Modeling Spatial and Temporal Textures

by

Fang Liu

B.S., Beijing University, China (1984)
M.S., Northeastern University (1989)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

Author ....
Program in Media Arts and Sciences
August 8, 1997

Certified by ....
Rosalind W. Picard
NEC Development Professor of Computers and Communications
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by ....
Stephen A. Benton
Chairman, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Modeling Spatial and Temporal Textures
by
Fang Liu

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 8, 1997, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Bridging the gap between the Wold random process decomposition theory and practical texture modeling, this research establishes Wold-based texture modeling as an important method for a wide range of applications that benefit from efficient and effective characterization of textural information.

- A robust and efficient algorithm is developed for spectral 2-D Wold decomposition of homogeneous or near homogeneous random fields.
- A psychophysical study is conducted to show that the Wold component energy of a texture pattern is a good computational measure for the most salient human texture perception dimension of repetitiveness vs. randomness.
- A shift, rotation, and scale invariant Wold-based texture model is presented. This model provides efficient and perceptually sensible features that are robust to many natural texture inhomogeneities.
- For model perspective invariance, a linear system characterization and a decomposition of image perspective transformations are proposed to form a basis for future algorithms to infer image perspective parameters from a single sample of harmonic texture data.
- Based on the Wold texture model, an algorithm is developed for textured image database retrieval. Compared to the state-of-the-art texture models, the new model appears to offer perceptually more satisfying retrieval results while matching or surpassing the best recognition performance of the others.
- A K-means-based image segmentation method is presented to demonstrate the use of Wold-based modeling in characterizing textured regions in natural scene images.
- Applying the principle of Wold decomposition to temporal texture modeling, a robust and efficient algorithm is developed for detecting and segmenting periodic motion. The use of periodicity templates is proposed for characterizing periodicity in space and time.

Thesis Supervisor: Rosalind W. Picard
Title: NEC Development Professor of Computers and Communications
Program in Media Arts and Sciences

Doctoral Committee

Thesis Advisor ....

Rosalind W. Picard
NEC Development Professor of Computers and Communications
Program in Media Arts and Sciences

Thesis Reader ....

Michele M. Covell
Ph.D., Member of Research Staff
Interval Research Corp.

Thesis Reader ....

Alex P. Pentland
Toshiba Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader ....

Tomaso A. Poggio
Uncas and Helen Whitaker Professor of Vision Sciences and Biophysics
Dept. of Brain and Cognitive Sciences

Acknowledgments

Grateful and sincere thanks to my advisor, Prof. Rosalind Picard, for her guidance and constant support. Her patience and the freedom she gave me in my research are also very much appreciated. Among other things, I learned a great deal from her about conducting research and pursuing a professional career.

Many thanks to the other members of my doctoral committee: Dr. Michele Covell, Prof. Alex Pentland, and Prof. Tomaso Poggio. Their comments and insightful suggestions have helped to strengthen the work and improve the clarity of the final document. Several long and informative conversations with Prof. Pentland have broadened my view on various aspects of this research.

Thanks also go to Prof. Edward Adelson, who led me into the area of visual perception, and Prof. Aaron Bobick, who has always been willing to help and offered many constructive critiques on my work. Thanks to Joseph Francos, for many early discussions on the 2-D Wold decomposition theory; to Thomas Minka, for benchmarking several texture models; and to Jocelyn Riseberg, for getting me started on human experiments.

The Vision and Modeling Group has been a stimulating and enjoyable place to work. Thanks to all: Ali Azarbayejani, Bea Bailey, Sumit Basu, Dave Becker, Judy Bornstein, Matthew Brand, Lee Campbell, Laureen Chapman, Trevor Darrell, Jim Davis, Irfan Essa, Raul Fernandez, Martin Friedmann, Bill Freeman, Monika Gorkani, Jennifer Healey, Bradley Horowitz, Stephen Intille, Yuri Ivanov, Tony Jebara, Jonathan Klein, Steve Mann, Baback Moghaddam, Kate Mongiat, Chahab Nastar, Nassir Navab, Karen Navarro, Nuria Oliver, Claudio Pinhanez, Kris Popat, Deb Roy, Stan Sclaroff, Alex Sherstinsky, Eero Simoncelli, Pawan Sinha, Flavia Sparacino, Thad Starner, Martin Szummer, Hong Z. Tan, Erik Trimble, Matthew Turk, Joshua Wachman, John Wang, Laurie Ward, Andy Wilson, Chris Wren, and many others.

