# Statistical Modeling in the Wavelet Domain and Applications

by
Roland Kwitt

A thesis submitted to the
Department of Computer Sciences at the
University of Salzburg
in partial fulfillment of the requirements
for the degree of *Dr. techn.*

April 2010

| | |
|---|---|
| *Supervisor:* | Ao. Prof. Dr. Mag. rer. nat. Andreas Uhl<br>Department of Computer Sciences<br>University of Salzburg, Salzburg, Austria |
| *External Reviewer:* | Prof. Dr. Nick G. Kingsbury<br>Department of Engineering<br>University of Cambridge, Cambridge, United Kingdom |

# Abstract

In this thesis, we study statistical models for transform coefficients of two different wavelet transform variants, the pyramidal Discrete Wavelet Transform (DWT) and the Dual-Tree Complex Wavelet Transform (DTCWT). The work is motivated by the high computational demand of many state-of-the-art modeling approaches, although a variety of applications require computationally efficient, yet accurate models which facilitate straightforward parameter estimation and possess an analytically tractable form. In case of the DTCWT, there is also very little literature on (joint) statistical modeling of complex wavelet coefficients, even though it is a well-established fact that complex wavelet transforms exhibit striking advantages compared to the DWT when it comes to image analysis applications. The statistical models we develop throughout this thesis are utilized in three different areas of image processing. We address the research branches of (probabilistic) texture image retrieval, medical image classification and image watermarking. For each particular field, we provide a brief introduction of the problem, then introduce our contribution and conclude with an extensive experimental section. This includes a comparative study to existing work in literature and, depending on whether computational effort is a crucial issue, a thorough computational analysis of the main building blocks. Our results reveal, that the proposed models are beneficial in the aforementioned areas and improve upon state-of-the-art work. In addition, application of statistical models is not limited to the presented fields. In fact, we presume that other areas of transform domain based image processing, such as denoising or segmentation, can benefit in a similar manner.

# Acknowledgements

This thesis is devoted to my parents, Richard and Monika, and my girlfriend Daniela. Special thanks goes to my advisor Andreas Uhl for providing magnificent support and guiding throughout my PhD studies. I also like to thank my second advisor Karl Entacher and my co-worker Peter Meerald for great collaborations and many productive discussions in the last years. Eventually, I like to give thanks to the Austrian Science Fund (FWF) for funding this thesis under project no. L366-N15.

# Contents

# Chapter 1

# Introduction

In many disciplines of scientific research, measurements or observations in general make up the basis for any processing step. It is not uncommon to assume that these measurements stem from some underlying stochastic process. Consequently, many problems can be formulated as problems of statistical inference. At the very core of inferential procedures, we identify suitable statistical models which capture certain characteristics of the observations. In this thesis, we are particularly concerned with statistical inference problems which arise in the context of image processing. To be more specific, we focus on the area of transform domain image processing, where the wavelet transform in all of its variants has proved to be highly beneficial. In fact, the wavelet transform resembles the way our visual system processes information which makes it very attractive from an image processing point of view. The basic motivation for leaving the pixel domain and switching to a transform domain representation of images is to facilitate any kind of processing operations. Throughout the last years, statistical models for wavelet transform coefficients have found application in many areas of image processing, such as denoising [16, 127, 155], coding [164, 110], compression [14], classification [18, 106], image retrieval [40] or watermarking [69, 139, 12].

The pyramidal Discrete Wavelet Transform (DWT) [113, 114] is by far the most prevalent transformation in the image processing community. In a similar manner, the coverage of statistical models for DWT coefficients is quite extensive [71, 155, 190, 109]. Nevertheless, the DWT is not tailored for image analysis applications (i.e. classification, denoising, etc.) and even has some well-known deficiencies in this context. To overcome the shortcomings of the DWT, many alternative transformations have been developed recently, however, only a small subset has gained substantial interest in the community. Two of these alternatives are the Steerable Pyramid of Simoncelli et al. [168] and the Dual-Tree Complex Wavelet Transform (DTCWT), proposed by Kingsbury [85]. Since the transform coefficient statistics of coefficients from the Steerable Pyramid resemble the statistics of DWT coefficients, many works have been devoted to the development of suitable statistical models for Steerable Pyramid coefficients as well [190, 39, 179]. In contrast, the number of publications dealing with statistical models for DTCWT coefficients is substantially lower [163, 154, 19, 151]. Most works focus on models for the magnitudes of complex wavelet coefficients, however, recently the phase has gained research interest as well [128, 129, 189].

The motivation for developing novel statistical models for wavelet coefficients has several facets. In case of the DWT, our motivation is strongly related to the area of transform domain watermarking [28]. We identify two topics which received little treatment in literature so far. First, the commonly-accepted model for DWT coefficients, the Generalized Gaussian distribution [22] has the disadvantage of computationally expensive and numerically cumbersome parameter estimation [97]. Unless the model parameters are set to predefined values (see, e.g. [69]) – which might have a negative impact on detector performance – this fact prevents the use of a GGD-based detector in computationally demanding scenarios. As an alternative, we seek a statistical model which allows to derive a watermark detector with an analytically tractable form and a computationally inexpensive way to estimate the model parameters. This model might be less appropriate in terms of Goodness-of-Fit, yet accurate enough to outperform the standard Gaussian distribution in terms of watermark detection performance. As a second point, we highlight the fact that statistical models tailored to capture the association between DWT coefficients are primarily based on Hidden Markov models [155, 109]. Although, this allows to model inter- and intra-scale dependencies, those models turn out to be analytically intractable in Likelihood-Ratio testing scenarios. Since we want to facilitate color image watermarking in the wavelet transform domain, extending the Hidden Markov model approach to color images [192] is unrewarding. Instead, we seek a joint statistical model which can capture coefficient dependencies across color channels and yet allows to derive closed-form expressions for Likelihood-Ratio tests.

Regarding the development of statistical models for DTCWT coefficients, our motivation stems from a completely different research area. We are concerned with a medical image classification problem which bears a strong relation to the field of texture image retrieval and classification. Our intention is to evaluate whether statistical approaches to capture coefficient characteristics are equally effective for our medical problem as they are in texture analysis applications, see e.g. [150, 40, 179]. Our objective is to advance existing statistical models for DTCWT coefficient magnitudes and to quantify the suitability of the models with respect to classification and texture retrieval performance. Since both problems have strong computational constraints, we aim for analytically tractable approaches and straightforward parameter estimation. In addition, we are further motivated to develop a computationally simple alternative to the Hidden Markov Tree approach of [19] in order to capture DTCWT coefficient dependencies, especially across color channels. This seems a promising idea in consideration of the fact that color information has shown to be beneficial in texture discrimination scenarios [35].

## 1.1 Contribution

The contribution of this thesis is split into several parts. Basically, we discuss several statistical models for DWT and DTCWT coefficients and their application in three different areas of image processing. In the context of DWT coefficient modeling, we briefly review the popular Generalized Gaussian model and the less often used Cauchy distribution. The latter model is then used in the context of image watermarking to derive a computationally efficient watermark detector which exhibits substantially better detection performance than several state-of-the-art detectors on a large set of natural images. In order to incorporate coefficient dependencies among the subbands of DWT decomposed color channels into the watermark detection process, we present a joint statistical model which can be considered as a multivariate extension to the Generalized Gaussian distribution. We deal with parameter estimation issues and suggest a novel Goodness-of-Fit test to quantify the suitability of the model. In an extensive size and power study we show that the desired significance levels can be met and that the test exhibits remarkable power against shape alternatives. Eventually, we derive a novel watermark detector based on the joint statistical model and demonstrate that our detector performs better than two state-of-the-art detectors in field of color image watermarking.

In the context of DTCWT coefficient modeling, we advance current research results to the effect that we present two novel models for subband coefficient magnitudes which are both accurate and admit straightforward parameter estimation. We quantify the suitability of the proposed models by means of an extensive Goodness-of-Fit study on four commonly-used texture image databases. The modeling results are then exploited for lightweight texture image retrieval where we propose a novel retrieval approach based on a probabilistic formulation of image retrieval [186]. We show that switching from computationally expensive Maximum-Likelihood parameter estimation procedures to moment matching approaches does not negatively affect the retrieval rates, however considerably lowers the computational burden of this step. A computational analysis of the main building blocks of the retrieval framework confirms that we can design a probabilistic approach with low computational complexity. In contrast to the majority of research papers on texture image retrieval, we conduct an extensive comparative retrieval study on four texture image repositories to evaluate the quality of our proposed approach with respect to several state-of-the-art approaches.

In a second step of modeling the DTCWT coefficient magnitudes, we present an alternative model to the Hidden Markov Tree approach of [19]. Since DTCWT coefficients exhibit a quite strong association structure, it appears reasonable to capture this association by a joint statistical model. For that purpose, we propose a copula-based approach which (i) allows to rely on existing knowledge about the DTCWT coefficient statistics and (ii) completely separates the task of finding a suitable model for the association structure. We show, that the copula-based model for DTCWT coefficients can be exploited for texture image retrieval and perfectly fits into the probabilistic framework we mentioned above. Again, we can demonstrate a considerable increase in retrieval performance, however, at the expense of computation time. To remedy this shortcoming, we suggest a simple data reduction strategy which only slightly affects the retrieval results, but allows to deploy the approach even on large databases.

As a third field of application, we tackle the medical image processing problem of predicting histologies from colonoscopy images based on the visual appearance of the mucosal surface patterns. We demonstrate, that a computer-assisted prediction system can be a serious diagnostic tool for *in vivo* staging of colorectal lesions. In particular, we consider two different strategies to cope with that problem. First, we take the straightforward way of using a discriminant classifier approach. Second, we consider the prediction problem from the viewpoint of image retrieval and discuss the advantages of a generative model based approach. In the former case, we exploit the statistical models for DTCWT coefficient magnitudes to construct feature vectors based on the estimated model parameters. Then, we extend the concept of co-occurrence matrices (see [65, 148]) to capture the joint occurrence of wavelet coefficients across different color channels and compute a set of commonly-used texture descriptors from these matrices. Eventually, we present an approach of decorrelating wavelet subbands from different color channels and using the variances of the decorrelated subbands as image features. In all three cases, classification is based on a nearest-neighbor principle and we demonstrate remarkable classification rates for two clinically relevant scenarios. In the context of generative models, we highlight potential disadvantages of discriminant classifier based approaches and emphasize the points where a retrieval oriented point of view can be beneficial. We present impressive prediction results for the image retrieval approaches with similar or higher rates compared to human-based studies.

## 1.2 Organization

In the remaining part of this introductory chapter, we include a brief discussion of the four image databases we use throughout the thesis. Further, we provide some notational conventions and address the topic of reproducible research. The remaining chapters are then organized into two major parts: in the first part, i.e. Chapter 2, we develop the statistical foundation of the following chapters. The second part is devoted to the areas of application of the different statistical models. Each application-specific chapter is structured in a similar way: first, we introduce the presentation of the problem, then we present our contribution and conclude with an experimental evaluation and a brief discussion of the results. Since the fields of application span different research areas, it is unrewarding to devote a separate chapter to related research work. We rather follow the strategy to establish connections to previous works as we progress from chapter to chapter. In Chapter 3, we revisit a recently proposed formulation of probabilistic image retrieval and then exploit the statistical models for DTCWT coefficients to develop two novel retrieval approaches. Chapter 4 is devoted to the medical image classification problem and Chapter 5 deals with the topic of image watermarking. Chapter 6 concludes the thesis with a confrontation of the original questions and the achieved results. Finally, we provide an outlook on open research problems and topics we could not cover in this thesis.

## 1.3 Image Databases

Image databases constitute the basis for all experimental results presented throughout this thesis. We use one database of natural images (UCID [159]) and three databases of texture images. The three texture databases consist of two commonly-known repositories (Outex [142] and Vistex [31]), and one real-world database of textures captured by the author and several coworkers[1] (Stex). We consciously exclude two other popular databases, the Brodatz album [13] and the CUReT [30] textures for several reasons: first, availability of the Brodatz album is limited to grayscale images[2] and the amount of available textures differs in literature (111 in [149], 112 in [186] or even 116 in [111]). Second, CURet[3] only provides a set of 61 different physical textures, however, under 205 different viewpoint and illumination combinations. Since we already use the Outex database which contains textures captured under artificial conditions, we choose not to include another database of this kind. Example images from all four databases are shown in Fig. 1.1 including some commonly-known example images (i.e. Fig. 1.1a) we often use for illustration purposes.

**UCID** Summarizing the description of Schaefer & Stich [159], the UCID image database consists of 1338 images in uncompressed form (TIFF format) captured by a Minolta Dimage 5 camera. All images are either 512 × 384 or 384 × 512 pixel and were captured using automatic settings which mostly resembles a real-world scenario.

**Outex** Since the test suite for texture retrieval in the Outex database only consists of grayscale images of size 128 × 128 pixel, we first fetched 316 color texture images in BMP format

[1] thanks to Heinz Hofbauer, Stefan Huber, Peter Meerwald and Daniela Wöckinger

[2] available from `http://www.ux.uis.no/~tranden/brodatz.html`

[3] available from `http://www.cs.columbia.edu/CAVE/software/curet`

with 600dpi under inca lightning conditions from the Outex website[4]. Two images, canvas007, canvas010 were missing, wallpaper015 was not accessible. The images were then cropped to 512 × 512 pixel starting from the top-left hand corner of the image.

**Vistex** We use the original 512 × 512 pixel versions of the texture images available from the MIT Vision Texture website[5]. There are 167 textures available, denoted by Vistex (full). We further select a subset of 40 textures, denoted by Vistex (small), since many approaches in various publications (see, e.g. [40, 101, 188]) use this limited subset. According to the information on the website, images in the Vistex database were captured under real-world conditions without studio lightning.

**Stex** The Stex database is a novel texture database consisting of 476 images of different textures captured in the area around Salzburg/Austria using three cameras: a Canon IXUS 70, a Canon EOS 450D and a Nikon D40. Similar to the Vistex database, our image set is intended to resemble a real-life scenario. Except for the Canon EOS 450D pictures which were captured in RAW format, all other textures were stored as JPEG images. Post-processing consisted of conversion to PNM format (using the ImageMagick's convert tool) and resizing to 512 × 512 pixels by means of bicubic interpolation (using MATLABs imresize routine).

## 1.4 Notational Conventions

To reach maximum notational consistency, we have to introduce some conventions. First, if not stated otherwise, uppercase letters (i.e. $X$) will be used to denote random variables. Lowercase letters (i.e. $x$) will denote observations. Accordingly, boldface uppercase letters denote random vectors (i.e. $\mathbf{X}$). In case $\mathbf{X}$ denotes a matrix, the meaning will be unambiguous from the context. Lowercase boldface letters (i.e. $\mathbf{x}$) will denote observation vectors. We adhere to the convention that $F_X$ denotes the cumulative distribution function (c.d.f.) of a random variable $X$ and $p_X$ denotes the corresponding probability density function (p.d.f) or the probability mass function (p.m.f.) in case of discrete random variables. Greek letters, such as $\alpha$ or $\boldsymbol{\alpha}$ denote parameters or parameter vectors, respectively. Entities, such as images, will be denoted by calligraphic letters (i.e. $\mathcal{I}$). When we speak of an image database, we mean a collection of images $\mathcal{I}_1, \ldots, \mathcal{I}_L$ of size $L$. Regarding the use of special functions, $\Gamma$ denotes the Gamma function and $\psi$ denotes the Digamma function [1]. All further notational conventions will be introduced at the corresponding locations.

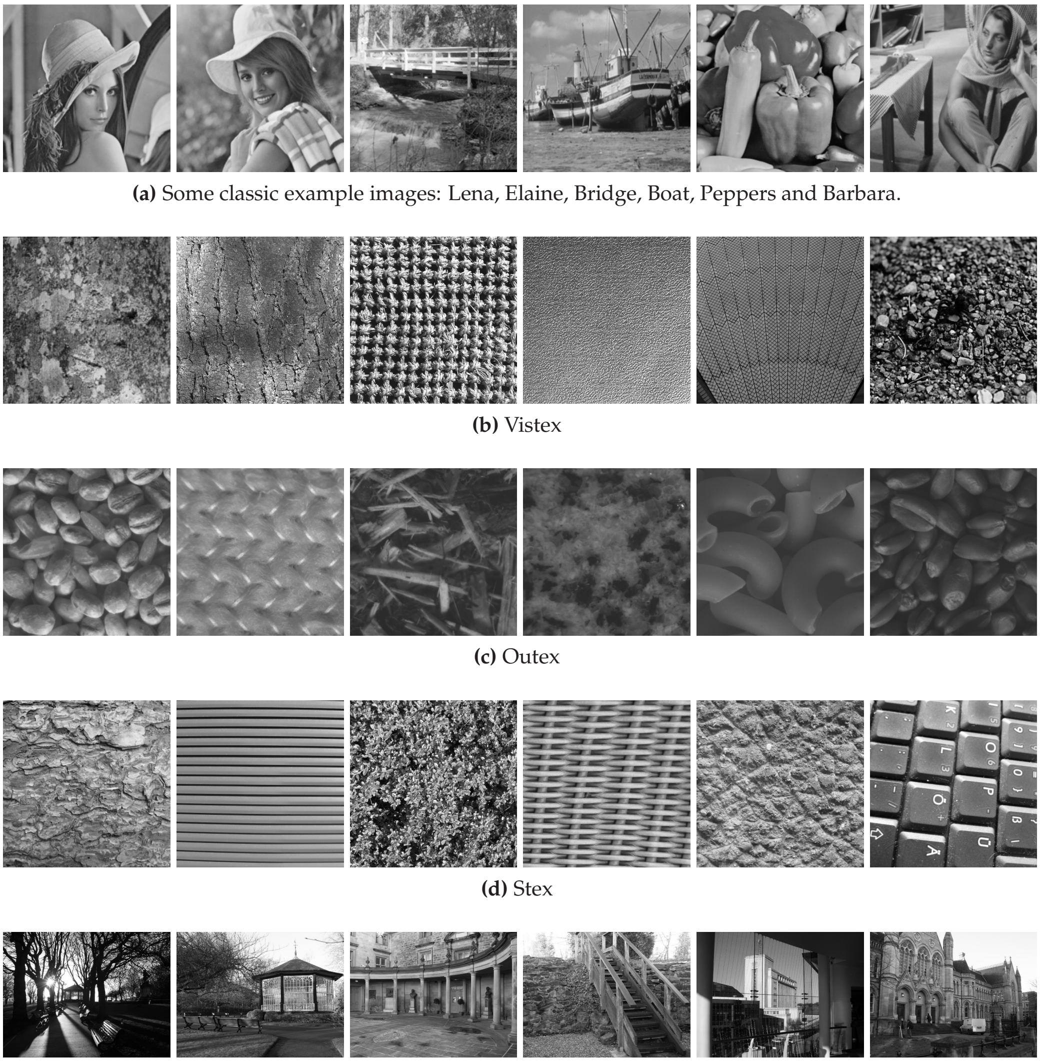
## 1.5 Some Notes on Reproducibility

As Vandewalle et al. (see [181] and references therein) recently pointed out, reproducible research is at the very core of every scientific discipline. In order to reach a certain degree of reproducibility of the results presented in this thesis, we provide reference implementations of all approaches as either C or MATLAB code[6]. Further, we provide access to the Stex database as another reference repository to evaluate texture analysis algorithms. Unfortunately, access to the medical database we use in Chapter 4 is restricted due to privacy issues.

[4] available from http://www.outex.oulu.fi/

[5] available from http://vismod.media.mit.edu/vismod/imagery/VisionTexture/

[6] available from http://www.wavelab.at/sources



**(a)** Some classic example images: Lena, Elaine, Bridge, Boat, Peppers and Barbara.

**(b)** Vistex

**(c)** Outex

**(d)** Stex

**(e)** UCID

**Figure 1.1:** Example images from different image databases.

# Chapter 2

# Statistical Modeling in the Wavelet Domain

In this chapter, we discuss the foundation of this thesis, namely the statistical models of wavelet coefficients from two different wavelet transforms. We start with an introduction of a set of statistical tools which we extensively use in the following sections. Other statistical procedures which are used in this thesis will be introduced when needed. After this brief introduction, the chapter is basically split into two parts: in the first part, we recapitulate the main results on statistical modeling of Discrete Wavelet Transform (DWT) coefficients, and in particular, we take a closer look at the characteristic distributions which arise in case of natural images. Then, we present a novel multivariate model to capture the dependencies across DWT detail subbands of different color channels and develop a novel Goodness of Fit test for this multivariate model. In the second part, we continue with a discussion of characteristic coefficient distributions which arise when we decompose images by means of a complex wavelet transform variant, known as the Dual-Tree Complex Wavelet Transform (DTCWT). We particularly focus on statistical models for DTCWT transform coefficient magnitudes of texture images. Finally, we present a multivariate extension to the univariate models in order to capture coefficient dependencies across subbands and color channels.

## 2.1 The Statistical Toolset

A commonly observed situation in the first stage of finding a suitable statistical model for a set of (univariate) observations is to analyze the frequency distribution. Usually, a classic histogram is used as a first choice where the range of observation values is divided into a certain number of bins (with equal bin width) and we count the number of observations falling into each bin. Plotting the bins against the bin count then conveys an impression about the frequency distribution. However, in case our objective is to highlight certain characteristics of the observations such as tail behavior for instance, other variants of the classic histogram are more reasonable. In situations where we expect heavy tails for example, it has become common practice to visualize the y-axis of the histogram on a logarithmic scale. We refer to this type of histogram as the *log-scale histogram*. In order to check the Goodness of Fit (GoF) of a selected statistical model, we employ Q-Q plots as a graphical tool and Chi-Square GoF tests to obtain a quantifiable measure of model fit. Basically, both the Q-Q plot and the Chi-Square GoF test are implemented according to the algorithmic description provided by Krishnamoorthy [89]. When not stated otherwise, the significance level $\alpha$ is set to 5%. Since the binning strategy is a crucial point when testing the GoF by means of a Chi-Square test, we adopt the bin width of $0.3s$ as the standard setting, where $s$ denotes the sample standard deviation. This setup is used in the software DATAPLOT [66]. In case of empty edge bins, the bins are combined with the next non-empty bin. In contrast to univariate GoF testing, statistical tests for the GoF of multivariate models are a neglected issue in literature. Chi-Square tests are computationally not feasible in general, since it is not trivial to choose a suitable binning of the possibly high-dimensional space. Even in three dimensions we expect many cells with cell counts of less than 5 observations, an empirical requirement of the Chi-Square test. Tests for multivariate normality are an exception to the rule, since some GoF tests (see [27, 170]) actually exist. In Section 2.2.3, we will take up the rather generic GoF test idea of Smith & Jain [170] and propose a novel GoF test for a special multivariate distribution. Last, we introduce a less commonly known graphical tool to assess the dependency structure between pairs of observations, such as pairs of wavelet coefficients from different subbands or different color channels. Besides the classic measures of association, i.e the linear correlation coefficient, Kendall's $\tau$ [84] or Spearman's $\rho$ [173], the so called Chi-plot of Fisher & Switzer [131] is a valuable visual tool. The basic idea of a Chi-plot is to transform the pairs of observations in such a way that the resulting pairs (residing in the interval $[-1, 1] \times [-1, 1]$) reveal the structure of association. Hence, it can be considered as an extension of the scatterplot which is usually employed to illustrate possible dependencies. In a Chi-plot, departures from independency are indicated by a deviation from the central region of the plot. A tolerance band is defined to allow slight scattering caused by sampling variability. Our implementation follows the description given in [131, 48, 47], with the tolerance region enclosed by horizontal lines at $\pm^{c_p}/\sqrt{n}$, where $c_p = 1.78$ and $n$ denotes the number of observations. This is a common setting, as it is noted in [48, 47] for example. In the Chi-plots, the tolerance band will always be shown as a gray-shaded region.

## 2.2 DWT Subband Models

The Discrete Wavelet Transform provides a convenient way to obtain a multiscale representation of an image which closely resembles the way the human visual system processes information [105, 152, 33]. It possesses some attractive properties of which the most important three are highlighted in [29]: first, *locality* denotes the fact that wavelets are localized in both space and frequency simultaneously. Second, *multiresolution* allows to analyze a signal at different scales, hereby allowing to capture both short- and long-term structures. Third, *compression* denotes the fact that we obtain a sparse representation of a signal which explains the highly non-Gaussian nature of the transform coefficients. Another interpretation of the compression property is that we obtain a large number of small coefficients containing little signal information and a small number of large coefficients representing significant signal information. From a computational point of view, the DWT is also very appealing since it provides a non-redundant representation of an image and it can be computed with linear complexity. The decomposition of an image by a 2-D DWT can be efficiently computed by separate row and column filtering and leads to four subbands per scale with one approximation subband and three detail subbands capturing image details oriented along the horizontal, vertical and diagonal (i.e. ±45°) direction. Hence, a $J$-scale 2-D DWT leads to $J \times 3 =: B$ detail subbands in total. Figure 2.1 shows all subbands (including the approximation subband) of a one-scale 2-D DWT of the test image Lena. To highlight the directional selectivity of the detail subbands, i.e. the important frequency information (e.g. edges) in the different directions, we only show the coefficients with absolute values above the 0.9 quantile (i.e. the largest 10% of all coefficients).

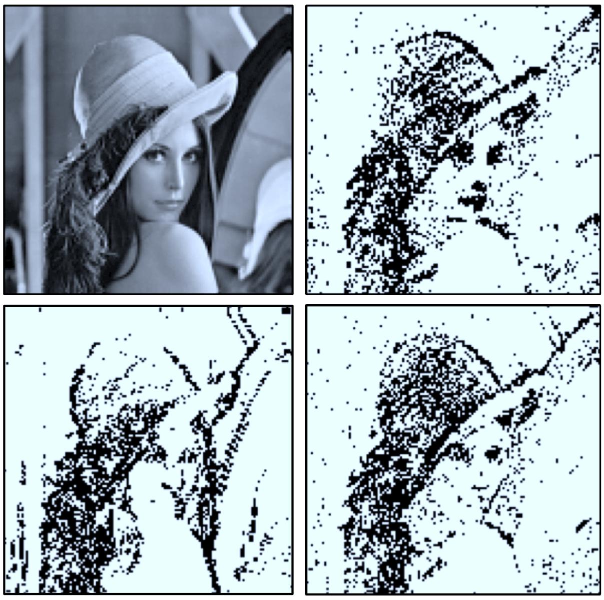


**Figure 2.1:** One-scale 2-D DWT decomposed test image Lena using a CDF 9/7 filter [32].

In order to visualize the non-Gaussian nature of the transform coefficients, Fig. 2.2 shows a collection of log-scale coefficient histograms obtained from different detail subbands. The plots include the p.d.f. of fitted Gaussian distributions as a reference model. Further, we list the kurtosis "excess" $\gamma_2$ [1] which is supposed to be zero in case the coefficients actually follow a Gaussian law. From the considerable deviation in the middle and tail region of the plot and the strong positive values of $\gamma_2$ (i.e. leptokurtic), we conclude that the Gaussian distribution is a bad statistical model for the coefficients.

Regarding the issue of intra- and inter-scale coefficient dependencies and implications for statistical modeling, we state three assumptions which often implicitly occur in literature. Basically, these assumptions are motivated by the fact that the 2-D DWT can be considered as an approximate Karhunen-Loéve transform [115] and hence acts as a *decorrelator*. However, as it is pointed out by Crouse et al. [29] or Liu & Moulin [109] this is only partially true.

**Assumption 1.** *The transform coefficients* $x_{b1}, \ldots, x_{bN_b}$ *of an arbitrary 2-D DWT detail subband* $b, 0 < b \leqslant B$ *are assumed to be a realization of* $N_b$ *i.i.d. copies* $X_{b1}, \ldots, X_{bN_b}$ *of a random variable* $X_b$*, where* $N_b$ *denotes the number of transform coefficients of that subband.*

This assumption neglects the *clustering* property of wavelet coefficients [29], i.e. that small /large coefficients tend to have small/large adjacent coefficients with high probability. This property is successfully exploited by LoPresto et al. in [110] for the purpose of wavelet-based image coding for example.

**Assumption 2.** *The transform coefficients of different subbands of the same scale are considered to be independent.*

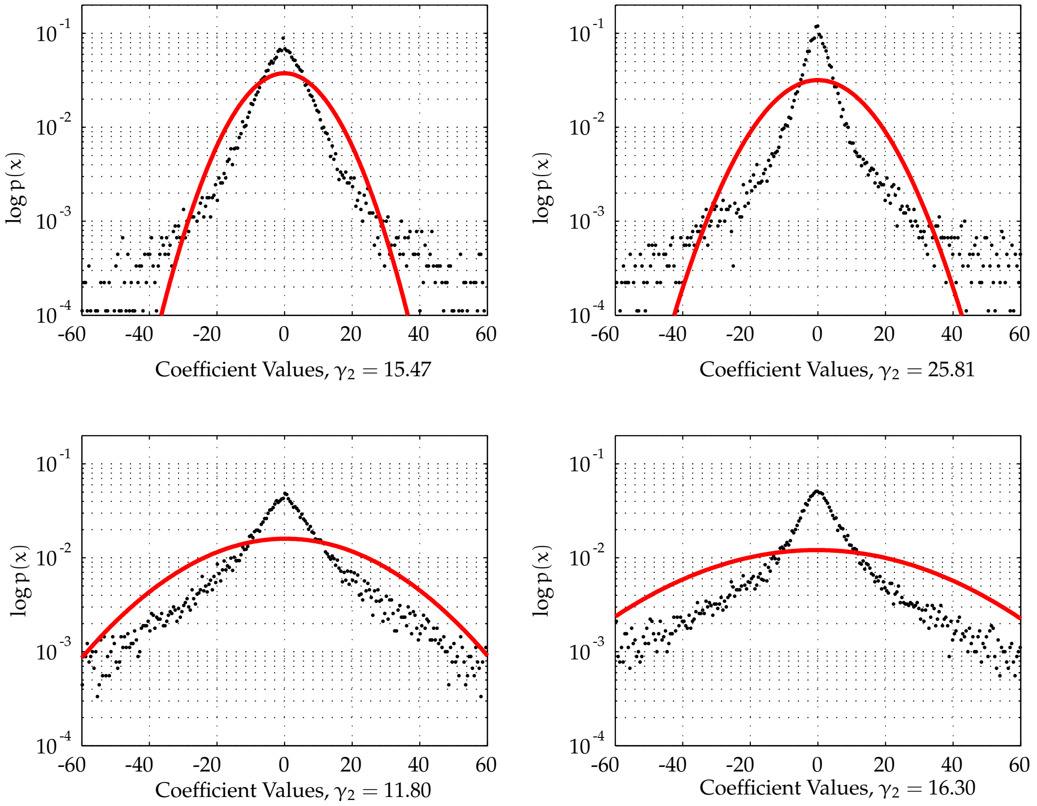


**Figure 2.2:** Log-scale histogram of the vertical DWT detail subband of four different natural images showing the coefficient values (black points) and the p.d.f.s of fitted Gaussian distributions ($\gamma_2$ denotes the sample kurtosis "excess").

Given that $h, v, d$ identify the horizontal, vertical and diagonal detail subband at an arbitrary decomposition level, then the joint p.d.f. of $\mathbf{X} = (X_h, X_v, X_d)$ be written as $p_{\mathbf{X}} = p_{X_h} \cdot p_{X_v} \cdot p_{X_d}$. Basically, this allows to estimate statistical model parameters separately for each subband on the same scale. We can quantify the validity of the assumption by using Chi-plots constructed from the coefficients of subband pairs on the same scale. Figure 2.3 shows a set of Chi-plots for a selection of such pairs where we can observe that the observations are located around the shaded region or even inside, especially in the central (i.e. $\lambda \approx 0$) part of the plot. This visual impression does not admit to postulate independence, however, the deviation from the central region is also not distinctive enough to claim the opposite. Further, the linear correlation coefficient $r$, Spearman's $\rho$ and Kendall's $\tau$ exhibit values close to zero which at least indicates no correlation.

**Assumption 3.** *The transform coefficients of subbands across different scales are considered to be independent.*

In combination with the previous assumptions, this allows to write the joint p.d.f. of the random vector $\mathbf{X} = (X_1, \ldots, X_B)$ as $p_{\mathbf{X}} = p_{X_1} \cdots p_{X_B}$. This assumption is definitively a very strong one, since inter-scale dependencies do exist and have been successfully exploited in the coding community by means of zero-trees [164] or for signal estimation and detection [29]. However, all three assumptions contribute to the same objective, namely to allow the use of simple and analytically tractable models which can be estimated in a computationally efficient and reliable way. In the following two sections, we present statistical models for the p.d.f. $p_{X_b}$ and rely on all three assumptions stated above. Regarding the notation, we follow the convention to omit the subband index $b$ in cases where there is no added value. Further, when we speak of the DWT we mean the 2-D variant from this point on. Another convention we follow is to identify the statistical model of a particular subband by indexing the parameter (vector) $\theta$ of the corresponding model.

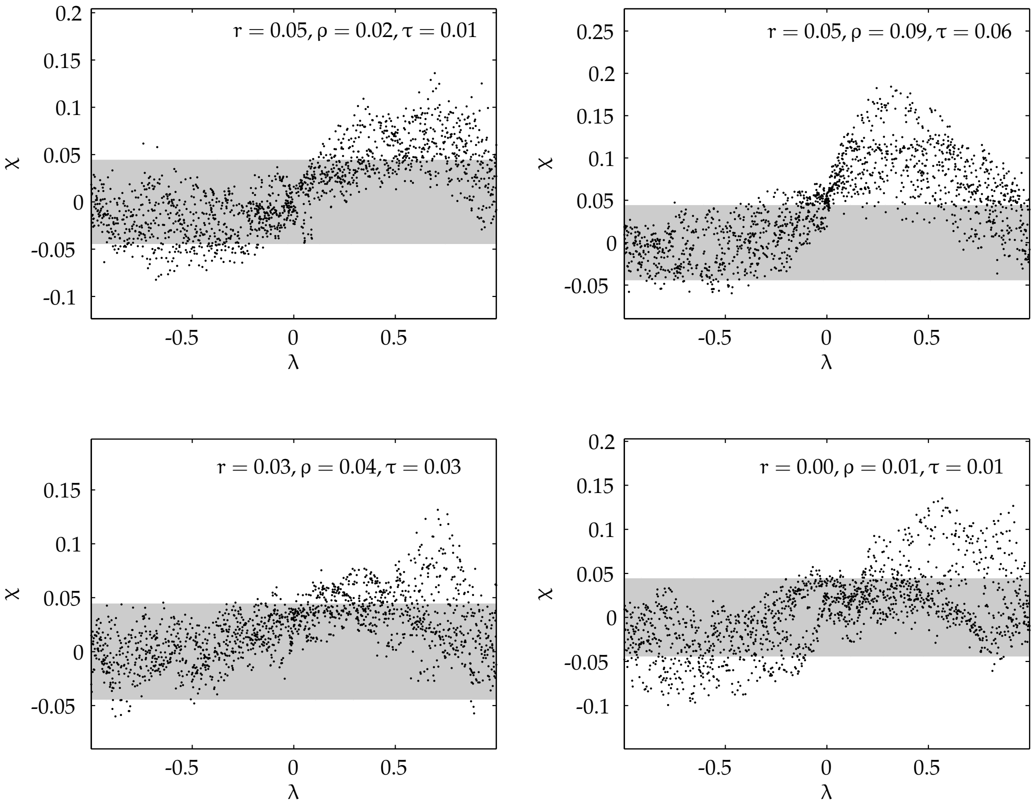


**Figure 2.3:** Exemplary Chi-plots of the vertical and horizontal DWT detail subband (level three) of four natural images to illustrate the *approximate* decorrelation of DWT coefficients across subbands of the same scale.

### 2.2.1 Generalized Gaussian Distribution (GGD)

The Generalized Gaussian distribution is by far the most popular statistical model for DWT detail subband coefficients and has been extensively used in literature. The GGD first appears in a textbook by Clarke [22] for modeling the AC coefficients of a Discrete Cosine Transform (DCT). In the context of DWT transform coefficients, Mallat [113] proposes the GGD as a reasonable model to capture the non-Gaussian nature of the transform coefficients. In this thesis, we use the GGD parametrization of Nadarajah et al. [133], where the p.d.f. with shape parameter $c > 0$, scale parameter $a > 0$ and location parameter $\mu \in \mathbb{R}$ is given by

$$p_X(x; a, c) = \frac{c}{2a\Gamma(1/c)} \exp\left(-\left|\frac{x-\mu}{a}\right|^c\right), \quad -\infty < x < \infty. \tag{2.1}$$

We can safely assume $\mu = 0$ in our case since the DWT transform coefficients theoretically sum to zero [115]. The Laplace distribution [89] arises as a special case of the GGD for $c = 1$ and the Gaussian distribution can be obtained by setting $c = 2$. The relation to the Gaussian distribution can be easily checked by using Euler's reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$ for $z = 0.5$ which gives $\Gamma(0.5) = \sqrt{\pi}$. Since the inverse c.d.f. (i.e. the quantile function $F^{-1}(u) = \inf_{x\in\mathbb{R}}\{F(x) \geqslant u\}, u \in [0, 1]$) is needed for the computation of the Q-Q plot, we briefly restate [133, 134]

$$F_X^{-1}(u; a, c) = \begin{cases} -a\left[P_u^{-1}\left({}^1\!/\!_c,\ 2u\right)\right]^{1/c} & \text{if } u \leqslant 0.5 \\ a\left[P_u^{-1}\left({}^1\!/\!_c,\ 2(1-u)\right)\right]^{1/c} & \text{if } u > 0.5 \end{cases}, \tag{2.2}$$

where

$$P_u(a, x) := \frac{1}{\Gamma(a)} \int_x^\infty t^{a-1} \exp(-t) dt \tag{2.3}$$

denotes the regularized (upper) incomplete Gamma function[1] [1]. Regarding the issue of parameter estimation based on an i.i.d. sample $x_1, \ldots, x_N$ , basically two methods are commonly used in literature: Moment Matching (MM) and Maximum Likelihood (ML) estimation. Moment matching is discussed by Mallat [113] and Birney et al. [10]. Unfortunately, computation of the moment estimates requires to find a numerical solution to a function inversion problem. A computationally fast way to approximate this function inversion problem is discussed by Krupinski [91], other authors commonly use a lookup-table approach (e.g. [40]). ML estimation is extensively covered by Varanasi et al. [182] and a Newton-Raphson algorithm to compute a numerical solution to the ML equations is introduced by Do & Vetterli [40]. Starting values for Newton-Raphson are obtained using moment estimates based on the lookup-table approach. Whenever we mention ML estimation for the GGD parameters in this thesis, we refer to the procedure given in [40]. Due to the computational and numerical difficulties related to parameter estimation of the GGD in general, Song [171] introduced a novel method based on a convex shape equation. In Section 3.3, we will revisit the computational demand of the various estimation methods in terms of required arithmetic operations. Fig. 2.4 shows the same log-scale coefficient histograms of Fig. 2.2 together with the p.d.f.s of fitted GGDs. To illustrate the GoF, Fig. 2.5 then shows some Q-Q plots for arbitrarily chosen subband coefficients from our test images. Although we observe slight deviations in the tail regions of the Q-Q plots, the points approximately follow the dashed line. In Section 2.2.4, we revisit the question of GoF by means a quantitative study using Chi-Square GoF tests conducted on the subband coefficients of the UCID images.

### 2.2.2 Cauchy Distribution

In [12], Briassouli et al. introduce the Cauchy distribution as a possible alternative for modeling the AC coefficients of DCT transformed images in the context of digital image watermarking.

[1] To avoid confusion, this function is implemented by the MATLAB routine `gammaincinv(x,a,'upper')` or by `InverseGammaRegularized[a,x]` in Mathematica.

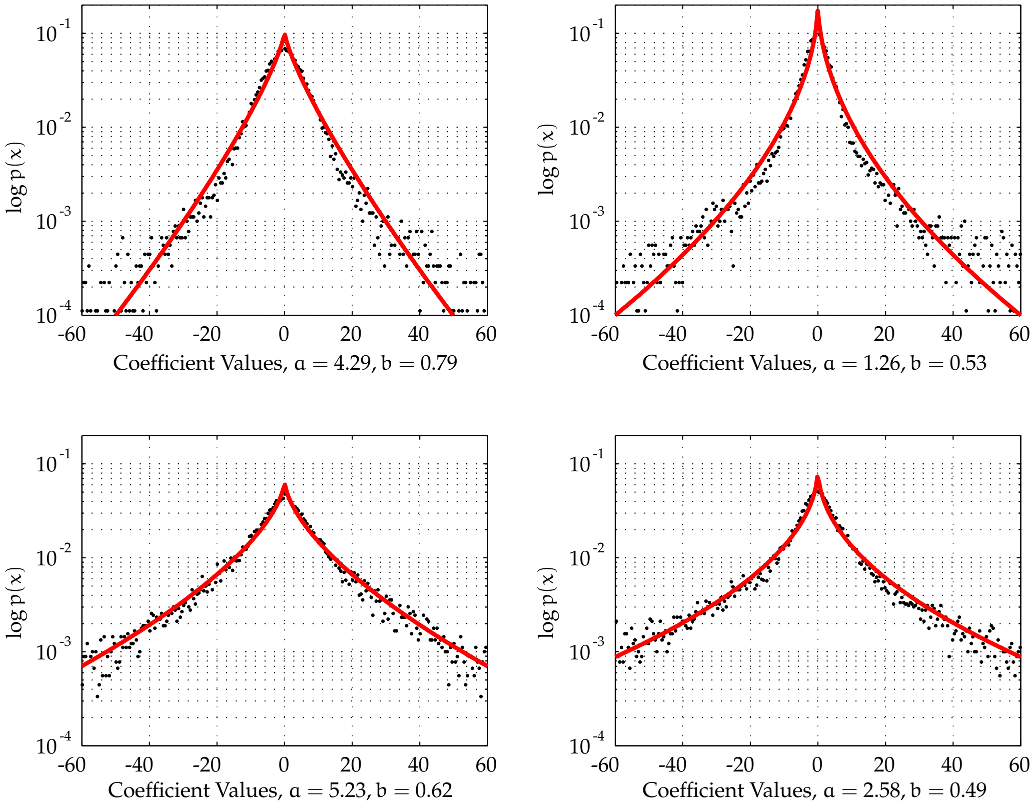


**Figure 2.4:** Log-scale histogram of the vertical DWT detail subband coefficients of four different natural images showing the coefficient values (black points) and the p.d.f.s of fitted Generalized Gaussian distributions (using ML estimation).

In [94], we exploited this model for modeling DWT detail subband coefficients for the purpose of image watermarking as well. The p.d.f. of the Cauchy distribution with location parameter $-\infty < \delta < \infty$ and shape parameter $\gamma > 0$ is given by [89]

$$p_X(x;\gamma,\delta) = \frac{1}{\pi}\frac{\gamma}{\gamma^2 + (x-\delta)^2}, \quad -\infty < x < \infty. \tag{2.4}$$

Again, we can safely assume that the location parameter $\delta$ is zero for the same reason explained in Section 2.2.1 and abbreviate the p.d.f. by $p_X(x;\gamma) := p_X(x;\gamma,0)$. In contrast to the Gaussian distribution, the tails of the Cauchy distribution decay at a rate slower than exponential, hence we observe heavy tails. The inverse c.d.f. which is needed to compute the Q-Q plots is given by

$$F_X^{-1}(u;\gamma) = \gamma \tan(\pi(u-0.5)), \quad 0 < u < 1. \tag{2.5}$$

It is worth noting that neither the mean nor the variance or any other higher moments are defined for the Cauchy distribution. To illustrate the shape of the p.d.f., Fig. 2.6 shows log-scale histograms of same DWT detail subband coefficients as in the previous section, together with fitted (ML estimation) Cauchy p.d.f.s. Note, that the case $\gamma = 1$ would indicate a standard Cauchy distribution.

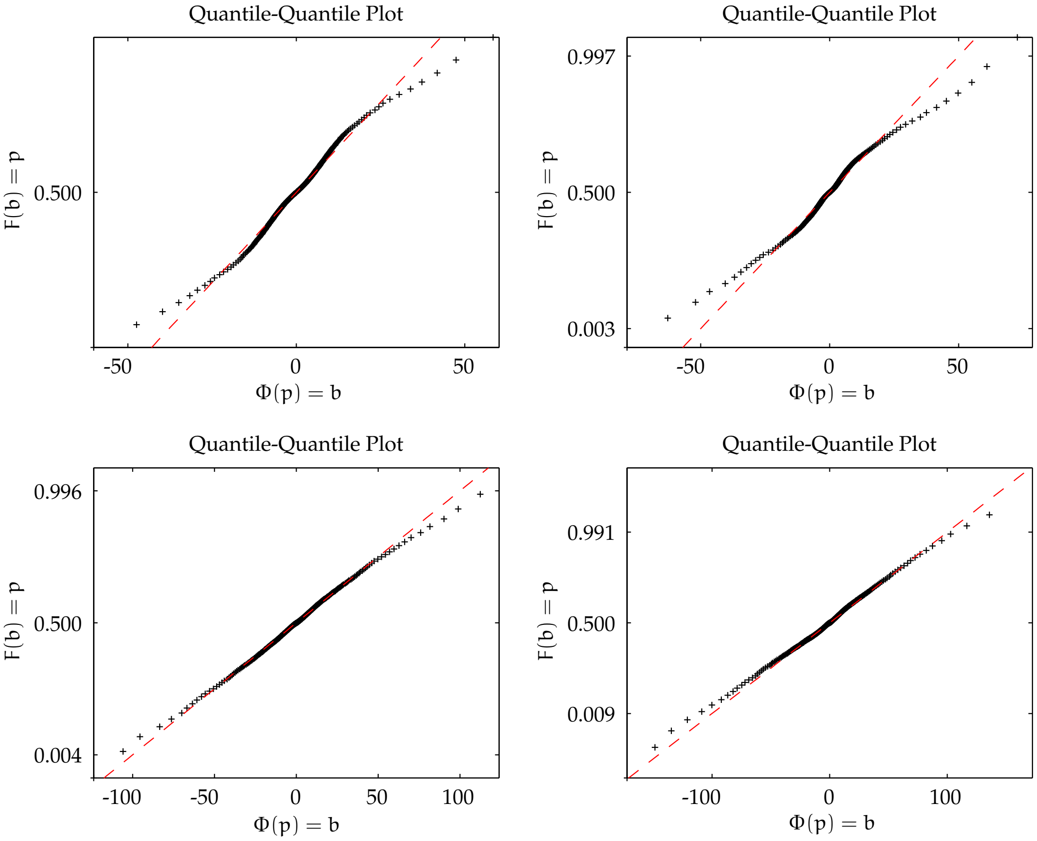


**Figure 2.5:** Exemplary Q-Q plots to visualize the GoF of the Generalized Gaussian distribution for the DWT transform coefficients of the vertical detail subband of four natural images (at DWT level two).

Regarding the estimation of the shape parameter $\gamma$ from an i.i.d. sample $x_1, \dots, x_N$, we can either rely on sample quantile estimation, direct ML estimation or the estimation approach proposed by Tsihrintzis & Nikias [176] for Symmetric $\alpha$ Stable (S$\alpha$S) distributions [138]. The last approach is particularly interesting, since the Cauchy distribution is a special case of an S$\alpha$S distribution for $\alpha = 1$ and the estimate of $\gamma$ can be computed with linear effort, i.e. $\mathcal{O}(N)$. Given the estimation setup of our problem, i.e. $\delta = 0$ and $\alpha = 1$, the shape estimator presented in [176] is

$$\hat{\gamma} = \left[ \frac{\frac{1}{N} \sum_{i=1}^{N} |x_i|^p}{C(p,1)} \right]^{1/p} \quad \text{with} \quad C(p,1) = \frac{1}{\cos\left(\frac{\pi}{2}p\right)} \tag{2.6}$$

for $0 < p < 1/2$. The parameter $p$ denotes the order of the fractional moment and can be chosen arbitrarily according to [176]. As it is pointed out by the authors, the choice $p \approx 1/3$ is reasonable and has shown good performance. Estimation based on the sample quantiles and ML estimation is given in [89]. The sample quantiles estimator is $\hat{\gamma} = 0.5(x_q - x_{1-q}) \tan[\pi(1-q)]$ where $x_q$ denotes the q-th sample quantile $(0.5 < q < 1)$ and the ML estimate of $\gamma$ is defined as

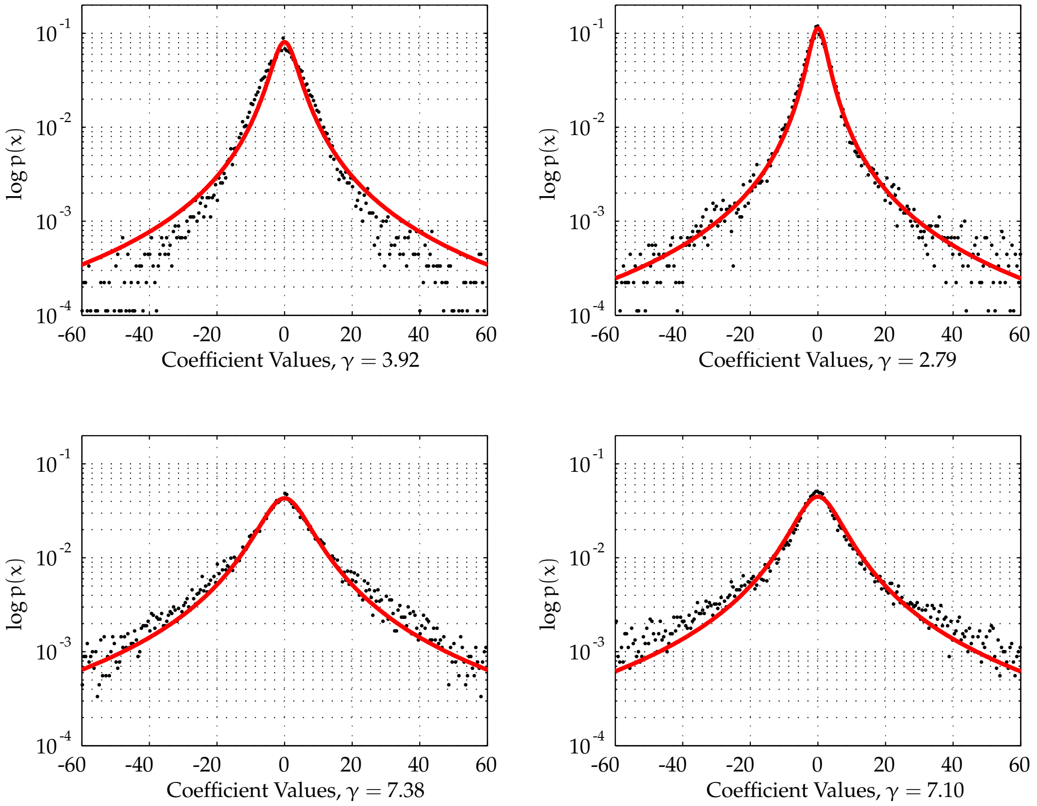


**Figure 2.6:** Log-scale histogram of DWT coefficients from the vertical detail subband of four different natural images, showing the coefficient values (black points) and the p.d.f.s of fitted Cauchy distributions (using ML estimation).

the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\frac{2}{1+(x_i/\gamma)^2}-1=0. \tag{2.7}$$

This equation has to be solved numerically, e.g. using the Newton-Raphson algorithm. The update steps can easily be derived [94]: first, we define the left-hand side of Eq. (2.7) as $g(\gamma)$ and then deduce

$$g(\gamma)' := \frac{\partial}{\partial\gamma}g(\gamma) = \frac{4\gamma}{N}\sum_{i=1}^{N}\frac{x_i^2}{(\gamma^2+x_i^2)^2}. \tag{2.8}$$

The update step follows as $\hat{\gamma}_{k+1} = \hat{\gamma}_k - g(\hat{\gamma}_k)/g'(\hat{\gamma}_k)$. A possible starting value $\hat{\gamma}_1$ is the sample quantile estimate for example. We illustrate the visual GoF by providing a series of Q-Q plots in Fig. 2.7 for the same subband coefficients we used in the previous section. The plots look almost equal to the ones shown in Fig. 2.5, again showing slight deviations in the tail regions. However, since the Q-Q plot just provides a first visual impression of the GoF, we conduct Chi-Square GoF tests on the transform coefficients of a collection of DWT decomposed test images in Section 2.2.4.

