Selecting and robustifying local image descriptors
Everts, I.

Citation for published version (APA):
Everts, I. (2014). Selecting and robustifying local image descriptors

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Selecting and Robustifying Local Image Descriptors

Ivo Everts
Selecting and Robustifying Local Image Descriptors

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 woensdag 17 september 2014, te 12:00 uur

door

Ivo Everts

geboren te Haarlem
Promotiecommissie:

Promotores: prof. dr. T. Gevers
prof. dr. ir. A. W. M. Smeulders

Overige leden: prof. dr. ir. F. C. A. Groen
prof. dr. ir. M. J. T. Reinders
dr. J. C. van Gemert
dr. I. Laptev
dr. N. Sebe

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Universiteit van Amsterdam

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.

This book was typeset by the author using \LaTeX\textsuperscript{2e}. The cover was designed by Maarten van den Berg. The scaling and spinning images on the front and back cover are the result of applying Eq. (3.3) on a photograph of the vinyl release of The Beatles / 1967-1970, side four.

Copyright © 2014 by Ivo Everts. All rights reserved.
This work appears in the following articles:

I. Everts, J. C. van Gemert and T. Gevers
*Per-patch descriptor selection using surface and scene attributes*
In European Conference on Computer Vision, 2012 [21].

I. Everts, J. C. van Gemert and T. Gevers
*Evaluation of Color STIPs for Human Action Recognition*
In Computer Vision and Pattern Recognition, 2013 [22].

I. Everts, J. C. van Gemert and T. Gevers
*Evaluation of Color Spatio-Temporal Interest Points for Human Action Recognition*
In IEEE Transactions on Image Processing, 2014 [23].

I. Everts, J. C. van Gemert, T. Mensink and T. Gevers
*Robustifying Descriptor Instability using Fisher Vectors*
In IEEE Transactions on Image Processing, 2014 [24].

*Principles for the Development of a Complete Mind:*
*Study the Science of Art.*
*Study the Art of Science.*
*Develop your senses - especially learn how to see.*
*Realize that everything connects to everything else.*

Leonardo Da Vinci
# Contents

1. Introduction ................................................. 1  
   1.1 Objectives ........................................... 5  

2. Robustifying Descriptor Instability using Fisher Vectors ........... 9  
   2.1 Introduction ......................................... 9  
   2.2 Relating Signal Strength to Descriptor Instability .......... 11  
   2.3 Fisher Vectors from Unstable Descriptors .............. 14  
      2.3.1 Fisher Vector Gradients ......................... 18  
   2.4 Experiments .......................................... 19  
      2.4.1 Experimental Setup ............................ 19  
      2.4.2 ALOI Object Matching ......................... 20  
      2.4.3 INRIA Holidays Image Retrieval .............. 22  
      2.4.4 Pascal VOC 2007 Object Category Recognition .... 23  
   2.5 Discussion ........................................... 26  

3. Per-Patch Metric Learning for Robust Image Matching ............ 29  
   3.1 Introduction ......................................... 29  
   3.2 Related Work ......................................... 31  
   3.3 A Single-patch Metric ................................ 33  
      3.3.1 Metric Learning ................................ 34  
      3.3.2 Direct Metric Estimation ...................... 35  
   3.4 Experiments .......................................... 37  
      3.4.1 Geometric Robustness ......................... 37  
      3.4.2 Photometric Robustness ....................... 41  
      3.4.3 Complexity Analysis ........................... 42  
      3.4.4 Extensive Analysis of the Proposed Method ...... 43  
   3.5 Discussion ........................................... 44
Introduction

“Learn how to see”

Try to imagine to learn how to see, as demanded by this quote of Leonardo da Vinci. Are you seeing any pictures?

As infants, most people learn how to see so well that it comes fully natural to us for the rest of our lives. From motion detection for predator evasion, to semantic visual interpretation for recognition: the ability of sight is of vital importance. Look, it is so firmly grounded in human language that it is common to appeal to the reader’s or listener’s ability to see.

The ease and flexibility with which humans see is accomplished by the highly complex visual processing system. In rough terms, this constitutes the transduction of photons that hit the retina in the back of the eyeball into signals sent through the optic nerve to the brain. Information at any layer of abstraction is extracted by an extremely large, complex and interrelated network of neurons. The system responsible for the processing of visual input (the visual cortex) is the largest functional system in the brain, which in turn is the largest energy consumer in the human body. Without such a vast and complex processing capacity, human vision is not possible.

When one tries to imagine to learn how to see, or to convey the ability of sight to someone who has never seen, one should start by accessing photoreceptors for determining the intensity and the wavelength (i.e. color) of the light entering the eye. Subsequently these signals should be filtered to detect visual phenomena such as edges and textures. Then, segments must be grouped together when
being similar in terms of location, intensity, color, texture and shape, before finally being interpreted on a semantic level. However, the mental experiment in which it is attempted to learn how to see reveals that it appears impossible to imagine to actively ‘use’ our cells, nerves and neurons that enable sight. Also, it appears difficult to communicate about the process without referring to any kind of visual phenomenon. In the development of computer vision this has been taken to the extreme, as machines have to be instructed exactly what to do. The question is how to convey the ability of sight step by step.

Because of the enormous amount of image data, online and in digital archives, security databases, in medical domains, numerous industrial settings and for human-computer interaction applications, the usefulness of sight granted to machines is evident. Structuring these large amounts of complex image data is possible only with the aid of effective and efficient computer vision, which is reflected at the time of writing by the recent installment of computer vision departments at Google and Facebook Inc. Computer vision is used to perform face detection and recognition to reveal links in social networks, to perform multimedia search based on the media content, and for the automatic categorization of photo collections made up of ordinary images like the ones in figure 1.1.

![Figure 1.1: Examples of ordinary images from a typical photo collection. All images may be categorized as containing ‘cars’, whereas image (A) and (B) are taken ‘outdoors’, (B) and (C) are both images of a ‘Saab’ and image (A) also contains a ‘van’ and ‘buildings’.](image)

The images in figure 1.1 reveal a few things. All images contain ‘cars’, which is easily observed despite many large variations in car manufacturer, car model, position, scale, viewpoint, illumination, color and camera hardware. Image (C) was photographed ‘indoors’, which renders illumination effects different to the ‘outdoor’ settings of (A) and (B). Image (A) contains several ‘buildings’, ‘cars’ and a ‘van’, which is a different instance in the class of ‘vehicles’ just like a ‘car’.
Also, it is worth noting that the cars in images (B) and (C) are both a ‘Saab’. The examples illustrate many aspects of the problem faced by the field of computer vision: there may be large differences between instances of the same visual category, whereas instances across categories might appear very similar.

In order to be able to explain why objects and scenes appear in images the way they do it is important to understand the process of image formation. Similar to the first steps of biological vision, images are formed by capturing the projection of the continuous 3D-world on a discrete 2D-lattice. The process involves many complex factors, which produce incidental variations in the appearance of objects and scenes. Geometric variations result from camera viewpoint, the object location, the object pose and the scene [36]: the images in figure 1.1 exhibit cars in the side-, rear- and frontal view, in close-up and far away, and from different makings. Illumination conditions will cause photometric variations [31], illustrated in figure 1.1 by the indoor and outdoor settings, highlights and shadows. The appearance of an object and scene is best understood as a specific geometric and photometric instantiation.

The information in a digital image is a list of pixels consisting of a location \((x, y)\) and color \((\text{Red, Green, Blue})\). Such a representation is affected by the variations in image formation. For computer vision, it is advised to formulate the problem in terms of invariance against factors of image formation and class variation that are disturbing the task at hand. For example, one way of correctly recognizing a ‘Saab’ in images 1.1 (B) and (C) is the detection of the Saab logo, despite viewpoint variations and the presence of strong highlights. Simultaneously, the representation should adequately capture the visual properties that are informative for the content: the highlights should somehow be ignored while preserving the structure and color of the Saab logo. The ability of doing so results in image representations exhibiting discriminative power. Computer vision has to consider trade-offs between invariance and discriminative power because informative content might be ignored due to invariant image measurement [93]. This trade-off is referred to as robustness [91].

The current mainstream approach to robust computer vision considers large amounts of measurements of local image features, known as local image descriptors, which are either learned [45] or modeled [55, 31]. Deep learning is more and more popular as a result of recent top performances in recognition benchmarks. A drawback is the large amount of parameters in the neural network, the robust estimation of which requires enormous amounts of data. Also, due to the complexity of the network, it is difficult to interpret the representations.
at the various layers and neurons.

This thesis is concerned with physical models of local image features and representations thereof [76]. Local image measurements are relatively insensitive to global geometric variations as well as to a range of photometric variations, and are used for tasks as diverse as object and scene recognition [76, 89], object tracking [81], motion estimation [53], panoramic stitching [8], depth estimation [62] and 3D reconstruction [36]. Thus, representations based on local image descriptors are suitable for modeling image content. The robustness of a computer vision system then depends on the robustness of the local descriptors themselves.

Currently, there exist two main paradigms for achieving descriptor robustness. First, invariance can be imposed by design [55, 31]. Invariant properties are derived analytically from physical models of image features conforming to a model of image formation, or the algorithm for feature extraction and aggregation may be engineered such that certain variations are ignored. Second, the building blocks of the descriptor extraction algorithm (and parameter settings thereof) may be globally optimized in order to identify the trade-off between invariance and discriminative power [7]. Both approaches result in a single descriptor exhibiting a fixed level of robustness. However, due to the enormous variety of local image content resulting from variations in image formation, there exists no single local image descriptor that is optimal under all circumstances [93, 21]. For example, textured image regions may be represented best by measurements of structure, where color could be most informative for homogeneous regions. Such a diversification of descriptors also applies to the discrimination of materials. Shiny materials may be best represented with a highlight-invariant, whereas the best representation of matte materials is in general a shading-invariant.

This thesis aims to identify the issues related to the diversification of local descriptors arising from the variations in image formation and object properties. In general, this requires an adaptation of the extraction and modeling of local descriptors based on properties of the underlying image signal itself. By doing so, a descriptor is selected dynamically. The dynamic selection of local image descriptors results in novel image representations with different robustness characteristics.
1.1 Objectives

The objectives of this thesis are illustrated by analyzing the stylized pictures in figure 1.2. The images 1.2(A) through 1.2(D) show a sequence of vinyl recorded under varying illumination conditions. Here, in addition to the ambient light, an illumination source is positioned at various locations relative to the vinyl, which renders photometric variations in the image data. The images 1.2(E) through 1.2(H) in the bottom row show a sequence of the vinyl record in rotation, recorded at 0°, 90°, 180° and 270°. These pictures exhibit one type of geometric variation.

Figure 1.2: Illustration of the chapters in this thesis. A vinyl record is photographed under varying illumination conditions (A-D) and in different poses (E-H), producing photometric resp. geometric variations of the object. Photometric variations are most apparent on the homogeneous, shiny vinyl, whereas geometric variations are best observed on structured and distinctively colored object regions. The way in which the object appearance responds to a variation in the recording conditions is dependent on properties of the object itself. This may impose conflicting demands on the way in which the image is best represented for a computer vision task.

Images 1.2(A) and 1.2(B) show that a photometric transformation has most effect on unstructured image regions such as those exhibiting the black vinyl, whereas structured regions containing e.g. the apple stem appear to not change at all. This illustrates that the degree to which a photometric change in the recording conditions affects the object appearance is in fact dependent on structure properties of the object itself. The relation is denoted as stability, which
propagates to any attempt to represent the image data for accomplishing some visual recognition task. As a consequence, the descriptors that are used for image representation may be very unstable on regions exhibiting little or no structure and could thus negatively impact recognition performance. Therefore, it is aimed in this thesis to explicitly account for the stability of an image descriptor such that the associated performance loss is prevented.

**Objective 1.** *Formulate and exploit measurements of photometric descriptor instability in standard image representations to improve robustness of visual recognition systems (chapter 2).*

In contrast with the observed effect of varying illumination conditions in images 1.2(A)-(B), it is shown in 1.2(E)-(F) that the appearance of the same image regions behave differently under a geometric transformation of the imaged content such as rotation. Here, the visual effect of the transformation is observable on structured regions, whereas no variation is observed on the black vinyl due its lack of structure (at this scale of observation). Thus, the observed effect of the transformation on the image depends on the imaged content itself, as is also the case for photometric transformations, however reversed in relation to image structure. Knowledge of the effect of transformations is valuable as it identifies what regions remain stable. This, in turn, permits an image matching scheme to be adaptable on-the-fly such that more importance is granted to stable image regions and as a consequence matching performance might be improved.

**Objective 2.** *Formulate and exploit the effect of local geometric image transformations for improved robustness of image matching (chapter 3).*

In addition to properties related to structure, other aspects of object appearance are material and albedo. The images 1.2(A)-(D) show that vinyl is a shiny material. A colorful object region is indicated in image 1.2(G). Structure, material and color together impose complex and potentially conflicting demands on the way images are being represented. The demands move beyond globally optimal, task-specific image representations. Instead, they focus on descriptor diversification at the level of image regions. The aim is to select the most representative and discriminative descriptor for a given image region, based on structure, material and color properties of that image region.

**Objective 3.** *Select from a diversified set of image descriptors the best descriptor based on properties of the underlying image content for robust matching (chapter 4).*
1.1 Objectives

Finally, it is observed from images 1.2(C)-(D) and 1.2(G)-(H) that there are multiple sources of apparent motion in an image sequence. On object regions without structure such as the black vinyl in figure 1.2(C)-(D), the only observed variations are not due to object motion but to a change in position of an illumination source. Structured and colorful regions such as the apple stem in figures 1.2(G)-(H) exhibit variations that comply with the actual object motion. Thus, for motion detection and description it is important to discriminate between deceptive motion of illumination effects and variations due to actual (object and camera) motion. Moreover, better motion estimates may result from considering color in addition to structure in the representation.

Objective 4. Reformulate local video descriptors in order to enable invariant color image measurement for improved robustness of visual recognition in the video domain (chapter 5).

In summary, the central question of this thesis is how to select (parameters of) local image descriptors, such that the robustness of visual recognition and matching is improved. The objectives 1-3 approach the problem by learning how to select or adapt a descriptor per image region based on visual properties of that region. Objective 4 aims to select the single best descriptor based on batch processing of video datasets.
Robustifying Descriptor Instability using Fisher Vectors *

2.1 Introduction

Computer vision tasks such as (object) recognition, image matching and retrieval typically depend on local image descriptors. Many robust image descriptors have been designed [89] or optimized [7] to deal with small changes in image geometry and photometry. For such robust descriptors, ideally, small geometric and photometric changes in the image recording conditions correspond to a negligible change in the image descriptor.

We focus on the popular descriptor family of gradient orientation histograms such as SIFT [55], HOG [16], SURF [6], or color SIFTS [89]. With a strong gradient signal, the SIFT descriptor is indeed robust to small perturbations in the image. However, if the gradient signal is weak, small changes in the image signal could result in huge variations in the local descriptor after \( \ell_2 \)-normalization. Such change in the descriptor after a small variation is what we denote as descriptor (in)stability. We show that there exists a parametric relation between descriptor stability and signal strength, which also applies to robust image descriptors and cannot be resolved by noise filtering.

To illustrate the problem, in Figure 2.1 we show for a few image patches the influence of adding small amounts of zero mean additive Gaussian noise to the image. Stable patches containing strong gradient signals are robust to the

*Published in the transactions on Image Processing [24].
Figure 2.1: Different image patches exhibit different behavior in SIFT descriptor space when subject to small image perturbations. Image perturbations for low-signal patches cause large jumps in descriptor space. We aim to model the descriptor instability (colored circles) based on the signal strength.

additive noise, and remain close to the original patch in descriptor space. In contrast, image patches with weak gradient signals make large shifts in the descriptor space after their distortion. This could severely influence the encoding scheme used to transform the local descriptors into an image representation.

There are two common approaches that address descriptor instabilities, either explicitly or implicitly. First, unstable patches may be identified based on a threshold on the gradient strength and subsequently mapped to a NULL descriptor [59, 94]. However, this may potentially lead to severe performance loss, whereas the optimal threshold is highly dataset dependent otherwise. Second, the instabilities may be taken for granted in the image representation and instead assumed to be modeled by a classifier as many variations are observed in a training set [104].

In contrast to fully relying on a classifier, thresholding on the gradient signal or changing the descriptor itself [33], we model descriptor instability in the Fisher Vector (FV) framework. The FV encodes local descriptors into a global image vector representation [66, 68] which can be used for image classification, retrieval or matching. We choose the FV framework for two reasons, (i) it has proven to be one of the most powerful encoding schemes for image classification and retrieval [12, 39], and (ii) it offers a principled way to model descriptor instabilities in the underlying graphical model. The FV is based on the Fisher Kernel [37], and it consists of characterizing a set of local image descriptors by its deviation measured by the gradient with respect to the log-likelihood
2.2 Relating Signal Strength to Descriptor Instability

of a generative Gaussian mixture model (GMM). The GMM corresponds to a probabilistic version of the visual dictionary as used in the bag-of-visual-words approach [83, 14, 48]. We will show in this chapter that descriptor instability modeling with FVs substantially improves recognition performance in comparison to signal thresholding for matching and classification tasks.

The rest of the chapter is organized as follows. Next, we relate the descriptor strength to the instability of the descriptors. In Section 2.3, we introduce our modification of the FV framework to incorporate descriptor instability, which we use in Section 2.4 for image matching, retrieval and classification. Finally, we summarize our contributions in Section 2.5.

2.2 Relating Signal Strength to Descriptor Instability

The stability of a descriptor is related to the signal strength of an image patch, as illustrated in figure Figure 2.1. Here, we aim to quantify the relation such that descriptor instability can be measured and interpreted as observational variance.

The signal strength of an image patch $I$ is measured by the $\ell_2$-norm of the gradient magnitudes $||\nabla I||_2$, since we describe image patches by the gradient-based SIFT descriptor. For gauging stability in descriptor space we use a near-copy of the same image patch, which is created by either (1) a stochastic change by adding a small amount of Gaussian noise, or by (2) a photometric change by a re-recording of the image patch under a slightly different light color. For each near-copy we extract its SIFT descriptor and compute the distance to the original descriptor. Ideally, these self-distances are close to zero since the underlying image content remains unaltered.

For the stochastic variant (1) to obtain a near-copy we use a small amount of i.i.d. zero-mean additive Gaussian noise, which is the standard model of amplifier noise [33, 63]. For these near-copies we also evaluate the impact of applying a noise reduction technique prior to descriptor extraction by an edge preserving anisotropic diffusion filter [65]. For the photometric variant (2), the near-copies are obtained by two recordings in the ALOI set [28] (see section 2.4) of the same object under a nearly imperceptible different illumination color temperature ($2975^\circ$K vs $3075^\circ$K) where cameras are white balanced at $3075^\circ$K. For the difference in illumination color, we also evaluated RGB-SIFT, which is invariant to changes in the illumination color [89]. Example patches are depicted
in Figure 2.2.

**Figure 2.2:** Example patches from ALOI. The original patch and its near-copies due to: changed illumination color, additive Gaussian noise with $\sigma_{\text{noise}}^2 = 10^{-3}$, and the noisy patch after applying the noise-reduction filter. See figure 2.7 for examples of full ALOI images.

**Figure 2.3:** Self-distance scattered against signal strength (x-axis) for color temperature change and additive i.i.d. Gaussian noise. The solid line is a least-squared fit of an exponential function. SIFT is computed from densely sampled 24x24 image patches from ALOI.

