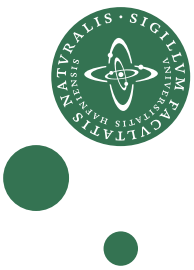


On Texture and Geometry in Image Analysis

David Karl John Gustavsson

The Image Group, Department of Computer Science
Faculty of Science, University of Copenhagen

2009

This thesis is dedicated to
Elin and Ludwig

Preface

In your hands you hold
the result of three years of hard labor in the science mine,
materialized in this PhD dissertation.

Acknowledgements

Many people have helped and supported me to get where I am today. A big thanks goes to my extended family - especially my mother Lena and my father Håkan - and to all my friends, of course.

Mads Nielsen, for excellent supervision and always putting ideas into a larger context, thank you so much! Kim S. Pedersen, for insightful supervision on almost daily basis and for great inspiration, you have my deepest gratitude and appreciation. Francois Lauze, for helping me to transform vague ideas into mathematics and for never hesitating to give a math lecture, thank you!

Anders Heyden and Niels-Christian Overgaard, for making my visit at Malmö University both enjoyable and scientifically fruitful, thank you! Christoph Schöerr, for making my visit at Heidelberg University enjoyable and for always sharing your knowledge, thank you!

I also want to thank all the PhD-students - former and current - at the former Image Group at ITU University of Copenhagen, the Image Group at DIKU and Applied Mathematics at Malmö University.

The biggest thanks goes to my wife Mariana, and our twins Ludwig and Elin, for always perfectly balancing my professional life with a perfect private life.

This research was funded by the VISIONTRAIN RTN-CT-2004-005439 Marie Curie Action within the EC's FP6.

Abstract

Images are composed of geometric structure and texture. Large scale structures are considered to be the geometric structure, while small scale details are considered to be the texture. In this dissertation, we will argue that the most important difference between geometric structure and texture is not the scale - instead, it is the requirement on representation or reconstruction. Geometric structure must be reconstructed exactly and can be represented sparsely. Texture does not need to be reconstructed exactly, a random sample from the distribution being sufficient. Furthermore, texture can not be represented sparsely.

In image inpainting, the image content is missing in a region and should be reconstructed using information from the rest of the image. The main challenges in inpainting are: prolonging and connecting geometric structure and reproducing the variation found in texture. The Filter, Random fields and Maximum Entropy (FRAME) model [213, 214] is used for inpaining texture. We argue that many 'textures' contain details that must be inpainted exactly. Simultaneous reconstruction of geometric structure and texture is a difficult problem, therefore, a two-phase reconstruction procedure is proposed. An inverse temperature β is added to the FRAME model. In the first phase, the geometric structure is reconstructed by cooling the distribution, and in the second phase, the texture is added by heating the distribution. Empirically, we show that the long range geometric structure is inpainted in a visually appealing way during the first phase, and texture is added in the second phase by heating the distribution.

A method for measuring and quantifying the image content in terms of geometric structure and texture is proposed. It is assumed that geometric structures can be represented sparsely, while texture can not. Reversing the argumentation, we argue that if the image can be represented sparsely then it contains mainly geometric structure, and if it cannot be represented sparsely then it contains texture. The degree of geometric structure is determined by the sparseness of the representation. A Truncated Singular Value Decomposition complexity measure is proposed, where the rank of a good approximation is defining the image complexity.

Image regularization can be viewed as approximating an observed image with a simpler image. The property of the simpler image depends on the regularization method, a regularization parameter and the image content. Here we analyze the norm of the regularized solution and the norm of the residual as a function of the regularization parameter (using different regularization methods). The aim is to characterize the image content by the content in the residual. Buades *et al.* [27] used the content in the residual - called 'Method

Noise' - for evaluating denoising methods. Our aim is complementary, as we want to characterize the image content in terms of geometric structure and texture, using different regularization methods.

The image content does not depend solely on the objects in the scene, but also on the viewing distance. Increasing the viewing distance influences the image content in two different ways. As the viewing distance increases, details are suppressed because the inner scale also increases. By increasing the viewing distance, the spatial lay-out of the captured scene will also change. At large viewing distances, the sky occupies a large region in the image and buildings, trees and lawns appear as uniformly colored regions. The following questions are addressed: How much of the visual appearance in terms of geometry and texture of an image can be explained by the classical results from natural image statistics? and how does the visual appearance of an image and the classical statistics relate to the viewing distance?

Contents

Chapter 1

1.1 Introduction

It is often claimed that 'everybody knows what texture is, but no one can define it'. This seems to be true both in daily discussions and in scientific papers. In image processing papers, texture is rarely defined, and if a definition is present it is often a problem specific definition that is solely valid in a specific setting. No generally accepted definition of texture exists. One reason for the absence of such a definition is the fact that texture should capture a large and partly contradictive set of concepts - regular, irregular, stochastic, stationary, outer scale and inner scale. Because of the large variation of concepts included in texture, it is fairly easy to construct negative examples that annihilate an attempt to define texture.

Images are often viewed as a composition of geometric structures and texture. The geometric structure is considered to be the large scale structure, while texture is considered to be the small scale details. Geometric structure is considered to be simple because of its rather homogenous appearance. Smooth objects under the same illumination will reflect the light in a similar way, resulting in smooth geometric structures. A scene is composed of independent, discrete and roughly uniformly colored objects occluding each other. Consider a concrete building viewed from a large distance or a person with a uniformly colored sweater viewed from a few meters, both the concrete building and the sweater will appear as smooth uniformly colored objects, i.e. geometric structure. Texture, on the other hand, is considered to be more complex, because it is composed of a large number of small scale elements. Under the same illumination, regions with large numbers of small scale elements will reflect light in different ways. The small scale elements can be either the roughness of the surface or small scale 'objects' such as leaves or hair. Texture can also be viewed as a composition of independent, discrete and roughly uniformly colored objects but on a smaller scale. This reveals one fundamental property of texture: it always contains some kind of varia-

tion. When the distance to the concrete building is decreased, the roughness of the concrete becomes visible and the smooth geometric structure therefore transforms into a textured object.

Viewing images as a composition of geometric structure and texture implies the possibility to decompose an image into its components, i.e.

$$I = I_{struct} \oplus I_{text}, \tag{1.1}$$

where $\oplus$ is some image composition operator, I_{struct} is the geometric structure component and I_{text} is the texture component. Two very different examples of image decomposition, which will be discussed later, are Total Variation image decomposition [169] and Primal Sketch [56, 57]. In Total Variation decomposition, an image is decomposed by minimizing an energy functional and the composition operator $\oplus$ is ordinary addition - the intensities of the structure and texture component are added to form the images. In Primal Sketch by Guo *et. al* [56, 57] - inspired by earlier work by Marr [128] - the operator $\oplus$ denotes a rather complex algorithm, which includes image segmentation and texture modeling. Here the geometric structure is formed by the boundaries of objects defined by edges over a fixed scale, and the object boundaries are represented by sparse coding. The texture component is formed by the remaining regions, which are not object boundaries, divided into regions containing stationary textures and are represented using a Markov random field model (FRAME).

Models for image decomposition also relate to image formation models. The Dead Leaves model is a well known formation model for natural images, introduced in a morphological setting by Matheron [131], further studied in connection with natural image statistics by Ruderman [168], for size distribution of homogenous regions by Alvarez *et. al* [1, 3] and for modeling scale invariance in natural images by Lee *et. al* [117]. The Dead Leaves model is based on the notion that a scene is composed of independent discrete objects of different sizes and different colors occluding each other. The scene is composed of planar templates T of the same shape, often squares or circles, but of different sizes and colors. Templates of random size and color are randomly located in the $3D$ world such that the (x, y) plane, i.e. the image plane, is totally covered. The z dimension is solely used to determine which of the template is closest to the image plane. The intensity in a location (x, y) is determined by the color of the template closest to the image plane (i.e. the template with the smallest z). As shown by Lee *et. al*, the Dead Leaves Model is a generative model that can reproduce statistical properties found in ensembles of natural images.

Grenander *et. al* [79] use 2D profiles called generators - gs - of 3D ob-

jects to analyze images. The gs are views or appearances of the 3D objects captured from a random viewing position. An image is composed of random selected generators of random size, color and location. Grenander *et. al* use a superposition model - instead of an occlusion model as the Dead Leaves model - where the intensity in a location x is the sum of intensities of the generator covering the location. Grenander *et. al* [78] and Srivastava *et. al* [184] use the generator-based image formation process to analytically find probability distributions for linear marginal distributions - i.e. the histogram of linear filter responses - called Bessell K-form.

Image decomposition methods often use an implicit definition of geometric structure and texture. Geometric structure in the image is the content in the structure component and texture is the content in the texture component. Geometric structure and texture is therefore defined in terms of how the method decompose the image, i.e. each decomposition method has its own implicit definition of structure and texture.

The notion of scale is almost always discussed in connection with texture. Indeed, texture is sometimes defined, as in the monograph on texture by Petrou and García [155], as: 'the details on a scale smaller than the scale of interest'. Two fundamental different types of scales, as pointed out by scale space theory [98, 202, 111, 122], are present: the inner scale and the outer scale. The inner scale is the smallest details captured by the camera, and the outer scale is the field of view, or the part of the scene, captured by the camera. As discussed earlier, texture always contains some kind of variation. The variation property of texture indicates that texture should be defined on larger regions - it does not make sense to call a pixel or a small patch a texture. A texture must be defined on a large enough region, such that the variation is present. The outer scale must be sufficiently large, such that the texture variation is present. Consider the brick wall shown in figure 1.2, the image to the left contains roughly one brick, while the image to the right contains a larger part of the wall. The bricks in the image to the right can be considered to form a texture; the outer scale is large enough to capture the variation of the brick wall. The bricks in the image to the left cannot be considered to be texture because the outer scale is too small to capture the variation of the brick wall. On the other hand, the variation on the bricks in the image to the right can be considered to be texture.

In this thesis, we will argue that the most important difference between geometric structure and texture is the exactness in which the content must be described or represented. Both the geometric structure and the texture are considered to be samples from stochastic processes. For the geometric part, a (the) fixed specific instance is required, i.e. the geometric structure needs to be described exactly. The texture component does not need to be described

exactly - instead, a random sample from the distribution is sufficient.

Furthermore, the details that must be described exactly in an image depend on the problem at hand. The image content that requires an exact representation depends on what it should be used for. The same image in another context may alter the content that needs an exact representation.

Consider shining stars on a dark sky - as shown to the left in figure 1.1: is this geometric structure or texture? As background in a romantic scene of a feature movie, the stars in the dark sky represent texture. In such a setting, the stars appear to be small lighted dots of different sizes randomly located in the dark sky. Any random sample from the same distribution would serve equally as good as background in the movie. On the other hand, if the image is used for orientation or for locating a constellation of stars, then the exact size and location of the star must be represented. In such a setting, the shining stars are no longer texture because a random sample from the distribution is not sufficient. Consider an aerial photo, as shown in the middle in figure 1.1: is this geometric structure or texture? Again, it depends on the context or the problem at hand. If the photo is used for finding roads or counting trees, then the details must be represented exactly. On the other hand, viewed as an aerial photo of a landscape, any random sample would serve equally well. To the right in figure 1.1, a gathering of people is shown - is this geometric structure or texture? If one is searching for a specific person, then each person in the gathering must be represented exactly and the content should be considered as geometric structure. Viewed as an example of a gathering of people at a football match, it can be considered to be texture. In all examples the contents are considered to be texture if pointing the camera at another location with the same type of content would serve equally well; pointing the camera at another part of the sky, aerial photo of the landscape at another location or another gathering of people (possible at another part of the grand-stand).

The insight that geometric structure must be represented exactly, while texture can be represented by some distribution leads to the question: 'How much geometric structure and texture does an image contain?'. Measuring and quantifying the image content in terms of geometric structure and texture is a challenging problem. Assuming that geometric structure can be represented exactly using some sparse representation, while texture cannot. Reversing the assumption leads to an approach to measuring the image content. If the image can be represented sparsely, then it contains geometric structures. Images that can not be represented sparsely mainly contain texture. Using this approach the image content can be quantified by the sparseness of the representation.

The approach proposed in this thesis started to arise while experimenting

with the inpainting problem (chapter 2 and papers [84, 86]), and evolved during the project to finally become the main theme in the thesis. In the inpainting problem, the image content (intensities) in a region Ω is missing and should be reconstructed using information in the rest of the image and constrained to the boundary of Ω. In image inpainting, one is facing two types of fundamentally different problems:

- Prolonging and connecting geometric structures.
- Reproducing the variation found in texture.

Prolonging and connecting geometric structures needs to be done exactly, and the result is almost binary: the number of visually appealing reconstructions for the geometric structure are few and in the extreme case just one. Texture, on the other hand, contains some degree of freedom, which influences the number of visually appealing reconstructions. At the first glance, it seems like the two problems are fundamentally different and not related at all. That is also the approach often found in the literature, geometric structures can be reconstructed by minimizing a suitable functional, such as Total Variation, while texture should be reconstructed using a suitable texture synthesizing method. The difference between image synthesizing and inpainting lies in the boundary conditions that put hard restrictions on the possible reconstructions in the latter case. Many textures also contain geometric structures at a smaller scale, which need to be reconstructed exactly. So geometric structures that require an exact reconstruction are present even in textures. This is sometimes referred to as textons [101, 209, 210].

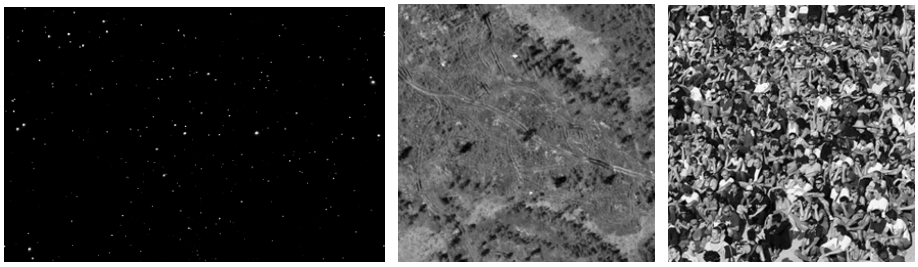


Figure 1.1: Shining stars on a dark sky, an aerial photo and a gathering of people. Texture or geometry?

In chapter 3, a newly collected database containing sequences of natural images containing the same scene, captured at different viewing distances, is presented. How does changing the viewing distance influence the image content in terms of geometric structure and the image content? Does the

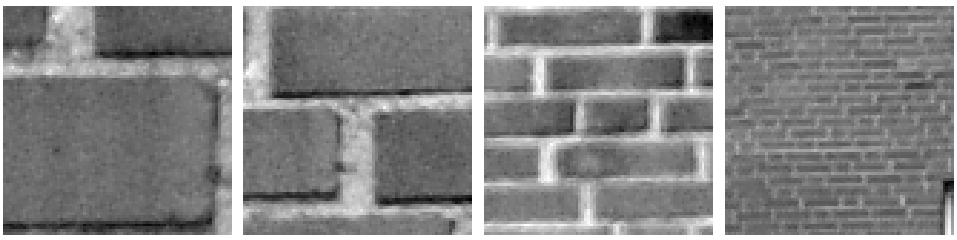


Figure 1.2: A brick wall captured at different viewing distance. Four 80×80 patches contain a brick wall captured at different viewing distances. What is the geometric structure and texture in the different images?

image composition in terms of geometric structure and texture depend on the viewing distance? The viewing distances influence the image content in two ways: the image composition ('outer-scale') and the level of captured details ('inner-scale'). Classical statistical properties of natural images are estimated on individual images and analyzed with respect to the viewing distance. The estimations are strongly linked with both the image content and the viewing distance. The results are further analyzed and discussed in section 1.3.

In chapter 4 (paper [85]), an image complexity measure is proposed. The basic idea presented in the paper is that an image is simple if it can be can be approximated with a small error using a sparse representation, and it is complex otherwise. Large scale geometric structures can be described using a sparse representation, while small scale stochastic texture cannot. In this sense, geometric structure is simple, while texture is complex. The proposed method is based on truncated singular value decomposition and the optimal rank-k property of such approximation. Furthermore, an image is composed of image patches and the complexity of the image should be determined by the patches in the image. The results are further analyzed and discussed in section 1.4.

In chapter 5 (paper [83]), image regularization methods are used to characterize the image content. Image regularization can be viewed as approximating the observed image with a simpler image, where simpler is defined by the regularization term and the regularization parameter. By increasing the regularization parameter, the regularized image gets simpler and more details in the observed image are suppressed. The residual image contains the details that have been suppressed during the regularization and the norm of the residual image is a measure of the suppressed details. The norm of the residual as a function of the regularization parameter, measures the amount

of details that are suppressed at different scales. Of interest is also the derivative with respect to the regularization parameter, which reveals the rate in which details are suppressed. The regularized solution contains the large scale geometric structure, while the residual image contains the texture. By measuring the content in the regularized and residual image the degree of geometric structure and texture can be quantified. The results are further analyzed and discussed in section 1.5.

In chapter 6 (papers [65, 82]), at first glance, a rather different topic was treated. The goal was to combine temporal inpainting with segmentation using shape prior. The research was done during a 3-month visit at Malmö University visiting Prof. Anders Heyden. By using the previous segmentation as a shape prior for the current segmentation, good segmentation will be achieved even if the object of interest is occluded or missing in some of the frames. How can object boundaries be used to estimate the motion of the object? The object boundaries, in form of simple connected curves, should solely be used for computing the motion, in the form of a displacement field mapping one image onto the other. The image content, or features based on the image content, cannot be used because they are assumed to be unreliable; for example, in inpainting where the image content is lost or in the case when the objects do not contain enough features, such as clouds. How can the geometric structure - the object boundary - be used to compute the motion and deformation of a non-rigid deformable object, including the motion of the interior of the object? The results are further analyzed and discussed in section 1.6.

1.1.1 The Bayesian Approach and MAP-solution

The Bayesian framework will be used in later sections in connection with inpainting and image decomposition and is introduced here in a general setting.

Let us introduce the Bayesian approach in a general signal processing setting. Let u_0 be an observed signal that is a degenerated version of a 'clean' signal u. The goal is to recover the 'clean' signal u, using information from the observed signal u_0. The *Posteriori distribution* $p(u|u_0)$ is the conditional probability distribution for the 'clean' signal u, given the observed signal u_0. Of special interest is the u, which maximizes the posteriori distribution - the 'clean' signal with the highest probability. This is the *Maximum A Posteriori (MAP)* solution defined as

$$u_{MAP} = \arg\max_u \{p(u|u_0)\}. \tag{1.2}$$

To compute the posteriori distribution and find the MAP-solution, Bayes' rule is often used. Bayes' rule states that

$$p(u|u_0) = \frac{p(u_0|u)p(u)}{p(u_0)}, \tag{1.3}$$

where $p(u_0|u)$ is the *data model term* (or *likelihood*), $p(u)$ is *prior term* and $p(u_0)$ is a normalization constant (which is usually ignored). Bayes' rule connects the posteriori distribution with likelihood and prior distributions. Using Bayes' rule to compute the MAP-solution

$$u_{MAP} = \arg\max_{u} \left\{p(u_0|u)p(u)\right\}. \tag{1.4}$$

the likelihood and prior term must be estimated to find the MAP-solution. It is often simpler to estimate the likelihood and prior term, than the posteriori distribution directly. The book by Kaipo and Somersalo [103] treats inverse problem from a (Bayesian) statistical point of view.

1.2 Inpainting using FRAME - Filter, Random fields And Maximum Entropy

1.2.1 Inpainting

Image inpainting - also known as image completion and hole filling - deals with the problem of reconstructing the image content in a missing region, using information in the rest of the image and constrained by the boundary of the missing region. The term 'image inpainting' is an artistic synonym for image interpolation and it comes from the restoration of paintings in museums. The term image/digital inpainting was first used by Bertalmio *et. al* [16].

The general inpainting problem can be formulated as: an image $u_0 = u_0(x)$ is defined on the image domain $x \in D$. For some reason, a subset $\Omega \subset D$ is missing or unavailable. The objective for image inpainting is to reconstruct the entire image u from the incomplete image u_0. (Figure 1.3 contains a visualization of the notation.) It is often assumed that $u(D \setminus \Omega) = u_0(D \setminus \Omega)$ - i.e. the content in the non-missing region should not be altered; in other cases, u_0 may be degraded due to noise and/or blur and the content in $u_0(D \setminus \Omega)$ may be altered. Let $\widetilde{\Omega}$ be an extended region including Ω such that $\Omega \subset \widetilde{\Omega} \subset D$.$\widetilde{\Omega}$ could, for example, be a rectangle covering Ω.

It is also worth noting that, in most cases, the main evaluation criteria is how visually appealing the inpainting is.

Applications of inpainting are, restoration of damaged images [39], restoration of damaged films [115, 67, 116], removing unwanted objects in images and sequences, super resolution [62, 106], lossy image compression [66], recover missing block in transmission and compression [163] and deinterlacing [10, 107, 108, 105].

In a Bayesian setting, the inpainting problem is stated as follows: the posteriori distribution $p(u|u_0)$ is the probability distribution of the inpainting u given the incomplete image u_0 and the MAP-solution is the most likely inpainting given u_0. To use the Bayes' rule to compute the MAP-solution, as stated in equation 1.4, the likelihood and the prior term must be estimated.

1.2.2 Related Work

Inpainting methods are often categorized into functional-based (or diffusion-based) and texture synthesis methods. The functional-based methods are considered to be able to reconstruct geometric structure, while they fail to reconstruct texture in a visually appealing way. Texture synthesis methods

Figure 1.3: The inpainting objective is to reconstruct the entire image $u(D)$ using the content in $u_0(D \setminus \Omega)$, where D is the image domain and Ω is the missing region. Two well-known example images often showed in inpainting papers: scratched photo of three girls and New Orleans covered with text.

are considered to reconstruct texture well, but fail to reconstruct geometric structures in a visually appealing way. The failure of the functional-based methods on texture is evident and has been shown many times. The failure of texture synthesis methods on geometric structure is less evident and, in fact, it has rarely been shown on realistic images.

Furthermore, the texture synthesis methods can be divided into parametric models and non-parametric/patch-based models. In parametric models, a parametric representation of the probability distribution is used and parameters are estimated using an observed image (or the non-missing part of the image). The missing content is reconstructed by sampling from the distribution. In non-parametric models, the probability distribution is represented non-parametrically by the patch samples from the known part of the image. The missing content is reconstructed by directly querying the patch samples.

In inpainting literature, it is often assumed that texture synthesis and texture inpainting are identical problems. If a method can synthesize a texture, it can directly be used for inpainting. The main difference between texture synthesis and texture inpainting is the presence of boundary conditions in the latter. The boundary conditions put hard constraints on the possible reconstructions.

PDE- and Functional-Based Methods

The recent monograph on image processing by Chan and Shen [40] contains an overview of the functional/PDE-based inpainting methods. The basic idea in most PDE-based methods is to prolong and connect the geometric structure present in the surroundings of the missing region Ω. The geometric

structures present in $\partial\Omega$ are the level lines, and the problem is to prolong and connect the level lines.

Masnou and Morel [130] (see also [129]) were the first to use variational image interpolation for edge completion. They proposed a disocclusion algorithm - inpainting algorithm - in which the elastic functional was minimized inside the missing region.

Bertilmio *et. al* [16] proposed a third order PDE solved only inside Ω with proper boundary condition given by $\partial\Omega$. The PDE is based on the transport equation, where the information transported is defined by the Laplacian transported orthogonal to the image gradient (i.e. along the isophotes).

Ballester *et. al* [11] used a joint interpolation of vector fields and intensities approach, Tschumperle [195, 194] proposed a tensor-driven PDE and Peyrés *et. al* [157] used a non-local regularization approach.

Total variation image decomposition has also been used for image inpainting [38, 39]. The Total Variation energy functional is defined as

$$E(u; u_0) = \int_{D\setminus\Omega} (u_0 - u)^2 dx + \lambda \int_D |\nabla u| dx \tag{1.5}$$

and the solution is a minimizer of energy functional. Minimizing the total variation energy functional using the calculus of variation leads to a second-order PDE. This second-order PDE prolongs and connects contours (geometric structure) with straight 'lines'. Geometric structures will be prolonged and connected using straight lines only if the missing region is smaller than the geometric structure.

TV inpainting uses the L^1-norm of the gradient as the smoothness term, resulting in a piecewise constant inpainting. In Harmonic inpainting, the L^2-norm of the gradient is used as the smoothness term, resulting in a smooth and blurred inpainting. Chan and Kang [37] use harmonic inpainting and total variation inpainting to analyze the error.

Total variation image decomposition has also been used for temporal inpainting combined with optical flow [116] and for deinterlacing [106].

Texture Synthesis - Parametric Models

In parametric models, an observed sample image is used to estimate the parameters in the image model, represented as parametric probability distribution. A sample is drawn from the probability model in order to synthesize an image. The observed data is used to estimate parameters in parametric probability distribution and the distribution is used for generating a sample. Ideally, a random sample from the parametric model should be drawn, but in most cases, a 'typical' sample is drawn instead.

Heeger and Berger [91] proposed an image pyramid approach for texture synthesis. It is based on the assumption that first order statistics of appropriately chosen linear filter responses capture the relevant information for characterizing the texture. A collapsing pyramid is used for matching the histogram of filter responses, resulting in an image with the same marginal distribution.

Portilla and Simoncelli [161] compute various correlations and use those correlations as constraints while synthesizing. An image is synthesized, subject to the constraints, by iteratively updating the image and projecting it onto the set of images satisfying the constraints.

Peyré [156] combined the sparse representation, using the non-orthogonal basis proposed by Olshausen and Field [149], with the non-negative matrix decomposition proposed by Lee and Seung [118, 119] to 'learn' the image basis - the image 'dictionary'. The variance and kurtosis of marginal distribution of the decomposed image, using the learned dictionary, are used to synthesise a texture. A 'typical' sample that matchs the marginal statistics is drawn, using a modified version of the sampling method proposed by Portilla and Simoncelli [161].

Texture Synthesis - Non-parametric Models

In non-parametric models, no model is specified a priori. Instead, the data, in terms of patch samples, is used directly for estimations. The probability distribution is represented non-parametrically by the patch samples.

A patch-based approach was proposed by Efros and Leung [52] for synthesizing textures. Instead of drawing a random sample from a statistical model, the sample image is used directly for synthesizing the texture. An image is synthesized in a pixel-by-pixel manner. A site x which should be synthesized is picked. Let $N(x)$ be a square window neighborhood of x, i.e. an image patch centered at x. The image patch N_{best} which is closest to $N(x)$ in the Sum-of-Square Distances (SSD) sense is found. The set of patches for sampling is given by $\Omega(x) = \{N : SSD(N, N_{best}(x)) \leq \epsilon\}$. The center pixels in each patch in $\Omega(x)$ form a histogram of intensities, with a neighborhood similar to $N(x)$. The intensity in site x is a random sample from the distribution. It is not only the size of the square window $N(x)$ that is crucial (as pointed out in the original paper), but also the visiting order - i.e. the order in which the intensities should be synthesized.

Criminisi *et al.* [47, 48] used a patch-based approach for image inpainting, but instead of an onion pealing visiting order they used a priority order. The priority order depends on two terms: the number of neighbors with known intensity (the amount of reliable information surrounding the site) and a

term which explicitly encourages geometric structures (often called isophotes in inpainting literature). Giving higher priority to sites containing geometric structures will prolong and connect geometric structures.

Efros and Freeman [51] proposed a fast and simple method for texture synthesis called image quilting. Square sized image patches from a sample image are placed in raster order in such a way that the boundaries are overlapping. In the overlapping boundary, the squared intensity difference is computed and a minimum boundary cut is computed. This will result in a ragged edge between the patches and the feature in the texture is better preserved. This is an early application of the well-known graph-cut using max-cut/min-flow algorithms (and dynamic programming) for solving image processing problems [113].

As reported by Cuzol *et. al* [49], the methods proposed by Efros and Leung [52] and Criminisi *et. al* [48] are one-sweep methods without any back-tracking or Gibbs-sampling. Once the intensity for a site has been determined, it will not be altered; this may lead to visual inconsistency. Cuzol et. al. [49] propose a particle filter-based approach for re-sampling in patch-space to overcome the 'one-sweep' problem.

Combining Geometric and Texture Inpainting

Some attempts to combine PDE/functional- and texture-based methods exist.

Bertalmio *et. al* [15, 17] decompose the image into a geometric and texture component using Meyers' G-norm [132, 7, 8]. The inpainting is done component-wise using different methods. The geometric component is inpainted using a third-order PDE developed by Bertalmio *et al.* [16] (briefly mentioned in the PDE methods section). The texture component is inpainted using a slightly modified version of the patch based texture synthesizing method proposed by Efros and Leung [51] (discussed in the texture inpainting section). A similar approach was proposed by Rane *et. al* [163] to recover missing image blocks in transmission and compression. Bugeau and Bertalmio [31] use a similar approach and evaluate different methods for the different components. Their results indicate that the method proposed by Tschumperle [195] is in general preferable in the geometric component.

Elad *et al.* [54] use the sparse representation-based image decomposition method Morphological Component Analysis (MCA)[54, 53] for inpainting. In MCA, as in TV or Meyer-decomposition, an image is decomposed using ordinary addition into a geometric and a texture component. In MCA, a sparse representation approach is used and two dictionaries are learned; T_t should sparsely decompose the texture component (and it should not be

able to represent the geometric component sparsely) and T_g should sparsely decompose the geometric component (and it should not be able to represent the texture component sparsely). Formally, the image is decomposed

$$I = I_g + I_t = T_t\alpha_t + T_g\alpha_g, \tag{1.6}$$

where α_t and α_g are the sparse coefficients in the geometric respective texture component. As an addition regularization term, total variation is used solely on the geometric component $T_g\alpha_g$. The dictionaries are learned using the non-missing part of the image and the missing information is inpainted simultaneously component-wise using the learned dictionaries.

1.2.3 FRAME

FRAME - Filter, Random fields And Maximum Entropy by Zhu *et. al* [213, 214] is a general framework for analyzing and synthesizing stationary textures. FRAME is based on two properties: (i) textures having the same marginal distributions - histogram of filter responses - are visually hard to discriminate and (ii) a probability distribution is uniquely determined by all its marginal distributions.

The basic idea behind FRAME is as follows: Let H be a set of statistics in the form of histograms of filter responses (marginal distributions) extracted from an observed image and let $\Omega(H)$ be the set of all probability distributions with the same (expected) marginal distributions as the observed image (i.e. H). Among the distributions - $\Omega(H)$ - that are consistent with the observed image, select the least committed distribution, that is the distribution that maximizes the entropy.

Even if the fundamental idea behind FRAME is rather straight forward, a detailed discussion requires a large amount of notations. Let $F = \{F^\alpha : \alpha \in K\}$ be a set of filters, $I^\alpha = I * F^\alpha$ be the filter response (an image) using filter α and $H^\alpha = \langle h_1^\alpha, \cdots, h_N^\alpha \rangle$ be the (normalized) histogram for filter α using N bins. Furthermore, let $\Omega(H) = \{p(I) : E_p(\mathcal{H}(I * F^\alpha)) = H^\alpha\}$, where E_p is the expectation and $\mathcal{H}$ is the histogram operator using N (fixed) bins. This is simply 'all probability distributions that have the same marginal distributions as the observed image'. Among the distributions $p(I) \in \Omega(H)$, the one that maximizes the entropy is selected (i.e. maximum entropy is the objective function), which leads to a constrained optimization problem which can be solved using the technique of Lagrange multipliers. The solution has the following form

$$p(I) = \frac{1}{Z(\Lambda)} \exp\left\{ -\sum_{\alpha}^{K} \sum_{i}^{N} \lambda_i^\alpha h_i^\alpha \right\}, \tag{1.7}$$

where $\Lambda = \{\lambda_i^\alpha\}$ is the Lagrange multipliers and $Z(\Lambda)$ is a normalization constant (that depends on Λ). λ_i^α is the Lagrange multiplier corresponding to h_i^α the histogram for filter α and bin i.

To synthesize an image, a random sample from the distribution is drawn. A random sample from $p(I)$ is generated by gibbs-sampling from the conditional distribution

$$p(I(x)|I(D\backslash x)) = p((I(x))|I(N_x)), \tag{1.8}$$

where N_x is the neighborhood of x (defined by the filter support). By repeatedly randomly selecting sites and sampling the intensity given the current intensities in the neighborhood N_x, a random sample from $p(I)$ will be generated.

1.2.4 Cooling and Heating - Inpainting using FRAME

Let $\widetilde{\Omega}$ be a square extended region of Ω such that $\Omega \subset \widetilde{\Omega} \subset D$, and of such a size that the filter is covered in $\widetilde{\Omega}$ for sites $x \in \Omega$. Inpainting is done by gibbs-sampling from the conditional distribution

$$p((I(x))|I(N_x)), \tag{1.9}$$

where $x \in \Omega$ and N_x is the neighborhood of x. N_x may contain sites $x \in \Omega$ and $x \notin \Omega$. Only sites $x \in \Omega$ are updated, while sites $x \in \widetilde{\Omega} \setminus \Omega$ are used as boundary condition. The intensities in the missing region are initialized by sampling from an (independent) uniformly distributed random distribution.

Synthesizing an image by sampling from the distribution showed visual similarity with the observed image. The features present in the observed image are also present in the sample from the distribution - the optimization process has converged visually. The visual convergency shows that the filters used caught the important visual features of the observed image. In contrast, inpainting by sampling from the distribution in the missing region Ω did not converge to a visually appealing solution. The failure was evident close to the boundary where the geometric structure was not prolonged in a visually appealing way. The problem of using FRAME for reconstructing large scale image structures such as edges was also observed in [57], where primal sketches were used to extract the edges. This supports the claim

made in this thesis that certain structures even on a smaller scale must be reconstructed exactly.

An inverse temperature $\beta = \frac{1}{T}$ was added to the distribution

$$p(I) = \frac{1}{Z(\Lambda)} \exp\left\{ -\beta \sum_{\alpha} \sum_{i} \lambda_i^{\alpha} H_i^{\alpha} \right\}. \tag{1.10}$$

Cooling the distribution - increasing β - will decrease the probability for images with low probabilities and increase the probability for images with high probabilities. Increasing β will narrow the probability distribution in the sense that a large part of the probability density will be located at a smaller subset of all images, and as $\beta \to \infty$ all the probability density will be located at the global maxima. In this sense, the distribution will be less 'stochastic' as β increases. Let N_{max} be the number of global maxima and let I_{max} be the set of all global maximizers, then as $\beta \to \infty$

$$p(I) = \begin{cases} 0 & \text{if } I \notin I_{max} \\ \frac{1}{N_{max}} & \text{if } I \in I_{max} \end{cases}$$

The probability mass is uniformly distributed over the set of global maximizers.

Cooling the Distribution - Adding the Geometry

The idea behind the cooling approaches is that the large scale geometric structures are brought out by cooling the distribution, while the small scale structures are suppressed. A more MAP-like solution is assumed to contain the large scale geometric structure, while suppressing the small scale details. As pointed out by Nikolova [144], the MAP-solution may not be smooth and may instead contain small scale structures.

To stress the inference of the geometric structure in the inpainting, three cooling approaches are proposed.

The first approach is to cool the distribution using a fixed $\beta > 1$. The motivation for this approach is to emphasize more likely structures, and fade out less likely structures. It is a redistribution of the probability mass in such a way that a larger part of the probability mass is located on the images with higher probabilities, while a smaller part of the probability mass is located on the images with lower probabilities.

The second approach is the so-called Iterated Conditional Mode (ICM), analyzed by Besag in [19] and by Kittler and Föglein in [110], which corresponds to setting $\beta = \infty$. When updating the intensity by gibbs-sampling from the conditional distribution (1.8), the ICM approach corresponds to

always selecting the intensity with the highest probability. The intensity at site $x \in \Omega$ is updated using

$$I(x) = \arg\max \{p(I(x)|N_x)\}, \tag{1.11}$$

i.e. the most likely intensity given the current neighborhood N_x. It is a site-wise greedy approach which, always locally selects the intensity with the highest probability. ICM depends both on the initialization of the missing region Ω and on the visiting order. If the missing region is initialized randomly and the visiting order is random, then repeating the inpainting will give different results. Winkler [201] and Li [121] contain a general discussion about ICM.

The third approach is a fast cooling scheme which gradually increases β. The fast cooling scheme has the following form

$$\beta_{n+1} = C_+ \cdot \beta_n = (C_+ \cdot \beta_0)^n, \tag{1.12}$$

where $C_+ > 1$ is the increment factor and $\beta_0 > 0$ is the initial inverse temperature. The fast cooling scheme was motivated by a simulated annealing approach for finding the MAP-solution of Markov Random Fields (MRF) [69]. By iteratively increasing β, the probability mass is gradually moved from low probability images to high probability images. The gradual increase of β - the annealing process - decreases the probability of getting stuck in a local optimum. Winkler [201], Bremaud [23] and Li [121] contain general treatment of simulated annealing and MAP-solutions for MRF's.