Special thanks to my family, a perpetual source of support and encouragement, and specially to Steven Rosenthal, for his love, understanding, and patience.

This research was conducted at the MIT Media Laboratory and was sponsored in part by BT, Interval Research Corp., IBM, and NEC.

Chapter 1

Introduction

1.1 Texture

The most obvious property of texture is perhaps its ubiquity. While two-dimensional visual and three-dimensional haptic textures are the most intuitive, "texture" can also be used to characterize audio signals and spatiotemporal events, such as music and motion.

Ubiquitous as it is, a formal definition of texture remains elusive. In the literature, various textual properties are often used to serve as the definition of texture or, more precisely, to constrain the domain of problems. Commonly seen descriptions of textures include coarseness, contrast, directionality, regularity, uniformity, roughness, busyness, *etc.*. Categorically, texture has also been considered as "stuff", as opposed to "things"; the latter of which is usually associated with identifiable objects.

In this dissertation, the term texture refers to signals that exhibit statistically certain degrees of homogeneity and can be regarded as stationary or near stationary one-dimensional (1-D) or two-dimensional (2-D) random processes. Other types of homogeneity exist. An example is self-similar texture. The modeling of such texture has been addressed by others using fractal models [54][73], and will not be considered here.

1.2 Texture Modeling

1.2.1 Texture Models

Texture models provide computational features that facilitate tasks such as image understanding, representation, and synthesis. Historically, texture models are categorized as statistical, structural, or statistical-structural (hybrid) [36][37][88][93].

Statistical Models

The statistical approach focuses on the statistical properties of textures. A texture pattern is characterized either by statistics of image pixel gray scale values or by a stochastic model. Early methods include autocorrelation [49], run length [33], and co-occurrence [38]. In the 1980's, a large volume of literature appeared in the area of texture modeling using Markov-type random field models. The work of Whittle [94], Woods [97], Besag [12], and Kashyap [52] made fundamental contributions to this development. Connections between image pixel value co-occurrence matrices and Markov/Gibbs random field modeling were established in [25]. An excellent review of work in Markov random field modeling can be found in [22]. Markov-type random field models typically are effective for the analysis and synthesis of micro and random looking textures, but not much so for that of larger scale and more structured patterns [21].

Structural Models

The structural methods represent a texture pattern by its textural primitives and their spatial placement rules. Examples for both texture analysis and synthesis using structural models can be found in [68][89][90]. The main deficiency of the structural methods is that they are incapable of capturing or generating the randomness that natural textures almost always possess.

Statistical-structural Models

Analogies have been made between the structural/statistical characterization of texture models and the attentive/pre-attentive dichotomy in low-level human vision [95]. However, this division of texture models is based on the functional properties of the models, and natural textures usually contain both structural and statistical components. Texture models capable of representing both structure and randomness have been studied. Garand and Weinman proposed a hybrid model which uses structured low frequency Fourier information as the initial state for a Gibbs random field model to synthesize cloud textures [34]. Picard discussed the parameter selection of this hybrid model and explored the use of an external field to introduce structure to a Gibbs random field [74]. More recently, Francos *et al.* proposed a unified texture model based on the 2-D Wold decomposition of homogeneous random fields [27]. Wold-based texture modeling (Wold-based modeling or Wold model for short) is the focus of this dissertation.