In Figure 2.3 we show the relation between signal strength and descriptor instability for 10K randomly sampled image patches from ALOI. As illustrated, there is a strong relationship between signal strength and image descriptor instability. Strong signal patches are stable, i.e., close to the near-copies in descriptor space. Low-signal patches, however, are unstable as illustrated by large self-distances. Moreover, any attempt to remove the differences between near-copies by either noise reduction (Filtered SIFT) or photometric invariance (RGB-SIFT) does not diminish the instability.

The experiment has also been repeated with filtered original patches to verify that the filtering routine is not influencing the observed relation. This results in essentially the same graphs (results not shown).
2.2 Relating Signal Strength to Descriptor Instability

We propose to use the signal strength to model descriptor instability and incorporate the descriptor instability in the Fisher Kernel model. We use signal strength to model the descriptor instability \( C(\cdot) \) as the average descriptor distance to itself, i.e., the variance of the descriptor distance. The relation between signal strength \( x = ||\nabla I||_2 \) and instability \( C(x) \) can be described with an exponential function,

\[
C(x) = \alpha e^{\beta x}.
\]  

(2.1)

We use least-squares to fit \( \alpha \) and \( \beta \), as illustrated by the black line in Figure 2.3.

Note that a larger difference between near-copies will influence the self-distances. A higher noise level or larger color temperature difference will also increase the self-distance of stronger signal patches. To take this into account, we consider several noise levels where the variance of the applied Gaussian noise \( \sigma^2_{\text{noise}} \) is a parameter to be optimized on a hold out set. Figure 2.4 illustrates the effect of different noise levels on the instability curve.

The advantage of using the relation in Eq. (2.1) is that merely computing the image gradient norm allows us to estimate the descriptor instability in terms of its variance as a single scalar value. A scalar variance suffices as there is no reason to assume a priori that the variance is not uniformly distributed over SIFT dimensions.

We consider the inferred variance in addition to the descriptor itself as image measurement. The Fisher Vector representation is next reformulated such that descriptor instability is incorporated as measurement variance. This offers a principled approach to dealing with descriptor instability, as opposed to thresholding on the gradient signal or fully relying on a classifier.
2.3 Fisher Vectors from Unstable Descriptors

The Fisher Vector (FV) approach for image classification [66, 68] models a visual word vocabulary by a Gaussian mixture model (GMM) and characterizes a set of...
of local image descriptors $X$ by their gradient w.r.t. the parameters $\theta$ of the GMM under a log-likelihood model, i.e. $\text{FV} \equiv F_\theta \nabla_\theta \log p(X)$, where $F_\theta$ is the Fisher Information Matrix. We also model the data using an GMM. However, the descriptor instabilities are incorporated while learning the parameters of the GMM. We analyze this by relating the responsibilities of the GMM components to the signal strength of the associated patches. As Fisher Vectors we extract gradients w.r.t. a different objective function to encode the noisy descriptors into the image representation.

Following [66, 68] we assume that our GMM has diagonal covariance and is defined as:

$$p(x; \theta) = \sum_k p(x|k)p(k) = \sum_k N(x; \mu_k, \sigma^2_k) w_k, \quad (2.2)$$

where $x$ is an arbitrary point in the $d$ dimensional descriptor space $\mathbb{R}^d$, $k$ denotes a mixture component, $N(x; \mu_k, \sigma^2_k)$ is the multi-variate Gaussian distribution of component $k$ with mean $\mu_k$ and variances $\sigma^2_k$, and $w_k$ is the mixing weight (with constraints $\forall k : w_k \geq 0$ and $\sum_k w_k = 1$). When a signal threshold $t_{\text{signal}}$ is used, $x$ is mapped to a NULL descriptor (i.e. containing only zero elements) if the gradient strength of the underlying image patch falls below the threshold:

$$x = \begin{cases} g(I) & \text{if } t_{\text{signal}} < ||\nabla I||_2 \\ 0 & \text{otherwise} \end{cases}. \quad (2.3)$$

Here, $I$ is the image patch and $g(\cdot)$ denotes the descriptor extraction algorithm (i.e. SIFT). Note that the proposed method does not rely on such a threshold. The set of parameters to be estimated is $\theta = \{w_k, \mu_k, \sigma_k\}_{k=1}^K$, for a $K$ component mixture.

The parameters of the GMM $\theta$ in the FV framework are usually learned on a set $\{x_1, \ldots, x_n\}$ of local descriptors using the EM algorithm to maximize the log-likelihood $\sum_j \log p(x_j)$. To gain insight in the influence of the descriptor instability on the FV encoding, we use again the patches from the ALOI dataset used in Section 2.2, and train a GMM with $k = 16$ components. In Figure 2.5 (top-row), we show the following. We compute the most likely component $k^*$ for the original patch $j$: $k^* = \arg\max_k q_{jk}$, where $q_{jk} \propto p(x_j|k)p(k)$ is the posterior of component $k$ for patch $j$. We show the difference $d_{jj'} = q_{jk^*} - q_{jk}$ between the posteriors of the original patch $j$ and its near-copy $j'$, for component $k^*$. We relate this difference to the signal strength, similar as in Figure 2.3. We observe that for all descriptors there is a clear relation between the signal strength and the difference in posterior. Especially for patches with a low signal strength there are substantial changes in the posterior values.
We now incorporate the descriptor instabilities derived from signal strength for learning the parameters $\theta$ of the GMM, to better model the uncertainties of the descriptors. We follow the EM approach of [95] to learn a GMM from noisy observations, which we coin N-EM for clarity in the rest of this chapter. Their method is summarized as follows. Using all descriptors, together with their (diagonal) covariance matrices $\{C_1, \ldots, C_n\}$, we can define a variable kernel density estimator as:

$$f(x) = \frac{1}{n} \sum_j f(x|j) = \frac{1}{n} \sum_j N(x; x_j, C_j).$$  \hspace{1cm} (2.4)

This kernel density estimator represents a non-parametric distribution over the descriptor space.

The learning problem is now expressed as the minimization of the Kullback-Leibler divergence between the kernel estimator and the unknown mixture, i.e. $\theta^* = \text{argmin}_\theta D_{\text{KL}}[f(x)||p(x; \theta)]$, which yields the following function to maximize:

$$L = \sum_j \int_x f(x|j) \log p(x; \theta) \, dx.$$  \hspace{1cm} (2.5)

Instead of directly maximizing $L$, an EM approach is considered to maximize a lower bound of $L$, which reads:

$$F = \sum_j \sum_k q_{jk} \left[ \int_x f(x|j) \log p(x|k) \, dx + \log p(k) - \log q_{jk} \right],$$  \hspace{1cm} (2.6)

where $q_{jk}$ is the posterior $p(k|x_j)$ between a descriptor $x_j$ and component $k$, also known as the responsibility. The integral in Eq. (2.6) is analytically solved as:

$$\int_x f(x|j) \log p(x|k) \, dx = \log N(x_j; \mu_k, \sigma_k^2) - \frac{1}{2} \langle \sigma_k^{-2}, C_j \rangle,$$  \hspace{1cm} (2.7)

where $\langle \cdot, \cdot \rangle$ denotes the dot-product between two vectors.

The N-EM update equations.

Iteratively maximizing $F$ results in update equations which are very similar to the EM algorithm for noise-free data. First, in the expectation step the responsibilities are computed as follows:

$$q_{jk} = \frac{N(x_j; \mu_k, \sigma_k^2) w_k \exp \left( -\frac{1}{2} \langle \sigma_k^{-2}, C_j \rangle \right)}{\sum_{k'} N(x_j; \mu_{k'}, \sigma_{k'}^2) w_{k'} \exp \left( -\frac{1}{2} \langle \sigma_{k'}^{-2}, C_j \rangle \right)}.$$  \hspace{1cm} (2.8)
Second, in the maximization step the mixture parameters are updated as follows:

\[ w_k = \frac{1}{n} \sum_j q_{jk} \quad (2.9) \]
\[ \mu_k = \frac{1}{nw_k} \sum_j q_{jk} x_j \quad (2.10) \]
\[ \sigma^2_k = \frac{1}{nw_k} \sum_j q_{jk} \left( (x_j - \mu_k)^2 + C_j \right) \quad (2.11) \]

where the exponentiation of a vector should be understood as a term-by-term operation.

Once more we use the patches from the ALOI dataset, to show the difference in the max-posterior in the GMM trained using N-EM in Figure 2.5 (middle-row). In this case \( q_{jk} \) is defined as in Eq. (2.8). The plot illustrates that indeed the posterior values of patch \( j \) and its near-copy \( j' \) are slightly more stable (i.e. similar to each other). This is also shown when plotting the difference between the distributions when using EM and N-EM in Figure 2.5 (bottom-row), where the blue region indicates higher mass for EM, and red for N-EM. Generally, more EM than N-EM mass is observed for larger differences in the posterior. This means that the probability of the most likely GMM component of a patch is less affected by distortion (illumination or noise) if N-EM is used (and thus the instability of the original descriptor is modeled). N-EM yields less variable assignments and appears more stable under distortions of the image.

![Figure 2.6: Illustration of the EM algorithm using noisy observations on synthetic data. We show the GMM, with \( K = 5 \), learned using standard EM (left), the observation noise (middle), and the GMM after learning with N-EM for noisy observations (right). Both models start from the same initialization.](image-url)
proposed Kullback-Leibler divergence, we also compare, in Figure 2.6, a \( k = 5 \) GMM learned with EM (left) and with N-EM (right), together with synthetic 2 dimensional noisy data (middle). The plot illustrates that the mixture components also try to model the uncertainties of the data, e.g. the variance of the mixture components becomes higher in areas where the data has a high variance c.f. the blue and red components.

### 2.3.1 Fisher Vector Gradients

We next detail on how we modify the Fisher Vector model to incorporate the descriptor instabilities. Instead of using the gradients w.r.t. the log-likelihood, \( FV \equiv \nabla_q \log \pi(X) \), we propose to use the gradient of the observations of an image w.r.t. the lower bound \( F \) in Eq. (2.6). In this model it is assumed that the responsibilities \( q_{jk} \) are given by Eq. (2.8), and that they are fixed, i.e. they do not yield a gradient signal w.r.t. \( \theta \) (see also [75]).

We follow [44] and use \( w_k = \frac{\exp \alpha_k}{\sum_{k'} \exp \alpha_{k'}} \), to obtain the following gradients with respect to \( \{\alpha_m, \mu_m, \sigma_m\} \):

\[
\nabla_{\alpha_m} F = \sum_j \sum_k q_{jk} (\mathbb{I}[k = m] - w_m),
\]

\[
\nabla_{\mu_m} F = \sum_j q_{jm} \frac{(x_j - \mu_m)}{\sigma_m^2},
\]

\[
\nabla_{\sigma_m} F = \sum_j q_{jm} \left( \frac{(x_j - \mu_m)^2}{\sigma_m^3} - \frac{1}{\sigma_m} + \frac{C_j}{\sigma_m^3} \right),
\]

where \( \mathbb{I}[z] \) denotes the Iverson brackets which is 1 if \( z \) is true and zero otherwise, and where the division between vectors or the exponentiation of a vector should be understood as term-by-term operations.

Intuitively, these gradient vectors include the descriptor instability both when computing the posterior \( q_{jk} \) according to Eq. (2.8), and in the gradients with respect to the variances.

**Comparison to Standard FVs.**

To compare our model to the normal FVs used in [68], lets assume that the kernel density function \( f(x|j) = \delta(x, x_j) \) is a Dirac delta function. In that case the integral of Eq. (2.7) is trivially solved as \( \int_x f(x|j) \log p(x|k) dx = \log p(x_j|k) \), and
the lower bound, Eq. (2.6), reads:

$$F_\delta = \sum_j \sum_k q_{jk} \left[ \log p(x_j|k) + \log p(k) - \log q_{jk} \right].$$ (2.15)

When assuming, as above, that the responsibilities $q_{jk}$ are fixed, it is easy to show that the gradients of Eq. (2.15) w.r.t. $\{\alpha_m, \mu_m, \sigma_m\}$ yield the normal FV equations, see e.g. Eq. (9)-(11) in [66]. Note that the responsibilities are now defined as usual as $q_{jk} \propto p(x_j|k)p(k)$.

Conforming to [68], as the final image representation we use $FV = [\nabla_{\mu_k} F \nabla_{\sigma_k} F]_{k=1}^K$, and we apply power-normalization $z \leftarrow \text{sign}(z)|z|^{-1/2}$, followed by $\ell_2$ normalization $z \leftarrow \frac{1}{|z|} z$. When multiple spatial pyramid levels are used, each pyramid cell is individually normalized.

Note that we consider the scalar variance $C_j = C(x)I$ from Eq. (2.1) which results from its definition in terms of descriptor instability. This is however not a restriction of the model.

### 2.4 Experiments

We compare our models to the standard FV framework for three different computer vision tasks: object matching, image retrieval and object category recognition. We use a classifier (SVM) for object category recognition whereas the other tasks are performed by directly matching the FVs (i.e. nearest neighbor classification and distance-based ranking). As we model SIFT instability directly in the FV representation, it is expected that matching-based approaches will especially benefit from the proposed method. Opposed to this, we expect learning-based approaches to already exhibit robustness due to the observed variations in the training data. The basic setup is the same for all tasks.

#### 2.4.1 Experimental Setup

Patch extraction proceeds by sampling 24x24 patches on a dense grid every 4 pixels. Images are processed on 5 scales by iterative down-sampling with a factor of $\sqrt{0.5}$. The signal strength for a patch is measured by the $\ell_2$-norm of its image gradient. SIFT descriptors [55] are extracted using VLFEAT [94] and we apply PCA to reduce the 128 dimensions of SIFT to 64, as commonly done in the FV framework [76].
GMM parameters are estimated from a set of 1M randomly sampled descriptors. The same initialization of the standard EM is used for our N-EM approach, taking into account the descriptor instabilities. For all experiments we estimate the parameters of the GMM and the instability curve on a separate dataset. We study the effect of instability modeling for different values of $k$, the number of GMM components. Instability curves are modeled separately per scale with additive Gaussian noise in $\{10^{-4}, 10^{-3}, 10^{-2}\}$. The gradient signal threshold is considered in $\{0.0025, 0.005, 0.01\}$ where 0.005 is the default setting of the SIFT implementation.

The reported performance measures are averaged over 3 runs using different seeds for training the GMM. We have observed standard deviations ranging from 0.2 to 0.5 percentage points. Based on unpaired t-tests, the best improvements on all datasets were found to be significantly different from the baseline at a standard 5% confidence level.

### 2.4.2 ALOI Object Matching

The Amsterdam Library of Object Images (ALOI) image set [28] contains 1000 objects with systematic variations in viewing angle, illumination angle, and illumination color. We use this set because it allows systematic evaluation of changing a single appearance variable. We focus on lighting arrangement change since this has proven to be difficult for SIFT [89]. We match the canonical image, i.e., with all lamps turned on, with a paired random illumination arrangement in a set of all other objects. Images are cropped such that no background is visible. Performance is measured by the percentage of correct closest object images using the Euclidean distance to the canonical image. See figure Figure 2.7 for some examples of ALOI’s illumination conditions.

![Example images from ALOI under varying illumination conditions.](image)

**Figure 2.7:** Example images from ALOI under varying illumination conditions. The left image is the canonical image, i.e., with all lamps turned on. For every canonical image, we have randomly picked a single match from a different illumination arrangement for creating pairs in the dataset. The dark background is ignored in the experiments.
Figure 2.8: Object matching results. This task is performed based on direct FV matching. The overall results generally improve due to instability modeling, of which several settings for the corresponding noise levels (variance) are plotted in black. Matching performance of several settings for the signal threshold are plotted in gray.

Observing the results in figure Figure 2.8, it appears that instability modeling has a considerable effect. That is, the already near-perfect baseline of 97.3% is improved to 98.9% for $k=64$, whereas performance improves with 2 percentage points for lower values such as $k=16$. Furthermore, SIFT mapping by signal thresholding may lead to incidental improvements (i.e. for $k = 16$, $t_{signal} = 0.005$), but performance degrades in general. Thus, despite the unstable behavior associated to low-signal descriptors, discriminative information may be lost when they are all mapped to the same NULL descriptor.

As object matching in the ALOI dataset is a simple problem, and the appearance changes are controlled and expected to be advantageous to our method, we conduct a more challenging image retrieval experiment in which the image content and recording conditions are more complex.
Figure 2.9: Example images from INRIA Holidays used in the image retrieval experiment. Matching image sets consist of 2-4 images.

INRIA Holidays Image Retrieval

![Graph showing image retrieval performance. The effect of instability modeling is similar to ALOI results (black plot). Special treatment of unstable descriptors by thresholding on the gradient signal may lead to drastic performance loss (gray plot).]

2.4.3 INRIA Holidays Image Retrieval

For the image retrieval task, we use the INRIA Holidays dataset [38], which consists of approximately 1500 images, see Figure 2.9. For each of the 500 queries,
the remaining images are ranked and average precision (AP) is computed. The final performance is measured as the mean AP over all queries (MAP).

The results for various $k$ values in Figure 2.10 show results similar to ALOI as performance increases along with $k$. Our baseline result of 72.5 MAP for $k = 256$ compares favorably to the 0.70 of the Fisher Vector approach in [67]. Instability modeling leads to substantial performance gains, where the improvement increases as $k$ decreases. The reason for this lies in the fact that compensating for the incidental location of descriptors with respect to the GMM clusters has most effect when the GMM is sparse: the responsibilities become even more stable as observations are ‘spread’ through the descriptor space by considering their associated instabilities as measurements of variance (which is also illustrated in Figure 2.5).

It is interesting to observe the dramatic performance degradation on the Holidays dataset when a threshold on the gradient signal is used for dealing with unstable descriptors. Also here, we conclude that discriminative information is ignored by mapping unstable SIFT descriptors to a NULL descriptor. The effect is more pronounced here than for the ALOI object matching task because the retrieval task is harder and thus suffers more from a cut down of discriminative information.

The matching and retrieval experiments showed that modeling the instability of SIFT descriptors effectively maintains discriminative power of low-signal patches in the FV representation. Another approach to (implicitly) dealing with descriptor instability is to model the variations in unstable image content by training a classifier on many examples.

### 2.4.4 Pascal VOC 2007 Object Category Recognition

Object recognition is evaluated on the Pascal VOC 2007 visual recognition benchmark [20]. This is a well known set for image categorization and consists of 20 categories such as Aeroplane, Bottle, Cat, Dog, etc. (see Figure 2.11), with train, validation and test sets of 2501, 2510, and 5011 images respectively. We train a linear SVM on the two variants of the FV and evaluate the effect of a 3x1 spatial pyramid [48]. Performance is evaluated by mean average precision (MAP), the area under the precision-recall curve.

The results in Figure 2.12(a) show the effect of varying the noise levels and signal thresholds on the Pascal VOC validation set. As with the other datasets, a high noise variance of $\sigma_{\text{noise}}^2 = 0.01$ for instability modeling decreases performance in comparison to lower values. High descriptor uncertainties may ren-
Figure 2.11: Example images from Pascal VOC 2007. This dataset exhibits much larger intra-class variation than the ALOI and Holidays datasets, and consists of train and test sets.

Pascal VOC 2007 Object Recognition

<table>
<thead>
<tr>
<th>Noise Variance</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_{\text{noise}}$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.47</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.49</td>
</tr>
<tr>
<td>0.001</td>
<td>0.51</td>
</tr>
<tr>
<td>0.01</td>
<td>0.53</td>
</tr>
</tbody>
</table>

With Spatial Pyramid

<table>
<thead>
<tr>
<th>Noise Variance</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_{\text{noise}}$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.57</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.55</td>
</tr>
<tr>
<td>0.001</td>
<td>0.53</td>
</tr>
<tr>
<td>0.01</td>
<td>0.51</td>
</tr>
</tbody>
</table>

(a) (b)

Figure 2.12: VOC 2007 validation scores in MAP for varying noise levels (black) and signal thresholds (gray). Without (Figure 2.12(a)) and with (Figure 2.12(b)) spatial pyramid.

