 2002.

[94] B. A. Olshausen and K. N. O'Conner. A new window on sound. *Nature Neuroscience*, 5(4):292–294, 2002.

[95] P. Olver, G. Sapiro, and A. Tannenbaum. Differential invariant signatures and flows in computer vision: A symmetry group approach. In B. M. ter Haar Romeny, editor, *Geometry Driven Diffusion in Computer Vision.* Kluwer Academic, Boston, 1994.

[96] PANTONE. ed. 1992-1993, groupe basf, paris, france. pantone is a trademark of patone inc.

[97] A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, New York, 4 edition, 2002.

[98] E. Pekalska and R. P. W. Duin. Classifiers for dissimilarity-based pattern recognition. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, page 2012, 2000.

[99] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *Proceedings of the European Conference on Computer Vision.* Springer Verlag, 2006.

[100] R. Polana and R. C. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, 23(3):261–282, 1997.

[101] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66:231–259, 2006.

[102] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *Proceedings of the European Conference on Computer Vision*, pages 414–131, 2002.

[103] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[104] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[105] F. J. Seinstra and D. Koelma. User transparancy: a fully sequential programming model for efficient data parallel image processing. *Concurrency and computation: practice and experience*, 16:611–644, 2004.

[106] S. A. Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985.

[107] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European Conference on Computer Vision*. Springer Verlag, 2006.

[108] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[109] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, 2003.

[110] D. Slater and G. Healey. The illumination-invariant recognition of 3d objects using local color invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(2):206–210, 1996.

[111] A. F. Smeaton, P. Over, and W. Kraaij. TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. *ACM Multimedia*, 2004.

[112] A. W. M. Smeulders, J. M. Geusebroek, and T. Gevers. Invariant representation in image processing. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 18–21. IEEE Computer Society, 2001.

[113] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[114] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM Multimedia*, pages 421–430, 2006.

[115] Y. Song, L. Goncalves, E. Di Bernardo, and P. Perona. Monocular perception of biological motion - detection and labeling. In *Proceedings of the International Conference Computer Vision*, pages 805–812. IEEE Computer Society, 1999.

[116] P. Suen and G. Healey. The analysis and recognition of real-world textures in three dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(5):491–503, 2000.

[117] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[118] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic motion detection for motion based recognition. *Pattern Recognition*, 27(12):1591–1603, 1994.

[119] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[120] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1-2):61–81, 2005.

[121] J. Venkateswaran, D. Lachwani, T. Kahveci, and C. Jermaine. Reference-based indexing of sequence databases. In *Proceedings of the International Conference on Very Large Data Bases*, pages 906–917, 2006.

[122] J. van de Weijer and Th. Gevers. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[123] J. van de Weijer, Th. Gevers, and J.-M. Geusebroek. Color edge and corner detection by photometric quasi-invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):325–630, 2005.

[124] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *European Conference on Computer Vision*, pages 334–348. Springer, 2006.

[125] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the International Conference Computer Vision*, pages 1800–1807. IEEE Computer Society, 2005.

[126] I. H. Witten, T. C. Bell, and A. Moffat. *Managing Gigabytes.: Compressing and Indexing Documents and Images*. Morgan Kaufmann, 1999.

[127] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5:1–7, 2004.

[128] I. T. Young and L. J. van Vliet. Recursive implementation of the gaussian filter. *Signal Processing*, 44(2):139–151, 1995.

[129] I. T. Young, L. J. van Vliet, and M. van Ginkel. Recursive gabor filtering. *IEEE Transactions on Signal Processing*, 50(11):2798–2805, 2002.

[130] R. A. Young. The gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles. Technical Report GMR-4920, General Motors Research Center, Warren, MI, 1985.

[131] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondance. In *Proceedings of the European Conference on Computer Vision*, pages 151–158, 1994.

[132] S. Di Zenzo. Note: a note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33(1):116–125, 1986.

[133] J. Zhang, M. Marsza, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, page to appear, 2006.

[134] A. Zisserman, M. Everingham, C. Williams, and L. Van Gool. The pascal visual object classes challenge 2006 (voc2006), 2006.