Heating the Distribution - Adding the Texture

The geometric structure is reconstructed by sampling from a cooled distribution or using a cooling scheme. In the cooling phase, the large scale geometric structure is added, while the small scale texture is suppressed. The result is prolonged and connected geometric structures that appear too smooth. In order to add the small scale texture, a second heating sampling phase was used. The 'initialization' for the second phase was the result of the first 'cooling' phase inpainting - the reconstructed geometric structure. The small scale texture should be added in this phase, without destroying the large scale geometric structure, reconstructed in the previous phase.

Two heating approaches were evaluated. The first approach was to use a fixed $\beta \geq 1$, where $\beta = 1$ corresponds to the original learned FRAME model.

The second approach was to use a simulated heating-like process, fast heating scheme, where β is gradually decreasing. The fast heating used was of the form

$$\beta_{n+1} = \beta_n \cdot C_- = (\beta_0 \cdot C_-)^n, \tag{1.13}$$

where $C_- < 1$ is the decrement factor and β_0 is the initial inverse temperature. The stopping criterion was $\beta_n < 1 + \epsilon$.

1.2.5 Discussion

Simple stationary textures containing structures on at least two scales are used. The coarse scale structure was rather large compared with the image resolution. Corduroy, birch tree bark and batten images were used from the KTH-TIPS2 database [63]. The corduroy is composed of rather large horizontal geometric structures with some intensity variation, the birch tree bark contains large scale geometric structures of different sizes distributed along the diagonal, and the batten contains connected vertical geometric structures which both merges and splits.

The missing region was rather large, relative to the larger scale details in the texture. The missing region was roughly 4 times the size of the large scale details. This implies that PDE-based methods such as TV would fail to prolong and connect the geometric structures.

The experiments on the corduroy, birch tree bark and batten images show that by cooling the distribution, the geometric structure is reconstructed better.

Adding a fixed $\beta > 1$, stressed the inference of the geometric structure. The geometric structure on the boundaries was prolonged and connected in a visually appealing way, while the small scale variation was suppressed. If β was too small or too large, then the geometric structures were not prolonged and connected in a visual appealing way.

The ICM approach which corresponds to $\beta = \infty$ did reconstruct the geometric structure in the corduroy image, but failed to reconstruct the geometric structure in the birch bark and batten image. In the birch bark and batten images, geometric structures were constructed 'randomly', depending on the random initialization and visiting order.

The fast cooling scheme reconstructs and prolongs the geometric structure found in the three images.

The geometric structure is reconstructed better by sampling from a cooled distribution or using a cooling scheme. Sampling from a distribution using a fixed, but not too large, β will often reconstruct the geometric structure. Using a fast cooling scheme will prolong and connect the geometric structure in a visually appealing way.

The texture is added, after the geometric structure has been reconstructed, by sampling from a heated distribution. Using the fixed β approach, where β is slightly larger than 1, added the texture without removing the geometric structure. If β is too large, no or very little texture is added. If β is too small, the geometric structure starts to degenerate.

Similar behavior was observed when using the fast heating scheme. For β_n much larger than 1, few intensities were altered; as β_n approached 1 more intensities were altered. If the fast heating scheme was stopped too early, then no or very little texture was added. On the other hand, if it was stopped too late - β_n was too small - then the geometric structure started to degenerate.

1.3 DIKU Multi-Scale Image Database

Image content does not solely depend on the object in the captured scene, but also on the viewing distance. Changing the viewing distance, either by physically moving the camera or by changing the focal length, will alter the image content. The visual appearance of a tree viewed from a few meters is rather different from viewing the same tree from 200 meters. At a few meters, the branches and even the individual leaves are visible. As the viewing distance increases, details are suppressed and the tree top appears as a uniform green region. Given an image, a coarse scale representation of the image can be generated using the linear gaussian scale space [98, 202, 111, 60, 122]. The coarse scale representation is generated by

$$I_t = G_t * I_0, \tag{1.14}$$

where $*$ denotes the convolution operator, I_0 the observed image and

$$G_t(x, y) = \frac{1}{2\pi t} exp\left\{-\frac{x^2 + y^2}{2t}\right\}. \tag{1.15}$$

In scale space theory, a coarse scale representation of the image is generated with the same resolution. By increasing the viewing distance, details will be suppressed, but the resolution will also decrease. The statistical behavior of natural images over scales has been studied in great detail [193, 205]. Here we consider images containing the environment - both nature and man-made structure - viewed from a normal human perspective (i.e. 'bird' and related perspectives are excluded).

Increasing the viewing distance will also alter the outer-scale of the image and the spatial layout of the captured scene. A cup on a table can be captured from almost all angles, a car on the street can be captured from many angles, while a building captured from 200 meters can be captured from a few angles. The distance to the main objects in the scene puts constraints on the spatial lay-out of the captured scene. As the viewing distance increases, the spatial layout will change - the sky will appear at the top of the image, houses will appear as uniformly colored blocks in the middle of the image and mountains will appear as a smoothly changing region in the middle of the image.

We propose to collect a new image database containing the same scene captured at different viewing distances (by adjusting the focal length). The database should contain sequences of the same scene captured using different focal length. Furthermore, the region present in all images in a sequence should be extracted, resulting in a set of sequences of images containing the same part of the scene captured at different scales (focal length in the

objective). The extracted region contains the same part of the scene captured at different scales and of different resolutions.

1.3.1 Related Work

In section 1.3.3, we discuss the scale invariant property of ensembles of natural images [94, 117, 58, 193, 168, 166, 166, 183], modeling the partial derivative of an ensemble of images using the generalized Laplacian distribution [117, 94, 126, 183] and the distribution of homogenous regions [76, 1, 3, 2, 77].

Torralba and Oliva [192, 193, 146] analyze the η in the power spectra power law for natural images as a function of viewing distance. They used a more complete model, where η depends on the orientation [9]. The spatial composition of an image - called the spatial envelope - is constrained by the viewing distance. Objects viewed from a small distance can be observed from almost any point of view. As the viewing distance increases, the possible point of view from which an object can be observed decreases. This property is strongly reflected in the η in the power spectra power law. The estimation of η can be used for estimating the distance to the objects in the scene.

Wu *et. al* [204] analyze the image content as a function of viewing distance using information theoretical tools, see also [199, 210]. They analyze how the compression rate and the entropy of the image gradient changes as a function of viewing distance. The starting point is the behavior of the Dead Leaves model [131, 168, 1, 117] as the distance to the 'image plane' increases. In the Dead Leaves model, images are formed by discrete objects of random size and color, occluding each other - a scene is composed of discrete objects occluding each other. The objects in the Dead Leaves model have the same shape (template) - often circles or squares - called leaves, while the size and the coloring are random. The components in the Dead Leaves model are:

- The size r of the leaves (template) follow a distribution $p(r) \propto \frac{1}{r^2}$, over a finite range $[r_{min}, r_{max}]$.
- The color of the objects is uniformly distributed over $[a_{min}, a_{max}]$.
- The position (x, y, z) is following a Poisson process with intensity λ, and the z-axis is solely used for occlusion detection.

Lee *et. al* [117] analyze the scale invariant property of the Dead Leaves model. They show that under the assumption $[r_{min}, r_{max}] \rightarrow [0, \infty]$ it is scale invariant. To analyze the behavior of the Dead Leaves model as a function of increasing viewing distance, $[r_{min}, r_{max}]$ is kept fixed and let $r = r_{max} - r_{min}$. An individual image contains objects of certain sizes.

Increasing the viewing distance involves two processes: smoothing and resolution reduction. The smoothing process is modeled by block averaging (using 2×2 blocks) and the resolution reduction is modelled by sub-sampling. This is similar to the classical image pyramid viewed from an image formation point of view [32]. By repeating the smoothing and sub-sampling procedure, images with increasing viewing distances will be generated. The viewing distance will be double in each iteration, and let s denote the iteration number.

Wu *et al.* study the statistical behavior of the Dead Leaves model using a fixed object size r distribution, with increasing viewing distance s. In the beginning, $s \ll r$ and the image contains large uniformly colored regions. As the distance increases, $s \approx r$. The average size of a leaf is roughly covering one pixel. The image contains small uniformly colored objects. As the viewing distance increases, $s \gg r$. And a pixel is the average of a large number of leaves. The visual appearance is close to white noise. By increasing the viewing distance the image content will transform from a rather large scale geometric structure to a highly stochastic appearance - from low entropy to high entropy. Wu *et al.* argue that different regimes are suitable for the different types of image content. Sparse representations such as wavelets are suitable for low entropy type of content, while Markov Random Field (MRF) is suitable in the high entropy case.

Yanulevskaya and Guesebroek [207] model the distribution of partial derivatives in natural images and image patches with the Weibull distribution (which essentially is the same as the generalized Laplacian distribution). They characterize the image content by the estimation of the parameters in the Weibull distribution. They propose three sub-models: the power law, the exponential and the Gaussian distribution. Akaike's information criterion (AIC) is used to determine an adequate sub-model. Images from the power law sub-model usually have a well-separated foreground and background. Furthermore, the background is often a rather uniform region. Images from the exponential sub-model, usually contain a lot of details at different scales. Images from the Gaussian sub-model, usually contain high frequency texture. Yanulevskaya and Guesebroek show that the image content for individual images can partly be explained by the parameters estimated in the Weibull distribution. Geusebroek and Smeulders used a similar approach to characterize stochastic textures [71, 72].

1.3.2 Collection Procedure and Content

We have collected a database of ensemble of image sequences containing the the same scene captured at different scales. The sequences contain natural images - both nature and man-made structure - viewed from a human

perspective.

The camera used to collect the database was a Nikon D40x, and 3 different objectives - $18-55$ mm, $55-200$ mm and $70-300$ mm - were used. The camera was placed on a tripod facing the scene. Each scene was captured at 15 different scales using different focal length - ranging from 18 mm to 300 mm or roughly 4 octaves. A 1×1 region in the least zoomed image corresponds to 16×16 region in the most zoomed image. The focal length was adjusted manually.

The image resolution was 2592×3872 pixels. The images were captured using Nikon's 14 bits raw format NEF, and were converted to 16 bits TIF images.

The database contains both man-made structures and natural scenes. The physical distance between the camera and the main object in the scene varies between 5 meters and a few kilometers. For most images, the distance between the camera and the main object in the scene is between 30 meters and 150 meters. The goal was that the sky should occupy a large region in the image at all scales in some sequences and the sky should be absent at all scales in other sequences.

The part of the scene present in all images in a sequence hs been extracted by use of registration techniques and by hand, resulting in a set of images containing the same part of the scene captured at different scales (different focal length). The resolution of the extracted regions ranges between 2592×3872 and 160×240 pixels.

1.3.3 Natural Image Statistics

Three classical results from natural image statistics are verified on the newly collected image database: the power spectra power law (scale invariance), the generalized Laplacian distribution of the partial distribution and the distribution of homogenous regions.

The statistics are estimated on three different 'sets':

- On the ensemble of images.
- On individual images.
- On all images captured at the same scale (same focal length setting).

Estimation on the ensemble of image sequences is performed to verify the soundness of the database which should be similar to previously reported estimation on other databases.

To analyze how far the visual appearance of an individual image is explained by the statistics, it was estimated on individual images.

Scale Invariance

One of the earliest result in natural image statistics is the apparent scale invariance [145, 94, 117, 58, 193, 168, 166, 166, 196].

The scale invariance can be stated as: the power spectra of an ensemble of image is following a power law in spatial frequencies given by

$$S(\omega) = \frac{A}{|\omega|^{2-\eta}}, \tag{1.16}$$

where $\omega = (\omega_x, \omega_y)$ is the spatial frequency, η is estimated on the ensemble and A is a constant that depends on the contrast in the ensemble of images. The power spectra power law can be formulated in the spatial domain using the correlation function [168, 94], and it has the following form

$$C(x) = C_1 + \frac{C_2}{|x|^{\eta}}, \tag{1.17}$$

where x is the distance between the pixels and C_1 and C_2 are constants. A large η implies that the intensities are less correlated, while a small η indicates higher correlation between the pixels. The intensity correlation decrease with the distances.

Ruderman and Bialek [166] reported $\eta = 0.19$ on their database collected in the woods. Ruderman [168] reported $\eta = -0.3$ for an ensemble of seashore images containing a lot of water and sky. Huang and Mumford [94] also reported $\eta = 0.19$ on the van Hateren database [197]. Lee *et al.* [117] estimated η on different types of environments; for an ensemble of images containing vegetation $\eta \approx 0.2$ and for an ensemble of images containing roads $\eta \approx 0.6$.

On our database, η is estimated to 0.202 on the ensemble of images. For individual images, η varies between -0.3 and 0.5. Images with a large η generally contain small scale details like texture and the distance to the main object in the scene is often small (often a few meters). Images with a small η contain large scale geometric structures and the distance to the main objects in the scene is rather large (often 100 meters or more). Hence, the estimation of η on individual images gives important information of statistical content of the image, especially whether η is large or small.

Estimations of η on all images captured at the same scale (focal length) show a tendency to increase as the viewing distance decreases. η increases rapidly for the first 3 capture scales. For the remaining capture scales, η has a tendency to increase but less rapidly and non-monotonically. As the viewing distance decreases, the region in the images occupied by the sky also has a tendency to decrease. The sky occupies a minor region in most of the

images after the first 3 capture scales. At a large viewing distance, buildings, lawns and trees appear to be rather uniform geometric structures, and as the viewing distance decreases, more details are brought out.

Laplacian Distribution of Partial Derivatives

It has been reported [172, 174, 173, 30, 207, 73, 72, 94, 183, 117, 126] that the distribution of partial derivatives of an ensemble of natural images can be modeled by a generalized Laplacien distribution

$$p(x) = \frac{1}{Z} e^{-|\frac{x}{s}|^{\alpha}}, \tag{1.18}$$

where α and s are estimated parameters. s is related to the width of the distribution (i.e. the variance) and α is related to the peakness of the distribution.

All computations are performed on a log-intensity scale (i.e. $log(I)$). Instead of using the intensity difference between two adjacent pixels - i.e. $log(I(x,y)) - log(I(x+1,y))$ - the normalized scale space derivatives are used. The partial scale space derivative in the x direction at scale t is

$$\frac{\partial}{\partial x}(G_t * I) = \frac{\partial G_t}{\partial x} * I, \tag{1.19}$$

where G_t is the gaussian function. The notation α_x denotes, estimation using the partial derivative in x direction.

Huang and Mumford [94] estimated α to 0.55 on the van Hateren [197].

On our database, α_x is estimated on the ensemble of images to 0.37 and 0.78 at scale $t = 1$ and $t = 64$, respectively. For individual images, α_x varies between 0.25 and 1.00 for $t = 1$ and between 0.55 and 2.00 for $t = 64$. The visual appearance of the images corresponds well with the estimation of α_x. Images with a large α_x contain small scale details, often high frequency texture. The t in the scale space derivative determines what is considered to be small scale details. The distance to the main object in the scene is often fairly small - a few meters. Images with a large α_x contain large scale geometric structures and the distance to the main objects in the scene is often large. (See also [207] and [73]).

The estimation of α_x on all images captured at the same scale (focal length) shows no clear tendency as the viewing distance decreases. Instead the estimation is rather stable with small variation over the capture scales.

Distribution of Homogenous Region

Alvarez *et. al* [2, 1, 3] and Gousseau *et. al* [76, 77] studied the distribution of homogenous regions in individual images and ensembles of natural images. They also relate the size distribution of homogenous regions to the question whether natural images belong to the function space of bounded variation (BV).

Following Alvarez *et al.* [2, 1, 3], the definition of homogenous regions is rather simple. (Gousseau *et. al* [77] use a different definition.) First, the intensity resolution is reduced to k levels such that each new intensity level contains the same number of locations. If I is a $N \times M$ image, then each new intensity level will contain $\frac{N \cdot M}{k}$ locations. A homogenous region is defined as a connected - using either 4 or 8 connectivity - set of locations with the same intensity. The size of a homogenous region is the number of locations it contains. They show that the size distribution of homogenous regions is following a power law

$$f(s) = \frac{A}{s^{\alpha}}, \tag{1.20}$$

where A and α are estimated parameters; α and A can be estimated using log-log regression. The size distribution is following a power law both on individual natural images and on ensembles of natural images. The estimation of α on individual images shows large variations. Images containing mainly small scale details have an $\alpha \approx 3$, while images containing mainly large scale geometric structures have an $\alpha \approx 1.6$. Alvarez *et al.* [2, 1, 3] reported α close to 2.0 on ensembles of natural images.

On our database, α was estimated to 2.11 on the ensembles of images. On individual images, α varied between 1.75 and 3.00. Images with small $\alpha \approx 1.75$ contain large scale geometric structures, and the distance to the main objects in the scene is rather large (hundreds of meters). Images with large $\alpha \approx 3.00$ contain small scale details, and the distance to the main objects in the scene is rather small (a few meters).

The estimation of α on all images captured at the same scale (focal length) shows no clear tendency as the viewing distance decreases. Instead the estimation is rather stable with small variation over the capture scales. The estimation using the different capture scales of α, varies between 2.15 and 2.22.

1.3.4 Discussion

Three classical and well-known statistical properties of natural images have been estimated on the newly collected ensemble of image sequences database. We study the statistical properties: the power spectra power law (scale invariance), the Laplacian distribution of the partial derivatives and the power law for size distribution of homogenous regions.

η in the power spectra power law - equation (1.16) - is estimated to 0.202 on the ensemble of images. Ruderman and Bialek [166] reported $\eta = 0.19$ on their database collected in the woods and Huang and Mumford [94] (also) reported $\eta = 0.19$ on the van Hateren database [197].

α in the generalized Laplacian distribution - equation 1.18 - is estimated to 0.37 and 0.78 at scale $t = 1$ respectively $t = 64$ on the ensemble of images. Huang and Mumford [94] estimated α to 0.55 on the van Hateren database.

α in the size distribution of homogenous region is estimated to 2.11 on the ensemble of images. Alvarez *et. al* [2, 1, 3] report α close to 2.

All three classical results could be verified on the new collected database.

The estimation on individual images in the database also verifies previous reported results. The estimation of η in the power spectra power law is large if the image mainly contains small scale details, and it is small if it mainly contains large scale geometric structures. The estimation of α in the generalized Laplacian distribution is large if the image mainly contains small scale details, and it is small if it mainly contains large scale geometric structures. The estimation of α in size distribution of homogenous regions is large if the image mainly contains small scale details, and it is small if the image mainly contains large scale geometric structures. The visual content is partly explained by the estimated statistical properties.

The estimation of η in the power spectra power law based on the capture scales increases as the viewing distance decreases. η increases rapidly for the first three capture scales and increases moderately for the remaining scales. The spatial layout - the spatial envelope - changes a lot at the first three capture scales. At the largest viewing distance, the sky occupies a large region in many of the images. As the viewing distance decreases, the region occupied by the sky shrinks and after the first three scales the sky occupies a small region in most images. As the viewing distance decreases, details are brought out. Sequences where the estimation of η is following a similar pattern as the capture scales based estimations often contain the sky at the large viewing distances. As the viewing distance decreases, the sky is absent or occupies a small region in the images. In sequences where η is large on all capture scales, the sky is often absent or occupying a small region, at all scales. Furthermore, the images contain small scale details and the viewing

distances are rather small at all capture scales. In sequences where η is small on all capture scales, the sky is occupying a large part of the image and the viewing distance is large at all capture scales. If the sequence contains a transition in distance, then the estimation of η has a tendency to increase with decreasing viewing distances. If the viewing distance in a sequence capturing a scene containing a building is between 100 and 6 meters, then the sequence contains a transition in distances. The spatial layout - spatial envelope - capturing a building from 100 meters is totally different from the spatial lay-out capturing a building from 6 meters. A picture captured from 100 meters is a distance photo, while that captured from 6 meter is a close-up photo. A sequence having a bush captured between 15 and 1 meters does not contain such a transition. The spatial layout does not change drastically over those viewing distances - both 15 meters and 1 meters are closeup photos. A sequence containing the sky and the ocean captured at a very large viewing distance does not contain such a transition. The spatial layout does not change over such viewing distances - all images are large distance or even panorama images.

Torralba and Oliva *et al.* [146, 192] use the estimation of η, as a function of orientation, in the power spectra power law to determine the distance to main objects in the scene. η depends on the spatial layout of the scene and the spatial layout of the scene constrained by the distance to the main object in the scene.

1.4 SVD as Content Descriptor

How can the image content be quantified in terms of geometric structures and texture? One approach is that the geometric structure can be described exactly using a sparse representation, while texture cannot be described exactly by a sparse representation.

Given an image I, a sparse approximation I_k should be constructed, where k is a sparseness/complexity parameter that measures the complexity of the approximation. As $k \to \infty$, $I_k \to I$ - i.e. as the sparseness decreases, the approximation gets closer to the observed image I.

1.4.1 Related Work

The 'approximation'-approach relates to the problem of finding the optimal base to represent the data. The book by Kirby [109] contains an overview of different best basis approaches - especially the SVD, PCA and wavelet bases.

One well-known and commonly used approach for representing data is the Principal Component Analyzis (PCA) - also known as the Karhunen-Loeven transformation or the Hotelling transformation. In PCA, the goal is to find an optimal orthonormal basis such that the variance of the data decreases as much as possible when an additional base vector is included. The PCA can be defined recursively in a natural way. The first normalized basis vector Ψ_1 is minimizing the variance of the data. An additional basis vector Ψ_{k+1} must be orthogonal to the previous basis vectors - i.e. $\langle \Psi_{k+1}, \Psi_i \rangle = 0$ for $i = 1, \cdots, k$ - and minimize the variance of the data. PCA is the optimal linear dimensionality reduction method in the mean square error sense [20]. Rather than finding the optimal basis for one observation, PCA is used for reducing the dimensionality of a set of observations.

Independent Component Analysis (ICA) was first formulated by Jutten and Herault in their seminal paper [102]. The concept of mutual independence is central in ICA. Let $X = (X_1, \cdots, X_n)$ be a set of stochastic variables and $p(X)$ be the joint probability distribution. The stochastic variables are independent if

$$p(X) = p(X_1, \cdots, X_n) = p(X_1) \cdot \cdots \cdot p(X_n), \tag{1.21}$$

where $p(X_i)$ is the marginal distribution for X_i. The objective in ICA is to find a transformation W such that

$$s = Wx, \tag{1.22}$$

where the components $s = \langle s_1, \cdots, s_n \rangle$ are as independent as possible (using some independence measure $F(s_1, \cdots, s_2)$). $x = \langle x_1, \cdots, x_n \rangle$ is a realization of X and x is generated using a linear model

$$x = As, \tag{1.23}$$

where A is the mixing matrix and s is the independent component. Given an x, the mixing matrix $A = W^{-1}$ and the independent component s should be found. See the tutorial on ICA by Hyvärinen [96] and Hyvärinen and Oja [97].

A well-known sparse representation was proposed by Olshausen and Field [149, 150], which relates to the models of the human visual front-end. An image is modeled as a linear superposition of (possibly) non-orthogonal basis functions $\phi_i(x, y)$

$$I(x, y) = \sum_i a_i \phi_i(x, y) \tag{1.24}$$

where ϕ_i form an over-complete basis for the image space and a_i are coefficients of the basis vectors. The a_i should be sparse, meaning that most of the a_i should be zero. The distribution $p(a)$ will be peaked at zero and will have 'heavy-tails'.

1.4.2 Optimal Rank Approximation and TSVD

One approach would be to approximate an image I in a lower dimensional subspace. An image is simpler if it can be approximated well in a subspace of low dimensionality, while an image is regarded as complex if a good approximation requires a subspace of dimensionality close to the dimension of the observed image. Let I_k be an approximation of I in a subspace of dimension k. As the dimension k increases towards the dimension of I, the approximation I_k gets closer to observed image I. Viewing images as matrices allows us to regard the dimensionality of subspaces as the matrix rank. The dimension of a matrix is the number of independent columns it contains or equally the dimension of the subspace spanned by the columns. This is captured by the rank of the matrix

$$Rank(A) = dim(span\{a_1, \cdots, a_n\}). \tag{1.25}$$

Given an image I with $rank(I) = k_0$, a rank k approximation I_k of I should be computed. The approximation I_k should be optimal in the sense that any other matrix B with rank k will have at least as large approximation error as I_k. Measuring the approximation error in terms of the 2-norm gives

$$I_k = \arg\min_{Rank(B)=k} \|A - B\|_2 . \tag{1.26}$$

The matrix B, with $Rank(B) = k$, that has the lowest approximation error in the 2-norm sense should be computed. The image residual $I - I_k$ contains the details that are suppressed in the approximation I_k, and $\|I - I_k\|_2$ is a measurement of the suppressed details.

Notice that any matrix A can be decomposed as

$$A = U\Sigma V^T \tag{1.27}$$

where U and V are orthogonal matrices, i.e. $UU^T = I$ and $VV^T = I$, where I is the identity matrix, and $diag(\Sigma) = (\sigma^1, \cdots, \sigma^n)$, where $\sigma^i \geq \sigma^{i+1} \geq 0$. This the well-known Singular Value Decomposition (SVD) [75, 109]; σ^i are called singular values, u^i left-singular vector and v^i right-singular vector. The set $\{\sigma^i, u^i, v^i\}$ is called the singular system of A. The rank of a matrix A is the number of singular values strictly larger than zero. Furthermore, the 2-norm is

$$\|A\|_2 = \sigma^1, \tag{1.28}$$

i.e. the largest singular value, and the squared Frobenious norm

$$\|A\|_F^2 = \sum_{i,j} a_{ij}^2 = \sum_i (\sigma^i)^2, \tag{1.29}$$

i.e. the sum of the squared singular values. The 2-norm is a vector induced norm defined as

$$\|A\|_2 = sup_{\|x\|_2 \neq 0} \left\{ \frac{\|Ax\|_2}{\|x\|_2} \right\}, \tag{1.30}$$

where the right-hand side is defined by the vector 2-norm, i.e. $\|x\|_2 = \sqrt{xx^T}$. The matrix 2-norm is an operator norm, which can be geometrically interpreted as how much A as a linear operator is scaling the vector x.
Let Σ_k be the matrix containing the k largest singular values on the diagonal, then the Truncated Singular Value Decomposition is defined as

$$A_k = U\Sigma_k V^T. \tag{1.31}$$

$Rank(A_k) = k$ and $Rank(A - A_k) = Rank(A) - k$. The 2-norm of the residual matrix is

$$\|A - A_k\|_2 = \sigma^{k+1} \tag{1.32}$$

and the squared Frobenious norm

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} (\sigma^i)^2. \tag{1.33}$$

The squared Frobenious norm corresponds to the Sum-of-Square Distance (SSD) between images often used for comparing images.

In fact, the TSVD approximation is the best rank k approximation in the sense that any other rank k approximation will have at least as large reconstruction error using either the 2-norm or the Frobenious norm. The TSVD is the solution to them minimization problem 1.26 (and it is also the solution if the Frobenious norm is used instead).

Damped Singular Value Decomposition (DSVD)

A common approach used in deblurring and denoising is to use a 'soft' threshold for the singular values. In TSVD, the k largest singular values are kept, while k+1 to Rank(A) singular values are set to zero. In Damped Singular Value Decomposition (DSVD), all singular values are damped using filter factors defined as

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \tag{1.34}$$

where λ is a problem dependent regularization parameter. The filtered singular values are $\phi_i = f_i \sigma_i$. TSVD can also be formulated using filter factors and the filter factors are

$$f_i = \begin{cases} 1 & \text{if } i \leq k \\ 0 & \text{if } i > k \end{cases}$$

.

The DSVD is a 'soft' threshold in the sense that the large singular values are kept almost unchanged and the small singular values are almost zero after filtering. This is because

$$f_i \approx \begin{cases} 0 & \text{if } \sigma_i \ll \lambda \\ 1 & \text{if } \lambda \ll \sigma_i \end{cases}$$

.

DSVD is related to the solution of Tikhonov regularization problems [142, 103, 87, 89, 88, 55]. The DSVD is the solution to the following Tikhonov regularization problem

$$u_\lambda = \arg\min_u \left\{ \|u_0 - u\|_2^2 + \lambda^2 \|u\|_2^2 \right\} \tag{1.35}$$

commonly studied in denoising and inverse problem.

1.4.3 Measuring the Complexity of Images - Singular Value Reconstruction Index

An image is considered to be simple if it can be approximated well in a subspace of low dimensionality, and it is considered to be complex if it can only be approximated well in a subspace of high dimensionality. Given an image I, an approximation I_k should be constructed such that the reconstruction error is smaller than σ^{err}. We define the complexity of the image as the lowest dimension of the subspace in which the approximation I_k can be constructed

$$\min \{k \ : \ \|A - A_k\|_2 \leq \sigma^{err}\} . \tag{1.36}$$

This is termed the Singular Value Reconstruction Index (SVRI) at level σ^{err} and tells how many singular values are required for an approximation with an error smaller than σ^{err}. First, the error level is determined - how well should the approximation fit the original image - then the number of singular values are determined (i.e. the dimension of subspace).

Another definition is: Given a subspace of dimension k, how well can the observed image be approximated? And use the 2-norm or the Frobenious norm of the residual image as the complexity measure.

Furthermore, an image is composed of image patches. We assume that the complexity of an image should be determined by the complexity of the patches that constitute the image. Rather than computing the global singular value reconstruction index at level σ^{err}, the complexity for each patch constituting the image is computed and the mean complexity of those patches gives the image complexity.

1.4.4 Discussion

In figure 1.4, the SVRI are shown as a filter applied to images of 100×100 pixels, containing the same scene captured at different viewing distances.

In figure 1.5, images with low/high SVRI value are shown. The top row shows images with low SVRI. The sky is covering a large part of the images. Furthermore, the distance to the main objects in the scenes are large, resulting in large scale geometric structures - such as buildings - with sharp boundaries. From this, we get the indication that images with a low SVRI mainly contain geometric structures.

In the second row of figure 1.5, images with high SVRI are shown. The images contain small scale details such as leaves and twigs, and the distance

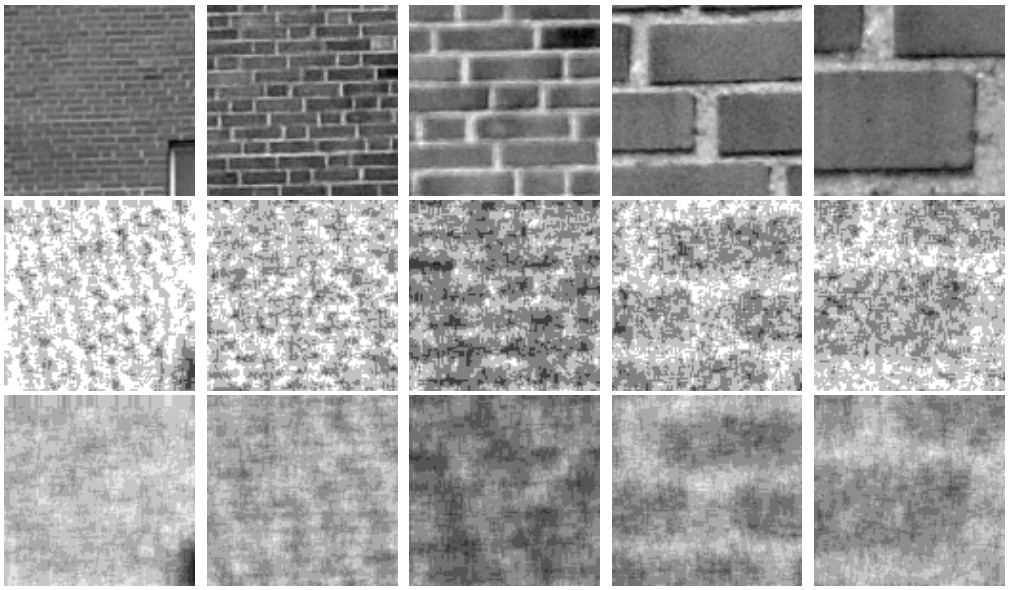


Figure 1.4: Example of SVRI used as a filter on 100×100 images containing the same scene captured at 5 different viewing distances. The top row contains the images, the second and third rows contain SVRI filter using patch size 5 and $\sigma^{err} = 0.05$ respective patch size 10×10 and $\sigma^{err} = 0.1$

to the main objects in the scene is rather small. Furthermore, the sky is absent (or covers a small region in the images) in all of the images. Hence, from this, we get the indication that images with high SVRI contain mainly texture.

Rank Distribution in Natural Images

Measuring the image complexity using TSVD and the proposed SVRI depends on the rank distribution and on the distribution of the singular values in natural image patches. Furthermore, the image content in natural image patches should depend on the size of the smaller singular values. The patch content should be different if the smaller singular values are small or large.

1000 randomly selected 25×25 image patches were selected from each image in the DIKU Multi-Scale Image database. (An experiment using 50×50 image patches gave similar results.) The singular values for each of the roughly 800,000 patches were computed. The first conclusion based on the experiment is that image patches in natural images are almost always of full rank - i.e. in the experiment, $\sigma_{25} > 0$ in all patches.

The condition number for a $n \times n$ matrix A is defined as

$$Cond(A) = \frac{\sigma_1}{\sigma_n} \tag{1.37}$$

and it measures how well-conditioned the matrix is [75]. A large condition number indicates that that matrix is ill-conditioned and that the columns are almost linearly dependent. For natural image patches, the condition number is always finite, because the patches are of full rank - i.e. $\sigma_{25} > 0$. The condition number is large, which indicates that columns are almost linearly dependent.

The distribution of σ_1 has a large variance and is not very peaked around the mode. The distribution of the σ_i's for $i > 1$ follows the same basic form. The distributions are very peaked at zero, which indicates that most singular values are very small. Still, the distributions have 'heavy tails' i.e. values (relatively) far from zero.

Visually comparing patches with large σ_n with patches of small σ_n clearly indicates a large difference in image content. Patches with a large σ_n contain small scale details, while patches with small σ_n contain geometric structure.

Figure 1.5: Example of images with low/high singular value reconstruction index (SVRI) at level 0.01 using 15×15 patches. The top row contains images with low SVRI, all images contain the sky and the viewing distances are rather large. The bottom row shows examples of images with high SVRI; all images contain small scale details and the viewing distances are small.

1.5 Image Description by Regularization

The problem of measuring and quantifying the image content in terms of geometric structure and texture can be approached in many ways. In section 1.4 and chapter 4, a matrix approach using TSVD is proposed, based on the property that the TSVD approximation is the optimal rank k estimation. In the following section and chapter 5, image regularization and decomposition in a continuous setting is used for measuring the image content.

Image regularization can be viewed as approximating an observed image u_0 with a simpler image u, called the regularized solution, such that some energy functional is minimized. Most energy functionals are composed of two terms, a data fidelity term and a regularization term. The simplicity of u is defined and determined by the regularization term and a regularization parameter λ. As the regularization parameter increases, the regularized solution gets simpler in the sense defined by the regularization term. To stress the dependence on the regularization parameter λ, u_λ will sometimes be used instead of u. The residual image (also called the image residual) is the difference between the observed image and the approximated image - $(u_0 - u_\lambda)$ - and it contains the details that were removed in the approximation.

Image regularization can also be viewed as decomposing the observed image u_0 into two components: a geometric and a texture component. By measuring the content in two components, the image content can be quantified in terms of geometric structure and texture. We analyze the squared L^2 norm of the regularized solution and the residual image as a function of the regularization parameter λ.

The L^2 norm of the regularized solution and the residual has been studied in connection with parameter selection in denoising, but not for describing the image content in terms of geometric structure and texture.

1.5.1 Related Work

Characterizing the image content by analyzing the norm of the scale space representation of the image as a function of scale/regularization parameter has not received a lot of research attention. The behavior of the norm as a function of scale/regularization parameter has been studied in denoising, for optimal parameter selection. Thompson *et. al* [189] contains a classical study of parameter selection in denoising.

Sporring [180] and Sporring and Weickert [181, 182] view images as distribution of light quanta and use information theory to study the image content in scale spaces. They show that the entropy of an image is an increasing function of the scale (in scale space). Empirically, they show that the derivative

of the entropy with respect to the scale is a good texture descriptor.