Under the responsibilities ambiguous. Opposed to this, the representations benefit substantially from instability modeling, especially those that are based on a sparser GMM ($k = 64$). However, the performance differences between FVs with and without instability modeling appear somewhat less pronounced than for the ALOI and Holidays datasets. This is because the SVM effectively ex-
exploits the variations in unstable image content by observing the training examples. Furthermore, in contrast with the substantial performance drops resulting from descriptor NULL-ling in the retrieval task, we observe that a signal threshold may slightly improve classification results. However, this highly depends on the settings for $k$ and the threshold $t_{signal}$, and also varies across datasets. Opposed to this, instability modeling consistently improves over the baseline on all datasets and $k$, where a noise variance $\sigma_{noise}^2$ of $10^{-2}$ or $10^{-3}$ has to be chosen.

Figure 2.12(b) shows the same effect as Figure 2.12(a), but with the use of a spatial pyramid level in the representation, which is commonly used to boost the performance. Here, the improvements also hold and even become more pronounced for the often used setting of $k=256$ [76].

Based on the observations made on the validation set, we conclude to not perform descriptor mapping based on signal thresholding, and to adopt a noise level of $\sigma_{noise}^2 = 10^{-3}$ for instability modeling. These settings are applied on the Pascal VOC 2007 test set, for which results are reported in Table 2.1. The results show improvements for every FV component and combinations thereof.

Table 2.1: Recognition performance on Pascal VOC 2007 test set, using spatial pyramids and $k=256$ GMM components. Included are the results for $0^{th}$ order ($w$), $1^{st}$ order ($\mu$) and $2^{nd}$ order ($\sigma$) statistics, and combinations thereof.

<table>
<thead>
<tr>
<th></th>
<th>$w$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\mu + \sigma$</th>
<th>$w + \mu + \sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard FVs</td>
<td>39.0%</td>
<td>55.4%</td>
<td>56.7%</td>
<td>58.9%</td>
<td>58.9%</td>
</tr>
<tr>
<td>Proposed FVs</td>
<td>40.0%</td>
<td>56.2%</td>
<td>57.1%</td>
<td>60.5%</td>
<td>60.7%</td>
</tr>
</tbody>
</table>

In summary, it is always beneficial to incorporate descriptor instability in the FV as long as the noise variance for instability modeling is not too high (i.e. $\leq 10^3$).

The Effect of Variance Estimation

We next present a number of recognition experiments on the Pascal VOC 2007 validation set in which variations of the proposed method are considered. First, instead of estimating the variance by instability modeling, we consider assigning the same variance to all descriptors. Second, we apply instability modeling either during GMM learning or FV coding in order to determine where it has most effect. The results are presented in Table 2.2.

Using the same non-zero variance for all descriptors has a marginal negative
Table 2.2: Variations of the proposed method (on the Pascal VOC 2007 validation set using $k = 256$ and spatial pyramids). Instead of estimating the variance per descriptor, a fixed variance can be used for all descriptors. Variance estimation by instability modeling can be performed either during GMM learning (N-EM) or FV coding (N-FV), or both. Using a fixed variance of 0 constitutes the baseline, whereas the proposed method is N-EM+N-FV.

<table>
<thead>
<tr>
<th>Fixed Variance</th>
<th>Estimated Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>N-EM</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>N-FV</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>N-EM+N-FV</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td>N-EM+N-FV</td>
</tr>
</tbody>
</table>

effect, because stable descriptors may be spread out too much whereas the opposite holds for unstable descriptors. Note that a fixed variance of 0 constitutes the standard FV. Furthermore, the table shows that per-descriptor variance estimation by instability modeling has most effect in the FV coding step, as compared to GMM learning. This illustrates that the enhanced stability of the GMM (as depicted in Figure 2.5) not necessarily implies very substantial performance improvements. More gain is obtained from the FV coding step because this directly affects the image representation.

Run-time Comparison

The proposed method is computationally somewhat more expensive than standard FVs. Computing the responsibilities from noisy observations requires an extra dot product (in the log domain) in Eq. (2.8). Furthermore, the instabilities propagate to the computation of second order statistics in Eq. (2.11) for GMM training and Eq. (2.14) for FV coding, involving, for all observations, an extra element-wise summation in both Eq. (2.11) and Eq. (2.14), and an extra division in Eq. (2.14). The proposed method is $8.9\%$ slower as compared to NULL-ing the low signal patches ($t_{signal} = 0.01$), which is determined by computing the total runtime of extracting all descriptors from the VOC2007 train set.

2.5 Discussion

In this chapter we make the observation that local image descriptors extracted from low-signal image patches are unstable in feature space. We fit an exponential relation between signal strength and descriptor instability and exploit
2.5 Discussion

the estimated instability as measurement variance in a novel Fisher Vector feature encoding scheme. The proposed framework allow to model the descriptor instability in a principled way, as opposed to employing a threshold on the gradient signal. In effect, the discriminative information of these unstable descriptors is better preserved. The results show improvements for image classification, retrieval and matching. The proposed method can be especially beneficial in settings where classification is performed by direct descriptor matching.
Per-Patch Metric Learning for Robust Image Matching *

3.1 Introduction

Viewing and lighting condition changes in real-world scenes cause substantial variations in image feature representations. Significant progress has been made in developing image representations that are invariant to visual changes due to photometric [30, 32] or geometric [5, 51, 58] transformations. These invariant image representations are beneficial for applications such as object recognition, image retrieval and scene recognition.

A full invariant representation, unfortunately, leads to a decrease in discriminative power [93]. This trade-off is due to discriminant variations that an invariant representation cannot capture. For example, under rotational invariance a ‘6’ is identical to a ‘9’. Another disadvantage of invariant image representations is that they negatively effect stability [33, 90]. This is due to their sensitivity to noise when the image signal is low or ambiguous. For example, a rotational invariant that chooses the dominant orientation [58] is unstable when multiple equally dominant orientations are present.

Current invariant methods are always ‘on’. One can either choose to use the invariance, or choose not to use it; there is no middle-ground. It is not possible to have invariance for only some shading or only slight rotations. These rigid properties of current invariants play a central role in the trade-off between in-

*Submitted to the Asian Conference on Computer Vision in extended form [40].
Figure 3.1: 2D PCA projection of SIFT (blue); 1000 affine transformations of the top-left image patch (red); same-class samples (yellow). The top row is the original space, the bottom row is after learning the metric.

variance and discriminative power. Here, we propose to replace these binary on/off invariants, by steering the invariance to a limited range of disturbances. Such a limited degree of invariance is called robustness. For example, in the case of rotation, the proposed method can learn that a ‘6’ is only invariant up to $\pm 45^\circ$ of (and thus robust to) rotation, therefore eliminating the confusion with ‘9’.

By allowing a degree of invariance, a single global image representation cannot be used as it essentially depends on the specific image content how the limited transformation range will take effect, which is illustrated by the ‘6’ and ‘9’ example. Robustness must be achieved on a per-patch basis. Figure 3.1 (top) illustrates that the feature distributions after a transformation depend on the patch content, since even instances within the same class behave differently (red versus yellow). The bottom row of Figure 3.1 illustrates the effect of steerable invariance applied to each patch.

In order to achieve robustness, a Mahalanobis metric is computed for each individual patch. In effect, the metric weights the subset of feature dimensions that require robustness. For this, a relevant subset of transformations is generated and a metric that is specific for only those transformations is learned. Two approaches are presented for learning the metric: (i) full and (ii) direct. In the full method, synthetic image patches are generated, descriptors are extracted for each patch, and the robustness is obtained through a metric that is learned on these descriptors. In the direct approach, we only once generate a transfor-
3.2 Related Work

Figure 3.2: The flow of the proposed method. A range of transformations are applied on each query patch. With the transformations we learn a robust patch-specific metric, and use this map to estimate the metric from the patch directly without explicitly generating any synthetic images. The approach is illustrated in Figure 3.2. For learning the metric we here focus only on geometric image transformations.

3.2 Related Work

Figure 3.3 relates the proposed approach of this chapter to existing literature. Approaches that aim to achieve (full) invariance either use a transformation model based on the laws of physics [30, 32] or a model of the observed variations [5, 51, 56, 58]. The two disadvantages of invariants — stability and discrim-
Invariance can be addressed by propagating camera noise parameters [33] or by deriving quasi-invariants [90]. Noise propagation requires proper noise estimation and the quasi-invariants are incomparable over different images and thus cannot be used for matching. In this chapter, we specifically focus on matching images by learning a distance metric that is not fixed.

<table>
<thead>
<tr>
<th>Invariance</th>
<th>Fixed/Binary</th>
<th>Learned/Steerable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Physics Based</td>
<td>Learned/Steerable</td>
</tr>
<tr>
<td></td>
<td>Model Based</td>
<td>Physical Models</td>
</tr>
<tr>
<td></td>
<td>No Models</td>
<td>No Models</td>
</tr>
</tbody>
</table>

In contrast to employing pre-determined models, (deep) learning methods learn invariant features from unsupervised training examples [41, 70, 77]. Such methods do not explicitly model invariance as they attain robustness from training examples. Therefore, learning methods require large amounts of training data which is hard to obtain. Moreover, pure learning approaches do not directly incorporate known physical laws of the world. In this work we use a hybrid approach of modelling robustness by learning from generated geometric training data.

Synthetically generated data can be used to directly create variation in the train and test samples [10, 26, 50]. Other brute-force methods like ASIFT [60] generate a full range of affine transformations for both training and testing images which are used in an exhaustive matching scheme. Similar to our work, Simard et al. [82] avoid brute-force approaches and use synthetically generated images to learn a robust distance metric which is tangent to the manifold that is spanned by the generated transformations. We also learn a robust metric, however, where Simard et al. [82] require pixel values to estimate a manifold, our method estimates a Mahalanobis distance, which is applicable to any feature representation such as SIFT or LBP. To improve discriminability of a local descriptor, Cai et al [10] also propose to learn a projection matrix for a limited range of affine transformations through generated data. However, it is important to note that the authors learn a global projection whereas it is proposed
in this chapter to learn patch specific projections. Figure 3.1 illustrates that the same transformations applied to even the same class instances has different effects for different patches. These variations are thus patch-dependent and might not reflect the appropriate effect on other patches from the same class.

In a supervised scenario, Mahalanobis metrics can be learned to allow class-specific weighting of feature dimensions [3, 43, 57, 100]. Such local approaches learn multiple distance metrics for various parts of the feature space [103, 100] or for every single image [25], as opposed to learning a single global metric. We derive inspiration from these local methods to learn a distance metric for each patch, however we do not use class labels. The proposed method learns either from generated instances (full metric), or directly from a patch, without explicitly generating any instances (direct metric). This chapter thus contributes by the development and evaluation of novel methods for per-patch metric learning.

### 3.3 A Single-patch Metric

We first develop a metric that is learned from synthetically generated geometric distortions.

#### Geometric Transformations

Images are subject to geometric distortions introduced by perspective effects caused by view point changes. For small patches, the perspective transformation \((x', y')^T\) can be approximated by an affine transformation for a given point \((x, y)^T\) as

\[
\begin{pmatrix}
x' \\
y' \\
1
\end{pmatrix} = \begin{pmatrix} s_x \cos \alpha & -\tau_x \sin \alpha & t_x \\ \tau_y \sin \alpha & s_y \cos \alpha & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\
y \\
1
\end{pmatrix},
\]

where \(s\) denotes scale, \(\tau\) represents shearing, \(\alpha\) is the rotation angle, and \(t\) denotes translation. Examples of the transformations are depicted in Figure 3.4. Sample generation involves repetitive random selection from appropriate parameter ranges.
3.3.1 Metric Learning

A Mahalanobis distance metric between image features $x_i$ and $x_j$ is parameterized by the matrix $M$

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)}, \quad (3.2)$$

where $M$ is a positive semi-definite matrix. Eq. (3.2) simplifies to the Euclidean distance when $M$ is the identity matrix. Otherwise, $M$ considers different weights for different feature dimensions.

A straightforward approach to compute $M$ is to use $M = C^{-1}$ where $C$ is the empirical feature covariance of the training data [3, 43]. The rationale behind using the inverse covariance as a metric is that for objects of the same class, a high variance in a feature value means that this feature is not very stable. The most informative feature dimensions are those that have low variance on training data.

For large-scale matching it is convenient to use fast indexing techniques such as trees [57]. Such techniques typically work with the Euclidean distance. To this end, the metric can be rewritten to $(x_i - x_j)^T W^T W (x_i - x_j) = ||Wx_i - Wx_j||^2$, where $W = M^{-\frac{1}{2}}$. This effectively scales the feature space with an affine transformation $W$ to allow the Euclidean distance to be used for metric $M$.

Robustness is obtained by generating a limited range of transformations for
a single patch. From these generated samples the covariance $C$ is estimated, which is then used to compute $M$.

We coin this simple method for metric learning the full approach, since it needs to fully generate a large number of transformations for a patch in order to extract the image features and estimate the covariance matrix.

### 3.3.2 Direct Metric Estimation

Instead of using the brute-force approach of computing a covariance matrix by explicitly generating all patches and extracting features for each of them, we propose a direct approach to estimate the metric per patch. In the direct approach, the covariance is estimated from a single patch.

In the following, let $x$ be a column vector, with dimensionality $N$. For clarity, we start with pixel values to explain the direct metric estimation, i.e. $x$ is a vector of image pixels.

**Pixel Values** The direct metric is a combination of two terms, a transformation probability $D_T$ that a pixel moves to a different position after transformation $T$, and the patch-specific covariance term $V$ of the feature values.

Let $D_T$ be a symmetric matrix of size $N \times N$, containing the probabilities $P_T(i|j)$ of a pixel at position $j$ affecting the position of pixel $i$ under a transformation $T$. Note that this transformation is independent of the actual feature values. Matrix $D_T$ represents the per-pixel transformation probability which is determined by simulating a large set of transformations and comparing the transformed patch with the ground truth location obtained through the homography in Eq. (3.1).

The matrix $V$ of size $N \times N$ is the variance matrix with elements $\sigma(i, j)$ representing the covariance of pixel values at position $i$ with respect to position $j$. To compute the covariance matrix, we need the expected average weighted pixel value of the pixel values in $x$ at position $i$ after the transformation $T$, which is given by

$$E[x_T(i)] = \sum_{j=1}^{N^2} P_T(i|j)x(j) = D_T x^T,$$

where $x_T$ represents the pixel vector $x$ after the transformation $T$. The expected value of the transformed image $x_T$ is denoted by $E[x_T(i)]$ and represents the average image of all transformations that are present in $D_T$. See the top row of Figure 3.5 for some example images of $E[x_T(i)]$. 

Figure 3.5: Top row: the average transformation $E[I_T(i)]$ after applying the motion map in the direct metric estimation. Middle row: visualization of independent pixel motion. Bottom row: dependent pixel motion, as used in the full metric estimation.

The covariance $\sigma(i, j)$ after transformation $T$ is then

$$\sigma(i, j) = E[(E[x_T(i)] - x_T(i))(E[x_T(j)] - x_T(j))]. \quad (3.4)$$

Note that in the transformation of $x_T(i)$ and $x_T(j)$ it is allowed for pixels to move independently to other pixels. To illustrate this, in the middle row of Figure 3.5 several images $x_T$ are created by sampling the distribution of motion probabilities $D_T$. None of these images are actually used in the direct metric estimation approach. Eq. (3.4) can be rewritten in matrix form to obtain $V$ directly

$$W_V = [[D_T > 0]] \bullet (D_T \times 1^T - (x^T 1)^T), \quad (3.5)$$

$$V = \frac{1}{N^2} D_T \bullet W_V W_V^T, \quad (3.6)$$

where $1$ is a column vector of all ones, $\bullet$ denotes element-wise multiplication and $[[\cdot]]$ indicate Iverson brackets which resolves a (matrix) element to 1 when the argument is true, and 0 otherwise. The metric is computed by $M = V^{-1}$, and the transformation by $W = V^{-\frac{1}{2}}$.

**SIFT** For other descriptors, the $D_T$ matrix of transformation probabilities can be reused and is not required to be recomputed. The $V$ matrix, however, has to be adapted to the specific form of the descriptor. In the case of SIFT, $D_T$ is converted to $D_T^{\text{sift}}$ of size 128x128. In contrast to generating all possible SIFT values, this transformation only has to be computed once.

The 128 dimensions of SIFT comprise of a 4x4 spatial grid and 8 angular bins (4x4x8). We use the pixel-based transformation probabilities $D_T$ to directly calculate the probability $P_T^{\text{sift}}(i|j)$ for spatial SIFT bins $i$ and $j$ with

$$P_T^{\text{sift}}(i|j) = \sum_{y=0}^{N^2} [[f(x) = i]][[f(y) = j]] P_T(x|y), \quad (3.7)$$
where the function $f(x)$ maps the pixel at location $x$ to the correct SIFT-bin $i$. For the angular transformation probabilities, we assume independence with the spatial bins. The unit circle is sampled with 360 vectors and transformations are applied to these vectors. Since the original orientation is known, this results in counting how often a vector switches bins after the transformations, as illustrated in Figure 3.6. The joint 128x128 matrix $D_{T}^{\text{sift}}$ is calculated by multiplying the angular probabilities with the spatial probabilities.

![Figure 3.6: Angular transformations. Original angular SIFT binning is depicted by the white lines, whereas the red colored segments show how orientation vectors switch bins.](image)

Finally, we note that SIFT is $L_2$ normalized, which obscures the true probabilities of the angular bins. Hence, $L_2$ normalization is performed after applying $W$ (Section 3.3.1). The computation of $M_V^{\text{sift}}$ is identical to the pixel value case in eq. Eq. (3.6). The input feature $I$ is now the unnormalized 128d SIFT vector.

### 3.4 Experiments

The full and direct metric estimation methods are evaluated on two different publicly available datasets for image matching: MNIST [49] and ALOI [29]. For both datasets we consider metric learning for achieving geometric robustness. Photometric robustness is assessed on ALOI only.

#### 3.4.1 Geometric Robustness

Here, we evaluate per-patch metric learning under geometric transformations. The performance of the proposed methods is compared against other existing methods, and we evaluate how different combinations of affine changes affect matching performance.

The optimal affine parameter ranges are obtained on separate validation sets for both MNIST and ALOI. Parameters are repetitively sampled (1500 times)
from these ranges to either generate and apply geometric transformations in the full approach or to estimate the transformation probabilities $P_T(i|j)$ in the direct approach.

Classification is performed by feature matching in a 1-NN classification scheme.

**MNIST Matching**

The MNIST dataset consists of 60,000 handwritten digits for training and another 10,000 for testing, see e.g. Figure 3.4 and Figure 3.5 for examples. Both raw pixels and SIFT are considered as image features.