Samenvatting[*]

In dit proefschrift is de kwaliteit van beeldkenmerken, elk met een specifieke invariantie voor effecten in het beeld, verkend. Het scala aan invariante eigenschappen betreft diegene die toe zijn te schrijven aan toevallige omstandigheden van de visuele scene zoals die in het beeld vastgelegd is. De resultaten van de evaluatie van beeldkenmerken worden besproken per hoofdstukken in de volgende paragrafen:

Hoofdstuk 2: Quality of Variant and Invariant Features for Color Image Processing. In dit hoofdstuk zijn kleurinvarianten geevalueerd op basis van eigenschappen die belangrijk zijn voor beeldverwerking. Voorafgaand aan de evaluatie is een kader opgesteld met daarin fysische parameters van de beeldformatie, Gaussische filters die kleur, vorm en textuur in het beeld meten, en invarianten gebaseerd op de filters om effecten in de beeldformatie tegen te gaan. De bijdrage van dit hoofdstuk is de evaluatie van invarianten.

Ondanks de nonlineaire bepaling van de invarianten, zijn de invarianten bijna net zo stabiel onder lage beeldintensiteit en JPEG compressie als de Gaussische filtermetingen, kortweg varianten genoemd. Verder hebben de invarianten, met uitzondering van de hue-gebaseerde invariant, inderdaad een lage respons onder de irrelevante variaties in het beeld, terwijl ze covarieren met relevante beeldinformatie.

Hoofdstuk 3: Performance Evaluation of Local Color Invariants. In dit hoofdstuk worden kleurinvarianten als lokale descriptors van het beeld vergeleken met grijswaarde descriptors. De evaluatie methodologie van Mikolayzcyk and Schmid [84, 85] wordt aangepast, met als doel de basis van descriptors, namelijk de onderliggende gradient-metingen, te vergelijken. Zowel het onderscheidend vermogen als invariante eigenschappen onder scene variatie en onder specifieke variatie zoals schaduw en speculariteit worden geevalueerd. Uitgangspunt is hier het nut van de beeldkenmerken voor het beschrijven van alledaagse, 3-dimensionale objecten in beelden.

Intensiteits-genormaliseerde metingen en schaduw invarianten blijken zowel zeer beschrijvend als robust tegen de meest voorkomende variaties in alledaagse opnames.

[*] Summary, in Dutch.

De schaduw invarianten zijn te prefereren als significante variatie van belichtingsrichting is te verwachten. Deze invariant is ook in de veelgebruikte SIFT descriptor [79] ingebouwd, waarmee aangetoond wordt dat het classificeren van objecten uit de uitdagende VOC dataset [134] significant verbeterd kan worden.

Hoofdstuk 4: Quasi-periodic Spatio-temporal Filtering Dit hoofdstuk presenteert, binnen de Gaussische set van beeldkenmerken, een *online* filter voor het meten van repeterende objectbeweging. Het doel is om zowel frequentie te meten als te detecteren, d.w.z. te lokaliseren in de tijdsdimensie. Kleur wordt meegenomen als extra kenmerk om onderscheidend vermogen te vergroten, en om kleurinvariantie te kunnen afleiden. Spatieele context wordt inherent beschouwd door het gebruik van Gaussische metingen, zodat het ook mogelijk wordt objectbeweging op verschillende schalen te analyseren.

Het voorgestelde filter is een online Gabor filter, en verschilt van het traditionele, offline Gabor filter. Het online filter reageert sneller en dooft sneller uit dan offline filters. Verder is het online filter beter in staat om een bepaalde frequentie te meten. Het nut van het online filter wordt aangetoond in verschillende video-opnames van diverse, zowel stilstaande als bewegende, objecten.