One of the oldest, and still often used, optimal parameter selection method in denoising is the *Morozov discrepancy principle* (or *discrepancy principle*) [138, 87, 198]. The noise is assumed to be additive

$$u_0 = u + e, \tag{1.38}$$

where u is the 'clean' image and e is the noise. Furthermore, the norm of the noise, $\|e\| = \varepsilon$, must be known or possible to estimate. The parameter should be selected such that

$$\|e\| = \|u_0 - u_\lambda\| = \varepsilon, \tag{1.39}$$

i.e. the residual norm should be equal to the norm of the noise.

Another common method for determining the optimal parameter in denoising is the L-curve studied by Hansen [89, 87, 88, 198]. The L-curve is a log-log plot of the norm of the regularized solution against the norm of the corresponding residual. It shows the trade-off between the size of the solution and the size of the residual, using suitable norms. In the log-log scale, it has a L-shape. According to the *L-curve criterion*, the optimal value for the parameter λ is the one with highest curvature (i.e. the corner of the L). The L-curve is related to the less formal trade-off curve discussed in [22].

Buades *et.al* [27] introduced the concept of 'Method Noise' for evaluating denoising methods. (See also [28, 29, 26].) The 'Method Noise' is simply the difference between the original image and the denoised image - i.e. the image residual. Optimally the image residual should only contain noise and no structure after denoising. The noise and only the noise has been suppressed in the denoising. For example, denoising a noise 'free' image should optimally result in an empty residual, while denoising an image corrupted by additive independent Guassian white noise should result in a residual containing Gaussian white noise. Furthermore, the residual should not contain any structure not caused by the noise. 'Method Noise' aims to characterize the denoising methods by analyzing the content in the residual image. While 'Method Noise' aims to characterize the behavior of the method by analyzing the residual, our aim is to characterize the image by analyzing the residual norm as a function of the regularization parameter. Our aim is, in some sense, complementary.

1.5.2 Image Decomposition

In image decomposition, an observed image u_0 is considered to be composed of two components: a smooth/geometric component and a noise/texture

component. Formally, an image decomposition may be written as

$$u_0 = u + v, \tag{1.40}$$

where u_0 is the observed image, u is the smooth/geometric component and v is the texture. Often, we are interested in the geometric component u, which can be found by minimizing a suitable energy functional

$$E(u;\lambda) = \int (u_0 - u)^2 dx + \lambda \int \Psi(Du) dx \tag{1.41}$$

Here λ is a problem dependent regularization parameter, D is a linear operator (often a differential operator) $\Psi(x)$ is a 'penalty' function (often absolute or squared absolute value). The noise/texture component v is also known as the residual (image), because $v = u_0 - u$, i.e. the difference between the observed image and the regularized solution, and it contains the details that have been suppressed in the regularization. Regularization in computer vision and image processing has a long history and is used for transforming ill-posed problems into well-posed problems [18, 159, 158], and for denoising. The energy functional is composed of two terms, a data term and a regularization term. The data term forces the solution u to be close in the L_2 sense to the observed data u_0, while the regularization term forces u to be smooth in the sense defined by the operator D.

In [83] the squared L_2-norm of the regularized solution and the residual using three regularization methods were used to analyze and characterize the image content. The regularization methods include first order Tikhonov regularization

$$E(u;\lambda) = \int (u_0 - u)^2 + \lambda |\nabla u|^2 dx, \tag{1.42}$$

Linear Guassian Scale Space [98, 202, 111, 122]

$$u_\lambda = u_0 * G_\lambda, \tag{1.43}$$

where * denotes the convolution operator and G_λ is the gaussian function with variance λ. Linear gaussian scale space is equivalent with an infinity order Tikhonov regularization [143], but it is more intuitive to use the convolution formulation.

Finally, the Total Variation Image Decomposition is also studied

$$E(u;\lambda) = \int (u - u_0)^2 + \lambda |\nabla u| dx. \tag{1.44}$$

1.5.3 The Bayesian Approach and MAP-Solution

The bayseian approach allows us a statistical interpretation to energy minimization [140]. The Bayes' rule and the MAP-solution were introduced in section 1.1.1. For the statistical interpretation, it is assumed that the pure signal u has been corrupted by additive gaussian white noise, resulting in a observed image u_0. The pure signal u should be recovered from the observed image u_0.

The additive gaussian white noise assumption gives $v = u_0 - u \in N(0, \sigma^2)$, and

$$p(u_0(x_0)|u(x_0)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u_0(x_0)-u(x_0))^2}{2\sigma^2}}. \tag{1.45}$$

Assuming that the pixel noise is independent, we have

$$p(u_0|u) = C_1 e^{-\sum_{x\in D} \frac{-(u_0(x)-u(x))^2}{2\sigma^2}}. \tag{1.46}$$

The prior term is harder to model and more assumptions are required. For smooth images without texture, it is feasible to assume small intensity variation. One may assume that $|\nabla u|$ is following a zero mean normal distribution with variance μ, which gives

$$p(u) = C_2 e^{-\sum_{x\in D} \frac{|\nabla u(x)|^2}{2\mu^2}}. \tag{1.47}$$

Another assumption would be that $|\nabla u|$ is following a Laplacian distribution, which gives

$$p(u) = C_2 e^{-\sum_{x\in D} \frac{|\nabla u(x)|}{2\mu^2}}. \tag{1.48}$$

The distribution of partial derivatives in natural images can be modeled with the generalized Laplacian distribution [184, 117, 126, 207]. Inserting the estimation in to the Bayes formulation gives

$$u_{map} = \text{ arg max } \{C_1 e^{-\sum_{x\in D} \frac{(u_0(x)-u(x))^2}{2\sigma^2}} C_2 e^{-\sum_{x\in D} \frac{|\nabla u(x)|^2}{2\mu^2}}\}, \tag{1.49}$$

where the prior term is estimated by assuming that the gradient magnitude is following a gaussian distribution. By taking the negative log ($-log$) of the MAP-solution, one can get rid of the exponential and the maximization problem turns into a minimizing problem, given by

$$E(u) = \sum_{x\in D} \frac{(u_0(x) - u(x))^2}{2\sigma^2} + \sum_{x\in D} \frac{|\nabla u(x)|^2}{2\mu^2}. \tag{1.50}$$

Switching to the continuous domain and renaming the parameters gives

$$E(u) = \int_D (u_0(x) - u(x))^2 dx + \lambda \int_D |\nabla u(x)|^2 dx, \tag{1.51}$$

which corresponds to the first order Tikhonov regularization energy functional. Instead, using the assumption that the gradient magnitude is following a Laplacian distribution as in equation (1.48) leads to the total variation image decomposition functional.

Using the Bayesian formulation, we see that the first order Tikhonov regularization and the total variation image decomposition are MAP solutions under different assumptions about the distribution of the prior term.

1.5.4 Regularized and Residual Norm

To analyze the image content with respect to geometric structure and texture, the squared L^2 norm of the regularized solution u_λ

$$s(\lambda) = \|u_\lambda\|_2^2 \tag{1.52}$$

and the squared residual norm

$$r(\lambda) = \|v_\lambda\|_2^2 = \|u_0 - u_\lambda\|_2^2 \tag{1.53}$$

as a function of the regularization parameter, are studied. Of interest is also the corresponding derivatives with respect to the regularization parameter λ. The derivative with respect to λ reveals the rate in which details are suppressed.

The normalized norm of the regularized solution is defined as

$$s_{norm}(\lambda) = \frac{\|u_\lambda\|_2^2}{\|u_\lambda\|_2^2 + \|v_\lambda\|_2^2} \tag{1.54}$$

and the normalized residual norm is defined as

$$r_{norm}(\lambda) = \frac{\|v_\lambda|_2^2}{\|u_\lambda\|_2^2 + \|v_\lambda|_2^2}. \tag{1.55}$$

The derivative of the normalized norm with respect to λ reveals the rate in which details are suppressed as the regularization parameter increases.

By the triangle inequality we have $\|u_0\|_2^2 \leq \|u\|_2^2 + \|v\|_2^2$. $t(\lambda) = \|u_\lambda\|_2^2 + \|v_\lambda\|_2^2$ will denote the total norm. The total norm $t(\lambda)$ is not constant, instead, it depends on the parameter λ. $t(0) = \|u_0\|_2^2$ and $t(\infty) = \|C\|_2^2 + \|u_0 - C\|_2^2$, where C is the mean intensity in the image. Normalizing the initial image

u_0 by subtracting the mean value C gives a simpler form for the limit case $t(\infty) = \|u_0\|_2^2$.

The sum of the two norms is one, i.e. $s_{norm}(\lambda) + r_{norm}(\lambda) = 1$. The normalized regularized norm $s_{norm}(\lambda)$ can be viewed as the degree of the total norm that is explained by the regularized solution, and the normalized residual norm $r_{norm}(\lambda)$ as the degree of the total norm that is explained by the texture component.

The squared L^2 norm of the regularized solution and the residual as a function of λ were studied in terms of convexity/concavity using three regularization methods. We show that for first order Tikhonov regularization, $s(\lambda)$ is a monotonically decreasing convex function, while $r(\lambda)$ is a monotonically increasing function but $r(\lambda)$ is neither concave nor convex. The same holds for the linear gaussian scale space if the parameter in the gaussian function is the variance, but fails when the parameter is the standard deviation. Empirically, we show that the squared L^2-norm of the residual using TV is an increasing non-concave function.

1.5.5 Discussion

We attempt to characterize the image contents in terms of geometric structure and texture using regularization. Image regularization can be viewed as approximating a given image with a simpler one. Analyzing and measuring the content that is preserved after regularization and the content that is removed can help in characterizing the original image content in terms of geometry and texture. Measure the content that is kept - i.e. the regularized solution - and the content that is removed - i.e. the residual - can be used to characterize content. Image regularization can also be viewed as image decomposition - the image is decomposed into a geometric component and a texture component. Again, the image content can be characterized by measuring the content in the components.

Buades *et.al*[27] used the content in the residual - termed 'Method Noise' - for evaluating denoising methods. They try to characterize the denoising methods by the content in the residual. Our goal is complementary: how to characterize the image by the content in its residual (using some regularization method).

1.6 Motion Estimation by Contour Registration

Image registration is the process of spatially overlaying two or more images containing the same objects. The images may contain a scene at different times or from different viewpoints, or they may contain the same objects or scene, but captured using different sensors (different modality). Image registration is a fundamental problem in image processing and is a crucial intermediate step in many applications. Some common applications are preprocessing for image classification [41, 206, 95], image stitching/mosaicing [50, 187, 178] and sensor fusion in medical applications [160, 164, 125, 127, 203].

The registration problem has been approached in a number of ways. One approach has been to find a transformation that overlays the images such that the sum of intensity differences is small. This approach is often called a direct method because the image intensities are used directly. Another approach has been to find interesting points such as corners and edges in the images and then find correspondences between these points. The transformation that overlays the images is then found by using the correspondence between the points. This approach is called feature-based and only a sparse representation - the interesting points - of the images is used to determine the transformation.

There is also a close relation between motion estimation and registration [92]. Registration of images in a time sequence - temporal registration - may be regarded as a motion estimation problem. Optical flow [93, 13, 25, 152] is often used for estimation of the apparent motion in a sequence and it has some similarities with the direct registration approach.

The problem addressed in this thesis has some similarities with image registration, contour registration, and motion estimation, but it is still quite different. How can the motion of a deformable moving object, such as a walking person or running horse, be estimated solely by the knowledge of the boundary of the object as seen in different images? The motion of the interior of the moving object should be estimated solely based on the knowledge of the boundary.

Let Γ_1 be a closed curve embedded in an image I_1 and Γ_2 another closed curve embedded in another image I_2. The problem is to find a geometric transformation Φ that overlays the two images such that Γ_1 is mapped on to Γ_2; the interior of Γ_1 should be transformed in a reasonable way and the transformation should be the "simplest" possible. The motion of the contour and the motion of the interior must be consistent and computed

simultaneously.

To add some intuition, one can think of the two closed contours as the boundary of a deformable moving object in a sequence and the problem is to simultaneously compute the motion of the contour and estimate the motion of the interior of the object. The motion of the entire object should be computed solely based on the contour. For example, the motion of the nose should be estimated solely based on the contours of the head in two images.

This can be useful when the boundaries of the object are available, but the image contents are not reliable. The boundary of the object may be available because shape priors are used in the segmentation process and the image contents are not reliable because the object has been occluded or the image contents may have been lost.

1.6.1 Image and Contour Registration

The general image registration problem may be defined as:

Definition 1 ***The Image Registration Problem***
Given two images T_1 and T_2 and a distance measure $D(T_1, T_2)$, that measures the difference between two images, find a geometric transformation $\Phi : R^2 \rightarrow R^2$ that minimizes $D(T_1, \Phi(T_2))$.

The registration problem has been approached in many different ways. See Browns' [24] rather old survey and Zitova and Flussers' survey [215].

Direct Method

In the direct approach towards the registration problem, the image intensities are directly used in the distance measure [99]. A geometric transformation Φ that minimizes the distance between the intensities of the image T_1 and the transformed image $\Phi(T_2)$ should be found. One common distance measure is the sum of squared distance

$$D^{SSD}(T_1, T_2) = \frac{1}{2} \parallel T_1 - T_2 \parallel^2_{L_2} = \int (T_1(x) - T_2(x))^2 dx, \tag{1.56}$$

and a geometric transformation Φ that minimizes $D^{SSD}(T_1, \Phi(T_2))$ should be found (see [136, 24] for some approaches). Often, Φ is a parametric transformation with parameters a and the problem is to find the optimal parameters for the transformation Φ_a. One example of a parametric geometric transformation is the intensity-based affine linear transformation.

The direct (or intensity-based) method is in general ill-posed in the sense that a small change in the input images may give a completely different transformation. By adding an additional smoothness (or regularization) term that penalizes certain transformations the problem becomes well-posed. The transformation Φ is a minimizer of the functional

$$E(\Phi) = D(T_1, \Phi(T_2)) + \alpha S(\Phi), \tag{1.57}$$

where $D(T_1, T_2)$ is the distance measure, $\alpha > 0$ is a positive smoothness parameter and $S(\Phi)$ is a smoothing term.

Feature-Based Method

In the feature-based approach, a number of feature points - also called control points, interest points and landmarks - are extracted from the images. A correspondence between the feature points detected in the two images is established and some feature points may be discarded. The correspondence between the feature points is used for finding a geometric transformation.(See e.g. [191, 136]).
Good feature points should be stable over time, spread over the whole image and efficiently detectable. Such features are not present in all types of images. Common and well-suited feature points are corners, edges and line intersections. A region can be represented as a feature point by the center of gravity and line segments can be represented as feature points by the two endpoints or the middle point. Common feature detection methods are the HARRIS detector [90], the scale invariant HARRIS detector [134], SUSAN [176, 175, 177] and SIFT [124]. (See also [135] and [133] for an evaluation of feature detectors.)

Given two sets of control points from two images, a correspondence between the control points should be established. One approach for establishing a correspondence between the set of control points is to look at the spatial relations. Another approach is to compute a descriptor locally around the control point and use that for establishing a correspondence. The simplest descriptor is image intensities locally around the control points, however, some form of filter responses is often used.

After a correspondence between the control points has been established, a geometric transformation that overlays the images should be constructed. The geometric transformation should be constructed such that the corresponding feature points are overlaid. Global linear geometric transformations such as similarity and affine transformation are common transformations. Non-parametric feature-based geometric transformations are also common, such as elastic and fluid-based registration.

Dense Contour Registration

Contour registration is a fundamental problem in image processing, with a lot of applications within shape analysis. In the contour registration problem, two contours Γ_1 and Γ_2 (i.e. object boundaries) are given and a transformation that overlays the contours should be found. Contour registration is a mapping between two contours and does not in general give an image registration.

One dense based approach to the curve matching problem is to minimize the "elastic energy" that is required to transform one curve into the other [208, 14, 74]. The curves are usually represented as simple connected parametric curves. Let Γ_1 be parameterized by the arch-length s. Let $t(s)$ be a function mapping arch-length to arch-length, $t(s)$ is a correspondence between Γ_1 and Γ_2. $\Gamma_1(s)$ is mapped to $\Gamma_2(t(s))$. Associated with each continuous correspondence function $t(s)$ is a cost - the elastic cost of the correspondence function $t(s)$. Given the two contours Γ_1 and Γ_2 and a fixed correspondence function $t(s)$, the cost function measures the "elastic energy" that $t(s)$ requires to transform Γ_1 into Γ_2. The cost function C is defined by

$$C(\Gamma_1, \Gamma_2, t(s)) = \int_{\Gamma_1} F(\Gamma_1, \Gamma_2, t(s))ds, \tag{1.58}$$

where the function F measures the "elastic" properties. The distance between two contours is then the minimum "elastic energy" over all correspondence functions $t(s)$

$$D(\Gamma_1, \Gamma_2) = \ min \ \left\{ \int_{\Gamma_1} F(\Gamma_1, \Gamma_2, t(s))ds \right\}. \tag{1.59}$$

Given the distance measure D between two contours, the contour registration problem becomes

$$R(\Gamma_1, \Gamma_2) = \ \arg\min_{t(s)} \ \left\{ \int_{\Gamma_1} F(\Gamma_1, \Gamma_2, t(s))ds \right\}. \tag{1.60}$$

The function F models the elastic properties and can depend on the physical properties of the subject being studied. F can depend on other curve properties such as the first derivative $\dot{\Gamma}$ and the curvature $|\ddot{\Gamma}|$.

This approach minimizes the elastic energy exactly on the contours. The cost for deforming the contour is explicitly formalized on the contour. Minimizing the elastic energy of deforming a contour gives an implicit cost of deforming the interior of the contour.

The relation between shape similarity measures and contour registration is very close. For example, the minimum 'elastic energy' that is required to

transform one curve into the other can be considered to be a shape similarity measure. In registration, the objective is to find the contour transformation, while for shape similarity measures the interest is the cost of the transformation.

1.6.2 Image Registration by Contour Matching

By using shape priors in the segmentation, good object boundaries can be found, even if the object boundary is occluded or the image contents have been destroyed at the boundary. Let $F_1, \cdots, F_n$ be an image sequence containing a non-rigid moving object that should be segmented. In some frames, the image contents may not be reliable either because the object is occluded or because the image contents are missing. The object boundary can still be found in many cases by using shape priors in the segmentation.(See e.g. [64, 44, 45, 46]). Often, one also wants to estimate the motion of the object between the frames. Because the image contents inside the object are missing, neither a direct registration approach nor a feature-based approach can be used directly. Furthermore, it is not the motion of the contour that should be computed; instead, it is the motion of the interior of the object that should be computed solely based on the object boundaries. The motion of the object should be computed based on the assumption that

- The object boundary is correct.
- Part of the image contents is not reliable.

A variational formulation of this problem is presented, which simultaneously computes a displacement field for the contour and interpolates the motion of the interior of the object. A good motion estimation should overlay the two boundaries in such a way that the motion of the interior is interpolated in a consistent way.

A Variational Approach

In this section, we are going to present a variational solution to the following contour matching problem: Suppose we have two simple closed curves Γ_1 and Γ_2 contained in the image domain Ω. Find the "most economical" mapping $\Phi = \Phi(x) : \Omega \to \mathbf{R}^2$ such that Φ maps Γ_1 onto Γ_2, i.e. $\phi(\Gamma_1) = \Gamma_2$.

The latter condition is to be understood in the sense that if $\alpha = \alpha(s) : [0, 1] \to \Omega$ is a positively oriented parametrization of Γ_1, then $\beta(s) = \Phi(\alpha(s)) : [0, 1] \to \Omega$ is a positively oriented parametrization of Γ_2 (allowing some parts of Γ_2 to be covered multiple times).

To present our variational solution of this problem, let $\mathcal{M}$ denote the set of twice differential mappings Φ, which maps Γ_1 to Γ_2 in the above sense. Let

$$\mathcal{M} = \{\Phi \in C^2(\Omega; \mathbf{R}^2) \,|\, \Phi(\Gamma_1) = \Gamma_2\}. \tag{1.61}$$

Moreover, given a mapping $\Phi : \Omega \to \mathbf{R}^2$, not necessarily a member of $\mathcal{M}$, then we express Φ in the form $\Phi(x) = x + U(x)$, where the vector valued function $U = U(x) : \Omega \to \mathbf{R}^2$ is called the *displacement field associated with* Φ, or simply the displacement field. It is sometimes necessary to write out the components of the displacement field; $U(x) = (u_1(x), u_2(x))^T$.

We now define the "most economical" map to be the member Φ^* of $\mathcal{M}$, which minimizes the following energy functional:

$$E[\Phi] = \frac{1}{2} \int_\Omega \|DU(x)\|_F^2 \, d\mathbf{x}, \tag{1.62}$$

where $\|DU(x)\|_F$ denotes the Frobenius norm of $DU(x) = [\nabla u_1(x), \nabla u_2(x)]^T$, which for an arbitrary matrix $A \in \mathbf{R}^{2\times 2}$ is defined by $\|A\|_F^2 = \mathrm{tr}(A^T A)$. The optimal transformation is given by

$$\Phi^* = \arg\min_{\Phi \in \mathcal{M}} E[\Phi]. \tag{1.63}$$

Using that $E[\Phi]$ can be written in the form

$$E[\Phi] = \frac{1}{2} \int_\Omega |\nabla u_1(x)|^2 + |\nabla u_2(x)|^2 \, d\mathbf{x}, \tag{1.64}$$

it can be seen that the Gâteaux derivative [170, 68, 6] of $E[\Phi]$ is given by

$$\begin{aligned} dE[\Phi; V] &= \int_\Omega \nabla u_1(x) \cdot \nabla v_1(x) + \nabla u_2(x) \cdot \nabla v_2(x) \, d\mathbf{x} \\ &= \int_\Omega \mathrm{tr}(DU(x)^T DV(x)) \, d\mathbf{x}, \end{aligned}$$

for any displacement field $V(x) = (v_1(x), v_2(x))^T$. After integration by parts, we find that the necessary condition for $\Phi^*(x) = x + U^*(x)$ to be a solution of the minimization problem (1.63) takes the form

$$0 = -\int_\Omega \Delta U^*(x) \cdot V(x) \, d\mathbf{x}, \tag{1.65}$$

for any *admissible* displacement field variation $V = V(x)$. Here $\Delta U^*(x) = (\Delta u_1^*(x), \Delta u_2^*(x))^T$ is the Laplacian of the vector valued function $U^* = U^*(x)$.

Since every admissible mapping Φ must map the initial contour Γ_1 onto the target contour Γ_2, it can be shown that any displacement field variation V must satisfy

$$V(x) \cdot \mathbf{n}_{\Gamma_2}(x + U^*(X)) = 0 \quad \text{for all } x \in \Gamma_1. \tag{1.66}$$

Notice that this condition only has to be satisfied precisely on the curve Γ_1, and that $V = V(x)$ is allowed to vary freely away from the initial contour. The interpretation of the above condition is that the displacement field variation at $x \in \Gamma_1$ must be tangent to the target contour Γ_2 at the point $y = \Phi(x)$. In view of this interpretation of (1.66), it is not difficult to see that the necessary condition (1.65) implies that the solution Φ^* of the minimization problem (1.63) must satisfy the following Euler-Lagrange equation:

$$0 = \begin{cases} \Delta U^* - (\Delta U^* \cdot \hat{\mathbf{n}}_{\Gamma_2})\, \hat{\mathbf{n}}_{\Gamma_2}, & \text{on } \Gamma_1, \\ \Delta U^*, & \text{otherwise,} \end{cases} \tag{1.67}$$

where $\hat{\mathbf{n}}^*_{\Gamma_2}(x) = \mathbf{n}_{\Gamma_2}(x + U^*(x))$, $x \in \Gamma_1$, is the pullback of the normal field of the target contour Γ_2 to the initial contour Γ_1. The standard way of solving (1.67) is to use the gradient descent method: Let $U = U(t, x)$ be the time-dependent displacement field which solves the evolution PDE

$$\frac{\partial U}{\partial t} = \begin{cases} \Delta U - (\Delta U \cdot \hat{\mathbf{n}}^*_{\Gamma_2})\, \hat{\mathbf{n}}^*_{\Gamma_2}, & \text{on } \Gamma_1, \\ \Delta U, & \text{otherwise,} \end{cases} \tag{1.68}$$

where the initial displacement $U(0, x) = U_0(x) \in \mathcal{M}$ specified by the user, and $U = 0$ on $\partial\Omega$, the boundary of Ω (Dirichlet boundary condition). Then $U^*(x) = \lim_{t\to\infty} U(t, x)$ is a solution of the Euler-Lagrange equation (1.67).

The PDE (1.68) coincides with the so-called *geometry-constrained diffusion* introduced by Andresen and Nielsen in [5]. Thus we have derived the energy functional that *geometry-constrained diffusion* is minimizing.

1.6.3 Relation to Feature-Based and Contour Registration

In our approach, image F_1 contains curve Γ_1 and image F_2 contains curve Γ_2, and an image registration - $\Phi(x)$ - that overlays the two contours and minimizes the functional (1.62) should be found.

The preceding segmentation can be a viewed as a feature extraction step. In the segmentation step, an accurate object boundary is extracted and it is viewed as an continuous planar curve. Feature points are not allocated along

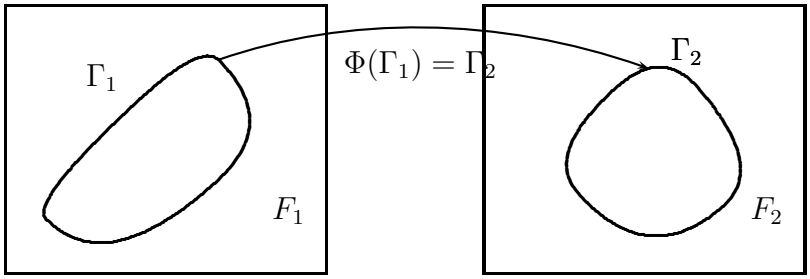


Figure 1.6: Given two closed curves Γ_1 and Γ_2 contained in two images F_1 and F_2, Φ maps F_1 onto F_2 such that Γ_1 is mapped onto Γ_2 (i.e. $\Phi(\Gamma_1) = \Gamma_2$).

the boundary; instead, the dense curve is used directly to determine the image registration. Instead of extracting control points on the curves - such as points with high curvature and zero crossings of the curvature - and then a correspondence between the control points, a continuous transformation that overlays the dense contours in the image domain is found. In the feature-based approach, a registration should be found using the correspondence between a discrete set of feature points; in our approach, a registration should be found based on a continuous feature set. The constraint on the geometric transformation is the mapping between the contours.

The segmentation of the object represents the feature extraction step. Given the segmentation - i.e. a continuous set of features, the feature correspondence step and the geometric transformation step are solved simultaneously. The dense correspondence between the contours restricts the set of possible transformations.

1.6.4 Applications

The contour-based motion estimation was combined with shape prior segmentation in image sequences. By using the previous segmentation as shape prior, accurate object segmentation was possible even if part of the object was missing or occluded. The contours of the object in two adjacent frames were used for estimating the displacement field. The intensity in the second frame can be predicted by applying the displacement field. This is temporal inpainting or transport of intensities between frames. By comparing the predicted intensity with the observed intensity, object occlusion can be detected. If the difference between the predicted and observed intensity is large, then the object is occluded. Temporal inpainting using solely the contour has

been used for texturing objects in image sequences [190].

1.6.5 Discussion

The estimation of the displacement field - the deformation and motion - of a non-rigid moving object solely based on the object boundary relates both to feature-based image registration and contour registration. The contours can be viewed as continuous sets of features that should be mapped onto each other. Simultaneously, the deformation field for the interior of the object should be estimated. From a contour registration point of view, the contours should be mapped onto each other, but the deformation cost is no longer solely on the deformation of the boundary. Instead, the deformation cost is the cost of deforming the interior of the contour. Simultaneously, the contours should be mapped onto each other, such that the cost of deforming the interior is minimized. The elastic energy of deforming one contour into the other is commonly used as a shape similarity measure. In a similar way, the deformation cost of the contour could instead be measured in terms of deforming the interior of the contour.

1.7 Scientific Contributions

1.7.1 Published Paper and Scientific Contribution

- David Gustavsson, Kim S. Pedersen, and Mads Nielsen. Geometric and texture inpainting by gibbs sampling. In *Proceedings of Swedish Symposium in Image Analysis 2007*, 2007.

  Filter Random fields And Maximum Entropy (FRAME) [213, 214] is used for inpainting missing regions in images containing stationary texture. The problem of prolonging and connecting geometric structures, by Gibbs-sampling directly from the learned FRAME model is observed. An inverse temperature term $\beta = \frac{1}{T}$ is added to the FRAME distribution. A two-phase inpainting procedure is proposed. In the first phase, the large scale geometric structure is inpainted by sampling from a cooled distribution using a fixed $\beta > 1$. By cooling the distribution, the probability mass is redistributed in such a way that a larger part of the probability mass is located on the images with high probability. It is assumed, and empirically verified, that by cooling the distribution, the large scale geometric structure will be brought out. In the second phase, the small scale texture is added by sampling from a heated distribution.

  The experiments show that the geometric structure is prolonged and connected in a visual pleasing way in the first phase, even if the missing region is much larger than the geometric structure. The heating phase adds texture to the inpainting without destroying the geometric structure.

  Theory developed in collaboration with all authors. All implementation and experiments were done by the author.

- David Gustavsson, Kim S. Pedersen, and Mads Nielsen. Image inpainting by cooling and heating. In *Proceedings of Scandinavian Conference on Image Analysis (SCIA) 2007*, 2007. Peer review

  The two-phase inpainting strategy using FRAME, proposed in the paper [86], is extended. In the first phase, which inpaints the geometric structure, a fast cooling scheme is proposed. Using the fast cooling scheme, a more MAP-like solution is found which prolongs and connects geometric structures. The fast cooling scheme is less sensitive to parameter settings and seems to perform better than the fixed temperature approach.

The iterated Conditional Mode (ICM) which corresponds to $\beta = \infty$ is also evaluated. ICM is a site-wise greedy strategy that depends on the initialization and the visiting order. ICM often fails to inpaint the geometric structure.

In the second phase, which adds the texture, a fast heating procedure is proposed. The improvement, by using the fast heating procedure, is minor.

Theory developed in collaboration with all authors. All implementation and experiments were done by the author.

- David Gustavsson, Ketut Fundana, Niels-Ch. Overgaard, Anders Heyden, and Mads Nielsen. Variational Segmentation and Contour Matching of Non-Rigid Moving Object. In *Proceedings of Workshop on Dynamical Vision WDV 2008*, 2008. Peer review

  Level set based segmentation, including shape priors, in image sequences is combined with registration by geometry-constrained diffusion [5, 4]. By using the previous segmentation of a moving non-rigid object, as shape prior for the current segmentation, accurate object segmentation is possible even if the object is partly occluded or missing. By using registration by geometry-constrained diffusion on the object boundaries, the complete deformation and motion of the object can be estimated. The estimated motion is used for occlusion detection and temporal inpainting.

  We show, by using calculus of variation, that the geometry-constrained diffusion equation proposed by Andresen and Nielsen [5, 4] is minimizing an energy functional. The Euler-Lagrange equation for the energy functional corresponds to the proposed diffusion equation.

  Theory developed in collaboration with all authors. The shape prior based level set segmentation was implemented by Ketut Fundana. The Registration by geometry-constrained diffusion method was implemented and evaluated by the first author.

- Ketut Fundana, Niels-Ch. Overgaard, Anders Heyden, David Gustavsson and Mads Nielsen.
  Nonrigid Object Segmentation and Occlusion Detection in Image Sequences. In *Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP) 2008*, 2008.
  Peer review

  Motion estimation using shape prior segmentation and registration by geometry-constrained diffusion is treated (as in paper [81]). An al-

gorithm for estimation of the deformation and motion of a non-rigid moving object using geometry-constrained diffusion is presented. An algorithm for occlusion detection using the contour based deformation and motion estimation is presented. The intensity inside the moving object can be predicted by applying the motion estimation. If the predicted intensity in a location is different from the observed intensity, then the object is occluded in that location. The experiments show that occlusion can be detected if the deformation and motion are mild. Estimation of the deformation and motion using solely the contour of the object is not possible under large self-occlusion.

Theory developed in collaboration with all authors. The shape prior based level segmentation is implemented by Ketut Fundana. The Registration by geometry-constrained diffusion method is implemented and evaluated by the author.

- David Gustavsson, Kim S. Pedersen and Mads Nielsen. Multi-Scale Natural Images: a database and some statistics In *Danish Conference on Pattern Recognition and Image Analysis (DSAGM) 2008* Extended abstract

  The new multi-scale image sequences database is presented. The procedure and equipment used for collecting the database are discussed. Natural images are defined as images containing 'natural' scenes - both nature and man-made structure - from a human perspective, which exclude bird perspective.

  Classical results from natural image statistics are computed and verified on the new database.

  Theory developed in collaboration with all authors. The database was collected by the author and Rabia Granlund. All implementation and experiments were done by the author.

- David Gustavsson, Kim S. Pedersen, and Mads Nielsen. A SVD Based Image Complexity Measure. In *Proceding of International Conference on Computer Vision Theory and Applications (VISAPP) 2009*, 2009. Peer review

  A Truncated Singular Value Decomposition image complexity measure is proposed, based on the assumption that simple images can be approximated well in a subspace of low dimensionality, while a complex image cannot. Using the well-known property that the truncated singular value decomposition is the optimal rank k estimation, using either the 2-norm or the Frobenious norm, the rank of an approximation with

a smaller error than σ^{err} is used as the complexity measure. It is termed Singular Value Reconstruction Index (SVRI) at level σ^{err} and it is the dimensionality of the subspace where the image can be approximated with an error smaller than σ^{err}. Geometric structure can be approximated well in subspace of low dimensionality, while stochastic texture require a subspace of high dimensionality. An image is composed of patches, and the complexity of the image should be determined by the patches constituting the image. The complexity of the image is the average SVRI at level σ^{err} of the patches constituting the image.

Empirically, the rank distribution of image patches in natural images is studied. Patches in natural images are almost always of full rank. The condition number is often very large, which indicates that columns are almost linear dependent. Visual inspection indicates that patches with large smallest singular values often contain small scale details, while patches with small smallest (σ_n) singular values often contain geometric structure.

Theory developed in collaboration with all authors. All implementation and experiments were done by the author.

- David Gustavsson, Kim S. Pedersen, Francios Lauze and Mads Nielsen. On the Rate of Structural Change in Scale Spaces. In *Proceedings of Scale Space and Variational Methods in Computer Vision (SSVM) 2009*, 2009. Peer review

  The squared L^2-norm of the regularized solution and residual, as a function of the regularization parameter λ, is studied, using first order Tikhonov regularization, linear Gaussian Scale Space and Total Variation (TV) image decomposition. The squared L^2-norm of the regularized solution is a monotonically decreasing convex function of λ, the squared L^2-norm of the residual is a monotonically increasing function of λ but for non-trivial images it is not concave, using first order Tikhonov regularization. The same holds for Linear Gaussian Scale Space when the parameter is the variance of the Gaussian, but fails when the parameter is the standard deviation. Experimentally, we have shown that the squared L^2 norm of the residual is not a concave function of the regularization parameter λ using TV-decomposition. We also show, on artificial images containing details of different sizes, that inflection points of the squared residual norm as a function of λ corresponds to λ where details are totaly suppressed.

  Theory developed in collaboration with all authors. All implementation and experiments were done by the author.

1.7.2 Discussion

In this dissertation, we treat geometric structure and texture from different points of view. We argue that the most important difference between geometric structure and texture is the requirement on the representation. Geometric structure must be represented exactly, while a random sample from a distribution is sufficient for texture. Often, but not always, this is related to scale. The geometric structure is the large scale structure, while the texture is the small scale details.

In the primal sketch by Guo *et al.* [56, 57], the geometry of the objects in an image - i.e. edges and blobs - are represented exactly. The remaining regions in the image are segmented into regions containing stationary texture. The textured regions are reconstructed by a random sample from a learned distribution using the FRAME model.

In information scaling by Wu *et al.* [205], the image content as a function of the viewing distance is studied using information theory. Statistical properties of an image (or in a region of the image) depend on the viewing distance and alter by changing the viewing distance. Two processes are involved when the viewing distance is increased: smoothing and sub-sampling. Wu *et al.* argue that different image processing methods are suitable for different image content. Two regimes are singled out: low entropy and high entropy. Wavelet - or some other sparse representation - is suitable in the low entropy case, while random markov field is suitable in the high entropy case. Sparse coding can encode geometric structure - low entropy - while it fails to encode small scale stochastic details - high entropy. Random markov field fails to reconstruct long range geometric structures - low entropy regime - but it can reconstruct small scale stochastic texture. Again, we see that geometric structure must be represented exactly and this can be done using a sparse representation. Texture, on the other hand, cannot and does not have to be represented exactly.