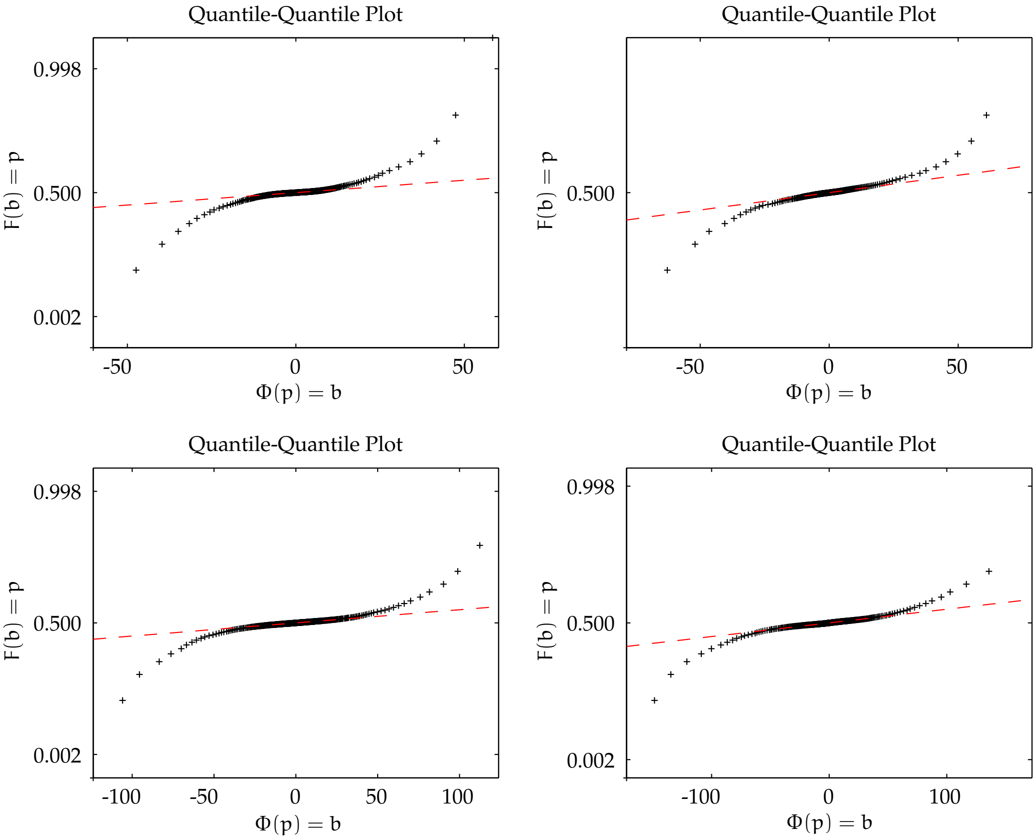


**Figure 2.7:** Exemplary Q-Q plots to visualize the GoF of the Cauchy distribution for the DWT transform coefficients of the vertical subband of four natural images.

### 2.2.3 Multivariate Power Exponential Distribution

Generally speaking, the Multivariate Power Exponential (MPE) distribution is a special case of Kotz-type distribution [132] and can be considered as a multivariate extension of the GGD [58]. Verdoolaege et al. [188] first employed this distribution as a statistical model to capture the dependencies of DWT detail subband coefficients across different color channels. In [95], we used the MPE for color image watermarking (see Section 5.3). To illustrate that it is reasonable to use a multivariate model to capture dependencies among subband coefficients of different color channels, Fig. 2.8 shows two exemplary Chi-plots for two subband combinations of the test image Lena. In case of independence, the points are supposed to lie in the central (shaded) region of the plot. Apparently, there is a quite strong dependency between the coefficients which is further confirmed by looking at the numbers for the linear correlation coefficient $r$, Spearman's $\rho$ and Kendall's $\tau$.

In consideration of the non-Gaussian nature of the DWT transform coefficients, the MPE model seems to be a good candidate to take the strong association structure into account. The

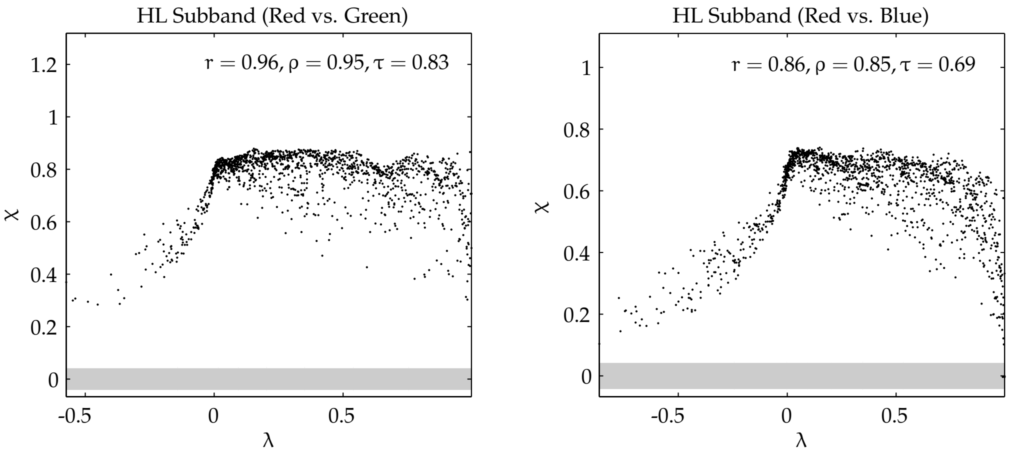


**Figure 2.8:** Exemplary Chi-plots of vertical DWT detail subband (level three of Lena) coefficients extracted from the red-green (left) and red-blue (right) color channel combination to illustrate the association among transform coefficients of equal subbands but different color channels.

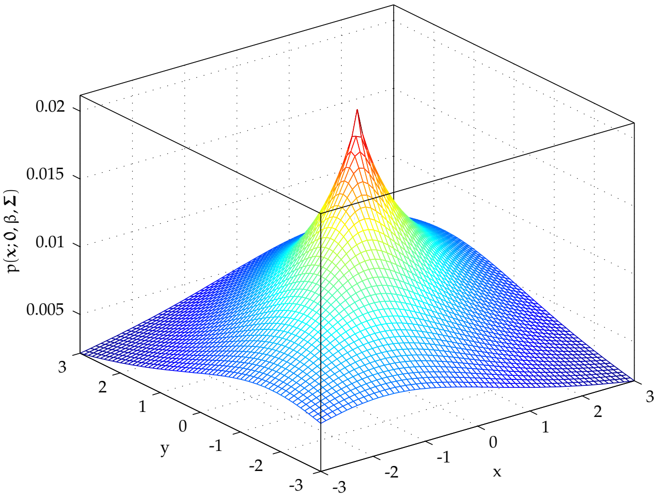


**Figure 2.9:** Exemplary p.d.f. of a MPE distribution.

p.d.f. of a $n$-variate MPE distribution is given by [58]

$$p_{X}(x; \mu, \Sigma, \beta) = \frac{n\Gamma\left(\frac{n}{2}\right)}{\pi^{\frac{n}{2}}\Gamma\left(1 + \frac{n}{2\beta}\right)2^{1+\frac{n}{2\beta}}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}\left[(x-\mu)^{T}\Sigma^{-1}(x-\mu)\right]^{\beta}\right\} \tag{2.9}$$

with $x \in \mathbb{R}^n$ and parameters $\beta > 0$ (shape), $\mu \in \mathbb{R}^n$ (location) and $\Sigma$ (positive definite symmetric $n \times n$ matrix). The p.d.f. of an exemplary bivariate MPE distribution with $\mu = \mathbf{0}$, $\beta = 0.4$ and $\Sigma = \left(\begin{smallmatrix} 1 & 0.6 \\ 0.6 & 1 \end{smallmatrix}\right)$ is shown in Fig. 2.9.

Since we only have three color channels, i.e. $n = 3$, and we can safely assume a zero location vector, we have to estimate a $3 \times 3$ matrix $\Sigma$ and the shape parameter $\beta$. Gomez et al. [58]

mention moment estimation as a suitable method, Verdoolaege et al. [188] propose a ML estimation strategy. However, the computational steps are neither listed in [58] nor [188]. In [95], we decided in favor of moment matching as a numerically stable and computationally inexpensive way. Nevertheless, we discuss both moment matching and ML estimation in the following paragraphs.

For the moment matching strategy, we match the variance and Mardia's multivariate kurtosis coefficient [193, 120] to their empirical estimates. Formally, let $\mathbf{X}$ denote a random variable following a MPE distribution with parameters $n$, $\beta$ and $\mathbf{\Sigma}$, i.e. $\mathbf{X} \sim \text{MPE}_n(\beta, \mathbf{\Sigma})$. We first determine $\hat{\beta}$ and then use this estimate to calculate $\hat{\mathbf{\Sigma}}$. Mardia's multivariate kurtosis coefficient $\gamma_2(X)$ is generally defined as

$$\gamma_2(X) = \mathbb{E}\left[\left((\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right)^2\right] - n(n+2) \tag{2.10}$$

which has a closed-form expression in case of Eq. (2.9)

$$\gamma_2(\mathbf{X}) = \frac{n^2\Gamma\left(\frac{n}{2\beta}\right)\Gamma\left(\frac{n+4}{2\beta}\right)}{\Gamma^2\left(\frac{n+2}{2\beta}\right)} - n(n+2). \tag{2.11}$$

Given an i.i.d. random sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from $\text{MPE}_n(\beta, \mathbf{\Sigma})$, we can calculate the sample version $\hat{\gamma}_2$ of $\gamma_2$ as

$$\hat{\gamma}_2(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i^T\mathbf{S}^{-1}\mathbf{x}_i\right)^2 - n(n+2), \tag{2.12}$$

where $\mathbf{S}$ denotes the classic sample covariance. By matching Eqs. (2.11) and (2.12) we can then compute the moment estimate $\hat{\beta}$[2]. Next, we can estimate $\mathbf{\Sigma}$ based on the theoretical expression for the variance $\mathbb{V}(X)$ [58]

$$\mathbb{V}(X) = \frac{2^{\frac{1}{\beta}}\Gamma\left(\frac{n+2}{2\beta}\right)}{n\Gamma\left(\frac{n}{2\beta}\right)}\mathbf{\Sigma}. \tag{2.13}$$

As we can see, $\mathbf{\Sigma}$ is proportional to the covariance matrix. To obtain $\hat{\mathbf{\Sigma}}$, we use the moment estimate $\hat{\beta}$ and the sample covariance $\mathbf{S}$ as an estimate of $\mathbb{V}(\mathbf{X})$. Then, it is straightforward to compute $\hat{\mathbf{\Sigma}}$ from Eq. (2.13).

In order to determine the ML estimates we first formulate the Likelihood equation as

$$l(\beta, \mathbf{\Sigma}; \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} \frac{\beta\Gamma(\frac{n}{2})}{\pi^{\frac{n}{2}}2^{\frac{n}{\beta}}|\mathbf{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{n}{2\beta}\right)} \exp\left\{-\frac{1}{2}\left[\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right]^{\beta}\right\}. \tag{2.14}$$

Taking the logarithm leads to

$$\begin{aligned} L(\beta, \mathbf{\Sigma}; \mathbf{x}_1, \ldots, \mathbf{x}_N) = N\log\Gamma\left(\frac{n}{2}\right) - N\log\Gamma\left(\frac{n}{2\beta}\right) + N\log(\beta) - \\ N\log\left(\pi^{\frac{n}{2}}\right) - \frac{Nn}{\beta}\log(2) - \frac{N}{2}\log(|\mathbf{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{\Sigma}^{-1}\mathbf{x})^{\beta} \end{aligned} \tag{2.15}$$

[2] In the actual implementation, we formulate moment matching as a numerical root-finding problem and then use MATLABs `fzero` function to solve it.

which can now be used to calculate the partial derivatives w.r.t. $\beta$ and $\boldsymbol{\Sigma}$ using basic algebra and matrix calculus, i.e.

$$\begin{aligned}\frac{\partial}{\partial\beta}L(\beta,\boldsymbol{\Sigma};\mathbf{x}_1,\ldots,\mathbf{x}_N) = \frac{1}{\beta}\left[N + \frac{Nn}{\beta}\left(\log(2) + \psi\left(\frac{n}{2\beta}\right)\right)\right] - \\ \frac{1}{2}\sum_{i=1}^{N}\log(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x})(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{\beta}\end{aligned} \tag{2.16}$$

and

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}L(\beta,\boldsymbol{\Sigma};\mathbf{x}_1,\ldots,\mathbf{x}_N) = -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{\beta}{2}\sum_{i=1}^{N}(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{\beta-1}\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}. \tag{2.17}$$

The solutions $\hat{\beta}$ and $\hat{\boldsymbol{\Sigma}}$ to both equations are the ML estimates. It is worth noting, that after setting the right-hand side of Eq. (2.17) to zero and performing some straightforward manipulations (i.e. multiplying two times by $\boldsymbol{\Sigma}$) we obtain

$$\boldsymbol{\Sigma} = \frac{\beta}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\left(\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right)^{\beta-1} \tag{2.18}$$

which allows to employ a fix-point iteration directly (e.g. Picard Iteration aka successive substitution). Since it is hard to prove that Eq. (2.18) actually is a contraction – which would guarantee convergence to the fixpoint – we follow an alternative technique to obtain the estimates. We directly try to minimize the negative Log-Likelihood, i.e. $-L(\beta,\boldsymbol{\Sigma};\mathbf{x}_1,\ldots,\mathbf{x}_N)$, using a gradient descent approach. This is an optimization problem with non-linear constraints, since we have to satisfy the requirements that $\boldsymbol{\Sigma}$ must be positive definite and symmetric and $\beta > 0$. We already have the derivatives of the log-likelihood function w.r.t. $\beta$ and $\boldsymbol{\Sigma}$, see Eq. (2.16) and (2.17). To take care of the positive definiteness criteria, we use the Sylvester criterion [126] which requires that all leading principal minors of $\boldsymbol{\Sigma}$ are positive. This is a necessary and sufficient condition to guarantee positive definiteness. Eventually, we have $(n+1)/2-1$ unknowns to solve[3] (since $\boldsymbol{\Sigma}$ is symmetric).

### 2.2.4 Quantifying the Goodness-of-Fit

In order to quantify the GoF of the presented GGD and Cauchy model, we conduct a series of Chi-Square GoF tests using the images of the UCID database. Each RGB channel is decomposed separately by a three-scale DWT and the test statistic is computed using the transform coefficients of each detail subband. In contrast to the Chi-Square tests we conducted in [101], we slightly modify the test setup here to account for different sample sizes on each decomposition level. The problem with the test in [101] can be formulated as follows: first, the type of GoF test setup we use here can be termed an *Accept-Support* testing setup. This means that the null-hypothesis represents what we actually believe (i.e. the observations stem from the distribution we assume). Second, we know that increasing the sample size likewise increases the power of a hypothesis test. Hence, if the sample size is *too large*, we will inevitably decide against the null-hypothesis even in cases when the model represents a good fit to the data. This happens because even minor deviations from the null-hypothesis are rigorously penalized in

[3] In the actual implementation of this estimation procedures, the non-linear optimization problem with non-linear constraints is solved by means of MATLAB's `fminbnd` routine.

| Database | Level | Model | |
|---|---|---|---|
| | | GGD | Cauchy |
| UCID | 1 | 36.64 | 62.04 |
| | 2 | 35.14 | 62.55 |
| | 3 | 34.73 | 71.62 |
| Stex | 1 | 0.82 | 41.63 |
| | 2 | 1.23 | 43.91 |
| | 3 | 2.34 | 32.79 |
| Vistex (small) | 1 | 0.57 | 44.64 |
| | 2 | 0.94 | 43.96 |
| | 3 | 2.34 | 34.74 |
| Vistex (full) | 1 | 0.87 | 42.68 |
| | 2 | 1.20 | 36.33 |
| | 3 | 2.26 | 28.60 |
| Outex | 1 | 1.76 | 70.95 |
| | 2 | 0.60 | 55.24 |
| | 3 | 0.76 | 32.06 |

**Table 2.1:** Percentage of rejected null-hypotheses of Chi-Square GoF tests (at 5% significance), averaged over all subbands of a DWT decomposition level.

case of large sample size. Due to subsampling, the number of DWT coefficients on successive scales differs by a factor of four. Hence, we have 16 times more coefficients on level one as we have on level three for example. The aforementioned sample size effect on the test power would therefore inevitably lead to more rejections of the null-hypothesis at lower decomposition levels. In order to deal with that problem, we modify the GoF setup such that we limit the sample size to N samples, randomly selected from each subband. In detail, we use uniform sampling without replacement. The percentage of rejected null-hypotheses on each DWT decomposition level for the GGD and Cauchy model is listed in Table 2.1 using $N = 500$. As expected, the GGD is a quite good model for the coefficients of DWT decomposed images. The Cauchy distribution on the other hand leads to higher rejection rates, however, we emphasize that this model is only supposed to be a better approximation to the coefficients than the Gaussian model. In contrast to [101], we further notice that the rejection rates are now rather stable over the decomposition levels.

### Testing the GoF of the MPE Distribution

To the best of our knowledge, there exists no published GoF test for the MPE distribution, although Gomez et al. [58] sketch a possible test strategy. We first discuss this idea and then introduce a novel GoF test which is based on a generic test for multivariate normality. The approach proposed by Gomez et al. is a three-stage strategy which relies on the stochastic representation of the MPE distribution. Unfortunately, no clear description of how to perform the three stages is given by the authors. In the following, we discuss a possible implementation of the test. We know that in case $\mathbf{X} \sim \text{MPE}_n(\mathbf{x}; \beta, \mathbf{\Sigma})$, then

$$\mathbf{X} \sim r\mathbf{A}^T\mathbf{u} \tag{2.19}$$

where $r$ is a realization of the random variable $R \sim f_R(r;\beta)$ with p.d.f.

$$f_R(r;\beta) = \frac{n}{\Gamma\left(1+\frac{n}{2\beta}\right)2^{\frac{n}{2\beta}}} r^{n-1} \exp\left\{-\frac{1}{2}r^{2\beta}\right\} \mathbf{1}_{(0,\infty)}(r). \tag{2.20}$$

The vector $\mathbf{u} \in \mathbb{R}^n$ is uniformly distributed on the unit sphere and $\mathbf{A}$ is a lower triangular matrix such that $\mathbf{\Sigma} = \mathbf{A}^T\mathbf{A}$. Based on this stochastic representation of the MPE distribution and the moments of $R$ [58], the first step of the GoF procedure is to test whether

$$Z = \left((\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{\beta} \tag{2.21}$$

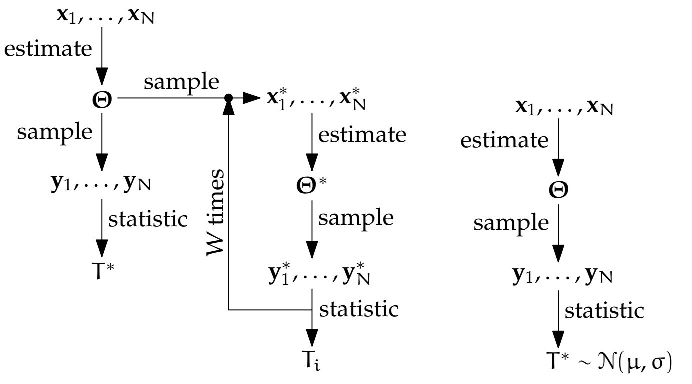
follows a Gamma distribution [89] with shape parameter 2 and scale parameter $n/2\beta$. This can easily be accomplished by means of a Chi-Square GoF test. In the second step, we have to test whether

$$\mathbf{u} = \frac{\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{x}}{\left\|\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{x})\right\|} \tag{2.22}$$

is uniformly distributed on the unit sphere in $\mathbb{R}^n$ (in fact, it is the unit ball in the $n$-dimensional Euclidean space). We perform this task by means of a Rayleigh test for uniformity on the sphere, originally proposed by Mardia and Rupp [121]. In the last step, we test if the random variable $R$ is independent of $\mathbf{u}$. For that purpose, we employ a very recently proposed test by Gretton et al. [60]. Probably the most crucial step is the fusion of the three test results. We choose the rather strict strategy to reject the overall null-hypothesis, in case one test shows evidence against its null-hypothesis. At the end of this section, we will assess the size and power of this test. To the best of our knowledge, no such study has been conducted so far.

As a second, novel alternative to assess the GoF of the MPE distribution, we propose a modification of the GoF test for a multivariate normality proposed by Smith & Jain [170]. The components of the test procedure are outlined in Fig. 2.10. The left part shows the Monte-Carlo variant of the test which is based on an estimate of the p-value. The right part shows the second variant which relies on the asymptotic distribution of the test statistic under the null-hypothesis. In [170], the null-hypothesis is that the observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are drawn from a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ with parameter vector $\mathbf{\Theta} = [\boldsymbol{\mu}\ \mathbf{\Sigma}]$. Consequently, the null-hypothesis of our MPE GoF test is that the data is drawn from a MPE distribution $\text{MPE}_n(\beta, \mathbf{\Sigma})$ with parameters $\beta$ and $\mathbf{\Sigma}$, hence $\mathbf{\Theta} = [\beta\ \mathbf{\Sigma}]$. According to Fig. 2.10, the critical parts of the GoF test are the estimation part, the sampling part and the computation of a suitable test statistic. Estimation and sampling in the multivariate Gaussian case is straightforward and a well covered topic in literature. Estimation of the MPE parameters has already been discussed in Section 2.2.3. Hence, the remaining parts are the sampling step in case of the MPE distribution and the definition of a test statistic. Both topics are covered next:

**Sampling from a MPE distribution** We can rely on the stochastic representation of the MPE distribution, given in Eq. (2.19). For our purpose, we assume $\boldsymbol{\mu} = \mathbf{0}$. In order to generate a random sample from a MPE distribution $\text{MPE}_n(\beta, \mathbf{\Sigma})$ we have to draw a random sample $\mathbf{u}_1, \ldots, \mathbf{u}_N$ from a uniform distribution on the $n$-dimensional unit sphere first. We



**(a)** "Monte-Carlo" Test **(b)** "Normal" Test

**Figure 2.10:** Outline of the generic GoF test setup proposed by Smith & Jain [170], originally intended to test for multivariate normality.

then perform a Cholesky decomposition of $\mathbf{\Sigma}$ to obtain $\mathbf{A}^T$ and generate another random sample $r_1, \dots, r_N$ from the distribution given by the p.d.f. in Eq. (2.20). Eventually, we use

$$\forall i, 0 < i < N : \mathbf{x}_i = r_i \mathbf{A}^T \mathbf{u}_i \tag{2.23}$$

to generate a MPE random sample $\mathbf{x}_1, \dots, \mathbf{x}_N$ of size N. To obtain $\mathbf{u}_1, \dots, \mathbf{u}_N$, several ways are possible. We choose the simple strategy of generating a random vector $\mathbf{u}_i$ from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$ and then normalize each element of the vector by $(\sum_j u_{ij}^2)^{1/2}$. Due to the radial symmetry of the multivariate Gaussian distribution, this gives a random vector which is uniformly distributed on the unit sphere in the $n$-dimensional Euclidean space. The process of generating the random sample $r_1, \dots, r_N$ is slightly more involved. In order to use the classic inversion method, we first need to determine the quantile function $F_R^{-1}$ (i.e. the inverse c.d.f.) corresponding to the p.d.f. given in Eq. (2.20). First, we derive the c.d.f. as

$$F_R(y; \beta) = \int_0^y f_R(x; \beta) dx = 1 - \frac{\Gamma\left(\frac{n}{2\beta}, \frac{y^{2\beta}}{2}\right)}{\Gamma\left(\frac{n}{2\beta}\right)}. \tag{2.24}$$

Inverting the c.d.f. gives the desired result

$$F_R^{-1}(u; \beta) = 2^{\frac{1}{2\beta}} \left[ P_u^{-1}\left(\frac{n}{2\beta}, 1 - u\right)\right]^{\frac{1}{2\beta}} \tag{2.25}$$

where $P_u(a, x)$ is defined as in Eq. (2.3). We can then generate $r_i$ by using $r_i = F_R^{-1}(u_i; \beta)$ with $u_i \sim \mathcal{U}(0, 1)$.

**Defining a suitable test statistic** In [170], Smith & Jain propose to test for multivariate normality by first computing the *Euclidean Minimum Spanning Tree (EMST)* of the pooled sample

$$\forall i, 0 < i \leqslant 2N : \mathbf{z}_i = \begin{cases} \mathbf{x}_i, & 0 < i \leqslant N \\ \mathbf{y}_i, & N < i \leqslant 2N \end{cases} \tag{2.26}$$

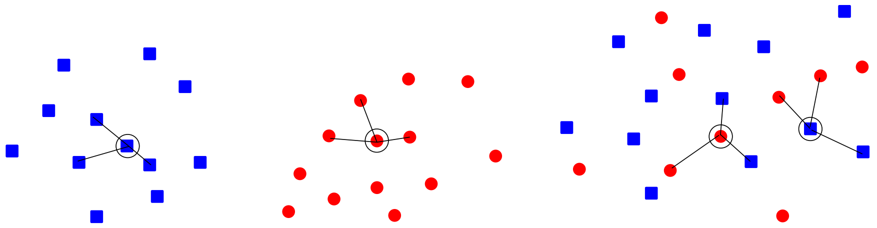

**(a)** $H \neq G, T = 6$ **(b)** $H = G, T = 2$

**Figure 2.11:** Illustration of the two-sample hypothesis test proposed by Henze [68] based on the number of nearest neighbor coincidences. In case the samples stem from the same population (i.e. $H = G$), we expect the test statistic $T$ to be low, while in case the samples stem from different populations (i.e $H \neq G$) we expect the test statistic to be high.

The sample $x_1, \ldots, x_N$ denotes the collection of original observations, whereas the sample $\mathbf{y}_i, \ldots, y_N$ is drawn from a multivariate Gaussian distribution with parameters fitted on the basis of $x_i$. The test statistic $T$ is defined as the number of edges connecting vertices from different samples. This idea was first introduced by Friedman & Rafsky [49] in the field of multivariate two-sample hypothesis testing, where the objective is to quantify whether two samples stem from the same population without making any assumptions about the distribution family. In the same context, a similar strategy is suggested by Henze [68], based on the computation of the number of nearest neighbor coincidences. A graphical visualization of the NN coincidences idea is shown in Fig. 2.11, where $H$ signifies the distribution of the first sample (marked as blue squares) and $G$ signifies the distribution of the second sample (marked as red discs). The value of the test statistic when we only consider two elements of each sample is given as $T$. The basic idea of the EMST and NN coincidences approach is the same: given that the null-hypothesis is true, we (i) expect that the number of EMST edges connecting vertices from different samples to be high and (ii) the number of nearest neighbor coincidences to be low.

When using the EMST approach for testing multivariate normality as in [170], we consequently expect high values of $T$ in case the observations $x_i$ actually follow a multivariate Gaussian distribution and vice versa. From Friedman & Rafsky [49], we know that in case the null-hypothesis is true, the test statistic $T$ follows a Gaussian distribution with mean $\mu$ and standard deviation $\sigma$. Hence, it is straightforward to compute a $p$-value and reject the null-hypothesis if the p-value is less than the fixed significance level $\alpha$. However, it is worth noting that the sampling procedure to generate $\mathbf{y}_i$ introduces bias, because sampling is based on the distribution parameters fitted on the basis of $x_i$. Since the EMST and NN coincidences approach rely on the assumption of independent random samples, the resulting GoF tests will inevitably loose power. A reasonable way to circumvent the independency problem is to estimate the critical region of the test using a Monte-Carlo approach, illustrated in Fig. 2.10a. The iteration in the right branch of Fig. 2.10a is repeated $W$ times and the $p$-value estimate is finally obtained by

$$\hat{p} = \frac{\#\{T_i \geqslant T^*\} + 0.5}{W + 1}. \tag{2.27}$$

To construct a GoF test for the MPE distribution similar to the one of Smith & Jain, we use (i) the gradient decent approach of minimizing the negative Log-Likelihood to estimate the MPE parameters, (ii) the MPE sampling procedure outlined above and (iii) the NN coincidences approach of [68] to obtain a test statistic T. To provide full detail, let $\mathbf{z}_1, \dots, \mathbf{z}_M$ denote the pooled sample (i.e. $M := 2N$); further, let $m$ denote a function returning the sample membership of $\mathbf{z}_i$ and let $NN_i(r)$ denote the r-th nearest neighbor of $\mathbf{z}_i$ (in the Euclidean norm). Then, the formal description of the NN coincidences test statistic is given by

$$T_{k,M} = \frac{1}{Mk} \sum_{i=1}^{M} \sum_{r=1}^{k} \mathbf{1}_i(r) \tag{2.28}$$

where $\mathbf{1}_i(r)$ denotes the indicator function of the event that $m(\mathbf{z}_i) = m(NN_i(r))$. According to Schilling [160], we have the asymptotic (i.e. $M \to \infty$) result that in case the null-hypothesis (denoted by $\mathcal{H}_0$) is true, the term

$$\sqrt{Mk} \left( \frac{T_{k,n} - \mu_{T_{k,M}|\mathcal{H}_0}}{\sigma_{T_{k,M}|\mathcal{H}_0}} \right) \sim \mathcal{N}(0,1) \tag{2.29}$$

follows a standard normal distribution with

$$\mu_{T_{k,M}|\mathcal{H}_0} = \lambda_1^2 + \lambda_2^2, \quad \sigma^2_{T_{k,M}|\mathcal{H}_0} = \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2 \left[ 1 - \binom{2k}{k} 2^{-2k} \right] \tag{2.30}$$

and $\lambda_i = N/M$ (i.e. in our case $\lambda_1 = \lambda_2 = 0.5$). By using Eqs. (2.29) and (2.30), the $p$-value can be calculated by determining $\mathbb{P}(T^* \geqslant T|\mathcal{H}_0)$, i.e. the probability of obtaining a test statistic at least as extreme as $T^*$. Adhering to the terminology of Smith & Jain, we denote the test variant based on the Monte-Carlo $p$-value estimation approach as the "Monte-Carlo" test and the second variant, based on the asymptotic normality of T, as the "Normal" test.

In order to assess the quality of the proposed GoF test and the test suggested by Gomez et al., we conduct a study on the size, i.e. the test's probability of falsely rejecting the null-hypothesis, and power of the test. Regarding the methodology, both size and power are evaluated by means of a Monte-Carlo strategy with $M = 500$ iterations for the case $n = 3$ (i.e. three-dimensional observations).

**Size Study** In each Monte-Carlo iteration, we sample N points from a $\text{MPE}_3(0.5, \mathbf{I})$ distribution and determine the percentage of rejected null-hypotheses. We let the sample size N be 200, 400 and 800. Since we do not obtain an overall $p$-value in case of the GoF test of Gomez et al., we have to decide when to reject the null-hypothesis based on the outcomes of the three stages. As mentioned before, we choose the strict way of rejecting the null-hypothesis in case just one stage rejects its own null-hypothesis. Formally, given that $H_i, i = 1, 2, 3$ denotes the outcome of stage $i$ (i.e. $H_i \in \{0, 1\}$), we reject the null-hypothesis if $\sum_i H_i > 0$. Regarding the "Monte-Carlo" variant of our proposed GoF test, we set the number of iterations $W$ to 1000. Tables 2.2 and 2.3 list the estimated significance level $\hat{\alpha}$ for different sample sizes. For the Gomez et al. GoF test, we observe that the estimated percentage of rejections $\hat{\alpha}$ is above the desired significance level $\alpha$ in all cases. Regarding the two variants of our proposed GoF approach, we can see that the "Monte-Carlo" test is quite conservative, i.e. the percentage of false positives is always below the fixed significance level. However, in case of the "Normal" test, the situation is different. Except for $N = 400$, the rejection rates are always slightly above the desired level.

| **Significance** | **Sample Size** N | $\sum_i H_i > 0$ |
|---|---|---|
| | | $\hat{\alpha}$ |
| $\alpha = 0.01$ | 200 | 0.030 |
| | 400 | 0.028 |
| | 800 | 0.014 |
| $\alpha = 0.05$ | 200 | 0.084 |
| | 400 | 0.118 |
| | 800 | 0.108 |
| $\alpha = 0.10$ | 200 | 0.194 |
| | 400 | 0.212 |
| | 800 | 0.196 |

**Table 2.2:** Rejection rates for the three-stage GoF test sketched by Gomez et al. in [58] for various levels of $\alpha$ and various sample sizes N.

| **Significance** | **Sample Size** N | **“Monte-Carlo”** | **"Normal"** |
|---|---|---|---|
| | | $\hat{\alpha}$ | $\hat{\alpha}$ |
| $\alpha = 0.01$ | 200 | 0.002 | 0.022 |
| | 400 | 0.001 | 0.002 |
| | 800 | 0.001 | 0.018 |
| $\alpha = 0.05$ | 200 | 0.022 | 0.063 |
| | 400 | 0.012 | 0.014 |
| | 800 | 0.053 | 0.069 |
| $\alpha = 0.10$ | 200 | 0.044 | 0.132 |
| | 400 | 0.026 | 0.048 |
| | 800 | 0.084 | 0.1520 |

**Table 2.3:** Rejection rates for the two variants of the proposed MPE GoF test for various levels of $\alpha$ and different sample sizes N.

**Power Study** To assess the power of the GoF tests, we sample from a two-component mixture of MPE distributions. Given that $p(\mathbf{x}; \beta_i, \boldsymbol{\Sigma}_i) := \text{MPE}_3(\mathbf{x}; \beta_i, \boldsymbol{\Sigma}_i)$, the mixture p.d.f. is given by

$$p(\mathbf{x}; \pi_1, \pi_2, \beta_1, \beta_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \sum_{i=1}^{2} \pi_i p(\mathbf{x}; \beta_i, \boldsymbol{\Sigma}_i) \quad \text{with} \quad \sum_i \pi_i = 1. \tag{2.31}$$

We start from an equal parameters $\beta_1 = \beta_2 = 0.5, \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and then move the shape parameter $\beta_2$ of the second mixture component away from the original choice, as illustrated in Fig. 2.12. The component weights are set to $\pi_1 = \pi_2 = 0.5$. For each parameter setting along the line we perform M Monte-Carlo iterations for each sample size $N \in \{200, 400, 800\}$ and determine the number of rejected null-hypotheses. Figures 2.13 and 2.14 show the corresponding power plots, where the x-axis shows the shape parameter value of $\beta_2$ and the y-axis shows the percentage of rejected null-hypotheses. In case

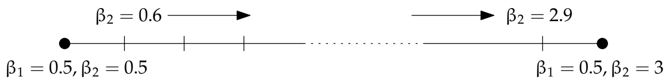


**Figure 2.12:** Illustration of the power study procedure for scale alternatives. The starting model is a mixture of two MPE distributions with $\beta_1 = \beta_2 = 0.5$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$ and equal weights $\pi_1 = \pi_2 = 0.5$. As we progress from left to right, the shape parameter $\beta_2$ of the second mixture component is increased by a stepsize of 0.1.

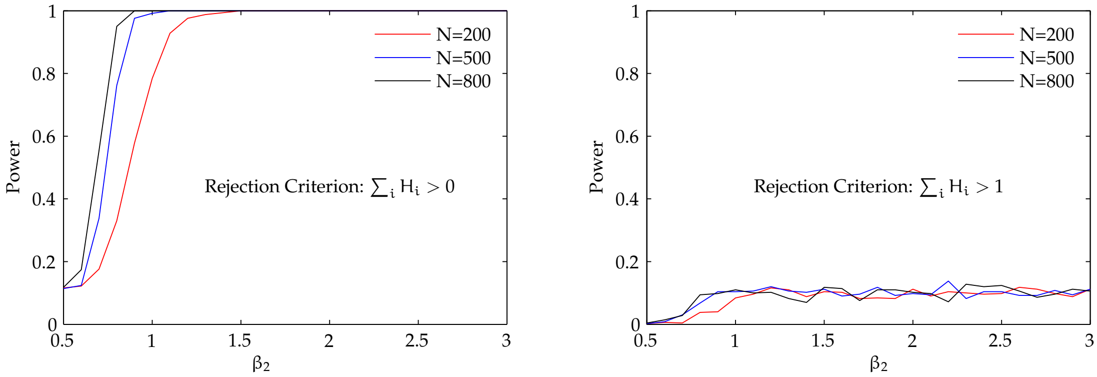


**Figure 2.13:** Power vs. $\beta_2$ for two choices of how to combine the three-stages of the Gomez et al. GoF test. The plot on the left-hand side shows the results of the rejection criterion we select for our tests.

of the GoF test of Gomez et al. (see Fig. 2.13), we observe that our fusion strategy of the three stages leads to reasonable power, even at moderate sample size, i.e. $N = 200$. For comparative reasons, we additionally show a power plot for the case of requiring evidence against the null-hypothesis in at least two of the three stages. In this case, the test exhibits almost no power at all and renders this setting useless. Regarding the two variants of our proposed GoF test, both exhibit reasonable power with the "Normal" test showing high power even at moderate sample size. The higher power can be explained by referring to Table 2.3, where the "Normal" test exhibits less conservative behavior than the "Monte-Carlo" test.

After completing the size and power study, we finally turn to the actual application of the GoF test. We apply the test to the DWT detail subband coefficients of our database images. To obtain the same power for each DWT decomposition level, we uniformly sample 500 coefficients from each subband and set the significance level to $\alpha = 0.05$. We choose the "Normal" GoF test variant in all cases. In addition to the estimation of both MPE parameters, we test against the fix choice of $\beta = 1$, i.e. multivariate Gaussian, for comparative reasons. The rejection rates are listed in Table 2.4. Apparently, the MPE distribution is a quite good model for textured images and slightly worse for natural images. However, compared to the GoF results for $\beta = 1$, the MPE distribution is definitely the more suitable statistical model to capture the non-Gaussian nature of the coefficients.

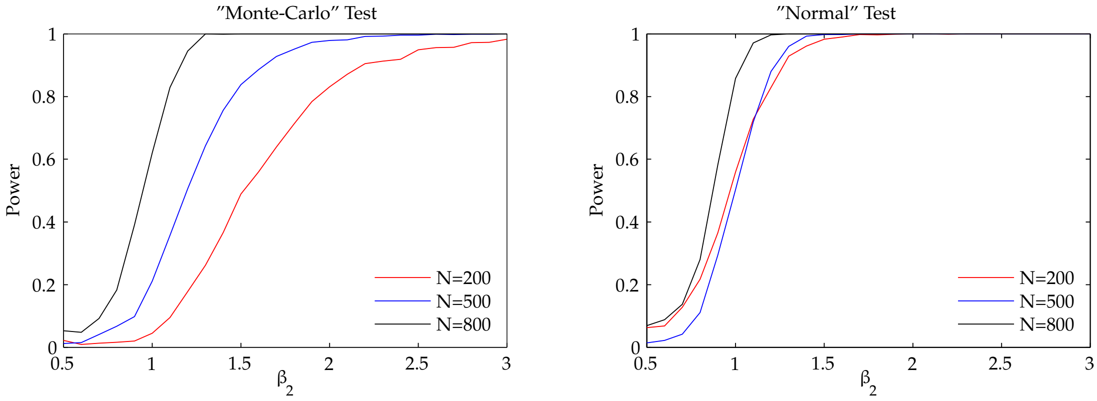


**Figure 2.14:** Power vs. $\beta_2$ for the two variants of the proposed GoF test, i.e. using either the Monte-Carlo approach to approximate the critical region or the Normal approximation of the test statistic.

| **Model** | **Databases** | | | |
|---|---|---|---|---|
| | Stex | Vistex (full) | Outex | UCID |
| MPE | 25.09 | 35.13 | 11.15 | 56.18 |
| Gaussian ($\beta = 1$) | 57.13 | 73.19 | 39.66 | 98.97 |

**Table 2.4:** Rejection rates of the MPE "Normal" GoF test at 5% significance for several image databases. The second row lists the GoF test results of the same test when we fix the shape parameter to $\beta = 1$, i.e. multivariate Gaussian.

## 2.3 Complex Wavelet Transform Subband Models

Since two major parts of this thesis, namely Chapters 3 and 4 are concerned with image analysis applications, we select the Dual-Tree Complex Wavelet transform [85, 86] (DTCWT) as a second wavelet transform variant due to its advantages over the DWT. In particular, the DTCWT overcomes two shortcomings of the DWT: lack of shift-invariance and lack of directional selectivity, as it is vividly illustrated and explained in [86] or [162]. These shortcomings are especially relevant for image analysis purposes. Lack of shift-invariance implies that singularities at different locations in an image lead to different representations in the wavelet domain (i.e. different coefficients). Hence, wavelet coefficients representing an edge along an object contour for example, are not necessarily large across all scales which causes ringing artifacts when reconstruction is performed using only a subset of the coefficients. Of course, the perfect reconstruction property guarantees that all artifacts are canceled when computing the reconstruction using all coefficients. The technical reason for the shift-dependency problem is that the wavelet and scaling filters which are used to implement the DWT have finite support and the coefficients are downsampled by two after each decomposition stage. As a matter of fact, shift-dependency is a severe deficiency in the context of image analysis. The second shortcoming – lack of directionally selectivity – is related to the fact that the filters of the DWT are real functions and are thus supported on both sides of the frequency axis. Since the 2-D DWT is usually implemented by

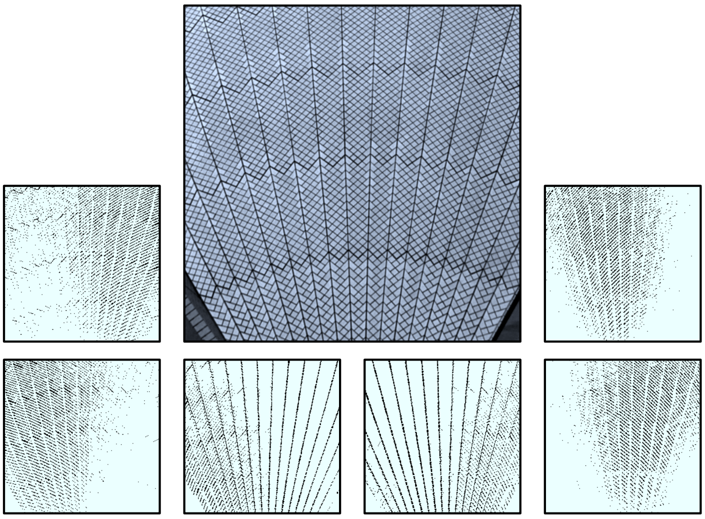


**Figure 2.15:** Exemplary texture image Tile.0000 including magnitude images of the six DTCWT detail subbands (±15°, ±45° and ±75° in counter clockwise order).

separate row- and column filtering (which is equivalent to using tensor-product wavelets), this causes ambiguities in distinguishing features oriented along ±45°. All other features oriented mostly along the vertical or horizontal direction are lumped in the vertical and horizontal detail subbands. Since orientation information can be an important characteristic for many texture images for example, better directional selectivity is desired. Both deficiencies are eliminated to a certain extent by using the DTCWT, at low computational overhead. The basic idea is to use complex wavelets which are composed of two real wavelets forming an approximate Hilbert transform pair. Since this construction ensures that negative frequencies are suppressed, aliasing effects are reduced and thus approximate shift-invariance is guaranteed. Further, a higher degree of directional selectivity is achieved with six complex detail subbands at each decomposition stage (compared to three in case of the DWT). The detail subbands are oriented along approximately ±15°, ±45° and ±75°. An exemplary texture image (Tile.0000 [31]) and the six magnitude images of the detail subbands at the first scale are shown in Fig. 2.15. To emphasize the image details captured by each subband, all coefficients with absolute values below the 0.9 quantile are set to zero.

In the following, we consider statistical models $p_X$ for the coefficient magnitudes $|x_i|$ of the DTCWT detail subbands and adhere to all three assumptions of Section 2.2. We then discard Assumptions 1 and 2 and introduce a joint statistical model which is flexible enough to even capture the association among coefficient of different color channels. A first, straightforward approach for modeling the detail subband coefficient magnitudes is proposed by Shaffrey et al. [163]. The authors employ the Rayleigh distribution together with Hidden Markov Trees (HMT) for the purpose of image segmentation. The theoretical reasoning of this model is that in case the real and imaginary part of a coefficient follow a zero-mean Gaussian distribution with equal variance $\sigma^2$, it is a well known fact that the magnitude follows a Rayleigh distribution with shape parameter $\beta := \sigma$. The p.d.f. of a Rayleigh distribution is given by [89]

$$p_X(x;\beta) = \frac{x}{\beta^2}\exp\left(-\frac{x^2}{2\beta^2}\right), \quad 0 < x < \infty \tag{2.32}$$

with $\beta > 0$. In Fig. 2.16 we illustrate the shape of the Rayleigh p.d.f. and the characteristic coefficient histograms which can be observed in case of texture images. ML parameter estimation of $\beta$ has a closed-form solution which can be found in [89]. In a very recent work, Rahman et al. [151] studied the statistics of DTCWT detail subband coefficients restricted to the decomposition of Gaussian distributed signals. The authors show that the real and imaginary part can actually be modeled by zero-mean Gaussian distributions for decomposition levels greater than one and hence allow to employ the Rayleigh model for the magnitudes. On the first level, however, they propose to use a Generalized Gamma distribution [174] instead. The reason for switching the statistical models is that on the first level of the DTCWT it is necessary to use different filter sets (e.g., see [162]) which violate the Hilbert transform property. As a result, the real and imaginary parts no longer show equal variances and hence prevent to employ the Rayleigh distribution to model the magnitudes. Although the results presented in [151] are theoretically interesting, they lack practical application, since we rarely observe a Gaussian distributed signal in image processing. The effect of the deviation from Gaussianity is apparent by looking at the bad fit of the Rayleigh model in Fig. 2.16. The idea of using a Generalized Gamma distribution to model the coefficient magnitudes is a good starting point, though. Apparently, candidate models are positively skewed distributions (i.e. skewed to the right) which are often used in reliability and life-span modeling [23]. Similar distributions are also employed in modeling the amplitude statistics of Synthetic Aperture Radar (SAR) data (e.g. see [130, 93]). The use of the Generalized Gamma distribution, however, is not widespread due to the difficulties in parameter estimation (e.g., see [172]). In the following, we present two reasonable statistical models which are both special cases of the Generalized Gamma distribution, allowing computationally efficient parameter estimation. Further, we show that the models are flexible enough to capture the magnitude distributions.

### 2.3.1 Weibull Distribution

The first model we consider is the two-parameter Weibull distribution which includes the Rayleigh distribution as a special case. This model is a reasonable choice since there are more degrees of freedom to adapt to the underlying data. In [98], we exploited the Weibull distribution parameters for the purpose of medical image classification and in [101, 61] this model was successfully employed in texture image retrieval. The p.d.f. and c.d.f. of a Weibull distribution, as given in [89], are

$$p_X(x;\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\}, \quad 0 < x < \infty \tag{2.33}$$

and

$$F_X(x;\alpha,\beta) = 1 - \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\} \tag{2.34}$$

with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. For $\alpha = 2$ and $\beta = \sqrt{2}\beta$, Eq. (2.33) reduces to the Rayleigh distribution. The inverse c.d.f. has the closed form expression $F^{-1}(u;\alpha,\beta) = \beta[-\log(1-u)]^{1/\alpha}$. Regarding parameter estimation of $\alpha$ and $\beta$, we discuss both moment matching and ML estimation. First, lets assume that we have an i.i.d. random sample

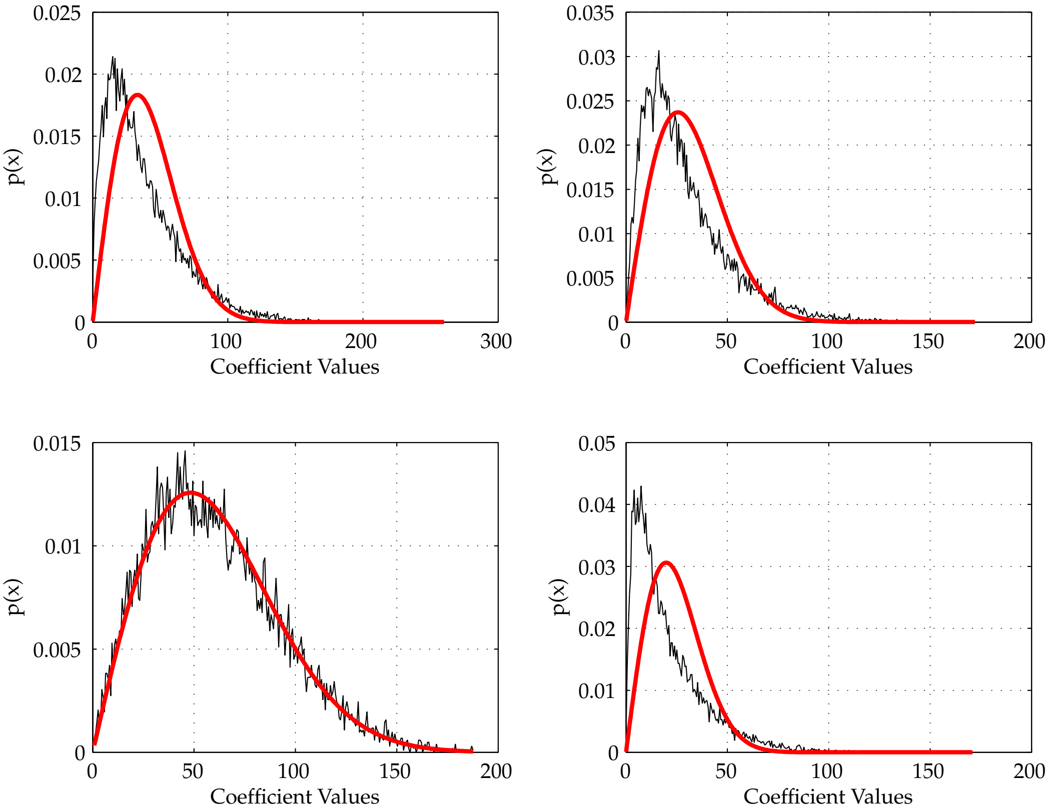


**Figure 2.16:** Exemplary DTCWT coefficient histograms (i.e. $|x_i|$) of the +75° subband on DTCWT level two of four texture images together with fitted Rayleigh p.d.f.s.