Hoofdstuk 5: Color Textons for Texture Recognition Modellen van textuurprimitieven, ook wel *textons* genoemd, zijn erg informatief gebleken voor het beschrijven van ruwe texturen. Om het onderscheidend vermogen van textonmodellen verder te vergroten, stellen we twee schema's voor om kleurinformatie toe te voegen. Eerst doen we dit direct op het niveau van textons, door textons te verkrijgen uit het kleurenbeeld. Maar, het is niet efficient om alle kleur- en vormcombinaties te moeten leren. Daartoe stellen we een tweede aanpak voor, die de texton-modellen a posteriori weegt met de kleurinformatie, zodat kleurinformatie niet de vormbeschrijving beinvloed maar slechts nader specificeert. Deze aanpak blijkt meer onderscheidend dan zowel het grijsmodel als het leren van kleur direct op het niveau van de textons. Vooral als de leerset klein is, zoals een set van slechts twee beelden, wordt een significante verbetering behaald van 10% tot een herkenningspercentage van 86% voor de CURET dataset van ruwe texturen.

Hoofdstuk 6: Material-specific Adaptation of Color Invariant Features. Voor het modelleren van materialen is een veelbeoefende aanpak om het beeld uit te drukken in typische materiaal-primitieven. Deze primitieven worden veelal uit filterbank-responses geleerd, vanwege het generieke karakter en de simpliciteit van een filterbank. De MR8-filterbank van Varma en Zisserman blijkt zeer discriminerend in een recente evaluatie op de CURET dataset [120]. In dit hoofdstuk construeren we verscheidene filterbanken met kleurinvariante eigenschappen. De voorgestelde filterbanken worden geevalueerd op de ALOT dataset van 250 alledaagse materialen.

De bijdrage van dit hoofdstuk is een framework wordt voorgesteld dat in staat is om zonder tussenkomst van een expert te leren welke selectie of combinatie van de voorgestelde filterbanken het beste een specifiek materiaal van andere onderscheid. Voor de grote dataset van materialen met verschillende fysische eigenschappen wordt gedemonstreerd dat met materiaal-specifieke filterbank-modellen betere resultaten behaald worden dan met een vooraf vastgelegde keuze.

Hoofdstuk 7: The Distribution Family of Similarity Distances. Om beelden te vergelijken is een maat voor similariteit nodig. Veelal wordt een beeldkenmerk gerepresenteerd als meerwaardige vector, en om twee of meerdere vectoren te vergelijken wordt een L_p-norm – een klasse van metrieken – als similariteitsmaat gehanteerd. In dit hoofstuk wordt afgeleid dat L_p-normen tussen (beeld)kenmerken die gerepresenteerd worden als meerwaardige vectoren een vaste statistische verdeling met slechts enkele parameters volgen. Aangetoond wordt dat als de waardes van de te vergelijken vectoren correleren en niet-identiek verdeeld zijn, de similariteiten tussen een en andere vectoren een Weibull-verdeling volgen. Ondanks dat deze eigenschappen gelden voor vele typen kenmerken – ook voor andere media dan beelden – illustreren we dat voor Weibull-verdeling optreedt voor similariteiten tussen veelgebruikte beeldkenmerken.

Dankwoord

Ik ben ontzettend blij dat dit onderzoek tot een mooi einde is gekomen. Het is spannend, intrigerend, zwaar, mooi, teleurstellend, maar vooral vormend geweest. Deze kenmerken opsommend, komt met name het diffuse karakter van onderzoek naar boven. Dat is wat me enorm aangesproken heeft toen ik ermee begon, in oktober 2002. Ik herinner me nog goed dat ik met een biertje op een grasveldje zat, met wat vrienden; samen waren we naar een band aan het luisteren. Ik raakte daar in gesprek met een promovendus, die me vertelde over zijn onderzoek naar hoe te meten aan video om uitspraken te kunnen doen over de semantiek ervan. Een prachtig onderwerp; inderdaad, later zijn we collega's geworden. Cees, dank voor je enthousiasme.

Eenmaal begonnen aan mijn eigen onderzoek, ben ik enorm geéfnspireerd door Jan-Mark, mijn dagelijks begeleider. Ik moest een behoorlijke sprong maken om van mijn afstudeerproject in Twente op het niveau van mijn nieuwe groep, ISIS, te komen. Jan-Mark, dank voor de tijd en moeite die je daarin hebt gestoken. Meer ben ik je dankbaar voor het verbreden van mijn blik – vaak heb je me interessante artikelen aangereikt om te laten zien wat er nog meer speelt in de wereld van de wetenschap. Natuurlijk heb ik ook genoten van je praktische insteek. Waren er foto's nodig, dan zaten we even later op de fiets, met een daklozenkrant in de hand op zoek naar verschillende bomen in Amsterdam om de verschijningsvormen van texturen vast te leggen. Schreef ik een paper over het meten van beweging, dan nam ik de camera van thuis mee, en zat ik er later mee in Artis om een anemoon te bewonderen. Jan-Mark, dank!