Parameter selection is performed on the first 200 images of the training set. The best transformation combination for MNIST dataset comprises moderate amounts of translation (in [-2:2]), rotation (in [-.1:.1]) and shearing (in [-.125:.125]) (see Eq. (3.1)) for SIFT features. Scale has no influence, because the digit size is almost the same throughout the dataset. Shearing and rotation help to learn variations such as italic writing while translation helps to learn small shifts of the digits from the center.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cl. Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>96.57</td>
</tr>
<tr>
<td>SIFT + Rotation invariance</td>
<td>95.61</td>
</tr>
<tr>
<td>SIFT + Affine invariance</td>
<td>96.55</td>
</tr>
<tr>
<td>SIFT + RCA [3]</td>
<td>96.16</td>
</tr>
<tr>
<td>SIFT + Proposed-full</td>
<td><strong>96.87</strong></td>
</tr>
<tr>
<td>SIFT + Proposed-direct</td>
<td>96.49</td>
</tr>
<tr>
<td>Raw Pixel</td>
<td>96.91</td>
</tr>
<tr>
<td>Raw Pixel + one sided TD [82]</td>
<td>98.12</td>
</tr>
<tr>
<td>Raw Pixel + two sided TD [82]</td>
<td><strong>98.61</strong></td>
</tr>
<tr>
<td>Raw Pixel + Proposed-direct</td>
<td>96.88</td>
</tr>
<tr>
<td>SIFT + SVM (linear)</td>
<td>97.40</td>
</tr>
<tr>
<td>SIFT + LDA + 1-NN</td>
<td>97.39</td>
</tr>
<tr>
<td>SIFT + LMNN 1-NN Mahalanobis [100]</td>
<td><strong>97.50</strong></td>
</tr>
</tbody>
</table>

**Table 3.1:** Classification rates for different methods on MNIST dataset.

Classification performance is compared against unsupervised and supervised
3.4 Experiments

state-of-the-art methods in Table 3.1. Best performance for this dataset is obtained with tangent distance (TD) [82]. As discussed in Section 3.2, Simard et al. invoke prior information by generating small global transformations. This supports our idea of exploiting prior information for steering the invariance. TD however requires pixel values as image features, which makes applicability of this approach harder for some other feature representation such as SIFT.

SIFT does not benefit from rotation invariance. This is due to the associated decrease in discriminative power. However, affine invariance does not affect the performance. The per-patch full metric preserves discriminative power while moderately benefiting from improved robustness, as for example illustrated by the ‘6’ vs ‘9’ example. The direct metric assumes independent pixel movements which makes it difficult to preserve shape information, as illustrated by the performance of the ‘Proposed-direct’ methods in Table 3.1. Note that ‘Raw Pixel + Proposed-full’ is not considered as the computational load associated to generating enough examples and computing covariance is prohibitively large.

Supervised methods generally outperform unsupervised methods on MNIST. To show the strength of the proposed metric learning approaches, we next consider a matching task in which no class labels are available and pixel-based methods such as TD fail due to increased visual complexity.

ALOI Matching

The ALOI dataset contains 1000 objects under varying imaging conditions. The variations are due to illumination direction, illumination color and camera viewpoint. Here, we focus on geometric transformations and thus only consider patches recorded from different viewpoints under canonical illumination conditions, i.e. with all light sources switched on. The recording setup consists of three cameras positioned next to each other. For the following matching experiments, patches are sampled from the two outer cameras (that are furthest apart). To find the matching pairs from two outer cameras, the same procedure is followed as in [21]. In total 8300 matching pairs are extracted of which 200 are used for validation. The patch size is $64 \times 64$, whereas SIFT is extracted from a $56 \times 56$ window around the center to avoid the artificial border edges due to the geometric transformations for the full metric method. Parameter selection is performed as described in Section 3.4.1 on the validation set.

The influence of sweeping each individual transformation range on the validation set is illustrated in Figure 3.8. The figures show that the direct and full metric agree more or less on their performance changes for different parameter
settings. A joint optimization of parameters on the ALOI dataset yields translation in [-2:2] and shearing in [-1:1] for SIFT descriptors. Scale and orientation do not affect performance as the positional difference between cameras do not yield large scale and orientation variations.

<table>
<thead>
<tr>
<th>SIFT</th>
<th>+Rot. Inv.</th>
<th>Full Metric</th>
<th>Direct Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.85%</td>
<td>13.86%</td>
<td>61.42%</td>
<td>61.07%</td>
</tr>
</tbody>
</table>

Table 3.2: Classification rates on the geometric ALOI dataset based on SIFT features.

The results in Table 3.2 show that the proposed full and direct SIFT-metric methods have significant improvement of 6.57% and 6.22% over regular SIFT performance respectively. The substantial performance increase for the ALOI dataset is due to the fact that viewpoint variations degrade SIFT performance for matching. Considering rotation invariance for SIFT leads to a dramatic performance loss as most discriminative information is ignored and due to the visual complexity it is harder to estimate a dominant gradient orientation.

Considering raw pixels and tangent distance on this set, we obtain 16.54% and 23.47% respectively. As expected, the raw pixel matching performance drastically drops for this dataset. Furthermore, as texture rather than shape is prominent in ALOI patches, the direct metric does not suffer substantially from its assumption with respect to independent pixel movement.

As the dataset contains no class labels, the supervised methods used for comparison in Section 3.4.1 cannot be included here.
3.4 Experiments

3.4.2 Photometric Robustness

We next evaluate the proposed direct method for photometric robustness. Matching patch pairs are selected from the same viewpoint but with different illumination. The same evaluation procedure is followed as in Section 3.4.1. With SIFT matching, we obtain a baseline result of 73.40% matching accuracy. Using the direct metric from the geometric ALOI dataset increases the performance to 76.88% accuracy.

The direct method does not rely on synthetic photometric data generation, but nonetheless substantially outperforms SIFT when confronted with photometric variations in the data. Thus, the direct approach, which is developed for achieving robustness against geometric transformations, is also applicable in the context of photometric variations. This positive side effect is due to the fact that the proposed method is robust against signal changes in general, which may also be caused by variations in illumination.

Figure 3.8: ALOI validation set parameter selection. Blue, Red and Green lines represent Direct, Full and SIFT performance, respectively.
3.4.3 Complexity Analysis

We perform an analysis of the computational complexity of the per-patch metric learning approaches that are considered, i.e. by either explicitly applying geometric transformations or directly per patch. The analysis is based on the steps involved in matching a query patch to a single database patch. The load associated to the simulation of geometric transformations is denoted by $T$. This involves applying the actual affine transformation to the query patch. Its contribution to the complexity depends on the number of repetitions, i.e. the size of the dataset from which to estimate the covariance, and the patch size. The computational load associated to estimating the full covariance $C_{\text{full}}$ is determined by the actual computation of covariance from the generated set of descriptors. Direct computation $C_{\text{direct}}$ involves one matrix multiplication of the descriptor with the displacement matrix, subtracting the result from the original and squaring the result. The displacement matrix is optimized offline on a holdout set and is therefore not relevant to the analysis of runtime complexity. Computing the metric $M$ involves inverting the covariance and decomposing the result. Applying the metric to a patch descriptor involves a matrix multiplication and is denoted by $A$. This is performed on both the query and the database patch. Finally, the computational load associated to distance computation is denoted by $D$. The computational complexity of metric learning based on full covariance is then $O(T + C_{\text{full}} + M + 2A + D)$, whereas the proposed direct approach is associated to a complexity of $O(C_{\text{direct}} + M + 2A + D)$. Thus, doing away with the necessity of actually applying the transformations and computing the covariance grants a big advantage to the direct approach. The graph in Figure 3.9 compares the full and direct approaches for increasing patch size.

![Figure 3.9: Full and direct metric efficiency comparison. 10 repeats, 100 random transformations.](image-url)
To illustrate the effect of per-patch metric learning, we consider the examples in Figure 3.10. Figure 3.10 shows matching pairs of image patches, and their difference image. These patches are extracted from two different viewpoints, by which a geometric distortion is introduced. We observe from Figure 3.10.a that the regions in the patch that remain stable across viewpoints (i.e. have zero difference) are those that lack structure. For example, the homogeneously red colored regions of the patches in Figure 3.10.a are unaffected by the transformation. By repetitively simulating such (affine) transformations and recording the resultant descriptors, a dataset is generated which exhibits low variance for those descriptor dimensions that remain stable under these image transformations. In mahalanobis metric learning setups, these variations are inverted and consequently translate to high weights in the distance metric. Considering again the example in figure Figure 3.10.a, the metric will thus emphasize structureless regions such as the red colored part at the bottom. More uncertainty is associated to regions near prominent structure such as edges, which is reflected by strong energy in the difference image. Consequently, a metric learnt from highly textured patches such as the ones in figure Figure 3.10.b will grant low weight to a large amount of descriptor dimensions (and thereby affecting the distance measure only marginally). At the other side of the ‘texturedness spectrum’, a metric learnt from a completely homogeneous patch such as in Figure 3.10.c will grant high weight to all descriptor dimensions, which leaves the distance measure basically unaltered. It is thus the structure-based localization of structureless regions that is performed by per-patch metric learning, where most effect is expected for patches that are neither homogeneous nor highly textured. Note that one such metric could be derived directly from image derivatives. That is, image derivatives can be thought of as the expected variance per pixel when the image is subject to small translations. Also note that it is essential to perform per-patch processing in order to obtain these kind of metrics.
3.5 Discussion

In this chapter, we propose a generic patch-specific robust metric learning method to improve matching performance of local descriptors. We show that a full invariant representation leads to a decrease in discriminative power of descriptors. Therefore, we propose a per-patch metric learning method that allows invariance to only an optimizable transformation range. Two approaches for learning the metric are presented: full and direct. The proposed approaches are validated on two publicly available datasets exhibiting geometric and photometric image transformations. It is shown that the proposed approaches outperform the original SIFT descriptor in terms of matching performance.
Per-patch Descriptor Selection using Surface and Scene Properties *

4.1 Introduction

Representing local image structures is important for many computer vision tasks such as (object) recognition, wide baseline matching and tracking. In these tasks, a generic image descriptor is typically chosen which should be well-suited for describing all possible image patches.

Image patch appearance is determined by combinations of material properties, such as color and texture, with accidental scene properties such as illumination conditions, viewpoint, scale, and so on. A successful image descriptor should have high discriminative power between material properties, while remaining invariant against disturbing instances of scene-accidental conditions.

Invariance, however, is inversely related with discriminative power [90, 93]. Many excellent image descriptors have been designed [89] or optimized [7] to find a good trade-off between invariance and discriminative power. Nevertheless, a single descriptor cannot be optimal in all cases. Consider for example a patch containing highlights in the top row of Figure 4.1a. Using a highlight-invariant descriptor would increase the matching score. On the other hand, consider the bottom row of Figure 4.1a. Using the highlight-invariant descriptor may actually remove the discriminating characteristic. A similar argument holds for material properties such as texture and color. For the example in the

*Published in the European Conference on Computer Vision [21].
46 Per-patch Descriptor Selection using Surface and Scene Properties

- Reflectance
- Photometric stability
- Response ratios
- Geometric stability
- Micro-texture
- Softness
- Texturedness
- Colorfulness

Figure 4.1: (a) Example input patches to our method. The rows contain corresponding patches under a range of photometric and geometric disturbances. (b) Schematic of the descriptor selection algorithm.

second row of Figure 4.1a it makes little sense to use a color descriptor since it becomes unstable with little color present [90]. The conflicting demands on the degree of invariance and between material representations cannot be resolved by a single descriptor.

In this chapter, we propose a method to select the best descriptor for a single patch. For example, if we can detect the difference between a strong scene-accidental highlight and a glossy material surface in Figure 4.1a, then we can select a suitable different descriptor in both cases. With this aim, we identify material properties [1, 52] and scene-accidental properties [90, 89] which we will use in a supervised learning scheme that can take mismatching costs into account. See Figure 4.1b for a schematic overview of our approach. While we are not aware of any articles in which physical properties are related to descriptors on a per-patch basis, we review relevant works in the following.

4.2 Related Work

A local image descriptor can be optimized discriminatively out of combined variations of atomic operations such as smoothing, angular quantization, spatial pooling and feature normalization [7]. Alternatively, a projection can be learned on (sift) features to reduce descriptor size and simultaneously improve matching performance [10] or visual word assignment in retrieval [69]. These methods take a descriptor and improve its overall performance, resulting in a better single-purpose descriptor for all patches. It is, however, not possible to tune a descriptor to a single patch, as we propose in this chapter.
4.3 Descriptor Selection using Surface and Scene Properties

For category-level image classification, the best image-level descriptor can be learned for each category. When such supervised information is available, various machine-learning techniques can learn the best descriptor combination by boosting [19, 27, 64], multiple kernel learning [93], topographic filter maps [41], dimension reduction [52, 15] or the Fisher criterion [35]. These methods cleverly exploit the intra- and inter-class variance between the image categories. However, when category labels are not available, as for example in feature tracking or wide-baseline matching, these methods cannot be applied. We propose a generic method that is suitable for such applications by selecting the best descriptor based on the material properties of a single patch.

Material recognition is a generalization of texture recognition, which is widely studied, see e.g. [11]. Recently, more generic material classes such as glass, metal and fabric have been proposed [52] and followed-up by [15]. These methods aim to find the named material class of a given image, such as wood, leather or stone. In this chapter, however, we are not as much interested in the class per se, as this would require a large database of common material classes. Alternatively, we propose to find the best image descriptor of a patch based on its structural and surface reflectance properties.

Surface reflection can be characterized by the bidirectional reflectance distribution function (BRDF). The BRDF represents the reflection ratio for all surface locations under all possible illumination and viewing directions. Despite the complexity of the BRDF, there are methods to estimate it under constraints on object shape or illumination direction [18, 73, 99]. In this work, we are interested in unconstrained shapes and illuminations and therefore focus on simpler features.

4.3 Descriptor Selection using Surface and Scene Properties

To select the best descriptor for an image patch, it is important to identify features that can estimate material properties such as colorfulness, roughness, shininess etc. Moreover, the pool of available descriptors to choose from has to be diverse enough to emphasize or ignore those properties that are important for recognition. For example, a smooth shiny patch from an apple will benefit from keeping the shininess and perhaps not focusing on edges too much. On the other hand, a cast shadow or the position of a strong highlight is scene accidental, and therefore better ignored. The material properties should be able
to measure and represent such effects from an image patch whereas the image descriptors should ideally be able to distinguish between various levels of invariance. Such levels of invariance apply to the object’s structure, such as edge-based vs. pixel-based, but also on photometric invariant properties such as highlights, shadows and shading.

### 4.3.1 Photometric Representations

Photometric invariance can be modeled by the dichromatic reflection model [78]. In this model, an RGB vector \( \mathbf{f} = (R, G, B)^T \) is the vector summation of the body reflectance with the specular interface reflectance

\[
\mathbf{f} = e(m^b c^b + m^i c^i),
\]

where \( e \) is the intensity of the light source, \( c^b \) is the color of the body reflectance, \( c^i \) the color of the interface reflectance, the scalars \( m^b \) and \( m^i \) depend on the surface orientation and represent the magnitude of the body and interface reflection respectively.

For representing image invariants we consider the transformation to the opponent color space [90, 89]. Save scaling factors, the transformation is given by

\[
\begin{pmatrix}
O_1 \\
O_2 \\
O_3
\end{pmatrix} = \begin{pmatrix}
R - G \\
R + G - 2B \\
R + G + B
\end{pmatrix}.
\]

The opponent color components are combined in four different representations. First, the chromatic components \( O_1 \) and \( O_2 \) are separated from the intensity component \( O_3 \). On itself, \( O_3 \) has no invariance properties but does generally contain most information regarding image structure. Due to the subtraction of \( RGB \) components, \( O_1 \) and \( O_2 \) are invariant with respect to shifts in illumination such as highlights. Nevertheless, \( O_1 \) and \( O_2 \) are still sensitive to illumination scalings such as shadow and shading. To this end, we also consider intensity-normalized chromatic components \( \frac{O_1}{O_3} \) and \( \frac{O_2}{O_3} \). These intensity-normalized invariants however, are again sensitive to illumination shifts. Therefore, the set of photometric representations is complemented with \( \text{hue} = \frac{O_1}{O_2} \), which is invariant to both illumination scalings and shifts. The full set of photometric representations that we consider in this chapter is given in Table 4.1. Patch descriptors and image attributes are extracted from these representations as detailed in the following.
4.3 Descriptor Selection using Surface and Scene Properties

<table>
<thead>
<tr>
<th>Representation</th>
<th>Invariant to</th>
<th>Descriptor name</th>
<th>Intensity</th>
<th>Chromatic</th>
<th>Normalized</th>
<th>Hue</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_3$</td>
<td>-</td>
<td>I.pix/I.grad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Highlights</td>
<td>C.pix/C.grad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shadows</td>
<td>N.pix/N.grad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hl. &amp; Sd.</td>
<td>H.pix/H.grad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Image representations and descriptor names.

4.3.2 Image Attributes

Low-level image attributes have been used to measure a degree of objectness in a bounding box [2], or in a functional extension of the spatial pyramid beyond spatial information towards more generic types of pooling [92]. Here we are interested in low-level features to identify surface structure and reflection. Surface reflectance properties such as shiny, matte or gloss have been found to correlate with simple image statistics [18, 61, 79]. Surface structures such as crinkles in leather or grains in paper have been proposed to be detectable by subtracting a bilateral filtered image from the original [52]. Further, the difference in edge types in e.g. metal, glass, or paper can be linked to the variance of the gradient magnitude or orientation [15]. We will evaluate and extend these low-level surface features to link structure and reflectance surface qualities to an image descriptor. To avoid any confusion between local image descriptors which are also often called features we will use the term attribute for low-level image features that measure structural and photometric surface/scene properties. We detail these attributes in the following.

(a) Interface reflectance. The dichromatic model in equation Eq. (4.1), has a term for the object reflectance $m^b c^b$ and a term for the interface reflectance, $m^i c^i$. The latter term, representing gloss, matte/shininess, has been found to correlate with the skew (third-moment) of the intensity histogram [61]. Other research also uses the standard deviation, 10th, and 90th percentile [79] and kurtosis [18] to represent the shape of the intensity histogram to predict interface reflectance. The amount of interface reflection is a valuable attribute for selecting between the highlight invariants ($[O_1, O_2]$, $\frac{\partial I}{\partial O_1}$, $\frac{\partial I}{\partial O_2}$) and the highlight variants ($O_3$, $[\frac{\partial I}{\partial O_1}$, $\frac{\partial I}{\partial O_2}$]). Therefore we use these intensity statistics over $O_3$ as patch attributes.

(b) Photometric stability. The invariant representations introduced in the previous section are insensitive to various photometric transformations. However, this comes at a price that is paid in numerical instability [90]. The $\text{hue} = \frac{\partial I}{\partial O_2}$ invariant is unstable for colors on the black-white axis (i.e. low saturation), whereas the intensity-normalized $\frac{\partial I}{\partial O_1}$ and $\frac{\partial I}{\partial O_2}$ invariants are unstable near zero.
intensity. The occurrence of this instability depends on the surface reflection, and varies per patch. Therefore, these features are well-suited attributes to determine if an invariant representation is suitable. To this end, we use the mean intensity $\mu(O_3)$, and mean saturation $\mu(\sqrt{O_1^2 + O_2^2})$ as photometric stability attributes. Moreover, to obtain a richer representation, we compute the same statistical values for saturation as we did for intensity in the previous paragraph.

(c) Photometric response ratio. To obtain attributes specifically tuned to each invariant representation, we relate the response in the full-color representation to the response for each invariant. Different invariant representations will respond differently to shadows, shading, gloss and highlights and consequently this difference allows descriptor diversification. To this end, we compute the average gradient ratio [34] for each invariant with respect to the full color gradient, $\frac{|\nabla O_3|}{|\nabla [R,G,B]|}$ and similarly for other representations.

(d) Geometric stability. We include a sense of the geometric stability of the patch under a viewpoint change. The basic idea is that a small geometric transformation on a stable patch should lead to a small differences in the descriptor. Large differences may indicate high sensitivity to the disturbance. This is also a structured attribute since it gives a sense of homogeneity. The sensitivity is measured by a set of self-dissimilarities after applying a geometric transformation to the patch. Specifically, we depart from a centered sub patch (80%) cropped out of the image. The region of interest is then up- and down- scaled and translated such that a set of geometrically transformed versions of the initial patch is obtained. We take the average descriptor distances over two scales and eight directions as geometric stability attributes for each image descriptor.