$x_1, \ldots, x_N$ drawn from a two-parameter Weibull distribution. According to [89], the MLE of $\alpha$ is the solution to $g(\alpha) = 0$ with

$$g(\alpha) := \sum_{i=1}^{N} x_i^{\alpha} \log(x_i) - K \sum_{i=1}^{N} x_i^{\alpha} - \frac{1}{\alpha} \sum_{i=1}^{N} x_i^{\alpha} \tag{2.35}$$

and $K := \frac{1}{N} \sum_{i=1}^{N} \log(x_i)$. In order to solve Eq. (2.35) using Newton-Raphson root finding, we first determine the first derivative $g'(\alpha)$ as

$$\begin{gathered} g'(\alpha) := \frac{\partial}{\partial \alpha} g(\alpha) = \sum_{i=1}^{N} x_i^{\alpha} \log(x_i)^2 - \\ K \left( \sum_{i=1}^{N} x_i^{\alpha} \log(x_i) \right) + \frac{1}{\alpha^2} \sum_{i=1}^{N} x_i^{\alpha} - \frac{1}{\alpha} \sum_{i=1}^{N} x_i^{\alpha} \log(x_i). \end{gathered} \tag{2.36}$$

The MLE is then obtained by using the update step $\hat{\alpha}_n = \hat{\alpha}_{n-1} - g(\hat{\alpha}_{n-1})/g'(\hat{\alpha}_{n-1})$ for $n \geqslant 2$. Subsequently, the MLE of $\beta$ has the explicit expression:

$$\hat{\beta} = \left( \frac{1}{N} \sum_{i=1}^{N} x_i^{\hat{\alpha}} \right)^{1/\hat{\alpha}} \tag{2.37}$$

The starting value $\hat{\alpha}_1$ is usually computed by moment matching. Unfortunately, even that requires a numerical procedure, since the moment parameter estimate $\hat{\alpha}$ is the solution to [23]

$$\frac{\Gamma_3 - 3\Gamma_2\Gamma_1 + 2\Gamma_1^3}{\left(\Gamma_2 - \Gamma_1^2\right)^{3/2}} - a_3 = 0, \tag{2.38}$$

where $\Gamma_k := \Gamma(1 + k/\alpha)$ and

$$a_3 := \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{3/2}} \tag{2.39}$$

denotes the sample skewness. A first approximation of $\hat{\alpha}$ to solve Eq. (2.39) can be obtained from a $\alpha$-versus-$a_3$ lookup-table and linear interpolation. The moment estimate of $\hat{\beta}$ is then computed by

$$\hat{\beta} = \frac{s}{\left(\hat{\Gamma}_2 - \hat{\Gamma}_1^2\right)^{\frac{1}{2}}} \tag{2.40}$$

where $s$ denotes the sample standard deviation and $\hat{\Gamma}_k$ signifies that we use the moment estimate $\hat{\alpha}$ to compute $\Gamma_1$ and $\Gamma_2$. Finally, it is worth noting that computational difficulties can arise for ML estimation in cases where $\alpha < 2.2$ [23].

We next present an alternative estimation method which is computationally more attractive than the direct ML estimation approach from above. This estimation strategy is based on the theoretical result, that if a random variable $X$ follows a Weibull distribution, then the random variable $Y = \log(X)$ follows an Extreme Value (EV) distribution of type I (i.e. Gumbel distribution) [118]. This result can easily be verified by exploiting the fact that the random variable transformation $t(X)$ is the natural logarithm which is monotonically increasing, continuous and differentiable. Hence, we have $F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(X \leqslant t^{-1}(x)) = F_X(t^{-1}(y))$ which in our case (i.e. $t(x) = \log(x)$ and $t^{-1}(x) = \exp(x)$) leads to

$$\begin{aligned} F_Y(y) &= F_X(\exp(y)) && (2.41)\\ &= 1 - \exp\left\{ - \left[ \frac{\exp(y)}{\beta} \right]^{\alpha} \right\} && (2.42)\\ &= 1 - \exp\left\{ - \left[ \exp\left\{ \frac{y-\mu}{\sigma} \right\} \right] \right\} && (2.43) \end{aligned}$$

using the substitution $\sigma := 1/\alpha$ and $\mu := \log(\beta)$ in Eq. (2.43). The last expression in this derivation is the c.d.f. of a Gumbel distribution. Given that we set $y_i := \log(x_i)$, the corresponding p.d.f. follows as

$$p_Y(y; \mu, \sigma) = \frac{1}{\sigma} \exp\left( \frac{y-\mu}{\sigma} \right) \exp\left\{ - \exp\left( \frac{y-\mu}{\sigma} \right) \right\}, \quad -\infty < y < \infty \tag{2.44}$$

with location parameter $0 < \mu < \infty$ and scale parameter $\sigma > 0$. This extreme-value distribution might be thought of as a log-Weibull distribution [23]. The MLE of $\sigma$ requires a numerical solution to $f(\sigma) = 0$ with

$$f(\sigma) := \overline{y} - \sigma - \frac{\sum_{i=1}^{N} y_i \exp\left(-\frac{y_i}{\sigma}\right)}{\sum_{i=1}^{N} \exp\left(-\frac{y_i}{\sigma}\right)} \tag{2.45}$$

where $\overline{y}$ denotes the sample mean of the observations. Again, to derive the update step of the Newton-Raphson algorithm, we first determine the derivative of $f(\sigma)$ w.r.t. $\sigma$ as

$$f'(\sigma) := \frac{\partial}{\partial \sigma} f(\sigma) = \frac{1}{\sigma^2} \sum_{i=1}^{N} y_i^2 \exp\left(-\frac{y_i}{\sigma}\right) + \\ \sum_{i=1}^{N} \exp\left(-\frac{y_i}{\sigma}\right) + \frac{1}{\sigma} \sum_{i=1}^{N} y_i \exp\left(-\frac{y_i}{\sigma}\right) - \overline{y} \frac{1}{\sigma^2} \sum_{i=1}^{N} y_i \exp\left(-\frac{y_i}{\sigma}\right) \tag{2.46}$$

which then allows to formulate the update step as $\hat{\sigma}_n = \hat{\sigma}_{n-1} - f(\hat{\sigma}_{n-1})/f'(\hat{\sigma}_{n-1})$ for $n \geqslant 2$. In contrast to the problematic computation of the starting value $\hat{\alpha}_1$ in Eq. (2.38) which we obtained by moment matching, the starting value $\hat{\sigma}_1$ can be easily obtained from the explicit expressions of the moment estimates [23]

$$\hat{\sigma} = \frac{1}{\pi}\sqrt{6}s \approx 0.779697s \text{ and } \hat{\mu} = \overline{y} - \gamma\hat{\sigma}, \tag{2.47}$$

where $\gamma$ denotes the Euler-Mascheroni constant, i.e. $\gamma \approx -0.57$. Eventually, we can use the moment estimate of $\sigma$ to start the Newton-Raphson algorithm and obtain the corresponding ML estimate. Inserting the ML estimate of $\sigma$ into

$$\hat{\mu} = \hat{\sigma} \log\left(\frac{1}{N} \sum_{i=1}^{N} \exp\left(\frac{y_i}{\hat{\sigma}}\right)\right) \tag{2.48}$$

gives the ML estimate of $\mu$. The only thing left to do is to transform the parameter estimates $\hat{\mu}$ and $\hat{\sigma}$ back to the estimates $\hat{\alpha}$ and $\hat{\beta}$ of the Weibull distribution. From the substitution we used in Eq. (2.43) we deduce

$$\hat{\alpha} = \frac{1}{\hat{\sigma}} \quad \text{and} \quad \hat{\beta} = \exp(\hat{\mu}). \tag{2.49}$$

To visualize the GoF of the Weibull distribution, Fig. 2.17 shows a set of Q-Q plots for the same DTCWT detail subband coefficients of Fig. 2.16. As we can see, the points approximately follow the dashed red line which indicates that the Weibull model is a reasonable choice here.

### 2.3.2 Gamma Distribution

A second, alternative model which also occurs in the literature of reliability and life span modeling is the two-parameter Gamma distribution. The Gamma distribution has been proposed as an alternative to the Rayleigh distribution for modeling the magnitudes of Gabor filter outputs [123] for instance. The p.d.f. and c.d.f., as given in [89], are

$$p_X(x; \alpha, \beta) = \frac{\beta^{-\alpha} x^{\alpha-1}}{\Gamma(\alpha)} \exp\left(-\frac{x}{\beta}\right), \quad x < 0 < \infty \tag{2.50}$$

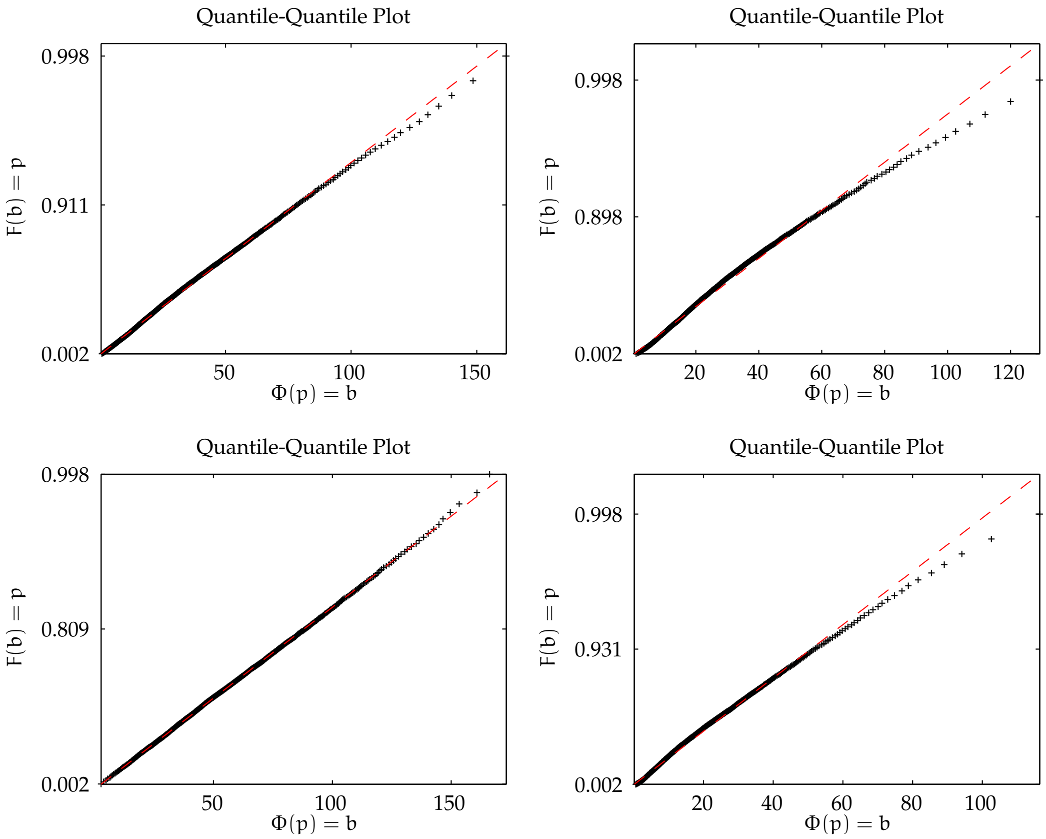


**Figure 2.17:** Exemplary Q-Q plots for GoF of the Weibull distribution.

and

$$F_X(x;\alpha,\beta) = P_l\left(a, \frac{x}{\beta}\right) \tag{2.51}$$

with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, respectively. The term $P_l(a,x)$ denotes the regularized (lower) incomplete Gamma function, i.e.

$$P_l(a,x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} \exp(-t)\,dt. \tag{2.52}$$

The inverse c.d.f. can be computed by $F_X^{-1}(u;\alpha,\beta) = \beta P_l^{-1}(\alpha,0,u)$ which is the numerical solution for $x$ to the equation $u = P_l(\alpha,0,x/\beta)$. In order to estimate the parameters $\alpha$ and $\beta$, we follow the approach presented by Choi & Wette [20]. The authors already provide the Newton-Raphson update step to compute the ML estimate of $\alpha$ as

$$\hat{\alpha}_n = \hat{\alpha}_{n-1} - \frac{\log(\hat{\alpha}_{n-1}) - \psi(\hat{\alpha}_{n-1}) - M}{1/\hat{\alpha}_{n-1} - \psi'(\hat{\alpha}_{n-1})}, \tag{2.53}$$

for $n \geqslant 2$. Here, $\psi$ and $\psi'$ denote the Digamma and Trigamma function [1], resp., and $M$ is defined as

$$M := \log(\overline{x}) - \frac{1}{N}\sum_{i=1}^{N} \log(x_i). \tag{2.54}$$

Given the ML estimate of $\alpha$, the ML estimate of $\beta$ has the closed-form expression

$$\hat{\beta} = \frac{\hat{\mu}}{\overline{x}}. \tag{2.55}$$

In order to reduce the computational overhead to evaluate the Digamma and Trigamma function we employ a lookup-table approach and linear interpolation. A starting value $\hat{\alpha}_1$ is obtained from the moment estimates [45]

$$\hat{\alpha}_1 := \hat{\alpha} = \left(\frac{\overline{x}}{s}\right)^2 \quad \text{and} \quad \hat{\beta} = \frac{s^2}{\overline{x}}. \tag{2.56}$$

We highlight the fact, that no computationally expensive operations have to be performed to estimate the starting values. To visualize the GoF of the Gamma distribution, Fig. 2.18 shows a set of Q-Q plots for the DTCWT detail subband coefficients we used in the previous sections. Apparently, the Q-Q plots are similar to the Weibull Q-Q plots in Fig. 2.17. In Section 2.3.4 we will show that the Gamma model is in many cases a more reasonable choice than the Weibull model.

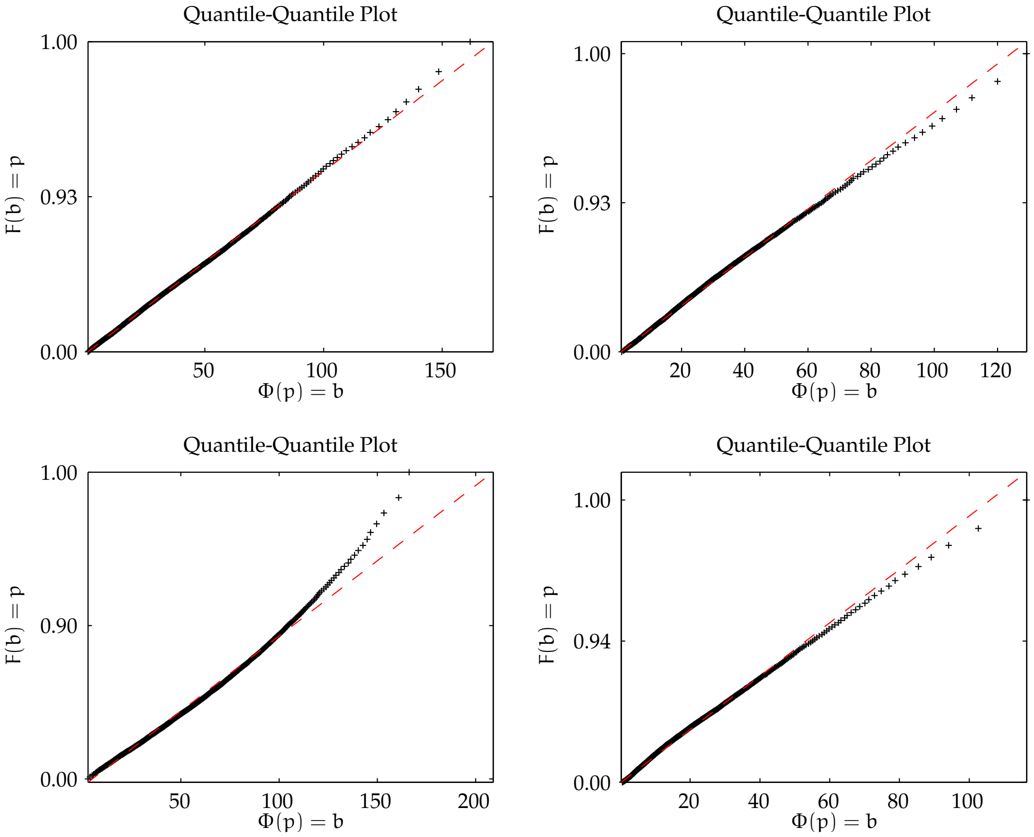


**Figure 2.18:** Exemplary Q-Q plots to visualize the GoF of the Gamma distribution.

### 2.3.3 Copula Modeling

As a last statistical model for the DTCWT detail subband coefficients, we present an approach which accounts for the association of transform coefficients between subbands of the same scale and between transform coefficients of subbands from different color channels. The only independency assumption of Section 2.2 we retain is the independency of transform coefficients across scales. Since we have already discussed the Weibull and Gamma distribution as suitable models for the transform coefficient magnitudes, we obviously favor a joint model which incorporates this information. A possible and elegant way to achieve this goal is to use the mathematical construct of copulas. Most of the following theoretical foundations are assembled from [43] and the classic textbooks on copulas by Joe [77] and Nelsen [137]. From a formal point of view a copula is a $n$-dimensional distribution function $C : [0,1]^n \rightarrow [0,1]$ with uniform marginals, satisfying the following requirements:

1. $\forall \mathbf{u} \in [0,1]^n : C(\mathbf{u}) = 0$, if at least one coordinate $u_b$ of $\mathbf{u}$ is 0
2. $\forall \mathbf{u} \in [0,1]^n : C(1,\dots,1,u_b,\dots,1) = u_b$
3. $\forall \mathbf{a},\mathbf{b} \in [0,1]^n, \mathbf{a} \leqslant \mathbf{b} : \sum_{\mathbf{c}} \operatorname{sgn}(\mathbf{c})C(\mathbf{c}) \geqslant 0$, where $\mathbf{c}$ is a vertex of the $n$-Box defined by the Cartesian product of the intervals $[a_1,b_1] \times [a_2,b_2] \times \cdots \times [a_n,b_n]$ and $\mathbf{a} \leqslant \mathbf{b} :\Leftrightarrow \forall b \in \{1,\dots,n\} : a_b \leqslant b_b$.

For our purpose, we will only consider random vectors $\mathbf{X} = (X_1,\dots,X_n)$ with continuous and strictly increasing marginal distribution functions. In [169], Sklar showed that given a $n$-dimensional distribution function $F_{\mathbf{X}}$ of $\mathbf{X}$ with marginal distribution functions $F_1,\dots,F_n$ there exists a $n$-dimensional copula $C$ such that

$$F_{\mathbf{X}}(x_1,\dots,x_n) = C(F_1(x_1),\dots,F_n(x_n)), \tag{2.57}$$

exploiting the fact that every random variable can be transformed to a uniform random variable by its probability integral transform [156], i.e. the mapping $\mathbb{R}^n \rightarrow [0,1]^n, (x_1,\dots,x_n) \mapsto (F_1(x_1),\dots,F_n(x_n))$. In other words, a copula can be considered as the distribution function of the *Probability Integral Transformed (PIT)* margins. Since we assume that the marginal distributions are absolutely continuous, the copula $C$ is uniquely determined on $[0,1]^n$. As a corollary of Sklar's theorem it follows that given a $n$-dimensional distribution function $F_{\mathbf{X}}$ with margins $F_1,\dots,F_n$ and copula $C$ we have the relation

$$C(\mathbf{u}) = F_{\mathbf{X}}(F_1^{-1}(u_1),\dots,F_n^{-1}(u_n)) \tag{2.58}$$

where $F_i^{-1}$ denotes the quantile functions and $\mathbf{u} = [u_1 \cdots u_n] \sim U([0,1]^n)$. Regarding the process of finding a suitable statistical model for multivariate observations, using the copula framework brings along a convenient simplification: the process of modeling the marginal distribution functions is completely decoupled from the process of modeling the association structure. This is a direct consequence of Sklar's theorem and allows to thoroughly adopt the findings we already obtained for the marginal distributions in Sections 2.3.1 and 2.3.2.

Before we discuss the choice of copula, we first assess the structure and strength of association across transform coefficients of subbands of the same scale and on different color channels by means of Chi-plots, shown in Fig. 2.19. We select a subset of all possible subband combinations to show the most prominent examples of association. In general, we observe three different types of association: (i) the weakest form of association occurs between coefficients of subbands capturing nearly orthogonal details on different color channels, shown in the bottom-left plot; (ii) on the contrary, the strongest association can be observed between coefficients of subbands oriented at the same angle but different color channels, shown in the top left-hand plot; (iii) coefficients of subbands oriented at opposite angles on different color channels exhibit association in between the two extremes, shown in the top and bottom right-hand plots.

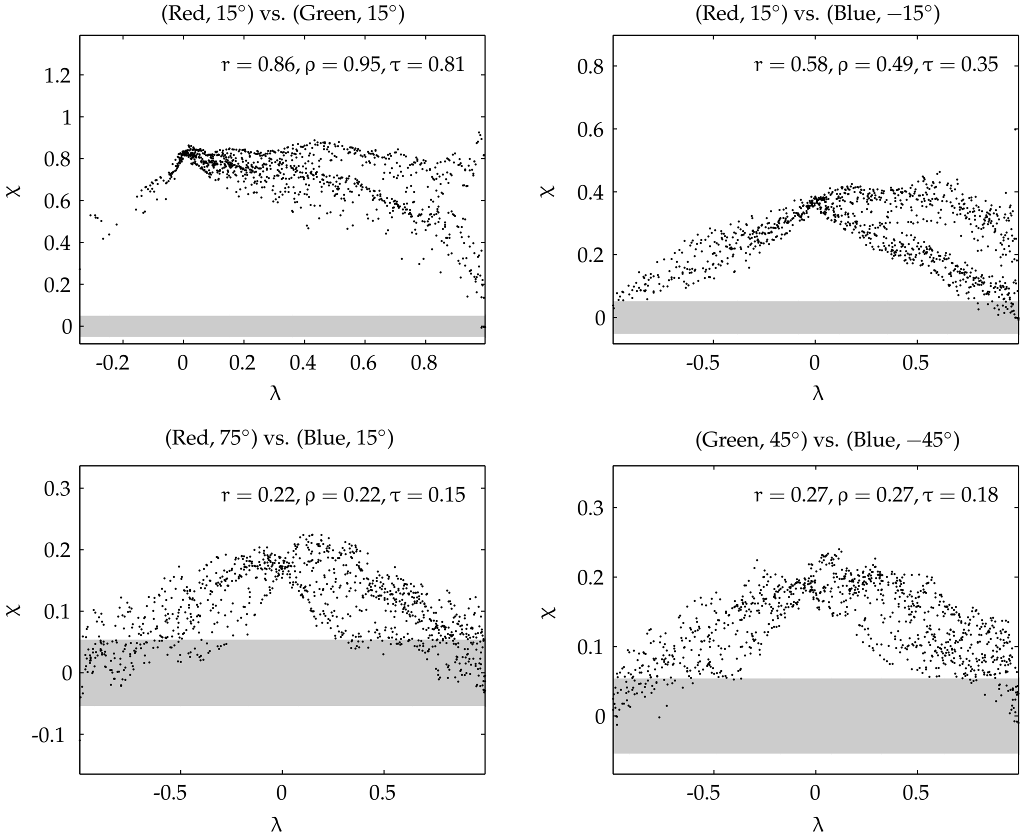


**Figure 2.19:** Chi-plots for coefficient magnitudes of various subband combinations of the texture image Bark.0008 (Vistex) on DTCWT level two.

We select two members of the family of elliptical copulas to capture the dependency structure between the transform coefficients: the Gaussian copula and the Student t copula. Elliptical copulas arise from the family of elliptical distributions. In fact, they are the copulas of elliptical distributions and inherit all the properties such as simple simulation of random numbers or well-known parameter estimation procedures for example. The copula of the multivariate Gaussian distribution with linear correlation matrix $\mathbf{R}$ (i.e. diag $\mathbf{R} = \mathbf{1}$) is defined as

$$C(u_1, \ldots, u_n; \mathbf{R}) = \boldsymbol{\Phi}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n); \mathbf{R}) \tag{2.59}$$

where $\boldsymbol{\Phi}$ denotes the standard multivariate Gaussian distribution function and $\Phi^{-1}$ denotes the quantile function of the standardized univariate Gaussian distribution. In the same manner, the Student t copula is defined as

$$C(u_1, \dots, u_n; \mathbf{R}, \nu) = T_{\mathbf{R},\nu}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_n)) \tag{2.60}$$

where $T_{\mathbf{\Sigma},\nu}$ denotes the standard multivariate Student *t* distribution, $\mathbf{R}$ is defined as before, $\nu$ denotes the degrees of freedom and $t_\nu^{-1}$ denotes the quantile function of the univariate Student *t* distribution. A crucial point for the copula modeling approach is the issue of parameter estimation. The setting is as follows: given a random vector $\mathbf{X} = (X_1, \dots, X_n)$ and the associated (parametric) copula model

$$F_{\mathbf{X}}(x_1, \dots, x_n; \theta_1, \dots, \theta_n, \mathbf{\Theta}) = C(F_1(x_1, \theta_1), \dots, F_n(x_n; \theta_n); \mathbf{\Theta}) \tag{2.61}$$

our objective is to estimate the parameter (vectors) $\theta_i$ of the marginal distributions and the copula parameter (vector) $\mathbf{\Theta}$. In the concrete example of a Gaussian copula and Weibull margins we have $\mathbf{\Theta} = \mathbf{R}$ and $\theta_i = [\alpha_i \ \beta_i]$. Since the p.d.f. of the copula can be deduced from

$$c(u_1, \dots, u_n) = \frac{\partial^d C(u_1, \dots, u_n)}{\partial u_1 \cdots \partial u_n} \tag{2.62}$$

we can write the joint p.d.f. of $\mathbf{X}$ as

$$p_{\mathbf{X}}(\mathbf{x}; \theta_1, \dots, \theta_n, \mathbf{\Theta}) = c(F_1(x_1; \theta_1), \dots, F_n(x_n; \theta_n); \mathbf{\Theta}) \cdot \prod_{i=1}^{n} f_i(x_i; \theta_i). \tag{2.63}$$

Eventually, given an i.i.d. sample $\mathbf{x}_1, \dots, \mathbf{x}_M$ we can write the log-likelihood function as

$$\begin{aligned} L(\theta_1, \dots, \theta_n, \mathbf{\Theta}; \mathbf{x}_1, \dots, \mathbf{x}_m) = \\ \sum_{i=1}^{M} \log c(F_1(x_{i1}; \theta_1), \dots, F_1(x_{in}; \theta_n); \mathbf{\Theta}) + \sum_{i=1}^{M} \sum_{j=1}^{n} \log f_j(x_{ij}; \theta_j). \end{aligned} \tag{2.64}$$

Due to the fact that it is computationally expensive and numerically cumbersome to jointly estimate the parameters of the marginal distributions and the copula parameters (denoted as the exact ML approach), we follow a commonly-used two-step procedure, termed the *Inference Functions from Margins (IFM)* method or *Canonical Maximum Likelihood (CML)* approach. The IFM approach refers to the situation where we have a parametric representation of the marginal distributions, whereas the CML approach refers to the situation where we rely on empirical c.d.f.s. We use the IFM method throughout this thesis. The basic idea was introduced by Joe [77] and is based on a very simple decoupling of the estimation procedure. First, we estimate the parameters of the parametric margins (e.g. Weibull, Gamma, etc.)

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{M} \log f_n(x_{in}; \theta) \tag{2.65}$$

using ML estimation. Second, we use the obtained estimates to perform the probability integral transform on the margins. Third, we estimate the copula parameters in a ML sense by maximizing

$$\hat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{M} \log c(F_1(x_{i1}; \hat{\theta}_1), \dots, F_n(x_{in}; \hat{\theta}_n); \mathbf{\Theta}). \tag{2.66}$$

To provide a concrete example, we consider the case of using a Gaussian copula with Weibull margins, a case which we will return to in Section 3.4 for the purpose of image retrieval. In a first step, we deduce the p.d.f. of the Gaussian copula. For that purpose we assume that $\mathbf{X}$ follows a standard multivariate Gaussian distribution with correlation matrix $\mathbf{R}$. We know that the marginal distributions are univariate standard Gaussians, i.e. $X_i \sim \mathcal{N}(0,1)$. Hence we can try to manipulate the p.d.f.

$$p_{\mathbf{X}}(\mathbf{x};\mathbf{R}) = \frac{1}{2\pi^{\frac{n}{2}}|\mathbf{R}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right) \tag{2.67}$$

such that we get an expression similar to Eq. (2.63). After some algebraic manipulations, it turns out that the p.d.f. of the Gaussian copula has the form

$$c(u_1,\ldots,u_n;\mathbf{R}) = |\mathbf{R}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\xi}^T(\mathbf{R}^{-1}-\mathbf{1})\boldsymbol{\xi}\right) \tag{2.68}$$

with $\boldsymbol{\xi} = [\Phi^{-1}(u_1)\ \cdots\ \Phi^{-1}(u_n)]$, or more precisely $\boldsymbol{\xi} = [\Phi^{-1}(F_1(x_i))\ \cdots\ \Phi^{-1}(F_n(x_n))]$. It is then straightforward to determine the ML estimate of $\mathbf{R}$ as

$$\hat{\mathbf{R}} = \sum_{i=1}^{M}\boldsymbol{\xi}_i^T\boldsymbol{\xi}_i \tag{2.69}$$

by taking the partial derivative w.r.t. $\mathbf{R}$ of the log-likelihood function corresponding to Eq. (2.68) and setting the resulting term to zero. The ML estimates of the Weibull distribution parameters $\alpha_i, \beta_i$ are given in Section 2.3.1. In a similar manner, we can determine the p.d.f. of the Student t copula, however the derivation is somewhat more involved. The p.d.f. of a $n$-variate Student t distribution is given as

$$p_{\mathbf{X}}(\mathbf{x};\mathbf{R},\nu) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma(\frac{\nu}{2})(\nu n)^{\frac{n}{2}}|\mathbf{R}|^{\frac{1}{2}}}\left(1+\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right)^{-\frac{\nu+n}{2}} \tag{2.70}$$

with correlation matrix $\mathbf{R}$ and $\nu$ degrees of freedom. By factorizing out the univariate standardized Student t distributions

$$p_X(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{x^2}{\nu}\right)^{\frac{-\nu+1}{2}} \tag{2.71}$$

we can finally deduce the p.d.f. of the Student t copula as

$$p(u_1,\ldots,u_n;\mathbf{R},\nu) = |\mathbf{R}|^{-1/2}\frac{\Gamma\left(\frac{\nu+n}{2}\right)\left[\Gamma\left(\frac{\nu}{2}\right)\right]^n}{\left[\Gamma\left(\frac{\nu+1}{2}\right)\right]^n\Gamma\left(\frac{\nu}{2}\right)}\frac{\left(1+\frac{1}{\nu}\boldsymbol{\xi}^T\mathbf{R}^{-1}\boldsymbol{\xi}\right)^{-\frac{\nu+n}{2}}}{\prod_{i=1}^{n}\left(1+\frac{\xi_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}} \tag{2.72}$$

with $\boldsymbol{\xi} = [t_\nu^{-1}(u_1)\ \cdots\ t_\nu^{-1}(u_n)]$ or again more precisely $\boldsymbol{\xi} = [t_\nu^{-1}(F_1(x_i))\ \cdots\ t_\nu^{-1}(F_n(x_n))]$. Unfortunately, the ML estimates of the Student t parameters $\mathbf{R}$ and $\nu$ have no explicit expression and have to be calculated by a numerical optimization algorithm. In this thesis we use MATLAB's `copulafit` routine to estimate $\nu$ and $\mathbf{R}$. Basically the routine employs numerical function minimization to find a minimum of the negative log-likelihood function corresponding to Eq. (2.72) w.r.t. $\nu$. During minimization, $\mathbf{R}$ is iteratively estimated using an algorithm proposed in a working paper by Bouyé et al. [11]. To visualize the shape of the p.d.f. and c.d.f of a Gaussian and Student t copula, Fig. 2.20 shows the corresponding plots for a correlation coefficient of $\rho = 0.5$.

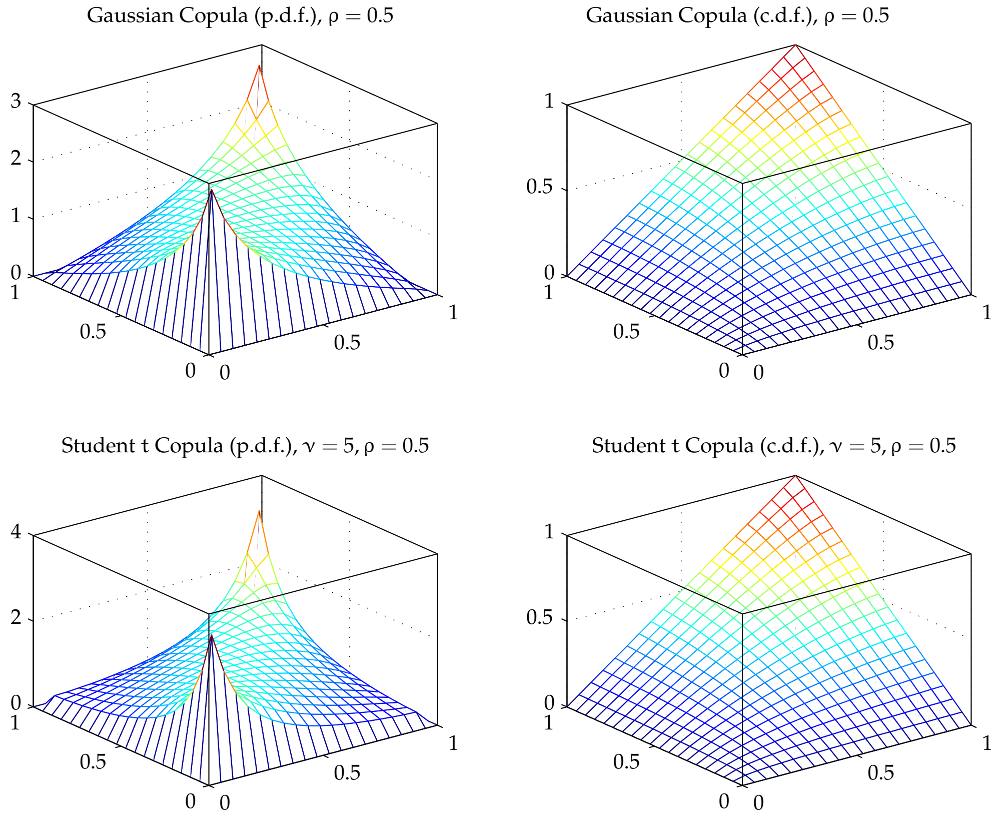


**Figure 2.20:** Visualization of the p.d.f. and c.d.f. of a Gaussian and Student t copula with correlation coefficient $\rho = 0.5$.

### 2.3.4 Quantifying the Goodness-of-Fit

In order to allow a quantitative statement about the GoF of the Rayleigh, Weibull or Gamma model for the DTCWT transform coefficient magnitudes, we conduct a series of Chi-Square GoF tests on subband coefficients from DTCWT decomposed Vistex [31], Outex, and Stex images. We decompose each RGB color channel separately and conduct a Chi-Square test for each subband on each decomposition level of a three-scale DTCWT at 5% significance. The percentage of rejected null-hypotheses per decomposition level (averaged over all subbands) is listed in Table 2.5. Apparently, the rates for all decomposition levels are consistent over all three databases. However, the reported rejection rates are different to the results presented in [101] or [104] where we reported quite high rejection rates for decomposition levels one and two. This effect can be attributed to our change in the GoF testing strategy, where we try to achieve the same test power by means of sampling 500 coefficients from each subband. In [101] or [104], we did not perform this correction and consequently the rejection rates were higher at lower decomposition levels. Further, the listed rejection rates are in accordance with our visual impression that the Gamma and Weibull distribution represent reasonable statistical models for the coefficient magnitudes. The rejection rates for both distributions range from ten to twenty percent across all scales with some exceptions in case of the Outex database where the rejection

| Database | Level | Model | | |
|---|---|---|---|---|
| | | Weibull | Gamma | Rayleigh |
| Vistex | 1 | 11.68 | **8.87** | 66.02 |
| | 2 | 14.37 | **13.48** | 66.34 |
| | 3 | **14.81** | 14.91 | 58.07 |
| Stex | 1 | 19.90 | **12.93** | 70.75 |
| | 2 | 19.63 | **15.79** | 62.63 |
| | 3 | 18.70 | **17.68** | 58.57 |
| Outex | 1 | 4.66 | **2.48** | 33.31 |
| | 2 | 12.86 | **8.68** | 42.08 |
| | 3 | 14.34 | **11.51** | 50.61 |

**Table 2.5:** Percentage of rejected null-hypotheses for each decomposition level of the DTCWT, averaged over all subbands using equal sample sizes (i.e 500 samples). The lowest rejection rates per level are marked bold.

rates are even lower. However, it is obvious that the Rayleigh distribution is a too rigid model for the coefficient magnitudes.

As a final point, we discuss the issue of copula model selection and GoF testing which we consider two particular weaknesses of the copula approach. Generally speaking, there exists no commonly-accepted or recommended method to accomplish these tasks. Nevertheless, several approaches have been proposed recently in literature (see Genest et al. [54] or Berg [9] and references therein). The variety of ideas ranges from the reduction of the multivariate GoF problem to an univariate one (mainly using the probability integral transform), to parametric bootstrap procedures [140] or even the exploitation of positive definite bilinear forms [146]. In [103], we choose a very pragmatic and straightforward approach, originally suggested by Genest and Favre [52] as a first step towards model selection. We plot the pairs of original DTCWT transform coefficient magnitudes against random samples from the fitted copula model. An example of such a plot is shown in Fig. 2.21, where we have fitted a Gaussian copula with Weibull margins to the same subband combinations we used in Fig. 2.19. The red points represent the scatter plot of the original subband coefficient magnitudes while the light-gray crosses represent the scatter plot of 500 points sampled from the statistical model. In fact, the light-gray crosses are obtained by sampling from the Gaussian copula and using the Weibull quantile functions to transform the margins.

However, the large number of possible subband combinations limits the applicability of this approach to a preliminary visual inspections of model fit. To overcome this shortcoming, we further experimented with the Akaike [2] and Schwarz Information Criterion [161] which both take into account the log-likelihood of the data under the given model and penalize additional parameters to avoid overfitting issues. Nevertheless, AIC and BIC are not an adequate tool to address the problem of model selection in a hypothesis testing sense. They are rather useful as a means for selecting among possible candidate models without caring whether the models can actually describe the underlying data. To re-evaluate our selection of the Student t copula in [103], we implement a GoF test recently proposed by Genest et al. [53, 54]. The test is based on

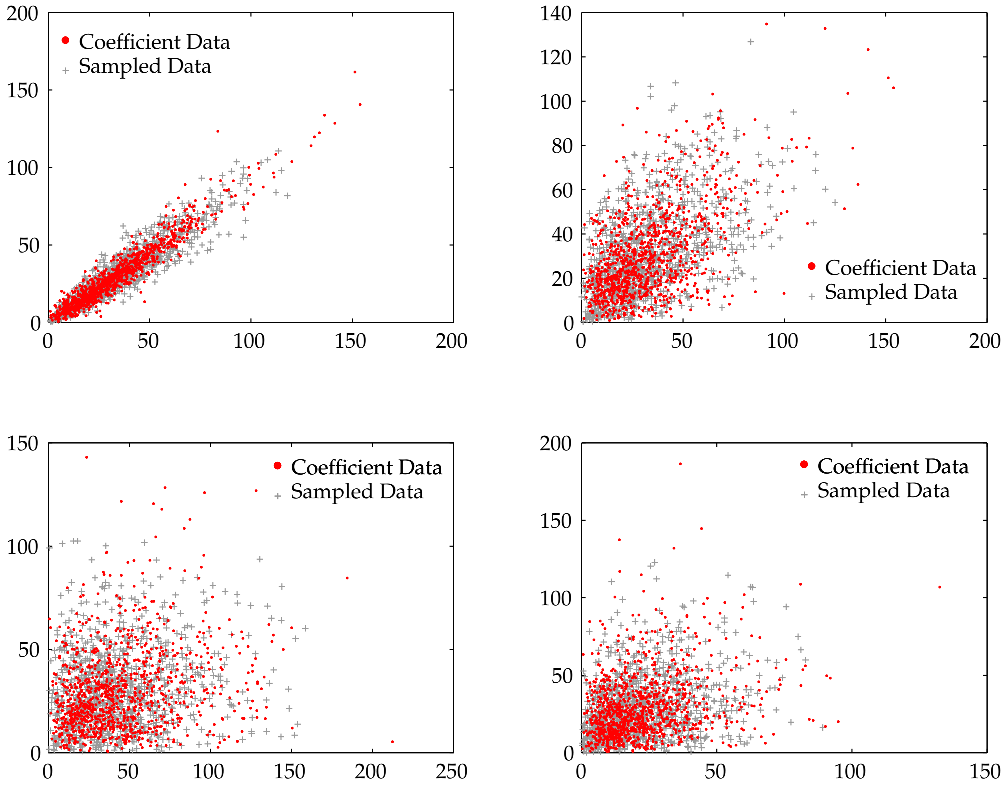


**Figure 2.21:** Scatter plots of original DTCWT transform coefficient magnitudes (red points) against 500 samples drawn from a fitted Gaussian copulas with Weibull margins.

the computation of the Cramer-von-Misés statistic

$$\int_{[0,1]^n} \mathbb{C}_n(\mathbf{u})^2 dC_n(\mathbf{u}), \quad \text{with} \quad \mathbb{C}_n = \sqrt{n}(C_n - C_{\theta_n}), \tag{2.73}$$

where $C_n$ denotes the empirical copula [137] and $C_{\theta_n}$ denotes the estimated parametric copula under the null-hypothesis (i.e. either Gaussian or Student t). Regarding the actual implementation of the GoF test, we adhere to the parametric bootstrap algorithm [42] given in Appendix A of [54]. We choose 1000 bootstrap samples for our test. The null-hypothesis is rejected whenever the estimated p-value is lower than the significance level of $\alpha = 0.05$. Due to the fact that the parametric bootstrap procedure includes the computation of $C_{\theta_n}$ in Eq. (2.73), we run into considerable computational problems since the test requires to compute multivariate Gaussian or multivariate Student t probabilities. This in turn requires computationally intensive multi-dimensional numerical integration for which we use the specifically-tailored algorithms presented by Genz [55] and Genz & Bretz [56]. As a consequence of the intensive computational demands, we limit our GoF study to the 200 example textures of the Vistex (full) database to get an impression of model fit. We select the subbands of DTCWT decomposition level three. Since we have three color channels and six subbands per scale, the joint statistical model is 18-dimensional. The rejection rates are listed in Table 2.6. The numbers are almost equal for both

| **Copula** | |
|---|---|
| Student t | Gaussian |
| 38.50 | 35.50 |

**Table 2.6:** Rejection rates of the GoF test proposed by Genest et al. [53, 54] for 18-dimensional coefficient magnitude vectors (DTCWT level three) of 200 texture images.

copulas. In Chapter 3, we will however see that the Gaussian copula is far more attractive from a computational point of view.

# Chapter 3

# Texture Image Retrieval

This part of the thesis is devoted to the first application scenario of the statistical models presented in Chapter 2. We deal with the problem of Content-Based Image Retrieval (CBIR) and particularly focus on texture images. Throughout the last years, we observed the trend that the amount of digital data stored in multimedia databases, such as image repositories, is constantly growing. In order to handle this huge amount of data, we are confronted with the need for systems which allow classification of content as well as sorting and searching. These three exemplary requirements share a common ground: in order to obtain reasonable results, we need to know how to represent or describe the content. In the context of searching in visual data, the ambitious goal of allowing semantic queries is still an issue of open research. Systems which solely perform image searches or queries by relying on textual annotations, are usually not capable of representing the visual content that is perceived by human beings. A popular, alternative CBIR strategy is to perform image queries by providing examples of the visual content we search for. This is a less ambitious, however, not necessarily less complex problem, since it requires to define a suitable similarity measure between images. In practice, a CBIR system will usually not return just one image but a set of potential results. This gives the user the freedom to decide which images to keep. The fields of application of a CBIR system range from searching in databases of natural images, e.g. holiday photos, to searching for images in repositories of medical content. In Chapter 4, we will discuss how the idea of CBIR can be exploited to predict the histological diagnosis of endoscopy images for example. In a more formal description, the objective of a CBIR system is to find the $K \ll L$ most similar images to a given query in an image repository of $L$ potential candidates. A schematic illustration of the CBIR building blocks is shown in Fig. 3.1.

The chapter is basically divided into two major parts: in the first part, we introduce the problem of CBIR as a problem of statistical inference. In this context, a probabilistic formulation of CBIR will serve as a basis for our work. In Section 3.2, we then review related research work in the field of (texture) image retrieval and especially focus on approaches which closely adhere to the probabilistic formulation of image retrieval. The second major part of the chapter is then devoted to our contribution. First, we motivate the need for a lightweight texture retrieval system by discussing two retrieval scenarios with different computational requirements. We

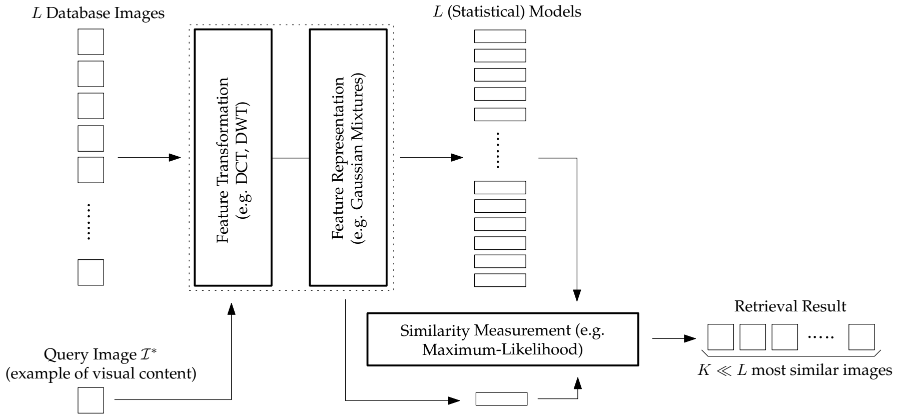

$L$ Database Images
Feature Transformation (e.g. DCT, DWT)
Feature Representation (e.g. Gaussian Mixtures)
$L$ (Statistical) Models
Similarity Measurement (e.g. Maximum-Likelihood)
Retrieval Result
$K \ll L$ most similar images
Query Image $\mathcal{I}^*$
(example of visual content)

**Figure 3.1:** Schematic illustration of a CBIR system with the critical parts marked bold.

then introduce a novel, lightweight retrieval approach for which we provide a thorough computational analysis of the main building blocks and a comparative study to popular approaches from literature. In the second part of the contribution, we develop a retrieval approach based on the theory of copula modeling. To evaluate the retrieval performance, we conclude with a large-scale comparative study on four texture image databases. As a guideline for the reader, we highlight that major parts of the following content recently appeared in:

[101] R. Kwitt and A. Uhl. Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'08)*, pages 933–936, San Diego, California, United States, October 2008

[103] R. Kwitt and A. Uhl. A joint model of complex wavelet coefficients for texture retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pages 1877–1880, Cairo, Egypt, November 2009

[104] R. Kwitt and A. Uhl. Lightweight probabilistic image retrieval. *IEEE Transactions on Image Processing*, 19(1):241–253, January 2010

## 3.1 Image Retrieval as Statistical Inference

To the best of our knowledge, Vasconcelos & Lippman [185, 186] first introduced a Bayesian formulation of CBIR, also referred to as *Minimum Probability of Error* retrieval. An image $\mathcal{I}$ consists of a number of pixel observations $(x_1, \dots, x_N) = \boldsymbol{x} \in \mathcal{X}$ residing in the space of observations $\mathcal{X}$. We assume that each image of the database belongs to one of $M$ image classes. Hence, the starting point of the probabilistic retrieval formulation resembles a standard classification scenario. Next, let $Y$ denote a random variable with realizations in $\{1, \dots, M\}$ and let $p_Y$ denote the probability mass function (p.m.f.) of $Y$. As a first building block of the CBIR system, Vasconcelos & Lippman identify a *feature transformation* stage which is a mapping $T : \mathcal{X} \to \mathcal{Z}$ from the space of observations to the so called *feature space* $\mathcal{Z}$. The key issue here is, to represent the image content in a domain which is more suitable for further processing. Accordingly, $\boldsymbol{z} = T(\boldsymbol{x})$ denotes a so called *feature vector*. The second building block of the CBIR system is a probabilistic model describing how the feature vectors populate the feature space with respect to their class membership. The corresponding class-conditional p.d.f. $p_{Z|Y}(z|y)$ constitutes the *feature representation*. The final part of the CBIR system deals with the task of assigning a novel image to one of the M image classes which leads to the question of how to define the so called retrieval function $g : \mathcal{Z} \rightarrow \{1, \ldots, M\}$. In the formulation of [186], the authors argue that the ulterior objective for designing this function is to minimize the probability of retrieval error or in classification terminology the probability of classification error. Given that the function $\omega : \mathcal{Z} \rightarrow \{1, \ldots, M\}$ returns the true class membership of a feature vector $z$, the objective is to minimize $\mathbb{P}(g(z) \neq y|\omega(z) = y)$, i.e. the probability of assigning $z$ to a class other than its true class $y$. From statistical classification theory (e.g. see [51]) we know that the function minimizing this criteria is the Bayes classifier

$$g(z) = \arg\max_{y} p_{Y|Z}(y|z). \tag{3.1}$$

Applying the Bayes rule and noting that the maximization is independent of $p_Z$, we obtain the equivalent formulation

$$g(z) = \arg\max_{y} p_{Z|Y}(z|y)p_Y(y) \tag{3.2}$$

which is substantially easier to handle than Eq. (3.1). We only have to estimate the class-conditional likelihood $p_{Z|Y}$ instead of the posterior probability $p_{Y|Z}$. One important element in the formulation of [186] is, that we can get rid of the p.m.f. $p_Y$ in Eq. (3.2) by assuming that each image belongs to its own class with equal prior probability, i.e. $\forall y \in \{1, \ldots, M\} : p_Y(y) = {}^1\!/\!_M$. In CBIR, this is a reasonable simplification since it is hard to establish a-priori probabilities of database images. As a consequence, Eq. (3.2) reduces to the Maximum Likelihood (ML) selection criterion.

In any practical scenario, we will have to estimate $p_{Z|Y}$ from a collection of feature vectors $z_1, \ldots, z_R$ and the actual retrieval process will be based on a collection of query feature vectors $z_1^*, \ldots, z_K^*$ extracted from the query image $\mathcal{I}^*$. As we will later see, it is computationally beneficial to choose K smaller than R. Assuming that the feature vectors are i.i.d. and conditionally independent given the true class membership, facilitates estimation of $p_{Z|Y}$ and allows to write the ML selection rule as

$$g(z_1^*, \ldots, z_K^*) = \arg\max_{y} \prod_{k=1}^{K} p_{Z|Y}(z_k^*|y). \tag{3.3}$$

Since each image belongs to its own class, we can omit the notation Z|Y from now on and instead indicate that a feature representation belongs to image $\mathcal{I}_j$ by indexing the model parameter $\theta_j$. As all feature representations we consider in this chapter belong to some parametric family, this is notationally more convenient. To conclude the recapitulation of probabilistic CBIR, we finally highlight the important relation between ML image retrieval according to Eq. (3.3) and retrieval by searching for the feature representation which minimizes the Kullback-Leibler (KL) divergence [26] to the feature representation of the query image. Given that $p_Z(z;\theta_1), \ldots, p_Z(z;\theta_L)$ denote the representations of the candidate images and $p_Z(z;\theta^*)$ denotes the representation of the query image, it can be shown that ML selection is asymptotically (i.e. $K \rightarrow \infty$) equivalent to

$$g(z) = \arg\min_{y} D(p_Z(z;\theta_y)||p_Z(z;\theta^*)) \tag{3.4}$$

where

$$D(p_Z(z;\theta_y)||p_Z(z;\theta^*)) := \int_{\Omega} p_Z(z;\theta_y) \log \frac{p_Z(z;\theta_y)}{p_Z(z;\theta^*)} dz \tag{3.5}$$

denotes the KL divergence and $\Omega$ denotes the domain of the p.d.f. $p_Z$. This relation can easily be verified by application of the weak law of large numbers (see [186] or [187] for a proof and some other interesting relationships). Note, that in case we rely on Eq. (3.4) as the retrieval function, we have to estimate $p_Z(z; \theta^*)$ from $z_1^*, \dots, z_R^*$ first. In situations where there exists a closed-form expression for the KL divergence between two feature representations, the additional estimation step pays off, since we can compute the measure of similarity by solely relying on the model parameters $\theta_1, \dots, \theta_L$ and $\theta^*$. In comparison, Eq. (3.3) requires to evaluate the p.d.f. $p_Z$ for each query feature vector $z_i^*$. Apparently, this implies a trade-off in choosing the number of query feature vectors $K$, since $K \approx R$ presumably reduces retrieval errors, but on the other hand increases computational demand as well. Nevertheless, using either ML selection or the KL divergence minimization strategy only requires model parameters to be stored. Hence, both strategies are quite efficient from a storage point of view. In the following section, we review research works which all more or less exploit the probabilistic CBIR formulation of Vasconcelos & Lippman.