In this thesis, we treat the problem of inpainting a missing region in a texture. Inpainting, in contrast to texture synthesizing, has boundary conditions that put constraints on it. Most texture contains details at different scales and certain details present on the boundary must be reconstructed exactly. Often, 'texture' contains geometric structure - details that must be reconstructed exactly - on a smaller scale. The classical division of inpainting methods into methods suitable for geometric structure and methods suitable for texture is rather artificial, because texture often contains geometric structure on a smaller scale and texture synthesizing methods can often prolong geometric structures. A more suitable division could be energy minimization methods and sampling methods. The failure of the energy minimization to

faithfully reconstruct stochastic texture is evident and has been shown many times. The failure of sampling methods on geometric structure is less evident and has rarely been shown on realistic images.

Empirically, we show that FRAME can prolong and connect geometric structure by cooling the learned distribution. The missing region is large, compared with the 'size' of the geometric structure, and in contrast geometric methods such as TV fails to connect the geometric structure in this case.

As pointed out by Wu *et al.* [205], the image content does not solely depend on the object in the scene, but also on the viewing distance. As the viewing distance changes, so do the image statistics. Wu *et al.* mainly study the statistical changes of different types of content, as a function of the viewing distance.

Torralba and Oliva [193, 192, 146, 147] study the image composition as a function of viewing distance. The spatial lay-out of a scene is termed *the spatial envelope* and is a function of the viewing distance. Considering images captured by a human (i.e. from human view point at ground position), the possible angles from which an object can be captured from is determined by the viewing distance. Small objects captured from a small distance can be captured from almost all angles, while large objects captured from a large distance can be captured from very few angles. A cup on a table can be captured from almost all angles, while a house captured from 200 meters can be captured from a few angles, e.g. we cannot see the house from above unless flying. The spatial lay-out of a scene captured at a large viewing distance is rather fixed; the sky occupies the top, buildings, forests and mountains occupy the middle, while the roads and lawns occupy the lower part of the image. The spatial envelope is constrained by the viewing distance. Torralba and Oliva show that the spatial envelope, and thereby the viewing distance, has a large influence on the η in the power spectra power law for natural images. They show that by estimating η on individual images, the distance to the main object in the scene can be estimated. They used a model where η depends on the angle and estimate η_θ using different angles.

Three classical results from natural image statistics are studied, using the new image sequence database containing images of the same scene captured at different scales. The power spectra power law (scale invariance), the Laplacian distribution of the partial differential and the distribution of homogenous regions (the size power law) are studied. We are facing the question: How much of the visual appearance of the image can be be explained by the statistical property of the image? How does the estimation relate to geometric structure and texture in the image, and to the viewing distance?

In general, images captured from a large distance - more than 100 meters - mainly contain geometric structure. At large viewing distances, the sky

is present, which often is a very large geometric structure. Furthermore, buildings, roads, lawns and trees appear as rather uniformly colored regions viewed from a large distance, i.e geometric structure. This is confirmed by the statistical estimations: η in the power spectra power law is rather small, which indicates that intensities are more correlated. Estimation of α in the generalized Laplacian distribution is small, which indicates a sharp peak at zero, but also large values which correspond to object boundaries. Estimation of α in the size power law of homogenous regions is small, which indicate the presence of larger homogenous regions. Images mainly containing geometric structure can be characterized as follows: the intensities are more correlated, they contain larger homogenous regions and the partial derivatives are in general small inside the regions, but rather large on the object boundaries.

In general, images captured from a small distance - less than 20 meters -, mainly contain texture. Trees capture from less than 100 meters also mainly contain texture. At small viewing distances the sky is absent and the small scale details on the object in the scene have been brought out. At such a small distance, details on trees, bushes and lawns are brought out. This is, again, confirmed by the statistics: estimation of η in the power spectra power law is large which indicates that the intensities are less correlated. Estimation of α in the generalized Laplacian distribution is rather large and estimation of α in the size distribution is large. Images mainly containing texture can be characterized by: the intensities are less correlated, they contain smaller homogenous regions and the distribution of partial derivatives is less peaked at zero.

In order to estimate the image content in terms of geometric structure and texture, an approximation approach is proposed. The approximation approach, again, relates to previous work by Wu *et al.* [205], where they argue that texture cannot be represented sparsely, while geometric structure can. The approximation approach can be viewed as reversing the argumentation: if the image content can be represented sparsely, then it contains geometric structure. And, if the image content can not be represented sparsely, then it contains texture. The truncated singular value decomposition is used and the rank of a good approximation is used as the complexity measure. The rank of the approximation is the number of basis vectors required for a good approximation.

The second approximation approach is based on image regularization in the continuous domain. Image regularization can be viewed as an approximation of an image with a simpler one, often in a different subspace of functions. Assuming that the observed image is in $u_0 \in L^2$, then first order Tikhonov regularization and linear Gaussian scale space map the image into a Sobolev space, while TV maps into the space of functions with bounded variation

(BV).

Buades *et al.*[27] introduced the 'Method Noise' for evaluating denoising methods. The 'Method Noise' is the image residual and should, after denoising, solely contain the noise. By analyzing the content in residual, the performance of the denoising method can be characterized. The proposed 'Method Noise' evaluation method is, in some sense, the complementary problem. Instead of evaluating and characterizing the method, our aim is to characterize the image content by the content in the image residual.

In the residual norm study, conclusions in the case of first order Tikhonov regularization and linear Gaussian scale space are made by analytically proving the properties. In the TV case, the conclusion is based on experiments. Finding a closed form expression for the residual norm and the derivative of the residual norm with respect to λ would be very rewarding. As it seems from the experiments, the residual norm has points of high curvature at a scale for which structure of a certain size is totally removed. The distribution of such a point of high curvature would be very important for describing the image content. It would also be very useful for optimal parameter selection.

Chapter 2

Image Inpainting by Cooling and Heating

This chapter contain a slightly re-formatted version of

David Gustavsson, Kim S. Pedersen, and Mads Nielsen.
Image inpainting by cooling and heating.
In *Proceedings of Scandinavian Conference on Image Analysis (SCIA) 2007*, 2007.

Image Inpainting by Cooling and Heating

David Gustavsson[1], Kim S. Pedersen[2], and Mads Nielsen[2]
[1]IT University of Copenhagen
Rued Langgaards Vej 7, DK-2300 Copenhagen S,Denmark
davidgsson@itu.dk
[2]DIKU, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark
{kimstp,madsn}@diku.dk

abstract

We discuss a method suitable for inpainting both large scale geometric structures and stochastic texture components. We use the well-known FRAME model for inpainting. We introduce a temperature term in the learnt FRAME Gibbs distribution. By using a fast cooling scheme a MAP-like solution is found that can reconstruct the geometric structure. In a second step a heating scheme is used that reconstruct the stochastic texture. Both steps in the reconstruction process are necessary, and contribute in two very different ways to the appearance of the reconstruction.
Keywords: Inpainting, FRAME, ICM, MAP, Simulated Annealing

2.1 Introduction

Image inpainting concerns the problem of reconstruction of the image contents inside a region Ω with unknown or damaged contents. We assume that Ω is a subset of the image domain $D \subseteq \mathbb{R}^2$, $\Omega \subset D$ and we will for this paper assume that D form a discrete lattice. The reconstruction is based on the available surrounding image content. Some algorithms have reported excellent performance for pure geometric structures (see e.g. [39] for a review of such methods), while others have reported excellent performance for pure

textures (e.g. [21, 51, 52]), but only few methods [17] achieve good results on both types of structures.

The variational approaches have been shown to be very successful for geometric structures but have a tendency to produce a too smooth solution without fine scale texture (See [39] for a review). Bertalmio et al [17] propose a combined method in which the image is decomposed into a structure part and a texture part, and different methods are used for filling the different parts. The structure part is reconstructed using a variational method and the texture part is reconstructed by image patch pasting.

Synthesis of a texture and inpainting of a texture seem to be, more or less, identical problems, however they are not. In [84] we propose a two step method for inpainting based on Zhu, Wu and Mumford's stochastic FRAME model (Filters, Random fields and Maximum Entropy) [214, 213]. Using FRAME naively for inpainting does not produce good results and more sophisticated strategies are needed and in [84] we propose such a strategy. By adding a temperature term T to the learnt Gibbs distribution and sampling from it using two different temperatures, both the geometric and the texture component can be reconstructed. In a first step, the geometric structure is reconstructed by sampling using a cooled - i.e. using a small fixed T - distribution. In a second step, the stochastic texture component is added by sampling from a heated - i.e. using a large fixed T - distribution.

Ideally we want to use the MAP solution of the FRAME model to reconstruct geometric structure of the damaged region Ω. In [84] we use a fixed low temperature to find a MAP-Like solution in order to reconstruct the geometric structure. To find the exact MAP-solution one must use the time consuming simulated annealing approach such as described by Geman and Geman [69]. However to reconstruct the missing contents of the region Ω, the true MAP solution may not be needed. Instead a solution which is close to the MAP solution may provide visually good enough results. In this paper we propose a fast cooling scheme that reconstruct the geometric structure and approaches the MAP solution. Another approach is to use the solution produced by the Iterated Conditional Modes (ICM) algorithm (see e.g. [201]) for reconstruction of the geometric structure. Finding the ICM solution is much faster than our fast cooling scheme, however it often fails to reconstruct the geometric structure. This is among other things caused by the ICM solutions strong dependence on the initialisation of the algorithm. We compare experimentally the fast cooling solution with the ICM solution.

To reconstruct the stochastic texture component the Gibbs distribution is heated. By heating the Gibbs distribution more stochastic texture structures will be reconstructed without destroying the geometric structure that was reconstructed in the cooling step. In [84] we use a fixed temperature to find

a solution including the texture component. Here we introduce a gradual heating scheme.

The paper has the following structure. In section 2.2 FRAME is reviewed, in section 2.2.1 filter selection is discussed and in section 2.2.2 we explain how FRAME is used for reconstruction. Inpainting using FRAME is treated in section 2.3. In section 2.3.1 a temperature term is added to the Gibbs distribution, the ICM solution and fast cooling solution is discussed in sections 2.3.2 and 2.3.3. Adding the texture component by heating the distribution is discussed in section 2.3.4. In section 2.4 experimental results are presented and in section 2.5 conclusion are drawn and future work is discussed.

2.2 Review of FRAME

FRAME is a well known method for analysing and reproducing textures [213, 214]. FRAME can also be thought of as a general image model under the assumptions that the image distribution is stationary. FRAME constructs a probability distribution $p(I)$ for a texture from observed sample images.

Given a set of filters $F^\alpha(I)$ one computes the histogram of the filter responses H^α with respect to the filter α. The filter histograms are estimates of marginal distributions of the full probability distribution $p(I)$. Given the marginal distributions for the sample images one wants to find all distributions that have the same expected marginal distributions, and among those find the distribution with maximum entropy, i.e. by applying the maximum entropy principle. This distribution is the least committed distribution fulfilling the constraints given by the marginal distributions. This is a constrained optimisation problem that can be solved using Lagrange multipliers. The solution is

$$p(I) = \frac{1}{Z(\Lambda)} \exp\{-\sum_i \sum_\alpha \lambda_i^\alpha H_i^\alpha\} \tag{2.1}$$

Here i is the number of histogram bins in H^α for the filter α and $\Lambda = \{\lambda_i^\alpha\}$ are the Lagrange multipliers which gives information on how the different values for the filter α should be distributed. The relation between λ^α:s for different filters F^α gives information on how the filters are weighted relative to each other.

An Algorithm for finding the distribution and Λ can be found in [214]. FRAME is a generative model and given the distribution $p(I)$ for a texture it can be used for inference (analysis) and synthesis.

2.2.1 The Choice of Filter Bank

We have used three types of filters in our experiments: The delta filter, the power of Gabor filters and Scale Space derivative filters. The delta, Scale Space derivative and Gabor filters are linear filters, hence $F^\alpha(I) = I * F^\alpha$, where $*$ denotes convolution. The power of the Gabor filter is the squared magnitude applied to the linear Gabor filter.

The Filters F^α are:

- Delta filter - given by the Dirac delta $\delta(x)$ which simply returns the intensity at the filter position.
- the power of Gabor filters - defined by $\mid I * G_\sigma e^{-i\omega x} \mid^2$, where $i^2 = -1$. Here we use 8 orientations, $\omega = 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{3\pi}{2}, \frac{7\pi}{4}$ and 2 scales $\sigma = 1, 4$, in total 16 Gabor filters have been used.
- Scale space derivatives - using 3 scales $\sigma = 0.1, 1, 3$ and 6 derivatives $G_\sigma, \frac{\partial G_\sigma}{\partial x}, \frac{\partial G_\sigma}{\partial y}, \frac{\partial^2 G_\sigma}{\partial x^2}, \frac{\partial^2 G_\sigma}{\partial y^2}, \frac{\partial^2 G_\sigma}{\partial x \partial y}$.

For both the Gabor and scale space derivative filters the Gaussian aperture function G_σ with standard deviation σ defining the spatial scale is used,

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) .$$

Which and how many filters should be used have a large influence on the type of image that can be modelled. The filters must catch the important visual appearance of the image at different scales. The support of the filters determines a Markov neighbourhood. Small filters add fine scale properties of the image, while large filters add coarse scale properties of the image. Hence to model properties at different scales, different filter sizes must be used. The drawback of using large filters is that the computation time increases with the filter size. On the other hand large filters must be used to catch coarse scale dependencies in the image.

Gabor filters are orientation sensitive and have been used for analysing textures in a number of papers and are in general suitable for textures (e.g. [20, 100]). By carefully selecting the orientation ω and the scale σ, structures with different orientations and scales will be captured.

It is well known from scale space theory that scale space derivative filters capture structures at different scales. By increasing σ in the Gaussian kernel, finer details are suppressed, while coarse structures are enhanced. By using the full scale-space both fine and coarse scale structures will be captured [188].

2.2.2 Sampling

Once the distribution $p(I)$ is learnt, it is possible to use a Gibbs sampler to synthesise images from $p(I)$. I is initialised randomly (or in some other way based on prior knowledge). Then a site $(x, y)_i \in D$ is randomly picked and the intensity $I_i = I((x, y)_i)$ at $(x, y)_i$ is updated according to the conditional distribution [123, 201]

$$p(I_i|I_{-i}) \tag{2.2}$$

where the notation I_{-i} denotes the set of intensities at the set of sites $\{(x, y)_{-i}\} = D\backslash(x, y)_i$. Hence $p(I_i|I_{-i})$ is the probability for the different intensities in site $(x, y)_i$ given the intensities in the rest of the image. Because of the equivalence between Gibbs distributions and Markov Random Fields given a neighbourhood system N (the Hammersley-Clifford theorem, see e.g. [201]), we can make the simplification

$$p(I_i|I_{-i}) = p(I_i|I_{N_i}) \tag{2.3}$$

where $N_i \subset D\backslash(x, y)_i$ is the neighbourhood of $(x, y)_i$. In the FRAME model, the neighbourhood system N is defined by the extend of the filters F^α.

By sampling from the conditional distribution in (2.3), I will be a sample from the distribution $p(I)$.

2.3 Using FRAME for inpainting

We can use FRAME for inpainting by first constructing a model $p(I)$ of the image, e.g. by learning from the non-damaged part of the image, $D\backslash\Omega$. We then use the learnt model $p(I)$ to sample new content inside the damaged region Ω. This is done by only updating sites in Ω. A site $(x, y)_i \in \Omega$ is randomly picked and updated by sampling from the conditional distribution given in (2.3). If the site $(x, y)_i$ is close (in terms of filter size) to the boundary $\partial\Omega$ of the damaged region, then the filters get support from both sites inside and outside Ω. The sites outside Ω are known and fixed, and are boundary conditions for the inpainting. We therefore include a small band region around Ω in the computation of the histograms H^α. Another option would have been to use the whole image I to compute the histogram H^α, however this has the downside that the effect of updates inside Ω on the histograms are dependent on the the relative size ratio between ω and D, causing a slow convergence rate for small Ω.

2.3.1 Adding a temperature term $\beta = \frac{1}{T}$

Sampling from the distribution $p(I)$ using a Gibbs sampler does not easily enforce the large scale geometric structure in the image. By using the Gibbs sampler one will get a sample from the distribution, this includes both the stochastic and the geometric structure of the image, however the stochastic structure will dominate the result.

Adding an inverse temperature term $\beta = \frac{1}{T}$ to the distribution gives

$$p(I) = \frac{1}{Z(\Lambda)} \exp\{-\beta \sum_{\alpha} \sum_{i} \lambda_i^{\alpha} H_i^{\alpha}\} \ . \tag{2.4}$$

In [84] we proposed a two step method to reconstruct both the geometric and stochastic part of the missing region Ω:

1. Cooling: By sampling from (2.4) using a fixed small temperature T value, structures with high probability will be reconstructed, while structures with low probability will be suppressed. In this step large geometric structures will be reconstructed based on the model $p(I)$.

2. Heating: By sampling from (2.4) using a fixed temperature $T \approx 1$, the texture component of the image will be reconstructed based on the model $p(I)$.

In the first step the geometric structure is reconstructed by finding a smooth MAP-like solution and in the second step the texture component is reconstructed by adding it to the large scale geometry.

In this paper we propose a novel variation of the above discussed method. We consider two cooling schemes and a gradual heating scheme which can be considered as the inverse of simulated annealing.

2.3.2 Cooling - the ICM solution

Finding the MAP solution by simulated annealing is very time consuming. One alternative method is the Iterated Conditional Modes (ICM) algorithm. By letting $T \to 0$ (or equivalently letting $\beta \to \infty$) the conditional distribution (2.3) will become a point distribution. In each step of the Gibbs sampling one will set the new intensity for a site $(x, y)_i$ to

$$I_i^{\text{new}} = \arg\max_{I_i} p(I_i \mid I_{N_i}) \ . \tag{2.5}$$

This is a site-wise MAP solution (i.e. in each site and in each step the most likely intensity will be selected). This site-wise greedy strategy is not

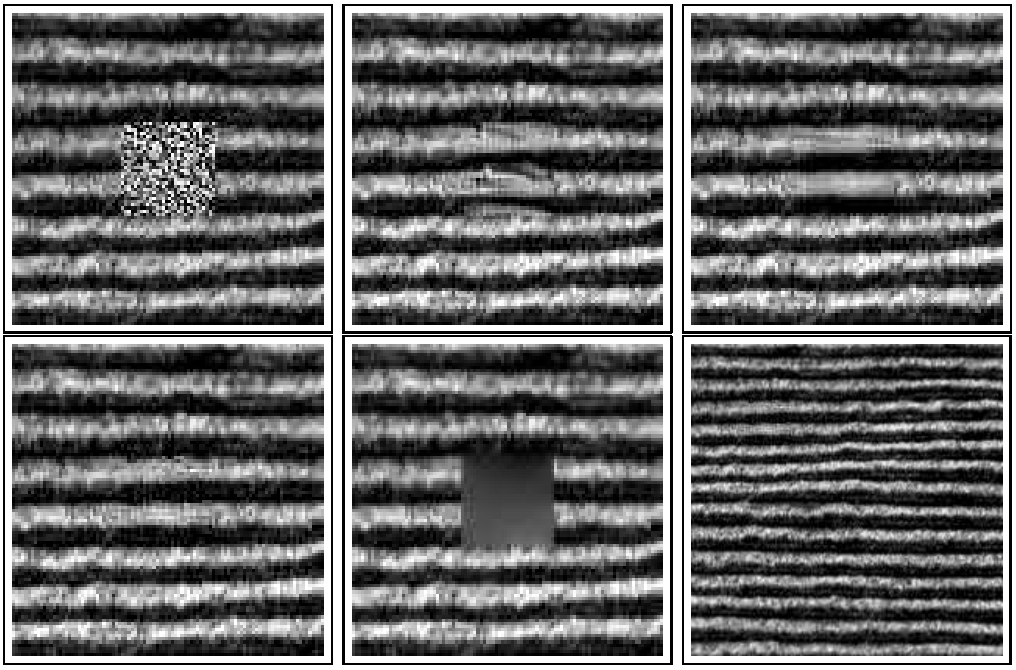


Figure 2.1: From top left to bottom right: a) the image containing a damaged region b) the ICM solution c) the fast cooling solution d) adding texture on top of the fast cooling solution by heating the distribution e) total variation (TV) solution and f) the reconstructed region in context (can you find it?).

guaranteed to find the global MAP solution for the full image. The ICM solution is similar but not identical to the high β sampling step described in [86]. The ICM solution depends on initialisation of the unknown region Ω. Here we initialise by sampling pixel values identically and independent from a uniform distribution on the intensity range.

2.3.3 Cooling - Fast cooling solution

The MAP solution for the inpainting is the most likely reconstruction given the known part of the image $D\backslash\Omega$,

$$I^{\text{MAP}} = \arg \max_{I_i \forall (x,y)_i \in \Omega} p(I \mid I(D\backslash\Omega), \Lambda) \ . \tag{2.6}$$

Simulated annealing can be used for finding the MAP solution. Replacing β in (2.4) with an increasing (decreasing) sequence β_n called a cooling (heating) scheme. Using simulated annealing one starts to sample using a high temperature T and slowly cooling down the distribution (2.4) by letting

$T \to 0$. If β_n is increasing slowly enough and letting $n \to \infty$ then simulated annealing will find the MAP solution (see e.g. [69, 201, 123]). Unfortunately simulated annealing is very time consuming.

To reconstruct Ω, the true MAP solution may not be needed, instead a solution which is close to the MAP solution may be enough. We therefore adopt a fast cooling scheme, that does not guarantee the MAP solution. The goal is to reconstruct the geometric structure of the image and suppress the stochastic texture.

The fast cooling scheme used in this paper is defined as (in terms of β)

$$\beta_{n+1} = C^+ \cdot \beta_n \tag{2.7}$$

where $C^+ > 1.0$ and $\beta_0 = 0.5$.

2.3.4 Heating - Adding texture

The geometric structures of the image will be reconstructed by sampling using the cooling scheme. Unfortunately the visual appearance will be too smooth, and the stochastic part of the image needs to be added.

The stochastic part should be added in such a way that it does not destroy the large scale geometric part reconstructed in the previous step. This is done by sampling from the distribution (2.4) using a heating scheme similar to the cooling scheme presented in previous section and using the solution from the cooling scheme as initialisation.

The heating scheme in this paper is

$$\beta_{n+1} = C^- \cdot \beta_n \tag{2.8}$$

where $C^- < 1.0$ and $\beta_0 = 25$.

By using a decreasing β_n, value finer details in the texture will be reproduced, while coarser details in the texture will be suppressed.

2.4 Results

Learning the FRAME model $p(I)$ is computational expensive, therefore only small image patches have been used. Even for small image patches the optimisation times are at least a few days. After the FRAME model has been learnt, inpainting can be done relatively fast if Ω is not to large.

The dynamic range of the images have been decreased to 11 intensity levels for computational reasons. The images that have been selected includes both large scale geometric structures as well as texture.

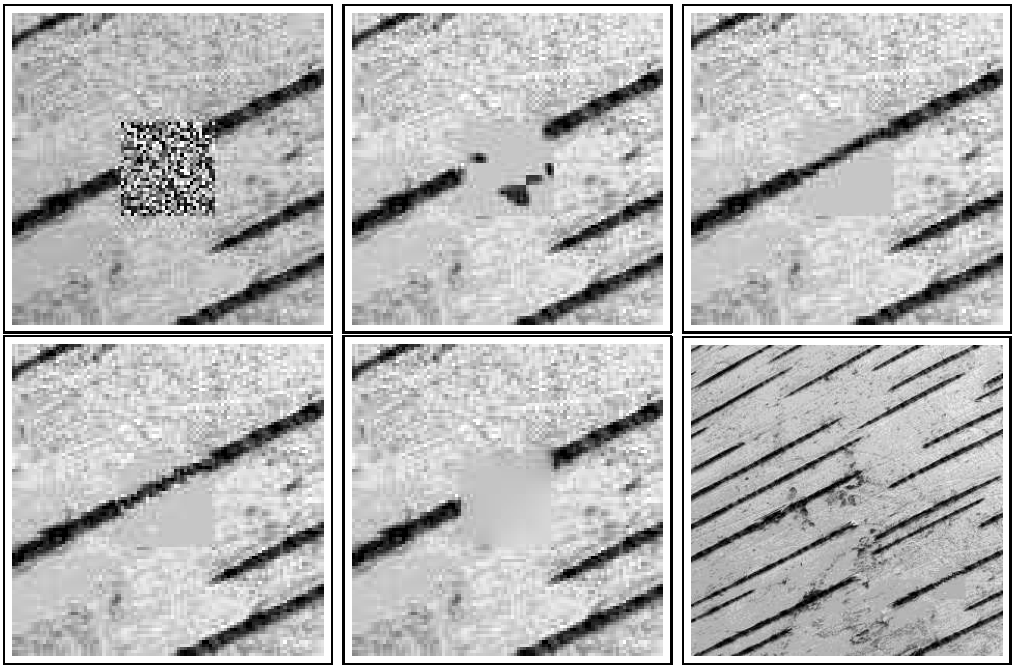


Figure 2.2: From top left to bottom right: a) the image containing a damaged region b) the ICM solution c) the fast cooling solution d) adding texture on top of the fast cooling solution by heating the distribution e) total variation (TV) solution and f) the reconstructed region in context (can you find it?).

The delta filter, 16 Gabor filters and 18 scale space derivative filters have been used in all experiments and 11 histogram bins have been used for all filters (see section 2.2.1 for a discussion).

In the cooling scheme (2.7), we use $\beta_0 = 0.5$, $C^+ = 1.2$ and the stopping criterion $\beta_n > 25$ in all experiments. In the heating scheme (2.8), we use $\beta_0 = 25$,$C^- = 0.8$ and the stopping criterion $\beta_n < 1.0$.

Each figure contains an unknown region Ω of size 30×30 that should be reconstructed. Figure 2.1 contains corduroy images, figure 2.2 contains birch bark images and figure 2.3 wood images. Each figure contains the original image with the damaged region Ω with initial noise, the ICM and fast cooling solutions and the solution of a total variation (TV) based approach [39] for comparison.

The ICM solution reconstruct the geometric structure in the corduroy, but fails to reconstruct the geometric structure in both the birch and the wood images. This is due to the local update strategy of ICM, which makes it very sensitive to initial conditions. If ICM starts to produce wrong large

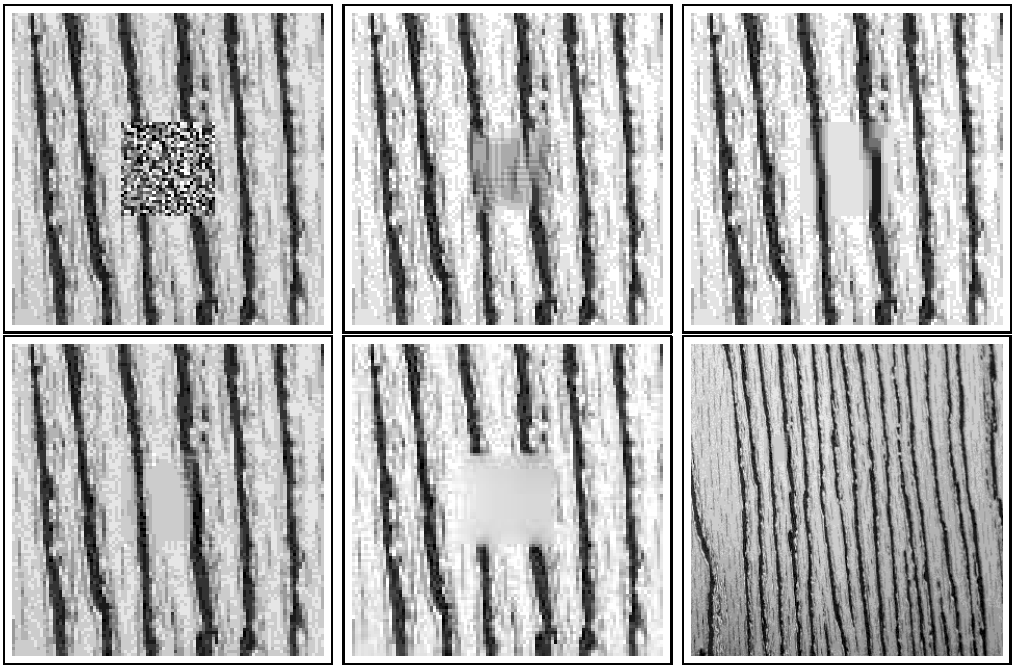


Figure 2.3: From top left to bottom right: a) the image containing a damaged region b) the ICM solution c) the fast cooling solution d) adding texture on top of the fast cooling solution by heating the distribution e) total variation (TV) solution and f) the reconstructed region in context (can you find it?).

scale geometric structures it will never recover.

The fast cooling solution on the other hand seem to reconstruct the geometric structure in all examples and does an even better job than the ICM solution for the corduroy image. The fast cooling solutions are smooth and have suppressed the stochastic textures. Because of the failure of ICM we only include results on heating based on the fast cooling solution.

The results - image d) - after the heating are less smooth Ω's, but it is still smoother than $I \backslash \Omega$. The total variation (TV) approach produce a too smooth solution even if strong geometric structures are present in all example.

2.5 Conclusion

Using FRAME to learn a probability distribution for a type of images gives a Gibbs distribution. The boundary condition makes it hard to use the learnt Gibbs distribution as it is for inpainting; it does not enforce large scale

geometric structures strongly enough. By using a fast cooling scheme a MAP-like solution is found that reconstruct the geometric structure. Unfortunately this solution is too smooth and does not contain the stochastic texture. The stochastic texture component can be reproduced by sampling using a heating scheme. The heating scheme adds the stochastic texture component to the reconstruction and decrease the smoothness of the reconstruction based on the fast cooling solution.

A possible continuation of this approach is to replace the MAP-like step with a partial differential equation based method and a natural choice is the Gibbs Reaction And Diffusion Equations (GRADE) [212, 211], which are build on the FRAME model.

We decompose an image into a geometric component and a stochastic component and use the decomposition for inpainting. This is related to Meyer's [8, 132] image decomposition into a smooth component and a oscillating component (belonging to different function spaces). We find it interesting to explore this theoretic connection with variational approaches.

Acknowledgements

This work was supported by the Marie Curie Research Training Network: Visiontrain (MRTN-CT-2004-005439).

Chapter 3

A Multi-Scale Study of the Distribution of Geometry and Texture in Natural Images

A Multi-Scale Study of the Distribution of Geometry and Texture in Natural Images

David Gustavsson, Kim S. Pedersen, and Mads Nielsen
DIKU, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark
{davidg,kimstp,madsn}@diku.dk

abstract

A new image database containing an ensemble of image sequences is presented. Each sequence contains 15 images of the same scene, captured at different viewing distances, termed capture scales. The scenes contain both nature and man-made structure, and the images are captured from a 'normal' human point-of-view. The part of the scene present at all capture scales has been extracted, resulting in sequences of images of increasing resolution with the same content.

Classical results from natural image statistics - scale invariance, Laplacian distribution of the partial derivative and the size distribution of homogenous region - are verified and analyzed on the database.

The classical natural image statistics are also estimated on individual images. The estimation on individual images can explain the visual appearance in terms of geometric structure and texture to some degree. We argue that estimation on individual images depends on the viewing distance in two different ways: the spatial lay-out of the scene and the suppression of details (inner scale). Images, captured from a human point of view, from large viewing distance contain the sky on top, houses or forests in the vertical middle and lawns or roads in the lower part. The spatial layout is constrained by the viewing distance. The sky, buildings, lawns and forests appear as rather uniformly colored regions viewed from a large distance. This is because the inner scale is too large to bring out the texture at such a distance.
Keywords: natural images, scale space, geometric structure, texture, scale invariance, power law, generalized Laplacian distribution, area distribution

3.1 Introduction

Images contain different types of information, from highly stochastic texture such as grass and fur to highly geometric structures, such as houses and cars. Furthermore, most images contain a mix of geometric structures and stochastic textures. The image content does not solely depend on the objects in the captured scene, but also on the scale that it was captured at. The

same object, captured at different scales, will have different appearance. For example, a tree viewed from 5 meters is very different from the same tree viewed from 100 meters. At a coarse scale, finer details - such as the leaves - are suppressed while the coarse scale structure - the tree top and the trunk - are brought out. At a finer scale, the coarse scale geometric structures are suppressed while the finer scale details are brought out. A coarse scale representation of an image can be generated artificially using linear gaussian scale space by convolving the original image with a gaussian function [98, 202, 111, 122, 60]. Generating a coarse scale representation of an image using linear gaussian scale space will increase the effective inner-scale, small details are suppressed, while keeping the resolution. Similarly, increasing the viewing distance increases the inner-scale, smaller details will not be captured, and the resolution will decrease. Scale space does not model the statistical changes of the image content when the viewing distance is altered.

For specific type of objects - such as houses or trees - how does the statistical property change as a function of viewing distance? By capturing the same part of a scene at different viewing distances, statical changes due to altering the inner-scale and resolution can be analyzed. Wu *et. al* [205] studied the 'entropy rate' and 'inferential uncertainty' as a function of viewing distance.

Changing the viewing distance will change the inner-scale, it will also have an effect on the outer scale of the image. Changing the viewing distance, either by moving the camera or adjusting the focal length in the objective, will alter the composition of the captured scene. Torralba and Oliva [146, 147] called the spatial lay-out of an image the *spatial envelope*, and they showed that it can be used for determining the distance to the main objects in the scene [192]. Here we only consider natural images, captured from a human point of view i.e. underwater and bird views are excluded. The distance to the object in the scene also puts hard constraints on the possible views that the object can be seen from. A cup on a table can be viewed from almost all angles, a car on street can be viewed from many angles, while a large building can be viewed from a few angles. How does the image statistics change when the viewing distance changes? Capturing the same scene at different viewing distances (and different outer-scale) may reveal statical changes due to changes in the spatial envelope.

The statistical change as a function of the viewing distance due to changes in the spatial envelope will be studied in this report.

The natural image database provided by van Hateren, also called the natural stimuli collection, is one of the most widely used databases [197]. The van Hateren database contains roughly 4000 images of resolution 1024×1536. The database contains scenes captured at different scales, but it does not

contain images of the same scene captured at different scales.

The KTH-TIPS2 database contains 11 materials captured under different illumination and scales ([63]). The database contains texture captured at different scales.

To be able to analyze how the image content changes, when the same scene has been captured at different scales, a firs step is to collect a new image database. A new database containing a rich variety of scenes and distances is introduced. The database contains natural scenes - both man-made and natural environments - captured at 15 different 'scales'. The viewing distance is altered by adjusting the focal length and the different focal lengths are called capture scales. The database and the collection procedure are presented in section 3.2.

Classical statistical properties, found in ensemble of natural images are computed on the database and the result is compared with previous reported results. The statistics are also computed based on the capture scale - estimation using images with the same capture scale. In section 3.4.1, the apparent scale invariance and power spectra power law found in ensemble of natural images are discussed. The distribution of partial derivatives computed on an ensemble of natural images can be modeled by a generalized Laplacian distribution, and is discussed in section 3.4.2. The distribution of homogenous regions in natural images, both computed on individual images and on ensemble of images, is following a power law in size, and is discussed in section 3.4.3.

3.2 Multi-Scale Geometry and Texture image Database (MS-GTI DB)

Images or rather scenes are considered to be 'natural' if they naturally appear in everyday life, from a human point of view. The human point of view excludes aerial and underwater images even if they in some sense are natural. Scenes containing both man-made structures and natural environments have been captured.

3.2.1 Collection procedure and equipment

The MS-GTI database contains images of the same scene captured at different scales. The camera that has been used is a Nikon D40x. The three different objectives that have been used are: 18-55 mm, 55-200 mm and 70-300 mm. The camera has been placed on a tripod stand facing the scene. A region of interest in the scene, of such a size that it is present at all capture

scales, has been selected. The scene, with the region of interest approximately in the center, is captured at different scales by adjusting/changing the objective. The scene is captured at 15 different scales, the focal length varies from 18 mm to 300 mm - roughly 4 octaves and 16 times magnification. Hence a 1×1 pixel region in the least zoomed image corresponds to a 16×16 region in the most zoomed image. The image resolution is 2592×3872 pixels.

The image content is of course determined by the distance from the camera to the scene. The distance between the camera and the scene varies, from a few meters to a distance of hundreds of meters - "panorama" distance .

Objective	Focal Length	number of Capture Scales
18-55 mm	18, 24, 35, 45	4
55-200 mm	55,70,85,105,135,165,200	7
70-300 mm	225,250,275,300	4

Table 3.1: The three objectives used to collect the database, together with the focal length used for the objectives. 15 images have been collected for each scene giving approximately 16x magnification.