(e) Micro-texture. The surface structure of a patch may be rough or smooth. Metal, for example, is typically smooth, whereas fabric is fine grained. To distinguish between rough and smooth surfaces, we follow the approach of Liu et al. [52] to detect micro-texture. Specifically, we subtract a bilateral smoothed version from the original image patch. As attributes we use the sum of the residual, and we do this for each of the four invariant representations separately.

(f) Softness. Material may also be soft as plastic, or hard as metal. As suggested in [15], we adopt the standard deviation of the gradient orientation, and the standard deviation of the gradient magnitude to measure material softness. The authors’ rationale is that soft materials have soft edges, with softly varying transitions in gradient orientation and magnitude. We compute this for each of the invariants and add the mean values to obtain a richer statistical representation.

(g) Texturedness. For a notion of material texturedness we build on the work of
4.3 Descriptor Selection using Surface and Scene Properties

[101] in which it is shown that a Weibull parameterization of images results in
textural diversification. Specifically, the contrast distributions of natural images
generally follow the 2-parameter integrated Weibull distribution. We compute
these two parameters, $\beta$ and $\gamma$ in each invariant, and also compute additional
statistics by counting the number of edges above a noise threshold.

(h) Colorfulness. As a measure of colorfulness we compute a single valued hue
entropy score, $-\sum (p \log_2 p)$ where $p$ is the histogram of the hue pixels $\frac{O_1}{O_2}$.

4.3.3 Image Descriptors

For constructing image descriptors, we use the photometric representations as
given in Table 4.1. Besides these photometric variation, we model structure
variation with multiple differential orders. Zeroth-order descriptors are his-
tograms extracted from pixels per color channel, whereas first-order (sift) de-
scriptors are based on the per-channel gradient orientations [89]. Note that the
order of differentiation affects the invariance properties of the descriptor. We
do not consider higher order representations. Spatial pooling of the descriptors
is obtained by aggregating features in a $4 \times 4$ cell grid as originally proposed
by Lowe. For zeroth-order descriptors we compute 8-bin histograms of pixel
values. For first order descriptors the gradient orientations are quantized in 8
bins. Furthermore, feature contributions are weighted by a Gaussian window
centered on the image patch. Finally, the descriptors are normalized to unit
length. The invariance properties and the names of the descriptors are given in
the bottom row of Table 4.1, where pix denotes zeroth-order, and grad indicates
first-order.

4.3.4 Descriptor Selection

We relate attributes to descriptors in a supervised learning setup. Our setup is
similar to [7]. However, where they learn the best single-descriptor parameters
over a training set, we leverage the patch attributes to learn the best descriptor
for a single patch. We start with a ground truth set which has for each patch
a corresponding transformed version of the same patch (under homography
or photometry, more details below) and 100 randomly sampled non-matches.
Such a set allows the computation of a matching score in average precision (AP)
for each descriptor type per patch. The attributes of the patch are the input for
our supervised setup whereas our goal is to select the descriptor that gives the
best average precision score.
Let \( X = \{x_1, x_2, \ldots, x_n\} \) be the patch ground truth data set containing \( n \) patches, where \( x_i \) is a \( p \)-dimensional vector containing the \( p \) attribute values. The corresponding average precision scores \( Y_i = \{y_{i1}, y_{i2}, \ldots, y_{id}\} \) for patch \( i \) are computed by ranking all retrieved patches according to each descriptor distance for all \( d \) descriptors. We aim to find a learning model \( L \) that maximizes the average precision for a patch, i.e., \( \mathbb{L}(x_i) = \arg\max_i(Y_i) \). Note that the cost of misclassification is not uniformly distributed over the descriptor classes since each descriptor typically gives a different average precision score. The misclassification cost \( c_i \) of selecting a descriptor for a patch \( i \) is the score of the selected descriptor minus the score of the best possible (oracle) descriptor in the pool, \( c_i = Y_{\mathbb{L}(x_i)} - Y_{\arg\max_i(Y_i)} \). To take non-uniform misclassification costs into account, we adopt the cost-sensitive support vector machine (SVM) approach by Zadrozny et al. [102]. This formulation incorporates the misclassification costs \( c_i \) directly in the SVM optimization problem. Note that the classes are descriptor types. For each binary \( g \)-vs-\( h \) sub-class problem (e.g. \textit{I.grad} vs \textit{H.pix}, see Table 4.1) this becomes

\[
\begin{align*}
\text{minimize}_{w, \xi, k} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^n c_i \xi_i \\
\text{subject to} & \quad b_i(w^T x_i + k) \geq 1 - \xi_i, \quad \xi_i > 0, \\
\text{where} & \quad b_i = \begin{cases} +1 & \text{if } y_{ig} > y_{ih} \\
-1 & \text{otherwise.} \end{cases}
\end{align*}
\] (4.3)

Thus, \( b_i \) denotes +1 if the average precision score of \( x_i \) of descriptor \( g \) is higher than the average precision of \( x_i \) of descriptor class \( h \); and −1 otherwise. If two patch descriptors have the same average precision score \( (y_{ig} = y_{ih}) \) in the training phase, we assign the patch to both descriptor classes (the misclassification cost \( c_i \) will subsequently be 0). We use class voting to obtain the multiclass label from all 1-vs-1 class-pairs. In the case of equal votes we assign the sample to the descriptor with the highest a-priori score on the training set.

Note that our 1-vs-1 setup allows us to utilize the full dataset for each descriptor class-pair. If we first assign the best descriptor to each patch, a binary descriptor-pair classifier could only train on those samples where the global maximum is obtained by one of the two classifiers. Because we use the pairwise maximum \( b_i \) for each descriptor-pair, we do not suffer from this problem.
4.4 Experiments

It is our hypothesis (Figure 4.1b) that the best descriptor is dependent on two factors: the accidental scene properties and the patch’s surface material properties. To evaluate this hypothesis we create a separate dataset for each factor. The influence of the surface material is tested by extracting attributes from a clean, canonical, patch without any distortions. This allows us to judge the influence of the surface alone. Alternatively, to evaluate the influence of the scene, we extract attributes from a photometric/geometric distorted patch. In this case, however, it cannot be helped that the surface will also have some influence. We refer to the patch used for prediction as the query patch, and the aim is to match the same patch in a set of non-matching patches.

In our experimental setup we use pairs of matching patches. For every match in the database, we sample 100 random non-matches, which is repeated 10 times. Retrieved patches are ranked based on the Euclidean distances to the query patch of the respective descriptors. From this we compute the average precision for measuring retrieval performance per descriptor. One half of the dataset is used for training, and the other half is used for testing. Note that a patch is exclusively in the train or in the test set.

4.4.1 Synthetic Dataset

We start with a synthetic dataset to evaluate the descriptors and attributes under controlled circumstances. The synthetic dataset is generated from the work of Barnard et al. [4] where measurements of 1995 surface reflection spectra and 287 illumination spectra are provided. The camera sensitivity function allows computation of $RGB$ values given the surface albedo and illumination spectrum. We create ‘Mondrian’-style images by inserting colored blocks of random size at random locations on a $64 \times 64$ image lattice. We keep the illuminant color fixed, rotate the colored blocks by a random degree, and introduce some skew and noise, so as to reflect a more diverse range of image content.
Per-patch Descriptor Selection using Surface and Scene Properties

Figure 4.3: Influence of geometric and photometric disturbances on descriptor matching performance. Mean average precision is plotted against the disturbance level on the x-axis. Pixel-based (pix) and gradient-based (grad) descriptors are extracted from intensity, chromatic, normalized chromatic and hue representations, denoted by I., C., N. and H (see Table 4.1). Figures a-f are the result of matching experiments in which the query patch is in canonical form. In figures g-l the query patch is distorted. Note that the per-descriptor performances are averaged over all disturbances against known numbers of foreground blocks to obtain the ‘complexity’ figures in f and l.

Pairs of matching patches are generated by applying a geometric or photometric disturbance to a synthetized image. Geometric disturbances encompass a translation or rotation. Photometric disturbances are achieved by applying a scaling (shadow and shading) or offset (highlights) to all RGB channels, see Eq. (4.1). The location and extent of the disturbance is governed by an anisotropic Gaussian with a random location and covariance. The disturbances are progressively increased, in five steps, where we generate a dataset consisting of 1000 matching image pairs per disturbance level. This is repeated for increasing amounts of foreground blocks, which denotes a basic notion of image complexity. See Figure 4.2 for some example patches.

Matching performance per descriptor under each of the disturbances is shown in Figure 4.3. The figure shows that pixel-based descriptors almost always outperform gradient-based descriptors on this dataset. This is partly due to the fact that the set of colors in the dataset is limited and distinct. However, as can be seen in Figure 4.3 a-b and g-h, it stands out that pixel-based descriptors are considerably less sensitive to geometric disturbances than gradient-based descriptors. Naturally, there is no difference between canonical or distorted query patches if the disturbance is purely geometric, as in Figure 4.3a, Figure 4.3g and Figure 4.3b, Figure 4.3h.

Pixel-based intensity-only (I.pix) descriptors fail when shadows and highlights are applied to the image (Figure 4.3 c-d and i-j) as expected. However, under additive gaussian noise, the I.pix descriptors are superior when the patch
is presented in distorted form. Gradient-based descriptors suffer more from noise in general. Furthermore, gradient-based opponent color descriptors are also sensitive to shadows, while normalized opponent color descriptors appear sensitive to highlights, which is in accordance with the respective photometric invariance classes. Gradient-based intensity descriptors also suffer from highlights because most edges comprise of color transitions which may become less prominent in the vicinity of highlights. It appears to be more difficult in general to retrieve the canonical form based on a distorted query patch than vice versa (see Figure 4.3 c-e and i-k). This is because the disturbance increases the average similarity to all patches (you may find a highlight if you look for it). Increased image complexity (Figure 4.3 f and l) generally results in improved matching performance for all descriptors. However, pixel-based descriptors (other than intensity) suffer significantly less from the absence of image structure.

In Table 4.2 we show classification rates when using attributes to predict a patch’s disturbance level on the synthetic data. We evaluate each disturbance individually, using an SVM trained on 90% of the data, and tested on 10% in 10 random folds. The results show that attributes can reasonably detect the disturbance levels (chance performance for the five disturbance levels is 20%).

### 4.4.2 Aloi Dataset

The Aloi dataset [28] consists of images of 1000 objects under a variety of imaging conditions. These include the illumination direction, illumination color and object viewpoint. The recording setup consists of five light sources, positioned on a hemisphere aslant to the object. Three cameras are positioned next to each other underneath the light sources. Here, we consider images from the two outer cameras that are furthest apart and use eight different light source combinations. Thus, we consider two recordings for every of eight illumination conditions, leading to a total of 16 variations of object appearance.

Patches are extracted in similar spirit to [7]. First, planar homographies between the cameras are computed based on correspondences between sift descriptors extracted from interest points detected by the harris-laplace detector on images from ‘canonical’ illumination condition l8 (all light sources switched on). For this, we use a standard ransac procedure. We impose additional constraints based on camera vicinity, i.e. the transformation should be small and near-translational, regardless of object geometry. Using the obtained homography,
we propagate the feature detections in the canonical image \( I_8 \) to all other images and extract rectangular image patches proportional to the detection scale. The non-planarity of most objects causes unbiased geometric variations in the image patches. The patches are resized to 64x64 pixels and patches outside the range of 64±20 pixels are discarded. The dataset consists of about 200K patches. See Figure 4.1(a) for an example.

The descriptor pool is augmented with Osift and Csift descriptors, as these have been shown to be the single optimal choice for a wide range of matching and recognition tasks [89]. These descriptors follow from the descriptor pool (Table 4.1) by concatenation: Osift= \([O_1, O_2, O_3]\) and Csift= \([O_1/O_3, O_2/O_3, O_3]\). Furthermore, we include combinations of all descriptors by (concat): concatenation of the full descriptor pool and (mul): multiplication of individual distances prior to ranking, both suggested by [27].

Descriptor selection is evaluated in separate settings. Each setting allows isolated analysis of attributes as representations of either object surface or scene-accidental properties. To this end, a query patch is presented in either canonical (can) or distorted (dis) form, where the canonical patch is extracted from the canonical image \( I_8 \). We distinguish between distortions of a geometric (g) or of a photometric nature (p), and both (pg). A pure photometric disturbance has all patches recorded by the same camera and has therefore no geometric differences. A purely geometric disturbance considers patches for different cameras, however only under uniform illumination. Note that for this geometric distortion there is no difference between a disturbed or a canonical query patch. To these distinct evaluation settings we add the (most realistic) setting in which query patches as well as database patches may arrive in either canonical or distorted form (all). See Table 4.3 for an overview. Matching results in terms of mean average precision are presented in Figure 4.4 and Figure 4.5.

**Individual Descriptor Performance.**
The results show that gradient-based descriptors generally perform (much) bet-

<table>
<thead>
<tr>
<th>Disturbance type</th>
<th>Canonical</th>
<th>Distorted</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photometric</td>
<td>can p</td>
<td>dis p</td>
<td>-</td>
</tr>
<tr>
<td>Geometric</td>
<td>can g</td>
<td>can g</td>
<td>-</td>
</tr>
<tr>
<td>Both</td>
<td>can pg</td>
<td>dis pg</td>
<td>all pg</td>
</tr>
</tbody>
</table>

*Table 4.3:* Named Aloi experiments, results in figures Figure 4.4 and Figure 4.5.
4.4 Experiments

(a) can p (b) dis p (c) can g (d) can pg (e) dis pg

Figure 4.4: Aloi matching results individually per descriptor (row 1) and per attribute (row 2). In row 3 we show the descriptor selection performance in comparison with the single best descriptor, all descriptors combinations (mul, concat) by multiplication or concatenation, and the best possible descriptor from the descriptor pool (oracle). Several scenarios are evaluated: the query patch may be in either canonical ('can') or distorted form ('dis'), while the disturbance is either photometric ('p'), geometric ('g'), or both ('pg'). See Table 4.3 for an overview. Pixel-based (pix) and gradient-based (grad) descriptors are extracted from intensity, chromatic, normalized chromatic and hue representations, denoted by I., C., N. and H as given in Table 4.1. For easy reference we mark the score of the single best descriptor with a gray bar.

Figure 4.5: Most realistic results where the query patch is either canonical or distorted, with photometric and/or geometric distortion. See the caption of Figure 4.4 for the explanation of the symbols.
Per-patch Descriptor Selection using Surface and Scene Properties

ter than pixel-based descriptors. However, if the disturbance is purely geometric (can g) pixel-based descriptors perform at their best. This is because calculating image gradients requires larger spatial support than a single pixel. In accordance to other work [89], Osift is often the best performing individual descriptor. When the distortion is purely photometric, a distorted query patch (dis p) gives better performance for most descriptors than an undistorted query patch (can p). When the distortion is both photometric and geometric (can pg and dis pg) the situation is reversed. Moreover, the combined photometric and geometric distortions (can pg, dis pg, all pg) perform significantly worse than their single-distortion counterparts (can p, dis p, can g).

Individual Attribute Performance.
Texture and geometric stability are the best performing attributes, whereas hue rarely helps. Overall, each attribute generally increases performance. This indirectly shows that they are helpful for descriptor selection. Interestingly, the attributes are able to predict which material will be sensitive to a photometric distortion (can p). However, they fail completely to recognize a photometric distortion (dis p) when it is present.

Descriptor Selection Performance.
Our descriptor selection method is always better than the best single descriptor (typically Osift). When comparing descriptor selection to feature combination methods, we perform equal, or better. Concatenating the full descriptor pool produces a very high-dimensional descriptor yet results in poor performance. Descriptor combinations by multiplication often improve over individual descriptors, however, it fails when there is high variance between individual descriptor performance (can p, dis p). The performance gain of descriptor selection is most prominent in the scenario in which patches appear in either canonical or distorted form under mixed disturbances (all pg) in Figure 4.5.

Analysis of Descriptor Selection.
In Figure 4.6 we display the confusion matrix between predicted and best descriptors. The diagonal shows that we outperform random classification performance which for 10 descriptors is 0.1. There is a slight bias towards intensity sift (I.grad), because this is often the best descriptor. In Figure 4.7(a-c) we give the distribution of the predicted and best descriptors. Pixel-based descriptors are frequently superior in both train and test-sets, but hard to select automatically by the classifier. The intensity sift (I.grad) is often the best descriptor, whereas on average Osift is slightly better, as shown in Figure 4.5.
4.4 Experiments

Figure 4.6: Predicted and best descriptors confusion (all pg).

Figure 4.7: Distribution of the best descriptors (all pg) on: (a) Train set (b) Test set (c) selected by our classifier.
4.5 Discussion

This chapter introduces a novel descriptor selection framework. Other methods can select a single feature for a whole image, or optimize a single feature over a dataset of patches. We, in contrast, show that the most appropriate descriptor alternates per patch. Therefore, we propose to select the descriptor on a per-patch basis. The selection method operates on attributes extracted from the image, through which object surface and scene properties are measured. These attributes are indicative for the appropriate descriptor. On a large dataset of colored object patches, the proposed selection method is shown to outperform existing sophisticated image descriptors that claim to be invariant to one or more imaging conditions.
Evaluation of Color Spatio-Temporal Interest Points

5.1 Introduction

Human activities play a central role in video data that is abundantly available in archives and on the internet. Information about the presence of human activities is therefore valuable for video indexing, retrieval and security applications. However, these applications demand recognition systems to operate in unconstrained scenarios. For this reason, research has shifted from recognizing simple human actions under controlled conditions to more complex activities and events ‘in the wild’ [54]. This requires the methods to be robust against disturbing effects of illumination, occlusion, viewpoint, camera motion, compression and frame rates.

High-level approaches for unconstrained human activity recognition aim at modeling image sequences based on the detection of high level concepts [74], and may build on low-level building blocks [88] which typically consider generic video representations based on local photometric features [42, 47, 98]. High-level approaches are based on complex, computationally expensive video processing operations but may be superior to low-level approaches in terms of recognition rates. However, high-level approaches are sensitive to local geometric disturbances such as occlusion, which limits their applicability [74]. Low-level approaches are conceptually simple, relatively easy to implement

*Published in the transactions on Image Processing [23]. Ideas appeared previously in the conference on Computer Vision and Pattern Recognition [22]
and potentially sparse and efficient. Due to the local nature of features on which low-level approaches are based, they are inherently robust against recording disturbances such as occlusion and clutter. Therefore, in this chapter, we focus on low-level representations for recognizing human actions in video.

Low-level action recognition approaches are often based on spatio-temporal interest points (STIPs). Here, image sequences are represented by descriptors that are extracted locally around STIP detections, see Figure 5.1 for example detections. The descriptors are vector quantized based on a visual vocabulary, and subsequent learning and recognition operates on these quantized descriptors, comprising the well known bag-of-(spatio-temporal)-features framework. The formulations of spatio-temporal feature detectors and descriptors available in literature are based on single-channel intensity representations of the video data. Due to the lack of photometric invariance of the intensity channel [89], current approaches are consequently sensitive to disturbing illumination conditions such as shadows and highlights. More importantly, discriminative information is ignored by discarding chromaticity from the representation.