## 3.2 Related Work

In the original work [186], Vasconcelos & Lippman present a first application of the probabilistic CBIR formulation based on the 2-D Discrete Cosine Transform (DCT) for feature transformation and multivariate Gaussian Mixture Modes (GMM) for feature representation. The authors employ a sliding-window approach to compute the 2-D DCT on each $8 \times 8$ pixel window and extract the first $D$ coefficients (including the DC coefficient) in MPEG zig-zag scan order to obtain $D$-dimensional feature vectors. A eight-component GMM is then fit to the feature vectors using the classic Expectation-Maximization (EM) algorithm [36], initialized by an adaption of Gray's codeword-splitting procedure (see [59] for the original algorithm and [186] for a description of the modification). Retrieval is accomplished by extracting query feature vectors in the same way, however, using a non-overlapping $8 \times 8$ block 2-D DCT. Hence, the amount of query feature vectors is significantly smaller than the number of feature vectors used for GMM estimation and the computational demand for similarity measurement is reduced. In another work by Vasconcelos [183], the author proposes an approximation of the KL divergence between mixture models for retrieval, denoted as the Asymptotic Likelihood Approximation.

In [40], Do & Vetterli present an CBIR approach which is based on the same idea of minimizing the retrieval error, however the configuration of the feature transformation and feature representation step is different. The authors base their approach on the DWT for feature transformation and follow the assumptions of Section 2.2 to construct an efficient feature representation based on the GGD. Although the independency assumptions potentially affect retrieval accuracy in a negative way [184], they allow computationally efficient retrieval as the authors show by deriving a closed-form expression for the KL divergence between two GGDs. Consequently, the retrieval task solely depends on the estimated GGD parameters. In [39], Do & Vetterli present an extension of this approach to achieve rotational invariance by relying on the Steerable Pyramid [168] for feature transformation and two particular forms of Hidden Markov Trees (HMT) for feature representation. Retrieval is accomplished by an approximation of the KL divergence between HMTs [38].

In [180], Tzagkarakis et. al. propose a similar idea but use the DWT for feature transformation and the family of Symmetric $\alpha$-Stable distributions (S$\alpha$S) for feature representation, again adhering to the assumptions of Section 2.2. Since there exists no closed-form solution for the KL divergence between two S$\alpha$S distributions in general form, the authors suggest to use the characteristic functions instead of the p.d.f.s to compute the KL divergence. In [179], this approach is carried forward by the same authors to achieve rotational invariance by means of a Steerable Pyramid together with $\alpha$-stable modeling of the subband coefficients and a "Gaussianization" procedure to obtain multivariate Gaussian distributed coefficients. In further consequence, this allows application of the KL divergence between multivariate Gaussian distributions (for which a closed-form expression exists).

Another interesting approach is presented by de Ves et. al [34], where the wavelet coefficients of the vertical and horizontal DWT detail subbands are considered as realizations of a bivariate random vector and the magnitude is modeled by a two-parameter Gamma distribution. The authors report good retrieval results using the Stationary Wavelet Transform (SWT, implemented by the à-trous algorithm) as a substitution for the DWT to get rid of the shift-dependency problem. Similar to previous works, the KL divergence minimization strategy is employed for image retrieval.

## 3.3 Lightweight Probabilistic Texture Retrieval

In this section, we introduce a novel texture image retrieval approach which is based on the probabilistic CBIR formulation of Section 3.1 and can be considered as a direct extension of the work of Do & Vetterli [40]. The ingredients of this approach are the DTCWT for feature transformation and the Weibull or Gamma distribution for feature representation. Image retrieval is based on the KL divergence minimization strategy for which we present closed-form expressions. Besides the development of a novel variant of probabilistic CBIR, a main concern of this section is computational complexity. Since most publications on CBIR solely aim at an improvement in retrieval accuracy and often neglect computational issues, solutions which are computationally inexpensive and minimize the retrieval error are rare. In the probabilistic framework where each image is represented by some statistical model and image similarity is measured by a function of these models, we have to deal with the trade-off between model complexity and computational performance. Increasing the model complexity to better capture image characteristics might lead to higher retrieval rates on the one hand, but it is very likely that the computational demand for feature transformation, representation or similarity measurement increases in a similar manner. In particular, we consider two scenarios which impose computational constraints on different building blocks of the CBIR framework. The scenarios differ in that possible performance bottlenecks arise at different locations. Both scenarios are sketched next:

**Retrieval Scenario A** This scenario is the *classic* retrieval scenario, where the model parameters of all images in the repository are calculated off-line and new images are added to the database at a slow rate. Hence, overall runtime performance is predominantly limited by similarity measurement which inherently depends on the size of the image repository L. The runtime impact of model parameter estimation and feature transformation is of secondary importance since both steps have to be performed only once (i.e. for each new query).

**Retrieval Scenario B** The second retrieval scenario we are concerned about has several facets and imposes additional requirements on the building blocks of the retrieval framework. First, we observe situations where new images arrive at a high rate and have to be stored in the database. At the same time, image queries are executed. The computational demand for similarity measurement is still the primary concern here, however the complexity of parameter estimation becomes an important issue. If the images are represented in a domain other than spatial, the feature transformation step possibly contributes a significant amount of additional runtime as well. Other challenging variants of this scenario occur when online texture similarity measurement is required, e.g. when the frames of an image stream have to be matched to a limited set of query templates. Real-world examples for that include video-controlled quality assurance in texture manufacturing, or the detection of cancerous tissue during video-colonoscopy. Computationally expensive parameter estimation or feature transformation can scale up to the limiting factors for production throughput or slow down the diagnostic process. In order to cover both retrieval scenarios, we need a low-complexity features transformation, a similarity measure which exclusively depends on the image model parameters and an efficient model parameter estimation procedure in the feature representation step.

In order to meet the requirements set by the two retrieval scenarios, we choose to adopt all three assumptions of Section 2.2. First, we establish the formal connection to the probabilistic CBIR formulation. Let $\mathcal{X}$ denote the space of pixel observations and let $T$ denote the feature transformation, i.e. the DTCWT. Given a $J$-scale DTCWT, we obtain $B := 6J$ detail subbands in case of single-channel (e.g. luminance) images. For the feature representation, we only consider the magnitudes of the complex-valued transform coefficients. A feature vector $\boldsymbol{z} = (z_1, \dots, z_B)$ consists of one coefficient magnitude per subband. Due to the independency assumption, we can write the joint p.d.f. $p_{\mathbf{Z}}$ of the random vector $\mathbf{Z}$ as

$$p_{\mathbf{Z}}(\boldsymbol{z}; \boldsymbol{\Theta}) = \prod_{b=1}^{B} p_{Z_b}(z_b; \boldsymbol{\theta}_b) \tag{3.6}$$

with $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_B]$. In case we take a Weibull or Gamma distribution as a basis, $\boldsymbol{\theta}_b = [\alpha_b \ \beta_b]$. In order to estimate $p_{\mathbf{Z}}$ we have to estimate the parameter vectors $\boldsymbol{\theta}_b$ from a collection of feature vectors. The assumption of i.i.d. transform coefficients allows to estimate $\boldsymbol{\theta}_b$ from all coefficients of subband $b$. In contrast to [186], we do not follow a sliding window approach to extract feature vectors. Nevertheless, due to subsampling by two after each decomposition level of the DTCWT, the subbands of two successive levels differ in size by a factor of $1/4$. As a consequence, we do not obtain vectors of equal lengths. Technically, this means that estimation of $\boldsymbol{\theta}_b$ is accomplished based on the $N_b$ coefficient magnitudes $z_{b1}, \dots, z_{bN_b}$ of subband $b$. For the actual retrieval process, we have $B$ query feature vectors $\boldsymbol{z}_1^*, \dots, \boldsymbol{z}_B^*$ where $\boldsymbol{z}_b^*$ consists of $V_b$ coefficients from a subband $b$. We intentionally use $V_b$ to signify that the number of transform coefficients in subband $b$ does not necessarily have to be equal to $N_b$ for the computation of the ML selection rule, i.e.

$$g(\boldsymbol{z}_1^*, \dots, \boldsymbol{z}_B^*) = \arg \max_{k \in \{1, \dots, L\}} \sum_{b=1}^{B} \sum_{j=1}^{V_b} \log p_{Z_b}(z_{bj}; \boldsymbol{\theta}_k). \tag{3.7}$$

Although, we compute the DTCWT on the whole image and obtain all coefficients anyway, limiting the amount of coefficients to $V_b$ might be of practical interest for very large images due do the reduced computational effort to evaluate the likelihood. Another consequence of assuming independency between $Z_1, \dots, Z_B$ is, that we can employ the chain-rule of entropy

[26] and obtain

$$g(\mathbf{z}) = \arg \min_{k \in \{1,\ldots,L\}} \sum_{b=1}^{B} D(p_{Z_b}(z;\theta_k) || p_{Z_b}(z;\theta^*)) \tag{3.8}$$

as an alternative, KL divergence based, retrieval strategy. For the Gamma and Weibull distribution the KL divergence in Eq. (3.8) has a closed-form expression [104]. Given that $p_i := p(z;\alpha_i,\beta_i)$ and $p_j := p(z;\alpha_j,\beta_j)$ denote the p.d.f.s of two Weibull distributions, we obtain

$$\begin{aligned} D(p_i||p_j) = \Gamma\left(\frac{\alpha_j}{\alpha_i}+1\right)\left(\frac{\beta_i}{\beta_j}\right)^{\alpha_j} + \log\left(\beta_i^{-\alpha_i}\alpha_i\right) - \\ \log\left(\beta_j^{-\alpha_j}\alpha_j\right) + \log\left(\beta_i\right)\alpha_i - \log\left(\beta_i\right)\alpha_j + \frac{\gamma\alpha_j}{\alpha_i} - \gamma - 1. \end{aligned} \tag{3.9}$$

and in case $p_i, p_j$ denote the p.d.f.s of two Gamma distributions we obtain

$$\begin{aligned} D(p_i||p_j) = \psi(\alpha_i)(\alpha_i - \alpha_j) - \alpha_i + \\ \log\left(\frac{\Gamma(\alpha_j)}{\Gamma(\alpha_i)}\right) + \alpha_j \log\left(\frac{\beta_j}{\beta_i}\right) + \frac{\alpha_i\beta_i}{\beta_j} \end{aligned} \tag{3.10}$$

Here, $\gamma = 0.577216$ denotes the Euler-Mascheroni constant [1]. Our formal description of the feature representation and similarity measurement step can be directly adapted to describe the approach by Do & Vetterli [40] in the framework of [187]. We only have to replace the DTCWT by the DWT and the Gamma or Weibull distribution by the GGD. The corresponding closed-form expression for the KL divergence is given in [40]. With respect to the approach of Do & Vetterli, we remark that the independency assumptions are a crude simplification in our setup, since coefficients of a redundant transform, such as the DTCWT, will inevitably exhibit dependencies (e.g., see Section 2.3.3). However, we will see that this simplification pays off in the sense that we obtain a simple and computationally efficient CBIR approach with good retrieval rates (see Section 3.3.2). Finally, we point out that although many research papers on CBIR do not adhere to the terminology of feature transformation, representation and similarity measurement to express the computational steps, the basic ideas are usually similar. The framework of probabilistic CBIR is flexible enough to capture a considerable subset of these approaches in a formally unified way.

### 3.3.1 Computational Analysis

In this section, we present an in-depth computational analysis for the main building blocks of our CBIR system (see Fig. 3.1) in terms of required arithmetic operations. This is a crucial step, since it allows to quantify the term *lightweight* and assess the practical usefulness of the approach in the context of the two retrieval scenarios we discussed in Section 3.3. In particular, we take a closer look at the feature transformation step, the feature representation step (which basically involves parameter estimation) and the similarity measurement or retrieval step. As a reference, we include a discussion of the computational steps of [40] since this is the closest relative to our approach. By the term *arithmetic operations*, we understand the number of additions & subtractions and multiplications & divisions (i.e. basic arithmetic operations) as well as the computationally expensive log, $e^x$ and $x^r$ operations with $x, r \in \mathbb{R}$. We further take into account any non-trivial operation, such as the evaluation of the Gamma $\Gamma$ or the Digamma $\psi$ function. To avoid numerical difficulties, we compute $\log \Gamma$ instead of $\Gamma$ at the cost of perhaps one additional exponentiation. The function values of $\log \Gamma$ and $\psi$ are obtained by employing a lookup-table approach with linear interpolation. Both, lookup and interpolation, can be performed with constant complexity and only require basic arithmetic (e.g. 5 additions & subtractions, 4 multiplications & divisions and 2 table-lookups in our implementation). Since we will also provide relative runtime measurements, all estimation methods as well as the similarity measurement routines are implemented in MATLAB to obtain comparable results. Runtime is measured on a Intel Core2 Duo 2.66Ghz system with 2GB of memory running MATLAB 7.6. We particularly emphasize that the focus is on relative runtime differences and not on absolute values.

#### Feature Transformation

Besides its advantages for image analysis (see Section 2.3), the DTCWT is appealing from a computational point of view since it can be implemented very efficiently by four parallel pyramidal DWTs using appropriate filter sets. Regarding memory requirements, the DTCWT is an overcomplete transform with a redundancy factor of four in case of images. In contrast to that, the DCT (e.g. used in [186]) is non-redundant, the Steerable Pyramid [168] (e.g. used in [39, 179]) is overcomplete by a factor of $4k/3$ ($k$ denotes the number of orientation subbands) and the Stationary Wavelet Transform (SWT) [136] (e.g. used in [34]) is overcomplete by a factor of $3J$, where $J$ denotes the maximum decomposition depth. The computational complexity of the DTCWT is linear $\mathcal{O}(N)$ in the number of input pixels $N$, since it basically requires computation of four parallel DWT decompositions which are of linear complexity. Hence, both DWT and DTCWT differ only by a constant factor. For comparison, the DCT, SWT, Steerable Pyramid and Gabor wavelets (when implemented in the frequency domain) have complexity $\mathcal{O}(N \log N)$. However, to be fair we have to note that in case of a block-based DCT with $8 \times 8$ blocks for example, the $\log N$ term carries no weight compared to a full-frame DCT.

#### Feature Representation/Parameter Estimation

Maximum-Likelihood parameter estimation for the Gamma and Weibull distribution requires a numerical root-finding algorithm to obtain estimates. Since we can determine the derivatives of the log-likelihood functions w.r.t. the relevant parameters in both cases, it is reasonable to use the Newton-Raphson algorithm due to its good convergence properties. However, optimal (i.e. quadratic) convergence is only possible if the starting value is close to the actual root. We attempt to fulfill this requirement by using moment estimates for the Gamma and Weibull model. Employing the Gumbel moment matching method with the corresponding parameter transformation in case of the Weibull distribution at least eliminates the issue of computationally intensive starting value calculation. We will refer to this approach as the *Weibull/Gumbel* approach and denote the direct ML estimation strategy as *Weibull (direct)*. In the latter case, we employ a $\alpha$-vs-$a_3$ lookup-table to obtain the starting value $\hat{\alpha}_1$. The exact computational requirements for moment matching will be discussed later. To get an impression of the computational demand in each iteration step of the Newton-Raphson algorithm, we determine the number of required arithmetic operations. For comparative reasons, we also provide the number of operations in case of the GGD ML estimation approach of [40] and the GGD estimation approach proposed by Song [171]. The starting value $\hat{c}_1$ for [40] is obtained by the method of Krupinski [91] and the starting value for the Newton-Raphson iteration of Song [171] is fixed to $\hat{c}_1 = 3$. We optimize computation in such a way, that terms (e.g. summations, logarithms, etc.) which occur repeatedly in an iteration step are only calculated once. Since many operations depend on the signal length N, we omit any additional constants for the sake of readability in these cases. The number of arithmetic operations per iteration and the runtime performance of the ML estimation procedures relative to the longest runtime (marked bold) are listed in Table 3.1. Further, Fig. 3.2 shows a boxplot of the mean estimation times over a set of reasonable parameter values for all ML estimation approaches. For each parameter value, ML estimation is repeated 100 times on $10^5$ random numbers drawn from the corresponding model.

| **Model** | $\pm$ | $\times, \div$ | $\lvert \cdot \rvert$ | $e^x, x^r$<br>log | $\psi, \psi'$ | Relative<br>Runtime |
|---|---|---|---|---|---|---|
| GGD, MLE [40] | 3N | 2N | N | 2N | 2 | 0.76 |
| GGD, Song [171] | 4N | 3N | N | 2N | | **1.00** |
| Weibull/Gumbel | 4N | 3N | | N | | 0.21 |
| Weibull (direct) | 4N | 2N | | 2N | | 0.62 |
| Gamma | 2N | 4 | | N | 2 | 0.21 |

**Table 3.1:** Number of arithmetic operations for one Newton-Raphson update step as a function of the signal length N.

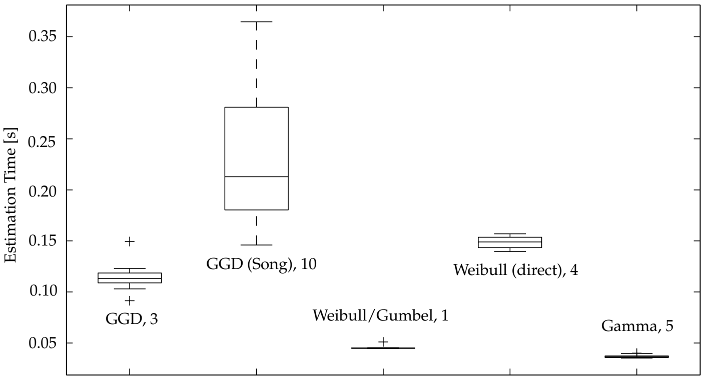


**Figure 3.2:** Boxplot of the mean ML estimation times over a set of parameter values. The y-axis shows the estimation time in seconds and the number in the annotation denotes the average iterations to reach convergence of the Newton-Rapshon algorithm.

As we can see, ML estimation using the Weibull/Gumbel approach shows the best performance, with only one iteration on average to reach convergence. The convergence criterion is met in case the absolute difference of two successive estimates is less than $10^{-6}$. In contrast, direct estimation of the Weibull parameters is less competitive, although we already use the $\alpha$-vs-$a_3$ lookup-table implementation. The higher number of iterations deteriorates the total runtime. The Gamma MLE procedure performs as good as the Weibull/Gumbel approach. Nevertheless, the number of iterations is the limiting factor again, since one Newton-Raphson update step in fact requires fewer arithmetic operations compared to the Weibull/Gumbel approach. As expected, the complex update step of the GGD ML estimation approach of [40] with more log, $x^r$, $e^x$ operations leads to an increase in computation time compared to the Weibull/Gumbel or Gamma case. Regarding the number of iterations, we confirm the results of [40] with three to four iterations on average to reach convergence. The estimation approach proposed by Song [171] exhibits the worst runtime performance of the experiment and a quite strong dispersion as well. A closer look at the number of iterations for each choice of the shape parameter $c$ reveals an average of 10 iterations for $c < 1.0$ which distorts the average. This seems reasonable, since the starting value of $\hat{c}_1 = 3$ is actually far-off the true value in these situations.

Next, we assess the number of arithmetic operations to compute moment estimates in case of the GGD, Gamma, Weibull/Gumbel and Weibull (direct) approach. As mentioned before, we use Krupinski's [91] fast approximation to obtain moment estimates for the GGD, an $\alpha$-vs-$\alpha_3$ lookup-table approach for Weibull (direct) moment estimates, Eq. (2.47) for Weibull/Gumbel moment estimates and Eq. (2.56) for Gamma moment estimates. A careful analysis of moment estimation is reasonable, since we use these estimates as a fast alternative to the MLEs in our retrieval experiments. The corresponding numbers of arithmetic operations are listed in Table 3.2. We emphasize, that this is the total effort to compute the parameter estimates. No iterative procedures are necessary and mostly basic arithmetic operations are performed. Only in case of Weibull/Gumbel moment estimation, the log operation is dependent on the signal length $N$. This is reflected in the relative runtime differences because log is an expensive operation compared to addition/subtraction or multiplication/division. The fast approximative GGD parameter estimation of [91] shows the best performance because the expensive computations like $\log \Gamma$, $e^x$ or log do not depend on the signal length $N$. Further, this approach apparently benefits from our lookup-table implementation of $\log \Gamma$. Regarding moment estimation of the Gamma parameters, we emphasize that this approach basically requires to compute the sample mean and sample standard deviation and hence performs at a competitive level compared to [91] as well.

| **Model** | $\pm$ | $\times, \div$ | $\lvert \cdot \rvert$ | $e^x, x^r$<br>log | $\log \Gamma$ | Relative Runtime |
|---|---|---|---|---|---|---|
| GGD [91] | 2N | N | N | 3 | 2 | 0.07 |
| Weibull (direct) | 4N | 2N | | | | 0.24 |
| Weibull/Gumbel, Eq. (2.47) | 3N | N | | N | | **1.00** |
| Gamma, Eq. (2.56) | 3N | N | | | | 0.17 |

**Table 3.2:** Number of arithmetic operations to obtain moment estimates for the model parameters as a function of the signal length $N$.

### Similarity Measurement/Retrieval

In the classic retrieval scenario, the similarity measurement part is most critical for runtime performance since each new query image requires computation of the similarity measure for all candidate images in the database. In case the statistical model parameters of the feature representations are estimated at the time of storage, the runtime performance of the whole retrieval operation is completely determined by the performance of the similarity measurement process. Although all presented KL divergences can be computed with constant complexity, it is worth taking a closer look at the required arithmetic operations. Given, that the statistical model pa-

| **Model** | $\pm$ | $\times, \div$ | $e^x, x^r$<br>$\log$ | $\log \Gamma$ | $\psi$ | Relative Runtime |
|---|---|---|---|---|---|---|
| GGD [40] | 6 | 10 | 3 | 4 | 0 | **1.00** |
| Gamma, Eq. (3.10) | 6 | 5 | 1 | 2 | 1 | 0.56 |
| Weibull, Eq. (3.9) | 8 | 9 | 8 | 1 | 0 | 0.31 |

**Table 3.3:** Number of arithmetic operations for KL divergence based similarity measurement.

rameters of an arbitrary wavelet subband are available for the query and all $L$ database images, we simulate a database search for $L = 10^4$. Table 3.3 lists the number of arithmetic operations for each KL divergence as well as the runtime relative to the longest runtime (marked bold). As we can see, the KL divergence for the GGD has the worst performance, due to the computations of $\log \Gamma$. The KL divergence of the Gamma model shows slightly worse runtime performance than the KL divergence for the Weibull model which can be attributed to computation of $\psi$ and the additional $\log \Gamma$. As a concluding remark, we note that since all KL divergences have a closed-form expression, no histogram computation and discrete version of the KL divergence is required. In practice, this is a huge advantage since we only have to store the model parameters and further avoid the search for a reasonable histogram binning.

### 3.3.2 Experiments

In this experimental section, we intent to cover three important issues: first, we address the impact of either using moment or ML estimates on the retrieval performance of the DTCWT based approaches. We additionally discuss this issue in the context of the approach of Do & Vetterli [40]. Second, we conduct a comparative study to three approaches from literature including the Gabor wavelet approach of Manjunath & Ma [117], the Local Binary Patterns proposed by Ojala et al. [141] and the popular MRSAR model of Mao & Jain [119]. In the following, we provide a brief description of these approaches as well as the exact parameter configuration we use for our experiments. Regarding the parameter configuration of our own retrieval approach, we use a three-scale DTCWT with Kingsbury's Q-Shift $(14, 14)$-tap filters for decomposition levels greater than two in combination with $(13, 19)$-tap near-orthogonal filters for the first decomposition level [87].

**Do & Vetterli, 2002** Basically, the idea of this approach is already explained in Section 3.3. Regarding the parameter configuration, we choose a three-scale DWT with the popular CDF 9/7 [32] filter. Parameter estimation is either accomplished by the fast moment matching method proposed by Krupinski [91] or the ML approach of Do & Vetterli [40]. In the labeling of our figures the approach is denoted by *DWT, GGD (Mom.)* or *DWT, GGD (MLE)*, resp., depending on the type of estimation method.

**Manjunath & Ma, 1996** The Gabor wavelets approach of Manjunath & Ma [117] is one of the pioneering approaches in the field of texture image retrieval. A Gabor wavelet decomposition is used to obtain a multi-resolution representation of an image at different scales and orientations. The important parameters of the Gabor wavelets are the upper $U_u$ and lower $U_l$ filter frequency which, in combination with the number of scales $J$ and orientations $O$, determine the exact filter configuration. The feature vector of an image consists of the mean and standard deviation of the transform coefficient magnitudes of each subband. Hence, a feature vector contains $J \times O \times 2$ elements. Image similarity is measured by the city-block distance between two feature vectors normalized by the standard deviations of the features. In our experiments, we use a configuration of $U_l = 0.04$, $U_u = 0.5$, $J = 3$ and $O = 6$.

**Ojala et al., 1996** In [141], Ojala et al. first introduce the concept of Local Binary Patterns (LBP) to capture texture information. The basic idea is to consider the pixel neighborhood of every pixel in an intensity image and extract a binary pattern from that. In a classic eight pixel neighborhood, we start from the top left-hand pixel (clockwise) and assign a '1' in case the intensity value is larger than the intensity value of the center pixel, or '0' otherwise. The resulting eight bits are then interpreted as a natural number in the range of $[0, 255]$ and a histogram over all LBPs is constructed. In our experiments, we use the standard eight pixel neighborhood and only consider those pixel as valid center pixel where all neighbors are inside the image boundary. No border extension is performed. Of course, other neighborhood definitions are possible as well and several extensions to the classic LBP approach have been proposed, e.g. see [112]. As a suitable distance measure between the LBP histograms of two images, the authors propose to use the histogram intersection metric.

**Mao & Jain, 1992** In [119], Mao & Jain introduce the Multiresolution Simultaneous Autoregressive (MRSAR) model to capture local pixel dependencies in an intensity image by a variant of Markov Random Fields. The basic idea is to estimate the intensity of a pixel from the local 8 pixel neighborhood by means of a Simultaneous Auto-Regressive (SAR) process. Four SAR parameters and the variance of the estimation error are estimated over a $N \times N$ pixel window, sliding by increments of $s$ pixel in the horizontal and vertical direction. Multiresolution is accomplished by increasing the neighborhood size (i.e. "pseudo" multiresolution) and repeating the estimation process. In our implementation we adhere to the neighborhood definition of [119]. Hence, given three resolution levels we finally obtain a 15-dimensional parameter vector per sliding window position. In the original work, the authors propose to determine the mean and covariance of the parameter vectors for feature representation and hence implicitly assume multivariate normality of the parameter vectors. The Mahalanobis distance is then suggested to measure the similarity between the feature representations of two images. We deviate from this setup and compute the Bhattacharya distance instead. In [191], Xu et al. have demonstrated superior retrieval performance using this metric. Regarding the parameter configuration, we use the resolution levels $1, 2$ and $3$, a sliding window of $21 \times 21$ pixel with $s = 4$ pixel increments and the method of least-squares to estimate the SAR parameters.

As a third and final issue of our experimental study, we take up the results of the computational analysis section and intend to give a guideline for lightweight retrieval. As an extension to the work of [104], we considerably enlarge our study to include experimental results for the Outex, Stex and Vistex (full) database (see Chapter 1). All images are first converted to the LUV colorspace and only the luminance (L) channel information is retained. The original $512 \times 512$ pixel versions of the textures are split into $B = 16$ non-overlapping subimages ($128 \times 128$ pixel) and each subimage is used as a query image once. The evaluation process of the retrieval system is discussed next.

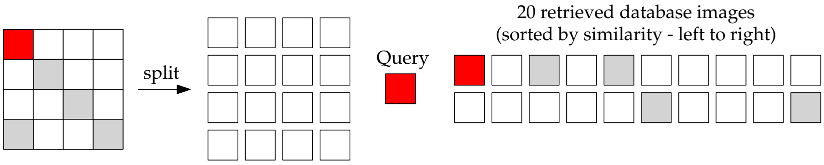

split
Query
20 retrieved database images
(sorted by similarity - left to right)

**Figure 3.3:** Procedure of splitting the 512 × 512 pixel images into 16 subimages of size 128 × 128; on the right, we see a query image (red) and the top 20 retrieval results. Those images belonging to the same parent as the query are marked light grey. In accordance to our evaluation criterion, the retrieval rate at operating point $K = 19$ is 31.25%.

### Evaluation Criterion

To evaluate the performance of the retrieval system we have to define a measure of retrieval *correctness*. We follow the common approach of counting the number of correct images among the top $K$ retrieved images, see [149, 111, 40, 191]. To capture this measure in a formal way, we let $\mathcal{P}_1, \dots, \mathcal{P}_N$ denote the $N$ parent images and let $\mathcal{I}_1, \dots, \mathcal{I}_L$ denote the images in the repository, obtained by the splitting process, i.e. $L = BN$. Further, we define a *parent indicator* function as

$$p : \{1, \dots, L\}^2 \to \{0, 1\}, \quad p(i,j) := \begin{cases} 1, \text{if } \mathcal{I}_i \text{ and } \mathcal{I}_j \text{ are splits of the same parent image} \\ 0, \text{else} \end{cases} \tag{3.11}$$

and let $R_j := \{r_1^{(j)}, \dots, r_L^{(j)}\}$ denote the index set of the sorted similarity values for the query image $\mathcal{I}_j$ to all $L$ candidate images (including the query itself). The percentage of correctly retrieved images for an arbitrary query image $\mathcal{I}_j$ at operating point $K$ can then be calculated as

$$s_K^{(j)} = \frac{1}{B} \sum_{i=1}^{K+1} p(j, r_i^{(j)}) \tag{3.12}$$

where the upper limit of the sum, $K+1$, accounts for the fact that the query image is not excluded from the set $R_j$. This of course assumes that the query is naturally defined to be most similar to itself (which is always the case in our setup). The final retrieval rate of the CBIR system at operating point $K$ – calculated on the basis that each database image is used as a query once – can then be determined by

$$S_K = \frac{1}{BL} \sum_{j=1}^{L} \sum_{i=1}^{K+1} p(j, r_i^{(j)}). \tag{3.13}$$

Since each image is split into 16 subimages in our setup, $B = 16$ for all reported results. Based on this evaluation setup, it is possible to construct Receiver Operating Characteristic (ROC) curves by plotting $K$ against $S_K$. This allows to study the retrieval behavior as we increase the number of retrieved images. For practical purposes, reasonable values of $K$ seem to be in the range of 16 to 40 images. To visualize the retrieval performance criterion, Fig. 3.3 illustrates the splitting process and shows an exemplary retrieval result for $K = 20$. As it is pointed out by Picard et al. [149], showing that a ROC curve of an approach lies above the ROC curve of another approach is a reasonable way to demonstrate a performance increase.

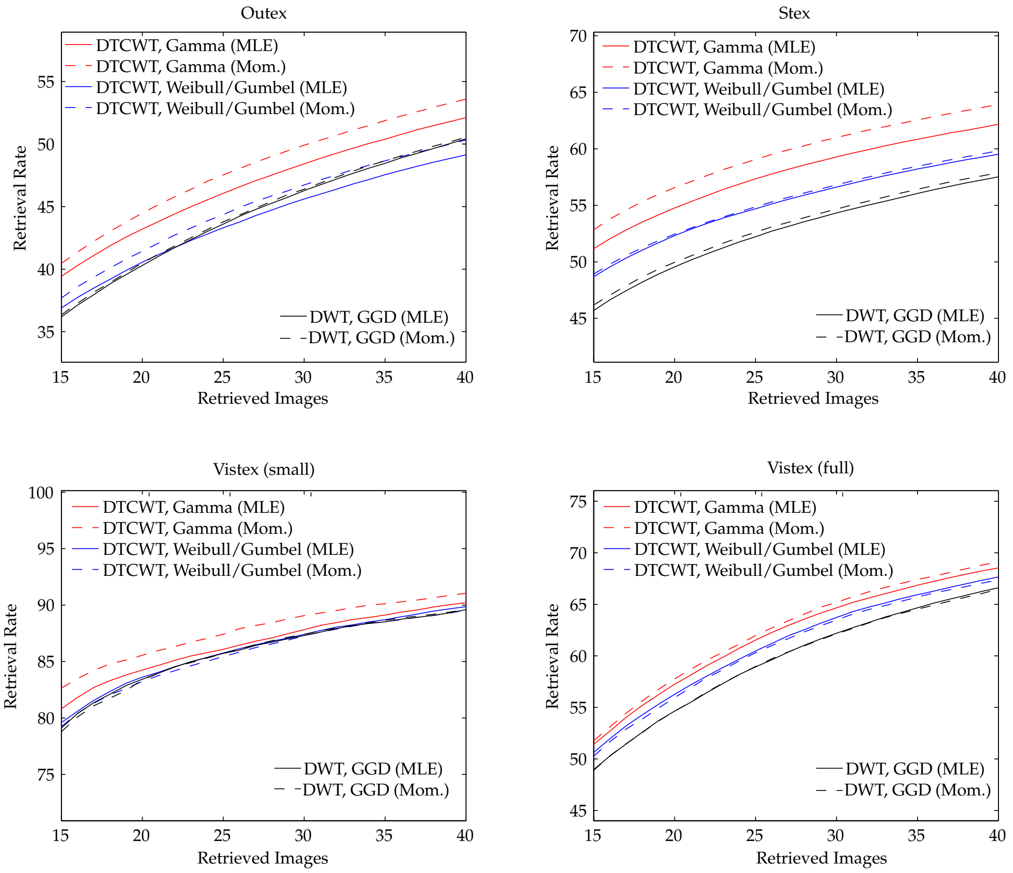


**Figure 3.4:** Retrieval rate comparison of the top 40 retrieved images for the DTCWT-based retrieval approaches and the strongly related DWT approach of Do & Vetterli [40]. The dashed lines denote the results obtained using moment estimates, the solid lines denote the results obtained by relying on ML estimates of the distribution parameters.

### Results

As a first experiment, we assess the retrieval performance of the DTCWT approaches w.r.t. the used estimation procedure. Fig. 3.4 shows the ROC curves for the Outex, Stex, Vistex (full) and Vistex (small) database. ROC curves corresponding to moment estimates are marked by a dashed line, whereas ROC curves corresponding to ML estimates are marked by a solid line. The first observation we make is that the Gamma model apparently leads to the top retrieval performance no matter whether we use moment matching or ML estimation. This can also be confirmed by taking a look at the top $K = 16$ retrieval results listed in Table 3.4. We further observe that moment estimation does in no case lead to notably worse retrieval performance. In some cases, the moment matching approach leads to even better retrieval performance. This result is consistent with the observations we made in [104]. From a computational point of view, this is a rather appealing observation, since it allows to replace the computationally demanding procedure of ML estimation by the considerably faster moment estimation approach without sacrificing retrieval rate.

| Approach | Outex | Stex | Vistex (small) | Vistex (full) |
|---|---|---|---|---|
| DTCWT, Gamma (MLE) | 39.41 | 51.16 | 80.82 | 51.42 |
| DTCWT, Gamma (Mom.) | **40.45** | **52.84** | **82.65** | **51.76** |
| DTCWT, Weibull/Gumbel (MLE) | 36.90 | 48.69 | 79.59 | 50.63 |
| DTCWT, Weibull/Gumbel (Mom.) | 37.69 | 48.91 | 79.25 | 50.27 |
| DWT, GGD (Mom.) | 36.35 | 46.19 | 78.79 | 48.90 |
| DWT, GGD (MLE) | 36.18 | 45.70 | 79.11 | 48.97 |

**Table 3.4:** Retrieval rates at the operating point of $K = 16$ retrieved images for the different statistical models (and estimation strategies) on four texture databases. The top results are marked bold.

| Approach | Outex | Stex | Vistex (small) | Vistex (full) |
|---|---|---|---|---|
| DTCWT, Gamma (Mom.) | $40.45_{(2)}$ | $52.84_{(2)}$ | $82.65_{(3)}$ | $51.76_{(3)}$ |
| Do & Vetterli, 2002 | $36.35_{(4)}$ | $46.19_{(5)}$ | $78.79_{(4)}$ | $48.90_{(4)}$ |
| Manjunath & Ma, 1996 | $26.86_{(5)}$ | $46.73_{(4)}$ | $67.86_{(5)}$ | $39.57_{(5)}$ |
| Mao & Jain, 1992 | $\mathbf{45.91}_{(1)}$ | $\mathbf{61.51}_{(1)}$ | $\mathbf{90.19}_{(1)}$ | $\mathbf{63.77}_{(1)}$ |
| Ojala et al., 1996 | $38.50_{(3)}$ | $52.24_{(3)}$ | $83.67_{(2)}$ | $55.05_{(2)}$ |

**Table 3.5:** Retrieval rates at the operating point of $K = 16$ retrieved images on four databases; the rank of each approach is listed in parentheses and the top approaches are marked bold.

As a next point, we take a closer look at the competitiveness of the DTCWT, Gamma (Mom.) approach in comparison to the approaches of [40, 141, 117] and [119]. Fig. 3.5 shows the corresponding ROC curves. As we can see, the top performance is achieved by the MRSAR approach of Mao & Jain [119] in all cases. However, the MRSAR approach is also the most computationally expensive one, both in terms of parameter estimation and similarity measurement. The least-squares procedure to estimate the 15 MRSAR parameters is rather time consuming and computation of the Bhattacharya divergence requires considerably more time compared to the other similarity measures we use here. Especially for retrieval scenario B, the prerequisite to compute the expensive model parameter estimation procedure for each query image limits the usability of the MRSAR approach. Regarding retrieval scenario A, estimation is a less critical issue and similarity measurement can be speed-up by using the approximation to the Bhattacharya divergence proposed by Comaniciu et al. [24]. In Fig. 3.5, we further observe that the standard LBP approach of Ojala et al. [141] is quite competitive in terms of retrieval performance. Computation of the LPBs can be performed very efficiently in the spatial domain and histogram intersection basically requires one pass through the one-dimensional LBP histogram. The DTCWT, Gamma (Mom.) approach exhibits almost the same retrieval rate as the LBP approach with slightly higher rates on Stex and Outex. Finally, we highlight that we consistently achieve better retrieval rates than the DWT, GGD (Mom.) approach and the Gabor wavelets, no matter which database we consider. The only true competitor in terms of computational performance and retrieval rate is the LBP approach of Ojala et al. The detailed retrieval results at the operating point of $K = 16$ retrieved images are listed in Table 3.5.

Another interesting observation can be made by looking at the results of Fig. 3.5. Apparently, it is inadvisable to judge the quality of a retrieval approach solely based on the results

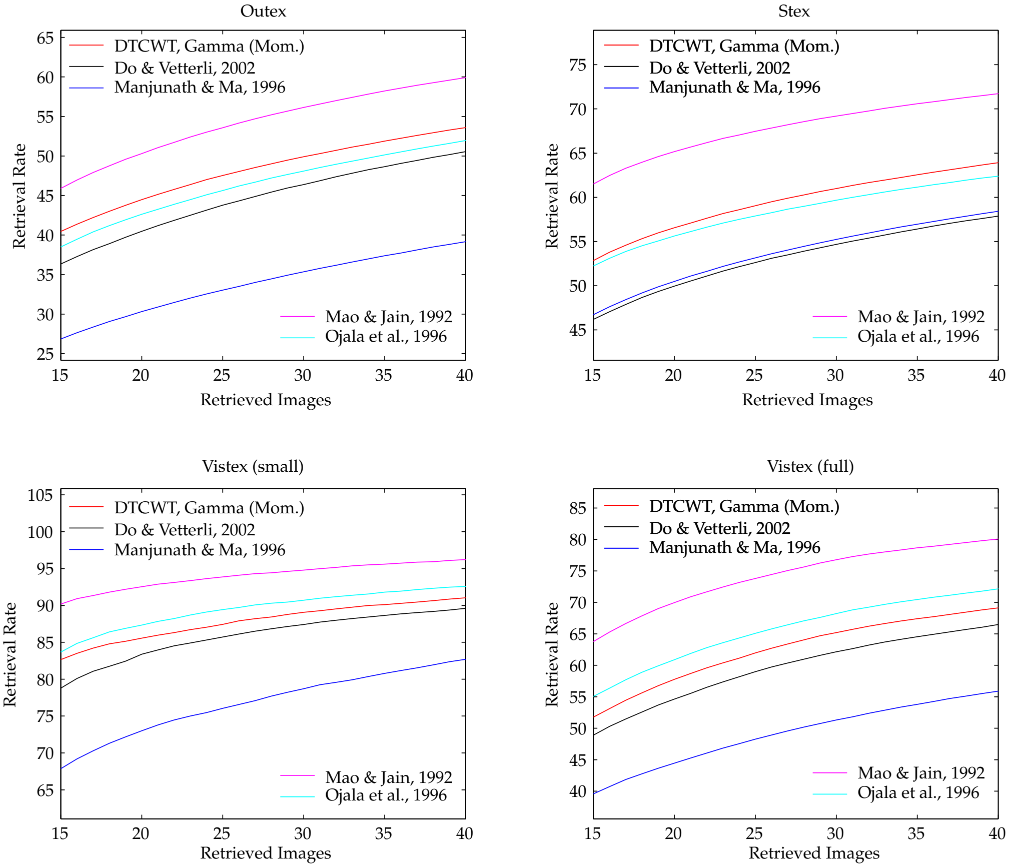


**Figure 3.5:** Retrieval rate comparison of the top 40 retrieved images.

obtained on just one image database, especially when the number of images is small. Although the Vistex (small) database is widely-used in the literature on texture image retrieval as a popular test set, we point out that the results might convey a wrong notion of total and relative retrieval performance. It is even possible that the overall ranking of the approaches changes from database to database. In Table 3.5, we highlight this fact by listing the ranks of the approaches in subscripted parentheses. As another example, consider the difference between the retrieval results obtained on Stex and Vistex (small). The margin between the DWT, GGD (Mom.) and DTCWT, Gamma (Mom.) approach is rather small on Vistex (small) while we observe a considerable margin of $\approx 7$ percentage points on Stex. We conclude, that statements about the ranking of different retrieval approaches are only convenient in case the study is conducted on at least two databases of reasonable size.

## 3.4 Copula-Based Retrieval

For the lightweight texture retrieval approach of the last section, we relied on the assumption of transform coefficient independency across subbands of the same scale and subbands of different scales. Further, the approach is tailored for singe-channel (e.g. grayscale) images, since the statistical models cannot capture information (e.g. association structure) between different subbands or channels. In this section, we present a novel retrieval approach which incorporates the association of DTCWT transform coefficients across subbands and color channels into the feature representation. The approach was first introduced in [103] and relies on the copula models of Section 2.3.3. The feature transformation is the DTCWT and we consider all available subbands of a specific decomposition level. In case of color images, a feature vector $\mathbf{z}$ contains $B = 18$ elements, where each element is a transform coefficient $z_i = |z_i|$ from one subband, i.e. $\mathbf{z} = (z_1, \ldots, z_B)$. Hence, according to Eq. (2.63) the joint p.d.f. of $\mathbf{Z}$ can be written as

$$p_{\mathbf{Z}}(\mathbf{z}; \theta_1, \ldots, \theta_B, \mathbf{\Theta}) = c(F_1(z_1; \theta_1), \ldots, F_B(z_B; \theta_B); \mathbf{\Theta}) \cdot \prod_{i=1}^{B} f_i(z_i; \theta_i) \tag{3.14}$$

where $c$ denotes the copula p.d.f. and $f_i$ denotes the p.d.f. of the $i$-th margin. In our setup, the type of copula is restricted to a Gaussian or Student t copula and the marginal distributions $F_i$ are limited to Weibull or Gamma. A concrete example of such a joint statistical model is a Gaussian copula with Weibull margins. The corresponding p.d.f. is given as

$$p_{\mathbf{Z}}(\mathbf{z}; \mathbf{R}, \alpha_1, \beta_1, \ldots, \alpha_B, \beta_B) = \\ \frac{1}{|\mathbf{R}|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^T(\mathbf{R}^{-1} - \mathbf{1})\boldsymbol{\xi}\right) \prod_{i=1}^{B} \frac{\alpha_i}{\beta_i}\left(\frac{z_i}{\beta_i}\right)^{\alpha_i - 1} \exp\left\{-\left(\frac{z_i}{\beta_i}\right)^{\alpha_i}\right\} \tag{3.15}$$

with $\boldsymbol{\xi} = [\Phi^{-1}(F_1(z_1; \alpha_1, \beta_1)) \cdots \Phi^{-1}(F_B(z_B; \alpha_B, \beta_B))]$. The parameters of the copula model are estimated by the IFM method we discussed in Section 2.3.3. Although, it is reasonable to incorporate as much information as we can into the feature representation of each image, we run into problems when it comes to similarity measurement. In the previous section, we have seen that the independency assumptions allowed to derive closed-form expressions for the Kullback-Leibler divergence between two feature representations. In case of copula-based models however, no such closed-form expressions exist and we have to rely on alternative strategies. A first pragmatic approach we employed in [103] is to exploit the "Monte-Carlo" approximation of the KL divergence. In particular, the KL divergence between two p.d.f.s $f$ and $\tilde{f}$ can be written as

$$D(f||\tilde{f}) = \mathbb{E}_f[\log f(x) - \log \tilde{f}(x)] \tag{3.16}$$

where $\mathbb{E}_f$ denotes the expectation w.r.t. $f$. Hence, we can approximate $D(f||\tilde{f})$ by drawing a random sample $x_1, \ldots, x_n$ from the model density $f(x)$ and then calculate

$$D_{MC}(f||\tilde{f}) \approx \frac{1}{n}\sum_{i=1}^{n}\left(\log f(x_i) - \log \tilde{f}(x_i)\right) \tag{3.17}$$

which converges to Eq. (3.16) as $n \to \infty$. Unfortunately, this approach has two inherent disadvantages: first, due to the "Monte-Carlo" nature of the approximation, the KL divergence will differ to a certain extent (depending on $n$) each time we compute the similarity between two feature representations. Second, the approach is computationally expensive since we need to estimate the joint statistical model for each query image, draw a random sample and compute the likelihood. As we have shown in [103], the Monte-Carlo approximation is rather stable even for small values of $n$ (e.g. $n = 10^3$). However, the computational burden of estimation and sampling still remains. As a second, and presumably more reasonable alternative to measure

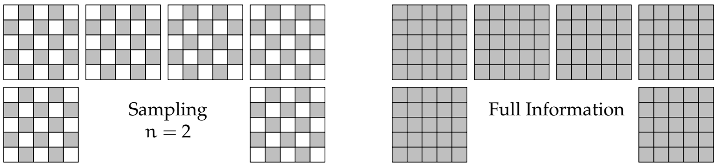


**Figure 3.6:** Information reduction by means of sampling every $n$-th coefficient.

similarity between two copula-based feature representations, we propose to employ the ML selection rule of the probabilistic CBIR framework, see Eq. (3.3). This is a natural choice, since it does not require sampling nor parameter estimation of the query image's feature representation. Given a collection of query feature vectors $\mathbf{z}_1^*, \ldots, \mathbf{z}_K^*$, the ML selection rule can be written as

$$g(\mathbf{z}_1^*, \ldots, \mathbf{z}_K^*) = \arg \min_{r \in \{1,\ldots,L\}} \sum_{i=1}^{K} \log p_{\mathbf{Z}}(\mathbf{z}_i^*; \theta_1^{(r)}, \ldots, \theta_B^{(r)}, \mathbf{\Theta}^{(r)}) \tag{3.18}$$

where $\theta_i^{(r)}$ denotes the parameter (vector) of the $i$-th marginal distribution of candidate image $\mathcal{I}_r$ and $\mathbf{\Theta}^{(r)}$ denotes the corresponding copula parameter (vector). The number of available query feature vectors $K$ depends on the number of subband coefficients. Due to the fact that we consider only the subbands of one particular decomposition level, $K$ is constant. However, for the computation of Eq. (3.18), we have to evaluate the marginal p.d.f.s as well as the multivariate copula p.d.f. for each query feature vector. Hence, it seems reasonable to limit the number of query feature vectors. Especially in case of the Student $t$ copula this can be a critical issue as we will see later on. Similar to the query feature vector extraction strategy presented by Vasconcelos & Lippman in [187], we suggest a coefficient reduction step by sampling every $n$-th transform coefficient (see Fig. 3.6 for a visualization of $n = 2$) and therefor reduce the data rate by a factor of $1/n$. This will speed up the ML selection process, however it might also negatively affect the retrieval rate.

### 3.4.1 Experiments

In order to compare the copula retrieval approach to existing approaches in literature, we select two approaches which were originally designed to deal with color (texture) images and do not have to be artificially extended (e.g. by feature vector concatenation). We test against the original CBIR approach of Vasconcelos & Lippman [186] and a very recently proposed approach by Verdoolaege et al. [188]. The general principles of both approaches are briefly discussed next, including the parameter configurations we use for our experiments.

**Vasconcelos & Lippman, 2000** To a large extent, this approach has already been discussed in Section 3.2. To handle color texture images, we implement the original interleaving strategy: first, the image is converted to YBR colorspace and the color channels are decomposed separately by a 2-D DCT. Then, the sliding window approach is used to extract the first $D$ coefficients of each window which are interleaved according to the pattern $YBRYBR\ldots$. Hence, we obtain $(3 \cdot D)$-dimensional feature vectors. The only point in which our implementation differs from the original work, is the actual retrieval part. Instead of employing the ML selection rule, we rely on an approximation of the KL divergence between Gaussian mixture models, proposed by Goldberger et al. [57]. Regarding the final parameter configuration, we use $C = 8$ mixture components and extract $D = 16$ coefficients. During the EM algorithm, the (diagonal) covariance matrices are regularized by a small positive constant $\epsilon > 0$ to ensure positive definiteness.

**Verdoolaege et al., 2008** An extension to the work of Do & Vetterli [40] is presented by Verdoolaege et al. [188] with the objective to allow color texture retrieval. The color channels of an RGB image are first decomposed separately by a J-scale DWT. Coefficients from corresponding subbands in the decomposition structure but from different color channels are then modeled by MPE distributions (see Section 2.2.3) with fixed shape parameter $\beta$. The authors assume independence of the horizontal, vertical and diagonal subband coefficients as well as independence across scales and hence obtain a $(J \times 9) \times (J \times 9)$ block-diagonal matrix $\boldsymbol{\Sigma}$ as the only parameter of the MPE model. Due to a missing closed-form expression for the KL divergence between two MPE distributions, Verdoolaege et al. derive a closed-form expression for the geodesic path between two MPE distributions on the corresponding statistical manifold. The parametrization of this approach for our experiments is as follows: we use a three-scale DWT decomposition and the parameter $\boldsymbol{\Sigma}$ is estimated by the method of moments, introduced in Section 2.2.3. As in the original work, $\beta$ is fixed to 0.5.