Waar Jan-Mark mijn blik verbreedde, heeft mijn andere begeleider, Arnold, mijn blik weer geconcentreerd om duidelijk te maken dat wetenschap ook hard werken is om bepaalde doelen te realiseren. Niet alleen in het schrijven van papers, of het bezoeken van conferenties, maar ook op persoonlijk vlak. Zo herinner ik me het zeer ongewone "sinaasappelpers" gesprek, waarvoor ik je erg dankbaar ben, Arnold – jouw begeleiding in het laatste jaar heeft me gevormd als collega-onderzoeker.

De onderzoeksgroep werd al even genoemd, ISIS, waar ik me erg thuis heb gevoeld. Goede mensen, leuke collega's, en ook al was ik naar het einde toe regelmatig afwezig bij de koffiepauzes, ik ben jullie erg dankbaar voor de prettige sfeer. Met Giang, Aristeidis, en mijn kamergenoot Frank heb ik met veel plezier de mijn eerste ASCI conferentie bezocht. Bij de andere Frank kon ik altijd terecht voor een vraag, opvallende was dat het vaak eindigde in gesprekken over muziek, kunst, of politiek. Jan, Cees, Dennis, Sumit, Michiel, Sander, en Arjan, dank voor de vele discussies over het werk maar vooral voor de gezellige pauzes. Enkele van mijn huidige collega's bij TNO heb ik leren kennen tijdens deze periode: Jeroen, Erik, en Andy, op een goede toekomst! Theo, Sennay, Minh, Ioannis, Arjan, and Thang, thanks for the nice soccer games – and remember, not bad, second over all! Virginie, diep dankbaar ben ik je dat je altijd aanspreekbaar bent geweest, en vooral de rust die je aanbracht op hectische momenten. Joost, je hebt me nog ingewijd in het meten in beelden, en van Cor heb ik regelmatig een inleiding gehad in de patroonherkenning, want met beide interessante onderwerpen was ik niet bekend toen ik bij ISIS kwam. Joost en Cor, bedankt!

Mijn goede vrienden Pascal en Joost, bedankt voor de enerverende avonden uit! Jullie volhardende initiatief om in Eindhoven, Amsterdam of Volendam een biertje te gaan drinken waardeer ik enorm. Ellen, Afke, Diana, Amy, Esther, Tineke, Fieke, Marijke, en Lisa, dank voor de ontspanning tijdens de gezellige verjaardagen, kermissen en housewarmings!

Mijn familie, jullie zijn heel belangrijk geweest. Het tweewekelijkse eten op donderdagavond met mijn gezin, en op vrijdag met mijn schoonfamilie, heb ik met name in de laatste maanden ervaren als ultiem rustpunt. Lekker een glas cointreau of Shiraz erbij, fantastisch! Mijn broer en zus, Nico en Marijke, jullie dank ik voor de getoonde interesse, en mijn zwager Michel voor het spelen (?) van de advocaat van de duivel. Mijn ouders, Sjaak en Carla – jullie zijn fantastisch, dank voor jullie warmte! Van de prachtige olijfboom die ik van jullie heb gekregen, zal ik nog lang genieten, en herinnert me aan een mooie periode in mijn leven.

Ina, mijn lieve vriendin, samen delen we de liefde voor zoveel: reizen, sport, gezelligheid, en het ondernemen van allerlei activiteiten. Voor het promoveren heb ik alleen gekozen, maar ik heb in alles het gevoel gehad dat ik ook dit met je kon en nog steeds kan delen. Voor je interesse, onvermoeibaarheid, liefde, en enthousiasme, ben ik je heel dankbaar. Maar niet alleen nu, dat zal ik altijd blijven, zoals ik hoop dat we samen nog van alles zullen beleven,

Gertjan
Edam