1.3 Criteria for Texture Models

Different criteria can be applied when evaluating a texture model, often biased by the particular application in hand. Two of the most common considerations are:

1. The ability to faithfully reconstruct the data. The quality of the reconstruction can be measured in two ways. One is by the pixel-level mean-squared error criteria. The other is by the perceptual resemblance. For example, two pictures of a grass lawn may have large mean-squared pixel difference but look very similar.

2. The model efficiency. Two types of efficiencies are involved. One is in data representation, i.e., the ratio between the model feature size and data size. The other is the level of complexity in model implementation and feature computation. A model can be easy to compute but produce large feature set or vice versa.

Emerging applications such as image and video database retrieval pose new challenges to texture modeling. In those applications, the computer system serves the purpose of saving human users the time and effort of browsing the entire database. It is often expected that the retrieved images resemble the visual properties of a given prototype. For such a system, it is important that the computational image features used for pattern comparison are meaningful to human perception. This motivates the following additional criterion:

3. The model perceptual properties, *i.e.,* the perceptual interpretation of model computational features.

1.4 Focus: Wold-based Texture Modeling

The focus of this dissertation research is on Wold-based Texture Modeling. The mathematical foundation of Wold-based texture modeling is the 2-D Wold decomposition of homogeneous random fields. The 2-D Wold theory allows a textured image to be decomposed into three mutually orthogonal components: harmonic, evanescent, and indeterministic (random). These component images can be characterized separately.

Wold-based texture modeling is capable of satisfying all three criteria discussed above. Francos *et al.* applied Wold-based models to image coding and reconstruction [26][27][85]. It was shown in their work that a handful of model parameters could reconstruct natural textures that are visually indistinguishable from the originals. The most important advantage of Wold-based texture modeling lies with the third criterion. An independent psychophysical study has identified the top three perceptual dimensions of human texture perception as "repetitiveness", "directionality", and "granularity and complexity" [79]. As shown by the Brodatz texture [15] examples in Figure 1-1, the perceptual properties of the Wold components can be described as "periodicity", "directionality", and "randomness", agreeing closely with the findings of the human study.

1.5 Goal and Contributions

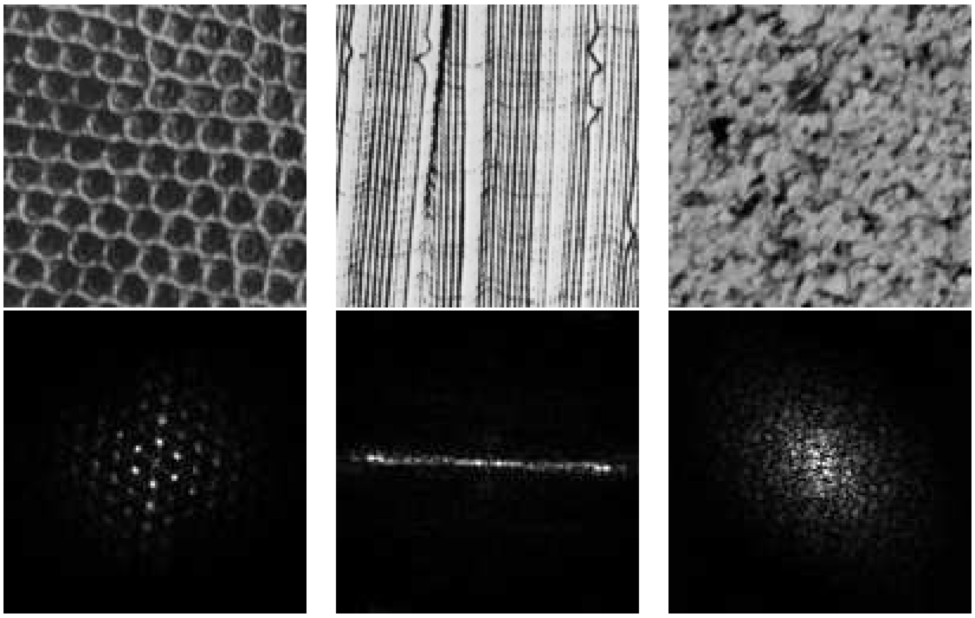
The goal of this research is to establish Wold-based texture modeling as an important method for a wide range of applications that benefit from efficient and effective characterization of textural information. This goal is achieved by bridging the gap between the Wold theory and practical texture modeling. The main contributions of this work are as follows:

- A robust and efficient spectral 2-D Wold decomposition algorithm is developed for homogeneous or near homogeneous random fields.