The RAW format used by the D40x camera is Nikons own 14 bits format NEF. The NEF images have been converted to 16 bits TIFF images, each TIFF image is 60 MB. The images in a sequence are indexed from the least zoomed image I_1 (smallest focal length) to the most zoomed image I_{15}, i.e. in increasing zoom order. This index is the *capture scale* used in the following sections. Increasing the capture scale corresponds to decreasing the viewing distance and the decreasing inner scale. The capture scale simply denotes the numbering of the focal length used.

Table 3.1 describes the used objectives and table 3.2 shows the focal length for each of the 15 images of a sequence. Examples can be found in figure 3.1

3.2.2 The different Scenes

The scenes selected for the database are mostly natural images containing both man-made environments - mostly buildings - and nature - trees, tree trunks and bushes. In many cases the same type of scenes have been captured but with different distance between the camera and the scene, which changed the captured image contents.

Figure 3.1: Example of captured scenes. The columns contain from left to right: the least zoomed image I_1 - 18 mm, I_3 - 35 mm,I_6 - 70mm, I_{10} - 165 mm, and the most zoomed image I_{15} - 300 mm. Row 1 (IS^1), 2 (IS^4) and 7 (IS^{18}) contain man-made environments, 4 (IS^8) and 9 (IS^{31}) contain a mixture of nature and man-made environments, and 3 (IS^6), 5 (IS^{10}), 6 (IS^{11}) and 8 (IS^{28}) contain nature environments. The distance between the main objects in the scene and the camera varies between the scenes - from a few meters to "panorama" distance. This gives a large variation in distance and image contents at all scales (zoom) - rows 7 and 9 contain a large portion of the sky even in the most zoomed images and row 8 contains small scale texture in the least zoomed image.

Image	Objective	focal length	Image	Objective	focal length
I_1	18-55	18	I_9	55-200	135
I_2	18-55	24	I_{10}	55-200	165
I_3	18-55	35	I_{11}	55-200	200
I_4	18-55	45	I_{12}	70-300	225
I_5	55-200	55	I_{13}	70-300	250
I_6	55-200	70	I_{14}	70-300	275
I_7	55-200	85	I_{15}	70-300	300
I_8	55-200	105			

Table 3.2: Summary of the objectives and focal lengths used for collecting the images in the sequences.

The depth of field is the portion of a scene that appears to be sharp in the image. A lens (objective) can be focused only on one specific distance (the focus plan), still objects in the scene on a distance close to the focus plan appear to be sharp. The depth of field is the distance range, where the objects in the scene are in acceptable focus. The depth of field varies with the objective, and is usually larger for normal objective, while it is smaller for zoom objective.

If objects in the scene, captured using a magnification objective, appear at different distances, some of the objects will be out-of-focus.

For example; a closeup picture of a scrub, the distance between the camera and the twigs is varying a lot. This will result in an image where some of the twigs are in focus, while others are out-of-focus. Examples of scenes are presented in figure 3.1. Scenes containing man-made structure - rows 1 and 2 - are often planar in the most zoomed image. Thereby all objects in the scene are in the depth of field. Nature images - row 3 - often contain objects at varying distance to the camera. Thereby, some objects are outside the depth of field and they are out-of-focus.

3.2.3 Region extraction

The part of the scene captured by the camera that is present in all capture scales, has been extracted. Resulting in sequences of images of different resolution, containing the same part of the scene at different capture scales. The resolutions of the different regions range from 2592×3872 to 160×240 and is summarized in table 3.3.

The regions are extracted by registration of the most zoomed image I_{15} in all of the other images. This is a very challenging registration problem, because

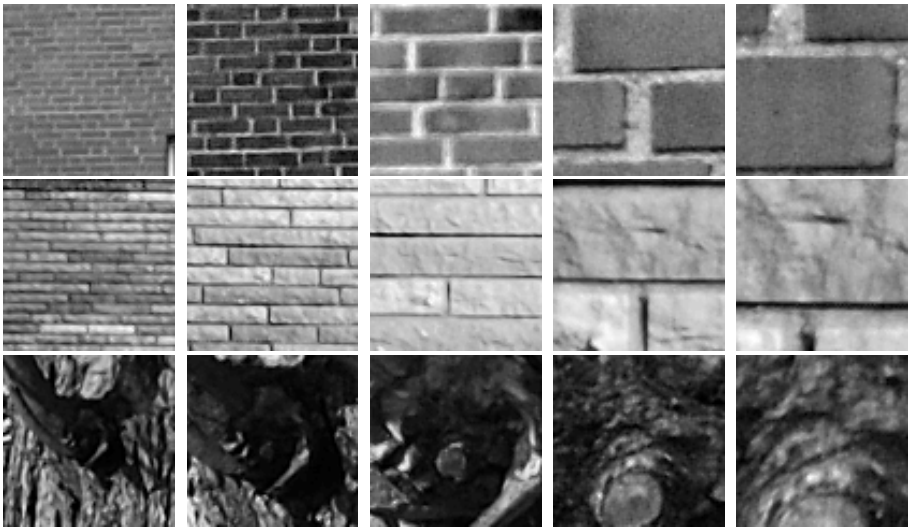


Figure 3.2: The figure contains 80×80 patches extracted from three different image sequences at different scales. Column 1 is extracted from I_1 (least zoomed), column 2 from I_3, column 3 from I_6, column 4 from I_{10} and column 5 from I_{15}. The image contents at the different scales are very different even if the captured object is the same. The first row contains part of a brick wall (row two in figure 3.1)- on the coarse scale the brick wall appears as texture that transforms into bricks at finer scale. The second row contains a part of a brick wall and the appearance at the different scale is similar to the other brick wall.

of the large range of scales. The problem has partly been solved using manual feature selection and affine registration, SIFT features [124] combined with RANSAC [59] for computing an affine registration, and manual registration.

Region Id	Region Size	Region Id	Region Size
R_{14}	2470×3690	R_7	760×1170
R_{13}	2230×3330	R_6	620×950
R_{12}	1980×2950	R_5	480×740
R_{11}	1740×2600	R_4	380×580
R_{10}	1520×2260	R_3	300×460
R_9	1200×1790	R_2	200×310
R_8	940×1440	R_1	160×240

Table 3.3: The extracted regions and where resolution, R_i is extracted from I_i and $R_{15} = I_{15}$.

3.2.4 Notation

The notation used for the sequences of images/regions is summarized in table 3.4. The captured full images are called 'images', and are denoted by I sometimes with an lower index, and the extracted part of the images are called 'regions', and are denoted by R, sometimes with an lower index. The sequences of images are denoted IS_i^j and the sequences of regions are denoted RS_i^j where the upper index indicates the sequence and the lower index indicates the image number (sometimes the indexes are omitted).

3.3 Point Operators and Scale Space

Comparing images containing the same part of a scene, captured at different scales by zooming, is a challenging problem that requires an understanding of the image formation process. A simple model [112, 80] and its relation to scale space is discussed.

Let $S(r)$ be a scene, and let

$$G_0(x, \sigma) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2}{2\sigma^2}) \tag{3.1}$$

be a linear detector, called a point operator, with weight σ. Applying the detector at a position will yield a point observation and by applying the

Abbreviation	Meaning
Image (I_i)	An (full) image what has been captured
Region (R_i)	An extracted region from an image
Patch	A part of a region or image
IS	Image Sequences i.e. all images in the db
IS_i^j	Image i from sequence j
IS_i	All images numbered i
	i.e one image from every sequence
IS^j	All images from sequence j
RS	Region Sequences i.e. all regions in the db
RS_i^j	Region i from sequence j.
RS_i	All regions numbered i,
	i.e. one region from each region sequence.
RS^j	All regions in sequence j.

Table 3.4: Summary of the terminology and notation used for the images in the database. 'Image' denotes the full captured image and 'region' denotes an extracted part of the image. The sequences of images are denoted IS and the sequences of regions RS, the upper index indicates the sequence number and the lower index indicates the image/region number (and are sometimes omitted).

detector at several positions an image is obtained. Formally this can be written

$$I(x, \sigma) = G_0(x, \sigma) * S(r) \tag{3.2}$$

here $*$ denotes the convolution operator. The σ is called the inner-scale and denotes the size of the point operator. One may think of a point operator as measuring the light comming from a point in the scene, but that is of course not true because zero size (zero-scale) observation does not exist. Instead the point operator should be viewed as a measurement over a small region, modeled with a guassian kernel, and the size of the region is determined by σ. So σ is the spatial resolution of the point operator and sets the limit of details that can be detected. The image captured by a point operator is always a "blurred version" of a point in the scene. By increasing σ the spatial resolution decreases and the point becomes more "blurred".

Given an image captured using a fixed inner-scale σ, images of lower spatial resolution can be studied using linear scale space (see e.g. [98, 111, 202, 122, 188]). Coarse scale representation of an image can be generated

using linear Gaussian scale space by convolving the observed image with a Gaussian function.

Capturing a scene at a different scale, connotes that σ in equation (3.2) has been changed, by adjusting the objective (zooming). By adjusting the objective, the scene will be captured using a different inner-scale and different levels of details will be suppressed. Furthermore, by changing the focal length, the sampling density will also be altered. Increasing the viewing distance will increase σ in the point operator and sampling points will be less dense. Wu *et al.* [205] use an image pyramid approach [32] for describing image content transformation when the viewing distance increases. They use block average 2×2 block to increase the inner-scale and subsampling is used to reduce the resolution.

3.4 Statistics of Natural Images

In the following sections some classical results regarding natural image statistics are reviewed and verified on the MS-GT database. By comparing the classical results with the results from the MS-GT DB the soundness of the images in the database will be verified and the classical results are verified on a new image database.

Most of the classical results in natural image statistics are based on empirical studies, on large image databases (often van Hateren's [197]) containing natural images. The MS-GT DB contains an ensemble of sequences of images. Each sequence contains the same scene captured at different scales.

The statistics in the following sections have been computed after transforming the RGB-images into gray value images.

3.4.1 Scale Invariance

One of the earliest result in the area of characterization of natural images is the (apparent) scaling invariance [145, 166, 167, 168, 94, 196]. The scaling invariance property was first formulated as: the power spectra of a large ensemble of natural images is follow a power law

$$S(\omega) = \frac{A}{|\omega|^{2-\eta}} \tag{3.3}$$

where ω is the spatial frequency, and A is a constant that depends on the overall contrast in the image. η is usually a small value and values close to 0.2 have been reported [196, 168, 94]. It should also be noted that η depends on the type of images and that small image databases with specific

contents - for example beaches and blue skies - may have η far from 0.2. Torralba and Oliva [192] use the distribution of the power spectra to estimate the distance between the scene and the camera in individual images [192] and to characterize the image in terms of man-made environment or nature environment [193].

The scale invariance property of natural images can also be expressed in the spatial domain using the correlation function (see [168]). The correlation function $C(x)$ where x is the separation distance between two pixels in an image is

$$C(x) = E(I(x_0)I(x_0 + x)) \tag{3.4}$$

and it reveals information about how intensities are correlated solely based on the distance between intensities. The correlation is computed by considering all images in the ensemble, all initial positions x_0 and all displacements vectors x. The power spectra power law (3.3) expressed in the spatial domain using the correlation function, takes the following form

$$C(x) = C_1 + \frac{C_2}{|x|^\eta} \tag{3.5}$$

where η is the same expositional as in (3.3). The intensity correlation decreases with the distance between the pixels.
On the MS-GT database η was estimated using log-log regression to 0.202. The highest estimation of η for a single image was $\eta = 0.52$ and the lowest was $\eta = -0.36$.

In the top row of figure 3.3, 4 images with small η are shown and in the bottom row 4 images with large η are shown. One striking difference between the images is the presence of the sky in the top row, while the sky is absent in the bottom row. The sky is a smooth and rather uniformly colored region, spatially extended especially in the y direction. The presence of the sky in an image has a large influence on he power spectra and the intensity correlation. The presence of the sky in an image implies a long range intensity correlation. The viewing distance, the distance to the main object in the captured scene, for the images in the top row is rather large, while the viewing distances for the bottom row images are rather small. Buildings, trees (forest) and lawns appear as rather homogenous regions viewed from a large distance.

In figure 3.4, η has been estimated on the different capture scales (i.e. IS_i where $i = 1, \cdots, 15$). Estimation of η is plotted against the capture scales. As the capture scale increases, η also increases. For the first four capture scales, η increases rather rapidly, while for the remaining capture scales the increase is less rapid (and not monotonic). The sky is present and

Figure 3.3: The power spectra for natural images follows a power law in spatial frequency. η is estimated to 0.202 for the images in the database, which is similar to the results reported by other researchers. Estimating the power low parameters for individual images shows large variation. The top row contains images with small η (≈ -0.3) and the bottom row contains images with large η (≈ 0.5). In the images shown in the first row a large part of each image is occupied by the sky, while the sky is absent in all images in the bottom row. Also note that the average distance to the main object in the scene is much larger in the images in the first row than in the second row.

occupies a large part of the image in many of the images at small capture scales (i.e. large viewing distances). As the viewing distance decreases, the sky is occupying a smaller region of the image. Furthermore, buildings, lawns and trees appear as rather uniformly regions at larger viewing distances, as the viewing distance decreases, more details appear and the regions appear to be less homogenous.

In figure 3.5, typical and untypical sequences are shown. The first three rows show three sequences where η has been estimated on the individual images. The estimations of η at the different viewing distance is following the same pattern as for the ensemble (shown in figure 3.4). In the sense that they follow the same pattern as the ensemble of images, they are considered to be 'typical' sequences. At large viewing distance the sky is present, and the objects in the scenes appear to be rather homogenous because of the viewing distance. As the distance decreases the sky is occupying a smaller region of the image and more details have emerged.

The subsequent three rows show three 'untypical' sequences. The birch tree bark sequence contains small scale details on all viewing distances, therefore are the estimations of η large on all viewing distances. In the bush sequence, estimations of η are rather stable and do not vary much. At larger viewing distances the bush and the lawn are rather homogenous regions, as

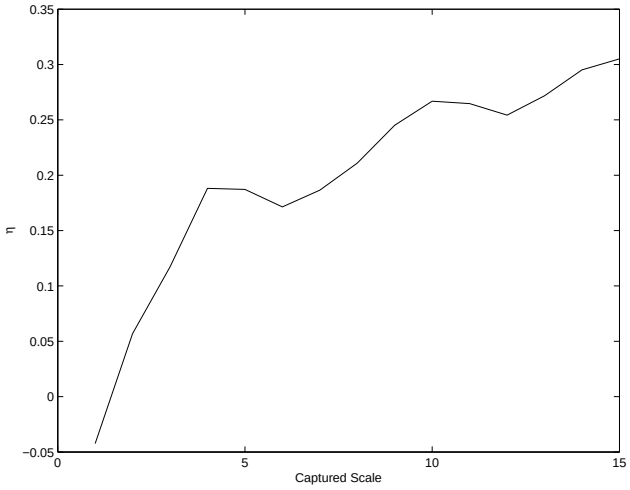


Figure 3.4: Estimation of η as a function of capture scale, where index 1 is the least zoomed and 15 is the most zoomed. Note that the capture scale is non-linear (see table 3.2), the increase in magnification is larger for the smaller index. η is increasing if the viewing distance decreases computed over an ensemble of images. In terms of intensity correlation, expressed in equation 3.5, the correlation decreases as the viewing distance decreases. This can partly be explained by the presence of the highly correlated sky at larger viewing distances. Furthermore, large scale objects such as trees and houses appear to be more homogeneous viewed from larger distances.

the viewing distance decreases more details emerge. Highly correlated leaves with sharp boundaries emerge, and as the viewing distance decreases details on the leaves emerge. In the Malmö harbor sequence the sky and the ocean is occupying a large region of the image at all viewing distances, therefore the estimations of η are small on all viewing distances.

3.4.2 Laplacian distribution of Linear Filter Responses

It has been reported, [30, 126, 173, 174, 94, 117], that the distribution of the partial derivatives of an ensemble of natural images can be modeled by an Generalized Laplacian Distribution

$$p(x) = \frac{1}{Z} e^{-|\frac{x}{s}|^{\alpha}} \tag{3.6}$$

where α and s are parameters estimated from an ensemble of natural images. The parameters s and α are related to the variance and kurtosis. The kurtosis κ and the skewness $\mathcal{S}$ for a random variable X is defined as

$$\kappa = \frac{E(X - m_x)^4}{\sigma^4} \text{ and } \mathcal{S} = \frac{E(X - m_x)^3}{\sigma^3} \tag{3.7}$$

where E is the expectation, m_x is the mean (an estimation of $E(X)$) and σ is the standard deviation (and σ^2 is the variance). The relation

$$\sigma^2 = \frac{s^2 \Gamma(\frac{3}{\alpha})}{\Gamma(\frac{1}{\alpha})} \text{ and } \kappa = \frac{\Gamma(\frac{1}{\alpha})\Gamma(\frac{5}{\alpha})}{\Gamma^2(\frac{3}{\alpha})} \tag{3.8}$$

can be used for estimation of s and α. A more model fitting approach, such as Kullback-Leibler divergence, Least Square Error (LSE) and Maximum Likelihood (ML), can also be used for estimating the parameters.
Natural images are in general not differentiable, therefore a (linear) scale space approach is adopted, the derivative of an image is the scale space derivative at a fixed scale. The scale space partial derivative in x is defined as

$$\frac{\partial}{\partial x}(G_t * I) = \frac{\partial G_t}{\partial x} * I \tag{3.9}$$

where $*$ denotes the convolution operator and G_t is the guassian function

$$G_t(x, y) = \frac{1}{2\pi t} \exp(-\frac{x^2 + y^2}{2t}). \tag{3.10}$$

Instead of using the scale space derivative at a fine scale, the intensity difference for adjacent pixels as an linear filter could have been used. The

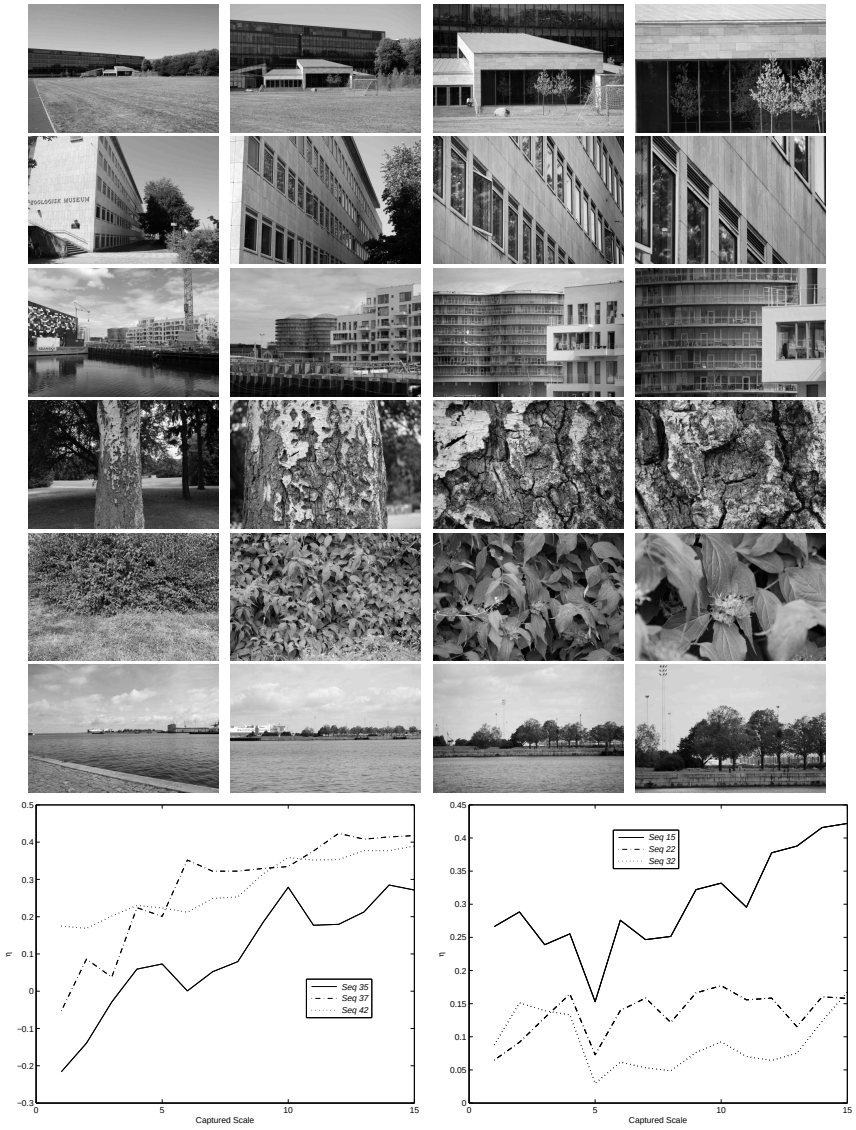


Figure 3.5: Example of η estimated on image sequences. The first three rows contain 'normal' sequences and subsequent three rows contain 'un-normal' sequences. The plots show η against the captured scales, for the 'normal' (left) and 'un-normal' (right) sequences. In the 'normal' sequences the sky occupies a large, but decreasing region in the photo. The 'un-normal' sequences contain either small scale details (tree trunk) or geometric structures (ocean) at all scales.

benefit of using the scale space derivative is the explicit scale formulation and the possibility to use different scales. The scale space derivatives have been computed on $log(I+1)$, and $t=1$ is the smallest scale.

Compared with the Gaussian distribution, the Generalized Laplacian distribution (usually) has a sharper peak at zero and 'heavy tails'. Most natural images contain homogenous regions, objects under the similar illumination, with similar or smoothly varying intensities, which correspond to the sharp peak at zero. At the object boundary the intensities change rapidly, which corresponds to the 'heavy tails'. The α parameter relates to the sharpness of the peak, while s relates to the width of the distribution.

Yanulevskaya and Geusebroek [207] analyzed the relation between α and the image content in (individual) images and image patches. (see also Geusebroek and Smeulders [72, 71]). Three sub-models are identified: power law, exponential and gaussian distribution. The appropriate image model is selected using Akaike's information criterion (AIC). Typically images with a well separated foreground and uniform background are following a power law, while images with a lot of details at different scales follow an exponential distribution, and images containing mainly high frequency texture are following a gaussian distribution.

In figure 3.6, α_x has been estimated on individual images. The first two rows contain images with large α_x value estimated using $t=1$ ($\alpha_x \approx 1.00$) and $t=64$ ($\alpha_x \approx 2.00$). The images contain small scale details, where small relates to the t used in the scale space derivative. The subsequent two rows show images with low α_x value, estimated using $t=1$ ($\alpha_x \approx 0.25$) and $t=64$ ($\alpha_x \approx 0.55$). The images contain large scale geometric structures such as the sky and the buildings.

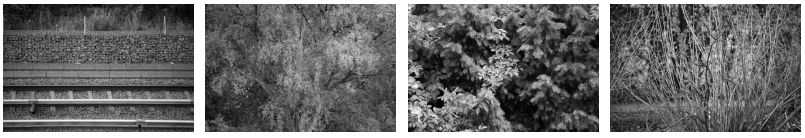
In the last row, in figure 3.6, the empirical distribution using the ensemble of images and the corresponding generalized Laplacian distribution is shown for $t=1$ and $t=64$. $\alpha_x = 0.37$ for $t=1$, and $\alpha_x = 0.78$ for $t=64$.

In figure 3.7, estimation of α_x using different capture scale (IS_i) are plotted. On the y-axis is α_x and on the x-axis is the capture scale. α_x is estimated using four different t in the scale space derivative; $t = 1, 4, 16$, and 64. As it seems α_x does not follow any trend (for any t). As a function of capture scale, α_x does neither increase or decrease, instead it seems to be stable with some variation.

Bessel K form for Natural Images

Related to the Generalized Laplacian Distribution is the so-called (statistical) Bessel K forms proposed by Grenander and Srivastava [78] and Srivastava *et al.* [184]. The Bessel K form is derived using the transport generator model

(a) Large α_x at $t = \sigma^2 = 1$



(b) Large α_x at $t = \sigma^2 = 64$

(c) Small α_x at $t = \sigma^2 = 1$

(d) Small α_x at $t = \sigma^2 = 64$

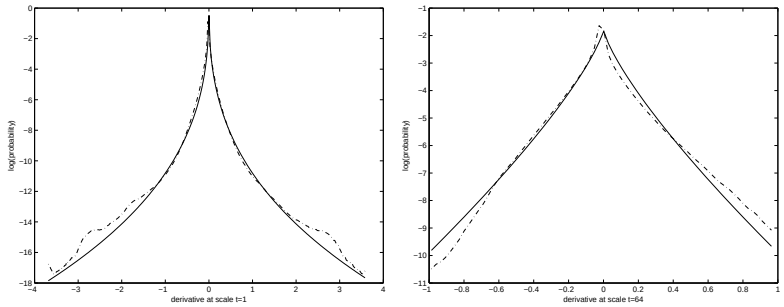


Figure 3.6: Estimation of α_x in the generalized Laplacian distribution using scale space derivatives at scale $t = 1$ and $t = 64$. The first two rows contain images with large α_x at scale $t = 1$ ($\alpha_x \approx 1.00$) respective $t = 64$ ($\alpha_x \approx 2$). The following two rows contain images with small α_x at scale $t = 1$ ($\alpha_x \approx 0.25$) respective $t = 64$ ($\alpha_x \approx 0.55$). α_x is large if the image contains small scale details and is small if it contains large scale geometric structures (where 'small' and 'large' are defined by the inner scale t). The last row shows plots of empirical distribution using the ensemble of images and the corresponding generalized Laplacian distribution at scale $t = 1$ (left) and $t = 64$ (right).

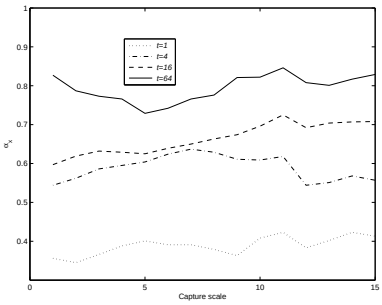


Figure 3.7: The plots show estimations of α_x in the Laplacian distribution estimated using IS_i, where α_x is on the y-axis and the captured scale is on the x-axis. $t = 1$, $t = 4$, $t = 16$ and $t = 64$ in the gaussian function are shown. No trend as a function of the capture scale can be found in the estimations, instead the estimations of α_x are rather stable at the different captured scales.

and it models the image formation process.

Images are generated by projecting 3D objects onto the 2D image, resulting in a set of so-called 2D profiles - g_i - representing the different objects in the scene. To make up an image, 2D profiles interact in a non-linear way - occlusion, scaling and superposition. Under some simplified statistical assumption on the distribution, scale and location of the generators g_i, the authors show that the marginal distribution of linear filters are following the **Bessel K Form** distribution

$$p(x; p, c) = \frac{1}{Z(p, c)}|x|^{p-0.5}\mathbf{K}_{(p-0.5)}(\sqrt{\frac{2}{c}}|x|) \tag{3.11}$$

where Z is a normalization constant and $\mathbf{K}$ is the modified Bessel function of second kind. The parameters p and c are called the Bessel Parameters and can be estimated using the following equations

$$p = \frac{3}{\mathcal{S} - 3} \tag{3.12}$$

and

$$c = \frac{\sigma}{p} \tag{3.13}$$

where σ and $\mathcal{S}$ are estimations from the filtered images.

As shown in [78, 184] the Bessel K form models the partial derivatives of individual images well. Furthermore the Bessel form parameter p relates to

the object present in the image. The p value depends on the distinctness of the edges and on the frequency of the edges. Images with large objects, with sharp boundaries have a low p value, while images with many objects have a large p value. Images containing large geometric structures will in general have a low p value, while images containing small scale textures will have a high p value.

The results of estimating p on the database is very similar to the estimations of α in the generalized Laplacian distribution. Yanulevskaya and Geusebroek [207] explain the relation between estimation of α and the visual content in a similar way as Grrenander and Srivastava connect the visual content with estimation of p.(See also [73].)

3.4.3 Size Distribution in Natural Images

Area Distribution in Natural images

As discussed, nearby pixel intensities are highly correlated and the correlation decrease with the distance according to a power law. It is natural to consider how the size of homogenous regions in natural images are distributed. Alvarez *et al.* [3, 1, 76, 77] analyze the size distribution of homogenous regions, in terms of area and perimeter, in natural images and they show that the size distribution of homogenous regions in natural images follow a power law. Alvarez *et al.* [3, 1], show that the size distribution of homogenous regions is following a power law, both estimated on individual images and on ensemble of images. Following Alvarez *et al.*, we will verify their result on the database and analyze the behavior as a function of capture scale.

A Homogenous region can be defined in many ways depending on the problem at hand. Our interest is to characterize natural images with respect to the distribution of size, therefore a very simple approach is suitable. Let I be a image of size $M \times N$ with intensities in $\{1, \cdots, G\}$ and let $k \in \{1, \cdots, G\}$. Histogram equalization such that the number of intensities are k and the number of pixels is (approximative) the same - $\frac{M \cdot N}{k}$ - for all intensities. After the histogram equalization a homogenous region is defined as the set of connected - using either 8 or 4 connectivity - pixels with the same intensity. The size of a homogenous region is defined as the number of pixels in the region.

The area distribution of homogenous regions are following a power law

$$f(s) = \frac{A}{s^{\alpha}} \tag{3.14}$$

where s is the area, A and α are an image dependent parameters. The

parameters α and A can be estimated by regression on the set

$$\{(log(f(s)), s) : s \in 1, \cdots, T_{max}\} \tag{3.15}$$

where T_{max} is the smallest size s for which $f(s)$ is zero. For an ensemble of natural images the T_{max} value is large and covering almost the full range of size distribution. For an individual image T_{max} value can be small and a large range of the size distribution will not be used in the regression.

For ensembles of natural images $\alpha \approx 2$, for individual images the α varies. For images containing larger geometric structures, α is often smaller, approximately 1.7, while for images containing small scale texture α is often larger, approximately 3.0.

In figure 3.8, images with the small $\alpha \approx 1.57$ (figure a) and large $\alpha \approx 3.00$ (figure b) are shown. The content difference is striking. The images with a small α mainly contain large scale geometric structure, and the distance to the main objects in the scene is large. The images with a large α contain small scale details (texture), and the distance to the main objects in the scene is small. α is estimated to 2.11 on the ensemble of images. Figure 3.8, also shows the empirical distribution and the estimated power law (in log-log scale) for a small α and a large α, and the fit is good in both cases.

Figure 3.9, shows estimations of α, estimated on different capture scale (IS_i). The α:s (y-axis) are plotted against the capture scales (x-axis). The lowest estimation of α is 2.15 and the largest is 2.22. Furthermore, no trend can be found in the estimations. α seems to neither decrease, nor increase as the viewing distance decreases.

Directional Homogenous Region Size

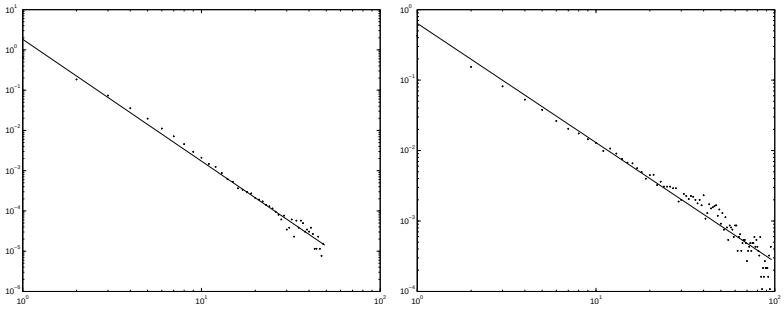
In previous section, the area distribution of homogenous region in individual images and an ensemble of image was shown to follow a power law - equation 3.14. The orientation of the homogenous region was not considered. In the following section the 'size' of homogenous region in the x and y directions are analyzed. Because natural images are more correlated in x direction, than in y direction the size distribution of homogenous regions in the different direction may be different.

The image intensity resolution is reduced to k intensities using histogram equalization, and the regions with the same intensity are considered to be homogenous.

The intersection length of a homogenous region along a direction (the x and y direction in our case) is the number of connected pixels with equal intensity. By collecting all intersection lengths of homogenous regions along

(a) Small $\alpha \approx 1.75$

(b) Large $\alpha \approx 3.00$



(c) The empirical distribution and the estimated regression

Figure 3.8: Distribution of homogenous regions in different types of images. Figure (a) contains examples of images with small α. The images contain mainly large scale geometric structures such as the sky and buildings viewed from distance. Figure (b) contains images with small α. The images contain small scale texture and the distance to the main objects in the scene is quite small. Figure (c) contains the empirical distributions and the estimated power law (in log-log scale) for a small α (left) and a large α (right). Estimated on the ensemble of images $\alpha = 2.11$.

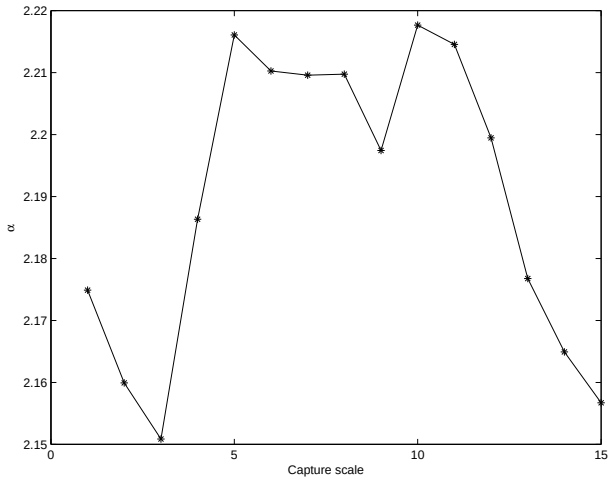


Figure 3.9: The plot shows estimations of α in the power law for the area distribution - equation 3.14 - using IS_i, where α is on the y-axis and the capture scales are on the x-axis. No trend in the estimation can be found - increasing or decreasing over the scales.

a fixed direction in an image, the distribution of intersection lengths of homogenous regions in the direction is computed. The x and y directions will be used but any other direction could also be used.

In figure 3.10, three different homogenous region which are covering the same area are shown. The distribution of intersection length in x and y directions is different for the region. One region is extended in x direction while the other region is extended in the y direction. The last region is connected in the 8-connectivity sense, but on the top it is not connected in the x direction. The intersection length distribution will therefore be two small intersections.

Analyzing the intersection length distribution for homogenous regions indicates that it follows a power law (as in 3.14) in intersection length, with different value for α in x and y direction. Estimating the α_x and α_y using log-log regression on all full images in the database gives $\alpha_x = 2.96$ and $\alpha_y = 3.55$ as shown in figure 3.11. Homogenous regions extend longer in the x direction than in the y direction. This supports the fact that natural images are more correlated in the x-direction than in the y-direction.

Figure 3.10: Three different homogenous regions with the same area but with different shape and/or orientation. The region to the left is longer in y-direction, than in the x-direction. The region to the middle is longer in the x-direction, than in the y direction. The distribution of intersection length in x and y direction for the two regions is different. The region to the right is not connected on the top in the x direction, the intersection length distribution will therefor be two short intersection.

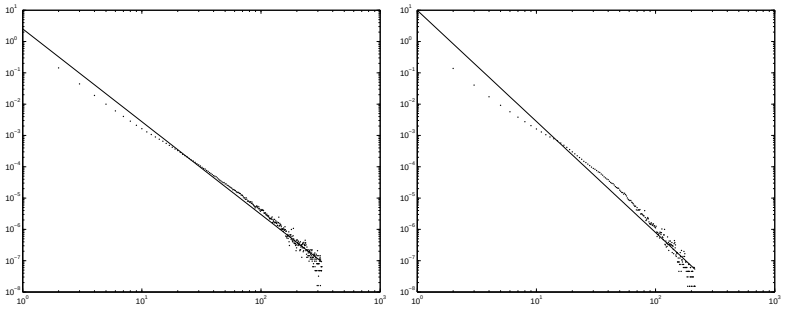


Figure 3.11: Log plot of the intersection length distribution in x (left) and y (right) direction, together with the regression line estimated using all full images in the database. $\alpha_x = 2.96$ and $\alpha_y = 3.55$ which shows that homogenous regions are longer in the x direction than in the y direction, which is also indicated by, and consistent with, the higher correlation in the x direction than in the y direction.

Figure 3.12: The two images with largest (left) and smallest (right) difference between α_x and α_y. For the image (IS_{11}^1) with the largest difference $\alpha_x = 2.14$ and $\alpha_y = 3.63$, the homogenous regions are longer in the x direction than in the y direction. For the image (IS_6^3) with the smallest difference $\alpha_x = 2.14$ and $\alpha_y = 2.13$, the homogenous regions have the same extension in x and y direction.