Regarding the parameter configuration of our copula-based approaches, we choose the transform coefficients of all detail subbands of DTCWT decomposition level three and use the standard RGB colorspace. Since the query images are $128 \times 128$ pixel, we obtain $R = 256$ coefficient vectors which are all used to perform image queries (i.e. $K = R$). Regarding the choice of copula, we have to make some restrictions for the following computational reasons: first, we note that estimation issues are no limiting factor for both the Student t or Gaussian copula since estimation can be performed offline in case of the classic retrieval scenario (i.e. scenario A). Estimation of the correlation matrix $\mathbf{R}$ in the Gaussian case is straightforward and can be computed efficiently. Estimation of the Student t copula parameters $\nu$ and $\mathbf{R}$ is somewhat more involved, but comparable to the effort required to estimate the Gaussian mixture model parameters of [186] or the MPE parameter $\boldsymbol{\Sigma}$ [188]. However, the bottleneck of Student t copula approach is the similarity measurement step, i.e. the computation of the ML selection rule. In particular, we face the problem to calculate the univariate Student t quantiles $t_\nu^{-1}$ (see e.g. [70]) for all elements of each query feature vector. Except for a few special cases of $\nu$, this computation is quite numerically involved and far more complex than the evaluation of the Gaussian quantile function $\Phi^{-1}$ (which basically requires evaluation of the inverse complementary error function [1]). Especially for the large image repositories Outex, Stex and Vistex (full), this computational disadvantage renders the Student t copula impractical. As a consequence, we restrict the presentation of the experimental results to the Gaussian copula and only exemplary show retrieval results of the Student t copula in case of Vistex (small). We further note, that the data reduction strategy we suggest in Fig. 3.6 does not remedy the computational problems of the Student t copula approach. In order to achieve comparable runtime behavior to the Gaussian copula model we have to reduce the number of query feature vectors to a point where the retrieval rate drops below reasonable levels.

### Results

First, we present a ROC curve comparison for the two types of copula and the two types of marginal distributions, see Fig. 3.7 (left). We observe that the ROC curves of the Student t and

| Approach | Outex | Stex | Vistex (small) | Vistex (full) |
|---|---|---|---|---|
| Copula (Gaussian, Weibull) | $44.03_{(2)}$ | $70.64_{(1)}$ | $89.54_{(2)}$ | $63.01_{(2)}$ |
| Copula (Gaussian, Gamma) | $43.35_{(3)}$ | $69.37_{(2)}$ | $89.12_{(3)}$ | $61.92_{(4)}$ |
| Verdoolaege et al., 2008 | $29.89_{(4)}$ | $63.66_{(3)}$ | $89.72_{(1)}$ | $62.29_{(3)}$ |
| Vasconcelos & Lippman, 2000 | $54.66_{(1)}$ | $65.44_{(4)}$ | $87.71_{(4)}$ | $65.12_{(1)}$ |

**Table 3.6:** Retrieval results at the operating point of $K = 16$ retrieved images on four texture databases.

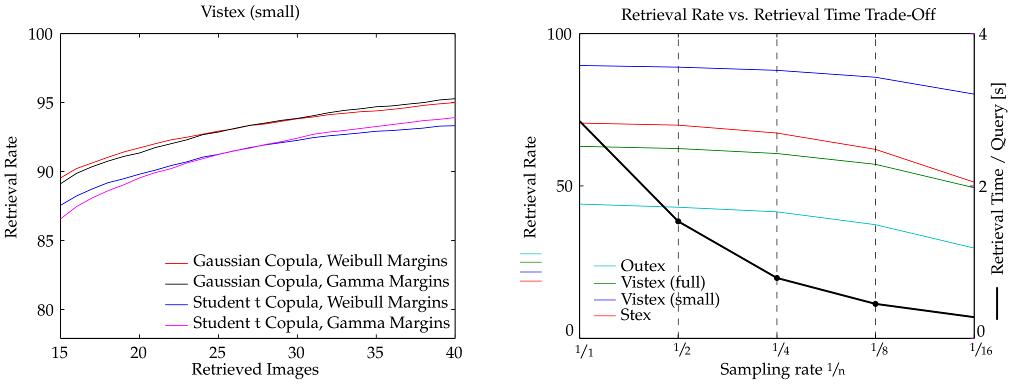


**Figure 3.7:** ROC curve comparison for joint statistical models relying on either a Gaussian or a Student t copula (left); Visualization of the trade-off between retrieval rate (using 16 retrieved images) and retrieval time as a function of the data reduction rate $1/n$ (right).

Gaussian copula are grouped together and there is a slight margin between the two groups. In accordance to the GoF analysis of Section 2.3.4, the models based on the Gaussian copula lead to better retrieval performance than the models relying on the Student t copula. This is also very convenient from a computational point of view, since the Gaussian copula is substantially easier to handle with respect to parameter estimation and likelihood computation.

In the second experiment, we fix the Gaussian copula and compute ROC curves (see Fig. 3.8) for a comparative study to [186] and [188]. From Fig. 3.8, we observe that the Gaussian copula with Weibull margins is consistently ranked among the top two approaches and performs best on Stex. The difference in using either Gamma or Weibull margins is neglectable in all cases. This is again a computational advantage for ML selection, since computation of the Weibull c.d.f., see Eq. (2.34), is straightforward due to a closed-form expression. In contrast, evaluation of the Gamma c.d.f., see Eq. (2.51), involves computation of the regularized incomplete Gamma function [1]. Regarding the overall ranking of the approaches, it is hard to identify a clear winner. We observe a situation similar to Fig. 3.5 where the ranking is not consistent over all databases. The approach of [186] for example is ranked first on Outex and Vistex (full), however exhibits worse retrieval performance on Stex and Vistex (small). The retrieval results at the operating point of 16 retrieved images are listed in Table 3.6, including the corresponding ranks of each approach (in subscripted parentheses). The fluctuations in the rankings highlight the requirement for large-scale tests on more than one database yet another time. Unfortunately, such studies often have to be omitted in research papers due to space limitations.

In a final experiment, we study the impact of reducing the number of query feature vectors

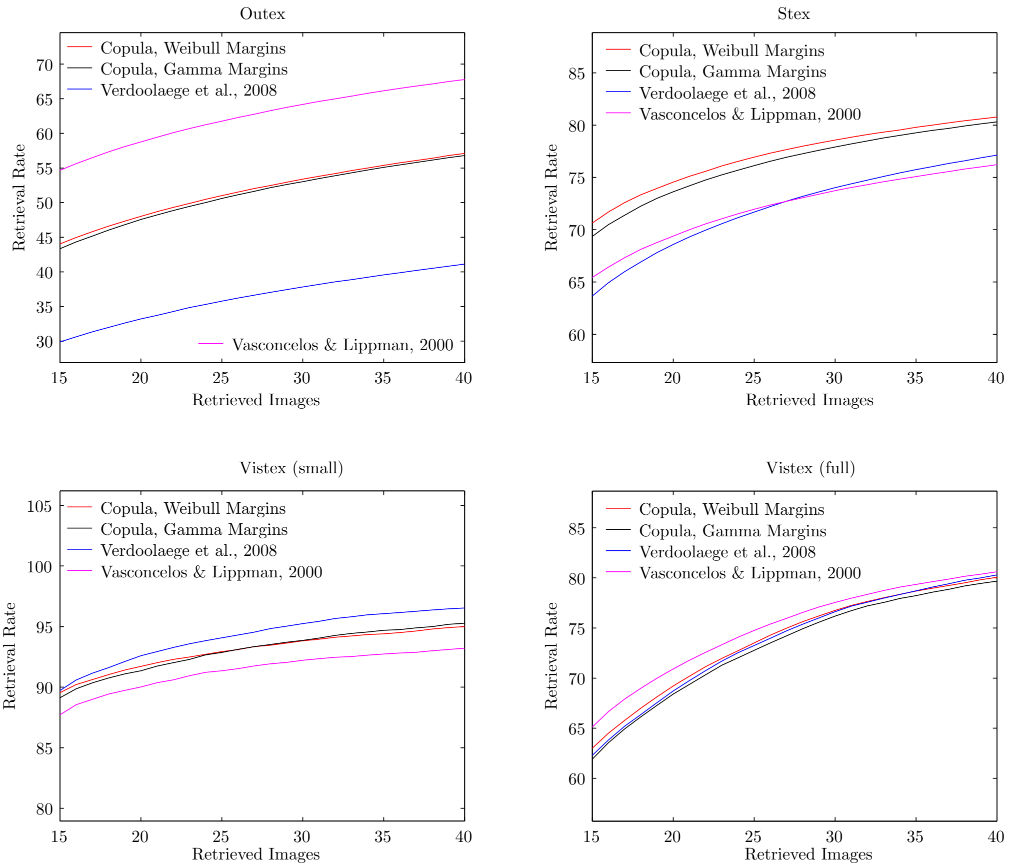


**Figure 3.8:** ROC curve comparison of the (Gaussian) copula-based CBIR approaches to the works of [188] and [186] for 40 retrieved images.

for the computation of the ML selection rule by a factor $^1/_n$. As noted before, the total number of available query feature vectors on DTCWT level three is $R = 256$ which corresponds to $n = 1$. We select the Gaussian copula with Weibull margins for the following experiment and let $n$ take on powers of two, i.e. $n \in \{1, 2, 4, 8, 16\}$. In order to illustrate the performance gain in retrieval time, we measure the time it takes to perform one query on a database of 1024 images. ML selection is implemented in ANSI C and runtime is measured on a 64bit Intel Xeon 2.27Ghz Quad-Core system with 24Gb of memory running Linux 2.6.18. The right part of Fig. 3.7 visualizes the retrieval rate at the operating point of 16 retrieved images in direct comparison to the retrieval time per query as a function of the data reduction rate $^1/_n$. The slope of the bold black line illustrates the decrease in retrieval time as we reduce the number of query feature vectors. One of the first things we notice is, that the slope of the retrieval rate is similar for all databases. We suppose that it could even be possible to fit a suitable function to the retrieval rate curve of one database and predict the retrieval rate decrease for the other databases. Next, we observe that the drop-off in retrieval rate for $n = 2$ is only $\approx 0.7$ percentage points, although we achieve a considerable performance gain in retrieval time of almost 50%. In consideration of the fact that retrieval time drops to $\approx 25\%$ for $n = 4$, even the decrease in retrieval rate of $\approx 2.5$ percentage points on average seems acceptable. In general, the final setting of $n$ will depend on the field of application. Nevertheless, we have shown that a significant speedup in computation time can be achieved by using fewer query feature vectors for similarity measurement while keeping the retrieval rate at a high level. Eventually, we eliminated the inherent disadvantage of our Monte-Carlo similarity measurement strategy suggested in [103]. This renders the copula-based CBIR approach applicable even on large databases.

## 3.5 Discussion

In this chapter, we introduced two novel retrieval approaches for texture images. In the first part, we focused on a lightweight approach to allow application in computationally demanding retrieval scenarios. In the second part, we showed that the incorporation of additional information about the association structure between DTCWT coefficients leads to a considerable increase in retrieval rate, however, at the cost of runtime performance. By introducing a simple data reduction strategy, we could enhance runtime performance while keeping the retrieval rate at almost the same level, at least for reasonable reduction rates. As a matter of fact, this enables deployment of the copula-based approaches in retrieval scenario A, even on large databases. Nevertheless, the complexity of the ML selection process still seems too high for scenario B. In contrast to that, the DTCWT, Gamma (Mom.) approach is perfectly suitable when runtime performance is a crucial issue. The computational complexity of the DTCWT is linear in the number of input pixels. Further, the feature representation only requires to determine the moment estimates (linear complexity) of the Gamma distribution and image similarity can be computed in constant time.

A second remarkable observation we made throughout all experiments is the low consistency of the rankings of the approaches with respect to the image databases. The relative difference in retrieval rate between two approaches tends to vary considerably as well. We conclude, that it is not reasonable to claim superiority of an approach in a comparative study by presenting results on just a few example textures of one database. These results can usually not be generalized to other databases. It is even possible that the situation might be completely different when changing the image set. For that reason, we strongly argue to adopt the strategy of testing on at least two or three texture databases in any experimental study. On the one hand, this enhances the quality of the presented results and on the other hand conveys an impression about the suitability of an approach with respect to different kinds of image sets.

As a last part of this discussion, we raise the question whether the criterion of splitting a set of texture images into equally sized parts and using each part as a query is the most suitable way to evaluate the quality of a texture retrieval system. Although, this has become the de-facto standard for evaluation, there is an inherent drawback: lets assume, that two parent images basically show the same visual content, e.g. the same surface material. In such a situation, it is possible that a retrieved image is perceptually almost identical to the query but stems from another parent and is hence classified as a wrong retrieval result. Taking a closer look at the images of the Outex database reveals, that this is exactly the reason for the rather low retrieval rates. As a consequence, the ROC curves might convey a wrong impression about the actual quality of an approach. The strategy of having a number of predefined categories might be a possible alternative here, although it seems hard to establish a categorization for a large number of images. At which point for example are the images of a category too *different* so that we have to create two separate categories? Further, category assignments will inevitably differ from user to user because of differences in visual perception. For these reasons, we consider the establishment of a suitable retrieval evaluation setup as an important issue of future research.

# Chapter 4

# Medical Image Classification

In this chapter, we discuss a classification problem in the field of medical image analysis. In particular, we are concerned with the prediction of the histopathological diagnosis of colorectal lesions, based on the mucosal surface structures which can be observed in High Magnification Chromoscopic Colonoscopy (HMCC). Our focus is on methods which employ wavelet coefficient statistics as a primary source to construct image features for classification. We will see, that this classification problem is strongly related to the texture retrieval setting of the previous chapter. Actually, towards the end of our discussion, we show that considering the classification problem from the viewpoint of probabilistic image retrieval leads to an elegant solution with respect to scalability and computational cost. Major parts of this chapter recently appeared in:

[98] R. Kwitt and A. Uhl. Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images. In *Proceedings of the IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'07)*, pages 1–8, Rio de Janeiro, Brasil, 2007

[102] R. Kwitt and A. Uhl. *Multi–Directional Multi-Resolution Transforms for Zoom–Endoscopy Image Classification (Best Paper Award at CORES 2007)*, volume 45 of *Advances in Soft Computing*, pages 35–43. Springer, 2008

[64] M. Häfner, R. Kwitt, A. Uhl, A. Gangl, F. Wrba, and A. Vecsei. Feature-extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images. *Pattern Analysis and Applications*, 12(4):407–413, December 2009

[99] R. Kwitt and A. Uhl. Color eigen-subband features for endoscopy image classification. In *Proceedings of the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 589–592, Las Vegas, Nevada, United States, 2008

[100] R. Kwitt and A. Uhl. Color wavelet cross co-occurrence matrices for endoscopy image classification. In *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP'08)*, pages 715–718, St. Julians, Malta, 2008

The chapter is structured as follows: in Section 4.1, we present the medical perspective of our problem and discuss related work on the topic of computer-assisted diagnosis of colorectal cancer. Section 4.2 then introduces three novel feature extraction approaches in a discriminant classifier setup. An alternative approach, based on the idea of generative models is proposed in Section 4.3. Eventually, we discuss the evaluation setup and present an extensive comparative study on classification/prediction performance in Sections 4.4 and 4.5. The chapter concludes with a discussion of the main contributions and an outlook on future research.

## 4.1 The Medical Presentation of the Problem

According to the statistics of the American Cancer Society[1], colorectal cancer is the third most commonly diagnosed cancer and the third leading cause of US cancer deaths in both men and women. Colorectal cancer is a paramount example where existing knowledge in combination with early screening procedures can prevent death and save lives. Computer-aided diagnosis systems have gained considerable research interest recently. A lot of work has been done on the automated discrimination between normal and cancerous tissue using microscopic imaging, mainly by means of texture analysis [44, 166]. While these studies work directly with tissue samples of resected specimen obtained from biopsies, other works have studied the versatility of endoscopic video-frame processing for the detection of colorectal polyps [79, 122] and the assessment of colorectal abnormalities [90, 78, 75, 74]. However, conventional white-light video colonoscopy as it is used in these studies has its limitations, especially with respect to the detection of flat and depressed lesions [73]. The emergence of high-magnification chromoscopic colonoscopy (HMCC) poses several advantages over white-light video colonoscopy. In HMCC, high-magnification endoscopes with zoom-factors of up to 150× are used to visualize the appearance of the colon mucosa. The high optical zoom and resolution reveal characteristic surface patterns (i.e. Pit Patterns) which can be analyzed by the experienced physician to predict the histological diagnosis. This visual inspection is guided by the Kudo criteria for Pit Pattern analysis. Usually, chromoagents such as indigo-carmine or methylene-blue are used during endoscopic examination to enhance the visual appearance of the observed tissue. As a matter of fact, HMCC is suggested as an *in vivo* staging tool to enhance the diagnostic process and guide therapeutic strategies.

### 4.1.1 Pit Pattern Analysis

Colorectal cancer predominantly develops from adenomatous polyps (adenomas), although adenomas do not inevitably become cancerous. Polyps of the colon are a frequent finding and are usually divided into metaplastic, adenomatous or malignant. Since the resection of all polyps is rather time-consuming, it is imperative that those polyps which warrant resection can be distinguished. The classification scheme presented by Kudo et al. [92] divides the mucosal crypt patterns into five types (Pit Patterns I–V). Fig. 4.1 provides a schematic illustration of the different Pit Patterns and Table 4.1 gives a textual description of their visual appearance. Exemplary HMCC images are shown in Fig. 4.2. While Pit Patterns I and II are characteristic of benign lesions and represent normal colon mucosa or hyperplastic polyps (i.e. non-neoplastic lesions), Pit Patterns III to V represent adenomatous and carcinomatous structures (i.e. neoplastic lesions).

[1] `http://www.cancer.org` (accessed on March, 19th, 2010)

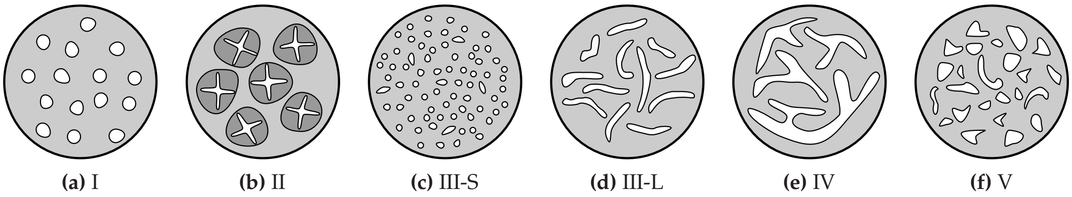


**Figure 4.1:** Schematic illustration of the six colorectal crypt architectures (i.e. Pit Patterns), according to the Kudo criteria [92].

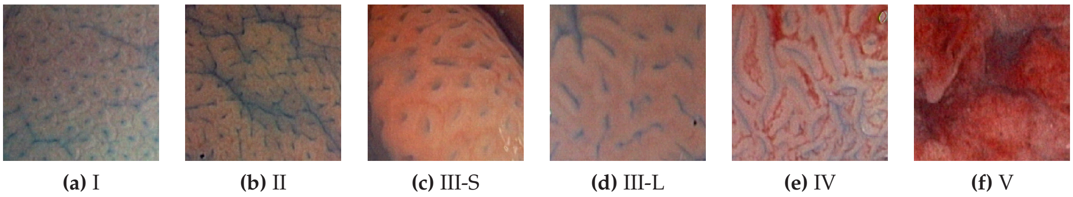


**Figure 4.2:** Representative HMCC images of the different Pit Patterns. Note that the types I and II show non-neoplastic lesions while III-L, III-S, IV and V show neoplastic disease.

| Pit Pattern | Visual Appearance |
|---|---|
| I | Round pit (normal pit) |
| II | Asteroid pit, stellar or papillary |
| III-S | Tubular or round pit, smaller than type I pit |
| III-L | Tubular or round pit, larger than type I pit |
| IV | Dendritic or gyrus-like pit |
| V | Irregular arrangement and sizes of III-S, III-L, IV |

**Table 4.1:** Description of the visual appearance of the colorectal crypt patterns observed during HMCC.

At first sight, the Kudo criteria seems to be straightforward and easy to be applied. Nevertheless, it needs some experience and exercising to achieve good results. Correct diagnosis very much relies on the experience of the gastroenterologist as the interpretation of the Pit Patterns may be challenging [72]. Computer-assisted diagnosis is motivated by the work of Kato et al. [81], where the authors state that assessing the type of mucosal crypt patterns can actually predict the histological findings to a very high accuracy. Regarding the correlation between the mucosal Pit Patterns and the histological findings, several (human-based) studies report good results for distinguishing non-neoplastic from neoplastic lesions, although with different diagnostic accuracies. A recent comparative study by Kato et al. [80] reports a prediction accuracy of 99.1% by means of HMCC and Pit Pattern analysis. Hurlstone et al. [73] claim a rate of approximately 95%, Tung et al. [178] claim 80.1%, however, at very low sensitivity of only 64.6%. In another work, Fu et al. [50] report 95.6% for HMCC compared to 84.0% using conventional white-light colonoscopy and 89.3% using chromoendoscopy without magnification. An even larger spread in prediction accuracy between HMCC and conventional white-light colonoscopy is listed by Konishi et al. [88] with 92% and 68%, respectively. In addition, inter-observer variability of HMCC-based diagnosis has been described at least for Barret's esophagus [124]. This inter-observer variability may to a lesser degree be also present in the interpretation of Pit Patterns of colonic lesions.

### 4.1.2 Objective

The objective of the computer-aided diagnosis system is two-fold: first, we intend to reliably discriminate Pit Patterns I and II from III to V, which amounts to identify non-neoplastic and neoplastic lesions. According to the medical literature, this is the clinically most relevant application scenario of the Pit Pattern analysis scheme. In the following, we will denote this problem as the *two-class* problem. Second, we focus on a more therapeutically relevant subcategorization in which neoplastic lesions are further discriminated into invasive and non-invasive types. We adhere to the Pit Pattern assignment of Hurlstone et al. [73] where the authors assign Pit Patterns III-S and V to the invasive class and III-L and IV to the non-invasive class. The classes differ in the treatment decision. Non-invasive neoplastic disease allows endoscopic mucosal resection (EMR), whereas invasive neoplasia may require surgical resection. We denote the more fine-grain classification setup as the *three-class* problem.

## 4.2 Prediction by Means of Discriminant Classifiers

As a first and straightforward way to cope with the prediction problem at hand is, to employ a discriminant classifier approach. The basic idea is, to determine some sort of decision boundary from the feature representation of each image and the known class membership in a separate training stage of the system. From a Bayesian point of view, this amounts to estimation of the posterior probability of each class based on a set of training images. In the following parts of this section, we introduce three approaches to determine discriminative image features for use in conjunction with a discriminant classifier. All three approaches are motivated by ideas from texture classification and retrieval since the Pit Pattern images exhibit strong texture characteristics such as regularity or homogeneity. A schematic system overview of a discriminant classifier based system is shown in Fig. 4.3 for the discrimination between non-neoplastic and neoplastic disease. Since we are primarily concerned with the development of image feature vectors and less with the classification side, we use a rather simple 1-Nearest-Neighbor classification strategy [41]. On the one hand, this allows a fair comparison of different feature sets and on the other hand, requires storage of the feature vectors in the classification/prediction step only.

### 4.2.1 Distribution Parameters as Image Features

In [102] and [64], we propose a feature extraction strategy that bears a close relation to the texture retrieval system we introduced in [99]. Motivated by the Gabor wavelet approach of Manjunath & Ma [117] and the shortcomings of the DWT w.r.t. to image analysis (see Section 2.3), we propose to use the DTCWT for feature transformation and to compute the mean and standard deviation of the complex coefficient magnitudes as features. In case of grayscale images, only the luminance channel is decomposed, in case of color images the channels are decomposed separately. The features are then arranged in feature vectors $\mathbf{z} = [\mu_{11}\ \sigma_{11} \dots \mu_{JB}\ \sigma_{JB}]$, where $\mu_{ij}$ denotes the mean of the coefficient magnitudes in subband $i$ at DTCWT level $j$. Given that $B$ denotes the total number of detail subbands, i.e. $B = 6J$ (grayscale) or $B = 18J$ (color), the feature vectors are $\mathbf{z} \in \mathbb{R}^{2B}$. We refer to this features as the *Energy* features. In [98], we present a refine-

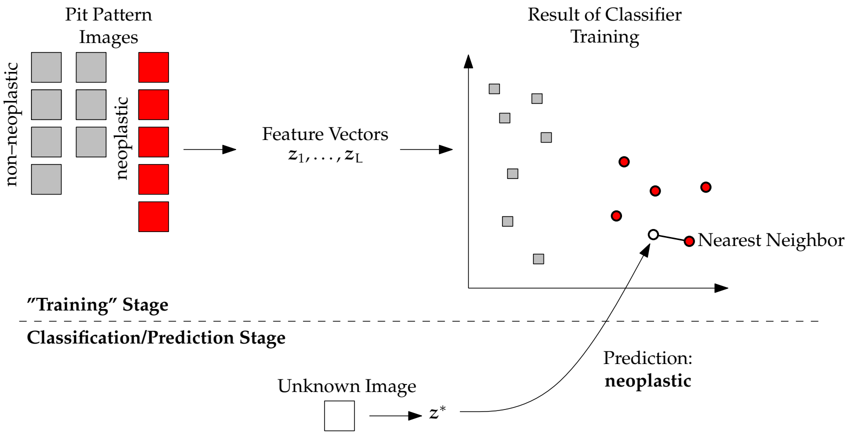


**Figure 4.3:** System overview of prediction based on discriminant classifiers for the non-neoplastic vs. neoplastic case. The illustration on the right-hand side shows the principle of assigning the class label of the nearest-neighbor.

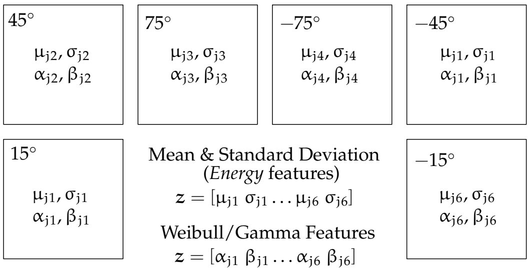
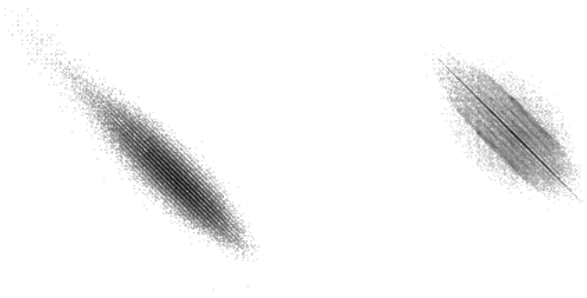

45°
$\mu_{j2}, \sigma_{j2}$
$\alpha_{j2}, \beta_{j2}$
75°
$\mu_{j3}, \sigma_{j3}$
$\alpha_{j3}, \beta_{j3}$
−75°
$\mu_{j4}, \sigma_{j4}$
$\alpha_{j4}, \beta_{j4}$
−45°
$\mu_{j1}, \sigma_{j1}$
$\alpha_{j1}, \beta_{j1}$
15°
$\mu_{j1}, \sigma_{j1}$
$\alpha_{j1}, \beta_{j1}$
Mean & Standard Deviation
(*Energy* features)
$z = [\mu_{j1}\ \sigma_{j1} \ldots \mu_{j6}\ \sigma_{j6}]$
Weibull/Gamma Features
$z = [\alpha_{j1}\ \beta_{j1} \ldots \alpha_{j6}\ \beta_{j6}]$
−15°
$\mu_{j6}, \sigma_{j6}$
$\alpha_{j6}, \beta_{j6}$

**Figure 4.4:** Extraction of a feature vector $z$ from DTCWT subband coefficient magnitudes on scale j, either by determining the sample mean $\mu_j i$ and sample standard $\sigma_{ji}$ deviation (as in [117]) or by determining Weibull distribution parameters $\alpha_{ji}, \beta_{ji}$.

ment of this approach by relying on the statistical models introduced in Chapter 2.3. Instead of computing the rather arbitrary features of sample mean and sample standard deviation (hence, implicitly assuming normality), the transform coefficient magnitudes of each subband are modeled by two-parameter Weibull or Gamma distributions. A feature vector is then composed of the fitted (e.g. ML estimation) distribution parameters, i.e. $z = [\alpha_{11}\ \beta_{11} \ldots \alpha_{JB}\ \beta_{JB}]$. The composition of feature vectors based on the mean/standard deviation and Weibull/Gamma distribution parameters is visualized in Fig. 4.4 for the detail subbands of an arbitrary DTCWT decomposition level j. Consequently, an admissible parameter configuration of this approach is the tuple $\Delta = (\mathit{Colorspace}, \mathit{Feature})$ where *Feature* either denotes the energy features or the distribution parameter features.

### 4.2.2 Cross Co-Occurrence Matrices in the Wavelet Domain

In [100], we extend the concept of classic co-occurrence matrices to capture the information between DWT detail subband pairs of different color channels. Several other studies have proposed to compute color-texture features in some transform domain as well. Karkanis et al. [79]



**(a)** Pit Pattern III-L **(b)** Pit Pattern II

**Figure 4.5:** Two exemplary co-occurrence matrices of different pit-pattern types using a quantization factor of $Q = 100$ and a displacement vector $\mathbf{d} = [-1\ 1]$.

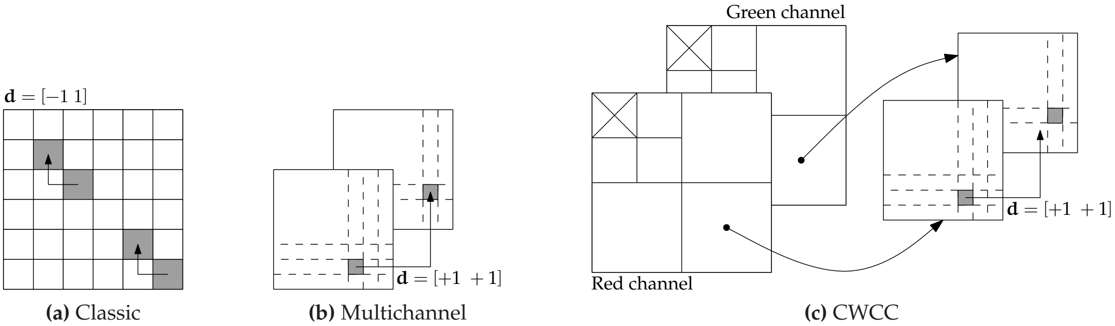
for example, compute co-occurrence matrices from second-level DWT detail subbands at various angles and then determine covariances between Haralick features [65]. Other approaches include *Wavelet Energy Correlation Signatures (WCS)* proposed by Van de Wouwer et al. [35] or Gabor opponent features proposed by Jain & Healey [76]. The latter two are very similar in nature, but reside in different transform domains. In the following, we introduce a novel set of color-texture descriptors, based on second-order statistics from cross co-occurrence matrices, a concept first suggested by Palm et al. [145, 143]. The cross co-occurrence matrices are computed between wavelet detail subbands of different color channels.

First, let us review the concept of co-occurrence matrices computed on intensity images. We assume that an image is given in matrix notation $\mathbf{C}^0 = \{c^0_{ik}\}_{0 \leqslant i,k < N}$ where $c^0_{ik}$ denotes the intensity value of the pixel at location $(i,k)$ and the superscript '0' signifies that we are working in the pixel domain. For simplicity, the location $(i,k)$ will be abbreviated by the lowercase variables $\mathbf{x}, \mathbf{y} \in \{0, \ldots, N-1\}^2$. In case of vector images, we extend this notation by another superscript $p$ or $p'$ to signify the image plane. Hence, in case of RGB images for instance, $p \in \{R, G, B\}$. The classic co-occurrence matrix $\mathbf{M}^p_{\mathbf{d}}(i,j)$ at position $(i,j)$ captures the joint occurrence of intensity values $i$ and $j$ separated by the displacement vector $\mathbf{d} \in \mathbb{N}^2$. The displacement vector thus implicitly defines the orientation and the distance of considered pixel pairs. Formally, $\mathbf{M}^p_{\mathbf{d}}$ is defined as

$$\mathbf{M}^p_{\mathbf{d}}(i,j) = \mathbb{P}\left(c^0_{\mathbf{x}} = i \wedge c^0_{\mathbf{y}} = j | \mathbf{x} - \mathbf{y} = \mathbf{d}\right). \tag{4.1}$$

This formulation of the co-occurrence matrix is specifically tailored for single-channel images, e.g. grayscale images. Depending on the type of texture in an image, we can observe characteristic patterns in the shape of $\mathbf{M}^p_{\mathbf{d}}$. To visualize this characteristic shape, two exemplary co-occurrence matrices are shown in Fig. 4.5. Unless any quantization step is employed, the final co-occurrence matrix has 256 × 256 entries and is sparsely populated in general. Hence, in any practical application the intensity values are mapped to $Q \ll 256$ values by using the mapping $g : \{0, \ldots, 255\} \rightarrow \{0, \ldots, Q-1\}, x \mapsto \lfloor x/255 \times (Q-1) + 0.5 \rfloor$.

A first extension of the classic co-occurrence matrix is proposed by Palm et al. [144] with the objective to capture the joint occurrence of intensity values between image planes $p$ and $p'$.



**(a)** Classic **(b)** Multichannel **(c)** CWCC

**Figure 4.6:** Illustration of three different types of co-occurrence matrices. On the left-hand side, we see the classic co-occurrence matrix for $\mathbf{d} = [-1\ 1]$. For the grayscale value $v$, $\mathbf{M}_{\mathbf{d}}^{p}(v, v)$ will be incremented by $+2$. In the middle, we see the extension to cross co-occurrence matrices between different image channels and on the right-hand side we see the principle of CWCC matrices between the diagonal DWT detail subbands (level one) of the red and green color channel.

Formally, this can be written as

$$\mathbf{M}_{\mathbf{d}}^{p,p'}(i,j) = \mathbb{P}\left(c_{\mathbf{x}}^{0,p} = i \wedge c_{\mathbf{y}}^{0,p'} = j | \mathbf{x} - \mathbf{y} = \mathbf{d}\right). \tag{4.2}$$

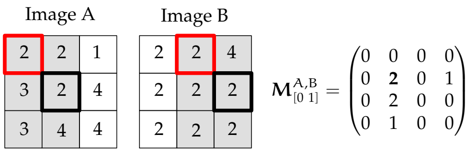
According to the terminology in [144], co-occurrence matrices as defined by Eq. (4.1) are denoted as *within* co-occurrence matrices and co-occurrence matrices as defined by Eq. (4.2) are denoted as *cross* co-occurrence matrices. Generally, the latter concept can be applied to all kinds of vector images. In [145] for example, images are analyzed on different scales, with the scale space generated by repeatedly applying Gaussian filters with varying variance.

Next, we leave the spatial domain and extend the concept of cross co-occurrence matrices to the wavelet domain. We refer to this extension as the *Color Wavelet Cross Co-occurrence (CWCC)* matrices. Let $\mathbf{D}_{k}^{s,p}$ denote the $k$-th DWT detail subband at scale $s$ and color channel $p$. The CWCC matrix $\mathbf{M}_{\mathbf{d},s,k,k'}^{p,p'}(i,j)$ at position $(i,j)$ between two arbitrary subbands $\mathbf{D}_{k}^{s,p}$ and $\mathbf{D}_{k'}^{s,p'}$ can be defined as

$$\mathbf{M}_{\mathbf{d},s,k,k'}^{p,p'}(i,j) = \mathbb{P}\left(c_{\mathbf{x}}^{s,k,p} = i \wedge c_{\mathbf{y}}^{s,k',p'} = j | \mathbf{x} - \mathbf{y} = \mathbf{d}\right). \tag{4.3}$$

The additional superscripts for the transform coefficients are necessary to completely specify their position in the decomposition structure. For our experiments, we follow the restriction $k = k'$, which means that only pairs of subbands at equal positions in the decomposition are considered. As with intensity images, Eq. (4.3) requires a quantization step before computation. We use three quantization factors $Q \in \{64, 128, 256\}$ for the experiments. We further point out, that by using Eqs. (4.2) and (4.3), it is now possible to have a zero-displacement vector $\mathbf{d} = \mathbf{0} = [0\ 0]^T$ as well. This bears a close relation to two-dimensional histograms [143]. The classic co-occurrence matrix approach together with the extensions of cross co-occurrence and CWCC matrices is visualized in Fig. 4.6.

The next imperative step we have to conduct is a dimensionality reduction step. We cannot directly use the entries of the co-occurrence matrices as inputs to a discriminant classifier for the following reason: even by using a quantization factor of $Q = 64$, we would end up with



$$\mathbf{M}^{A,B}_{[0\ 1]} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathbf{2} & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

**Figure 4.7:** A real example of a cross co-occurrence matrix between images A and B. The fields contributing to $\mathbf{M}^{A,B}_{[0\ 1]}(2,2) = 2$ are marked bold red and bold black. Those pixels which are taken into consideration for the computation of the co-occurrence matrix are marked light-gray.

$64^2$-dimensional feature vectors. According to [41], the number of samples needed to train a classifier grows exponentially with the number of input dimensions (known as the *curse of dimensionality*). Since the number of image samples we have is rather small, using the CWCC matrices directly is computationally infeasible. To remedy this problem, we compute a subset of the popular Haralick [65] second-order statistics from the CWCC matrices which are then assembled into feature vectors. We define the Haralick features *Contrast*, *Correlation*, *Homogeneity* and *Energy* as

**Contrast**

$$F_1 = \sum_{i=0}^{Q-1} \sum_{j=0}^{Q-1} |i-j|^2 \mathbf{M}^{p,p'}_{\mathbf{d},s,k,k'}(i,j) \tag{4.4}$$

**Correlation**

$$F_2 = \frac{\sum_{i=0}^{Q-1} \sum_{j=0}^{Q-1} (i-\mu_i)(j-\mu_j) \mathbf{M}^{p,p'}_{\mathbf{d},s,k,k'}(i,j)}{\sigma_i \sigma_j} \tag{4.5}$$

**Homogeneity**

$$F_3 = \sum_{i=0}^{Q-1} \sum_{j=0}^{Q-1} \frac{\mathbf{M}^{p,p'}_{\mathbf{d},s,k,k'}(i,j)}{1+|i-j|} \tag{4.6}$$

**Energy**

$$F_4 = \sum_{i=0}^{Q-1} \sum_{j=0}^{Q-1} \left( \mathbf{M}^{p,p'}_{\mathbf{d},s,k,k'}(i,j) \right)^2 \tag{4.7}$$

where $\mu_i, \sigma_i$ denote the horizontal mean and variance and $\mu_j, \sigma_j$ denote the vertical mean and variance, respectively. In order to signify that the features depend on the particular type of co-occurrence matrix, we adhere to the notation $F_i(\mathbf{M}^p_\mathbf{d})$ to denote that feature $F_i$ is computed based on $\mathbf{M}^p_\mathbf{d}$. In our experiments, we evaluate the discriminative power of the different features separately. Regarding the dimensionality of the final feature vectors, we note that a J-scale DWT produces a 3J-dimensional feature vector $\mathbf{z}^{(p,p')}$ for a given combination $(p,p')$. The final feature vector $\mathbf{z}$ for an image is constructed as a concatenation of all (i.e. six in case of three color channels) possible combinations. To provide a concrete example, consider the case of RGB images: we have $\mathbf{z}^{(R,G)}, \mathbf{z}^{(R,B)}$ and $\mathbf{z}^{(B,G)}$ which leads to the final 9J-dimensional feature vector $z = [z^{(R,G)}\ z^{(R,B)}\ z^{(B,G)}]$. An admissible parameter configuration $\Delta$ of the CWCC approach is the five-tuple $\Delta = (\mathit{Transform}, \mathit{Colorspace}, \mathbf{d}, Q, F_i)$.

### 4.2.3 Color "Eigen-Subbands"

The last feature extraction approach we discuss in the context of discriminant classifiers, is the *Color Eigen-subband (CES)* approach we proposed in [99]. In order to overcome the shift-dependency problem of the DWT – caused by downsampling the filter outputs by two – we replace the DWT by a non-subsampled variant known as the Stationary Wavelet Transform (SWT) [147]. This transform is implemented by the undecimated á-trous algorithm [165] and has a redundancy factor of 3J, where J denotes the maximum decomposition depth. In the terminology of Palm [143], we aim for an integrative color-texture feature extraction approach. By *integrative* we mean a technique which directly incorporates information among color channels. In contrast to that, it is always possible to artificially incorporate color channel information by means of feature vector concatenation, see Section 4.2.1. However, the problem of feature vector concatenation is, that we neglect the association structure between the wavelet detail subbands of different color channels. At least for the RGB colorspace, we have shown that the DWT/DTCWT transform coefficients exhibit a considerable degree of association, see Sections 2.2.3 and 2.3.3. We strongly presume, that the situation is similar in case of the SWT.

To avoid the loss of information, we propose to compute statistics of PCA [41] decorrelated detail subbands as image features. Decorrelation of color channels in the pixel domain is exploited by Heeger & Bergen [67] in the context of texture synthesis. The reason why we perform decorrelation in the wavelet domain is rooted in the fact that decorrelation of the color channels does not guarantee decorrelation of the transform coefficients, as Simoncelli et al. showed in [167]. However, Simoncelli et al. further point out, that decorrelation in the wavelet domain by means of PCA does not lead to decorrelated subband in all cases either. Instead of using PCA, the authors propose to use Independent Component Analysis (ICA) as an alternative. Nevertheless, we retain the PCA approach, since the setup of [167] differs from our setup in the following sense: in [167], the coefficient matrix is composed of transform coefficients from all levels and all color channels by randomly selecting a collection of coefficients. In our setup, however, the coefficient matrix is constructed by selecting just the transform coefficients of the same subband but on different color channels. Our experiments show that decorrelation of the transform coefficients is acceptable in this special case, e.g. see Fig. 4.9. Further, computation of the PCA is less expensive than computation of ICA.

We next explain PCA-based decorrelation by means of an example: we assume RGB images and that each color channel is decomposed separately by a J-scale DWT. Without loss of generality, we consider the $k$-th detail subband on decomposition level $j$, denoted by $\mathbf{D}_k^{j,p}$ (the superscript $p$ denotes the color channel). The transform coefficients are denoted by $c_i^p, 1 \leqslant i \leqslant N$ using linear indexing. We omit the subband and scale specifiers $k$ and $j$ for readability. The construction of the coefficient matrix $\mathbf{X}$ is illustrated in Fig. 4.8. Each row of the matrix $\mathbf{X}$ is an observation vector $\mathbf{c}_i \in \mathbb{R}^3$. To decorrelate the components of the observation vector, PCA works by diagonalizing the sample covariance matrix $\mathbf{S}$, using the projection

$$\tilde{\mathbf{S}} = \mathbf{\Phi}^T \mathbf{S} \mathbf{\Phi} \tag{4.8}$$

where $\mathbf{\Phi}$ denotes the matrix of eigenvectors corresponding to the eigenvalues of $\mathbf{S}$ (sorted in ascending order). Since the sample covariance matrix $\mathbf{S}$ can be written as the product

$$\mathbf{S} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi} \quad \text{with} \quad \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \tag{4.9}$$

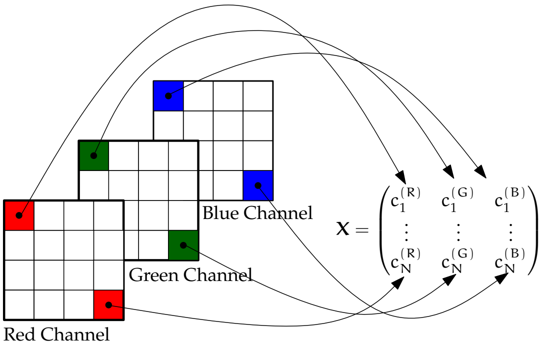

Blue Channel
Green Channel
Red Channel
$\mathbf{X} = \begin{pmatrix} c_1^{(R)} & c_1^{(G)} & c_1^{(B)} \\ \vdots & \vdots & \vdots \\ c_N^{(R)} & c_N^{(G)} & c_N^{(B)} \end{pmatrix}$

**Figure 4.8:** Arranging DWT detail subband transform coefficients of different color channels into a data matrix $\mathbf{X}$.

it is evident that $\tilde{\mathbf{S}} = \mathbf{\Lambda}$. Hence, the sample covariance between two dimensions is zero. The variance along the principal axis is given by the eigenvalues $\lambda_i$. We can now directly use the eigenvalues as features for classification. In fact, all the variance information is now packed into the eigenvalues $\lambda_i$. To obtain the decorrelated sample $\mathbf{y}_i$, we first conduct the transformation $y_i = \mathbf{\Phi}^T \mathbf{c}_i$ and then arrange the vectors $\mathbf{y}_i^T$ as rows of a new data matrix $\mathbf{Y}$. We finally obtain the *Color Eigen-Subbands (CES)* by reshaping the columns of $\mathbf{Y}$ to three $N \times N$ matrices.

Given that we use the variances along the principal axis (i.e. the eigenvalues) as image features, we obtain a 9J-dimensional feature vector for each image. Since the eigenvalues have completely different ranges, we have to be careful when computing the Euclidean distance between feature vectors, though. We remedy that problem by normalizing the elements of the feature vectors by subtracting the sample mean and dividing by the standard deviation. As an extension to the work of [99], we adopt the CES approach to work with the complex detail subbands of the DTCWT using only the magnitude information. Due to the larger number of subbands per scale, the dimensionality of the feature vectors is doubled. An admissible parameter configuration of the CES approach is the tuple $\Delta = (\mathit{Transform}, \mathit{Colorspace})$.

## 4.3 Prediction by Means of Generative Models

In the context of our classification problem, we identify three critical issues related to discriminant classifier approaches: first, classifier training usually requires a sufficiently large number of training samples. Unless this can be guaranteed, we inevitably run into overtraining issues. Second, most classifiers additionally require balanced class distributions. Unfortunately, we cannot guarantee this requirement either. Since some Pit Patterns (e.g. III-S) occur very rarely, the image distribution tends to be highly unbalanced. Neglecting this fact leads to overtraining in favor of classes with a large number of samples (e.g., see [135]). Third, we want to ensure that images with an already assigned histopathological diagnosis can be added to the image database at any time without effort. This avoids presumably time-consuming and unnecessary maintenance operations which might prevent the actual deployment in clinical practice. Since discriminant classifiers usually need re-training in case new samples are added, this requirement cannot be met either.

As a possible solution to these disadvantages, we propose to employ a prediction strategy

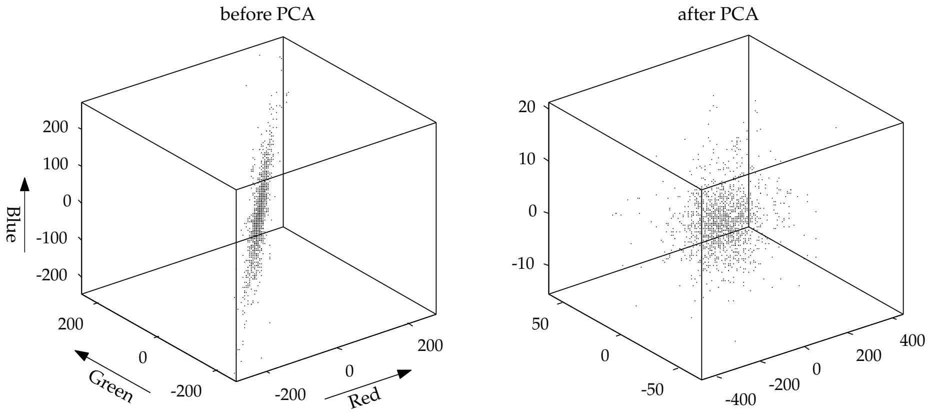


**Figure 4.9:** Scatterplot of DWT subband transform coefficients (horizontal detail subband on DWT level two of a Pit Pattern II image) before and after applying PCA. The pairwise (linear) correlation between each component is approx. zero after PCA.

based on generative models. The baseline for this proposal is the framework of Bayesian image retrieval [186] which we already discussed in Chapter 3. Considering the classification problem from the viewpoint of image retrieval brings along several advantages which correspond to the requirements stated above. An unknown HMCC image is considered as a query image in the probabilistic framework and classification is performed by first searching for the most similar image in the database of available HMCC images with an assigned histological diagnosis. Next, the class of the retrieved image is used as a prediction for the class of the unknown image. In classification terminology, this resembles a nearest neighbor classifier. Fig. 4.10 shows the two possible strategies for class prediction: (i) by searching for the feature representation $p_Z(z; \mathbf{\Theta}_r)$ which minimizes the KL divergence to the feature representation $p_Z(z; \mathbf{\Theta}^*)$ of the unknown image (left branch), or (ii) by searching the feature representation which maximizes the (log) likelihood of the unknown image's coefficient data (right branch). We don't want to go into detail too much at this point, since the theoretical foundations are given in Chapter 3 to which the reader is referred for further information. Eventually, we highlight the two striking advantages of the generative model based approach: first, no classifier training is required at all. Depending on which prediction strategy we use, it might not even be necessary to estimate the model parameters of the query image (i.e. for likelihood maximization). Hence, we consequently avoid overtraining issues and are not tempted to overly optimize feature sets by means of feature subset selection for example. Second, images with an existing histopathological diagnosis can be added to the database at any time and are immediately available for future image queries.

## 4.4 Classification Setup

The classification setup for the discriminant classifier approaches is as follows: we restrict our study to the one Nearest-Neighbor (1-NN) classifier we used in [98, 102, 64, 99, 100], since we focus on a comparison of the various feature extraction approaches and do not conduct a study on the performance of different classifiers. We further omit any feature subset selection or other

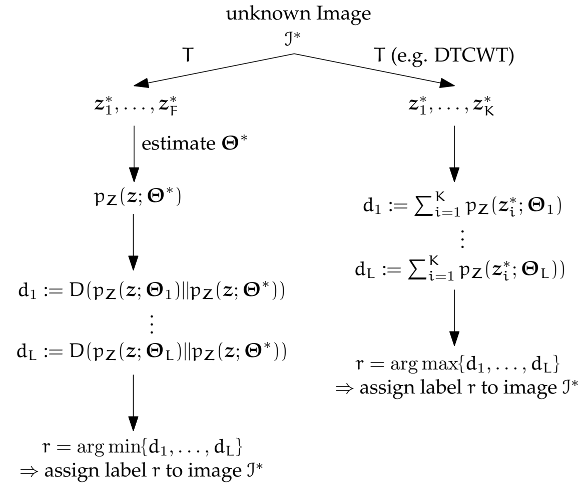

unknown Image
$\mathcal{I}^*$
T
T (e.g. DTCWT)
$\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_F^*$
estimate $\boldsymbol{\Theta}^*$
$p_Z(\boldsymbol{z}; \boldsymbol{\Theta}^*)$
$d_1 := D(p_Z(\boldsymbol{z}; \boldsymbol{\Theta}_1) \| p_Z(\boldsymbol{z}; \boldsymbol{\Theta}^*))$
$\vdots$
$d_L := D(p_Z(\boldsymbol{z}; \boldsymbol{\Theta}_L) \| p_Z(\boldsymbol{z}; \boldsymbol{\Theta}^*))$
$r = \arg\min\{d_1, \ldots, d_L\}$
$\Rightarrow$ assign label r to image $\mathcal{I}^*$
$\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_K^*$
$d_1 := \sum_{i=1}^K p_Z(\boldsymbol{z}_i^*; \boldsymbol{\Theta}_1)$
$\vdots$
$d_L := \sum_{i=1}^K p_Z(\boldsymbol{z}_i^*; \boldsymbol{\Theta}_L))$
$r = \arg\max\{d_1, \ldots, d_L\}$
$\Rightarrow$ assign label r to image $\mathcal{I}^*$

**Figure 4.10:** Prediction of the class label of an unknown image $\mathcal{I}^*$ by means of (i) finding the feature representation with the smallest KL divergence to the query image's feature representation (left branch) or (ii) by searching for the feature representation which maximizes the log-likelihood of the query image data (right branch).

tuning steps to avoid overtraining issues. In case of the generative model based approaches, the classification strategy is straightforward, since the class label of the retrieved image determines the class label of the unknown image anyway.