- A computational measure for the most salient human texture perception dimension of pattern repetitiveness vs. randomness is proposed and validated by a psychophysical study.

- A shift, rotation, and scale invariant Wold-based texture model is presented. This model provides efficient and perceptually sensible features that are robust to many natural texture inhomogeneities. The superior performance of the model is demonstrated in comparison to state-of-the-art texture models in a textured image database retrieval experiment.

- For model perspective invariance, a linear system characterization of image perspective transformation and its decomposition into affine and chirp transformations are presented. The relation between geometric and spectral descriptions of perspective transformation is formulated to form a basis for future algorithms to infer image perspective parameters from a single sample of harmonic texture data.



(a) (b) (c)

Figure 1-1: Examples of Brodatz textures with different prominent Wold components. Top row: originals. Bottom row: corresponding Fourier magnitude images. (a) D3: Reptile skin, having a prominent harmonic component (spectral peaks supported by point-like regions). (b) D105: Cheesecloth, having a strong evanescent component (spectral peaks supported by line-like regions). (c) D29: Beach sand, having mostly an indeterministic component (relatively smooth spectrum).

- Based on the new texture model, an image retrieval algorithm is developed for textured image databases. Compared to other well-known models, the Wold model appears to offer perceptually more satisfying results in the image retrieval experiments while matching or surpassing the best recognition performance of state-of-the-art texture models.
- Applying the principle of Wold decomposition to temporal texture modeling, a robust and efficient algorithm is developed for detecting and segmenting periodic motion. The use of periodicity templates is also proposed for characterizing periodicity in space and time.

1.6 Organization

The rest of the dissertation is organized as follows.

Chapter 2 A concise but comprehensive review of the 2-D Wold decomposition theory for random fields is presented. This mathematical framework is the theoretical foundation of this dissertation. Certain approximations to the theory are also discussed for practical applications.

Chapter 3 A spectral 2-D Wold decomposition algorithm for homogeneous or near homogeneous random fields is presented. This algorithm detects the Fourier spectral harmonic and evanescent frequencies of a textured image and decomposes the image by extracting these frequency components from the image spectrum.

Chapter 4 The perceptual property of Wold-based texture modeling is investigated by conducting a human texture ranking experiment. Human subjects and a computer order a set of texture samples along the perceptually most salient dimension of human texture perception. The ranking scores are analyzed to examine the concordance of the human rankings and the correlation between the computer behavior and that of the humans.

Chapter 5 A Wold-based texture model is constructed. The model emphasizes the perceptually most salient harmonic structures in a texture pattern and is designed for use in large collections of natural textures. Based on this new model, an image retrieval algorithm is developed for textured image databases. The model invariance study leads to the decomposition of perspective transformation. A K-means-based image segmentation method is also presented to demonstrate how the Wold model can be used to characterize textured regions in natural scenes.

Chapter 6 Based on the principle of 1-D Wold decomposition, an algorithm is developed to model temporal textures for image sequence analysis. This robust and computationally efficient method allows the detection, segmentation, and characterization of periodic motion to be accomplished simultaneously. The use of periodicity templates is also proposed for characterizing periodicity in space and time.

Chapter 7 Conclusions.

Chapter 8 Future research directions related to this work are suggested.

Chapter 2

Theoretical Background

2.1 Introduction

This chapter reviews the mathematical foundation for this dissertation work. The intent is to make a concise but comprehensive presentation of a theoretical framework to which many people have contributed over the years. References are given throughout the chapter for details and proofs.