Figure 5.1: Examples of STIP detections in the sequence depicted above. For illustration purposes we have polled the detectors for the 55 strongest STIPs in the original 55-frame sequence, and show the detections on frame 48.
5.1 Introduction

In the spatial (non-temporal) domain, color descriptors outperform intensity descriptors in a variety of image matching and object recognition tasks [9, 89]. The reason for this improved balance between photometric invariance and discriminative power is illustrated in Figure 5.2a by an estimate of the joint distribution of spatial intensity and color partial derivatives, being the image features based on which descriptors are formed. The figure shows that every intensity derivative is associated with a distribution over color derivatives and vice versa. Thus, information is lost when either intensity or chromatic representations are considered in isolation. For effective feature detection and extraction based on multi-channel differential representations in the spatio-temporal domain, it is thus a precondition that similar conclusions hold for the joint distribution of temporal intensity and color derivatives. This is verified by observing Figure 5.2b, in which the joint distribution of temporal color and intensity derivatives is shown to strongly resemble the distribution of spatial derivatives in Figure 5.2a.

In this chapter, we propose to incorporate chromatic representations in the spatio-temporal domain. The aim is to reformulate STIP detection and description for multi-channel video representations. Videos are represented in a variety of color spaces exhibiting different levels of photometric invariance. By this enhanced appearance modeling, we aim to increase the quality (robustness and discriminative power) of STIP detectors and descriptors for recognizing human activities in video. This is validated through a set of repeatability and recognition experiments on challenging video benchmarks. A previous version of this work appeared in [22].

Figure 5.2: Joint distributions of partial intensity and color (hue) derivatives for the spatial (a) and temporal (b) domain. The distributions are estimated from 5M pixels of one sequence of the FeEval dataset [87].
5.2 Related Work

In the spatial domain, multi-channel photometric invariant formulations of feature detectors are reported in e.g. [86, 90, 91]. These articles report increased repeatability, entropy, and object categorization results as compared to intensity-based detections. For descriptors, multi-channel formulations [9, 89] propose various color SIFT variants. Most notably, OpponentSIFT considerably improves the performance. Based on this, we formulate a family of increasingly invariant photometric representations which are incorporated in multi-channel formulations of spatio-temporal feature detectors and descriptors.

5.2.1 Spatio-Temporal Detectors

In the spatio-temporal domain, pioneering work by Laptev [46] extends the Harris function to 3D. Alternatively, the Gabor STIP detector proposed by Dollár et al. [17] applies a Gabor filter along the temporal axis and is not based on differential image structure. The authors [17] argue that differential based STIP detectors are incapable of detecting subtle and periodic motion patterns. Gabor STIPs are therefore essentially different from Harris STIPs and we develop multi-channel formulations for both detectors to study differential as well as raw spatio-temporal image data.

As an alternative to STIP-based sampling, local descriptors may be extracted along motion trajectories [97]. Here, densely sampled points are tracked from frame to frame based on optical flow. As the method involves tracking and dense multi-scale optical flow computation, the associated computational complexity is typically higher than that of STIP-based approaches. Depending on the descriptor(s) that are subsequently extracted, this sampling method may compare favorably in terms of recognition rates. In this chapter, we focus on the sparser STIP-based approach for studying color in the spatio-temporal domain.

Other color STIPs have been proposed earlier in [85]. However the formulation of the multi-channel spatio-temporal structure tensor for the 3D Harris function is somewhat erroneous. Also, the proposed color STIP descriptor is a concatenation of a color histogram, an intensity-based gradient (HOG) and optical flow (HOF) descriptor, which is not shown to produce performance improvements with respect to other existing STIP-based recognition methods. In this chapter, we extend the multi-channel structure tensor of [91] in a principled manner.
5.2 Related Work

to the spatio-temporal domain and investigate various methods to incorporate color gradients in the HOG3D descriptor.

5.2.2 Spatio-Temporal Descriptors

Among the local spatio-temporal descriptors available in literature, the HOG3D descriptor [42] appears well-suited for large scale video representation and multi-channel extensions. In contrast to e.g. HOG/HOF [47], MoSIFT [13] or MBH [97] descriptors, the HOG3D algorithm serves as an integrated and efficient approach, as it excludes optical flow which is computationally expensive [71, 80]. Also, good results in a STIP-based bag-of-features recognition framework using the HOG3D descriptor have been achieved, especially in combination with the Gabor STIP detector [98]. Moreover, motion-based descriptors are shown in [71] to suffer from scalability issues. Therefore, we derive several multi-channel variants of the HOG3D descriptor and evaluate their performance for realistic human action recognition.

Discriminability issues associated to motion descriptors in large scale action recognition are shown in [71] to be addressed by the motion boundary histograms (MBH) of [96]. As opposed to a direct motion description, MBH is based on differential optical flow, which greatly reduces the confusion between action categories. In recent work by Wang et. al. [97], MBH descriptors extracted along motion trajectories and modeled in a multiple kernel learning framework have achieved state-of-the-art results on a large number of datasets.

Another recently proposed video descriptor for human action recognition is Gist3D [84]. This is a global descriptor based on a 3D filter bank and describes the spatio-temporal ‘gist’ of a video. Reasonable recognition performance is achieved in combination with STIPs.

The works mentioned comprise low/medium level approaches to action recognition. Higher level approaches such as Action Bank by Sadanand et al. [74] give good results on some datasets. However, such high-level approaches are typically not scalable. In contrast, low-level approaches are widely applicable, conceptual simple, sparse and exhibit reasonable computational complexity. Moreover, they may serve as powerful building blocks for higher level methods [88]. We contribute by considering a variety of photometric representations for STIP detection and description for enhancing low-level approaches to action recognition.
5.3 Photometric Representations

We model the formation of images by the dichromatic reflection model [78],

\[ f = e(m^b c^b + m^i c^i), \]  \hspace{1cm} (5.1)

where \( f = (R, G, B)^T \) is the sum of the body reflectance color \( c^b \) with the interface reflection color \( c^i \). The contributions of these reflectance colors are weighted by their respective magnitudes \( m^b \) and \( m^i \), that depend on the surface orientation and illumination direction. Additionally, the specular reflection \( m^i \) is viewpoint dependent. The intensity of the light source is represented by \( e \).

Invariance against highlights (shifts in the signal) can be achieved by representations that cancel out the additive interface reflection term \( m^i c^i \). Signal scalings, such as those caused by shadows and shading, are ignored by dividing out the light source intensity \( e \). Here, we consider the transformation of the RGB image to the opponent color space [9, 21, 89, 90]

\[
\begin{bmatrix}
O_1 \\
O_2 \\
O_3
\end{bmatrix} =
\begin{bmatrix}
R - G \\
R + G - 2B \\
R + G + B
\end{bmatrix}.
\]  \hspace{1cm} (5.2)

The transformation approximately decorrelates the image channels, resulting in intensity \( O_3 \) and chromatic components \( O_1, O_2 \). Based on these formulations, several photometric properties can be derived.

**Highlights.** Due to subtraction of RGB components in Eq. (5.2), the reflection term from Eq. (5.1) is subtracted in the formulations of \( O_1, O_2 \). Hence, the chromatic opponent components are invariant to signal shifts such as those caused by (white) highlights.

**Shadow-shading.** The chromatic components are normalized by intensity \( O_3 \), canceling out the light source intensity term from Eq. (5.1). This yields the shadow and shading invariants \( \frac{O_1}{O_3}, \frac{O_2}{O_3} \).

**Shadow-shading-highlights.** Invariance against both scalings and shifts in the signal is achieved by considering the ratio of chromatic components: \( \frac{O_1}{O_2} \). This results in the shadow-shading-highlight invariant hue representation.

We refer to these photometric image representations as \( I \)(intensity), \( C \)(chromatic), \( N \)(normalized chromatic) and \( H \)(hue). These can be ordered with respect to their invariance level: \( H \succ N \succ C \succ I \). The intensity \( I \) preserves most image structures, which is the most discriminative representation. Therefore the intensity-
5.4 Multi-Channel STIP Detection

Table 5.1: Photometric image representations. Chromatic combinations with the intensity channel yield IC, IN and IH.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Intensity</th>
<th>Chromatic</th>
<th>N-Chromatic</th>
<th>Hue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invariant to</td>
<td>-</td>
<td>Highlights</td>
<td>Shadows</td>
<td>Hl. &amp; Sh.</td>
</tr>
<tr>
<td>Reference</td>
<td>I</td>
<td>C</td>
<td>N</td>
<td>H</td>
</tr>
</tbody>
</table>

normalized representations N and H have a higher level of photometric invariance than C, in which the light source intensity is preserved. We summarize the representations and their properties in Table 5.1.

The lack of discriminative power associated with the chromatic representations C, N and H typically renders them unsuitable for matching and recognition tasks. Combinations of intensity and chromatic channels result in IC, IN and IH. Note that the three-channel representation IC comprises the original opponent channels \([O_1, O_2, O_3]\). These representations are established first, i.e., prior to any subsequent processing. All channels are min-max normalized using the theoretical extremal values per channel based on the transformations in Eq. (5.2) and Table 5.1 so as to weight them equally a-priori.

5.4 Multi-Channel STIP Detection

Multi-channel Harris STIPs. Harris STIPs are local maxima of the 3D Harris energy function based on the structure tensor \([46]\). A multi-channel formulation of the structure tensor has been developed in e.g. \([91]\) which prevents opposing color gradient directions to cancel each other out. Here, we incorporate multiple channels in the spatio-temporal structure tensor \([46]\).

The multi-channel volume V consisting of \(n_c\) channels is denoted by \(V = (V^1, V^2, ..., V^{n_c})^T\). The individual channels are represented in scale space \(V^j = g(\cdot; \sigma_o, \tau_o) * f^j(\cdot)\), where \(g(\cdot; \cdot, \cdot)\) is the 3D Gaussian kernel with equal scales along the spatial dimensions, \(\sigma_o\) and \(\tau_o\) are the spatial and temporal observation scales and \(f^j : \mathbb{R}^3 \rightarrow \mathbb{R}\) is the imaging function of channel \(j\). The multi-channel spatio-temporal structure tensor is then defined by

\[
S = g(\cdot; \sigma_t, \tau_t) * \begin{pmatrix}
V_x \cdot V_x & V_x \cdot V_y & V_x \cdot V_t \\
V_y \cdot V_x & V_y \cdot V_y & V_y \cdot V_t \\
V_t \cdot V_x & V_t \cdot V_y & V_t \cdot V_t
\end{pmatrix},
\] (5.3)
where \( \sigma_i \) and \( \tau_i \) denote the spatial and temporal integration scale respectively. In Figure 5.3 we illustrate the response per representation. Incorporating increasingly invariant photometric representations clearly has an effect on the Harris energy. The highlight on the shiny heart-shaped object surface part triggers a strong response for the original \( I \)-based energy functions. This effect is clearly dampened in the \( C \) representation. However, the reflected illumination by the colored matte-shiny (left) object part still triggers a response, as the nature of the local object surface causes signal changes that are not captured by a simple shift. Intensity normalization of the chromatic components (\( N \)) then causes this response to be dampened, while emphasizing colorful transitions on the object surface. Finally, the scaling- and shift-invariant \( H \) representation eliminates essentially all responses except for salient color transitions.

### Multi-channel Gabor STIPs

The Gabor STIP detector is based on a Gabor filtering procedure along the temporal axis [17]. Invoking multiple channels is straightforward because the energy function is positive by formulation. Hence, no additional care has to be taken to account for conflicting response signs between channels

\[
R = \sum_{j=1}^{nc} (g(\cdot; \sigma_o) * h_{ev}(\cdot; \tau_o) * V^j)^2 + (g(\cdot; \sigma_o) * h_{od}(\cdot; \tau_o) * V^j)^2.
\]

(5.4)

Here, the 2D Gaussian smoothing kernel \( g(\cdot; \cdot) \) is applied spatially, whereas the
5.5 Multi-Channel STIP Description

Gabor filter pair \( \{ h_{ev}(\cdot; \cdot), h_{od}(\cdot; \cdot) \} \) measures the periodicity of the observed signal along the temporal dimension. As illustrated in Figure 5.3, the I-Gabor energy is mainly clustered around an incidental highlight, whereas the response-triggering local photometric events become increasingly rare and colorful along with the level of photometric invariance level of the representation.

5.5 Multi-Channel STIP Description

The HOG3D descriptor [42] is formulated as a discretized approximation of the full range of continuous directions of the 3D gradient in the video volume. That is, the unit sphere centered at the gradient location is approximated by a regular \( n \)-sided polyhedron with congruent faces. Tracing the gradient vector along its direction up to intersection with any of the polyhedron faces identifies the dominant quantized direction. Quantization proceeds by projecting the gradient vector on the axes running through the gradient location and the face centers with a matrix multiplication of the 3D gradient vector \( g \),

\[
q = (q_1, ..., q_n)^T = \frac{P \cdot g}{||g||_2}, \tag{5.5}
\]

where \( P \) is the \( n \times 3 \) matrix holding the face center locations and \( q \) is the projection result (i.e. the histogram of 3D gradient directions). Note that the contribution is distributed among nearby polyhedron faces. Descriptor dimensionality may be reduced by allocating opposing gradient directions to the same orientation bin. The descriptor algorithm proceeds by centering a cuboid at the STIP location, which is tessellated into a spatio-temporal grid. Histograms are computed for every grid cell and concatenated to form the final descriptor [42].

Chromaticity is incorporated in the HOG3D descriptor by considering the representations from Section 5.3 in a multi-channel formulation of the gradient vector \( g \) in Eq. (5.5). We follow the standard practice of concatenation of the per-channel descriptors [9, 89, 21]:

\[
g' = \{ g'_j \}, j = 1, ..., n_c. \tag{5.6}
\]

We also compute a single gradient variant where we prevent the effect of opposing color gradient directions by using tensor formulations. In tensors, opposing directions reinforce each other by summing the gradient orientations as opposed to their directions [91],
Table 5.2: Multi-channel HOG3D variants. \( C \) denotes some photometric representation comprising \( n_c \) channels. The dimensionality of an integrated direction-based descriptor is considered default (1D, which is 360 here), based on which we derive the dimensionality of the other descriptor variants. Variants of \( C \) are denoted by subscript flags, indicating channel combination (integration/concatenation) and gradient quantization (orientation/direction).

<table>
<thead>
<tr>
<th>Channel Integration</th>
<th>Gradient Orientation</th>
<th>( \varepsilon_{1,1} : D/2 )</th>
<th>( \varepsilon_{1,0} : 1D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel Concatenation</td>
<td>Gradient Direction</td>
<td>( \varepsilon_{0,1} : n_c D/2 )</td>
<td>( \varepsilon_{0,0} : n_c D )</td>
</tr>
</tbody>
</table>

\[
g'' = \sum_{j=1}^{n_c} g^j \cdot g^j. \tag{5.7}\]

This formulation of the gradient defines half of the full sphere of directions which is one of the HOG3D flavors in [42]. Here, it naturally follows from a tensor formulation of the multi-channel 3D gradient.

We formulate another variation as the summation of per-channel full direction descriptors. Together with the tensor-based approach, we call this descriptor **integration** as opposed to **concatenation**. The variant benefits from the expressiveness associated with the full set of multi-channel directions while maintaining the same dimensionality as a single channel descriptor. Note that the differences between integration and concatenation of channels do not apply to single-channel descriptors. The descriptor variants and their associated dimensionalities are summarized in Table 5.2.

### 5.6 Experiments

We evaluate the multi-channel STIP detectors and descriptors through a set of repeatability and action recognition datasets.

#### 5.6.1 Implementation Details and Notation

We base our implementation of STIP detectors on the activity recognition toolbox by Dollàr et al. [17] while re-implementing the HOG3D descriptor of Kläser et al. [42].
5.6 Experiments

STIP scale: For the Gabor detector, we set the spatial scale $\sigma_o = 2$ and the temporal scale $\tau_o = \sqrt{8}$ in Eq. (5.4). Note that this setting for $\tau_o$ is in conflict with e.g. [98], but we have found that the proposed default setting of $\tau_o = 4$ is too large for descriptor extraction in short sequences. For the Harris detector, we consider a reduced set of spatial scales with respect to prior work, as we have found this to be satisfactory in terms of discriminative power and computational load. Specifically, for computing the Harris energy based on Eq. (5.3), we consider $\sigma_o = \sqrt{2^i}, i \in \{2, 3, 4\}$ and $\tau_o = \sqrt{2^j}, j \in \{1, 2\}$. As in e.g. [98, 47], we do not perform STIP scale selection because of its high computational costs and decreased recognition performance [46].

Cuboids: Descriptors are extracted from cuboids centered at STIP locations. The spatio-temporal extent as well as the grid layout of these cuboids may be discriminatively optimized such as in [42]. In this chapter, we refrain from such an optimization scheme in order to maintain focus on the integration of chromatic channels. Instead, we consider one particular setting (from e.g. [98]) in which the extent of a cuboid is defined as $\Delta_x = \Delta_y = 18\sigma_o$ and $\Delta_t = 8\tau_o$. For feature aggregation, we employ a 3x3x2 spatio-temporal pooling scheme. This grid layout is attractive due its compactness, whereas we have not found significant dependencies of our results on these settings for our purpose.

Descriptors: We consider the four variants of the multi-channel HOG3D descriptor as summarized in Table 5.2. The variants are denoted by flagging the descriptor names. The first flag denotes whether the descriptor channels are integrated (or otherwise concatenated), whereas the second flag denotes the usage of gradient orientations (as opposed to directions). For example, $IC_{0,1}$ denotes the concatenated orientation-based Opponent-HOG3D descriptor. Integrated, orientation-based descriptors such as $IN_{1,1}$ follow from the tensor-based approach in Eq. (5.7). There is no difference between $I_{0, \cdot}$ and $I_{1, \cdot}$ as $I$ comprises a single channel.

We use integral video histograms for aggregating features over grid cells. We refrain from gradient approximation based on integral video representations of the partial derivatives as in [42], because this affects the information that we wish to study. For descriptor normalization, we adopt the method proposed by Brown et al. [7] in which the normalization cut-off threshold is a discriminatively optimized function of the descriptor dimensionality. By this, we discard the time consuming task of determining the optimal normalization parameters per descriptor variant.

In summary, apart from the photometric representations, our HOG3D implementation differs slightly from the original version [42] by 1) exact gradient computation, 2) descriptor normalization and 3) spatio-temporal pooling.
Recognition. Based on the multi-channel STIP detectors and descriptors, we perform action recognition in a standard bag-of-features learning framework. Unless stated otherwise, we closely follow the setup of [98]. Here, codebooks are created by clustering 200K randomly sampled HOG3D descriptors using k-means in 4000 clusters. A sequence is then represented by quantizing the extracted HOG3D descriptors based on the learned codebook. An SVM is trained based on the $\chi^2$ distance between codebook descriptors. Evaluation of the learned classifier is usually performed in a leave-n-out cross validation setup. Every experiment is repeated three times for different codebooks, which produces typical standard deviations between 0.2 and 1 percentage point (depending on dataset size and the number of STIP detections).

5.6.2 Datasets

We measure STIP repeatability and descriptor entropy for videos taken from the FeEval dataset [87]. This dataset consists of 30 videos taken from television series, movies and lab recordings where each video is artificially distorted by applying different types of photometric and geometric transformations. Every transformation type is associated to a challenge, in which the distortion is applied in increasingly severe steps. We consider the videos from the television series up to the first occurring shot boundary. That is, we do not aim at studying STIP behavior in controlled settings, cartoons or in typical movie settings for which editing effects are frequent. We consider the full set of challenges: blur, compression, darken, lighten, median filter, noise, sampling rate and scaling and rotation. Some examples are shown in Figure 5.4.

Figure 5.4: Examples from FeEval dataset. From left to right: original, noise, darken.