To obtain an estimate of classification accuracy, we use the method of Leave-One-Out Cross-validation (LOOCV) [41]. Given a dataset of L samples, LOOCV works by successively leaving out one sample of the whole dataset and performing the training procedure on the remaining $L - 1$ samples. The classification accuracy is then estimated as the number of times the left-out sample is correctly classified. Note, that in case of nearest neighbor classification paradigm, the training procedure just involves storage of the feature vectors. For the discriminant classifier approaches, we rely on the Euclidean distance $d(\boldsymbol{v}_, \boldsymbol{v}_j) = \|\boldsymbol{v}_i - \boldsymbol{v}_q\|$ between two feature vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. As mentioned before, it is reasonable to conduct a normalization step before computing the Euclidean distance. This is accomplished by subtracting the mean and dividing by the standard deviation. Formally, the normalized j-th element of feature vector $\boldsymbol{v}_i$ is computed as $v_{ij} = (v_{ij} - \overline{v}_j)/\sigma_j^2$. Of course, the standard deviation $\sigma_j$ and the mean $\overline{v}_j$ have to be repeatedly computed in each LOOCV iteration to ensure that no information of the left-out sample is included.

## 4.5 Experiments

We perform a comparison of the feature extraction approaches introduced in the context of discriminant classifiers and the approaches introduced in the context of generative models. In the former case, we include the approaches of Gabor wavelets of Manjunath & Ma [117] (see Section 3.3.2), the WCS features of Van de Wouwer et al. [35] and the color histograms proposed by Swain & Ballard [175] as a reference. Gabor wavelet features are commonly used in texture classification and retrieval literature, WCS features have been successfully employed in the context of endoscopic video frame processing [79] and the method of color histograms recently appeared in context of computer-assisted Pit Pattern classification [63]. The latter two approaches are described below. In case of the generative model based approaches, we compare our retrieval approaches of Sections 3.3 and Section 3.4 to the approaches of Vasconcelos & Lippman [186] and Verdoolaege et al. [188]. We refer the reader to Chapter 3 for a detailed description of the retrieval approaches. Again, we adhere to the convention to identify an approach by the names of the authors and the year of publication.

**Van de Wouwer et al., 1997** In [35], Van de Wouwer et al. introduce the approach of *Wavelet Energy Correlation Signatures (WCS)*. The authors propose to decompose the color channels of an image by a J-scale DWT and then calculate the correlation between all combinations of subband pairs on different channels. In particular, given a RGB image and a three-scale DWT, we obtain 27-dimensional feature vectors. Since this approach can be easily extended to work with the SWT and DTCWT, we will also consider these cases in our experiments. Note that in case of the DTCWT, the size of the feature vectors is doubled.

**Swain & Ballard, 1991** The method of color histograms was introduced by Swain & Ballard [175] in an effort to evaluate whether color information can effectively capture image characteristics. The authors compute three-dimensional histograms from the intensity values of each color channel. Since a full color histogram would consist of $256^3$ bins (very sparsely populated), intensity values are uniformly quantized to obtain a $N_1 \times N_2 \times N_3$ bin color histogram with $N_i \ll 256$. This eventually allows computationally efficient similarity measurement using the histogram intersection as a similarity measure. For our experiments, we use the RGB color space and a quantization setting of $N_1 = N_2 = N_3 = 8$.

### 4.5.1 Image Acquisition

Our original set of images consists of 269 RGB images (53 patients, either 624 × 533 or 586 × 502 pixel) acquired in 2005–2009 at the Department of Gastroenterology and Hepatology of the Medical University of Vienna using a zoom-endoscope (Olympus Evis Exera CF-Q160ZI/L) with a magnification factor of 150×. All images were selected by the gastroenterologist conducting the colonoscopy with special emphasis to provide images with similar lightning conditions at approximately the same camera angle. To enhance the visual appearance of the mucosa, dye-spraying with indigo-carmine was applied and biopsies or mucosal resections were taken to obtain a histopathological diagnosis (*our ground truth*). The histology was obtained by a pathologist blinded to the colonoscopic procedure. Table 4.2 lists the histologies for the observed Pit Patterns as well as the corresponding occurrences.

In order to increase the number of samples, we create an extended dataset by extracting 256 × 256 pixel subwindows from the original images such that the Pit Patterns are clearly distinctive and the subwindows contain a minimum number of specular reflections (see Fig. 4.11). This resembles the clinical methodology during colonoscopy, since the gastroenterologist will typically look at more than one region of an image. Finally, the extended dataset contains 627 HMCC images distributed according to column #*(extended)* in Table 4.2.

In this thesis, we differ to the originally published works in one particular point. Up to now, the medical presentation of the problem was considered from a purely classification oriented point of view. In such a setup, it does not matter which image is selected as the one to predict the class of an unknown image. The results, however, only convey an impression of how well an approach captures image information relevant for discrimination. The classification rates are less meaningful from a medical point of view, though. This becomes obvious when we consider

| Pit Pattern | # | #(extended) | Histology | # |
|---|---|---|---|---|
| **I** | 36 | 114 | Normal | 36 |
| **II** | 26 | 64 | Hyperplasia | 26 |
| **III-S** | 12 | 18 | serrated Adenoma | 4 |
| | | | tubular Adenoma | 8 |
| **III-L** | 44 | 119 | tubular Adenoma | 43 |
| | | | tibulovillous Adenoma | 1 |
| **IV** | 120 | 232 | tibulovillous Adenoma | 115 |
| | | | Adenoma | 2 |
| | | | tubular Adenoma | 3 |
| **V** | 31 | 80 | Lymphoma | 6 |
| | | | Carcinoma | 6 |
| | | | Adenocarcinoma | 19 |
| $\sum$ | 269 | 627 | | 269 |

**Table 4.2:** Pit Patterns with corresponding histopathological diagnosis. The second column, i.e. #, lists the number of original images, while the third column, i.e. # (extended), lists the number of images in the extended dataset. The last two columns list the histologies and the corresponding occurrences.

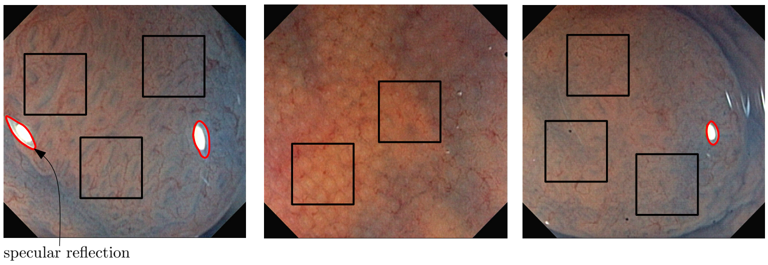


**Figure 4.11:** Extraction of 256 × 256 pixel subwindows (black squares) from the original HMCC images with the objective to increase the dataset. Specular reflections in the first and third image are marked red.

our dataset extension technique and the fact that there is no restriction on the type of nearest neighbor simultaneously. In fact, during the LOOCV process, it is possible that the nearest neighbor stems from the same parent as the unknown image. In case we are only interested to find images with similar visual content, this case does not pose a serious problem. Actually, the evaluation of texture retrieval systems works in the same way (see Section 3.3.2). Nevertheless, we can construct a clinically more relevant evaluation strategy by imposing a constraint on the type of nearest neighbor. We say, that images are only admissible as nearest neighbors in case they do not stem from the same parent as the unknown image. We refer to this setup as the *constrained* NN setup, whereas the setup in the original works will be referred to as the *unconstrained* NN setup. To visualize the difference, both types are illustrated in Fig. 4.12.

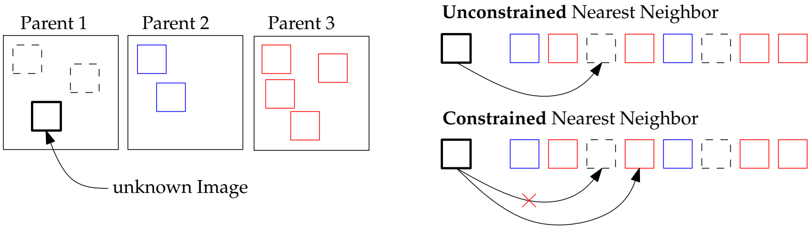

Parent 1
Parent 2
Parent 3
unknown Image
**Unconstrained** Nearest Neighbor
**Constrained** Nearest Neighbor

**Figure 4.12:** Illustration of the constrained and unconstrained nearest neighbor principle. In the unconstrained case, the nearest neighbor is allowed to stem from the same parent as the unknown image, whereas in the constrained case this is prohibited.

### 4.5.2 Parameter Configurations

In order to make our results reproducible, we have to define the parameter configurations for the approaches. We first report the common parameters and then discuss the specific parameter settings. Regarding the choice of colorspace, we perform experiments using the RGB, HSV, YBR and YIQ colorspace. The conversions based on the RGB model are accomplished using the colorspace conversion routines of MATLAB. For all wavelet transform variants (including Gabor wavelets) the decomposition depth is fixed to $J = 3$ levels and no image preprocessing steps are conducted. The Gabor wavelet settings (i.e. filter configurations) are listed in Section 3.3.2.

Regarding the distribution feature approach of Section 4.2.1, color image processing is implemented by means of feature vector concatenation. In [64], our experiments showed that feature vector concatenation leads to competitive classification results compared to other, more advanced, combination strategies. Further, we use moment and ML estimates for the Gamma and Weibull parameters. This is a reasonable choice, since it allows to assess whether the estimation methods have an impact on the classification results. This is similar to the retrieval scenario of Chapter 3, where we could show that the choice of parameter estimation method did not have any effect at all.

In case of the CWCC features of Section 4.2.2, we have several parameters which can be adjusted. Regarding the quantization levels Q, we decided to conduct experiments using $Q = 32$ and $Q = 64$ in order to keep the computational effort at a reasonable level. We strongly believe that this is a reasonable setting, since Karakanis et al. [79] reported no gain in classification rates using a higher number of quantization levels. The displacement vectors for the computation of the cross co-occurrence matrices are set to $\mathbf{d} = [0\ 0]$ (zero-displacement), $\mathbf{d} = [-1\ 1]$, $\mathbf{d} = [-1\ 0]$ and $\mathbf{d} = [-1\ -1]$. We do not consider displacements that are farther away then one coefficient. In fact, we presume that a zero-displacement vector which corresponds to a multidimensional coefficient histogram will outperform any other displacement setting.

The WCS approach of Van de Wouwer et al. [35] and the Color-Eigen Subband (CES) approach of Section 4.2.3 do not have any remaining free parameters.

### 4.5.3 Assessing Statistically Significant Differences

In any reasonable comparative study on classification performance, we face situations where classification rates between two approaches seem very similar and do not allow to make any statements whether one approach performs better than the other. To evaluate whether the class assignments of two approaches show statistically significant differences, we employ a McNemar test [46]. Besides the 5x2 cross-validation test, this is one of the most popular and recommended [37] tests for our purpose. The test statistic is based on counting the number of samples where approach A assigns the right class label and approach B fails (denoted by $n_{10}$) and vice versa (denoted by $n_{01}$). Based on these counts, the test statistic is defined as

$$T = \frac{(|n_{10} - n_{01}| - 1)^2}{n_{10} + n_{01}}. \tag{4.10}$$

In case the null-hypothesis of no statistically significant difference is true, T follows as Chi-Square distribution with one degree of freedom, i.e. $T \sim \chi_1^2$. Hence, given a fixed significance level $\alpha$, we decide against the null-hypothesis if $T > F^{-1}_{\chi_1^2}(1-\alpha)$, i.e. when T is larger than the $(1-\alpha)$ quantile of the Chi-Square distribution with one degree of freedom. For example, given $\alpha = 0.05$, we decide against the null-hypothesis if $T > 3.84$. Consequently, we can also say that there is enough evidence against the hypothesis of no significant differences.

Another important point in the context of evaluating differences between classifiers in our context, is the issue of multiple comparisons. We perform multiple pairwise comparisons in two variants. In order to assess whether a certain parameter configuration of an approach leads to better LOOCV accuracy than other configurations, we select the best (i.e. the one with the highest LOOCV accuracy) configuration and compare against the results obtained with all other configurations. Hence, we consider a LOOCV run with one particular parameter configuration as a separate experiment. This is in accordance with the guidelines of Salzberg [157], where the author suggests that different parameter settings should be considered as a special case of *repetitive tuning*. The second variant of the multiple comparisons scenario occurs when we compare several distinct approaches to each other. In order to establish a reasonable ranking, we need to know if the results of two approaches are significantly different. As Salzberg [157] points out, such experimental settings require a correction of the significance level $\alpha$ of each test. To highlight the problem, lets consider the case where we perform $n$ pairwise comparisons. The chance of identifying a statistically significant result is $1-(1-\alpha)^n$. It is straightforward to check that it only requires $n = 45$ comparisons in order to reach a probability $> 90\%$ of making a false discovery (given independent experiments). The classic strategy to control the so called *Familywise Error Rate (FWER)* (i.e. the probability of making one or more Type I errors) is to use the Bonferroni correction, i.e. $\alpha$ is corrected to $\hat{\alpha} = \alpha/n$, or a variant of the Bonferroni method known as the Šidàk correction (see [157]). For our experiments, we implement the latter method which corrects the significance level $\alpha$ to $\hat{\alpha} = 1-(1-\alpha)^{1/n}$. For the example of $\alpha = 0.05$ and $n = 45$ we obtain $\hat{\alpha} = 0.0011$. Although this correction is based on the assumption of independent tests – which might be violated in a practical scenario – it is still a reasonable strategy to reduce the chance of making false conclusions. A second, alternative strategy to cope with the problem of multiple comparisons is to control the *False Discovery Rate (FDR)* instead of the FWER. The general difference to the aforementioned approach of using Bonferroni or Šidàk correction is, that the FDR focuses on the concept of discoveries, i.e. a statistically significant experiment. The FDR is designed to control the rate of false discoveries which is a more natural view of the problem in many situations. In this work, we implement the FDR control algorithm proposed by Benjamini & Hochberg [7]. Formally, given a set of $n$ hypotheses with associated p-values $p_1, \ldots, p_n$, we first sort the p-values to get $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(n)}$ and then determine

$$\hat{k} = \max\{k : p_{(n)} \leqslant \alpha \cdot k/n\}. \tag{4.11}$$

Next, given that $\hat{k}$ exists, we reject the hypotheses corresponding to $p_{(1)}, \ldots, p_{(\hat{k})}$. Otherwise, no hypothesis is rejected at all. In the following, we refer to this procedure as the *Benjamini-Hochberg* procedure to control the FDR. Although, originally intended in situations where the hypotheses are independent, Benjamini & Yekutieli [8] show that in case the subset of test statistics corresponding to true null-hypothesis are positively dependent, then the Benjamini-Hochberg procedure still controls the FDR at a level less than or equal to the desired level $\alpha$. For our experiments, we make the assumption that the required condition can be met in case of our test statistics. When providing significance results, we list the McNemar test outcome for (i) controlling the FWER by means of the Šidàk correction and (ii) controlling the FDR by means of the Benjamini-Hochberg procedure.

### 4.5.4 Results

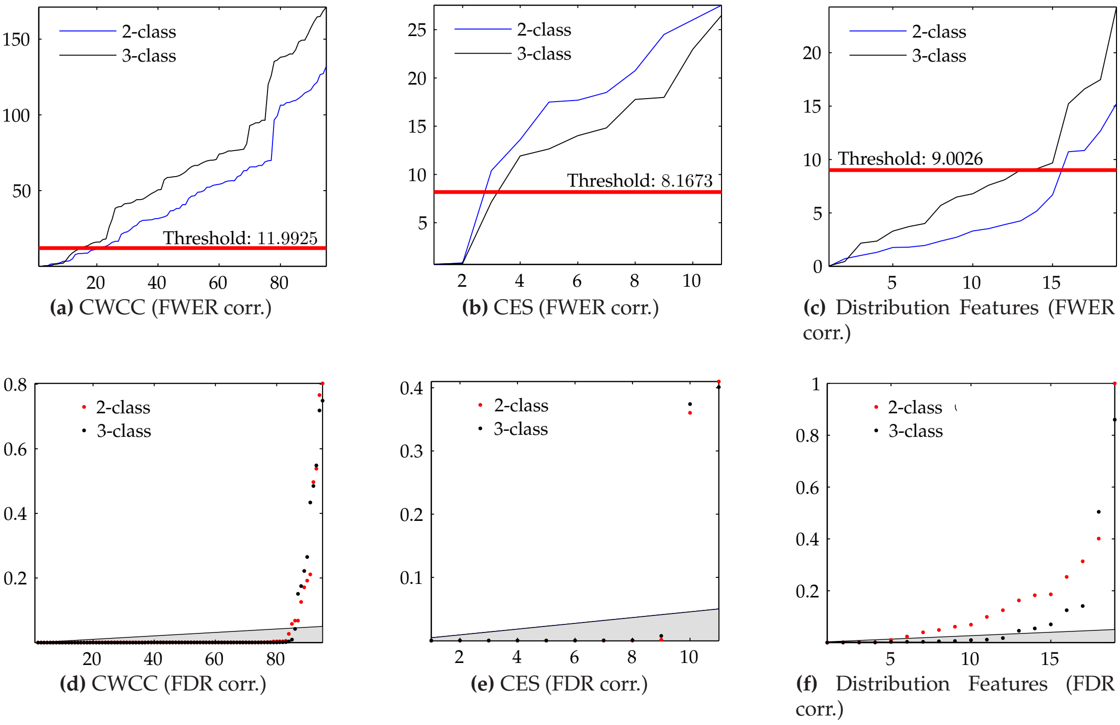
As a starting point for our results section, we take a closer look at the three particular feature extraction approaches we discussed in the context of discriminant classifiers, see Section 4.2. We indent to identify the parameter configurations which lead to the top LOOCV rates and then discuss whether we can claim superiority w.r.t. to the other configurations by means of searching for statistically significant differences.

First, we consider the CWCC features since this approach has the largest number of free parameters. The LOOCV accuracies as well as the detailed classifier performance measures of sensitivity/specificity as well as positive/negative predictive value (abbreviated by PPV and NPV, resp.) are listed in Table 4.3. We report, that the parameter configuration of $\Delta = (\mathit{DTCWT}, \mathit{RGB}, \mathbf{d} = [0\ 0], Q = 32, \mathit{Correlation})$ leads to the highest LOOCV accuracy of 89.63% in the two-class case. In case of the more fine grain discrimination of the three-class problem, the parameter configuration $(\mathit{SWT}, \mathit{YIQ}, \mathbf{d} = [0\ 0], Q = 32, \mathit{Correlation})$ leads to the top rate of 84.05%.

| Problem | Accuracy | Sensitivity | Specificity | PPV | NPV | not sig.? (FWER/FDR) |
|---|---|---|---|---|---|---|
| 2-class | 89.63 | 84.83 | 91.54 | 79.89 | 93.84 | 22.11/11.6 |
| 3-class | 84.05 | - | - | - | - | 15.8/9.5 |

**Table 4.3:** Top LOOCV rates for the CWCC approach.

As we can see, specificity and the NPV are remarkably higher than the sensitivity and the PPV. This signifies that neoplastic disease can be diagnosed more reliably. Next, we fix the top parameter configurations and perform pairwise comparisons of the top results to the results obtained by the remaining parameter configurations. In Fig. 4.13a, we plot the sorted values of the McNemar test statistic T against the number of pairwise comparisons. The bold red-line signifies the threshold (using FWER correction) above which we can claim statistically significant differences. Accordingly, Fig. 4.13d shows a plot of the sorted p-values against the number of pairwise comparisons when relying on the Bejamini-Hochberg correction. The shaded area signifies the region above which there is evidence against the null-hypothesis of the McNemar test. The percentage of non-significant differences among all pairwise comparisons is listed in the last column of Table 4.3. When tracing back the parameter configurations where there is no evidence against the null-hypothesis, we observe the following situation: in almost any case (no matter if we consider the two- or three-class problem) the Haralick feature *Correlation*



**(a)** CWCC (FWER corr.)

**(b)** CES (FWER corr.)

**(c)** Distribution Features (FWER corr.)

**(d)** CWCC (FDR corr.)

**(e)** CES (FDR corr.)

**(f)** Distribution Features (FDR corr.)

**Figure 4.13:** Illustration of the McNemar test outcomes for pairwise comparisons between the top parameter configuration of each approach and the remaining parameter configurations. In the top row, we plot the sorted McNemar test statistics T against the number of pairwise comparisons. The threshold (determined using FWER control) above which we have evidence against the null-hypothesis is marked by a bold red line. In the bottom row, we plot the sorted p-values against the number of comparisons. The region (determined using FDR control) of non-significant differences is marked gray.

and zero-displacement $\mathbf{d} = \mathbf{0}$ are fix elements in the configuration. Only the colorspace and wavelet transform actually change. Practically, this allows to draw the conclusion that the *Correlation* feature together with a zero-displacement vector are the key elements to achieve good classification (and hence good prediction) results when using the CWCC approach.

We next turn to the results of the CES features. The free parameters of this approach are the type of wavelet transform and the colorspace, hence there are twelve possible combinations. Table 4.4 lists the top LOOCV results for both classification problems. In either case, the highest LOOCV accuracy, i.e. 93.14% and 88.84%, is obtained using the parameter configuration $\Delta = (DTCWT, YBR)$. When relying on control of the FWER, this configuration leads to significantly better classification results than any other configuration in $\approx$ 19% and $\approx$ 27% of all cases. The Benjamini-Hochberg procedure for FDR control is less strict, with $\approx$ 18% of non-significant results for both problems. Again, identifying the pairwise comparisons where there is no evidence against the null-hypothesis, reveals that switching the colorspace from YBR to YIQ or RGB does not lead to a significant change in the classification results compared to the top parameter configuration. Hence, the key parameter element of the CES approach is the choice of wavelet transform, i.e. the DTCWT.

| Problem | Accuracy | Sensitivity | Specificity | PPV | NPV | not sig.? (FWER/FDR) |
|---|---|---|---|---|---|---|
| 2-class | 93.14 | 86.10 | 96.14 | 90.45 | 94.21 | 18.8/18.2 |
| 3-class | 88.84 | - | - | - | - | 27.27/18.2 |

**Table 4.4:** Top LOOCV rates for the CES approach.

Finally, we take a look at the distribution features of Section 4.2.1. We include the mean & standard deviation of the subband coefficient magnitudes as features [102, 64] (denoted as *Energy* features). Since we can either use moment or ML estimation for the Gamma and Weibull distribution parameters, we have 20 possible parameter configurations. For both classification problems, the configuration of $\Delta = (YBR, Energy)$ leads to the highest LOOCV accuracies of 93.30% and 89.47%, resp., see Table 4.5. Similar to Tables 4.3 and 4.4, the high rates for specificity and NPV indicate better prediction performance of neoplastic disease. In the two-class case, there is almost no significant difference in the results obtained by different parameter configurations. When we rely on FWER correction, this is also true for the three-class problem. However, in case of the FDR control approach, we report only $\approx$ 16% of non-significant differences for the three-class problem. The detailed results reveal that the *Energy* features generally lead to higher discrimination rates, no matter which colorspace we choose. We attribute this effect to the bad choice of similarity measure (i.e. the Euclidean distance) for the distribution features. As we will later see, the refinement of the similarity measure in favor of the KL divergence considerably improves the results.

| Problem | Accuracy | Sensitivity | Specificity | PPV | NPV | not sig.? (FWER/FDR) |
|---|---|---|---|---|---|---|
| 2-class | 93.30 | 91.01 | 94.21 | 86.17 | 96.36 | 78.95/73.68 |
| 3-class | 89.47 | — | — | — | — | 57.89/15.8 |

**Table 4.5:** Top LOOCV rates for the distribution features approach.

After the fine-grain analysis of the feature extraction approaches, we go on to a comparative study of the CWCC, CES and distribution features (using the top parameter configurations) to the Gabor wavelet features [117], color histograms [175] and the WCS features [35]. For the three reference approaches, we do not perform a detailed study whether there are statistically significant differences in the results obtained by different parameter configurations. We simply pick out the best parameter configuration in each case. Tables 4.6 and 4.7 summarize the achieved LOOCV accuracies for the two- and three-class problem. In Table 4.7, we additionally list the classifier performance measures of sensitivity, specificity, PPV and NPV for the discrimination of non-invasive vs. invasive disease. Accordingly, high values for specificity and NPV indicate good prediction performance for invasive neoplastic disease.

From Tables 4.6 and 4.7, we first notice that there is a considerable difference in LOOCV accuracies between the top rates of 93.30%/89.47% and the worst rates of 84.34%/78.31% which is equivalent to $\approx$ 60 more misclassified images. Further, specificity is higher than sensitivity in the two-class case which suggests that the diagnostic accuracy of neoplastic disease is generally higher than for non-neoplastic disease (at least for our dataset). In case of the discrimination between non-invasive and invasive neoplasia, the situation is reversed, see Table 4.7, with higher

| Approach | Accuracy | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| Distribution Features | **93.30** | 91.01 | 94.21 | 86.17 | 96.36 |
| CES | 93.14 | 90.45 | 94.21 | 86.10 | 96.14 |
| Van de Wouwer et al., 1997 | 90.75 | 89.89 | 91.09 | 80.00 | 95.78 |
| Manjunath & Ma, 1996 | 90.27 | 84.27 | 92.65 | 81.97 | 93.69 |
| CWCC | 89.63 | 84.83 | 91.54 | 79.89 | 93.84 |
| Swain & Ballard, 1991 | 84.37 | 74.16 | 88.42 | 71.74 | 89.62 |

**Table 4.6:** Comparison of the LOOCV rates for the 2-class problem.

| Approach | Total | Non-Invasive vs. Invasive | | | | |
|---|---|---|---|---|---|---|
| | | Accuracy | Sensitivity | Specificity | PPV | NPV |
| Distribution Features | **89.47** | 94.33 | 96.66 | 86.17 | 96.07 | 88.04 |
| CES | 88.84 | 93.62 | 96.08 | 84.62 | 95.80 | 85.56 |
| Van de Wouwer et al., 1997 | 85.96 | 92.67 | 95.27 | 83.70 | 95.27 | 83.70 |
| CWCC | 84.05 | 92.42 | 96.57 | 77.27 | 93.94 | 86.08 |
| Manjunath & Ma, 1996 | 81.18 | 86.30 | 92.64 | 63.33 | 90.15 | 70.37 |
| Swain & Ballard, 1991 | 78.31 | 90.43 | 96.09 | 71.11 | 91.90 | 84.21 |

**Table 4.7:** Comparison of the LOOCV rates for the 3-class problem and the subproblem of discriminating non-invasive vs. invasive neoplasia.

values for sensitivity than for specificity. Consequently, the diagnostic accuracy of non-invasive neoplastic disease is better.

An interesting question which remains to be answered is, whether the ranking presented in Tables 4.6 and 4.7 is actually reliable. In particular, we require that there has to be a statistically significant difference in the classification results between two approaches in order to assign different ranks. Table 4.8 lists the McNemar test statistic values for all pairwise comparisons and highlights those cases where the McNemar test shows evidence against the null-hypothesis. These significant differences are either marked gray (when relying on FWER control) and/or by a '*' (when relying on FDR control). The results indicate that taking the ranking as it is can be elusive, since the pairwise comparisons of the top four approaches do not show any evidence against the null-hypothesis at all. We attribute this effect to the significance level correction. Without this correction, the threshold of the McNemar test would be lowered to 3.84 for example.

As a final part of this section, we present the classification results of the generative model based approaches from CBIR. The LOOCV rates for both classification problems are listed in Tables 4.9 and 4.10. We report that the top parameter configuration remains the same for all approaches on both problems. The copula retrieval strategy exhibits the highest rates, using the parameter configuration of a Gaussian copula, Gamma margins and the RGB colorspace. The approach of Kwitt & Uhl performs at a competitive level using the YIQ colorspace and the Gamma distribution. This particularly emphasizes the point of using a suitable similarity measure. In comparison to the discriminant classifier strategy of using the distribution param-

| | | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|
| A | Distribution Features | - | | | | | |
| B | CES | 0 | - | | | | |
| C | Van de Wouwer et al, 1997. | 4.50 | 2.92 | - | | | |
| D | Manjunath & Ma, 1996 | 4.10 | 3.70 | 0.05 | - | | |
| E | CWCC | 6.68* | 5.63* | 0.31 | 0.04 | - | |
| F | Swain & Ballard, 1991 | 26.07* | 24.92* | 11.35* | 10.36* | 8.64* | - |
| A | Distribution Features | - | | | | | |
| B | CES | 0.23 | - | | | | |
| C | Van de Wouwer et al, 1997. | 6.30* | 3.28 | - | | | |
| D | CWCC | 8.38* | 8.24* | 1.06 | - | | |
| E | Manjunath & Ma, 1996 | 19.41* | 16.99* | 6.37 | 2.18 | - | |
| F | Swain & Ballard, 1991 | 30.92* | 28.54* | 12.99* | 7.65* | 1.70 | - |

**Table 4.8:** McNemar test statistic values T for pairwise comparisons of the classification results for the 2-class (top) and 3-class (bottom) problem in the context of the discriminant classifier based approaches. Test results, showing evidence against the null-hypothesis are marked shaded gray (when controlling the FWER) or by a '*' (when controlling the FDR).

eters in conjunction with the Euclidean distance, the CBIR strategy relies on the well-founded KL divergence and exhibits considerably better LOOCV rates. Regarding the approaches of Vasconcelos & Lippman [186] and Verdoolage et al. [188], we identify the YBR and RGB colorspace as the most suitable configurations, respectively. A follow-up study on whether there are significant differences in the classification results, however, reveals that there is no evidence against the null-hypothesis for the majority of pairwise comparisons in the two-class case, see Table 4.11. In fact, only the differences between the first and fourth approach in Table 4.9 and 4.10 are significant. Based on the high classification rates we infer that only a few images are misclassified. Consequently, the terms $n_{10}$ and $n_{01}$ in the computation of the McNemar test statistic are rather small. This leads to low values of T which eventually explains the results of Table 4.11. In less technical terms, there is very little space for one approach to produce a notably different classification result compared to the other approaches. The final comparison we make is to compare the classification results achieved by the discriminant classifier based approaches to the top approach of the generative models, see Table 4.12. As we can see, all pairwise comparisons show evidence against the null-hypothesis of the McNemar test. Hence, it is safe to claim that the copula approach is at least superior to any of the approaches listed in Tables 4.6 and 4.7.

| **Approach** | **Accuracy** | **Sensitivity** | **Specificity** | **PPV** | **NPV** |
|---|---|---|---|---|---|
| Copula | 96.65 | 94.94 | 97.33 | 93.37 | 97.98 |
| Kwitt & Uhl, 2008 | 95.06 | 93.26 | 95.77 | 89.73 | 97.29 |
| Vasconcelos & Lippman, 2000 | 94.74 | 84.27 | 98.89 | 96.77 | 94.07 |
| Verdoolaege et al., 2008 | 92.98 | 91.01 | 93.76 | 85.26 | 96.34 |

**Table 4.9:** Comparison of the LOOCV rates for the 2-class problem.

| Approach | Total | Non-Invasive vs. Invasive | | | | |
|---|---|---|---|---|---|---|
| | | Accuracy | Sensitivity | Specificity | PPV | NPV |
| Copula | 93.46 | 95.42 | 97.95 | 86.32 | 96.26 | 92.13 |
| Vasconcelos & Lippman, 2000 | 92.50 | 97.28 | 98.26 | 93.81 | 98.26 | 93.81 |
| Kwitt & Uhl, 2008 | 91.07 | 94.19 | 95.55 | 89.25 | 96.99 | 84.69 |
| Verdoolaege et al., 2008 | 88.52 | 93.35 | 96.36 | 82.42 | 95.21 | 86.21 |

**Table 4.10:** Comparison of the LOOCV rates for the 3-class problem and the subproblem of discriminating non-invasive vs. invasive neoplasia.

## 4.6 Discussion

Summarizing the results of this chapter, we make the following remarks: as with every discriminant classifier approach, the discriminative power of the feature set is the crucial factor for classification performance. The most advanced classifiers can only find correct decision boundaries if the feature set captures the information that is essential to discriminate the classes. Further, discriminant classifiers usually depend on a user-supplied measure of similarity between two feature vectors. Finding a reasonable similarity measure is not a trivial task in many situations, since feature vectors tend to be composed of many different kinds of features. Although the generic Euclidean distance works well in practice, it lacks a reasonable interpretation of the resulting value. Actually, we have seen that using the Euclidean distance for the Weibull/Gamma distribution features is suboptimal and performance can be considerably improved by a theoretically well-founded dissimilarity measure such as the KL divergence. As a matter of fact, the various degrees of freedom we have to cope with in a discriminant classifier scenario often lead to trial and error strategies in finding the most suitable configuration of feature set, similarity measure and classifier. In a generative model based approach, however, the degrees of freedom are more restricted in a certain sense. Basically, only the choice of feature transformation

| | | A | B | C | D |
|---|---|---|---|---|---|
| A | Copula | - | | | |
| B | Kwitt & Uhl, 2008 | 2.02 | - | | |
| C | Vasconcelos & Lippman, 2000 | 2.88 | 0.01 | - | |
| D | Verdoolaege et al., 2008 | 11.25* | 3.34 | 1.58 | - |
| A | Copula | - | | | |
| B | Vasconcelos & Lippman, 2000 | 0.39 | - | | |
| C | Kwitt & Uhl, 2008 | 3.69 | 0.75 | - | |
| D | Verdoolaege et al., 2008 | 15.25* | 6.06* | 3.51 | - |

**Table 4.11:** McNemar test statistic values T for pairwise comparisons of the classification results for the 2-class (top) and 3-class (bottom) problem in the context of the generative model based approaches. Test results, showing evidence against the null-hypothesis are marked gray (when controlling the FWER) or by a '*' (when controlling the FDR).

| **Approach** | **2-class** | **3-class** |
|---|---|---|
| **Copula** | – | – |
| Distribution Features | $9.30^{*}$ | $10.10^{*}$ |
| CES | $8.48^{*}$ | $11.36^{*}$ |
| Van de Wouwer et al, 1997 | $24.45^{*}$ | $28.21^{*}$ |
| CWCC | $26.41^{*}$ | $35.41^{*}$ |
| Manjunath & Ma, 1996 | $25.35^{*}$ | $49.36^{*}$ |
| Swain & Ballard, 1991 | $53.98^{*}$ | $63.56^{*}$ |

**Table 4.12:** McNemar test statistic values for a pairwise comparison of the copula approach to all discriminant classifier approaches. Test results, showing evidence against the null-hypothesis are marked shaded gray (when controlling the FWER) or by a ′$^{*}$′ (when controlling the FDR).

and feature representation is up to the user. Once we have a suitable transformation and an analytically tractable feature representation, we can at least follow the guidelines for selecting the most similar image by relying on the Bayesian formulation of CBIR. Measuring similarity in terms of the maximum likelihood or the minimal KL divergence has a reasonable interpretation in this framework. Regarding a recommendation which strategy to choose in a clinical application, it is hard to make a definitive statement. Although we tend to argue in favor of the generative model based strategy, we have also observed that significant differences in the classification results are rare. It is possible that on another dataset, the margin between the copula approach and the discriminant classifier approaches shrinks and some significant differences vanish. But this is a general issue of any classification problem when there is a lack of available data to perform a large scale study. Consequently, we argue that the conducted experiments should be considered as a prospective evaluation to select a collection of suitable approaches for a final fusion stage, where various predictions of the histology are fused together to a final decision. This fusion might be based on weighting the different predictions by their reliability for instance. However, this is topic of future research and beyond the scope of this thesis.

# Chapter 5

# Watermarking

In this chapter, we address the research topic of image watermarking, a branch of multimedia security where suitable statistical models of wavelet coefficients prove to be highly beneficial. Watermarking has been proposed as a technology to ensure copyright protection by embedding an imperceptible, yet detectable signal in digital multimedia content such as images or video. According to Barni et al. [3], there is a strong resemblance between a watermarking system and a communication system. Embedding watermark information into some host (e.g. an image) asset resembles a transmission process. Any processing steps (e.g. compression, resizing) along the path of the watermarked asset to the receiver can be modeled as a communication channel. Eventually, recovery of the embedded watermark signal corresponds to the receiving side in the communication scenario. In order to identify and delineate the work of this chapter in the wide field of image watermarking, Fig. 5.1 shows a schematic overview of the watermarking system configuration we rely on. As we can see, our focus is on the data recovery side and in particular on *blind* recovery of the watermark signal, i.e. when detection is performed without reference to the unwatermarked host asset $A$. Further, our study is limited to the case of *detectable* watermarks (signified by the yes/no decision in Fig. 5.1) in contrast to *readable* watermarks. In our configuration, the host interferes with the watermark signal. Hence, informed watermark embedding and modeling the host signal are crucial for detection performance [116, 25]. Transform domains – such as the DCT or the DWT domain – facilitate modeling human perception and permit selection of significant signal components for watermark embedding. We follow the embedding strategy of *additive* embedding throughout this chapter, a technique which has spawned many research articles in the last years. A plethora of different detectors has been proposed which all basically improve upon the particular statistical model for the host transform coefficients [69, 139, 17, 125, 12].

The chapter is structured as follows: we start-off with a brief recapitulation of watermarking as a statistical signal detection problem in Section 5.1. In Section 5.2, we introduce a novel watermark detector based on the Cauchy distribution for DWT coefficients. Section 5.3 then introduces another novel detector, specifically tailored for color image watermarking. For both detectors, we conduct an extensive experimental study on the UCID image database [159] and compare against a set of well-known watermarking approaches from literature. Regarding the

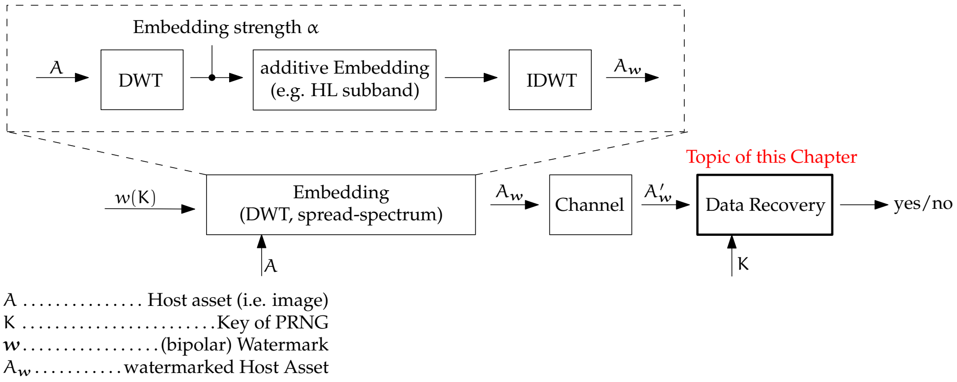

Embedding strength α
A
DWT
additive Embedding (e.g. HL subband)
IDWT
$A_w$
Topic of this Chapter
$w(K)$
Embedding (DWT, spread-spectrum)
$A_w$
Channel
$A'_w$
Data Recovery
yes/no
A
K
A .............. Host asset (i.e. image)
K ........................Key of PRNG
w................(bipolar) Watermark
$A_w$..........watermarked Host Asset

**Figure 5.1:** Configuration of the watermarking system we use in this chapter (adapted form [3]).

contribution of this chapter, we highlight that major parts of the content recently appeared in the following publications:

[94] R. Kwitt, P. Meerwald, and A. Uhl. A lightweight Rao-Cauchy detector for additive watermarking in the DWT-domain. In *Proceedings of the ACM Multimedia and Security Workshop (MMSEC '08)*, pages 33–41, Oxford, UK, September 2008. ACM

[96] R. Kwitt, P. Meerwald, and A. Uhl. Color-image watermarking using multivariate power-exponential distribution. In *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pages 4245–4248, Cairo, Egypt, November 2009. IEEE

## 5.1 Watermarking as a Signal Detection Problem

Regarding our description of the theoretical foundations, we closely adhere to the textbooks of Kay [83], Barni et al. [3] and Cox et al. [28]. The objective of this recapitulation is, to work out the prerequisites to deploy different signal detection strategies for additive spread-spectrum watermarking. We start from classic Neyman-Pearson detection and then successively loosen the requirements on the specification of the host signal noise model. This leads to the idea of Generalized Likelihood Ratio testing (GLRT) and finally, to an asymptotically equivalent formulation of the GLRT known as the Rao hypothesis test. We follow the convention that the transform coefficients of an arbitrary DWT detail subband are referred to as the *host* transform coefficients $x_1, \ldots, x_N$. The *watermark* signal $w_1, \ldots, w_N$ is a realization of N i.i.d. copies of a random variable $W$. For the purpose of additive spread-spectrum watermarking it is convenient to assume that $W$ follows a discrete uniform distribution with equiprobable values in $\{+1, -1\}$, hence, the corresponding p.m.f. of $W$ is given by

$$p_W(x) = \begin{cases} 0.5, & \text{if } x = +1 \\ 0.5, & \text{if } x = -1 \\ 0, & \text{else.} \end{cases} \tag{5.1}$$

The watermark signal is generated by a pseudo-random number generator (PRNG) seeded by some secret key $K$. The rule for additive embedding can be formulated as

$$\forall i : y_i = x_i + \alpha w_i \tag{5.2}$$

where $\alpha > 0$ denotes the embedding strength and the $y_i$ denote the watermarked transform coefficients. The detection problem can be formulated as the detection of a deterministic signal (i.e. the watermark) of unknown amplitude in incompletely specified noise. In terms of hypothesis testing, we can state the null- ($\mathcal{H}_0$) and alternative hypothesis ($\mathcal{H}_1$) as

$$\mathcal{H}_0 : y_i = x_i \quad \text{(not/other watermarked)}, \tag{5.3}$$

$$\mathcal{H}_1 : y_i = x_i + \alpha w_i \quad \text{(watermarked)} \tag{5.4}$$

which is equivalent to the (two-sided) parameter test

$$\mathcal{H}_0 : \alpha = 0, \tag{5.5}$$

$$\mathcal{H}_1 : \alpha \neq 0. \tag{5.6}$$

In the rare case that the p.d.f.s under both hypotheses can be completely specified, we can easily construct a Neyman-Pearson (NP) detector which is optimal in the sense that it maximizes the probability of detection $P_d$ for a fixed probability of false-alarm $P_f$. Given that $p(x; \mathbf{\Theta}_{\mathcal{H}_0})$ and $p(x; \mathbf{\Theta}_{\mathcal{H}_1})$ denote the p.d.f.s under $\mathcal{H}_0$ and $\mathcal{H}_1$, then the Neyman-Pearson theorem states that the optimal detector decides in favor of $\mathcal{H}_1$ if

$$T_L(\mathbf{x}) = \frac{p(\mathbf{x}; \mathbf{\Theta}_{\mathcal{H}_1})}{p(\mathbf{x}; \mathbf{\Theta}_{\mathcal{H}_0})} > \gamma. \tag{5.7}$$

The terms $\mathbf{\Theta}_{\mathcal{H}_0}$ and $\mathbf{\Theta}_{\mathcal{H}_1}$ denote the fully-specified parameter vector(s) of the noise model under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Eq. (5.7) is known as the *Likelihood-Ratio Test (LRT)* with threshold $\gamma$ [83]. In case we can deduce the distribution of the detection statistic $T_L(\mathbf{x})$ under $\mathcal{H}_0$, it is straightforward to determine a suitable threshold for a fixed probability of false-alarm $P_f$ as

$$\gamma = \inf_x \{(1 - F(x)) > P_f\} \tag{5.8}$$

where $F(x)$ denotes the distribution function of $T_L(\mathbf{x})$ under $\mathcal{H}_0$. For example, in case of a standard Normal distribution, i.e. $T_{L|\mathcal{H}_0} \sim \mathcal{N}(0, 1)$, the threshold can be expressed as

$$\gamma = Q^{-1}(P_f) \tag{5.9}$$

where $Q^{-1}$ denotes the inverse Q-function to determine right-tail probabilities of the standard Normal distribution.

In order to constrain the probability of false-alarm, the NP test requires that the distribution of the detection statistic under $\mathcal{H}_0$ does not depend on any unknown parameters. In cases where the noise model p.d.f.s under $\mathcal{H}_0$ and $\mathcal{H}_1$ cannot be fully specified, this requirement is usually violated. In fact, the embedding strength $\alpha$ as well as the distribution parameters of the assumed noise model might be unknown to the detector. Hence, in practice it is more realistic that we have to estimate the unknown parameters from the received signal. Nevertheless, a special case occurs when we assume that the host transform coefficients follow a Gaussian distribution with parameters $\mu$ and $\sigma$. In that case, it is possible to design a NP test as if all parameters were known and obtain a LRT detection statistic which does not depend on the unknown parameters. This detector is commonly referred to as the linear-correlation (LC) detector [83]. In the general case though, it is not feasible to get rid of the unknown parameters.