The original decomposition theory due to Wold applies to the analysis of one-dimensional stationary random processes [96]. It provides a general representation of such processes, as well as an interpretation of the representation in terms of linear prediction. To analyze the structure of a 2-D discrete random field, a two-dimensional linear prediction problem can be similarly formulated.

In the following discussion, it is assumed that the 2-D random field $\{y(m,n)\}$, $(m,n) \in \mathcal{Z}^2$, is real and zero mean. In addition, the second-order moments of $y(m,n)$ are assumed to be finite,

$$\sup_{(m,n)\in\mathcal{Z}^2} E\left[y^2(m,n)\right] < \infty, \tag{2.1}$$

and $E\left[y^2(m,n)\right] > 0$ for at least one $(m,n) \in \mathcal{Z}^2$. Symbol $E\left[\cdot\right]$ denotes the expected value. The objective is to find the minimum-norm linear predictor of $y(m,n)$ as the projection of $y(m,n)$ on the Hilbert space spanned by all the field samples that are in the "past" relative to the "present".

Since there is no natural definition of "past" and "future" in a 2-D plane, different order definitions can lead to different orthogonal decompositions of a 2-D random field. Various choices of the "past", such as symmetric half-plane, non-symmetric half-plane (NSHP), quarter-plane, "vertical" and "horizontal" half-planes, etc., have been used and resulted in two-fold, three-fold and four-fold Wold-like decomposition of 2-D homogeneous random fields [18][19][42][43] [50][51][57]. It is shown by Francos *et al.* [28] that, by considering a countably infinite set of "total-order" and rational non-symmetric half-plane (RNSHP) support (see Section 2.2.1), a corresponding countably-infinite-fold Wold-like decomposition of 2-D homogeneous random field can be obtained and the two-fold, three-fold, and four-fold Wold-like decompositions are special cases of this countably-infinite-fold decomposition. Francos *et al.* also generalized the NSHP-based decomposition to the case of 2-D non-homogeneous random fields [31].

In the following, first the 2-D linear prediction problem is formulated. Then, the Wold-like decomposition of 2-D non-homogeneous random fields is presented, followed by the Wold-like decomposition of homogeneous random fields and the corresponding spectral decomposition. To apply

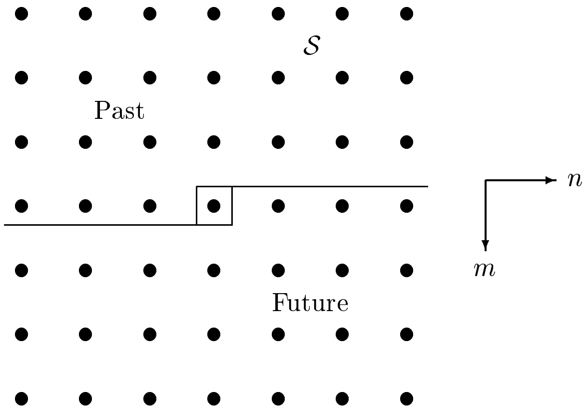


Figure 2-1: Totally ordered non-symmetric half-plane (NSHP) support $\mathcal{S}$ of a 2-D plane. The center location is "present".

---

the 2-D Wold decomposition to practical problems, certain approximations of the theory are made. Finally, examples are given for illustration.

2.2 Linear Prediction Formulation

In this section, the problem of 2-D linear prediction of a random field is formulated based on both infinite and finite supports.

2.2.1 Total-order and Support in 2-D Plane

Definition 1 *In the 2-D plane, a* **total-order** *can be defined for the samples of a random field* $\{y(m,n)\}$, $(m,n) \in \mathcal{Z}^2$, *in a raster-scan manner: row after row, from left to right and top to bottom. The order* $\prec$ *is*

$$(i,j) \prec (s,t) \text{ \textbf{iff} } (i,j) \in \{(k,l) \mid k = s, l < t\} \cup \{(k,l) \mid k < s, -\infty < l < \infty\} \tag{2.2}$$

and the order $\preceq$ *is*

$$(i,j) \preceq (s,t) \text{ \textbf{iff} } (i,j) \prec (s,t) \text{ or } (i,j) = (s,t). \tag{2.3}$$

Based on the total-order definition, a **totally ordered, non-symmetric half-plane (NSHP) support** $\mathcal{S}$ *can be defined as follows. Given the* (m,n)*-th sample as "present", all* $(i,j) \prec (m,n)$ *are in the "past", and the rest are in the "future".*

This NSHP support is illustrated in Figure 2-1, where the "present" is at the center.