3.5 Discussion

A new database containing an ensemble of sequences of natural images is collected. Each sequence contains the same scene captured at different scales by adjusting the focal length. Natural images, or rather natural scenes, are vaguely defined as everyday scenes observed by a human from a human perspective. The definition includes both nature and man-made structures, but exclude 'bird views', because they are not considered to be from a human perspective.

Three classical and well known results from natural image statistics are verified on the database.

The apparent scale invariance of an ensemble of natural images, which can be expressed as the power spectra of an ensemble of natural images, follows a power law in spatial frequencies. We estimated $\eta = 0.202$ in the power law on the database. Ruderman and Bialek [166] estimated $\eta = 0.19$ on their database collected in the woods, Huang and Mumford [94] estimated also $\eta = 0.19$ on the van Hateren image database [197] and van der Schaaf and van Hateren [196] estimated $\eta = 0.12$ on their natural image database. The estimation of η on the database is similar to the previously reported results.

The distribution of partial derivatives of an ensemble of natural images can be modeled with a generalized Laplacian distribution. The partial derivative of the image was defined as the scale space derivative at scale t. We estimated $\alpha_x = 0.37$ and $\alpha_x = 0.78$ at scale $t = 1$ and $t = 64$. For comparison, Huang and Mumford estimated $\alpha = 0.55$ on the van Hateren database [197].

The size distribution of homogenous regions in natural images follows a

power law. We estimated $\alpha = 2.11$ in the size power law. Alvarez *et al.* [3, 1, 2] reported α close to 2. Again, our result is similar to other reported results.

Estimation of the three statistics on individual images can to some degree explain the visual content of the image. Furthermore, the estimations depend highly on the viewing distance.

Estimation of η in the power spectra power law is in general large if the image is captured at small viewing distance and the scene contains small scale details. η is small if the viewing distance is large and the scene contains large, uniformly colored regions such as trees, lawns and buildings.

Estimation of α_x in the Laplacian distribution of the partial derivatives, is usually large if the viewing distance is small and the scene mainly contains small scale details. α_x is small if the viewing distance is large and the scene mainly contains geometric structures.

Estimation of α in the distribution of homogenous regions is large if the viewing distance is small and the scene contains small scale details. α is small if the viewing distance is large and the scene contains large scale geometric structures.

The relation between the estimation and the viewing distance can be explained by two different factors: the spatial composition - the spatial envelope - of the scene and the inner scale. In images captured at large viewing distances, the sky is often occupying a large region in the image (i.e spatial composition) and buildings, trees and lawns often appear as uniformly colored regions (i.e. inner scale is rather large). At a small viewing distance the sky is absent or occupying a small region in the image (i.e. spatial composition) and the details on the trees, the bushes and the lawns are brought out (i.e. the inner scale is smaller).

Chapter 4

A SVD-Based Image Complexity Measure

This chapter contain a slightly re-formatted version of

David Gustavsson, Kim S. Pedersen, and Mads Nielsen.
A SVD Based Image Complexity Measure.
In *Proceding of International Conference on Computer Vision Theory and Applications (VISAPP) 2009*, 2009.

A SVD Based Image Complexity Measure

David Gustavsson, Kim S. Pedersen, and Mads Nielsen
DIKU, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark
{davidg,kimstp,madsn}@diku.dk

abstract

Images are composed of geometric structures and texture, and different image processing tools - such as denoising, segmentation and registration - are suitable for different types of image contents. Characterization of the image content in terms of geometric structure and texture is an important problem that one is often faced with. We propose a patch based complexity measure, based on how well the patch can be approximated using singular value decomposition. As such the image complexity is determined by the complexity of the patches. The concept is demonstrated on sequences from the newly collected DIKU Multi-Scale image database.
Keywords: Image Complexity Measure, Geometry, Texture, Singular Value Decomposition, SVD, Truncated Singular Value Decomposition, TSVD, Matrix Norm

4.1 Introduction

Images contain a mix of different types of information, from highly stochastic textures such as grass and gravel to geometric structures such as houses and cars. Different image processing tools are suitable for different type of image contents and most tools are very image content dependent. The definition of what is texture and geometry is not particularly agreed upon in the computer

vision community. Our hypothesis is that the separation between geometry and texture is defined through the purpose of the method and the scale of interest. What may be considered an unimportant structure / texture in one application may be considered important in another.
For example, segmentation of an image containing objects with clear geometric structures forming boundaries calls for edge-based or geometry-based methods such as watersheds [148], the Mumford-shah model [141], level sets [171], or snakes [104]. While segmentation of an image containing objects only discernable by differences in texture calls for texture based segmentation methods [162]. That is, the type of objects we are attempting to segment defines our scale of interest, i.e. what type and scale of structure we include in the model of a segment.
In denoising an image containing geometric structures calls for e.g. an edge preserving method such as anisotropic diffusion [200] or total variation image decomposition [169]. For images containing small scale texture, a patch based denoising method such as non-local mean filtering may be more appropriate [29]. Again we see that depending on the purpose we include structures at finer scales into the model of the problem as needed.
As a final example, we mention that total variation (TV) image decomposition, and other functional base methods, are very successful for inpainting images containing geometric structures [39]. Unfortunately the functional based methods fails to faithfully reconstruct regions containing small scale structures, however texture based methods manage to reconstruct such images [52, 48, 86, 49]. In the functional approaches the focus is solely on large scale structures or geometry, whereas in the texture methods small scale texture is included in the model.
Prior knowledge about the methods and the image content are therefore essential for successfully solving a task. A natural question is: "For a given type of images, which type of methods are suitable?" Often one wants to characterize the methods by analyzing the type of images that it is (un)suitable for. To be able to characterize the methods in this way, the images must be characterized with respect to the image contents. An image complexity measure is needed, i.e. a measure that quantify the image contents with respect to geometric structure and texture or scale of interest.
A patch based complexity measure using Singular Value decomposition (SVD) is presented. The complexity for the patch is determined by the number of singular values that are required for good approximation - the matrix rank of a good approximation. The number of singular values that are required for approximating an image patch is used for characterizing the patch content. The global complexity measure for the image is computed as the mean complexity of all patches in the image. The proposed complexity measure is

evaluated on the baboon image and on the newly collected DIKU Multi-Scale image sequence database.

4.2 Complexity Measure

In the following section images are viewed as matrices, hence the image complexity measure transforms into a matrix complexity measure. Basic matrix properties are used extensively in the following section, which can be found in e.g. [75]. One obvious approach is to approximate a matrix A with a simpler matrix A_k and measure the error (residual) between the original matrix A and the approximation A_k. Here k is a parameter used for computing the approximation A_k. We assume that, as the parameter k increases the error between A and A_k decrease (or at least not increase) and as $k \to \infty$ the error becomes 0. The approximation A_k should also be simpler than A. To be able to use this approach, an error measure between matrices and a matrix complexity measure must be defined.

4.2.1 Error Measure - Matrix Norms

To measure the difference between the original image A and a simpler approximation A_k of I, it is natural to use a matrix norm $\|A - A_k\|$. One of the most commonly used matrix norms is the Frobenius norm (which corresponds to the L^2-norm). Let A be a $m \times n$ matrix with elements a_{ij}, the Frobenius norm of A is defined as

$$\|A\|_F = (\sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}|^2)^{\frac{1}{2}}. \tag{4.1}$$

Another common type of matrix norms are the so-called induced matrix norms. Let A be a $m \times n$ matrix and $x \in R^n$ a colon vector (i.e $x = (x_1, \cdots, x_n)^T$), the matrix norm induced by the vector norm $\|x\|$ is defined as

$$\|A\| = \sup_{\|x\|=1} \frac{\|Ax\|}{\|x\|} \tag{4.2}$$

(or in words the smallest number α such that $\frac{\|Ax\|}{\|x\|} \leq \alpha$ for all x). The matrix norm is here defined in terms of a vector norm $\|x\|$. The induced matrix norm can be viewed as how much the matrix A expands the vectors and is actually an operator norm. Different vector norms can be used to induce different matrix norms, most common are the p-norms defined as

$$\|x\|_p = (\sum_{i=1}^{n} \mid x_i \mid^p)^{\frac{1}{p}} \tag{4.3}$$

and especially the 2-norm $\|x\|_2 = (x^T x)^{\frac{1}{2}}$. The matrix norm induced by the 2-norm is

$$\|A\|_2 = \sup_{\|x\|_2=1} \frac{\|Ax\|_2}{\|x\|_2} \tag{4.4}$$

Both the The Frobenius matrix norm and the matrix 2-norm are invariant under orthogonal transformation and will be used in the following sections.

4.2.2 Matrix Complexity Measure - Matrix Rank

Given a matrix A, a simpler matrix approximation A_k of A should be constructed. But first one must define what 'simpler' means. A natural approach to quantify complexity of a matrix is by the rank of the matrix, and a simpler approximation of a matrix can be viewed as a matrix with lower rank.
Let A be a $m \times n$ matrix then the rank of A can be viewed as the dimension of the subspace spanned by the columns of $A = (a_1, \cdots, a_n)$,

$$rank(A) = dim(\ span\{a_1, \cdots, a_n\}\ ). \tag{4.5}$$

4.2.3 Optimal Rank k Approximation

It is well known from matrix theory that a $m \times n$ matrix A can be decomposed into

$$A = U\Sigma V^T \tag{4.6}$$

where U is a $m \times m$ orthogonal matrix, V is a $n \times n$ orthogonal matrix and Σ is a $m \times n$ diagonal matrix with elements $\sigma^1, \cdots, \sigma^l$ where $l = min\{m, n\}$. This is the so-called Singular Value Decomposition (SVD), where the σ^i:s are called singular values and the column vectors u^i and v^i, of U and V are called singular vectors. The entries in Σ is ordered such that $\sigma^1 \geq \sigma^2 \geq \cdots \geq \sigma^l \geq 0$.
Using the fact that the Frobenious norms are invariant under multiplication by orthogonal matrices gives

$$\|A\|_F^2 = \|\Sigma\|_F^2 = \sum_{i=1}^{l} (\sigma^i)^2. \tag{4.7}$$

Let Σ_k be the $m \times n$ matrix containing the k largest singular values on the diagonal and let

$$A_k = U\Sigma_k V^T. \tag{4.8}$$

A_k is the so-called Truncated Singular Value Decomposition (TSVD) approximation of A where the first k singular values are used, and if $rank(A) \geq k$ then $rank(A_k) = k$. The image approximation residual is defined as $A - A_k$ and if, again, $rank(A) \geq k$ then $rank(A - A_k) = rank(A) - k$.
The reconstruction error or the residual error for the Frobenious norm is

$$\|A - A_k\|_F = (\sum_{i=k+1}^{l} (\sigma^{(i)})^2)^{\frac{1}{2}} \tag{4.9}$$

and for the 2-norm

$$\|A - A_k\|_2 = \sigma_{k+1}. \tag{4.10}$$

The $rank(A_k) \leq rank(A)$, so A_k is simpler in the sense that its' rank is not larger (and usually the rank is lower). Furthermore A_k is the best $rank - k$ approximation of A in the sense that

$$A_k = arg\ min_{rank(B)=k} \|A - B\|_2 \tag{4.11}$$

So any matrix B with rank k has at least as large reconstruction error using the 2-norm as A_k. A_k is also the best rank k approximation using the Frobenious norm. Singular Value Decomposition can be viewed as a method for finding the optimal basis and is related to other optimal basis methods such as Independent Component Analysis (ICA) [96] and Karhunen-Love Expansion [109].
There are two possibilities to compare images by comparing the norm of the residual. Either the number of singular values, k, are fixed and the reconstruction error $\|A_k - A\|$ using k singular values are compared. The other possibility is to keep the reconstruction error fixed, σ^{err}, and use as many singular values that are required for the reconstruction error to be lower than σ^{err}. Either the rank k or the reconstruction error σ^{err} is kept fixed.
Let k_0 be the number of singular values that should be used in the reconstruction. The residual error (using either the 2-norm or Frobenious norm) is

$$\|A - A_{k_0}\| = \sigma_{k_0}^{err} \tag{4.12}$$

and $\sigma_{k_0}^{err}$ is called the singular value reconstruction error using k_0 singular values.
Let σ^{err} be a fixed reconstruction error and let k be the smallest integer such that

$$\|A - A_k\| \leq \sigma^{err} \tag{4.13}$$

k is called the singular value reconstruction index (SVRI) at level σ^{err}. The SVRI state the smallest number of singular values that are required to get a reconstruction with a reconstruction error smaller than σ^{err}.

4.2.4 Global Measure

Instead of computing an approximation of the full image, which is not feasible for high resolution images, a patch based approach is adopted. The singular value reconstruction error at level σ^{err} is computed for each $p \times p$ patch in the image.
Based on the patch complexities an image complexity measure should be computed. The obvious candidate is the mean or the mode complexity computed over all patches in the image. The mean patch complexity is used as the complexity measure for the image. The interpretation of the mean, is simply the average number of singular values that are required for an approximation, such that the reconstruction error is less than σ^{err}, of the patches in the image.

Figure 4.1: Image sequences - 02, 05 and 08 - from the DIKU Multi- Scale image database (used in the experiments) at three capture scales.

4.3 DIKU Multi-Scale Image Database

The newly collected DIKU Multi-Scale image database [83], contains sequences of the same scene captured using varying focal length - called capture scales -, will be used to analyze the distribution of singular values in natural image patches and analyze how the image content changes over different capture scales.
The database contains sequences of natural images - both man-made and natural environment - with a large variety of scenes and distances to the main object in the scene. Each sequence contains 15 high resolution images of the same scene captured using different focal length. The zoom factor is roughly 16x and the naming convention is that image 1 is the least zoomed and 15 the most zoomed. Three examples of sequences are shown in figure 4.1.
Furthermore, the part of the scene that is present at all capture scales has been extracted, resulting in a sequence of region containing the same part of scene captured at different capture scales. The part of the scene present in the image to the right in figure 4.1, has been extracted from the remaining 14 images (of which two are shown in the figure).
Three sequences - 02 building with windows, 05 building without windows and 08 tree trunk - shown in figure 4.1 are used in the experiments. The image contents are very different on the different capture scales that can be seen in the 80×80 extracted patches shown in figure 4.2. For example, in the most zoomed image a brick is almost covering the whole 80×80 patch, while in the least zoomed image a large part of the brick wall is contained in the patch. (The 80×80 patches are only shown for visualization of the contents differences, while the complete regions are used in the experiments.)

4.4 Singular Value Distribution in Natural Images

The proposed method depends on the distribution of singular values in natural image patches. The distribution of principal component and independent components in natural images has received a lot of attention for some years, partly because its relation to the front-end vision [197].
To analyze the distribution of singular values in natural image patches, 1000 randomly selected 25×25 patches from each image in the DIKU Multi-Scale image database have been selected - approximately 800000 patches - and the corresponding singular values have been computed.

Figure 4.2: 80×80 patches extracted from the three sequence shown in figure 4.1 at 3 different scales (index 1, 6 and 15). The patches show the contents different at the different capture scales.

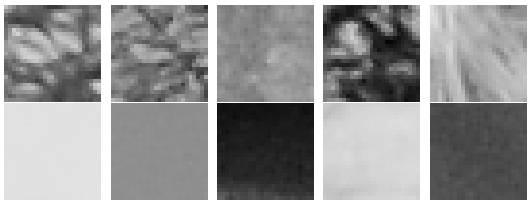


Figure 4.3: Each column show the patch with the largest (top) and smallest (bottom) σ_{25} in the same image. The content difference is striking and clearly indicate the importance for the small singular values for characterize the image content.

The first, not so surprising, conclusion is that patches in natural images almost always have full rank - i.e. the singular values are almost always strictly larger than 0.
The distribution of singular values σ_1 and σ_2 are shown in figure 4.4. The variance for the distribution of σ_1 is large, and it is interesting that many patches have values close to 25. The distribution for σ_2 is peaked at zero but also have 'heavy tails' - values relatively far from zero. This is also the case for σ_i where $i > 2$.
In figure 4.3 the patches with the largest σ_{25} (top) and smallest σ_{25} (bottom) in five different images are shown. The contents difference in the different patches are striking - the patches with the largest σ_{25} all contain large variations, while the patches with the lowest σ_{25} contain no or very little visible variations.
The distribution of the small singular values are peaked at zero, but also show some variation and 'heavy tails'. Visual comparison of patches with high and low σ_{25} clearly indicates a content difference, which implies that singular value reconstruction index is suitable for measuring image content.

4.5 Experiments

4.5.1 The baboon image

The baboon image is used only for demonstrating the method. The baboon is a good test image because it contains both very complex texture and large regions with geometric structures. In figure 4.5 the spatial distribution of complexity is shown using different patch sizes and error levels. White regions indicating high complexity and black indicating low complexity. The highly stochastic texture returns high complexity values at all scales and error levels, while the geometric structures return low complexity. As the patch size grows larger the spatial distribution of complexity gets smoother.

4.5.2 DIKU Multi-Scale Image Database

The image complexity measure is computed over the different capture scales using different patch sizes and error levels. The results are shown in figure 4.6.
The plot to the left and right, in figure 4.6, has the same error level 0.35, but different patch sizes, 15 respective 25 pixels. Still the shape of the curves are very similar. On the other hand the plot in the middle and to the right have same patch sizes - 25 pixels -, but different error level - 0.05 and 0.35 -

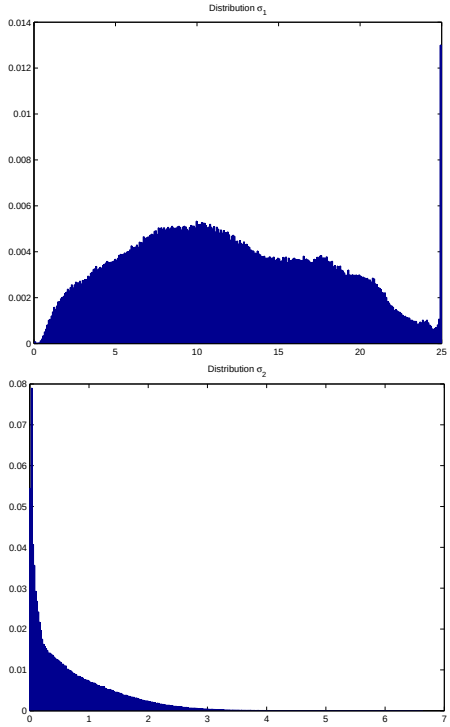


Figure 4.4: The distribution of singular values σ_1 and σ_2 for natural images patches of size 25×25. The variance for the distribution of σ_1 is large (as expected), the distributions for σ_2 is peaked at zero but also have 'heavy-tails'.

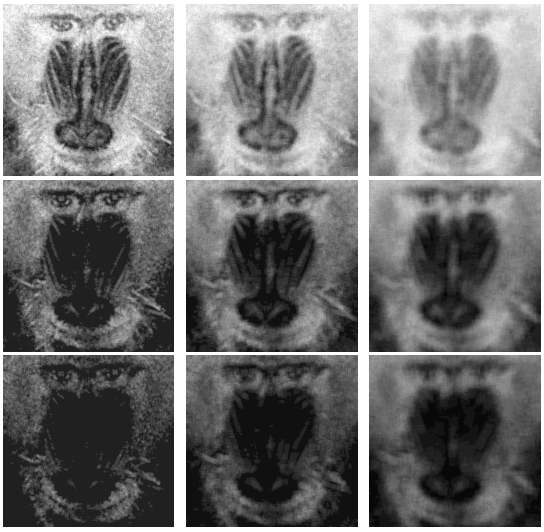


Figure 4.5: Patch based complexity measure of the baboon image. Different patch size are used in the colon, from left to right, the sizes are 9,15 and 25 pixels, and different reconstruction errors are used in the rows, from top to bottom, 0.1, 0.3, and 0.5.

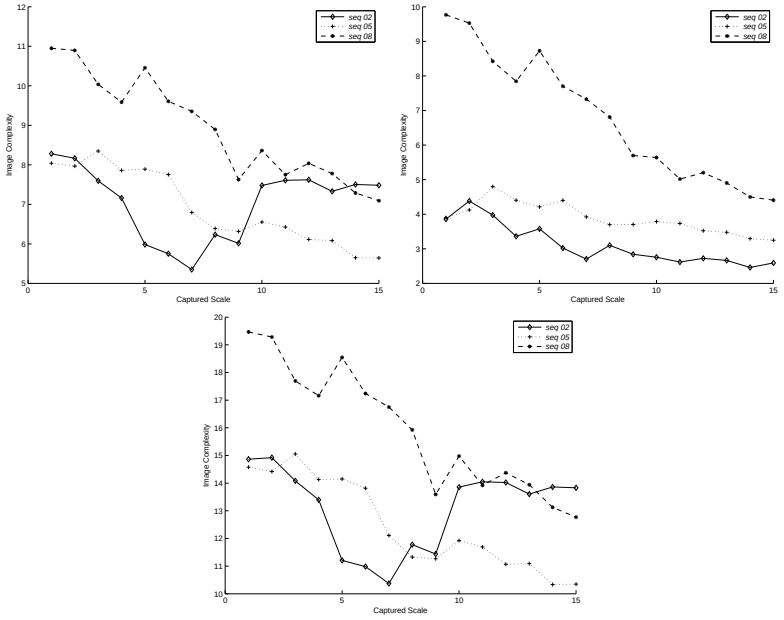


Figure 4.6: Complexity measure (y-axis)computed over different capture scales (x-axis) using different patch sizes and error levels. From left to right: patch size 15 and $\sigma^{err} = 0.05$, patch size 25 and $\sigma^{err} = 0.35$, and patch size 15 and $\sigma^{err} = 0.05$.

and the curves are very different which indicate that the error level is more important than the patch size.
For sequence 02 the complexity at error level 0.05 first decreases roughly for the first 7 capture scales, and then increases for the last 7 capture scales. For sequence 08 the complexity at error level 0.05 decrease quite rapidly at the first scales and then decreases slower for the remaining capture scales. For sequence 05 the complexity decreases with increasing capture scale.
The average number of singular values required for an approximation at a fixed error level varies a lot over the different capture scale. This indicate that the contents in terms of complexity, change over the capture scales which is clearly visiable from figure 4.2.

4.6 Conclusion

A patch based image complexity measure based on the number of singular values that are required to approximate a patch at a given error level is presented. The number of singular values is used to characterize the image content in terms of geometric structures and texture.
The proposed method is motivated by the optimal rank-k property of the truncated singular value approximation. The distribution of singular values in patches from natural images seems to be peaked at zero and have 'heavy-tails'. The image content in patches with relatively large smallest singular value are very different from the patches with relatively small smallest singular value.

ACKNOWLEDGEMENTS

This research was funded by the EU Marie Curie Research Training Network VISIONTRAIN MRTN-CT-2004- 005439 and the Danish Natural Science Research Council project Natural Image Sequence Analysis (NISA) 272-05-0256. The authors wants to thank prof. Christoph Schnörr (Heidelberg University) and PhD. Niels-Christian Overgaard (Lund University) for sharing their knowledge.

Chapter 5

On the Rate of Structural Change in Scale Spaces

This chapter contain a slightly re-formatted version of

David Gustavsson, Kim S. Pedersen, Francios Lauze and Mads Nielsen.
On the Rate of Structural Change in Scale Spaces.
In *Proceedings of Scale Space and Variational Methods in Computer Vision (SSVM) 2009*, 2009.

On the rate of structural change in scale spaces

David Gustavsson, Kim S. Pedersen, Francois Lauze and Mads Nielsen
DIKU, University of Copenhagen
Universitetsparken 1, DK-2100 Copenhagen Ø, Denmark
{davidg,kimstp,francois,madsn}@diku.dk

abstract

We analyze the rate in which image details are suppressed as a function of the regularization parameter, using first order Tikhonov regularization, Linear Gaussian Scale Space and Total Variation image decomposition. The squared L^2-norm of the regularized solution and the residual are studied as a function of the regularization parameter. For first order Tikhonov regularization it is shown that the norm of the regularized solution is a convex function, while the norm of the residual is not a concave function. The same result holds for Gaussian Scale Space when the parameter is the variance of the Gaussian, but may fail when the parameter is the standard deviation. Essentially this imply that the norm of regularized solution can not be used for global scale selection because it does not contain enough information. An empirical study based on synthetic images as well as a database of natural images confirms that the squared residual norms contain important scale information.
Keywords: Regularization, Tikhonov Regularization, Scale Space, TV, Total Variation, Geometric Structure, Texture

5.1 Introduction

Images contain a mix of different type of information - from fine scale stochastic textures to large scale geometric structures. Image regularization can be viewed as approximating the observed original image with a simpler image,

where simpler is defined by the regularization (prior) term and the regularization parameter λ. Here an image is considered to be simpler if it is smoother (or piece-wise smoother). Regularization can also be viewed as decomposing the observed image into a regularized (smooth) component and a small scale texture/noise component (called the residual, because it is the difference between the regularized solution and the observed image). By increasing the regularization parameter λ smoother and smoother approximations are generated. The rate in which image details are suppressed as a function of the regularization parameter depends on the image content and regularization method. The image residual contains the details that are suppressed during the regularization and the norm of the residual is a measurement of the amount of details that are suppressed. The norm of the residual as a function of the regularization parameter gives important information about the image content. For images containing small scale structure a lot of details are suppressed even for small λ and the norm of the residual will be large for small λ. For images containing solely large scale geometric structures few details will be suppressed for small λ and the norm of the residual will be small. The rate in which details are suppressed can be viewed as the derivative of the norm of the residual with respect to the regularization parameter, and reveals the amount of details that are suppressed if the regularization parameter increases.

First order Tikhonov regularization, Gaussian linear scale space (which is equivalent to infinite order Tikhonov regularization [143]) and Total Variation image decomposition are studied. The squared L^2-norm of the regularized solution and the residual are studied as functions of the regularization parameter. Of special interest is the convexity/concavity of those norms viewed as functions, because it relates to the possibility that the rate in which details are suppressed can increase/decrease. In section 5.2, first order Tikhonov regularization is revisited and it is shown that the norm of the regularized solution is a convex function, while the norm of the residual is not a concave function. In section 5.3, linear Gaussian Scale Space is revisited, and it is shown that the norm of the regularized solution is convex as a function of the Gaussian variance, or equivalently diffusion time, but may fail to be convex when the parameter is the Gaussian standard deviation. The squared norm of the residual is in general not a concave function of its parameter. In section 5.4, Total Variation (TV) image decomposition is revisited. In section 5.5 experimental results are presented, the norm of the Sinc function, synthetic image containing image structures at different scales and natural images are studied.

These studies tend to show that the square residual norm contains scale information, particularly at values where local convexity/concavity behavior

changes.

5.1.1 Related work

Characterization of images by analyzing the behavior of the norm of the regularized solution and the residual as functions of the regularization parameter has not received much research attention. Sporring and Weickert [181, 182] view images as distributions of light quanta and use information theory to study the structure of images in scale space. The entropy of an image as a function of the scale (in scale-space) is analyzed and shown to be an increasing function of the scale. The result holds both for linear Gaussian scale space and non-linear scale-space. Furthermore the derivative of the entropy with respect to the scale is shown, empirically, to be a good texture descriptor. The derivative of the scale-space entropy function with respect to the scale is a global measure of how much the entropy of an image changes at different scale. Where Sporring and Weickert studies monotone functions of images across scale, we study norms of the scale space image and residual.

Buades et.al [27] introduced the concept of Method Noise in denoising. The Method Noise is the image details that are removed in the denoising - i.e. the residual image - and the content is used for comparing denoising methods. The residual image has often been used for determine the optimal regularization parameter. (See Thompson et.al [189] for a classical study.) Selection of the optimal stopping time for diffusion filter was studied by Mrazek and Navara [139], which also relate to the Lyapunov functionals studied by Weickert [200].

5.1.2 Convexity, Fourier Transforms, Power Spectra

Recall that a function $f(x)$ defined on a convex set C is convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for all $0 \leq \lambda \leq 1$ and for all $x, y \in C$. If $f(x)$ is convex on a convex set C then $-f(x)$ is said to be *concave* on C. When $f(x)$ is twice-differentiable, a necessary and sufficient condition for convexity is

$$\forall x \in C, \quad f''(x) \geq 0 \tag{5.1}$$

(in the multidimensional case a the Hessian matrix is positive semi-definite). Two elementary facts will be used in the sequel: 1) let $h(\lambda)$ be a function of the form

$$h(\lambda) = \int d(\lambda, x)s(x)dx \tag{5.2}$$

where $d(\lambda, x)$ is convex in λ and $s(x) \geq 0$ then $h(\lambda)$ is convex. 2) Assume that $f(x) = h(g(x))$ where $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$. Then

- if h is convex and non-decreasing and g is convex, then f is convex,
- if h is convex and non-increasing and g is concave, then f is concave.

The Fourier transform of a function f is denoted with $\hat{f}$. Parseval's theorem asserts that this is an isometry of L^2: $\|f\|_{L^2} = \|\hat{f}\|_{L^2}$ where

$$\|f(x,y)\|_2^2 = \iint |f(x,y)|^2 dxdy. \tag{5.3}$$

The frequency domain variables are denoted $(\omega_x, \omega_y) =: \omega$. The power spectrum function of a function f is the function $\omega \mapsto |\hat{f}(\omega)|^2$. f is said to follow a (α-)power law if $|\hat{f}(\omega)| \sim C/|\omega|^\alpha$, where C and α are some constants. It is well known that the power spectra computed over a large ensemble of natural image approximate a power law in spatial frequencies with α around 1.7 or at least in $(0, 2)$ [166, 58].

We use often implicitly the following classical result from Calculus. Let $B := B(0,1)$ the unit ball of $\mathbb{R}^n$ and B^c its complement. Let g a positive function defined on $\mathbb{R}^n$. Assume that $g \sim \|x\|^{-\alpha}$ in B (resp B^c). Then $\int_B g\, dx < \infty$ if and only if $\alpha < n$ (resp. $\int_{B^c} g\, dx < \infty$ if and only if $\alpha > n$).

Finally, to conclude this paragraph, given a regularization, the functions $s(\lambda)$ and $r(\lambda)$ will denote the squared L^2-norm of respectively the the regularized solution and of the residual as a function of the regularization parameter λ.

5.2 Tikhonov Regularization

The first order Tikhonov regularization is defined as the minimizer of the energy functional

$$E_\lambda[f] = \iint (f-g)^2 + \lambda |\nabla f|^2 dxdy \tag{5.4}$$

where g is the observed data and λ is the regularization parameter. The energy functional is composed of two terms: the data fidelity term $\|f-g\|_2^2$ and the regularization term $\|\nabla f\|_2^2$. Note that Wiener filter can be regarded as a Tikhonov regularization method applied to the Fourier domain. Thanks to Parseval's theorem all calculation can be performed in the Fourier domain where this energy becomes

$$\hat{E}_\lambda[f] = \iint (\hat{f} - \hat{g})^2 - \lambda(\omega_x^2 \hat{f}^2 + \omega_y^2 \hat{f}^2) d\omega_x d\omega_y. \tag{5.5}$$

Using the Calculus of Variations, a necessary condition for a function f to minimize the functional (5.4) is given by its Euler-Lagrange equation: $(f-g)-\lambda\Delta f=0$. In the Fourier domain, it becomes

$$\hat{f}-\hat{g}+\lambda(\omega_x^2\hat{f}+\omega_y^2\hat{f})=0 \quad \text{i.e } \hat{f}=\frac{\hat{g}}{1+\lambda|\omega|^2} \tag{5.6}$$

that is, the original signal multiplied with the filter function $F(\lambda,\omega)=\frac{1}{1+\lambda|\omega|^2}$ which is a non-increasing convex function w.r.t λ (for $\lambda\geq 0$). Set $d(\lambda,\omega)=F(\lambda,\omega)^2$. It is important to remark that defining the regularization in frequency domain by $\lambda\to F(\lambda,\omega)\hat{g}(\omega)$ extends Tikhonov regularization beyond the case where $g\in W^{1,2}(\mathbb{R}^2)$, the Sobolev space of L^2 functions with L^2 weak derivatives, which is the natural space for Tikhonov regularization as defined by minimization of (5.4). Indeed, the corresponding function $s(\lambda)$ is given by

$$s(\lambda)=\|F(\lambda,\omega)\hat{g}\|_2^2=\iint d(\lambda,\omega)|\hat{g}|^2 d\omega. \tag{5.7}$$

This is the integral of the squared filter function times the power spectrum of the original signal g, and we have the following result:

Proposition 1 *The squared L^2-norm $s(\lambda)$ of the minimizer of the Tikhonov regularization functional as a function of the regularization parameter λ is, for non-trivial images, a monotonically decreasing convex function (for $\lambda\in(0,\infty)$), when it exists.*

If g follows an α-power law, then from the Calculus fact recalled in the previous section, $g\notin L^2(\mathbb{R}^n)$, however $s(\lambda)$, $s'(\lambda)$ and $s''(\lambda)$ exist and are finite for $\lambda>0$ if and only if $\alpha\in(0,2)$ (which is the case for natural images). Both s' and s'' diverge for $\lambda\to 0^+$.

The square of a non-increasing convex function is a convex function, and from Section 5.1.2 we have the first part of the proposition. Now

$$d_\lambda(\lambda,\omega)=-\frac{2|\omega|^2}{(1+\lambda|\omega|^2)^3},\quad d_{\lambda\lambda}(\lambda,\omega)=6\frac{2|\omega|^4}{(1+\lambda|\omega|^2)^4}.$$

$s'(\lambda)=\iint d_\lambda(\lambda,\omega)|g|^2\,d\omega$ and $s''(\lambda)=\iint d_{\lambda\lambda}(\lambda,\omega)|g|^2\,d\omega$ and the rest of the proposition follows by elementary analysis. □

Set $R(\lambda,\omega)=1-F(\lambda,\omega)$ and $e(\lambda,\omega)=R(\lambda,\omega)^2$. The Fourier image residual is $R(\lambda)\hat{g}$ and its squared norm is

$$r(\lambda)=\|R(\lambda,\omega)\hat{g}\|^2=\iint e(\lambda,\omega)|\hat{g}|^2\,d\omega$$

An elementary calculation gives $e_\lambda(\lambda,\omega) = 2\lambda|\omega|^2/(1+\lambda|\omega|^2)^3$ and this function, is for λ fixed, bounded in ω while it satisfies

$$\forall\omega, \quad \lim_{\lambda\to 0+} e_\lambda(\lambda,\omega) \to 0, \lim_{\lambda\to\infty} e_\lambda(\lambda,\omega) \to 0$$

The same holds for $r'(\lambda)$ when it is finite and therefore by the mean value theorem, as it is positive, it must have a maximum and $r''(\lambda)$ must change sign and we can state the following:

Proposition 2 *Assume first that $g \in W^{1,2}(\mathbb{R}^2)$ is non trivial. Then, although $s(\lambda)$ is convex and decreasing, the squared norm residual $r(\lambda)$ of Tikhonov regularization, while increasing from 0 to $\|g\|_2^2$, is neither concave nor convex.*

Note that when g is a $\alpha-$power law with $\alpha \in (0,2)$, $g \notin L^2(\mathbb{R})$ while its regularization g_λ is when $\lambda > 0$, thus $g - g_\lambda \notin L^2(\mathbb{R}^2)$ and $r(\lambda) = \|g - g_\lambda\|_2^2 = +\infty$.