When we cannot completely specify the noise distribution under both hypotheses, we have to resort to *composite* hypothesis testing approaches. A common strategy to tackle the detection problem is to use a *Generalized Likelihood Ratio Test (GLRT)*. This test replaces the unknown parameters by the corresponding ML estimates conditioned on either $\mathcal{H}_0$ or $\mathcal{H}_1$. In the context of watermarking, this practically means that we have to estimate the embedding strength $\alpha$ from the received signal. For many noise models, however, estimation of $\alpha$ turns out to have no explicit solution. In addition, the noise model parameters under $\mathcal{H}_1$ depend on $\alpha$ which further complicates the estimation task. In the terminology of composite hypothesis testing, the noise model parameters are referred to as the *nuisance parameters*. Although, the focus is on testing $\alpha = 0$ vs. $\alpha \neq 0$, the parameters affect the detection statistics under both hypotheses as well. We follow the convention, that $\boldsymbol{\theta}_{s,\mathcal{H}_0}$ and $\boldsymbol{\theta}_{s,\mathcal{H}_1}$ denote the nuisance parameters under the null- and alternative hypothesis, respectively. The GLRT decides in favor of $\mathcal{H}_1$ if

$$T_G(\mathbf{x}) = \frac{p(\mathbf{x}; \hat{\alpha}, \hat{\boldsymbol{\theta}}_{s,\mathcal{H}_1})}{p(\mathbf{x}; 0, \hat{\boldsymbol{\theta}}_{s,\mathcal{H}_0})} > \gamma \tag{5.10}$$

since $\alpha = 0$ in case of $\mathcal{H}_0$. It is well-known, that the detection statistic $2 \log T_G(\mathbf{x})$ asymptotically (i.e. $N \to \infty$) follows

$$2 \log T_G(\mathbf{x}) \overset{a}{\sim} \begin{cases} \chi_1^2, & \text{under } \mathcal{H}_0 \\ \chi_1^2(\lambda), & \text{under } \mathcal{H}_1 \end{cases} \tag{5.11}$$

where $\chi_1^2$ denotes a Chi-Square distribution with one degree of freedom and $\chi_1^2(\lambda)$ denotes a non-central Chi-Square distribution with one degree of freedom and non-centrality parameter $\lambda$, given by [82]

$$\lambda = \alpha^2[\mathbf{I}_{\alpha\alpha}(0, \boldsymbol{\theta}_s) - \mathbf{I}_{\alpha\boldsymbol{\theta}_s}(0, \boldsymbol{\theta}_s)\mathbf{I}^{-1}_{\boldsymbol{\theta}_s\boldsymbol{\theta}_s}(0, \boldsymbol{\theta}_s)\mathbf{I}_{\boldsymbol{\theta}_s\alpha}(0, \boldsymbol{\theta}_s)]. \tag{5.12}$$

Two examples of the detection statistic p.d.f.s under $\mathcal{H}_0$ and $\mathcal{H}_1$ are shown in Fig. 5.2. By taking a closer look at Eq. (5.11), we see that the GLRT leads to a *Constant False-Alarm Rate* (CFAR) detector since the detection statistic distribution under $\mathcal{H}_0$ does not depend on any parameters at all. Hence, no matter which noise model we choose, the threshold needs to be calculated just one time. The terms $\mathbf{I}_{\alpha\alpha}, \mathbf{I}_{\alpha\boldsymbol{\theta}_s}, \mathbf{I}_{\boldsymbol{\theta}_s\alpha}$ and $\mathbf{I}_{\boldsymbol{\theta}_s\boldsymbol{\theta}_s}$ in Eq. (5.12) denote partitions of the Fisher information matrix, given by:

$$\mathbf{I}_{\alpha\alpha} = \mathbb{E}\left[\frac{\partial \log p}{\partial \alpha}\frac{\partial \log p}{\partial \alpha}\right] \quad 1 \times 1 \tag{5.13}$$

$$\mathbf{I}_{\alpha\boldsymbol{\theta}_s} = \mathbb{E}\left[\frac{\partial \log p}{\partial \alpha}\frac{\partial \log p}{\partial \boldsymbol{\theta}_s}\right] \quad 1 \times s \tag{5.14}$$

$$\mathbf{I}_{\boldsymbol{\theta}_s\alpha} = \mathbb{E}\left[\frac{\partial \log p}{\partial \boldsymbol{\theta}_s}\frac{\partial \log p}{\partial \alpha}\right] \quad s \times 1 \tag{5.15}$$

$$\mathbf{I}_{\boldsymbol{\theta}_s\boldsymbol{\theta}_s} = \mathbb{E}\left[\frac{\partial \log p}{\partial \boldsymbol{\theta}_s}\frac{\partial \log p}{\partial \boldsymbol{\theta}_s}\right] \quad s \times s \tag{5.16}$$

To show a practical example of how to derive a CFAR detector relying on the GLRT, we assume that the DWT detail subband coefficients can be modeled by a Gaussian distribution with zero mean and variance $\sigma^2$, i.e. $X \sim \mathcal{N}(0, \sigma^2)$. The example is similar to the one presented

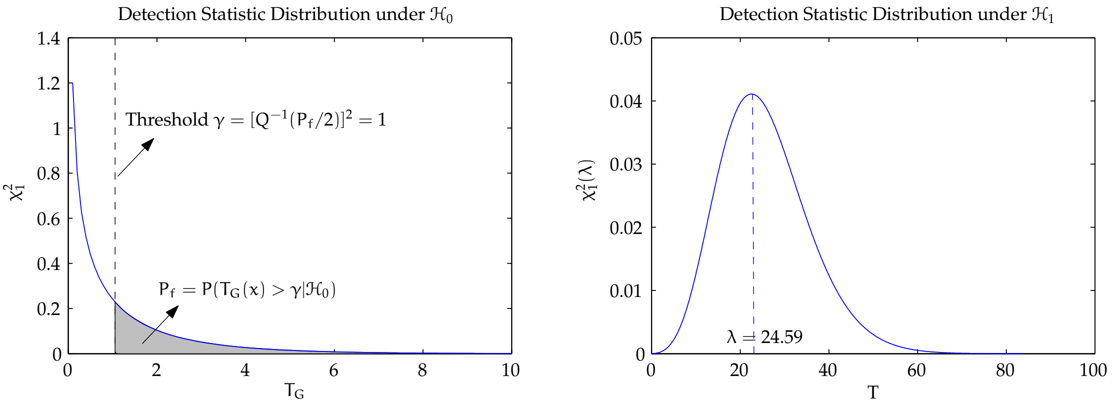


**Figure 5.2:** Illustration of the detection statistic distributions of the GLRT under $\mathcal{H}_0$ and $\mathcal{H}_1$ as well as the probability of false-alarm $P_f$. The threshold is calculated for $P_f \approx 0.3$.

in [82]. First, we need to determine the ML estimates of $\alpha$ and $\sigma^2$ under both hypotheses. To obtain the *restricted* MLE of $\alpha$ and $\sigma^2$, i.e. the ML estimates under $\mathcal{H}_1$, we formulate the log-likelihood function as

$$L(\alpha, \sigma; y_1, \dots, y_N) = \log\left(\frac{1}{(2\pi\sigma^2)^{N/2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \alpha w_i)^2. \tag{5.17}$$

Taking the derivative w.r.t. $\alpha$ and setting the resulting equation to zero gives

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i w_i \tag{5.18}$$

as the MLE of $\alpha$. The restricted MLE $\hat{\sigma}_1^2$ is obtained by taking the partial derivative of Eq. (5.17) w.r.t. $\sigma$ and setting the corresponding term to zero. This gives

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\alpha} w_i)^2 \tag{5.19}$$

which, in combination with Eq. (5.18), finally allows to write the host signal noise p.d.f. under $\mathcal{H}_1$ as

$$p(\mathbf{x}; \hat{\alpha}, \hat{\sigma}_1^2) = \frac{1}{2\pi\hat{\sigma}_1^2}\exp\left(-\frac{N}{2}\right). \tag{5.20}$$

Under the null-hypothesis, we know that $\alpha = 0$ and the MLE of $\sigma^2$ – denoted as the *unrestricted* MLE $\hat{\sigma}_0^2$ – is the sample variance of $y_1, \dots, y_N$. Eventually, the detection statistic of the GLRT for the CFAR detector is

$$2\log T_G(\mathbf{x}) = N \log\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}. \tag{5.21}$$

This detector is only asymptotically equivalent to the LC detector. However, we highlight that the threshold $\gamma$ can be set to a predefined value and does not have be determined for each new signal. Since the number of DWT coefficients N is usually quite large in case of images, we don't have to worry about signal length issues related to the asymptotic performance of the GLRT. Letting $Q^{-1}_{\chi^2_1}$ denote the Q-function to express right-tail probabilities of the Chi-Square distribution with one degree of freedom, $\gamma$ can be set according to

$$\gamma = Q^{-1}_{\chi^2_1}(P_f) \tag{5.22}$$

where $P_f$ denotes the desired probability of false-alarm, e.g. $P_f = 10^{-3}$. A presumably more convenient way to express $\gamma$ is to rely on the relation that right-tail probabilities of the $\chi^2_1$ distribution can also be expressed by means of the Q-function to compute right-tail probabilities of the Gaussian distribution, i.e. $Q_{\chi^2_1}(x) = 2Q(\sqrt{x})$. Hence,

$$\gamma = \left[Q^{-1}\left(\frac{P_f}{2}\right)\right]^2 \tag{5.23}$$

which is usually easier to handle due to existing implementations of the Gaussian Q-function. Fig. 5.2 illustrates a threshold (dashed line) of $\gamma = 1$ for $P_f \approx 0.32$. In order to determine the non-centrality parameter $\lambda$, we can rely on a theorem of Kay [82] which considers the special case of $p(x; \alpha, \boldsymbol{\theta}_s) = -p(x; \alpha, \boldsymbol{\theta}_s)$. The theorem states that the symmetry of the noise model leads to $\mathbf{I}_{\alpha\boldsymbol{\theta}_s} = \mathbf{0}$ which reduces the expression in Eq. (5.12) to

$$\lambda = \alpha^2[\mathbf{I}_{\alpha\alpha}(0, \boldsymbol{\theta}_s)]. \tag{5.24}$$

In our example of a Gaussian host signal, we thus have $\lambda = \alpha^2[\mathbf{I}_{\alpha\alpha}(0, \sigma^2)]$. The corresponding partition $\mathbf{I}_{\alpha\alpha}(\alpha, \sigma^2)$ of the Fisher information matrix can be derived from (see [82])

$$\mathbf{I}_{\alpha\alpha}(\alpha, \sigma^2)] = \mathbb{E}\left[\left(\frac{\partial \log p(x_i - \alpha w_i; \alpha, \sigma^2}{\partial \alpha}\right)^2\right] = \sum_{i=1}^{N} w_i^2 \int_{-\infty}^{\infty} \left[\frac{p'(n; \alpha, \sigma^2)}{p(n; \alpha, \sigma^2)}\right]^2 p(n; \alpha, \sigma^2) dn \tag{5.25}$$

where we have set $n_i = x_i - \alpha w_i$ and $p'(n; \alpha, \sigma^2)$ denotes the first partial derivative of $p(n; \alpha, \sigma^2)$ w.r.t. $n$. After some calculus, we obtain

$$\mathbf{I}_{\alpha\alpha}(0, \sigma^2) = \sum_{i=1}^{N} w_i^2 \frac{1}{\sigma^2} = \frac{N}{\sigma^2} \tag{5.26}$$

and the non-centrality parameter $\lambda$ takes the form

$$\lambda = \alpha^2 \frac{N}{\sigma^2}. \tag{5.27}$$

An alternative approach to tackle the problem of composite hypothesis testing is to rely on the asymptotic equivalence of the GLRT and the Rao hypothesis test [153]. The compelling advantage of the Rao hypothesis test is that it does not require to compute ML estimates for $\alpha$ and $\boldsymbol{\theta}_s$ under $\mathcal{H}_1$. Only the ML estimates under $\mathcal{H}_0$ are required for detection. Since we know that $\alpha = 0$ in case of $\mathcal{H}_0$, the Rao test is particularly useful in situations where the embedding side does not want to inform the detector about the choice of embedding strength. As pointed out by Barni et al. [3], this is an important degree of freedom, since it allows the embedding side to adjust the embedding strength to the signal at hand. The Rao test decides $\mathcal{H}_1$ in case

$$T_R(\mathbf{x}) = \left.\frac{\partial \log p(\mathbf{x}; \boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}}\right|^T_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \left[\mathbf{I}^{-1}(\hat{\boldsymbol{\Theta}})\right]_{\alpha\alpha} \left.\frac{\partial \log p(\mathbf{x}; \boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}}\right|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} > \gamma \tag{5.28}$$

where $\hat{\Theta} = [\hat{\alpha}\ \hat{\theta}_{s,\mathcal{H}_0}]$ denotes the ML estimates under $\mathcal{H}_0$, e.g. $\hat{\Theta} = [0\ \hat{\sigma}_0^2]$ for our previous problem or $\hat{\Theta} = [0\ \hat{\theta}_{s,\mathcal{H}_0}]$ for a general nuisance parameter vector. Further, the term $[\mathbf{I}(\Theta)]^{-1}_{\alpha\alpha}$ is given by

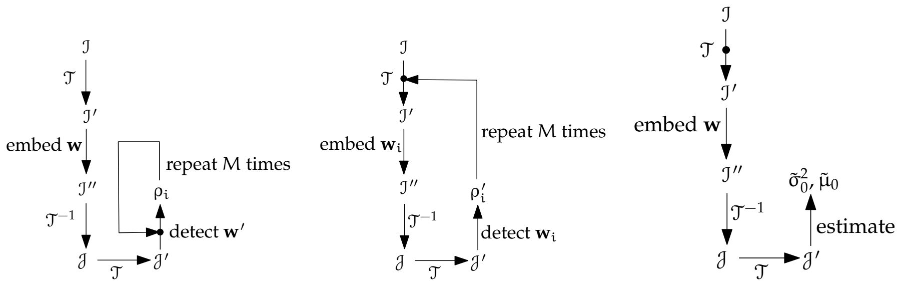
$$[\mathbf{I}(\Theta)]^{-1}_{\alpha\alpha} = \left(\mathbf{I}_{\alpha\alpha}(\Theta) - \mathbf{I}_{\alpha\theta_s}(\Theta)\mathbf{I}^{-1}_{\theta_s\theta_s}(\Theta)\mathbf{I}_{\theta_s\alpha}(\Theta)\right)^{-1} \tag{5.29}$$

where the partitions of the Fisher information matrix are defined in Eqs. (5.13) to (5.16). Due to the asymptotic equivalence to the GLRT, the Rao hypothesis test inherits the distribution of the detection statistic, i.e. $T_R(\mathbf{x}) \sim 2\log T_G(\mathbf{x})$, see Eq. (5.11). Consequently, we obtain a CFAR detector with the advantage to avoid ML estimation of the embedding strength $\alpha$. In [139], Nikolaidis et al. first exploit this test to derive a watermark detector for additive spread-spectrum watermarks in the DWT domain, based on a Generalized Gaussian noise model. In Section 5.2, we introduce a Rao hypothesis test conditioned on a Cauchy host signal noise model.

As a final remark of this section, we highlight that the *Neyman-Pearson criterion* is a quite overused term in watermarking literature. It is customary to derive a LRT-based detector for some noise model and refer to the Neyman-Pearson criterion for threshold selection. This however implies that we can actually constrain the probability of false-alarm which basically requires that the detection statistic under $\mathcal{H}_0$ does not depend on any unknown parameters. Taking a closer look at popular detectors in literature, e.g. [69, 12], reveals that this is usually not the case due to unknown noise parameters or unknown embedding strength. Given that we assume knowledge of the embedding strength at the detector, the noise parameters are still unknown and have to be estimated from the received signal. Consequently, the resulting detectors are not NP detectors but rather *estimate-and-plug* detectors, as pointed out by Kay [83]. The threshold will be biased because the watermark may be present in the received signal. Nevertheless, estimate-and-plug detectors are a reasonable choice in situations where the noise model leads to intractable expressions for the GLRT or Rao hypothesis test.

### 5.1.1 Evaluation of Detector Performance & ROC curves

A critical issue with any watermarking system is how to evaluate the performance of the detector. A convenient strategy is to construct Receiver Operator Characteristic (ROC) curves. Although, we will later see that the ROC curve plots are disadvantageous when evaluating detection performance on a large number of images, the general construction principle is worth a discussion. Usually, we plot the probability of detection $P_d$ (or miss $P_m$) in dependence of the probability of false-alarm $P_f$. In order to infer conclusions about the detector performance based on ROC curves, we first have to ensure that the detector retains the desired $P_f$. This is an important point, since in any practical situation we expect the actual host signal noise to deviate from the theoretical model to some extent. In case of the GLRT or Rao test for instance, we have to check whether the detection responses under $\mathcal{H}_0$ in fact follow a Chi-Square distribution with one degree of freedom. Other detectors, e.g. the LC detector, require to verify that the detection statistic follows a Gaussian law. A reasonable way to perform these checks is to skip the embedding step and to call the watermark detector $M$ times on unwatermaked transform coefficients. This gives $M$ detector responses, say $\rho_1, \dots, \rho_M$, which we can use in a GoF test to check if there is evidence against the null-hypothesis. In case there is no evidence, it is safe to set $P_f$ to the desired level. Of course we could also count the number of false detections among the $M$ detector responses and compare against the expected number of false detections. However, this is computationally not feasible for small values of $P_f$ (e.g. $P_f = 10^{-10}$) which is why we favor the former strategy. The next step is, to determine $P_m$ (or $P_d$). For that purpose,



**(a)** other watermark ($\mathbf{w} \neq \mathbf{w}'$) **(b)** watermarked ($\mathbf{w}_i = \mathbf{w}_i$) **(c)** estimate $\mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$

**Figure 5.3:** Schematic process description of how to determine the detection responses under $\mathcal{H}_0$ (i.e. $\rho_1, \ldots, \rho_M$), $\mathcal{H}_1$ (i.e. $\rho_1', \ldots, \rho_M'$) and the calculation of the detection statistic parameters $\tilde{\mu}_0$ and $\tilde{\sigma}_0^2$ from the received signal.

we successively embed and detect M watermarks to obtain M responses, say $\rho_1', \ldots, \rho_M'$ ,under $\mathcal{H}_1$. We can then estimate the corresponding detection statistic parameters under $\mathcal{H}_1$ and finally plot the ROC curves. In case of a GLRT for instance, we can exploit the relationship [83]

$$P_d = Q(Q^{-1}(1/2 P_f) - \sqrt{\lambda}) + Q(Q^{-1}(1/2 P_f) + \sqrt{\lambda}) \tag{5.30}$$

to express $P_d$ as a function of $P_f$ and $\lambda$. The non-centrality parameter can be estimated by remembering that given $X \sim \chi_1^2(\lambda)$, it can be shown that $\sqrt{X} \sim \mathcal{N}(\sqrt{\lambda}, 1)$ and thus

$$\hat{\lambda} = \left( \frac{1}{N} \sum_{i=1}^{N} \sqrt{\rho_i'} \right)^2 . \tag{5.31}$$

Inserting $\hat{\lambda}$ in Eq. (5.30) gives the *semi-experimental* ROC curve. We use the term *semi-experimental* since a fully experimental ROC curve would imply counting of the number of missed detections. The general expression to determine the semi-experimental probability of miss $\hat{P}_m$ is

$$\hat{P}_m = \mathbb{P}(T(\mathbf{x}) < \gamma) = F(\gamma; \hat{\mathfrak{b}}) \tag{5.32}$$

where F denotes the c.d.f. of the detection statistic T under $\mathcal{H}_1$, parametrized by $\hat{\mathfrak{b}}$. The semi-experimental evaluation strategy is of particular relevance when it comes to measuring the performance of a detector under attacks. Due to the vast number of possible attacks on the watermarked image, there is no way we could incorporate the attack characteristics into the host signal model in a tractable manner. As a matter of fact, evaluation of the watermark detection performance amounts to an experimental study. The semi-experimental way allows to plot ROC curves even for low values of $P_f$. A graphical visualization of the whole strategy is shown in Figs. 5.3a and 5.3b where $\mathcal{T}$ denotes the transformation of an image $\mathcal{I}$ to a suitable transform domain representation $\mathcal{I}'$, e.g. by a DWT. The watermarked image in the transform domain is denoted by $\mathcal{I}''$ and $\mathcal{T}^{-1}$ denotes the corresponding inverse transformation.

## 5.2 A Rao Hypothesis Test for Cauchy Host Signal Noise

One main motivation for deriving a novel watermark detector for host signal noise distributed other than Generalized Gaussian, is the fact that ML estimation of the GGD parameters is computationally expensive and requires a numerical root-finding procedure (see Chapter 3). Since the Cauchy distribution is a reasonable model for DWT transform coefficients and parameter estimation can be performed efficiently, chances are high that we can derive a computationally simple and effective watermark detector. While other approaches such as [15] aim for a reduction in watermark sequence length to enhance computational performance, we try to reduce the computational effort per step in the detection process. We start by deriving the first part of the detection statistic of Eq. (5.28)

$$\left[\frac{\partial \log p(\mathbf{x};\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}}\right]^2 = \left[\sum_{i=1}^{N} \frac{\partial \log p(y_i - \alpha w_i;\gamma)}{\partial \alpha}\right]^2 \tag{5.33}$$

with $\boldsymbol{\Theta} = [\alpha\ \gamma]$. Inserting the p.d.f. of the Cauchy distribution leads to

$$\sum_{i=1}^{N} \frac{\partial \log p(y_i - \alpha w_i;\gamma)}{\partial \alpha} \overset{(2.4)}{=} \sum_{i=1}^{N} \frac{2w_i(y_i - \alpha w_i)}{\gamma^2\left(1 + \frac{(y_i-\alpha w_i)^2}{\gamma^2}\right)}. \tag{5.34}$$

We next evaluate this expression at the ML estimate $\hat{\boldsymbol{\Theta}} = [0\ \hat{\gamma}]$ and take the power of two to obtain

$$\left[\frac{\partial \log p(\mathbf{x};\boldsymbol{\Theta})}{\partial \boldsymbol{\alpha}}\right]^2 \Bigg|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} = 4\left[\sum_{i=1}^{N} \frac{y_i w_i}{\hat{\gamma}^2 + y_i^2}\right]^2 . \tag{5.35}$$

In the second step, we need to derive an expression for

$$[\mathbf{I}(\boldsymbol{\Theta})]^{-1} = (\mathbf{I}_{\alpha\alpha}(\boldsymbol{\Theta}))^{-1} \tag{5.36}$$

which is the only term that is left over from Eq. (5.29), since we know that $\mathbf{I}_{\alpha\theta_s} = \mathbf{0}$ in case of a symmetric p.d.f. We modify Eq. (5.25) accordingly to obtain

$$\mathbf{I}_{\alpha\alpha}(\alpha,\gamma) = \sum_{i=1}^{N} w_i^2 \int_{-\infty}^{\infty} \left[\frac{p'(n;\alpha,\gamma)}{p(n;\alpha,\gamma)}\right]^2 p(n;\alpha,\gamma)dn = \frac{1}{2\gamma^2}\sum_{i=1}^{N} w_i^2 = \frac{N}{2\gamma^2}. \tag{5.37}$$

By using Eq. (5.28) and inserting the ML estimate $\hat{\boldsymbol{\Theta}} = [0\ \hat{\gamma}]$ under $\mathcal{H}_0$, we obtain the following expression for the detection statistic of our Rao hypothesis test conditioned on Cauchy host signal noise

$$T_R(\mathbf{y}) = \left[\sum_{i=1}^{N} \frac{y_i w_i}{\hat{\gamma}^2 + y_i^2}\right]^2 \frac{8\hat{\gamma}^2}{N}. \tag{5.38}$$

Based on Eq. (5.37) it is then straightforward to deduce the expression for the non-centrality parameter of the detection statistic under $\mathcal{H}_1$ as

$$\lambda = \alpha^2 \mathbf{I}_{\alpha\alpha}(0,\gamma) = \frac{N\alpha^2}{2\gamma^2}. \tag{5.39}$$

We will next test our theoretical expressions by means of artificially generated data and then go on to an experimental evaluation of the watermark detector on real data. Our test works as follows: we generate Cauchy distributed host signal noise samples $x_1, \ldots, x_N$ as realizations of N i.i.d. copies of a random variable $X \sim \mathcal{C}(\gamma)$, where $\mathcal{C}(\gamma)$ denotes a Cauchy distribution with shape parameter $\gamma$. We set $\gamma = 5$, $N = 10^5$ and generate the bipolar watermark sequence

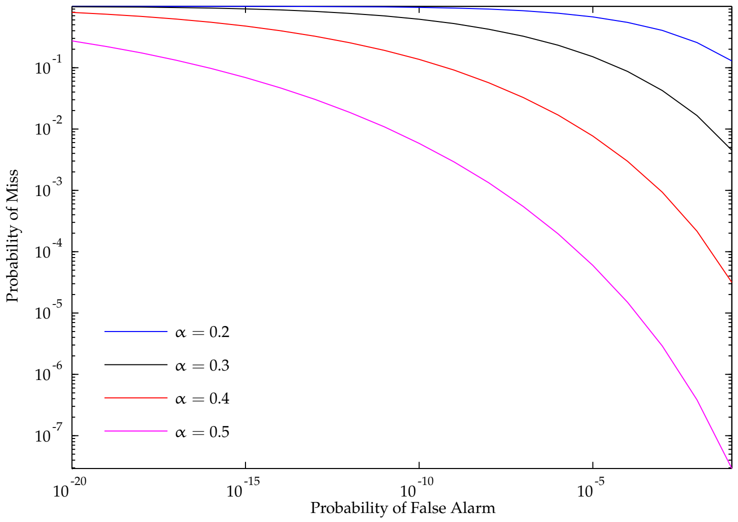


**Figure 5.4:** ROC curves for different embedding strengths of additively embedded (bipolar) watermarks in artificially generated Cauchy host signal noise samples with $\gamma = 5$, $N = 10000$.

$w_i$ as mentioned above. Further, we let the embedding strength $\alpha$ vary between 0.2 and 0.5 with a stepsize of 0.1. From Eq. (5.39) we expect $\lambda = 8, 18, 32, 50$ for this setup. The detector responses under $\mathcal{H}_0$ and $\mathcal{H}_1$ are determined as illustrated in Figs. 5.3a and 5.3b with $M = 1000$. The ROC curves are shown in Fig. 5.4. Obviously, increasing the embedding strength leads to better detector performance. The estimated values for the non-centrality parameter $\lambda$ are $8.68, 16.35, 32.47$ and $50.94$, resp., which conforms the validity of our derivation. For a practical watermarking scenario, however, it is not a smart choice to set the watermarking strength arbitrarily. In additive spread-spectrum watermarking, $\alpha$ is usually determined based on the Data-to-Watermark (DWR) ratio, expressed in decibel (dB). In our context, the term *Data* refers to the DWT detail subband coefficients which we use for embedding. According to [3], the DWR is given by the expression

$$DWR = 10 \log_{10} \left( \frac{\sigma_x^2}{\alpha^2 \sigma_w^2} \right) \tag{5.40}$$

where $\sigma_x^2$ denotes the variance of the DWT detail subband coefficients and $\sigma_w^2$ denotes the variance of the watermarking sequence which in our case (i.e. bipolar watermark) equals 1. Hence, we can express the embedding strength $\alpha$ as a function of the DWR and the variance of the host signal as

$$\alpha = \sqrt{\frac{\sigma_x^2}{\exp\left(\frac{\log(10) \cdot DWR}{10}\right)}}. \tag{5.41}$$

The embedding strengths of the previous example (i.e. $\alpha = 0.2, 0.3, 0.4, 0.5$) correspond to DWRs of 67.05dB, 61.53dB, 59.03dB and 57.09dB. As we can see, the DWR is rather high; reasonable DWRs for image watermarking are usually set to achieve a PSNR of 30dB to 50dB, i.e. the DWR is in the range of 12dB to 20dB.

### 5.2.1 Experiments

To conduct a comparative study of detector performance, we have to introduce the experimental setup first. We use all images from the UCID image database. Since the original images are color images, we first conduct a conversion to grayscale images by means of MATLAB's `rgb2gray` routine. Then, we extract a 256 × 256 pixel block from each image, starting in the top-left hand corner. Finally, all images are resized to 128 × 128 pixel using MATLAB's `imresize` routine which basically performs bicubic interpolation.

We implement the following detectors for additive spread-spectrum watermarks: the naming convention is, that the first part of the name denotes the host signal noise model and the second part denotes the type of hypothesis test. For example, *Cauchy-LRT* signifies that the host signal noise is modeled by a Cauchy distribution and the hypothesis test is a LRT. We highlight that all mentioned LRT detectors are estimate-and-plug detectors and assume that the embedding strength $\alpha$ is known at the detection side.

**Linear Correlator (Gaussian-LRT, LC)** This detector arises when we derive a LRT for Gaussian host signal noise, assuming that all parameters are known a-priori. It turns out, that the resulting detection statistic

$$T_1(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i w_i \tag{5.42}$$

does neither depend on the noise distribution parameters nor on the embedding strength and hence the resulting detector is a NP detector. The expressions for the mean and variance of the detection statistic under $\mathcal{H}_0$ and $\mathcal{H}_1$ are given in [3].

**Generalized Gaussian LRT (GGD-LRT)** This detector is introduced by Hernandez et al. [69], based on the LRT and a Generalized Gaussian host signal noise model. The detection statistic is given by

$$T_2(\mathbf{y}) = \hat{a}^{\hat{c}} \sum_{i=1}^{N} \left( |y_i|^{\hat{c}} - |y_i - \alpha w_i|^{\hat{c}} \right) \tag{5.43}$$

where the distribution parameters $a$ and $c$ are estimated from the received signal without caring whether a watermark is present or not.

**Cauchy LRT (Cauchy-LRT)** This detector is introduced by Briassouli et al. [12] as an extension of the GGD-LRT detector. The host signal noise is modeled by a Cauchy distribution and the detection statistic is given as

$$T_3(\mathbf{y}) = \sum_{i=1}^{N} \log\left( \frac{\hat{\gamma}^2 + y_i}{\hat{\gamma}^2 + (y_i - \alpha w_i)^2} \right). \tag{5.44}$$

**Generalized Gaussian Rao (GGD-Rao)** Nikolaidis et al. [139] first propose to use a Rao hypothesis test as a replacement of the plug-and-estimate detectors based on the LRT. Their work is motivated by the problem of informing the detector about the choice of $\alpha$ and the bias introduced by estimating the noise distribution parameters from the received signal. Based on the results of Kay [82], the authors derive a Rao test assuming Generalized Gaussian host signal noise with the detection statistic given by

$$T_4(\mathbf{y}) = \frac{\left( \sum_{i=1}^{N} \operatorname{sgn}(y_i) w_i |y_i|^{\hat{c}-1} \right)^2}{\sum_{i=1}^{N} |y_i|^{2\hat{c}-2}} \tag{5.45}$$

| **Detector** | $\hat{\mu}_0 \leqslant \tilde{\mu}_0$ | $\hat{\sigma}_0^2 \leqslant \tilde{\sigma}_0^2$ | $\rho \sim \mathcal{N}(\hat{\mu}_0, \tilde{\sigma}_0^2)$ | $\rho \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ | $\rho \sim \chi_1^2$ | FP |
|---|---|---|---|---|---|---|
| GGD-LRT | 98.36 | 15.40 | 88.57 | 2.54 | - | $0.1 \cdot 10^{-3}$ |
| Cauchy-LRT | 98.43 | 41.03 | 96.83 | 2.39 | - | $0.1 \cdot 10^{-3}$ |
| Gaussian-LRT | 62.63 | 59.34 | 97.16 | 95.07 | - | $0.9 \cdot 10^{-3}$ |
| Cauchy-Rao | - | - | - | - | 82.88 | $1.7 \cdot 10^{-3}$ |
| GGD-Rao | - | - | - | - | 99.48 | $0.8 \cdot 10^{-3}$ |

**Table 5.1:** Evaluation whether the detection statistic distributions under $\mathcal{H}_0$ conform to the expected distributions computed on the basis of the received signal. The numbers represent the percentage of UCID images where the test (given as column title) does not fail. The FP column lists the number of observed false positives.

where $\operatorname{sgn}(\cdot)$ denotes the Signum function.

First, we verify that the detectors actually exhibit the theoretically stated detection statistic distribution under the null-hypothesis, since this allows to set the detection threshold to a given probability of false-alarm. We verify that (i) the detector responses under $\mathcal{H}_0$ follow a Gaussian distribution for all LRT detectors and a $\chi_1^2$ distribution for the Rao detectors; (ii) the detection statistic parameters $\tilde{\mu}_0, \tilde{\sigma}_0^2$ calculated from the received signal (see Fig. 5.3c) correspond to the detection statistic parameters $\hat{\mu}_0, \hat{\sigma}_0^2$ estimated from the detection responses $\rho_1, \dots, \rho_M$ under $\mathcal{H}_0$. Due to the fact that we perform tests on all UCID images, we cannot directly compare the detection statistic parameters by listing them in a table, such as in [69]. As an alternative, we choose the following strategy: We perform a Chi-Square GoF test to check whether $\rho_1, \dots \rho_M \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ or $\rho_1, \dots, \rho_M \sim \chi_1^2$, respectively. The results are listed in column five of Table 5.1. Except for the Gaussian-LRT detector, however, there is evidence against the null-hypothesis in more than 90% of all cases. A closer look at the data reveals, that $\hat{\mu}_0$ and $\tilde{\mu}_0$ differ considerably in some cases. Nevertheless, the variances $\hat{\sigma}_0^2$ and $\tilde{\sigma}_0^2$ coincide to a large extent. Based on this observation, we need to know whether the difference in mean has any negative effect on the probability of false-alarm, i.e. whether the actually observed detection statistic distribution is shifted to the right. To rule out such a negative effect, we check if (i) $\hat{\mu}_0 \leqslant \tilde{\mu}_0$ and (ii) if $\rho_1, \dots, \rho_M \sim \mathcal{N}(\hat{\mu}_0, \tilde{\sigma}_0^2)$, see columns one and three of Table 5.1. Given that both tests show no evidence against the null-hypothesis, the detection threshold based on $\tilde{\mu}_0$ and $\tilde{\sigma}_0^2$ is conservative in the sense that the probability of false-alarm will be lower than expected (see Fig. 5.5). In the last column of Table 5.1 we additionally list the number of observed false positives. The numbers are in quite good accordance to the predefined probability of false-alarm of $P_f = 10^{-3}$.

**Performance without attacks** Due to the large amount of images, we cannot present classic ROC curves to evaluate the performance of the detectors. Further, our objective is to assess the detector performance on the whole image database and not only on a selected set of images. In particular, we are more interested in the ranking of the detectors in critical conditions, i.e. when $P_m$ is high. This is reasonable, since in practice we are not concerned about detector performance when $P_m \approx 10^{-100}$ for example. To provide such a comparative study, we fix the probability of false-alarm at a specified level, say $10^{-6}$, and construct a c.d.f. plot of the corresponding $P_m$ values on a logarithmic scale for each detector. We then *zoom-in* on our Region of Interest (ROI), i.e. where $P_m$ is high. To warrant this strategy, Fig. 5.7 shows the original c.d.f. plot on the left-hand side and a zoomed-in plot to

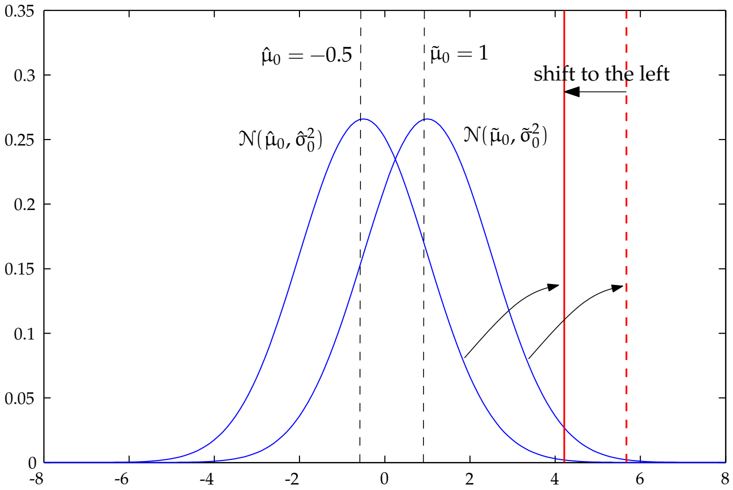


**Figure 5.5:** Illustration of the difference in mean between the theoretically expected detection statistic distribution $\mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ and the empirically observed $\mathcal{N}(\hat{\mu}_0, \hat{\sigma}_0^2)$. The red lines signify that the detection threshold is actually shifted to the left.

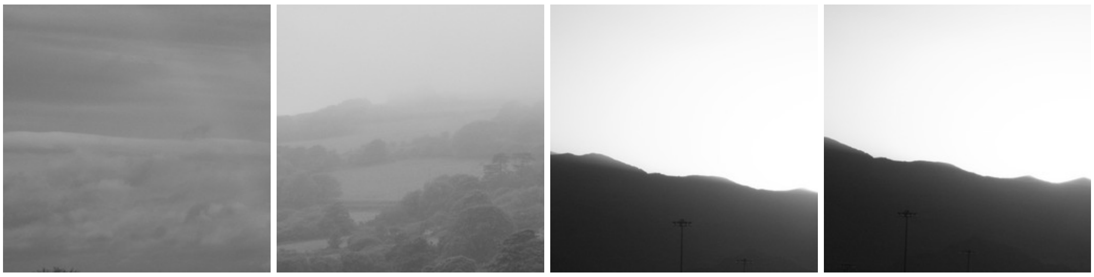


**Figure 5.6:** Four UCID images, `ucid00246`, `ucid00444`, `ucid01059` and `ucid01060` where all five detectors fail to detect the watermark when embedding at a DWR of 12dB.

our ROI on the right-hand side. The DWR is set to 12dB for watermark embedding which leads to an average PSNR of ≈ 42dB. Comparing both plots illustrates the point that the ranking of the detectors changes as we move towards a higher probability of miss. In detail, we observe that although the GGD-LRT and Cauchy-LRT detectors exhibit far better performance then both Rao detectors when $P_m < 10^{-100}$, the situation changes considerably when $P_m > 10^{-4}$. We observe, that the LRT detectors perform poorly and even fail in some cases while the Rao detectors still exhibit $P_m$ values of $< 10^{-2}$. The plots further highlight, that presenting ROC plots for a small selection of images cannot provide full insight into the ranking of detectors. It is even possible, that the ROC curves for a few images might convey a completely wrong impression about detector performance. We further note, that we can identify four UCID images where all detectors fail to detect the watermark in our setup. These images were excluded from the plots and are shown in Fig. 5.6. Detection failure occurs, since the embedding strength corresponding to a DWR of 12dB is too low for these images, resulting in no watermark presence.

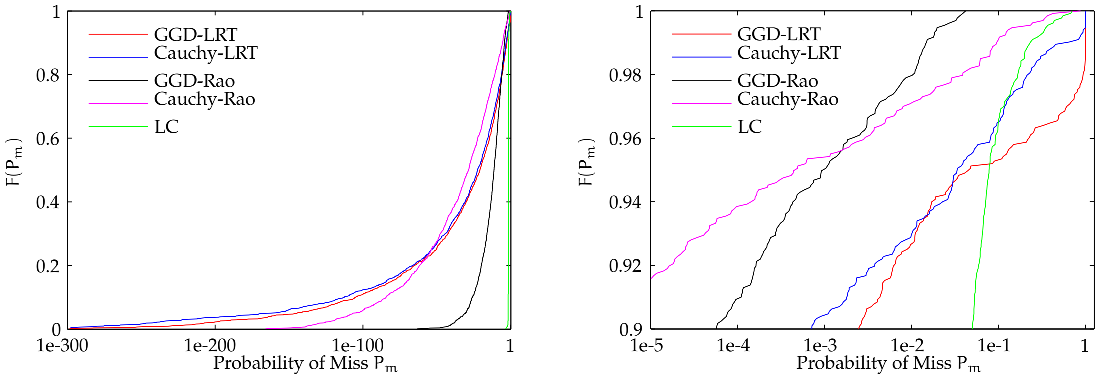


**Figure 5.7:** C.d.f. plots of the probability of miss $P_m$ for a fixed probability of false-alarm $P_f = 10^{-6}$ and a DWR of 12dB over 1334 UCID images including a zoomed-in version of the region of interest, i.e. where $P_m > 10^{-5}$.

**Performance under attacks** To evaluate the performance of the Cauchy-Rao detector under an attack, we choose JPEG compression with quality factors $Q = 30$ and $Q = 70$. The mean PSNR over the whole UCID image database for quality factor $Q = 30$ is $\approx 30$dB, whereas we obtain $\approx 33$dB for quality factor $Q = 70$. In both cases, we fix the probability of false-alarm to $P_f = 10^{-3}$, since for lower $P_f$ values we do not get any reasonable results w.r.t our low image size of $128 \times 128$. As in the previous experiments, the watermark embedding strength is set to obtain a DWR of 12dB. Table 5.2 lists the test results when evaluating the detection statistic distribution under $\mathcal{H}_0$. We observe the interesting effect, that the actual detection statistic distributions are pretty close to the theoretical ones. Comparing Table 5.2 to Table 5.1 shows that there is no evidence against the null-hypothesis $\mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ in almost all cases. Only for JPEG quality factor 30, the number of images where we observe evidence against $\chi_1^2$ increases slightly. The c.d.f. plots of $P_m$ over (almost) all UCID images are shown in Fig. 5.8. We again excluded the images shown in Fig. 5.6, since all detectors failed to detect the watermark. The left-hand side of Fig. 5.8 shows the unscaled versions of the c.d.f. plots and the right-hand side shows a zoomed-in version, where we focus on the most interesting region. In contrast to Fig. 5.7, the performance of the GGD-Rao detector strongly deteriorates and the LC detector starts to show acceptable performance for quality factor $Q = 30$. A possible explanation for the poor results of the GGD-Rao detector is the negative impact of JPEG compression on the ML parameter estimation procedure of the GGD. Regarding the Cauchy-Rao detector, we observe stable behavior over the whole range of $P_m$ values. Even when $P_m$ is high, the Cauchy-Rao detector exhibits acceptable performance.

### 5.2.2 Some Computational Considerations

As a final part of this section, we take a closer look at the computational requirements of each detector. This is a necessary step, since we originally proposed the Cauchy-Rao detector as a *lightweight* alternative to the GGD-based detectors. In fact, low computational complexity was a key motivation to derive a novel watermark detector. In particular, we consider the number of arithmetic operations to calculate the detection statistics, briefly discuss parameter estima-

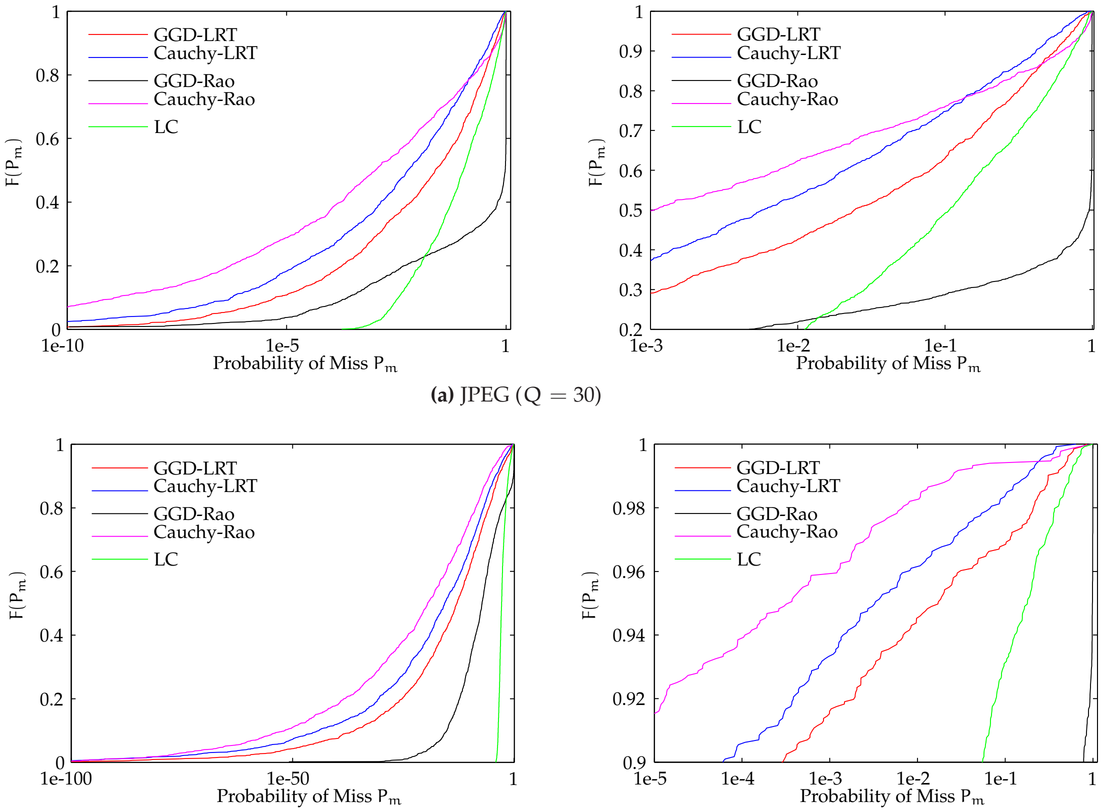
| **Detector** | $\hat{\mu}_0 \leqslant \tilde{\mu}_0$ | $\hat{\sigma}_0^2 \leqslant \tilde{\sigma}_0^2$ | $\rho \sim \mathcal{N}(\hat{\mu}_0, \tilde{\sigma}_0^2)$ | $\rho \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ | $\rho \sim \chi_1^2$ | FP |
|---|---|---|---|---|---|---|
| GGD-LRT [69] | 66.44 | 54.11 | 97.53 | 95.94 | - | $0.96 \cdot 10^{-3}$ |
| Cauchy-LRT [12] | 66.26 | 53.51 | 97.76 | 95.44 | - | $0.94 \cdot 10^{-3}$ |
| Gaussian-LRT | 56.88 | 52.47 | 96.56 | 94.62 | - | $0.92 \cdot 10^{-3}$ |
| GGD-Rao [139] | - | - | - | - | 72.50 | $0.66 \cdot 10^{-3}$ |
| Cauchy-Rao | - | - | - | - | 74.96 | $0.45 \cdot 10^{-3}$ |
| GGD-LRT [69] | 77.06 | 58.52 | 97.76 | 95.37 | - | $4.7 \cdot 10^{-3}$ |
| Cauchy-LRT [12] | 78.33 | 59.04 | 97.98 | 95.81 | - | $1.5 \cdot 10^{-3}$ |
| Gaussian-LRT | 59.49 | 51.12 | 97.46 | 96.79 | - | $0.9 \cdot 10^{-3}$ |
| GGD-Rao [139] | - | - | - | - | 92.75 | $0.07 \cdot 10^{-3}$ |
| Cauchy-Rao | - | - | - | - | 96.79 | $0.05 \cdot 10^{-3}$ |

**Table 5.2:** Evaluation whether the detection statistic distributions under $\mathcal{H}_0$ conforms to the expected distributions computed on the basis of the received signal under the influence of JPEG compression with quality factors $Q = 30$ (top) and $Q = 70$ (bottom). The numbers represent the percentage of UCID images where the test (given as column title) does not fail. The column FP lists the number of observed false-positives.

tion issues and highlight the advantages of the Rao detectors w.r.t. threshold determination. By *arithmetic operations*, we understand the number of additions & subtractions $(+, -)$, multiplications & divisions $(\times, \div)$, logarithms & exponentiations $(\log, \text{pow})$ as well as computation of sgn and $|\cdot|$. In Table 5.3, we provide the number of operations as a function of the input vector length $N$. From these numbers it is obvious, that the LC detector is by far the simplest one in terms of arithmetic operations, since it involves only summations and multiplications of floating point numbers. Only the watermarked coefficients and the watermark sequence itself are involved. However, the Cauchy-Rao detector is only slightly more expensive, since the exponentiations in Eq. (5.38) merely involve integer exponents. The remaining operations are just additions and multiplications which can be very efficiently performed with few CPU cycles. In contrast to that, the Cauchy-LRT detector requires $N$ computations of the logarithm and the GGD-LRT as well as the GGD-Rao detector even require exponentiations with floating point numbers, which is very expensive in terms of CPU cycles.

Regarding parameter estimation issues, the LC detector is again the simplest one, since it requires no parameter estimation at all, followed by the Cauchy-Rao and Cauchy-LRT detector, which both require to estimate the shape parameter $\gamma$ of the Cauchy distribution, see Section 2.2.2. In case of the detectors based on the GGD, we know that ML estimation of the shape $c$ and scale parameter $a$ requires to find the roots of a transcendental equation, see Section 2.2.1. Our estimation experiments confirm that the ML estimation of $\gamma$ is faster by a factor of four than ML estimation of the GGD shape parameter $c$. When relying on the estimation procedure suggested by Tsihrintzis et al. [176], see Eq. (2.6), estimation of $\gamma$ is even linear in $N$.

Finally, we cover the effort for the determination of detection thresholds. In case of the LC, GGD-LRT and Cauchy-LRT detector, we have to compute the mean and variance of the normally distributed detection statistic under $\mathcal{H}_0$ from the received signal $y_i$ to determine a suitable threshold. In contrast to that, the Cauchy-Rao and GGD-Rao detectors do not require to compute detection statistic parameters at all, since they are CFAR detectors. A detection threshold needs to be computed only once from Eq. (5.23).



**(a)** JPEG ($Q = 30$)

**(b)** JPEG ($Q = 70$)

**Figure 5.8:** C.d.f. plots of the probability of miss $P_m$ for a fixed probability of false-alarm $P_f = 10^{-3}$ over 1334 UCID images under JPEG compression with quality factors $Q = 30$ and $Q = 70$. The right-hand side shows a zoomed-in version of the region of interest, i.e. where $P_m > 10^{-3}$ and $P_m > 10^{-5}$, respectively.

| **Detector** | **Operations** | | | |
|---|---|---|---|---|
| | $\pm$ | $\times, \div$ | $\lvert\cdot\rvert, \text{sgn}$ | pow , log |
| Gaussian-LRT (LC), Eq. (5.42) | $N$ | $N + 1$ | | |
| Cauchy-Rao, Eq. (5.38) | $2N$ | $3N + 4$ | | |
| GGD-LRT [69], Eq. (5.43) | $3N$ | $N + 1$ | $2N$ | $2N + 1$ |
| Cauchy-LRT [12], Eq. (5.44) | $3N$ | $3N + 2$ | | $N$ |
| GGD-Rao [139], Eq. (5.45) | $2N + 1$ | $3N + 2$ | $2N$ | $N$ |

**Table 5.3:** Number of arithmetic operations to compute the detection statistic.

## 5.3 Color Image Watermarking

Most of the watermarking research focuses on grayscale images. The extension to color image watermarking is usually accomplished by marking only the luminance channel or by processing each color channel separately [6]. However, it is well known that the human visual system is least sensitive to the yellow-blue channel in the opponent representation of color, thus the

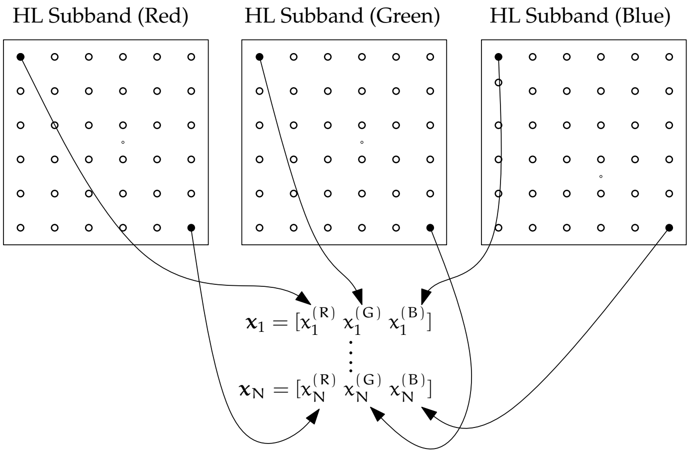

HL Subband (Red)
HL Subband (Green)
HL Subband (Blue)
$\mathbf{x}_1 = [x_1^{(R)}\ x_1^{(G)}\ x_1^{(B)}]$
$\mathbf{x}_N = [x_N^{(R)}\ x_N^{(G)}\ x_N^{(B)}]$

**Figure 5.9:** Extraction of DWT coefficient vectors $\mathbf{x}_i$ from three subbands (here HL) of different color channels.

watermark signal should be allocated to that band [158, 177]. In this section, we derive a novel watermark detector for color image watermarking. We propose to use a multivariate statistical model to capture the association structure between wavelet detail subbands across RGB color channels. Our objective is to show that watermark detection performance is improved, compared to decorrelating the color bands [4] or exploiting the correlation based on a Gaussian host signal model [6]. We highlight that we do not focus on perceptual shaping of the watermark signal but on detecting the watermark in highly correlated color channels where the watermark is embedded with constant strength.

### 5.3.1 A LRT detector for MPE host signal noise

We introduce an estimate-and-plug detector based on the LRT to detect an additively embedded watermark in host signal noise which follows a MPE distribution (see Section 2.2.3). For the following derivation of the detector, we will rely on the convention that $\mathbf{x}$ denotes a 3-dimensional vector of DWT coefficients, constructed by selecting one coefficient from the same detail subband of each color channel (illustrated in Fig. 5.9). We write $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to refer to the coefficient vectors. Our watermark will be a realization of N i.i.d. copies of a random variable $W$ following a discrete uniform distribution on $\{+1, -1\}$, see Eq. (5.1). The watermark sequence is denoted by $w_1, \ldots, w_N$. We follow the strategy to embed the same watermark in all three detail subbands. Given that $\mathbf{1} = [1\ 1\ 1]$ denotes a vector of ones, then the watermark vector to mark $\mathbf{x}_i$ can be written as $\boldsymbol{w}_i = w_i\mathbf{1}$. According to the rule of additive spread-spectrum watermarking, it follows that

$$\forall i : \mathbf{y}_i = \mathbf{x} + \alpha\mathbf{w}_i \tag{5.46}$$

where $\alpha > 0$ denotes the embedding strength and $\mathbf{y}_i$ denotes a watermarked DWT coefficient vector. We could choose a separate embedding strength for each signal dimension, but for the sake of readability we focus on the most simple case here. The embedding process is completed by computing the inverse DWT, followed by a quantization step to limit the pixel values to $[0, 255]$. Based on this watermarking setting, we can formulate the two hypothesis for our signal detection problem as

$$\mathcal{H}_0 : \mathbf{y} = \mathbf{x} \quad \text{(no/other watermark)}, \tag{5.47}$$

$$\mathcal{H}_1 : \mathbf{y} = \mathbf{x} + \alpha\mathbf{w} \quad \text{(watermarked)}. \tag{5.48}$$

Since we assume that $\alpha$ is known at the detection stage (i.e. the embedder has informed the detector about its choice of $\alpha$), we end up with the problem of detecting a known signal in incompletely specified noise. We proceed by constructing a NP detector as if all parameters were known for both $\mathcal{H}_0$ and $\mathcal{H}_1$ and see how far we can get. Assuming independence of the observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ allows to formulate a LRT which decides $\mathcal{H}_1$ in case

$$T(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \frac{\prod_{i=1}^{N} p(\mathbf{y}_i - \alpha\mathbf{w}_i)}{\prod_{i=1}^{N} p(\mathbf{y}_i)} > \gamma. \tag{5.49}$$

After taking the logarithm and inserting the p.d.f. of the MPE distribution, see Eq. (2.9), we obtain the test statistic

$$T(\mathbf{y}_1, \ldots, \mathbf{y}_N) = -\frac{1}{2}\sum_{i=1}^{N} \left((\mathbf{y}_i - \alpha\mathbf{w}_i)^T \mathbf{\Sigma}^{-1} (\mathbf{y}_i - \alpha\mathbf{w}_i)\right)^{\beta} + \frac{1}{2}\sum_{i=1}^{N} \left(\mathbf{y}_i^T \mathbf{\Sigma}^{-1} \mathbf{y}_i\right)^{\beta} \tag{5.50}$$

where we have used the fact that non signal-dependent terms are absorbed into the threshold $\gamma$. As we can see, the detection statistic depends on the host signal noise parameters $\mathbf{\Sigma}$ and $\beta$. In case of a GLRT approach, we would have to estimate both parameters under $\mathcal{H}_0$ and $\mathcal{H}_1$ which is analytically intractable. The estimate-and-plug detector, however, simply estimates the parameters from the received signal $\mathbf{y}_i$. If we consider all terms of the summation in Eq. (5.50) as independent, we can apply the central limit theorem and conclude that T follows a Normal distribution under $\mathcal{H}_0$ and $\mathcal{H}_1$ with parameters $(\mu_0, \sigma_0^2)$ and $(\mu_1, \sigma_1^2)$, respectively. Another difficulty arises, since we cannot compute the expected value of T w.r.t. $\mathbf{y}_i$ in closed-form. Alternatively, it is possible to consider $\mathbf{y}_i$ as fixed and average over the watermark signal $\boldsymbol{w}_i$. This strategy is also followed by Hernandez et al. [69] to derive the detection statistic parameters of the GGD-LRT. The expected value $\mu_0$ under $\mathcal{H}_0$ (note that $\mathbf{y}_i = \mathbf{x}_i$) then takes the form

$$\mu_0 = -\frac{1}{4}\sum_{i=1}^{N} \left((\mathbf{x}_i - \alpha)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \alpha)\right)^{\beta} + \left((\mathbf{x}_i + \alpha)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i + \alpha)\right)^{\beta} + \frac{1}{2}\sum_{i=1}^{N} \left(\mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{x}_i\right)^{\beta}. \tag{5.51}$$

To derive the variance $\sigma_0^2$ of T under $\mathcal{H}_0$, we exploit the following relation: given that X denotes a random variable and $k = \textit{const.}$, we know that $\mathbb{V}(\sum X) = \sum \mathbb{V}(X)$ and that $\mathbb{V}(X + k) = \mathbb{V}$. It follows that

$$\mathbb{V}(T|\mathcal{H}_0) = \mathbb{V}\left(-\frac{1}{2}\sum_{i=1}^{N} \left((\mathbf{x}_i - \alpha\mathbf{w}_i)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \alpha\mathbf{w}_i)\right)^{\beta}\right) \tag{5.52}$$

using $\mathbf{y}_i = \mathbf{x}_i$. We further know, that $\mathbb{V}(kX) = k^2\mathbb{V}(X)$ which leads to

$$\mathbb{V}(T|\mathcal{H}_0) = \frac{1}{4}\sum_{i=1}^{N} \mathbb{V}\left(\left((\mathbf{x}_i - \alpha\mathbf{w}_i)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \alpha\mathbf{w}_i)\right)^{\beta}\right). \tag{5.53}$$

To deduce the expression for $\mathbb{V}(T(\mathbf{y}_1, \ldots, \mathbf{y}_N))$, we remember that $\mathbf{w}_i$ is our variable term and that the elements of $\boldsymbol{w}_i$ follow a discrete uniform distribution on $\{+1, -1\}$. The variance of a random variable $W$ with a discrete uniform distribution is given by

$$\mathbb{V}(W) = \frac{1}{2}\left(\sum_{i=1}^{2} w_i^2 - \frac{1}{2}\left(\sum_{i=1}^{2} w_i\right)^2\right) = 1 \tag{5.54}$$

with $w_1 = -1$ and $w_2 = +1$. The variance of the detection statistic $T$ under $\mathcal{H}_0$ then follows as

$$\sigma_0^2 = \frac{1}{16}\sum_{i=1}^{N}\left(\left((\mathbf{x}_i + \alpha)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i + \alpha)\right)^{\beta} - \left((\mathbf{x}_i - \alpha)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \alpha)\right)^{\beta}\right)^2. \tag{5.55}$$

Given that $\hat{\beta}, \hat{\boldsymbol{\Sigma}}$ denote the estimates of the MPE parameters – computed from the received signal $\mathbf{y}_i$ – we can insert these estimates into Eqs. (5.51) and (5.55) to obtain $\hat{\mu}_0$ and $\hat{\sigma}_0^2$. Based on a chosen probability of false alarm $P_f$, it is then straightforward to set the detection threshold $\gamma$ as

$$\gamma = \text{erfc}^{-1}(2P_f)\sqrt{2\hat{\sigma}_0^2} + \hat{\mu}_0. \tag{5.56}$$

We consciously avoided the term Neyman-Pearson criterion to highlight that we cannot guarantee to constrain the probability of false-alarm due to the reliance on the host signal noise parameters (which are estimated). We can only say, that the threshold is selected in a Neyman-Pearson sense. We have to perform an empirical evaluation to ensure that we can constrain probability of false-alarm. Regarding the detection statistic parameters $(\mu_1, \sigma_1^2)$ under the alternative hypothesis $\mathcal{H}_1$, it can easily be shown that $\mu_1 = -\mu_0$ and $\sigma_1^2 = \sigma_0^2$.