For an in-depth evaluation of detector and descriptor settings we use the UCF sports dataset [72]. The dataset exhibits 10 sports action categories in 150 videos, all of which are horizontally flipped to increase the dataset size. Performance is evaluated in a leave-one-out cross validation scheme, in which the flipped version of the considered test video is removed from the training set. The best performing experimental settings are applied to the UCF11 [54] and UCF50 [71]
5.6 Experiments

Figure 5.5: Examples from UCF sports, UCF11 and UCF50 datasets (images are cropped).

datasets. The datasets contain 11 and 50 human action classes in about 1200 and 6700 videos respectively; UCF50 is a superset of UCF11. These challenging datasets comprise youtube videos exhibiting real human activities. Here, performance is evaluated through a leave-one-group-out cross validation scheme over 25 groups, in which we exactly follow the authors’ guidelines†. See Figure 5.5 for some examples of the datasets.

5.6.3 STIP Repeatability

We poll the detectors for an average number of 10 STIPs per frame of the FeE-val videos. A repeatability score is obtained by considering the detections in the challenge sequence, and computing the relative overlap of the cuboid around the detected STIP location with the corresponding location in the original sequence. We take the spatio-temporal extent of the cuboid to be equal to the observation scale. The repeatability scores averaged over all sequences and challenges are presented in Table 5.3.

Table 5.3: STIP repeatability for multi-channel Harris and Gabor detectors based on the considered photometric representations.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>IC</th>
<th>IN</th>
<th>IH</th>
<th>C</th>
<th>N</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris</td>
<td>61.3%</td>
<td>61.6%</td>
<td>61.3%</td>
<td>37.0%</td>
<td>45.6%</td>
<td>40.5%</td>
<td>28.7%</td>
</tr>
<tr>
<td>Gabor</td>
<td>43.6%</td>
<td>43.6%</td>
<td>43.6%</td>
<td>24.4%</td>
<td>25.4%</td>
<td>22.9%</td>
<td>19.3%</td>
</tr>
</tbody>
</table>

Harris STIPs are more stable than Gabor STIPs. Nonlinear differential spatio-temporal signal changes are more distinctive than temporal fluctuations only. As the representation becomes increasingly invariant, repeatability progressively decreases. Also, combining the invariants with intensity does not increase repeatability with respect to using intensity only (marginal improvements for the

†http://crcv.ucf.edu/data/UCF50.php
Moreover, the IH representation attains lower repeatability scores than I. The reason for these reduces scores is that, as disturbing conditions are effectively ignored, so are spatio-temporal image structures on which stable STIPs are detected. Adding C or N to the intensity I basically leaves the repeatability unaltered for this dataset. However, the STIP discriminability experiments will show different recognition scores for these representations.

From here on, the pure chromatic representations are discarded from the experimental batch due to the associated lack of discriminative power.

**Figure 5.6:** Entropy of descriptor variants extracted around STIPs from several detector variants. Multi-channel descriptors are associated to higher entropies than their single-channel counterparts. This holds for both integration and concatenation of channels. The figure looks similar for Harris and Gabor STIPs.

### 5.6.4 Descriptor Entropy

Here, we study the amount of information contained in each of the considered descriptors. For this, we extract unnormalized descriptors from the cuboids around STIP detections in the set of undistorted FeEval videos. The descriptors $D_i$ are then $L_1$-normalized to allow for the computation of entropy:

$$\text{entropy}(D_i) = -\sum_{j=1}^{\lvert D_i \rvert} D_i^j \log_2(D_i^j). \quad (5.8)$$

The above is illustrated in Figure 5.6 for Gabor STIPs. Entropies are averaged over all descriptors and sequences. The figure is essentially similar for descriptors extracted around Harris STIPs.

Standing out from the figure is the high entropy associated to the $IC_{0,0}$ descriptor (i.e. concatenated direction-based Opponent-HOG3D). This is partly explained by its high dimensionality due to concatenation. Note however the increased entropy with respect to $IN_{0,0}$ which has the same dimensionality. In that respect it also stands out that the entropy associated to the 2-channel descriptor $IH_{0,0}$ is higher than that of the 3-channel descriptor $IN_{0,0}$. We conclude
from this that the chromatic ratio constituting $H$ exhibits more (differential) variation than the intensity-normalized channels in $N$, whereas most variance is associated with $C$.

The single-channel descriptor $I_{0,0}$ is associated with a considerable lower entropy than its multi-channel counterparts. These differences are dampened when the channels are integrated instead of concatenated, by which the multi-channel dimensionality is equalized to that of a single channel. However, the integrated descriptors $IC_{1,0}$ and $IH_{1,0}$ are still clearly associated to higher entropies, whereas the difference between $IN_{1,0}$ and $I_{0,0}$ is marginal.

Orientation-based descriptors exhibit lower entropies than direction-based descriptors. This follows from their definition: two opposing gradient directions are indistinguishable in terms of their orientation. Observations regarding photometric representations and channel integration with respect to the direction-based descriptors also hold for orientation-based descriptors.

With respect to varying photometric representations in the detector, we observe a considerable drop in entropy for the $IH$ detector as compared to the other representations. This is explained by the fact that $H$ causes the detector to fire on signal fluctuations that do not necessarily correspond to strong structures in the intensity profile. There appears no substantial differences between the other representations, although slightly higher entropies are attained for $IC$ detections.

### 5.6.5 Color STIP Detector Discriminability

For evaluating action recognition performance on the UCF sports dataset, we consider the photometric variants of both the Harris and Gabor detectors. Direction-based intensity HOG3D ($I_{0,0}$) descriptors are extracted around multi-channel STIP detections, so as to separate the analyses regarding STIP detection and description. Recognition accuracy is computed for an average of $\{10, 20, 30, 40, 50\}$ STIPs per frame by varying the detection threshold. Results are given in Figure 5.7a,b.

We first validate our implementation by comparing recognition accuracies with the evaluation reported on intensity in [98]. Here, the average number of Harris STIPs is 33, for which an accuracy of 79.9% is attained. We obtain 80.4% for 30 STIPs per frame. As for the Gabor detector, [98] reports an accuracy of 82.9% for 44 STIPs. This is comparable to our performance of 83.4% for 40 STIPs.
Figure 5.7: Recognition performance on the UCF sports dataset per photometric representation for varying amounts of Harris (a) and Gabor (b) STIPs. Influence of the photometric representations on descriptor variants (c). Combinations of the top-performing $IC$-Gabor STIPs and $IC_{1,0}$-descriptors for varying codebook sizes (d).

Color STIPs

It is shown in Figure 5.7a,b that discriminative power is severely hampered by integrating $H$ in the energy functions. This is expected because $H$ is associated to the highest level of photometric invariance. As more detections are requested, however, performance converges to that of $I$-STIPs. Considering Harris STIPs in Figure 5.7a, integrating the $C$ and $N$ representations leads to marginal performance differences compared to $I$. For small to moderate amounts of STIPs, recognition accuracy is somewhat improved, in particular for $IN$. The primary characterization of Harris STIPs in terms of distinctiveness and sparsity is mainly due to nonlinear fluctuations in the spatio-temporal intensity signal. Adding chromatic components to the formulation of the energy function does not drastically alter this characterization.

Regarding the multi-channel Gabor detector in Figure 5.7b), discriminative STIPs are detected for the $C$ and especially $N$ channels as compared to using $I$ alone. While $I$ by itself contains the most important information regarding spatio-temporal signal fluctuations, invariants may prevent the detector to fire on disturbing factors such as highlights and shadows. Also, we assume the specific colorfulness of local spatio-temporal events associated to certain actions to be informative (e.g. ‘Diving’ (skin color, blue water) and ‘Riding-Horse’ (brown horse, green field and trees)).

Discussion on sparsity, distinctiveness and scale

Harris STIPs are more discriminative than Gabor STIPs for a relatively small number of detections. This relative performance difference reverses as more STIPs are considered. The reason for this is related to sparsity, distinctiveness
As can be derived from Figure 5.3, the Harris function is sparser than the Gabor energy. The Harris function fires only on relatively rare events - nonlinear signal changes in both space and time - which are also distinctive in scale space, and are usually caused by human activity rather than background and/or camera motion. As a consequence, Harris STIPs are highly discriminative, but very sparse: there resides a large and indifferent gap between the thresholds of a good quality Harris STIP detector and a noise detector. Opposed to this, the Gabor detector is more generic and covers the image sequences more densely. This results in improved recognition results as more STIPs are requested, whereas the performance of the Harris detector as a function of the number of STIPs quickly plateaus and even degrades.

Whereas the Harris function is typically computed over multiple scales, the Gabor detector (as originally proposed) operates at a single scale. In fact, we have found in the recognition experiments in which we poll the detectors for a fixed number of interest points, that a multi-scale Gabor implementation seriously hampers the recognition performance (results not shown). The reason for this is that the across-scale Gabor responses are highly correlated. This results in overly redundant overlapping detections for local volumes exhibiting strong periodic signal fluctuations, whereas other discriminative local volumes may not be detected at all. Applying the Gabor filters at a single scale only is therefore not so much a choice of design; it is rather instrumental to the method. These arguments do not apply to the Harris detector due to its associated sparsity, i.e. single scale Harris STIPs are insufficient for effective recognition. The unnecessity of multi-scale processing grants a large advantage to the Gabor detector over the Harris detector in terms of computational efficiency. The experimental summary over all datasets in Table 5.6 shows the effectiveness of the Gabor detector.

### 5.6.6 Color STIP Descriptor Discriminability

For the action recognition experiments on the UCF sports dataset, descriptors are extracted around Gabor STIPs as these have shown superior recognition performance over Harris STIPs in Figure 5.7a,b. The detector representation is fixed to $I$. We adopt the detection threshold that yields 50 STIPs per frame on average. Recognition accuracies are reported in Figure 5.7c.

General conclusions about photometric invariance relate to the discriminative power of the descriptors. That is, the $IC$-based descriptors typically outperform
IN descriptors, which in turn are favored over IH. Multi-channel descriptors usually outperform the I-based descriptor. We observe a general preference for direction-based descriptors over orientation-based descriptors (Table 5.2). This is due to the associated wider range of expressiveness. Most apparent in this respect is the IC representation, i.e. IC0,0 improves over IC0,1 by almost 4 percentage points, whereas IC1,0 attains 2 percentage points more than IC1,1. Thus, every channel exhibits discriminative power in the full range of gradient directions. It may even be the case that the (implicit) preservation of opposing gradient directions between channels is informative. Furthermore, IC-based descriptors favor channel integration over concatenation, which is not the case for IN- and IH-based descriptors. In fact, one would expect concatenation-based descriptors to perform better in general due to the enhanced expressiveness associated with multiple channels and increased dimensionality. This is also the most widely adopted approach to multi-channel descriptors, e.g. [9, 89, 21]. However, we obtain the positive side-effect of increased recognition performance against reduced descriptor dimensionality. That is, the multi-channel descriptor dimensionality remains equal to that of a single channel. Although the difference with IC0,0 is marginal, we report a top performance of 85.6% for IC0,0 against 1) our I,0 baseline of 83.4% and 2) 82.9% reported in [98]. A summary over all datasets in Table 5.6 illustrated the power of IC.

We conduct a final experiment on the codebook size. We consider ‘Opponent STIP’ combinations of I and IC Gabor STIPs with I,0 and IC1,0 HOG3D descriptors. We drop the orientation-based descriptors for now. Recognition results for varying codebook sizes are depicted in Figure 5.7d. We observe that the I-IC (detector-descriptor) combination performs best up to a codebook size of 4000. Top performance is marginally improved to 85.7% by the IC-IC combination for a codebook size of 8000. The computational load associated to such a vocabulary is not worth the effort, considering the performance of 85.5% attained by the I-IC combination for a much smaller codebook size of 1000. We have not observed a relationship between descriptor dimensionality and codebook size.

In contrast to these low/medium level action recognition approaches, the high level Action Bank approach of [74] reaches an accuracy of 95% on UCF sports. Here, we focus on low-level approaches, and our best performance for 50 STIPs per frame is on par with the performance of 85.6% for densely sampled I-HOG3D descriptors in [98], which on average yields over 600 descriptors per frame. Based on a combination of HOG, HOF and MBH descriptors extracted along dense motion trajectories, a performance of 88% is achieved in [97]. Compared to this, our STIP-based approach does a good job considering that it outperforms all reported individual features on UCF sports.
5.6 Experiments

5.6.7 UCF11

Based on the in-depth evaluations on UCF sports, we select the \(I\), \(IC\) and \(IN\) representations for both STIP detection and description for evaluation on the UCF11 and UCF50 datasets. Results are presented in Table 5.4 and summarised in Table 5.6.

Table 5.4: Color STIP action recognition results on UCF11 and UCF50 datasets. The first 5 columns show results for direction-based descriptors, whereas results for orientation-based descriptors are shown in the remaining columns.

<table>
<thead>
<tr>
<th>Det./Desc.</th>
<th>(I) - Gabor</th>
<th>(IC_{1,0})</th>
<th>(IC_{0,0})</th>
<th>(IN_{1,0})</th>
<th>(IN_{0,0})</th>
<th>(I_{1})</th>
<th>(IC_{1,1})</th>
<th>(IC_{0,1})</th>
<th>(IN_{1,1})</th>
<th>(IN_{0,1})</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF11</td>
<td>73.8% 77.5% 78.2% 76.0% 76.4% 71.6% 75.8% 74.2% 73.8% 74.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCF50</td>
<td>68.3% 71.7% 70.9% 71.2% 72.1% 68.8% 72.6% 69.7% 71.8% 72.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Differences between performance in the detectors are again small, but we observe a consistent top-performing combination of \(IN\)-Gabor STIPs with \(IC\)-based HOG3D. Thus, we conclude that a certain amount of invariance against local photometric events is beneficial for STIP detection, whereas the descriptor should be extracted from the most discriminative representation.

We achieve a baseline result of 73.8% on the UCF11 dataset for the intensity-based STIP variant. Adding chromaticity increases the recognition accuracies substantially. Also here, best performance is achieved by the direction-based \(IC\) descriptors: 78.4% for \(IC_{1,0}\) on \(IC\)-Gabor STIPs and 78.6% for \(IC_{0,0}\) on \(IN\)-Gabor STIPs. The representation of the detector appears to be more influential on this dataset, although its contribution is marginal on average.

The results compare favourably to the trajectory-based harvesting of HOG and HOF features in [97], for which 72.6% and 70% is achieved respectively. However, they report a superior performance of 84.1% for their motion boundary histograms.

Discussion on inter-class confusion

For a detailed analysis of the results on UCF11 we have included a confusion-difference matrix in Figure 5.8. The usage of color causes most performance
Figure 5.8: Confusion difference matrix between UCF11 categories. Depicted is the element-wise difference between the confusion matrices of (best performing) color and intensity STIPs.

The confusion difference matrix highlights the gain for the category ‘basketball’. Corresponding videos in the dataset exhibit mostly practicing individuals, whereas considerable variations are observed in other facets such as indoor/outdoor, solid/shaking camera work and clothing. These observations are supportive for the argument that multi-channel processing is useful for feature extraction in general, irrespective of the actual color itself. In addition to this, category-specific motion patterns are more accurately described by using color. For example, a basketball generally has the same orange color, which makes the description of its associated motion (bouncing) more accurate. Furthermore, the usage of color decreases the confusion between ‘basketball’ and ‘horse riding’, and especially ‘tennis swing’. The initial confusion (i.e. based on intensity-STIPs) between ‘basketball’ and ‘tennis swing’ is comprehensible, as most videos of both categories exhibit, in general, an individual performing the activities in isolation. Specific information associated to e.g. the colors of the basketball and tennis courts alleviate much of the confusion. The same line of reasoning applies to the confusion between ‘tennis swing’ and ‘golf swing’, and to a lesser extent ‘basketball’ and ‘volleyball spiking’, as the associated videos exhibit a single, sudden burst of activity performed by an individual. Less evident is the reason for resolved confusion between ‘basketball’
and ‘horse riding’. Videos associated to the latter exhibit a walking or galloping horse, which is characterized by a periodic motion pattern resembling that of a person shooting a basketball. It is probably the case that a bouncing basketball also renders similar motion patterns, while its color then provides the power to discriminate. Opposed to this, it stands out that color STIPs increase the confusion between ‘tennis swing’ and ‘soccer juggling’. This is mainly due to the fact that in one ‘soccer juggling’ video group the activity is performed on a typical tennis hardcourt, which renders similar patterns in all color channels.

### 5.6.8 UCF50

Considering the results on UCF50 in Table 5.4, we observe that best performance is achieved with orientation-based descriptors, as opposed to the direction-based descriptors that are favoured for UCF sports and UCF11. As the number of categories increases, descriptor robustness becomes more important. We observe a baseline result of 68.8% for $I_{1,1}$. This is substantially higher than the results reported in [74] for Action Bank (57.9%) and Harris STIP + HOG/HOF (47.9%) (see Table 5.5) for an overview of recent results on UCF50). We conclude that the Action Bank method is not scalable and suffers from increased geometric variations. As for Harris STIP + HOG/HOF, we conclude that the

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[97]</td>
<td>Trajectory(All)</td>
<td>84.5%</td>
</tr>
<tr>
<td></td>
<td>Trajectory(MBH)</td>
<td>82.2%</td>
</tr>
<tr>
<td></td>
<td>Trajectory(HOF)</td>
<td>68.2%</td>
</tr>
<tr>
<td></td>
<td>Trajectory(HOG)</td>
<td>68.0%</td>
</tr>
<tr>
<td>[80]</td>
<td>Dense(All)</td>
<td>83.3%</td>
</tr>
<tr>
<td></td>
<td>Dense(MBH)</td>
<td>80.1%</td>
</tr>
<tr>
<td></td>
<td>Dense(HOG3D)</td>
<td>72.4%</td>
</tr>
<tr>
<td></td>
<td>Dense(HOF)</td>
<td>69.7%</td>
</tr>
<tr>
<td></td>
<td>Dense(HOG)</td>
<td>58.6%</td>
</tr>
<tr>
<td>[71]</td>
<td>Scene context + STIP(MBH)</td>
<td>76.9%</td>
</tr>
<tr>
<td></td>
<td>Scene Context</td>
<td>47.6%</td>
</tr>
<tr>
<td></td>
<td>STIP(MBH)</td>
<td>71.9%</td>
</tr>
<tr>
<td>[84]</td>
<td>Gist3D + STIP(HOG/HOF)</td>
<td>73.7%</td>
</tr>
<tr>
<td></td>
<td>Gist3D</td>
<td>65.3%</td>
</tr>
<tr>
<td></td>
<td>STIP(HOG/HOF)</td>
<td>54.3%</td>
</tr>
<tr>
<td>[74]</td>
<td>Action Bank</td>
<td>57.9%</td>
</tr>
<tr>
<td></td>
<td>STIP(HOG/HOF)</td>
<td>47.9%</td>
</tr>
<tr>
<td>Here</td>
<td>Color STIP(HOG3D)</td>
<td>72.9%</td>
</tr>
</tbody>
</table>
high degree of distinctiveness of spatio-temporal corners limits generalization capacity for these descriptors. A performance of 76.9% is reported in [71] for a combination of scene context and spatio-temporal descriptors. Here, the best performing spatio-temporal descriptor is MBH on Harris STIPs, which achieves 71.9%. This shows the generalization capacity of differential optical flow descriptors, as well as the capacity of MBH to differentiate between video content around Harris STIPs, as opposed to HOG and HOF descriptors. It should however be noted here that MBH performance comprises a complex multiple kernel combination of a horizontal MBHx and vertical MBHy component. In [84], a recognition accuracy of 73.7% is reported for a combination of Gist3D and Harris STIP + HOG/HOF descriptors. However, performance of the individual descriptors is at most 65.3%. In the recent work of Wang et. al. [97], trajectory-based HOG, HOF and MBH attain 68%, 68.2% and 82.2% respectively, while a multiple kernel combination yields state of the art performance of 84.5%. Finally, in [80] a result of 72.4% is obtained based on dense random sampling of HOG3D descriptors, whereas 83.3% is achieved with a multiple kernel combination of HOG, HOF, HOG3D and MBH descriptors.