5.3 Linear Scale-Space and Regularization

Linear scale-space theory [111, 202, 98] deals with simplified coarse scale representation of an image g, generated by solving the diffusion (heat) equation with initial value g:

$$\frac{\partial f}{\partial t} = \triangle f, \quad f(-,0) = g(-) \tag{5.8}$$

where $\triangle = \partial_{xx} + \partial_{yy}$ is the Laplacian. Equivalently, this coarse scale representation can be obtained by convolution with a Gaussian kernel:

$$f_\sigma = g * G_\sigma, \quad G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5.9}$$

and the link between the two formulations is given by $f_\sigma = f(-, 2\sigma^2)$. A third formulation of Linear Scale-Space is obtained as "infinite order" Tikhonov regularization, the 1-dimensional case was introduced by Nielsen *et al.* in [143]. In dimension 2, one defines for $\lambda > 0$

$$E[f] = \iint (f-g)^2\, dxdy + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \iint \sum_{\ell=0}^{k} \left(\binom{k}{\ell} \frac{\partial^k f}{\partial x^\ell \partial y^{k-\ell}} \right)^2 dxdy \tag{5.10}$$

where $\binom{k}{\ell}$ is the (ℓ,k)-binomial coefficient. By a direct computation, its associated Euler-Lagrange equation is given by

$$f - g + \sum_{k=1}^{\infty} \frac{(-1)^k \lambda^k}{k!} \triangle^k f = 0$$

where $\triangle^k$ is the k-th iterated Laplacian

$$\triangle^k = \underbrace{\triangle \circ \cdots \circ \triangle}_{k \text{ times}} = \sum_{\ell=0}^{k} \binom{k}{\ell} \frac{\partial^{2k}}{\partial x^{2\ell} \partial y^{2(k-\ell)}}.$$

Via Fourier Transform, the Laplacian operator becomes the multiplication by $-|\omega|^2$ operator and as in 1st order Tikhonov regularization, the solution is given by filtering:

$$\hat{f} = \frac{\hat{g}}{1 + \sum_{k=1}^{\infty} \frac{\lambda^k |\omega|^{2k}}{k!}} = e^{-\lambda|\omega|^2} \hat{g}. \tag{5.11}$$

The solution of the filtering problem for a given $\lambda > 0$ is the same as solving (5.8) with $t = \lambda$. By setting $\lambda = 2\sigma^2$ and applying the convolution theorem to (5.9) one gets the above equation. Using the Fourier formulation, the squared norm of the solution at λ of (5.11) $s(\lambda)$ the squared-norm residual $r(\lambda)$ are given by

$$\begin{aligned} s(\lambda) &= \|e^{-\lambda|\omega|^2} \hat{g}\|_2^2 = \iint e^{-2\lambda|\omega|^2} |\hat{g}(\omega)|^2 \, d\omega, \\ r(\lambda) &= \|(1 - e^{-\lambda|\omega|^2}) \hat{g}\|_2^2 = \iint \left(1 - e^{-\lambda|\omega|^2}\right)^2 |\hat{g}(\omega)|^2 \, d\omega. \end{aligned}$$

If one defines $d(\lambda, \omega) = e^{-2\lambda|\omega|^2}$ and $e(\lambda, \omega) = (1 - e^{\lambda|\omega|^2})$, they have with respect to convexity/concavity, the same properties as their Tikhonov counterpart defined in the previous section and one can state the following, in term of heat equation / Gaussian variance

Proposition 3 *1. The squared L^2-norm $s(t)$ of the solution of heat equation as a function of the diffusion "time" t (or equivalently the convolution by the Gaussian kernel in function of the kernel* variance*) is, for non-trivial images, a monotonically decreasing convex function (for $t \in (0, \infty)$), when it exists.*

2. The squared norm residual $r(t)$ of the solution of the heat equation at time t, while increasing from 0 to $\|g\|_2^2$, is neither concave nor convex.

If, instead of using the diffusion time / variance as parameter, one uses the *standard deviation* σ of the Gaussian kernel, the resulting solution squared norm function $s(\sigma)$, although increasing, may fail to be convex as the function $\sigma \mapsto e^{-\sigma^2|\omega|^2}$ is not convex in σ, this is a half Gaussian bell. A simple example showing the convexity failure is provided by the band limited function b whose

Fourier transform is $\hat{b}(\omega) = 1$ if $|\omega| \leq 1$ and $\hat{b}(\omega) = 0$ otherwise. A direct calculation gives

$$s(\sigma) = \frac{\pi}{\sigma}\left(1 - e^{-\sigma^2}\right)$$

which is neither convex nor concave. In the other hand, for a function g following a $\alpha-$power law with $\alpha < 2$, $s(\sigma)$, this seems to be convex (for instance if $\alpha = 0$, $s(\sigma) = \pi/\sigma$, if $\alpha = 1$, $s(\sigma) = \pi^{3/2}/\sigma^2$).
If,again, the power spectrum of the image g is following a power law in spatial frequencies, while its regularized L^2- norm is finite, the residual norm is not as the initial datum is not square-integrable.

5.4 Total Variation image decomposition

Bounded Variation image modeling was introduced in the seminal work of Rudin *et al.* in [169], where the following variational image denoising problem is considered. Given an image g and $\lambda > 0$, find the minimizer of the following energy

$$E(f; g, \lambda) = \int (g - f)^2 dxdy + \lambda \iint |\nabla f| \, dxdy \tag{5.12}$$

The regularized image f_λ can be interpreted as a denoised version of g, but also as the "geometric" content of g while the residual $\nu_\lambda = g - f_\lambda$ contains the "noise/fine texture" component. Several methods have been proposed to solve the above equation, by solving a regularized form of the Euler-Lagrange equation of the functional

$$f - g - \lambda \nabla \cdot \frac{\nabla g}{|\nabla g|} = 0$$

where $\nabla\cdot$ denote the divergence operator, but also for instance the non linear projection method of Chambolle ([34]), which we have used in this work. λ is a regularization parameter that determines the level of details that ends up in the (noise/texture) component ν_λ. As λ increases ν_λ will contain details of larger and larger scale, that will not appear in f_λ.
Again it is interesting so see how the image content changes as λ increases. The component v_λ is the residual of the regularization and contains the details that are suppressed in the cartoon component f_λ and we set

$$r(\lambda; g) = \|v_\lambda\|_2^2 = \|g - f_\lambda\|_2^2 \tag{5.13}$$

i.e. the squared L^2-norm of the residual image as a function of the regularization parameter λ. Related to the norm of the residual is the norm of the

cartoon component as a function of λ

$$s(\lambda; u_0) = \|u_\lambda\|_2^2 \tag{5.14}$$

$s'(\lambda)$ encodes the rate in which details are suppressed in the cartoon component u_λ. Due to the high non linearity of the TV-regularization problem, there is no relatively simple expression for $s(\lambda)$, $r(\lambda)$ and their respective derivatives.

A norm study for the dual norm of the TV norm was done by Meyer in [132]. A more direct behavior for the 2-norm can be computed in a few cases. For instance Strong and Chan [186] showed that if g is the function $g(x) = 1$ if $x \in B(0,1)$ the unit disk, $g(x) = 0$ if $x \notin B(0,1)$, then its regularization has the form cg, where $c \in (0,1)$ is a constant, therefore attenuating the contrasts of the image.

In general situation, we cannot expect these type of simple results. We have instead decided to study the behavior of these functions experimentally on an image database.

5.5 Experiments

5.5.1 Sinc in Scale Space

Let $g(x) = \sin(x)/x$ be the Sinc function where $x \in [-\infty, \infty]$. The squared L_2 norm of the residual as a function of the regularization parameter is in the Tikhonov case

$$r(\lambda) = \int_{-1}^{1} (\frac{\lambda x^2}{1 + \lambda x^2})^2 dx \tag{5.15}$$

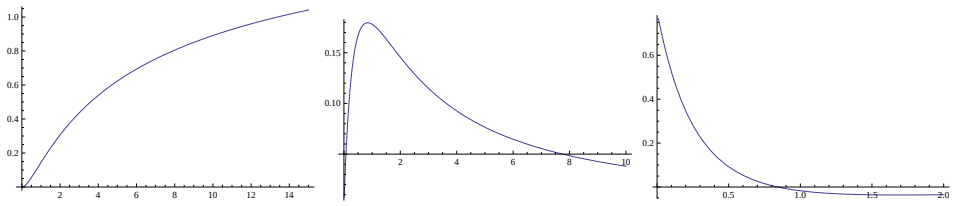
and in the scale space case

$$r(\sigma) = \int_{-1}^{1} (1 - e^{\frac{-\omega^2\sigma^2}{2}})^2 d\omega. \tag{5.16}$$

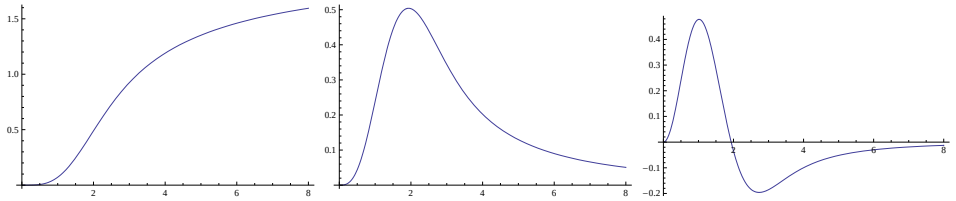
The result is presented in figure 5.1. The plots clearly indicate that the residual norm - in both cases -is not concave.

5.5.2 Black squares with added Gaussian noise

The first experiment is done on an artificially generated 100×100 image containing four 3×3 black squares, one 20×20 black square and added Gaussian white noise with $\sigma^2 = 12$. The white background has intensity 125



(a) Tikhonov regularization: Residual norm, first and second order derivative



(b) Scale Space: Residual norm, first and second order derivative

Figure 5.1: The residual norm as a function of the regularization parameter for $g(x) = \frac{sin(x)}{x}$. The plots clearly indicate that residual norm function are, in both case, increasing functions, but not concave.

and the black square 10, after the noise has been added the image is zero mean normalize.

In figure 5.2 the regularized and residual image are shown for increasing regularization using first order Tikhonov Regularization. As the small scale noise is suppressed, the large scale geometric structures are also smoothed out. The norm of the residual is an increasing function of the scale and it seems to be concave, and in fact it can be concave for the shown λ. However λ may be small at the inflection point.

In figure 5.3 the regularized and residual images are shown for increasing regularization using linear gaussian scale space. The results for the linear Gaussian scale-space is similar to the result using first order Tikhonov regularization.

In figure 5.4 the regularized and residual images are shown for increasing regularization using Total Variation image decomposition. The different structures are suppressed at using different λ while the large scale structures are well preserved. At $\lambda = 12$ the gaussian white noise is suppressed, and at $\lambda = 210$ is the small boxes remove and finally the large box is suppressed at $\lambda = 550$. The residual norm as a function of the regularization parameter is not a concave function of λ.

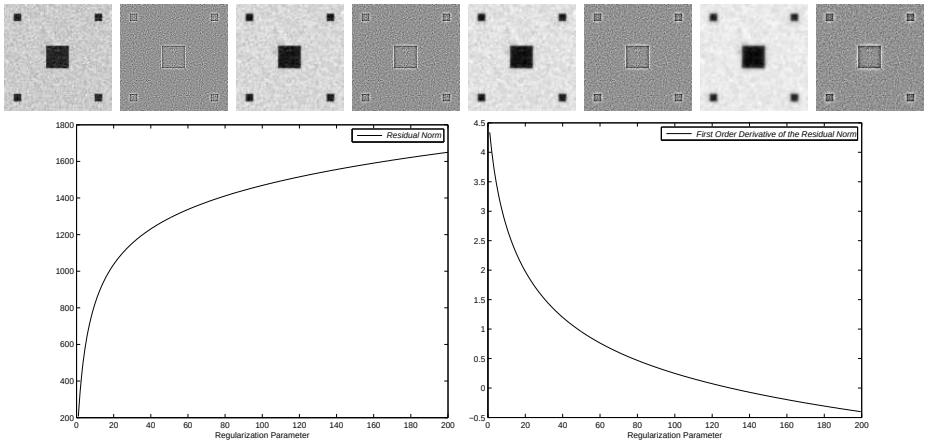


Figure 5.2: Result for the squares and noise image using first order Tikhonov regularization. On the first row the regularized and the residual images for $\lambda = 3, 10, 20$ and 50 are shown. The plots contain the L^2-norm of the residual as a function the scale λ, followed by the first order derivative in log-scale.

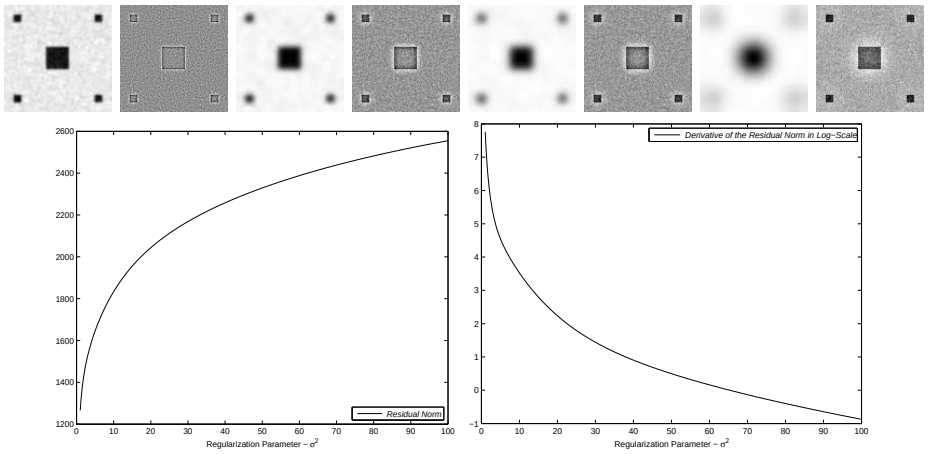


Figure 5.3: Result for the squares and noise image using linear scale space. On the first row the regularized and the residual images for $\sigma^2 = 1, 7, 13$ and 64 are shown. The plots contain the L^2-norm of the residual as a function the scale σ, followed by the first order derivative in log-scale.

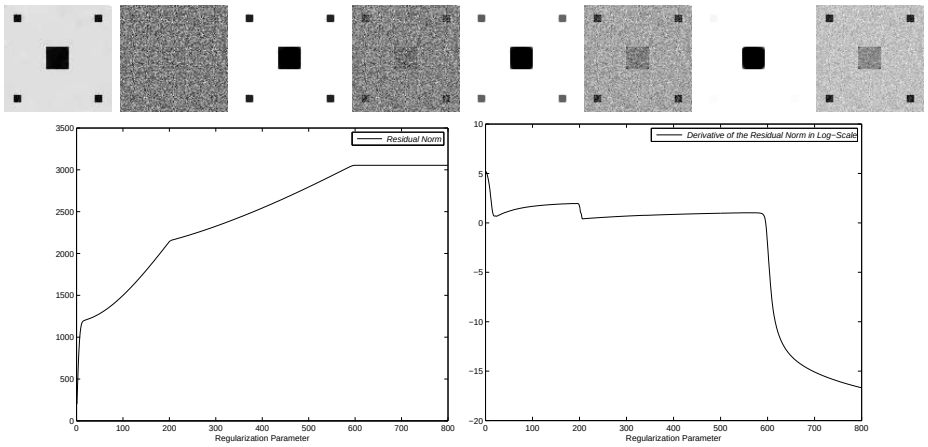


Figure 5.4: Result for the squares and noise image using TV-decomposition. On the row regularized and the residual images for $\lambda = 12, 38, 100$ and 200 are shown. The plots contain the L^2-norm of the residual as a function the scale λ, followed by the first order derivative in log-scale. The residual norm seems to be a monotonically increasing non-concave function. The residual norm has three points of 'high' curvature: one at $\lambda = 12$ - the noise is suppressed - and $\lambda = 210$ - the small squares are suppressed, and $\lambda = 580$ - the large square is suppressed.

5.5.3 DIKU Multi Scale Image Sequence Database I

The newly collected DIKU Multi-Scale image sequence database [85], contains sequences of the same scene captured using varying focal length. The sequences contain both man-made structures and nature, the distance to the main objects in the scenes also show a large variation (from a few meters to a few kilometers).

Each image has first been normalized by an affine intensity range change so that that the intensity range becomes $[0, 1]$, followed by subtracting the mean value (i.e. the mean intensity is 0 in each image).

The mean residual norm was computed on the normalized images in the database, using fixed scales $\sigma = 2^i$ where $i = 0, \cdots, 12$, using linear gaussian scale space. The result is a feature vector $\langle \bar{r}(0), \cdots, \bar{r}(12) \rangle$ containing

$$\bar{r}(i) = \frac{1}{N} \sum_{I \in F} r(i; I) \tag{5.17}$$

where F is the set of all N normalized images in the database.

The (signed) distance function $d(I_0)$ of a normalized image $I_0 \in F$ to the mean is defined as

$$d(I_0) = \sum_{i=0}^{12} r(i; I_0) - \bar{r}(i) \tag{5.18}$$

The (signed) distance to the mean has been computed for all images in the DIKU database. Images with large positive values have a larger than average residual and images with large negative values have a smaller than average residual.

The first row in figure 5.5 contains the 4 images with the largest positive distance to the mean, on the second row the 4 images with the largest negative distance to the mean. The image contents difference is striking and clearly indicate that the residual norm contains important contents information. The same experiment was performed using first order Tikhonov regularization with similar, but not identical, result.

Figure 5.5: The top row show images where $f(\sigma)$ is much larger than the average and bottom row show images where $f(\sigma)$ is much smaller than the average. The contents difference is striking! The images in the first row contain small scale details (texture), while the images in the bottom row contain large scale geometric structures.

5.6 Conclusions

For square-integrable images, the squared L^2-norms of the regularized images in first order Tikhonov regularization and linear Gaussian Scale Space are, in general decreasing convex functions of the regularizing parameter. This may fail for Linear Scale space when Gaussian standard deviation is used as a parameter. Their squared residual norm are however not concave functions. For the the Total Variation regularization too, it is shown empirically that the squared norm of the residual is not concave.

This confirms that the squared norm of the residual may be an indicator of image structure, both for 1st order Tikhonov regularization, Gaussian

Scale Space as well as Total variation regularization. The behavior of the latter will be studied further in future research.

ACKNOWLEDGEMENTS

This research was funded by the EU Marie Curie Research Training Network VISIONTRAIN MRTN-CT-2004- 005439 and the Danish Natural Science Research Council project Natural Image Sequence Analysis (NISA) 272-05-0256. The authors wants to thank Christoph Schnörr (Heidelberg University), Niels-Christian Overgaard (Lund University) and Vladlena Gorbunova (Copenhagen University) for charing their knowledge.

Chapter 6

Variational Segmentation and Contour Matching of Non-Rigid Moving Object

This chapter contain a slightly re-formatted version of

David Gustavsson, Ketut Fundana, Niels-Ch. Overgaard, Anders Heyden, and Mads Nielsen.
Variational Segmentation and Contour Matching of Non-Rigid Moving Object.
In Proceding of Workshop on Dynamical Vision WDV 2008, 2008.

Variational Segmentation and Contour Matching of Non-Rigid Moving Object

David Gustavsson[1,2], Ketut Fundana[3], Niels Chr. Overgaard[3], Anders Heyden[3], and Mads Nielsen[1]
[1]DIKU, University of Copenhagen
Universitetsparken 1,DK-2100 Copenhagen Ø, Denmark
{davidg,madsn}@diku.dk
[2]IT University of Copenhagen
Rued Langgaards Vej 7,DK-2300 Copenhagen S, Denmark
[3]Applied Mathematics Group,School of Technology and Society, Malmö University
Östra Varvsgatan 11A, SE-205 06 Malmö, Sweden
{ketut.fundana,nco,heyden}@ts.mah.se

abstract

In this paper we propose a method for variational segmentation and contour matching of nonrigid objects in image sequences which can deal with the occlusions. The method is based on a region-based active contour model of the Chan-Vese, augmented with a frame-to-frame interaction term which uses the segmentation result from the previous frame as a shape prior. This method has given good results despite the presence of minor occlusions, but can not handle significant occlusions. We have extended this approach by adding a registration step between two consecutive contours. This registration step is based on a novel variational formulation and gives also a mapping of the intensities from the interior of the previous contour to the next. With this information occlusions can be detected from deviations from predicted

intensities and the missing intensities in the occluded areas can then be reconstructed. The performance of the method is shown with experiments on synthetic and real image sequences.

6.1 Introduction

Segmentation is an important and difficult process in computer vision, with the purpose of dividing a given image into one or several meaningful regions or objects. This process is more difficult when the objects to be segmented are moving and nonrigid and even more when there are severe occlusions. The shape of nonrigid, moving objects may vary a lot along image sequences due to, for instance, deformations or occlusions, which puts additional constraints on the segmentation process. In particular we would like to distinguish *real* shape deformations of the object from *apparent* shape deformations due to occlusions.

There have been a number of methods proposed and applied to this problem. Active contours are powerful methods for image segmentation; either boundary-based such as geodesic active contours [33], or region-based such as Chan-Vese models [35], which are formulated as variational problems. Those variational formulations perform quite well and have often been applied based on level sets. Active contour based segmentation methods often fail due to noise, clutter and occlusion. In order to make the segmentation process robust against these effects, shape priors have been proposed to be incorporated into the segmentation process. In recent years, many researchers have successfully introduced shape priors into segmentation methods such as in [36, 44, 46, 43, 42, 165, 120].

We are interested in segmenting nonrigid moving objects in image sequences. When the objects are nonrigid, an appropriate segmentation method that can deal with shape deformations should be used. The application of active contour methods for segmentation in image sequences gives promising results as in [137, 153, 154]. These methods use variants of the classical Chan-Vese model as the basis for segmentation. In [137], for instance, it is proposed to simply use the result from one image as an initializer in the segmentation of the next.

Another major problem for segmentation methods for image sequences is the presence of occlusions. Minor occlusions can usually be handled by some kind of shape prior. However, major occlusions is still a big problem. In order to improve the robustness of the segmentation methods in the presence of occlusions, it is necessary to detect the occlusions. The occluded area can then either be excluded from segmentation process or reconstructed [185, 70, 114].

The main purpose of this paper is to propose and analyze a novel variational segmentation method for image sequences, that can both deal with shape deformations and at the same time is robust to noise, clutter and occlusions. The proposed method is based on minimizing an energy functional containing the standard Chan-Vese functional as one part and a term that penalizes the deviation from the previous shape as a second part. The second part of the functional is based on a transformed distance map to the previous contour, where different transformation groups, such as Euclidean, similarity or affine, can be used depending on the particular application. This variational framework is then augmented with a novel contour flow algorithm, giving a mapping of the intensities inside the contour of one image to the inside of the contour in the next image. Using this mapping, occlusions can be detected by simply thresholding the difference between the transformed intensities and the observed ones in the novel image.

This paper is organized as follows: in Sect. 6.2 we discuss the proposed segmentation of image sequences. The variational contour matching is described in Sect. 6.3 and how this can be used to detect and locate the occlusion is described in Sect. 6.4. Experimental results of the model are presented in Sect. 6.5 and we end the paper with some conclusions.

6.2 Segmentation of Image Sequences

In this section, we describe the region-based segmentation model of Chan-Vese [35] and a variational model for updating segmentation results from one frame to the next in an image sequence.

6.2.1 Region-Based Segmentation

The idea of the Chan-Vese model [35] is to find a contour Γ such that the image I is optimally approximated by a gray scale value μ_{int} on $\text{int}(\Gamma)$, the *inside* of Γ, and by another gray scale value μ_{ext} on $\text{ext}(\Gamma)$, the *outside* of Γ. The optimal contour Γ^* is defined as the solution of the variational problem,

$$E_{CV}(\Gamma^*) = \min_{\Gamma} E_{CV}(\Gamma), \tag{6.1}$$

where E_{CV} is the Chan-Vese functional,

$$E_{CV}(\Gamma) = \alpha|\Gamma| + \beta\left\{\frac{1}{2}\int_{\text{int}(\Gamma)}(I(\mathbf{x}) - \mu_{\text{int}})^2\,d\mathbf{x} + \frac{1}{2}\int_{\text{ext}(\Gamma)}(I(\mathbf{x}) - \mu_{\text{ext}})^2\,d\mathbf{x}\right\}. \tag{6.2}$$

Here $|\Gamma|$ is the arc length of the contour, $\alpha, \beta > 0$ are weight parameters, and

$$\mu_{\text{int}} = \mu_{\text{int}}(\Gamma) = \frac{1}{|\operatorname{int}(\Gamma)|} \int_{\operatorname{int}(\Gamma)} I(\mathbf{x})\, d\mathbf{x}, \tag{6.3}$$

$$\mu_{\text{ext}} = \mu_{\text{ext}}(\Gamma) = \frac{1}{|\operatorname{ext}(\Gamma)|} \int_{\operatorname{ext}(\Gamma)} I(\mathbf{x})\, d\mathbf{x}. \tag{6.4}$$

The gradient descent flow for the problem of minimizing a functional $E_{CV}(\Gamma)$ is the solution to initial value problem:

$$\frac{d}{dt}\Gamma(t) = -\nabla E_{CV}(\Gamma(t)), \qquad \Gamma(0) = \Gamma_0, \tag{6.5}$$

where Γ_0 is an initial contour. Here $\nabla E_{CV}(\Gamma)$ is the L^2-gradient of the energy functional $E_{CV}(\Gamma)$, cf. e.g. [179] for definitions of these notions. Then the L^2-gradient of E_{CV} is

$$\nabla E_{CV}(\Gamma) = \alpha\kappa + \beta\big[\frac{1}{2}(I - \mu_{\text{int}}(\Gamma))^2 - \frac{1}{2}(I - \mu_{\text{ext}}(\Gamma))^2\big], \tag{6.6}$$

where κ is the curvature.

In the level set framework [151], a curve evolution, $t \mapsto \Gamma(t)$, can be represented by a time dependent level set function $\phi : \mathbf{R}^2 \times \mathbf{R} \to \mathbf{R}$ as $\Gamma(t) = \{\mathbf{x} \in \mathbf{R}^2 \ ; \ \phi(\mathbf{x}, t) = 0\}$, $\phi(\mathbf{x}) < 0$ and $\phi(\mathbf{x}) > 0$ are the regions inside and the outside of Γ, respectively. The normal velocity of $t \mapsto \Gamma(t)$ is the scalar function $d\Gamma/dt$ defined by

$$\frac{d}{dt}\Gamma(t)(\mathbf{x}) := -\frac{\partial\phi(\mathbf{x}, t)/\partial t}{|\nabla\phi(\mathbf{x}, t)|} \qquad (\mathbf{x} \in \Gamma(t)). \tag{6.7}$$

Recall that the outward unit normal $\mathbf{n}$ and the curvature κ can be expressed in terms of ϕ as $\mathbf{n} = \nabla\phi/|\nabla\phi|$ and $\kappa = \nabla \cdot \big(\nabla\phi/|\nabla\phi|\big)$.

Combined with the definition of gradient descent evolutions (6.5) and the formula for the normal velocity (6.7) this gives the gradient descent procedure in the level set framework:

$$\frac{\partial\phi}{\partial t} = \Big(\alpha\kappa + \beta\big[\frac{1}{2}(I - \mu_{\text{int}}(\Gamma))^2 - \frac{1}{2}(I - \mu_{\text{ext}}(\Gamma))^2\big]\Big)|\nabla\phi|,$$

where $\phi(\mathbf{x}, 0) = \phi_0(\mathbf{x})$ represents the initial contour Γ_0.

6.2.2 The Interaction Term

The interaction $E_I(\Gamma_0, \Gamma)$ between a fixed contour Γ_0 and an active contour Γ may be regarded as a shape prior and be chosen in several different ways, such as the pseudo-distances, cf. [43], and the area of the symmetric difference of the sets $\text{int}(\Gamma)$ and $\text{int}(\Gamma_0)$, cf. [36].

Let $\phi_0 : D \to \mathbf{R}$ denotes the *signed distance function* associated with the contour Γ_0 and $\mathbf{a} \in \mathbf{R}^2$ is a group of translations. We want to determine the optimal translation vector $\mathbf{a} = \mathbf{a}(\Gamma)$, then the interaction $E_I = E_I(\Gamma_0, \Gamma)$ is defined by the formula,

$$E_I(\Gamma_0, \Gamma) = \min_{\mathbf{a}} \int_{\text{int}(\Gamma)} \phi_0(\mathbf{x} - \mathbf{a})\, d\mathbf{x}. \tag{6.8}$$

Minimizing over groups of transformations is the standard devise to obtain pose-invariant interactions, see [36] and [43].

Since this is an optimization problem $\mathbf{a}(\Gamma)$ can be found using the gradient descent procedure. The optimal translation $\mathbf{a}(\Gamma)$ can then be obtained as the limit, as time t tends to infinity, of the solution to initial value problem

$$\dot{\mathbf{a}}(t) = \int_{\text{int}(\Gamma)} \nabla\phi_0(\mathbf{x} - \mathbf{a}(t))\, d\mathbf{x}, \qquad \mathbf{a}(0) = 0. \tag{6.9}$$

Similar gradient descent schemes can be devised for rotations and scalings (in the case of similarity transforms), cf. [36].

6.2.3 Using the Interaction Term in Segmentation of Image Sequences

Let $I_j : D \to \mathbf{R}$, $j = 1, \ldots, N$, be a succession of N frames from a given image sequence. Also, for some integer k, $1 \leq k \leq N$, suppose that all the frames $I_1, I_2, \ldots, I_{k-1}$ have already been segmented, such that the corresponding contours $\Gamma_1, \Gamma_2, \ldots, \Gamma_{k-1}$ are available. In order to take advantage of the prior knowledge obtained from earlier frames in the segmentation of I_k, we propose the following method: If $k = 1$, i.e. if no previous frames have actually been segmented, then we just use the standard Chan-Vese model, as presented in Sect. 6.2.1. If $k > 1$, then the segmentation of I_k is given by the contour Γ_k which minimizes an *augmented* Chan-Vese functional of the form,

$$E^A_{CV}(\Gamma_{k-1}, \Gamma_k) := E_{CV}(\Gamma_k) + \gamma E_I(\Gamma_{k-1}, \Gamma_k), \tag{6.10}$$

where E_{CV} is the Chan-Vese functional, $E_I = E_I(\Gamma_{k-1}, \Gamma_k)$ is an *interaction term*, which penalizes deviations of the current active contour Γ_k from the

previous one, Γ_{k-1}, and $\gamma > 0$ is a coupling constant which determines the strength of the interaction.

The augmented Chan-Vese functional (6.10) is minimized using standard gradient descent (6.5) described in Sect. 6.2.1 with ∇E equal to

$$\nabla E^A_{CV}(\Gamma_{k-1}, \Gamma_k) := \nabla E_{CV}(\Gamma_k) + \gamma \nabla E_I(\Gamma_{k-1}; \Gamma_k), \tag{6.11}$$

and the initial contour $\Gamma(0) = \Gamma_{k-1}$. Here ∇E_{CV} is the L^2-gradient (6.6) of the Chan-Vese functional, and ∇E_I the L^2-gradient of the interaction term, which is given by the formula,

$$\nabla E_I(\Gamma_{k-1}, \Gamma_k; \mathbf{x}) = \phi_{k-1}(\mathbf{x} - \mathbf{a}(\Gamma_k)), \qquad (\text{for } \mathbf{x} \in \Gamma_k). \tag{6.12}$$

Here ϕ_{k-1} is the signed distance function for Γ_{k-1}.

We use the Chan-Vese model to segment a selected object with approximately uniform intensity and apply the proposed method frame-by-frame. First we compute the optimal translation vector (6.9) based on the previous contour, we then use this vector to translate the previous contour until it is aligned to the optimal position (6.12). Then the minimum of the functional (6.10) is obtained by the gradient descent procedure (6.11) implemented in the level set framework outline in Sect. 6.2. This procedure is iterated until it converges.

6.3 A Contour Matching Problem

In this section we are going to present a variational solution to the following contour matching problem: Suppose we have two simple closed curves Γ_1 and Γ_2 contained in the image domain Ω. Find the "most economical" mapping $\Phi = \Phi(x) : \Omega \to \mathbf{R}^2$ such that Φ maps Γ_1 onto Γ_2, i.e. $\phi(\Gamma_1) = \Gamma_2$. The latter condition is to be understood in the sense that if $\alpha = \alpha(s) : [0, 1] \to \Omega$ is a positively oriented parametrization of S_1, then $\beta(s) = \Phi(\alpha(s)) : [0, 1] \to \Omega$ is a positively oriented parametrization of Γ_2 (allowing some parts of Γ_2 to be covered multiple times).

To present our variational solution of this problem, let $\mathcal{M}$ denote the set of twice differential mappings Φ which maps Γ_1 to Γ_2 in the above sense. Loosely speaking

$$\mathcal{M} = \{\Phi \in C^2(\Omega; \mathbf{R}^2) \,|\, \Phi(\Gamma_1) = \Gamma_2\}.$$

Moreover, given a mapping $\Phi : \Omega \to \mathbf{R}^2$, not necessarily a member of $\mathcal{M}$, then we express Φ in the form $\Phi(x) = x + U(x)$, where the vector valued function $U = U(x) : \Omega \to \mathbf{R}^2$ is called the *displacement field associated with*

Φ, or simply the displacement field. It is sometimes necessary to write out the components of the displacement field; $U(x) = (u_1(x), u_2(x))^T$.

We now define the "most economical" map to be the member Φ^* of $\mathcal{M}$ which minimizes the following energy functional:

$$E[\Phi] = \frac{1}{2} \int_\Omega \|DU(x)\|_F^2 \, d\mathbf{x}, \tag{6.13}$$

where $\|DU(x)\|_F$ denotes the Frobenius norm of $DU(x) = [\nabla u_1(x), \nabla u_2(x)]^T$, which for an arbitrary matrix $A \in \mathbf{R}^{2\times 2}$ is defined by $\|A\|_F^2 = \mathrm{tr}(A^T A)$. That is, the optimal matching is given by

$$\Phi^* = \arg\min_{\Phi \in \mathcal{M}} E[\Phi]. \tag{6.14}$$

Using that $E[\Phi]$ can be written in the form

$$E[\Phi] = \frac{1}{2} \int_\Omega |\nabla u_1(x)|^2 + |\nabla u_2(x)|^2 \, d\mathbf{x}, \tag{6.15}$$

it is easy to see that the Gâteaux derivative of $E[\Phi]$ is given by

$$\begin{aligned} dE[\Phi; V] &= \int_\Omega \nabla u_1(x) \cdot \nabla v_1(x) + \nabla u_2(x) \cdot \nabla v_2(x) \, d\mathbf{x} \\ &= \int_\Omega \mathrm{tr}(DU(x)^T DV(x)) \, d\mathbf{x}, \end{aligned}$$

for any displacement field $V(x) = (v_1(x), v_2(x))^T$. After integration by parts we find that the necessary condition for $\Phi^*(x) = x + U^*(x)$ to be a solution of the minimization problem (6.14) takes the form

$$0 = -\int_\Omega \Delta U^*(x) \cdot V(x) \, d\mathbf{x}, \tag{6.16}$$

for any *admissible* displacement field variation $V = V(x)$. Here $\Delta U^*(x) = (\Delta u_1(x), \Delta u_2(x))^T$ is the Laplacian of the vector valued function $U^* = U^*(x)$. Since every admissible mapping Φ must map the initial contour Γ_1 onto the target contour Γ_2, it can be shown that any displacement field variation V must satisfy

$$V(x) \cdot \mathbf{n}_{S_2}(x + U^*(X)) = 0 \quad \text{for all } x \in \Gamma_1. \tag{6.17}$$

Notice that this condition only has to be satisfied precisely on the curve Γ_1, and that $V = V(x)$ is allowed to vary freely away from the initial contour. The interpretation of the above condition is that the displacement field

variation at $x \in \Gamma_1$ must be tangent to the target contour Γ_2 at the point $y = \Phi(x)$. In view of this interpretation of (6.17) it is not difficult to see that necessary condition (6.16) implies that the solution Φ^* of the minimization problem (6.14) must satisfy the following Euler-Lagrange equation:

$$0 = \begin{cases} \Delta U^* - (\Delta U^* \cdot \mathbf{n}^*_{\Gamma_2})\, \mathbf{n}^*_{\Gamma_2}, & \text{on } \Gamma_1, \\ \Delta U^*, & \text{otherwise,} \end{cases} \tag{6.18}$$

where $\mathbf{n}^*_{\Gamma_2}(x) = \mathbf{n}_{\Gamma_2}(x + U^*(x))$, $x \in \Gamma_1$, is the pullback of the normal field of the target contour Γ_2 to the initial contour Γ_1. The standard way of solving (6.18) is to use the gradient descent method: Let $U = U(t, x)$ be the time-dependent displacement field which solves the evolution PDE

$$\frac{\partial U}{\partial t} = \begin{cases} \Delta U - (\Delta U \cdot \mathbf{n}^*_{\Gamma_2})\, \mathbf{n}^*_{\Gamma_2}, & \text{on } \Gamma_1, \\ \Delta U, & \text{otherwise,} \end{cases} \tag{6.19}$$

where the initial displacement $U(0, x) = U_0(x) \in \mathcal{M}$ specified by the user, and $U = 0$ on $\partial\Omega$, the boundary of Ω (Dirichlet boundary condition). Then $U^*(x) = \lim_{t\to\infty} U(t, x)$ is a solution of the Euler-Lagrange equation (6.18).

Notice that the PDE (6.19) coincides with the so-called *geometry-constrained diffusion* introduced in [5]. Thus we have incidentally found a variational formulation of the non-rigid registration problem considered there.