### 5.3.2 Experiments

All following results are obtained on the whole UCID image database. Similar to the previous experiments of Section 5.2.1, all images are cropped to 256 × 256 pixel, followed by a downscaling stage to 128 × 128 pixel. The watermark is embedded in the HL subband on DWT decomposition level two. Biorthogonal CDF 9/7 filters are used for DWT decomposition. To compare the performance of the proposed detector against two state-of-the-art detectors for color image watermarking, we implement the approaches proposed by Barni et al. in [4] and [6]. These approaches are briefly described next, including the parameter configuration we use in the experiments.

**DCT-LRT** In [4], Barni et al. propose to embed a watermark sequence into the mid-frequency DCT coefficients obtained by computing a full-frame DCT on each color channel. In more detail, the $(k+1)$-th to $(k+n)$-th DCT coefficients are selected in MPEG zigzag-scan order for watermark embedding, as shown in Fig. 5.10b. At the detection stage, the classic LC detector is extended to the multichannel case. We point out, that the authors propose to use different embedding strengths for each channel, motivated by a study on how the Human Visual System (HVS) perceives color stimuli at different wavelengths. The detection statistic is given by

$$T_1(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \frac{1}{N}\sum_{i=1}^{N} w_i(y_{R,i} + y_{G,i} + y_{B,i}) \tag{5.57}$$

where $y_{R,i}$ denotes the $i$-th watermarked DCT coefficient of the red color channel and $w_i$ denotes the $i$-th element of the watermark sequence. The watermark is a realization of $N$

i.i.d. copies of a random variable $W \sim \mathcal{N}(0,1)$. The authors show, that the detection statistic under $\mathcal{H}_0$ follows a zero mean Gaussian distribution with variance given in [4]. This parameter is estimated from the received signal. To choose the embedding strengths of each channel, we fix the total strength $\alpha$ and use the relations $\alpha_R + \alpha_G + \alpha_B = \alpha$, $\alpha_R/\alpha_G = 1.37$ as well as $\alpha_B/\alpha_G = 3.24$ to solve for $\alpha_R, \alpha_G$ and $\alpha_B$.

**FFT-LRT** In [6], Barni et al. propose a different watermarking strategy based on decorrelation of the RGB color channels by means of the Karhunen-Loéve Transform (KLT). The decorrelated color channels are then transformed by the FFT. The basic idea is, that decorrelation allows to assume independency (at least in the Gaussian case) of the channels and leads to an analytically tractable joint statistical model for the magnitudes of the FFT coefficients. However, we point out that some caution is advisable here, since decorrelating the color channels does not guarantee that the transform domain coefficients across color bands are mutually decorrelated as well [107]. Basically, the approach is an extension of the work presented in [5] where the authors suggest a Weibull model for the magnitudes of FFT coefficients and derive a corresponding watermark detector based on the LRT. The watermark sequence is embedded in a diamond shaped region of the FFT domain (see Fig. 5.10a), defined by the $(k+1)$-th to $(k+n)$-th diagonal of the first FFT quadrant. Separate embedding strengths per channel are proposed to take into account that the KLT leads to decorrelated channels with decreasing variance. Lets assume for a moment that $\gamma$ denotes the embedding strength and $A, B, C$ denote the decorrelated color bands, then the detection statistic is given as

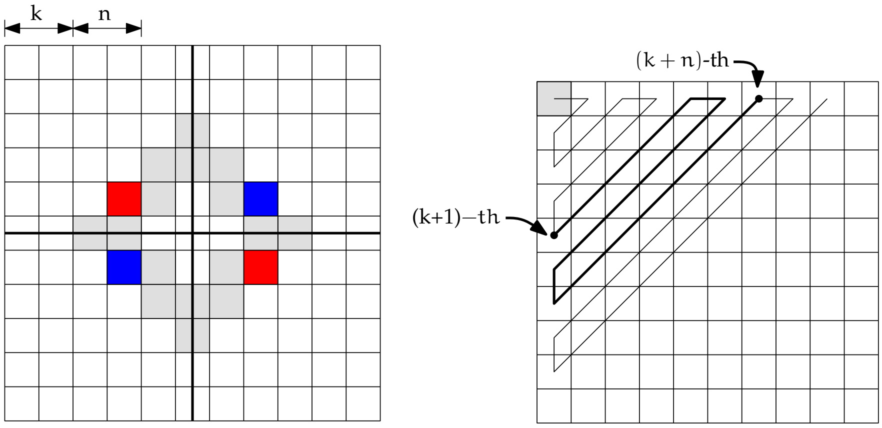
$$T_2(y_1, \ldots, y_N) = \sum_{c \in \{A,B,C\}} \sum_{i=1}^{N} \frac{y_{c,i}^{\alpha_c}[(1+\gamma_c w_i)^{\alpha_c} - 1]}{[\beta_c(1+\gamma_c w_i)]^{\alpha_c}} \tag{5.58}$$

where $\alpha_c, \beta_c$ are the Weibull parameters estimated from the received FFT coefficient magnitudes of the c-th decorrelated color band.

For all following results, the embedding strength of each approach is chosen such that we obtain a mean PSNR of $\approx$ 50dB across the three RGB channels of an image. Further, we set $k = n = 8000$ in case of [4] and $k = 30$, $n = 60$ in case of [6]. This gives $\approx$ 8000 marked coefficients for both the DCT and FFT approach (due to the symmetry of the FFT). For the proposed MPE-LRT detector, we choose the DWT HL subband on decomposition level two, resulting in $\approx$ 4000 marked coefficients in each channel.

Before we present the comparative study of the detection performance, we have to verify two important assumptions in order to ensure reasonable threshold selection. First, we verify that the detector responses under both hypotheses follow a Gaussian law for all three detectors by employing a Lilliefors test [108] at the 5% significance level. We report, that in no case the test shows evidence against the null-hypothesis. Second, we have to ensure, that the detection statistic parameters $\tilde{\mu}_0$ and $\tilde{\sigma}_0^2$ can be determined on the basis of the received signal. For this purpose, we conduct a Monte-Carlo study with $M = 1000$ runs to obtain Table 5.4[1]. The parameters $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ again denote the detection statistic parameters estimated from the experimental responses under $\mathcal{H}_0$, i.e. $\rho_1, \ldots, \rho_M$ (i.e. by sample mean and variance). Further, the column *FP* lists the number of actually observed false positives. As we can see, the detection statistic parameters of the MPE-LRT and DCT-LRT detector can be fairly well estimated from the received

[1] The remaining GoF tests are performed using a Chi-Square GoF test.



**(a)** FFT domain embedding [4]

**(b)** DCT domain embedding [6]

**Figure 5.10:** Watermark embedding location of the approaches [4] and [6]. In both works, the watermark is embedded in mid-frequency coefficients of either the FFT [4] or DCT [6]. The symmetry of the FFT quadrants is indicated by marking coefficients with equal values in red and blue.

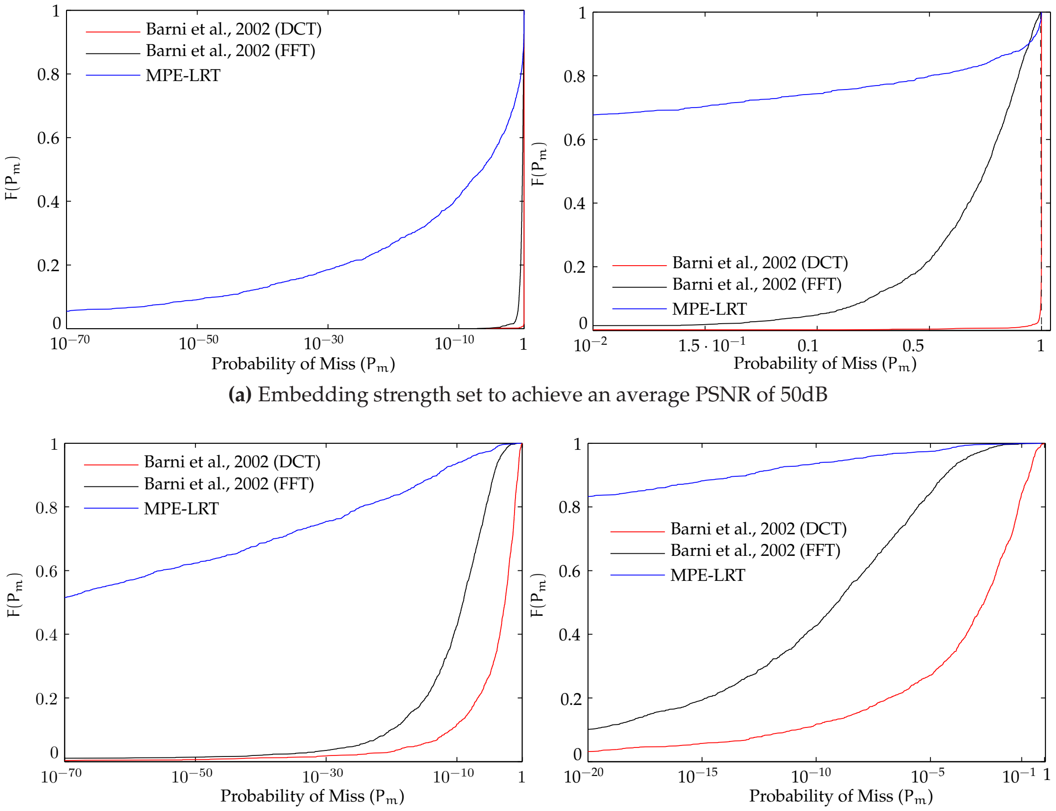
| **Detector** | $\hat{\mu}_0 \leqslant \tilde{\mu}_0$ | $\hat{\sigma}_0^2 \leqslant \tilde{\sigma}_0^2$ | $\rho \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ | $\mathcal{N}(\hat{\mu}_0, \tilde{\sigma}_0^2)$ | $\rho \sim \mathcal{N}(\tilde{\mu}_0, \hat{\sigma}_0^2)$ | FP |
|---|---|---|---|---|---|---|
| MPE-LRT | 52.17 | 49.25 | 71.60 | 100.0 | 71.75 | $2.5 \cdot 10^{-3}$ |
| DCT-LRT [4] | 52.09 | 48.61 | 99.93 | 100.0 | 99.93 | $1.0 \cdot 10^{-3}$ |
| FFT-LRT [6] | 48.28 | 0.00 | 0.009 | 0.005 | 56.50 | $8.8 \cdot 10^{-2}$ |

**Table 5.4:** Evaluation of the detection statistic distribution under $\mathcal{H}_0$ conforms to the expected distribution computed on the basis of the received signal. Then numbers represent the percentage of UCID images where the test (given as column title) does not fail. The probability of false-alarm is set to $P_f = 10^{-3}$.

signal. The percentage of observed false positives is in accordance with the fixed $P_f$ value of $10^{-3}$. In case of the FFT-LRT detector, however, the actually observed variance is larger than expected, resulting in a slightly higher number of false positives. In Fig. 5.11a, we show c.d.f. plots of the probability of miss over the whole UCID image database with a fixed $P_f$ of $10^{-3}$. Compared to the DCT-LRT and FFT-LRT, the MPE-LRT shows superior performance, especially in the critical region where $P_m$ is high. However, we note that there is a considerable number of images where all detectors fail to detect the watermark. Due to the relatively low resolution of the images (128 × 128) and the low embedding power to reach a mean PSNR of 50dB, this result is not unexpected, though. Fig. 5.11b shows the same plots, but choosing the DWR such that we achieve a mean PSNR of 40dB over the color channels of each image. We can see that the detectors of Barni et al. [4, 6] perform considerably better, but still the LRT-MPE shows the best performance, even for high values of $P_m$.

## 5.4 Discussion

In this chapter, we introduced two novel detectors for additive spread-spectrum watermarking in the DWT domain. After a careful recapitulation of the prerequisites to deploy certain signal



**(a)** Embedding strength set to achieve an average PSNR of 50dB

**(b)** Embedding strength set to achieve an average PSNR of 40dB

**Figure 5.11:** C.d.f. plots of the probability of miss $P_m$ for a fixed probability of false-alarm $P_f = 10^{-3}$ over 1338 UCID images including a zoomed-in version on the ROI. The embedding strength was set to achieve a mean PSNR (over the color channels) of 50dB (top) and 40dB (bottom).

detection strategies, we motivated the Rao hypothesis test as a lightweight alternative which requires very little knowledge about unknown parameters. We then derived a Rao hypothesis test conditioned on a Cauchy host signal noise model and showed that the detector exhibits quite good performance compared to the state-of-the art detectors in this field. The Cauchy-Rao detector is also attractive from a computational point of view, since computation of the detection statistic is comparable to the LC detector, estimation of the Cauchy shape parameter is less expensive than estimation of the GGD shape parameter and the computational demand for threshold calculation vanishes at all.

In the second part of this chapter, we focused on the problem of color image watermarking. By relying on a multivariate model for DWT detail subband coefficients, we could derive a novel estimate-and-plug detector based on the LRT. A comparative study to two state-of-the art detectors revealed quite competitive performance of the novel detector on the whole UCID image database. The results of the MPE-LRT detector show that the association between the DWT coefficients can be efficiently exploited to enhance detection performance. Nevertheless, estimation of the MPE parameters is a computationally expensive operation which prevents deployment of the MPE-LRT detector in computationally demanding scenarios.

In the experimental sections of this chapter, we have further introduced a novel visual tool to evaluate detector performance. Motivated by the shortcomings of the classic ROC plots – which only allow to visualize detector performance on one image – we suggested a c.d.f. plot of the probability of miss at a fixed probability of false-alarm. We strongly believe that this is a suitable way to study detector performance over a large set of images. To the best of our knowledge, such a plot has not appeared in literature so far. Based on our experimental results, we come to a conclusion similar to Chapter 3. In general, it is not advisable to thoroughly rely on ROC curves to judge the quality of a detector. We rather suggest to fix the probability of false-alarm, estimate the probability of miss and focus attention on the difficult cases.

Finally, we like to point out that a lot of questions remain unanswered and are topic of future research. It seems promising to take a closer look at noise parameter estimation for example. In consideration of the variety of possible attacks, the question arises whether it is possible to use fixed parameter settings instead of ML estimation to stabilize detector performance. This might negatively affect performance in case of no attacks, but could be beneficial in situations where the attack distorts the coefficient statistics. Hernandez et al. [69] already suggested a fixed setting of $c = 0.8$ for example. In addition to that, fixing the host noise parameters would also contribute to the idea of lightweight detection and allow application of a detector in scenarios where real-time performance is required, e.g. real-time detection of watermarks in video frames. Finally, the two novel detectors have to be evaluated under the influence of common attacks. Since we strongly focused on the theoretical signal detection part, we omitted the attack evaluation here and consider that as a topic for future work.

# Chapter 6

# Concluding Remarks

In this last part of the thesis, we recapitulate the main contributions and highlight future research directions. A general conclusion we draw from our studies is that there is still very much potential in developing novel statistical models for wavelet transform coefficients. In the context of this thesis, we could at least show that the models of Chapter 2 led to improvements upon state-of-the-art work in texture image retrieval, medical image classification and watermarking research.

In particular, we showed that the proposed models for DTCWT coefficient magnitudes led to a very lightweight probabilistic texture retrieval approach with remarkable retrieval performance. Incorporating coefficient dependencies across DTCWT subbands even further improved the retrieval results. However, the improvements in retrieval accuracy came at the cost of degraded runtime which highlights the trade-off between model complexity and computational performance. Surprisingly, the same statistical models turned out to be equally useful for medical image classification. We introduced a set of novel image features by refining existing ideas from texture classification literature and demonstrated a high accuracy in predicting histological diagnostic results from the visual appearance of colorectal lesions. Eventually, we pointed out the versatility of the statistical models by deriving two novel watermark detectors for luminance channel and color image watermarking. The detection experiments on a large set of images revealed competitive or even superior detection performance to current state-of-the-art detectors.

We summarize, that the particular field of application will eventually determine which statistical model to use. In situations where computation time is a crucial factor for example, the most suitable model is useless in case the key processing steps become too complicated, either analytically or computationally. In the context of watermarking for instance, a too complex model might substantially complicate the derivation of the detection statistic or even prevent to deduce a closed-form expression. In addition, parameter estimation issues might arise as well. A similar situation occurs in the context of probabilistic image retrieval, as we have pointed out by introducing two scenarios with different computational requirements. For a large number of applications, the following rule of thumb remains valid: the more information we incorporate into a statistical model, the higher the price we pay in terms of runtime performance. There is usually no such thing as a win-win situation in this context.

## 6.1 Future Research Directions

Finally, we remark that there are obviously many interesting topics we could not cover, or even address, in this thesis and which remain part of future research. In consideration of the three fields of application we discussed in this work, we like to highlight possible directions for future studies:

- In the context of texture image retrieval, we see great potential in a copula-based approach which incorporates the Generalized Gamma distribution [174] as a model for the margins. Since Choy & Tong [21] recently demonstrated better texture retrieval performance than the GGD based retrieval approach of Do & Vetterli [40], there is good reason to believe that a copula-based model would perform even better. Nevertheless, computational considerations will definitely play a key role for any practical application, since estimation of the Generalized Gamma model is computationally quite involved (e.g., see [172]).
- Concerning our particular watermarking setup of Chapter 5, the issue of how to combine detection responses from different detection processes is a neglected topic in literature. The problem occurs, when we embed watermarks in more than just one DWT subband and then try to combine the detection statistics into an overall detector response. Although a sum of i.i.d. Normal or Chi-Square random variables still follows a Normal or Chi-Square distribution, application of the additivity property is only reasonable in case of i.i.d. detection statistics. Since coefficients exhibit dependencies across subbands and the detection statistic depends on the coefficients, this prerequisite is obviously violated. Consequently, this raises the question of how to constrain the probability of false-alarm. In fact, we presume that this problem calls for a flexible multivariate coefficient model, since this would remedy the fusion problem by shifting complexity to the modeling and detector derivation stage.
- From our point of view, dealing with the fusion problem is a key issue for any further development of the computer-aided diagnosis system of Chapter 4 as well. Up to now, the majority of research work has focused on improving classification rates of standalone approaches. The only steps in the direction of combining prediction results were made in [62] or [61] with promising preliminary results. However, problems like overtraining issues or generalization quality still remain untreated. Finally, we suggest to further pursue the generative model based prediction strategy because of the advantages with respect to the aforementioned problems.

Although we have good reason to believe that other application areas, such as denoising or segmentation will benefit from the proposed statistical models in a variety of ways, this remains to be shown in future research work.

# References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[3] M. Barni and F. Bartolini. *Watermarking Systems Engineering*. Marcel Dekker, 2004.

[4] M. Barni, F. Bartolini, and A. Piva. Multichannel watermarking of color images. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(3):142–156, March 2002.

[5] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. A new decoder for the optimum recovery of non-additive watermarks. *IEEE Transactions on Image Processing*, 10(5):1–11, May 2001.

[6] M. Barni, F. Bartolini, A. De Rosa, and A. Piva. Color image watermarking in the Karhunen-Loeve transform domain. *Journal of Electronic Imaging*, 11(1):87–95, January 2002.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society – Series B*, 57(1):289–300, 1995.

[8] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[9] D. Berg. Copula Goodness-of-Fit testing: an overview and power comparison. *The European Journal of Finance*, 15(7):675–701, October 2009.

[10] K. A. Birney and T. R. Fischer. On the modeling of DCT and subband image data for compression. *IEEE Transactions on Image Processing*, 4(2):186–193, February 1995.

[11] E. Bouyé, V. Durrelman, A. Nikeghbali, G. Riboulet, and T. Roncalli. Copulas for finance – a reading guide and some applications. Working Paper, March 2000.

[12] A. Briassouli, P. Tsakalides, and A. Stouraitis. Hidden messages in heavy-tails: DCT-domain watermark detection using alpha-stable models. *IEEE Transactions on Multimedia*, 7(4):700–715, August 2005.

[13] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover Publications, New York, 1966. Pictures downloaded from *http://www.ux.his.no/~tranden/brodatz.html* (Trygve Randen).

[14] Robert W. Buccigrossi and Eero P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, December 1999.

[15] R. Chandramouli and N. D. Memon. On sequential watermark detection. *IEEE Transactions on Signal Processing*, 51(4):1034–1044, April 2003.

[16] S. Chang, B. Yu, and Martin Vetterli. Spatially adaptive wavelet thresholding with content modeling for image denoising. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'98)*, Chicago, IL, USA, October 1998.

[17] Q. Cheng and T. S. Huang. An additive approach to transform-domain information hiding and optimum detection structure. *IEEE Transactions on Multimedia*, 3(3):273–284, September 2001.

[18] H. Choi and R. Baraniuk. Multiscale texture segmentation using wavelet-domain hidden markov models. In *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, pages 1692–1697, Pacific Grove, CA, United States, 1998.

[19] H. Choi, J.K. Romberg, R.G. Baraniuk, and N. G. Kingsbury. Hidden markov tree modeling of complex wavelet transforms. In *Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000)*, Istanbul, Turkey, June 2000.

[20] S. C. Choi and R. Wette. Maximum likelihood estimation of the parameters of the Gamma distributon and their bias. *Technometrics*, 11(4):683–690, November 1968.

[21] S. K. Choy and C. S. Tong. Statistical wavelet subband characterization based on generalized gamma density and its application in texture retrieval. *IEEE Transactions on Image Processing*, 19(2):281–289, February 2010.

[22] R. J. Clarke. *Transform Coding of Images*. Academic Press, 1985.

[23] A. C. Cohen and B. J. Whitten. *Parameter estimation in reliability and life space models*. Marcel–Dekker, 1988.

[24] D. Comaniciu, P. Meer, P. Kun Xu, and D. Tyler. Retrieval performance improvement through low rank corrections. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'99)*, pages 50–54, Fort Collins, CO, USA, 1999.

[25] M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, May 1983.

[26] T. C. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[27] D. R. Cox and N. T. H. Small. Testing multivariate normality. *Biometrika*, 65(2), August 1978.

[28] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann, 2007.

[29] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based signal processing using Hidden Markov models. *IEEE Transactions on Signal Processing, Special Issue on Wavelets and Filterbanks*, 46(2):886–902, April 1998.

[30] K. J. Dana, B. V. Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real world surfaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 151–157, San Juan, Puerto Rico, 1997.

[31] MIT Vistion Texture Database. MIT vision and modeling group. [Online]. Available from: `http://vismod.media.mit.edu/vismod/`.

[32] I. Daubechies. *Ten Lectures on Wavelets*. Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM Press, Philadelphia, PA, USA, 1992.

[33] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, 2(7):1160–1169, July 1985.

[34] E. de Ves, A. Ruedin, D. Acevedo, C. Benavent, and L. Seijas. A new wavelet-based texture descriptor for image retrieval. *Lecture Notes in Computer Science, Computer Analysis of Images and Patterns*, 4673:895–902, August 2007.

[35] G. Van de Wouwer, S. Livens, P. Scheunders, and D. Van Dyck. Color Texture Classification by Wavelet Energy Correlation Signatures. In *Proceedings of the 9th International Conference on Image Analysis and Processing (ICIAP'97)*, pages 327–334, Florence, Italy, 1997. Springer.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society – Series B*, 39(1):1–38, 1977.

[37] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, October 1998.

[38] M. Do. Fast approximation of Kullback-Leibler distance for dependence trees and Hidden Markov models. *IEEE Signal Processing Letters*, 10(4):115–118, April 2003.

[39] M. Do and M. Vetterli. Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models. *IEEE Transactions on Multimedia*, 4(4):517–527, December 2002.

[40] M. Do and M. Vetterli. Wavelet-based texture retrieval using Generalized Gaussian density and Kullback-Leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158, February 2002.

[41] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley & Sons, 2nd edition, November 2000.

[42] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[43] P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. In S. Rachev, editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384. Elsevier, 2003.

[44] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colon mucosa. *IEEE Transactions on Information Technology in Biomedicine*, 2(3):197–203, September 1998.

[45] M. Evans and N. Hastings B. Peacock. *Statistical Distributions*. Wiley Series in Probability and Statistics. Wiley, 3rd edition, 2000.

[46] B. Everitt. *The Analysis of Contingency Tables*. Chapman and Hall, 1977.

[47] B. Everitt. *An R and S–Plus Companion to Multivariate Analysis*. Springer, 2005.

[48] N. Fisher and P. Switzer. Graphical assessment of dependence: Is a picture worth 100 tests? *The American Statistican*, 55(3):233–239, August 2001.

[49] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Annals of Statistics*, 7(4):697–717, 1979.

[50] K.-I. Fu, Y. Sano, S. Kato, T. Fuji, F. Nagashima, T. Yoshino, T. Okuno, S. Yoshida, and T. Fujimori. Chromoendoscopy using indigo carmine dye spraying with magnifying observation is the most reliable method for differential diagnosis between non-neoplastic and neoplastic colorectal lesions: a prospective study. *Endoscopy*, 36(12):1089–1093, December 2004.

[51] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Morgan Kaufmann, 2nd edition, 1990.

[52] C. Genest and A. C. Favre. Everything you always wanted to know about Copula modeling and were afraid to ask. *Journal of Hydrological Engineering*, 12(4):347–368, July 2007.

[53] C. Genest and B. Rémillard. Validity of the parametric bootstrap for Goodness-of-Fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré*, 44(6):1096–1127, 2008.

[54] C. Genest, B. Rémillard, and D. Beaudoin. Goodness–of–fit tests for copulas: A review and a power study. *Mathematics and Economics*, 44:199–213, 2009.

[55] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, June 1992.

[56] A. Genz and F. Bretz. Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4):950–971, December 2002.

[57] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of the KL-divergence between two Gaussian mixtures. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'03)*, pages 487–493, Nice, France, 2003.

[58] E. Gomez, M-A. Gomez-Viilegas, and J. M. Marin. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics – Theory and Methods*, 27(3):589–600, 1998.

[59] R. M. Gray. Vector quantization. *IEEE Transactions on Acoustic Signal and Speech Processing*, 1:4–29, April 1984.

[60] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A.J. Smola. A kernel statistical test of independence. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'07)*, pages 585–592, Vancouver, Canada, 2007.

[61] M. Häfner, A. Gangl, R. Kwitt, A. Uhl, A. Vecsei, and F. Wrba. Improving pit-pattern classification of endoscopy images by a combination of experts. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'09)*, pages 247–254, London, UK, 2009.

[62] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba. Pit pattern classification using multichannel features and multiclassification. In D.I. Fotiadis T.P. Exarchos, A. Papadopoulos, editor, *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*, pages 335–350. IGI Global, Hershey, PA, USA, 2009.

[63] M. Häfner, C. Kendlbacher, W. Mann, W. Taferl, F. Wrba, A. Gangl, A. Vécsei, and A. Uhl. Pit pattern classification of zoom-endoscopic colon images using histogram techniques. In Johannes R. Sveinsson, editor, *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG 2006)*, pages 58–61, Reykavik, Iceland, June 2006. IEEE.

[64] M. Häfner, R. Kwitt, A. Uhl, A. Gangl, F. Wrba, and A. Vecsei. Feature-extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images. *Pattern Analysis and Applications*, 12(4):407–413, December 2009.

[65] R. M. Haralick, Dinstein, and K. Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:610–621, November 1973.

[66] N. A. Heckert and J. J. Filliben. *NIST Handbook 148: DATAPLOT Reference Manual*, volume 1. National Institute of Standards and Technology Handbook Series, 2003.

[67] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'95)*, pages 229–238, Los Angeles, USA, 1995. ACM.

[68] N. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Annals of Statistics*, 16(2):772–783, 1988.

[69] J. R. Hernández, M. Amado, and F. Pérez-González. DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure. *IEEE Transactions on Image Processing*, 9(1):55–68, January 2000.

[70] G. W. Hill. Algorithm 396: Student's t-quantiles. *Communications of the ACM*, 13(10):619–620, 1970.

[71] J. Huang and D. Mumford. Statistics of natural images and models. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99)*, pages 1541–1547, Fort Collins, Colorado, United States, 1999.

[72] D. P. Hurlstone. High-resolution magnification chromoendoscopy: Common problems encountered in "pit pattern" interpretation and correct classification of flat colorectal lesions. *American Journal of Gastroenterology*, 97:1069–1070, 2002.

[73] D. P. Hurlstone, S. S. Cross, I. Adam, A. J. Shorthouse, S. Brown, D. S. Sanders, and A. J. Lobo. Efficacy of high magnification chromoscopic colonoscopy for the diagnosis of neoplasia in flat and depressed lesions of the colorectum: a prospective analysis. *Gut*, 53(2):284–290, February 2004.

[74] D. K. Iakovidis, D. E. Maroulis, and S. A. Karkanis. An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy. *Computers in Biology and Medicine*, 36(10):1084–1103, October 2006.

[75] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos. A comparative study of texture features for the discrimination of gastric polyps in endoscopic video. In *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, 2005 (CBMS'05)*, pages 575–580, Dublin, Ireland, June 2005.

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

[77] H. Joe. *Multivariate Models and Dependence Concepts*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1997.

[78] S. A. Karkanis, D. Iakovidis, D. Karras, and D. Maroulis. Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures. In *Proceedings of the IEEE International Conference in Image Processing, 2001 (ICIP'01)*, pages 833–836, Thessaloniki, Greece, October 2001.

[79] S.A. Karkanis. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Transactions on Information Technology in Biomedicine*, 7(3):141–152, September 2003.

[80] S. Kato, K.-I. Fu, Y. Sano, T. Fujii, Y. Saito, T. Matsuda, I. Koba, S. Yoshida, and T. Fujimori. Magnifying colonoscopy as a non-biopsy technique for differential diagnosis of non-neoplastic and neoplastic lesions. *World Journal of Gastroenterology: WJG*, 12(9):1416–1420, March 2006.

[81] S. Kato, T. Fujii, I. Koba, Y. Sano, K. Fu, A. Parra-Blanco, H. Tajiri, S. Yoshida, and B. Rembacken. Assessment of colorectal lesions using magnifiying colonoscopy and mucosal dye spraying: Can significant lesions be distinguished? *Endoscopy*, 33:306–310, April 2001.

[82] S. M. Kay. Asymptotically optimal detection in incompletely characterized non-gaussian noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(5):627–633, May 1989.

[83] S. M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*, volume 2. Prentice-Hall, 1998.

[84] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, June 1938.

[85] N. G. Kingsbury. The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters. In *Proceedings of the IEEE Digital Signal Processing Workshop, DSP '98*, pages 9–12, Bryce Canyon, USA, August 1998.

[86] N. G. Kingsbury. Image processing with complex wavelets. *Phil. Trans. Royal Society London A, a Discussion Meeting on "Wavelets: the key to intermittent information?"*, September 1999.

[87] N. G. Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis*, 10(3):234–253, May 2001.

[88] Kazuo Konishi, Kazuhiro Kaneko, Toshinori Kurahashi, Taikan Yamamoto, Miki Kushima, Akira Kanda, Hisao Tajiri, and Keiji Mitamura. A comparison of magnifying and nonmagnifying colonoscopy for diagnosis of colorectal polyps: a prospective. *Gastrointestinal Endoscopy*, 57:48–53, 2003.

[89] K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. Chapman & Hall, 2006.

[90] S. M. Krishnan, X. Yang, K. L.Chan, S. Kumar, and P. M. Y. Goh. Intestinal abnormality detection from endoscopic images. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1998 (EMBS'98)*, Hong Kong, China, October 1998.

[91] R. Krupinski and J. Purczynski. Approximated fast estimator for the shape parameter of Generalized Gaussian distribution. *Signal Processing*, 86(2):205–211, February 2006.

[92] S. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu. Colorectal tumours and pit pattern. *Journal of Clinical Pathology*, 47:880–885, 1994.

[93] E. Kuruoglu and J. Zerubia. Modeling SAR images with a generalization of the Rayleigh distribution. *IEEE Transactions on Image Processing*, 13(4):527–, April 2004.

[94] R. Kwitt, P. Meerwald, and A. Uhl. A lightweight Rao-Cauchy detector for additive watermarking in the DWT-domain. In *Proceedings of the ACM Multimedia and Security Workshop (MMSEC '08)*, pages 33–41, Oxford, UK, September 2008. ACM.

[95] R. Kwitt, P. Meerwald, and A. Uhl. Blind DT-CWT domain additive spread-spectrum watermark detection. In *Proceedings of the 16th International Conference on Digital Signal Processing, DSP '09*, Santorini, Greece, July 2009.

[96] R. Kwitt, P. Meerwald, and A. Uhl. Color-image watermarking using multivariate power-exponential distribution. In *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pages 4245–4248, Cairo, Egypt, November 2009. IEEE.

[97] R. Kwitt, P. Meerwald, and A. Uhl. Efficient detection of additive watermarking in the DWT-domain. In *Proceedings of the 17th European Signal Processing Conference, EUSIPCO '09*, pages 2072–2076, Glasgow, UK, August 2009. EURASIP.

[98] R. Kwitt and A. Uhl. Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images. In *Proceedings of the IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'07)*, pages 1–8, Rio de Janeiro, Brasil, 2007.

[99] R. Kwitt and A. Uhl. Color eigen-subband features for endoscopy image classification. In *Proceedings of the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 589–592, Las Vegas, Nevada, United States, 2008.

[100] R. Kwitt and A. Uhl. Color wavelet cross co-occurrence matrices for endoscopy image classification. In *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP'08)*, pages 715–718, St. Julians, Malta, 2008.

[101] R. Kwitt and A. Uhl. Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'08)*, pages 933–936, San Diego, California, United States, October 2008.

[102] R. Kwitt and A. Uhl. *Multi–Directional Multi-Resolution Transforms for Zoom–Endoscopy Image Classification (Best Paper Award at CORES 2007)*, volume 45 of *Advances in Soft Computing*, pages 35–43. Springer, 2008.

[103] R. Kwitt and A. Uhl. A joint model of complex wavelet coefficients for texture retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP '09)*, pages 1877–1880, Cairo, Egypt, November 2009.

[104] R. Kwitt and A. Uhl. Lightweight probabilistic image retrieval. *IEEE Transactions on Image Processing*, 19(1):241–253, January 2010.

[105] T. S. Lee. Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, October 1996.

[106] J. Li, R. M. Gray, and R. A. Olshen. Multiresolution image classification by hierarchical modeling with two-dimensional Hidden Markov models. *IEEE Transactions on Information Theory*, pages 1826–1841, August 2000.

[107] Y. Liang, E. Simoncelli, and Z. Lei. Color channels decorrelation by ICA transformation in the wavelet domain for color texture analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVRP '00*, volume 1, pages 606–611. IEEE, June 2000.

[108] H. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, June 1967.

[109] J. Liu and P. Moulin. Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. *IEEE Transactions on Image Processing*, 10(11):1647–1658, November 2001.

[110] S. LoPresto, K. Ramchandran, and T. Orchard. Image coding based on mixture modeling of wavelet coefficients and a fast estimation–quantization framework. In *Proceedings of the Data Compression Conference (DCC'97)*, pages 221–230, Snowbird, Utah, USA, 1997.

[111] W.-Y. Ma and H.J. Zhang. Benchmarking of image features for content-based retrieval. In *Proceedings of the Asilomar Conference on Signals, Systems & Computers*, pages 253–257, Pacific Grove, California, United States, 1998.

[112] T. Mäenpää. *The Local Binary Pattern Approach to Texture Analysis - Extensions and Applications*. PhD thesis, University of Oulu, 2003.

[113] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[114] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1997.

[115] S. Mallat. *A Wavelet Tour Of Signal Processing*. Academic Press, 2nd edition, 1999.

[116] H. S. Malvar and D. A. F. Florencio. Improved spread spectrum: A new modulation technique for robust watermarking. *IEEE Transactions on Signal Processing*, 51(4):898–905, April 2003.

[117] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, August 1996.

[118] N. R. Mann and K. W. Fertig. *Methods for Statistical Analysis of Reliability and Life Data*. Wiley, 1974.

[119] J. Mao and A. K. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25(2):173–188, February 1992.

[120] K. V. Mardia. Measures of multivariate Kurtosis and Skewness. *Biometrika*, 57:519–530, 1970.

[121] K. V. Mardia and P. Rupp. *Directional Statistics*, volume 2nd. John Wiley and Sons Ltd., 2000.

[122] D. E. Maroulis, D. K. Iakovidis, S. A. Karkanis, and D. A. Karras. CoLD: a versatile detection system for colorectal lesions in endoscopy video-frames. *Computer Methods and Programs in Biomedicine*, 70(2):151–66, February 2003.

[123] J. R. Mathiassen, A. Skavhaug, and K. Bo. Texture similarity measure using Kullback-Leibler divergence between Gamma distributions. In *Proceedings of the European Conference on Computer Vision (ECCV'02)*, pages 133–147, Copenhagen, Denmark, 2002.

[124] A. Meining, T. Rösch, R. Kiesslich, M. Muders, F. Sax, and W. Heldwein. Inter- and intraobserver variability of magnification chromoendoscopy for detecting specialized intestinal metaplasis at the gastroesophageal junction. *Endoscopy*, 36(2):160–164, February 2004.

[125] N. Merhav and E. Sabbag. Optimal watermark embedding and detection strategies under limited detection resources. *IEEE Transactions on Information Theory*, 54(1):255–274, January 2008.

[126] C. D. Meyer. *Matrix Calculus and Applied Linear Algebra*. Society for Applied and Industrial Mathematics, 2000.

[127] M. Mihcak, I. Kozintsev, and K. Ramchandran. Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99*, pages 3253–3256, Phoenix, AZ, USA, March 2009. IEEE.

[128] M. A. Miller and N. G. Kingsburg. Statistical image modelling using interscale phase relationships of complex wavelet coefficient. In *Proceedings of the IEEE International Conference on Accoustics, Speech and Signal Processing (ICASSP'06)*, pages 789–792, Toulouse, France, 2006.

[129] M. A. Miller and N. G. Kingsburg. Image modeling using interscale phase properties of complex wavelet coefficients. *IEEE Transactions on Image Processing*, 17(9):1491 – 1499, September 2008.

[130] G. Moser, J. Zerubia, and S. Seprico. SAR amplitude probability density function estimation based on Generalized Gaussian model. *IEEE Transactions on Image Processing*, 15(6):1429–1442, June 2006.

[131] P. Switzer N. Fisher. Chi–plots for assessing dependence. *Biometrika*, 72:253–265, August 1985.

[132] S. Nadarajah. The Kotz–type distribution with applications. *Statistics*, 37(4):341–358, July 2003.

[133] S. Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32:685–694, September 2005.

[134] S. Nadarajah and S. Kotz. On the generation of gaussian noise. *IEEE Transactions on Signal Processing*, 55(3):1172, March 2007.

[135] B. Mac Namee, P. Cunningham, S. Byrne, and O.I. Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, 24(1):51–70, January 2001.

[136] G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. *Lecture Notes in Statistics*, 103:281–300, 1995.

[137] R. B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, second edition, 2006.

[138] C. L. Nikias and M. Shao. *Signal Prcocessing with Alpha–Stable Distributions and Applications*. Wiley–Interscience, 1995.

[139] A. Nikolaidis and I. Pitas. Asymptotically optimal detection for additive watermarking in the DCT and DWT domains. *IEEE Transactions on Image Processing*, 12(5):563–571, May 2003.

[140] A. K. Nikoloulopoulos and D. Karlis. Copula model evaluation based on parametric bootstrap. *Computational Statistics and Data Analysis*, 52:3342–3353, March 2007.

[141] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, January 1996.

[142] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution Gray-Scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, July 2002.

[143] C. Palm. Color texture classification by integrative co–occurrence matrices. *Pattern Recognition*, 37(5):965–976, May 2004.

[144] C. Palm, T. M. Lehmann, and K. Spitzer. Color texture analysis of moving vocal cords using approaches from statistics and signal theory. In *Proceedings of the 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research*, pages 49–56, 2000.

[145] C. Palm, V. Metzler, B. Mohan O. Dieker, T. M. Lehmann, and K. Spitzer. *Bildverarbeitung für die Medizin*, chapter Co–Occurrence Matrizen zur Texturklassifikation in Vektorbildern, pages 367–371. Springer, 1999.

[146] V. Panchenko. Goodness–of–fit tests for copulas. *Physica A*, 355(1):1–232, September 2005.

[147] J. C. Pesquet, H. Krim, and H. Carfantan. Time invariant orthonormal wavelet representations. *IEEE Transactions on Signal Processing*, 44(8):1964–1970, August 1996.

[148] Maria Petrou and Pedro Garcia Sevilla. *Image Processing. Texture: Dealing with Texture*. Wiley John and Sons, 1st edition, 2006.

[149] R. Picard, T. Kabir, and F. Liu. Real-time recognition wih the entire Brodatz texture database. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'93)*, pages 638–639, New York, United States, 1993.

[150] J. Portilla and E. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, October 2000.

[151] S. M. Mahbubur Rahman, M. Omair Ahmad, and M. N. S. Swamy. Statistics of 2-d dt-cwt coefficients for a gaussian distributed signal. *IEEE Transactions on Circuits and Systems*, 55(7):2013–2025, August 2008.

[152] T. Randen and J.H. Husoy. Filtering for texture classification: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310, April 1999.

[153] C. R. Rao. *Linear Statistical Inference and Its Applications*. Probability and Mathematical Statistics. Wiley, 1973.

[154] J. Romberg, H. Choi, R. Baraniuk, and N. G. Kingsbury. Multiscale classification using complex wavelets. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'00)*, volume 2, pages 371–374, Vancouver, Canada, 2000.

[155] J. K. Romberg, H. Choi, and R. G. Baraniuk. Bayesian wavelet domain image modeling using hidden markov trees. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'99)*, Kobe, Japan, October 1999.

[156] M. Rosenblatt. Remarks on multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[157] S. L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–327, September 1997.

[158] E. Sayrol, J. Vidal, S. Cabanillas, and S. Santamaría. Optimum watermark detection in color images. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'99)*, volume 2, pages 231–235, Kobe, Japan, October 1999.

[159] G. Schaefer and M. Stich. UCID - an uncompressed colour image database. In *Proceedings of SPIE, Storage and Retrieval Methods and Applications for Multimedia*, volume 5307, pages 472–480, San Jose, CA, USA, January 2004. SPIE.

[160] M. F. Schilling. Two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, September 1986.

[161] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978.

[162] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The dual-tree complex wavelet transform - a coherent framework for multiscale signal and image processing. *IEEE Signal Processing Magazine*, 22(6):123–151, November 2005.

[163] C. Shaffrey, N. G. Kingsbury, and I. Jermyn. Unsupervised image segmentation via markov trees and complex wavelets. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'02)*, volume 3, pages 801–804, Rochester, New York, United States, 2002.

[164] J.M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. on Signal Process.*, 41(12):3445–3462, December 1993.

[165] M. J. Shensa. Wedding the à trous and Mallat algorithms. *IEEE Transactions on Signal Processing*, 40(10):2464–2482, October 1992.

[166] J. K. Shuttleworth, A. G. Todman, R. N. G. Naguib, B. M. Newman, and M. K. Bennett. Colour texture analysis using Co–Occurrence matrices for classification of colon cancer images. In *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'02)*, volume 2, pages 1134–1139, Winnipeg, Manitoba, Canada, 2002.

[167] E. Simoncelli and E. P. Zhibin Lei. Color channels decorrelation by ica transformation in the wavelet domain for color texture analysis and synthesis. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'00)*, pages 606–611, South Carolina, USA, 2000.

[168] E.P. Simoncelli and W.T. Freeman. The Steerable Pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'95)*, volume 3, pages 444–447, Washington, DC, USA, October 1995.

[169] M. Sklar. Fonctions de répartition à $n$ dimensions et leurs marges. *Publications de l'Institute de Statistique de l'Université de Paris*, 8:229–231, 1959.

[170] S. P. Smith and A. K. Jain. A test to determine the multivariate normality of a data set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(5), September 1988.

[171] K.-S. Song. A globally convergent and consistent method for estimating the shape parameter of a Generalized Gaussian distribution. *IEEE Transactions on Information Theory*, 52(2):510–527, February 2006.

[172] K.-S. Song. Globally convergent algorithms for estimating Generalized Gamma distributions in fast signal and image processing. *IEEE Transactions on Image Processing*, 17(8):1233–1250, August 2008.

[173] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1904.

[174] E. W. Stacy. A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192, 1962.

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

[176] G.A. Tsihrintzis and C.L. Nikias. Fast estimation of the parameters of alpha–stable impulsive interference. *IEEE Transactions on Signal Processing*, 44(6):1492–1503, June 1996.

[177] T. K. Tsui, X.-P. Zhang, and D. Androutsos. Color image watermarking using multidimensional fourier transforms. *IEEE Transactions on Information Forensics and Security*, 3(1):16–28, March 2008.

[178] S.-Y. Tung, C.-S. Wu, and M.-Y. Su. Magnifying colonoscopy in differentiting neoplastic from nonneoplastic colorectal lesions. *American Journal of Gastroenterology*, 96:2628–2632, 2001.

[179] G. Tzagkarakis, B. Beferull-Lozano, and P. Tsakalides. Rotation-invariant texture retrieval with Gaussianized Steerable Pyramids. *IEEE Transactions on Image Processing*, 15(9):2702–2718, September 2006.

[180] G. Tzagkarakis and P. Tsakalides. A statistical approach to texture image retrieval via alpha-stable modeling of wavelet decompositions. In *Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '04)*, Lisbon, Portugal, April 2004.

[181] P. Vandewalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing - What, why, and how. *IEEE Signal Processing Magazine*, 26(3):37–47, March 2009.

[182] M. Varanasi and B. Aazhang. Parameteric Generalized Gaussian density estimation. *Journal of the Acoustical Society of America*, 86(4):1404–1415, October 1989.

[183] N. Vasconcelos. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Transactions on Information Theory*, 50(7):1482–1496, July 2004.

[184] N. Vasconcelos and G. Carneiro. What is the role of independence for visual recognition. In *Proceedings of the European Conference in Computer Vision (ECCV'02)*, pages 297–311, Copenhagen, Denmark, 2002.

[185] N. Vasconcelos and A. Lippman. Library-based coding: A representation for efficient video compression and retrieval. In *Proceedings of the Data Compression Conference (DCC'97)*, pages 121–130, Snowbird, Utah, USA, 1997.

[186] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'00)*, pages 1216–1221, Hilton Head, South Carolina, United States, 2000.

[187] N. Vasconcelos and A. Lippman. A unifying view of image similarity. In *Proceedings of the International Conference on Pattern Recognition (ICPR'00)*, pages 38–41, Barcelona, Spain, 2000.

[188] G. Verdoolaege and P. Scheunders S. De Backer. Multiscale colour texture retrieval using the geodesic distance between multivariate Generalized Gaussian models. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'08)*, pages 169 – 172, San Diego, California, USA, 2008.

[189] A. Vo and S. Oraintara. A study of relative phase in complex wavelet domain: Property, statistics and applications in texture image retrieval and segmentation. *Signal Procesing: Image Communication*, 25(1):28–46, January 2010.

[190] M. J. Wainwright and E. P. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems (NIPS'99)*, volume 12, pages 855–861, Cambridge, MA, 2000. MIT Press.

[191] K. Xu, B. Georgescu, D. Comaniciu, and P. Meer. Performance analysis in content-based retrieval with textures. In *Proceedings of the International Conference on Pattern Recognition (ICPR'00)*, pages 4275–4279, Washington, DC, USA, 2000.

[192] Q. Xu, J. Yang, and S. Ding. Color texture analysis using the wavelet-based hidden markov model. *Pattern Recognition Letters*, 26(11):1710–1719, August 2005.

[193] K. Zografos. On Mardia's and Song's measures of Kurtosis in elliptical distributions. *Journal of Multivariate Analysis*, 99(5):858–879, May 2008.