We report a top performance of 72.9% for $IC_{1,1}$-HOG3D extracted around $IN$-Gabor STIPs. This result constitutes the best performing STIP-based approach to action recognition, while state of the art results are achieved by trajectory-based harvesting or dense sampling of MBH descriptors and multiple kernel modeling thereof.

**Figure 5.9:** Example frames from UCF50 dataset. The top row contains samples from the categories for which recognition performance based on color STIPs has improved the most over intensity STIPs. The bottom row shows examples from the 6 categories for which recognition performance has decreased. The samples are sorted from left to right based on the difference in recognition rates.
Figure 5.10: Per-class recognition performances on UCF50 dataset. Color-STIP ($IN$-Gabor+$IC_{1,1}$) performance is depicted in red, intensity-STIP ($I$-Gabor+$I_{,1}$) in black and their difference in yellow.
Discussion on per-class results

The results on UCF50 are further analyzed based on the per-category results in Figure 5.10. The recognition performance for 44 out of 50 action categories is improved by using color. The largest improvement is observed for ‘BenchPress’. The main reason for this is that the barbell weights are often (red) colored and thus render discriminative periodic motion patterns, see Figure 5.9 for examples. Another influential factor is the associated typical indoor setting (gym), which often consists of solidly colored walls contrasting with the motion patterns in the foreground. Apart from that, we observe a large variety in terms of, for example, the specific background color or the clothing of the actors. Another action category with large recognition improvement is ‘TaiChi’. We observe from corresponding examples that the activity is often performed outdoors on green grass by individuals wearing colorful clothes. Furthermore, it turns out that two ‘TaiChi’ video groups are composed of the same person performing the activity in the same pink clothes, which provides an obvious advantage to color based methods. A similar line of reasoning applies to the decreased recognition performance of ‘PlayingTabla’ activity, as one of the video groups contains grayscale samples only (in which all \(RGB\) channels are consequently identical). The subtraction of \(RGB\) channels in the transformation to ‘chromatic’ opponent space in Eq. (5.2) then yields NULL channels. However, it is also possibly the case that the cast shadows of the fingers on the tabla exhibit discriminative motion patterns which may be better detected by an unnormalized (intensity-only) STIP detector. Another category for which intensity STIPs perform better is ‘JumpingJack’. Also here, there is one video group containing essentially black/white footage which influences the results. We conclude from these observations that the usage of color for action recognition provides a performance boost in general, while the extremal result cases exhibit rather trivial characteristics.

5.6.9 Discussion on Entropy and Discriminative Power

Consider the descriptor with the highest entropy: \(IC_{0,0}-\text{HOG3D}\). This is the best performing descriptor on UCF11, suggesting that high entropy is an indicator for discriminative power. On UCF50, however, \(IC_{1,1}-\text{HOG3D}\) is the best performing descriptor, which has considerably lower entropy compared to most other descriptors. When larger datasets exhibiting higher intra-class variability and lower inter-class variability are considered, it becomes more important for descriptors to be robust, as opposed to discriminative only. Another illustrative example of this phenomenon would be a raw pixel descriptor (list of pixel
values) which typically has high entropy and is very discriminative but not at all robust. Another high-entropy descriptor is the 2-channel $IH_{0,0}$-HOG3D. This is remarkable at first sight because its dimensionality is lower than e.g. the 3-channel $IN_{0,0}$-HOG3D descriptor. That is, entropy is generally expected to increase along with dimensionality. Furthermore, the results on UCF sports show that $IH$ descriptors perform worse than other descriptors in general which can be attributed to the instability of the hue representation for unsaturated colors resulting in high entropy in the extracted descriptor.

In conclusion, high descriptor entropy indicates either discriminative power or instability of the underlying representation. Discriminative power does not guarantee best performance because descriptor robustness becomes more important as the problem becomes more difficult.

<table>
<thead>
<tr>
<th>Table 5.6: Summary of best recognition results over all datasets.</th>
</tr>
</thead>
<tbody>
<tr>
<td># videos</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td># actions</td>
</tr>
<tr>
<td>Best Detector</td>
</tr>
<tr>
<td>Best Descriptor</td>
</tr>
</tbody>
</table>

5.7 Discussion

We have reformulated STIP detectors and descriptors to incorporate multiple photometric channels in addition to image intensities, resulting in color STIPs. The enhanced modeling of appearance results in an improved balance between photometric invariance and discriminative power, as chromaticity provides more information, based on which better representations are formed. Color STIPs are thoroughly evaluated and shown to significantly outperform their intensity-based counterparts for recognizing human actions on a number of challenging video benchmarks. In Table 5.6 we show an overview of the best results over all datasets. The best detector is consistently $IN$, although differences between $I$ and $IN$ are small. Consistent across all results is the superior performance of descriptors extracted from the unnormalized opponent representation $IC$. Differences are observed between variations of the $IC$ descriptor in terms of channel integration/concatenation and gradient orientation/direction, where the best descriptor choice depends on the difficulty and size of the dataset. For a small to moderate amount of visually relatively distinct categories such as
in the UCF11 dataset, it is best to use a discriminative descriptor such as $IC_{0,0}$ (channel concatenation + gradient direction). For larger datasets such as UCF50 it is better to use the robust descriptor $IC_{1,1}$ (channel integration + gradient orientation), which has the additional advantage of low dimensionality.
Summary and Conclusions

The individual chapters in this thesis are summarized before conclusion.

6.1 Summary

Chapter 2: The current most popular local image descriptors represent the gradient structure of the underlying image [55]. This family of descriptors is unstable on image regions with low signal strength (i.e. without structure). This instability arises from the fact that perturbations in the image signal due to photometric variations, noise and compression have a relatively large effect on regions with little structure. As a consequence, region descriptors also suffer from these instabilities. The standard approach to dealing with unstable descriptors is to ignore them all together [59, 94]. This, however, is an ad hoc treatment of which the optimal setting is data dependent and may lead to performance loss. Therefore, it is proposed in this chapter to explicitly model the instability in the descriptor coding scheme for image representation. This is accomplished by designing a measure of descriptor instability which can be interpreted as the variance of the observed descriptor and can be computed on-the-fly. The descriptor variance is incorporated in the Fisher Vector encoding framework, which changes the learning problem because the observations are now represented as high dimensional densities instead of vectors. These proposed Fisher Vectors are shown to increase recognition performance in image matching, retrieval and object recognition tasks.

Chapter 3: Local image descriptors are generally designed and optimized to
be robust against small geometric variations. To this end, it is common to simulate the sets of relevant transformations, and to analyze and exploit the resulting descriptor distributions \[60, 10\]. While this generally leads to a global optimum of descriptor settings, it is shown in this chapter that the quality of descriptor matching is dependent on the distributions per patch, as the specific effect of the transformations is determined by the image content itself. However, the computational complexity involved in such a simulation strategy at run time is prohibitively large. Therefore, it is proposed in this chapter to formulate the effect of a given image transformation in closed form. In essence, it is aimed to compute the covariance of an image sample using prior knowledge about the expected image transformations. A metric is derived from the thus obtained covariance and used in an image matching task, which performs comparably to the case in which the image transformations are actually repetitively applied to the image for covariance estimation.

Chapter 4: Whereas the chapters 2 and 3 focus on photometric and/or geometric disturbances, additional sources of appearance variation are due to object material and albedo. All these sources of variation are relevant to the formation of images in practice. Therefore, many image descriptors have been designed to ideally be robust against all disturbances \[55, 31, 89\]. Despite advances in this respect, the optimal descriptor characteristics vary per patch and as a consequence there exists no single best descriptor under all circumstances. This statement is supported by the experimental results in chapter 4. Here, it is shown in an ‘oracle’ experiment that the best image descriptor varies on a per-patch basis, and as a consequence that selecting the best descriptor per patch for matching performs better than using the globally best single descriptor for all patches. It is further proposed to perform the selection automatically by a classifier that predicts the best descriptor for a given image patch based on visual properties related to structure, material and color of the imaged content.

Chapter 5: It is well known from studies in the image domain that descriptors extracted from invariant image representations contribute substantially to the robustness of many computer vision systems. Another well known result is that the best performing descriptor varies per dataset \[89\], and even per image or image region, as shown in this thesis. In the video domain, however, such studies have not yet been performed. Here, local descriptors have been designed based on measurements of the luminance signal only \[98\]. As a consequence, representations based on these descriptors are sensitive to photometric variations and ignore potentially discriminative features from the color channels. Therefore, in chapter 5, existing approaches for (human action) recogni-
tion in video based on spatio-temporal interest points (STIPs) are reformulated in order to operate on multi-channel video representations. This essentially involves extending the existing formulations to handle vector observations instead of scalars only. It is shown that this may result in compact and powerful representations that substantially outperform their single-channel counterparts.

6.2 Conclusion

This thesis is based on the observation that different types of image content require different representations. Variations in image content apply globally as well as locally, ranging from image level annotations to the appearance of local structures. Because of the wide applicability of local image descriptors in computer vision, this thesis contributes by focusing on variations in local image content. The central question is whether local measurement types can be dynamically selected or adapted based on properties of the image signal such that the robustness of visual recognition methods is improved. This thesis concludes with three flavors of the central answer: concept, practice and future.

From a conceptual point of view, it is concluded from the results in this thesis that it is indeed possible to select or adapt a local descriptor based on properties of the image content itself. The thesis proposes to 1) incorporate a measure of descriptor instability in the image encoding scheme based on measurements of the underlying signal itself, 2) adapt the local image matching scheme based on the expected covariance per descriptor under a given image transformation and 3) dynamically select an image descriptor based on visual properties of the underlying image content. The implementations of these ideas are shown to improve the robustness of image matching, retrieval and recognition methods.

The practical applicability of the concepts in this thesis depends on the degree to which they are intrusive for the image representation or recognition method. In chapter 2, the measure of descriptor instability can be computed from the descriptor itself and can be modeled as the descriptor variance in the image encoding scheme. This contrasts with the proposed methods in chapters 3 and 4, in which the metric that is derived from the estimated descriptor covariance (chapter 3) resp. the selected descriptor (chapter 4) has to be applied to the to-be-matched descriptors as well. This is not viable for large scale application. Note that chapter 5 deviates from the central question in the sense that it does not propose to dynamically select or adapt a local descriptor. Instead, a set of novel video descriptors is introduced from which the best one is selected for practical usage.
Future work with respect to the specific content of this thesis should focus, where possible, on the efficiency of the proposed methods in chapters 3 and 4 in particular. The aim should be to improve the trade-off between the merits of per-patch analysis regarding matching performance and the advantage of global image analysis regarding processing efficiency.

In a broader sense, the exploitation of image descriptors that are engineered based on physical models as advocated in this thesis will remain a very important aspect of computer vision in the future. There are two main reasons for this, both of which relate to the current upcoming popularity of deep learning. Consider the case of deep learning resulting in the best performance for accomplishing some visual recognition task according to some measure. First, while the learned representation is then apparently very robust and powerful, it can not be the case that this representation subsumes the set of all engineerable descriptors. Thus, there can always be image descriptors engineered that add to the information as contained in some learned representation. Second, and more importantly, in our mission to enrich industry and society with computer vision and machine learning capability, it is of critical importance that the techniques are conveyed in ways that are easy to understand. In practice, it is seldomly the case that a company or individual has large amounts of relevant and clean data available to train models for solving exactly their well defined problem. Instead, it is essential to infiltrate the various domains in which ‘data science’ appears important to actively cooperate and discover the value of data together. In the case of visual data, this means being able to clearly express the type of image measurements that are being used, i.e. ‘shape’ and ‘color’, while ignoring ‘shadows’ and ‘highlights’. Without such communication it is not possible do trials, make errors, try again and accomplish joint success.
Bibliography

[29] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The ams-
Bibliography

93


V. Nedovic, A. W. M. Smeulders, A. Redert, and J. M. Geusebroek. Stages


[95] N. Vlassis and J. Verbeek. Gaussian mixture learning from noisy data.


De hoofdstukken in dit proefschrift zijn alhier in het Nederlands samengevat.

**Hoofdstuk 2:** De meest populaire hedendaagse representaties van lokale beeld-informatie zijn gebaseerd op de gradient structuur (ook wel ‘textuur’) van het onderliggende beeld. Dit soort representaties is instabiel op gebieden met weinig tot geen structuur omdat verstoringen in het beeld als gevolg van ruis, compressie en belichting daar het grootste effect hebben. De standaard methode van omgaan met instabiele representaties is om ze simpelweg te negeren. Dit kan echter dramatische gevolgen hebben voor de prestaties van een visueel herkenningsysteem. In dit hoofdstuk wordt een methode ontwikkeld waarin de instabiliteit van de lokale beeldinformatie expliciet wordt gmodeleerd tezamen met de representatie ervan. Dit wordt bereikt met een functie die de verwachte instabiliteit van de beeldbeschrijving op efficiënte wijze relateert aan de gradient structuur. Deze verwachte instabiliteit kan worden geinterpreteerd als de variantie van de lokale beeldrepresentatie en wordt als zodanig compact gmodeleerd in de krachtige ‘Fisher Vectors’ voor globale beeldrepresentatie. Het blijkt uit de experimenten in dit hoofdstuk dat de voorgestelde uitbreiding een consistente verbetering van visuele herkenningsystemen teweegbrengt.

**Hoofdstuk 3:** Beeldrepresentaties worden over het algemeen zo ontworpen dat ze robuust zijn tegen betrekkelijk kleine geometrische variaties, zoals rotatie en schaling, teneinde beter te kunnen vergelijken met andere beelden. Hiertoe is het gangbaar om de relevante variaties te simuleren en de resulterende statistische verdelingen van beeldinformatie te analyseren en exploiteren. Dit leidt over het algemeen tot de optimale parameters van de beeldrepresentatie in globale zin. Het wordt echter in dit hoofdstuk aangetoond dat de optimale parameters verschillen per beeldregio, omdat het specifieke effect van een geometrische variatie afhankelijk is van lokale beeldinhoud. Aangezien het simuleren van de variaties in de praktijk te veel tijd kost, wordt in dit hoofdstuk een methode on-
twikkeld waarmee de simulatie met een enkele matrixvermenigvuldiging kan worden uitgevoerd. Het doel is om de verwachte covariantie van een beeldregio te berekenen, onder de aannamer dat het soort variaties en de parameters ervan bekend zijn. De (inverse) covariantie wordt gebruikt om de beeldrepresentatie te wegen zodat de meest stabiele delen van de beeldregio de meeste invloed hebben in een vergelijking met een andere beeldregio. De resultaten laten zien dat de voorgestelde ‘simulatie simulatie’ (in gesloten vorm) net zo goed maar vele malen sneller werkt dan het geval waarin de simulaties daadwerkelijk worden uitgevoerd.

**Hoofdstuk 4:** De nadruk van de hoofdstukken 2 en 3 ligt op respectievelijk photometrische en geometrische vervormingen van het beeld. Andere bronnen van variaties in de verschijningssvorm van beeldinhoud worden veroorzaakt door materiaal en albedo. In de praktijk zijn al deze facetten relevant bij de totstandkoming van beelden, en om die reden zijn er vele beeldrepresentaties ontworpen die robuust zijn tegen zoveel mogelijk bronnen van variatie. Ondanks enorme vooruitgang op dit terrein, blijkt in dit hoofdstuk dat de kenmerken van de optimale beeldrepresentatie varieert per beeldregio. De consequentie hiervan is dat het bij het vergelijken van beelden altijd beter is om de beste representatie per beeldregio te hanteren in plaats van alleen de representatie die globaal het beste is. Daarom wordt in dit hoofdstuk voorgesteld om de selectie van de beste beeldrepresentatie automatisch uit te voeren op basis van eigenschappen van het beeld zelf zoals structuur, materiaal en kleur.

**Hoofdstuk 5:** Het is alom bekend in de computer vision dat het hanteren van robuuste beeldrepresentaties een substantiële verbetering van visuele herkenningsystemen teweegbrengt, in vergelijking met bijvoorbeeld representaties op basis van alleen de beeldintensiteit. Tevens is het zo dat de best presterende beeldrepresentatie varieert per dataset, en zelfs per beeld en beeldregio, zoals aangetoond in dit proefschrift. Dergelijke studies zijn echter nog niet uitgevoerd in het video domein, waar de beeldrepresentaties tot nu toe zijn gebaseerd op alleen de intensiteit van het beeld. De consequentie hiervan is dat de representaties gevoelig zijn voor photometrische variaties, alsook dat waardevolle informatie in de kleurkanalen wordt genegeerd. In dit hoofdstuk wordt voorgesteld om de bestaande videorepresentaties te herformuleren opdat meer dan één visueel kanaal kan worden beschouwd. Dit resulteert in robuuste video-representaties waarmee substantieel betere resultaten worden bereikt in het herkennen van menselijke activiteiten in videos.
Acknowledgements

The act of writing these last words of my PhD thesis resonates with the first track of the record on the cover: Here comes the sun! What a long and winding road it was. And somewhere down that road, I actually thought that I could finish a paper, have a baby, do a tour with my band, write another paper, run my own company, and be a good partner and father all at the same time within half a year or so. How much can you learn.

Brechtje, love of my life, thank you for your abiding persistence. Because of you and our children Dieuwertje en Boudewijn, it has become simple to make choices and focus on what is important in life. We did this together: we can be proud of it. Harvesting time!

Theo, I was already more than half way down in my PhD when you took over supervision. You managed to make me focus on the ball and run in the right direction. Your passes were sparse in the beginning, and I would miss in front of an open goal. But as we got more and more accustomed, you taught me how to position myself, approach the opponent and be creative while obeying the tactics, resulting in a number of brilliant goals. Your creativity in developing scientific ideas is impressive. I have enjoyed our close collaboration very much, and I want to thank you for backing me in the jobs at SightCorp and UvA, and for guiding me to our achievements in recent prestigious conferences and journals - that made it all worthwhile. I look forward to frequently bumping in to eachother in Maxwell, Erik’s and Alibaba.

Jan, what a pleasant surprise when you returned to the research group. It was a relief for me to be able to collaborate ‘in the mud’ with such a bright mind and kind man. We have meant a lot to each other in our work and the successful results thereof. That feels good.
Arnold, I will never forget our meetings - it has been a privilege. I sometimes describe you as the 'Johan Cruijff' of computer vision, which is thus both advantageous and disadvantageous. In my experience you have always valued the holistic Ivo Everts - this has given me strength at times when I needed it. Your secretary feels like a second home: thank you Virginie and Suna!

Roberto, we have spend a lot of hours in the same office together and we collaborated in virtually all possible ways: in the classroom, in academia, in industry, with family matters, at parties and numerous festivals. Uniquely good times! Nicu, you gave me the opportunity to do a PhD. I want to thank you for that, and for arranging everything a PhD student could ever dream of. The first years were certainly fun - focus came later. Thanks Thomas for saving chapter 2 :) and Sezer for our collaboration in chapter 3 - at least Eq. (3.3) made it to the cover :) Thanks Vlado for always being there for me. I could be sitting here with just pee-stains on my rug. Mom and dad, because of you I possess the perpetual self-confidence that has been needed for accomplishing this work. There have been times in which you were my only source of peace. And big thanks to all members and associates of the research group throughout the years for constructive and less constructive discussions in the office, at brouwerij 't Ij, in the canteen, on the football pitch, at the tennis table or bowling lane (or any other arbitrary location, as long as The Big Lebowski is on the screen).