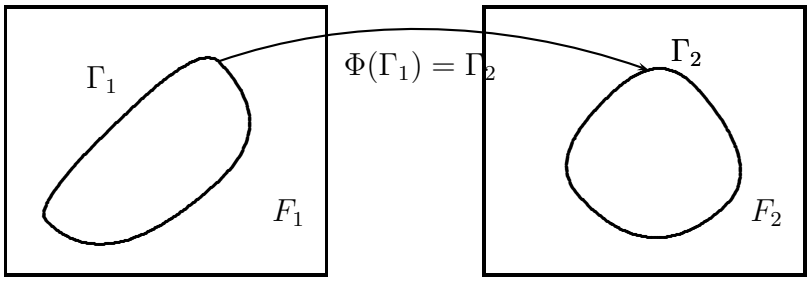


Figure 6.1: Given two closed curves Γ_1 and Γ_2 contained in two images F_1 and F_2, Φ maps F_1 onto F_2 such that Γ_1 is mapped onto Γ_2 (i.e. $\Phi(\Gamma_1) = \Gamma_2$).

6.4 Detect and Locate the Occlusion

The mapping $\Phi = \Phi(x) : \Omega \to \mathbf{R}^2$ such that Φ maps Γ_1 onto Γ_2 is an estimation of the displacement (motion and deformation) of the boundary of

an object between two frames. By finding the displacement of the contour, a consistent displacement of the intensities inside the closed curve Γ_1 can also be found. Φ maps Γ_1 onto Γ_2 and pixels inside Γ_1 are mapped inside Γ_2. This displacement field which only depends on displacement - or registration - of the contour (and not on the image intensities) can then be used to map the intensities inside Γ_1 into Γ_2. After mapping, the intensities inside Γ_1 and Γ_2 can be compared and then be classified as the same or different value. Since we can still find the contour in the occluded area, therefore we can also compute the displacement field even in the occluded area.

After the occlusion has been detected, the segmentation can be further improved by again employing the previously described Chan-Vese-method augmented with an interaction term. However, in this second stage, the integration is only performed over the area of the image where no occlusion has been detected. This procedure treats the occluded area in the same way as a part of the image with missing data as in [12], which is reasonable .

6.5 Experiments

6.5.1 Segmentation

In this section we present the results obtained from experiment using synthetic image sequence. We use the Chan-Vese model to segment a selected object with approximately uniform intensity and apply the proposed method frame-by-frame. The minimization of the functional is obtained by the gradient descent procedure (6.11) implemented in the level set framework. See also [151].

The classical Chan-Vese method will have problems segmenting an object if occlusions appear in the image which cover the whole or parts of the selected object. In Fig. 6.2 and Fig. 6.5, we show the segmentation results for a non-rigid object in a synthetic image sequence and for a walking human in a real image sequence (available at http://homepages.inf.ed.ac.uk/rbf/CAVIAR/), respectively, where occlusions occur. The classical Chan-Vese method fails to segment the selected object when it reaches the occlusion (Left column). Using the proposed method, which uses the frame-to-frame interaction term, we obtain much better results (Right column).

In both experiments the coupling constant γ is varied to see the influence of the interaction term on the segmentation results. The contour is only slightly affected by the prior if γ is small. On the other hand, if γ is too large, the contour will be close to a similarity transformed version of the

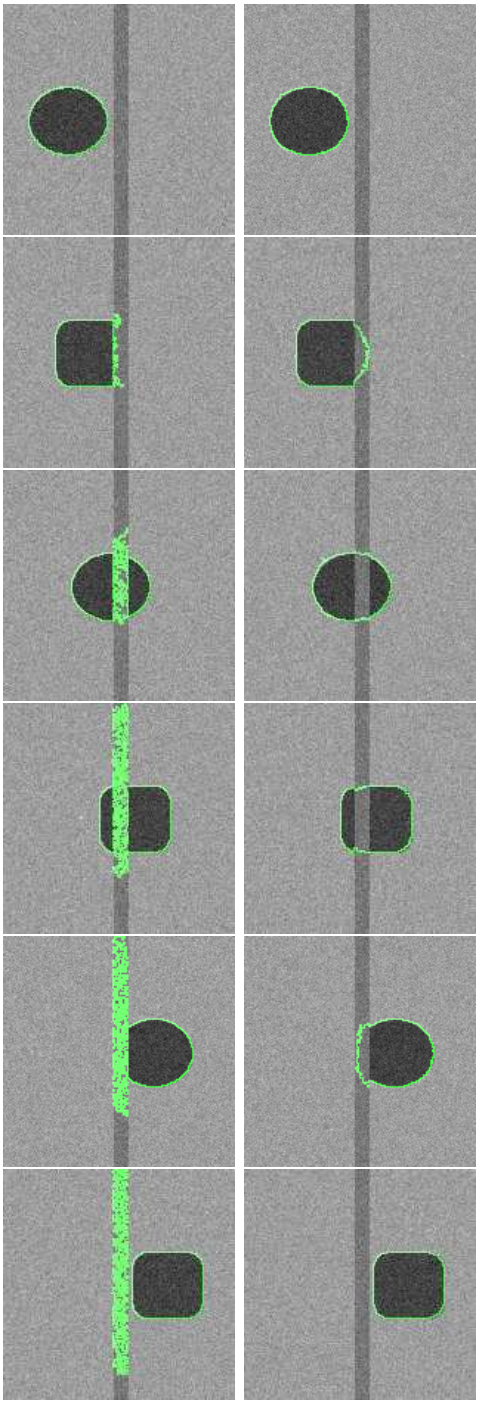


Figure 6.2: Segmentation of a non-rigid object in a synthetic image sequences with additive Gaussian noise (Frame 1-7). Without the interaction term, noise in the occlusion is captured (Left column). This is avoided when the interaction term is included (Right column).

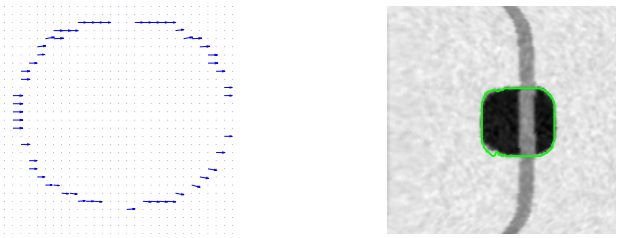


Figure 6.3: Left: Deformation field. Right: Frame 4 after deformation according to the displacement field onto Frame 5.

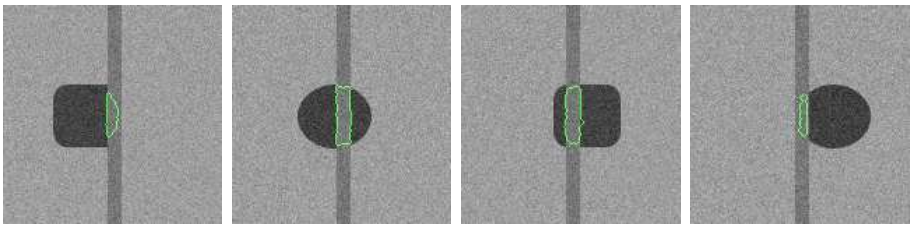


Figure 6.4: The occluded regions of the Frame 3-6 of Fig. 6.2 can be detected and located

prior.

6.5.2 Contour Matching and occlusion detection

As described in Sect. 6.3 and Sect. 6.4, occlusion can be detected and located by deforming the current frame according to the displacement and compare the deformed frame with the next frame (inside the contour Γ_2). First we compute the displacement field based on the segmentation results of two frames. In Fig. 6.3, we show the displacement field of Frame 4 and 5. With this displacement field, we can do full deformation of the Frame 4 onto Frame 5 (Fig. 6.3 right) and then compare the intensities between Frame 5 and deformed Frame 4. By comparing, we can then classify the intensities as having the same or different value by thresholding. The results for the artificial sequence are presented in Fig. 6.4 and for the walking person sequence in Fig. 6.6.

Figure 6.5: Segmentation of a person covered by an occlusion in the human walking sequence. Left column: without interaction term, and Right column: with interaction term

Figure 6.6: The occluded regions of the Frame 3 and 4 of Fig. 6.5 are detected and located by predicting the intensities inside the contour of the walking person.

6.6 Conclusions

We have presented a new method for segmentation and contour matching of image sequences containing nonrigid, moving objects, that also can handle occlusions. The proposed segmentation method is formulated as variational problem, with one part of the functional corresponding to the Chan-Vese model and another part corresponding to the pose-invariant interaction with a shape prior based on the previous contour. The optimal transformation as well as the shape deformation are determined by minimization of an energy functional using a gradient descent scheme. This segmentation method is augmented with a contour flow estimation algorithm based on a novel variational formulation. The estimated contour flow makes it possible to extract occluded areas and then further refine the segmentation. Preliminary results are shown and its performance looks promising both in terms of segmentation and occlusion detection.

Acknowledgements.

This research is funded by the VISIONTRAIN RTN-CT-2004-005439 Marie Curie Action within the EC's FP6.

Bibliography

[1] L. Alvarez, Y. Gousseau, and J.-M. Morel. Scales in natural images and a consequence on their bounded variation norm. In *Proceedings of Scale Space Methods in Computer Vision SS*, pages 247–258, 1999.

[2] L. Alvarez, Y. Gousseau, and J.-M. Morel. The size of objects in natural and artificial images. *Advances in Imaging and Electron Physics*, (111):167–242, 1999.

[3] L. Alvarez, Y. Gousseau, and J.-M. Morel. The size of objects in natural images. Technical Report CMLA9921, CMLA, 1999.

[4] P. R. Andresen and M. Nielsen. Non-rigid registration by geometry-constrained diffusion. *Medical Image Analysis*, 5(2):81–88, 2001.

[5] Per R. Andresen and Mads Nielsen. Non-rigid registration by geometry-constrained diffusion. In *MICCAI '99: Proceedings of the Second International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543, London, UK, 1999. Springer-Verlag.

[6] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations (second edition)*, volume 147 of *Applied Mathematical Sciences*. Springer-Verlag, 2006.

[7] Jean-Franois Aujol, Gilles Aubert, Laure Blanc-Féraud, and Antonin Chambolle. Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision*, 22(1):71–88, January 2005.

[8] Jean-Franois Aujol, Guy Gilboa, Tony Chan, and Stanley Osher. Structure-texture image decompositionmodeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, April 2006.

[9] R. Baddeley. The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science*, 21:351–372, 1997.

[10] C. Ballester, B. Bertalmio, V. Caselles, L. Garrido, A. Marques, and F. Ranchin. An inpainting- based deinterlacing method. *IEEE Transactions On Image Processing*, 16(10):2476–2491, October 2007.

[11] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions On Image Processing*, 10(8):1200–1211, August 2001.

[12] C. Ballester, V. Caselles, and J. Verdera. A variational model for disocclusion. In *Proceeding ICIP (3)*, pages 677–680, 2003.

[13] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, February 1994.

[14] R. Basri, L. Costa, D. Geiger, and D. Jacobs. Determining the similarity of deformable shapes. In *Proceedings of ICCV Workshop on Physics-Based Modeling in Computer Vision*, pages 135–143, 1995.

[15] M. Bertalmio, L.A. Vese, G. Sapiro, and S.J. Osher. Image filling-in in a decomposition space. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages I: 853–856, 2003.

[16] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[17] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous strurcture and texture image inpainting. *IEEE Transactions On Image Processing*, 12(8):882–889, August 2003.

[18] M Bertero, Tomaso Poggio, and Vincent Torre. Ill-posed problems in early vision. Technical Report A.I. Memo 924, MIT, May 1987.

[19] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48:259–302, 1986.

[20] Josef Bigun. *Vision with Direction - A Systematic Introduction to Image Processing and Computer Vision*. Springer-Verlag, 2006.

[21] J. S. De Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In *Computer Graphics*, pages 361–368. ACM SIGGRAPH, 1997.

[22] Stephen Boyd and Lieven Vandenberge. *Convex Optimization*. Cambridge University Press, 2004.

[23] Pierre Brémaud. *Markov Chains - Gibbs Field, MonteCarlo Simulation, and Queues*. Number 31 in TAM. Springer-Verlag, 1999.

[24] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376, 1992.

[25] Thomas Brox, Andres Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In Tomas Pajdla and Jiri Matas, editors, *Proc. 8th European Conference on Computer Vision (ECCV 04)*, volume 4, pages 25–36. Springer-Verlag, May 2004.

[26] Antoni Buades, A. Chien, Jean-Michel Morel, and Stanley Osher. Topology preserving linear filtering applied to medical imaging. *SIAM Journal on Imaging Sciences*, 1(1):26–50, 2008.

[27] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 60–65, Washington, DC, USA, 2005. IEEE Computer Society.

[28] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Neighborhood filters and pde’s. *Numer. Math.*, 105(1):1–34, 2006.

[29] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Nonlocal image and movie denoising. *International Journal of Computer Vision*, 76(2):123–139, February 2008.

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

[31] Aurelia Bugeau and Marcelo Bertalmio. Combining texture synthesis and diffusion for image inpainting. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009.

[32] P.J. Burt. Fast filter transforms for image processing. *Computer Vision Graphics and Image Processing*, 16(1):20–51, May 1981.

[33] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[34] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.

[35] T. Chan and L. Vese. Active contour without edges. *IEEE Transactions On Image Processing*, 10(2):266–277, 2001.

[36] T. Chan and W. Zhu. Level set based prior segmentation. Technical Report 03-66, Department of Mathematics, UCLA, 2003.

[37] Tony F. Chan and Sung Ha Kang. Error analysis for image inpainting. *Journal of Mathematical Imaging and Vision*, 26(1-2):85–103, 2006.

[38] Tony F. Chan and Jianhong Shen. Mathematical models for local nontexture inpaintings. *SIAM Journal of Applied Mathematics*, 62(3):1019–1043, 2001.

[39] Tony F. Chan and Jianhong Shen. Variational image inpainting. *Communications on Pure and Applied Mathematics*, 58, February 2005.

[40] Tony F Chan and Jianhong Shen. *Image Processing and Analysis - variational, PDE, wavelet, and stochastic methods*. SIAM, 2006.

[41] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[42] D. Cremers and G. Funka-Lea. Dynamical statistical shape priors for level set based sequence segmentation. In *3rd Workshop on Variational and Level Set Methods in Computer Vision*, LNCS 3752, pages 210–221. Springer Verlag, 2005.

[43] D. Cremers and S. Soatto. A pseudo-distance for shape priors in level set segmentation. In O. Faugeras and N. Paragios, editors, *2nd IEEE Workshop on Variational, Geometric and Level Set Methods in Computer Vision*, 2003.

[44] Daniel. Cremers. *Statistical Shape Knowledge in Variational Image Segmentation*. Phd thesis, Department of Mathematics and Computer Science, University of Mannheim, July 2002.

[45] Daniel Cremers. Dynamical statistical shape priors for level set-based tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1262–1273, 2006.

[46] Daniel Cremers, Nil Sochen, and Christoph Schnörr. Towards recognition-based variational segmentation using shape priors and dynamic labeling. In *Scale Space 2003*, LNCS 2695, pages 388–400. Springer Verlag, 2003.

[47] A Criminisi, P Perez, and K Toyama. Object removal by exemplarbased inpainting. In *Conference on Computer Vision and Pattern Recognition, CVPR 03*, volume 2, pages 721–728, 2003.

[48] A. Criminisi, P. Prez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions On Image Processing*, 13(9):1200–1212, september 2004.

[49] Ann Cuzol, Kim S. Pedersen, and Mads Nielsen. Field of particle filters for image inpainting. *Journal of Mathematical Imaging and Vision*, 31(2-3):147–156, July 2008.

[50] P. Dani and S. Chaudhuri. Automated assembling of images: Image montage preparation. 28(3):431–445, March 1995.

[51] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of SIGGRAPH*, Los Angeles, California, USA, August 2001.

[52] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1033–1038, Corfu, Greece, September 1999.

[53] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions On Image Processing*, 15(12):3736–3745, December 2006.

[54] M. Elad, J. Starck, P. Querre, and D. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis*, 19(3):340–358, November 2005.

[55] Lars Eldén. *Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[56] Cheng en Guo, Song-Chun Zhu, and Ying Nian Wu. Towards a mathematical theory of primal sketch and sketchability. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, volume II, pages 1228–1235, 2003.

[57] Cheng en Guo, Song-Chun Zhu, and Ying Nian Wu. Primal sketch: Integrating structure and texture. *Comput. Vis. Image Underst.*, 106(1):5–19, 2007.

[58] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4:2379–2394, December 1987.

[59] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[60] Luc Florack. *Image Structure*, volume 10 of *Computational Imaging and Vision*. Kluwert Academic Publishers, 1997.

[61] Luc Florack, R Duits, and J Bierkens. Tikhonov regularization versus scale space: A new result. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 271–274, 2004.

[62] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning low-level vision. *Int. J. Comput. Vision*, 40(1):25–47, 2000.

[63] Mario Fritz, Eric Hayman, Barbara Caputo, and Jan-Olof Eklundh. the kth-tips (textures under varying illumination, pose and scale). http://www.nada.kth.se/cvap/databases/kth-tips/index.html, 2004.

[64] K. Fundana, N.C. Overgaard, and A. Heyden. Variational segmentation of image sequences using region-based active contours and deformable shape priors. *International Journal of Computer Vision*, 80(3), December 2008.

[65] Ketut Fundana, Niels Chr. Overgaard, Anders Heyden, David Gustavsson, and Mads Nielsen. Nonrigid object segmentation and occlusion detection in image sequences. In *3rd International Conference on Computer Vision Theory and Applications (VISAPP 08)*, 2008.

[66] Irena Galić, Joachim Weickert, Martin Welk, Andrés Bruhn, Alexander Belyaev, and Hans-Peter Seidel. Image compression with anisotropic diffusion. *J. Math. Imaging Vis.*, 31(2-3):255–269, 2008.

[67] A. Gangal and B. Dizdaroglu. Automatic restoration of old motion picture films using spatiotemporal exemplar-based inpainting. In *Advanced Concepts for Intelligent Vision Systems ACIVS*, pages 55–66, 2006.

[68] I. M. Gelfand and S. V. Fomin. *Calculus of variations*. Dover, 1963.

[69] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[70] C. Gentile, O. Camps, and M. Sznaier. Segmentation for robust tracking in the presence of severe occlusion. *IEEE Transactions On Image Processing*, 13(2):166–178, 2004.

[71] J.M. Geusebroek. The stochastic structure of images. In *Proceedings of Scale Space Methods in Computer Vision SS*, pages 327–338, 2005.

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

[73] J.M. Geusebroek and A.W.M. Smeulders. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62(1-2):7–16, April 2005.

[74] Chris A. Glasbey and Kanti V. Mardia. A review of image-warping methods. *Journal of Applied Statistics*, 25(2):155–171, April 1998.

[75] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins, 3rd edition, 1996.

[76] Y. Gousseau and J-M Morel. Are natural images of bounded variation? *SIAM Journal of Mathematical Analysis*, 3(33):634–648, 2001.

[77] Y. Gousseau and F. Roueff. Modeling occlusion and scaling in natural images. *SIAM Journal of Multiscale Modeling and Simulation*, 1(6):105–134, 2007.

[78] Ulf Grenander and Anuj Srivastava. Probability models for clutter in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):424–429, 2001.

[79] Ulf Grenander, Anuj Srivastava, and Michael Miller. Asymptotic performance analysis of bayesian object recognition. *IEEE Transactions on Information Theory*, 46(4):1658–1666, 2000.

[80] Lewis D. Griffin. Scale-imprecision space. *Image Vision Comput.*, 15(5):369–398, 1997.

[81] David Gustavsson. Multi-scale texture and geometric structure image database - ms-gti-i. to be written... DIKU-666, DIKU, 2008. Will contain collection procedure and contents....

[82] David Gustavsson, Ketut Fundana, Niels-Ch. Overgaard, and Mads Nielsen. Variational segmentation and contour matching of non-rigid moving object. In *Workshop on Dynamical Vision 2007*, 2007.

[83] David Gustavsson, Kim S. Pedersen, Francois Lauze, and Mads Nielsen. On the rate of structural change in scale spaces. In *Proceedings of Scale Space and Variational Methods in Computer Vision SSVM*, 2009.

[84] David Gustavsson, Kim S. Pedersen, and Mads Nielsen. Geometric and texture inpainting by gibbs sampling. In *SSBA-2007*, 2007.

[85] David Gustavsson, Kim S. Pedersen, and Mads Nielsen. A SVD based image complexity measure. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009.

[86] David Gustavsson, Kim Steenstrup Pedersen, and Mads Nielsen. Image inpainting by cooling and heating. In Bjarne Ersbøll and Kim Steenstrup Pedersen, editors, *Scandinavian Conference on Image Analysis (SCIA '07)*, volume 4522 of *Lecture Notes in Computer Science*, pages 591–600. Springer Verlag, June 2007.

[87] Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems. Numerical Aspects of Linear Inversion.* SIAM, Philadelphia, 1998.

[88] Per Christian Hansen. The l-curve and its use in the numerical treatment of inverse problems. In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

[89] Per Christian Hansen and Dianne Prost O'Leary. The use of the l-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503, 1993.

[90] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[91] David J. Heeger and James R. Bergen. Pyramid-based texture analysis/synthesis. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 229–238, New York, NY, USA, 1995. ACM.

[92] Ellen C. Hildreth. Computations underlying the measurement of visual motion. pages 99–146, 1987.

[93] Berthold K P. Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[94] Jinggang Huang and David Mumford. Statistics of natural images and models. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 01:1541, 1999.

[95] X.S. Huang, S.Z. Li, and Y.S. Wang. Evaluation of face alignment solutions using statistical learning. In *Proceedings of International Conference on Automatic Face and Gesture Recognition AFGR*, pages 213–218, 2004.

[96] Aapo Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[97] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[98] T. Iijima. Basic theory on normalization of a pattern. *Bulletin of Electrical Laboratory*, 26:368–388, 1962. In Japanese.

[99] M. Irani and P. Anandan. All about direct methods. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Workshop on Vision Algorithms: Theory and practice*. Springer-Verlag, 1999.

[100] Anil K. Jain and Farshid Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern Recogn.*, 24(12):1167–1186, 1991.

[101] B. Julesz and R. Bergen. Textons, the elements of texture perception, and their interactions. *Nature*, 290:91–97, 1981.

[102] Christian Jutten and Jeanny Herault. Blind separation of sources, part 1: an adaptive algorithm based on neuromimetic architecture. *Signal Process.*, 24(1):1–10, 1991.

[103] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Series*. Springer, Berlin, 2004.

[104] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, V1(4):321–331, January 1988.

[105] S.H. Keller, F Lauze, and M. Nielsen. Deinterlacing using variational methods. *IEEE Transactions On Image Processing*, 17(11):1–14, November 2008.

[106] Sune Keller, Francoise Lauze, and Mads Nielsen. Motion compensated video super resolution. In *Proceedings of Scale Space and Variational Methods in Computer Vision SSVM*, pages 801–812, 2007.

[107] Sune H. Keller, Francoise Lauze, and Mads Nielsen. A total variation motion adaptive deinterlacing scheme. In *Proceedings of Scale Space Methods in Computer Vision SS*, pages 408–418, 2005.

[108] Sune Hogild Keller. *Video Upscaling Using Variational Methods*. PhD thesis, University of Copenhagen, 2007.

[109] Michael Kirby. *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[110] Josef Kittler and J. Föglein. Contextual classification of multispectral pixel data. *Image and Vision Computing*, 2(1):13–29, 1984.

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

[112] Jan J. Koenderink and Andrea J. Van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2-3):159–168, 1999.

[113] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, February 2004.

[114] J. Konrad and M. Ristivojevic. Video segmentation and occlusion detection over multiple frames. In *Image and Video Communications and Processing 2003*, SPIE 5022, pages 377–388. SPIE, 2003.

[115] G. Laccetti, L. Maddalena, and A. Petrosino. Removing line scratches in digital image sequences by fusion techniques. In *International Conference on Image Analysis and Processing CIAP*, pages 695–702, 2005.

[116] Francois B. Lauze. *Computational Methods For Motion Recovery, Motion Compensated Inpainting and Applications*. PhD thesis, IT University of Copenhagen, 2004.

[117] Ann B. Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1-2):35–59, 2001.

[118] Daniel D. Lee and Sebastian H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[119] Daniel D. Lee and Sebastian H. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of Conference on Neural Information Processing Systems NIPS*, volume 13, pages 556–562, 2001.

[120] M.E. Leventon, W.E.L. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 316–323, 2000.

[121] Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Computer Science Workbench. Springer, 2001.

[122] Tony Lindeberg. *Scale-Space Theory in Compter Vision*. Kluwert Academica Publishers, 1994.

[123] Jun S Liu. *Monte Carlo Strategies in Scientific Computing*. Springe Series in Statistics. Springer-Verlag, 2004.

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

[125] J.B.A. Maintz, P.A. van den Elsen, and M.A. Viergever. 3d multimodality medical image registration using morphological tools. *Image and Vision Computing*, 19(1-2):53–62, January 2001.

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

[127] P. Markelj, D. Tomazevic, F. Pernus, and B. Likar. Robust gradient-based 3-d/2-d registration of ct and mr to x-ray images. *IEEE Transations On Medical Imaging*, 27(12):1704–1714, December 2008.

[128] David Marr. *VISION: a computational investigation into the human representation and processing of visual information.* W. H. Freeman, San Francisco, 1982.

[129] S Masnou. Disocclusion: a variational approach using level lines. *IEEE Transactions On Image Processing*, 11(2):68–76, February 2002.

[130] S Masnou and Jean-Michel Morel. Level lines based disocclusion. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pages 259–263, 1998.

[131] S. G. Matheron. *Random Sets and Integral Geometry.* John Wiley and Sons, New York, 1975.

[132] Yves Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations: The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures.* American Mathematical Society (AMS), Boston, MA, USA, 2001.

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

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

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

[136] Jan Modersitzki. *Numerical Methods for Image Registration*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2004.

[137] M. Moelich and T. Chan. Tracking objects with the chan-vese algorithm. Technical Report 03-14, Department of Mathematics, UCLA, March 2003.

[138] V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

[139] Pavel Mrázek and Mirko Navara. Selection of optimal stopping time for nonlinear diffusion filtering. *International Journal of Computer Vision*, 52(2-3):189–203, 2003.

[140] David Mumford. Bayesian rationale for the variational formulation. In Bart M. ter Haar Romeny, editor, *Geometry-Driven Diffusion in Computer Vision*, volume 1 of *Computional Imaging and Vision*, pages 135–146, 1994.

[141] David Mumford and Jayant Shah. Boundary detection by minimizing functionals. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22–26, San Fransisco, 1985.

[142] Arnold Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40:636–666, 1998.

[143] Mads Nielsen, Luc Florack, and Rachid Deriche. Regularization, scale-space, and edge detection filters. *International Journal of Computer Vision*, 7(4):291–307, October 1997.

[144] Mila Nikolova. Counter-examples for bayesian map restoration. In *Proceedings of Scale Space and Variational Methods in Computer Vision SSVM*, pages 140–152, 2007.

[145] S Nishikawa, R Massa, and J Mott-Smith. Area properties of television pictures. *IEEE Transactions on Information Theory*, 11(3):348–352, July 1965.

[146] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[147] Aude Oliva and Antonio B. Torralba. Scene-centered description from spatial envelope properties. In *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 263–272, London, UK, 2002. Springer-Verlag.

[148] Ole Fogh Olsen and Mads Nielsen. Multi-scale gradient magnitude watershed segmentation. In *ICIAP'97 - 9th International Conference on Image Analysis and Processing*, volume 1310 of *Lecture Notes in Computer Science*, pages 6–13, Florence, Italy, September 1997.

[149] B A Olshausen and D J Field. Natural image statistics and efficient coding. In *Network: Computation in Neural Systems*, number 7, 1996.

[150] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997.

[151] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer-Verlag, New York, 2003.

[152] Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.

[153] N. Paragios and R. Deriche. Geodesic active contours and level set methods for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):266–280, 2000.

[154] N. Paragios and R. Deriche. Geodesic active regions and level set methods for motion estimation and tracking. *Computer Vision and Image Understanding*, 97:259–282, 2005.

[155] Maria Petrou and Pedro García Sevilla. *Dealing with Texture*. Wiley, 2006.

[156] Gabriel Peyré. Non-negative sparexe modeling of textures. In *Proceedings of Scale Space and Variational Methods in Computer Vision SSVM*, LNCS. Springer, 2007.

[157] Gabriel Peyré, Sébastien Bougleux, and Laurent Cohen. Non-local regularization of inverse problems. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Proceedings of European Conference on Com-*

puter Vision (ECCV), volume 5304 of *LNCS*, pages 57–68. Springer, 2008.

[158] Tomaso Poggio and Vincent Torre. Ill-posed problems and regularization analysisin early vision. Technical Report A.I. Memo 773, MIT, April 1984.

[159] Tomaso Poggio, H Voorhees, and A Yuille. A regularized solution to edge detection. Technical Report A.I. Memo 833, MIT, May 1985.

[160] B.C. Porter, D.J. Rubens, J.G. Strang, J. Smith, S. Totterman, and K.J. Parker. Three-dimensional registration and fusion of ultrasound and mri using major vessels as fiducial markers. *IEEE Transations On Medical Imaging*, 20(4):354–359, April 2001.

[161] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.

[162] Trygve Randen and John Hakon Husoy. Filtering for texture classification: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310, 1999.

[163] S.D. Rane, G. Sapiro, and M. Bertalmio. Structure and texture filling-in of missing image blocks in wireless transmission and compression applications. *IEEE Transactions On Image Processing*, 12(3):296–303, March 2003.

[164] A. Roche, X. Pennec, G. Malandain, and N.J. Ayache. Rigid registration of 3-d ultrasound with mr images: A new approach combining intensity and gradient information. *IEEE Transations On Medical Imaging*, 20(10):1038–1049, October 2001.

[165] M. Rousson and N. Paragios. Shape priors for level set representations. In *Proceedings of European Conference on Computer Vision (ECCV)*, LNCS 2351, pages 78–92. Springer Verlag, 2002.

[166] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, August 1994.

[167] Daniel L. Ruderman. Statistics of natural images. *Network: Computation in Neural Systems*, (5):517–548, 1994.

[168] Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997.

[169] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[170] Hans Sagan. *Introduction to the calculus of variations*. DOVER, 1992.

[171] J. A. Sethian. *Level Set Methods and Fast Marching Methods:Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press, 1999.

[172] E. P. Simoncelli. Bayesian denoising of visual images in the wavelet domain. In P Müller and B Vidakovic, editors, *Bayesian Inference in Wavelet Based Models*, volume 41 of *Lecture Notes in Statistics*, pages 291–308. Springer-Verlag, 1999.

[173] E P Simoncelli and E H Adelson. Noise removal via bayesian wavelet coring. In *Procedings of Third International Conference on Image Processing*, volume I, pages 379–382, Lausanne, 1996. IEEE Signal Processing Society.

[174] E.P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *Asilomar Conference on Signals, Systems and Computers*, 1997.

[175] Stephen M. Smith. Flexible filter neighbourhood designation. In *ICPR '96: Proceedings of the 1996 International Conference on Pattern Recognition (ICPR '96) Volume I*, page 206, Washington, DC, USA, 1996. IEEE Computer Society.

[176] Stephen M. Smith and J. M. Brady. SUSAN - a new approach to low level image processing. Technical Report TR95SMS1c, Chertsey, Surrey, UK, 1995.

[177] Stephen M. Smith and J. Michael Brady. SUSAN - a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, 1997.

[178] Pierre Soille. Morphological image compositing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):673–683, May 2006.

[179] J. E. Solem and N. Chr. Overgaard. A geometric formulation of gradient descent for variational problems with moving surfaces. In *Scale-Space 2005*, LNCS 3459, pages 419–430. Springer Verlag, 2005.

[180] Jon Sporring. The entropy of scale-space. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, volume I, Washington, DC, USA, 1996. IEEE Computer Society.

[181] Jon Sporring and Joachim Weickert. On generalized entropies and scale-space. In *SCALE-SPACE '97: Proceedings of the First International Conference on Scale-Space Theory in Computer Vision*, pages 53–64, London, UK, 1997. Springer-Verlag.

[182] Jon Sporring and Joachim Weickert. Information measures in scale-spaces. *IEEE Transactions on Information Theory*, 45:1051–1058, 1999.

[183] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003.

[184] Anuj Srivastava, Xiuwen Liu, and Ulf Grenander. Universal analytical forms for modeling image probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1200–1214, 2002.

[185] C. Strecha, R. Fransens, and L. V. Gool. A probabilistic approach to large displacement optical flow and occlusion detection. In *Statistical Methods in Video Processing*, LNCS 3247, pages 71–82. Springer Verlag, 2004.

[186] D. Strong and T. F. Chan. Exact solutions to total variation problems. Technical Report 96-41, UCLA, Ca., 1996.

[187] Richard Szeliski. Image alignment and stitching: a tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1):1–104, 2006.

[188] Bart M. ter Haar Romeny. *Front-End Vision and Multi-Scale Image Analysis: Multi-Scale Computer Vision Theory and Applications, written in Mathematica*, volume 27 of *Computional Imaging and Vision*. Kluwer Academic Publishers, 2003.

[189] Alan M. Thompson, John C. Brown, Jim W. Kay, and D. Michael Titterington. A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):326–339, 1991.

[190] N. P. Tiilikainen, A.E.Bartoli, and S. Olsen. Contour-based registration and retexturing of cartoon-like videos. In *Proceedings of British Machine Vision Conference (BMVC)*, 2008.

[191] Philip H. S. Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Workshop on Vision Algorithms*, pages 278–294. Springer-Verlag, 1999.

[192] Antonio Torralba and Aude Oliva. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9), September 2002.

[193] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391 – 412, August 2003.

[194] David Tschumperle. Curvature-preserving regularization of multi-valued images using pde's. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages II: 295–307. Springer-Verlag, 2006.

[195] David Tschumperlé. Fast anisotropic smoothing of multi-valued images using curvature-preserving pde's. *International Journal of Computer Vision*, 68(1):65–82, 2006.

[196] A. van der Shaaf and J.H. van Hateren. Modelling the power spectra of natural images : Statistics and information. *Vision research*, 36(17):2759–2770, 1998.

[197] J. H. van Hateren and A. vad der Schaaff. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Soc. Lond. B*, 265:359–366, 1998.

[198] Curtis R. Vogel. *Computational Methods for Inverse Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

[199] Y.Z. Wang and S.C. Zhu. Perceptual scale-space and its applications. *International Journal of Computer Vision*, 80(1), October 2008.

[200] Joachim Weickert. *Anisotropic Diffusion in Image Processing*. ECMI. Teubner-Verlag, 1998.

[201] Gerhard Winkler. *Image Analysis, Random Fields, and Markov Chain Monte Carlo Methods*. Number 27 in Stochastic Modelling and Applied Probability. Springer-Verlag, 2006.

[202] Andrew P. Witkin. Scale-space filtering. In *Proceedings 8th International Joint Conference on Artificial Intelligence*, volume 2, pages 1019–1022, Karlsruhe, August 1983.

[203] A. Wong and W. Bishop. Efficient least squares fusion of mri and ct images using a phase congruency model. *Pattern Recognition*, 29(3):173–180, February 2008.

[204] Fei Wu, Changshui Zhang, and Jingrui He. An evolutionary system for near-regular texture synthesis. *Pattern Recogn.*, 40(8):2271–2282, 2007.

[205] Ying Nian Wu, Cheng-En Guo, and SongChun Zhu. From information scaling of natural images to regims of statistical models. *Quarterly of Applied Mathematics*, 2007.

[206] S.C. Yan, C. Liu, S.Z. Li, H.J. Zhang, H.Y. Shum, and Q.S. Cheng. Face alignment using texture-constrained active shape models. *Image and Vision Computing*, 21(1):69–75, January 2003.

[207] Victoria Yanulevskaya and Jan-Mark Geusebroek. Significance of the Weibull distribution and its sub-models in natural images. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2009.

[208] Laurent Younes. Computable elastic distances between shapes. *SIAM Journal on Applied Mathematics*, 58(2):565–586, 1998.

[209] S.C. Zhu, C.E. Guo, Z.J. Xu, and Y.Z. Wang. What are textons? In *Proceedings of European Conference on Computer Vision (ECCV)*, page IV: 793 ff., 2002.

[210] S.C. Zhu and Y.Z. Wang. Perceptual scale-space and its applications. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages I: 58–65, 2005.

[211] Song Chun Zhu and David Mumford. Grade: Gibbs reaction and diffusion equation - a framework for pattern synthesis, denoising, image enhancement, and clutter removal. *IEEE Trans. PAMI*, 19(11):1627–1660, november 1997.

[212] Song Chun Zhu and David Mumford. Prior learning and gibbs reaction-diffusion. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250, November 1997.

[213] Song Chun Zhu, Ying Nian Wu, and David Mumford. Minimax entropy principle and its application to texture modelling. *Neural Computation*, 9(8):1627–1660, 1997.

[214] Song Chun Zhu, Ying Nian Wu, and David Mumford. Filters, random fields and maximum entropy FRAME: To a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

[215] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, October 2003.