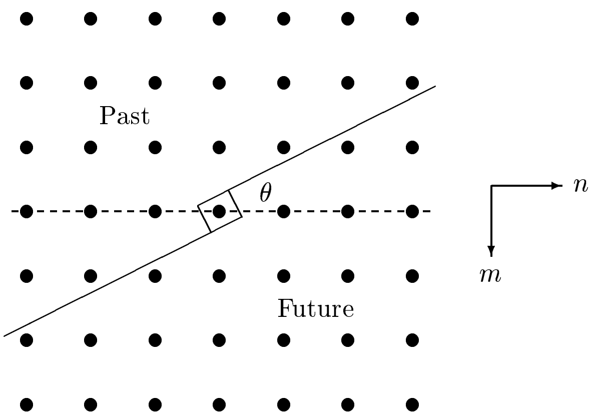


Figure 2-2: Example of a totally ordered rational non-symmetric half-plane (RNSHP) support, rotated from $\mathcal{S}$ by an angle $\theta = tan^{-1}(1/2)$. The center location is "present".

Obviously, the total-order and the NSHP support of Definition 1 is not unique on the 2-D lattice. Keeping the structure of the 2-D discrete sampling grid the same, multiple definitions of total-order and NSHP support can be made.

Definition 2 *Let α and β be co-prime integers and $\alpha \neq 0$. A new total-order and NSHP support can be defined on the original 2-D grid by rotating the NSHP total ordering $\mathcal{S}$ of Definition 1 counterclockwise by an angle*

$$\theta = tan^{-1}\left(\frac{\beta}{\alpha}\right)$$

about the origin of its coordinate system. This new support is called the **rational non-symmetric half-plane (RNSHP) support** *since its boundary line is of rational slope. Denote the set of all possible total order and RNSHP support defined in this manner by $\mathcal{O}$,*

$$\mathcal{O} = \{o \,|\, o = (\alpha, \beta);\ \alpha, \beta \ are\ co\text{--}prime\ integers\}\,.$$

Note that $\mathcal{O}$ is a countably infinite set.

By Definition 2, the total-order and NSHP support $\mathcal{S}$ of Definition 1 can be denoted as $o = (1, 0)$. Figure 2-2 shows an example of RNSHP total ordering, with $\alpha = 2$ and $\beta = 1$.

In the following, certain definitions and theorems are stated with respect to (w.r.t.) a particular total-order and NSHP support definition $o \in \mathcal{O}$. In places, this total-order and NSHP dependency is noted by the superscript or the subscript o.

2.2.2 Linear Predictor Based on Infinite Support

Let $\mathcal{H}$ denote the **Hilbert space** formed by the random variables $y(m,n), (m,n) \in \mathcal{Z}^2$, with the inner product of any two random variables $y(m,n)$ and $y(s,t)$ defined as $E[y(m,n)y(s,t)]$. Then the closed linear manifold spanned by the set $\{y(s,t)\}, (s,t) \preceq (m,n)$ by the total-order $o \in \mathcal{O}$, is a subspace of $\mathcal{H}$:

$$\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n)} = \overline{Sp}\{y(s,t)|(s,t) \preceq (m,n)\} \subset \mathcal{H}. \tag{2.4}$$

Note that this definition implies the **nesting property**:

$$\overset{o}{\mathcal{H}}{}^{y}{}_{(s,t)} \subset \overset{o}{\mathcal{H}}{}^{y}{}_{(m,n)}, \qquad \textit{if } (s,t) \preceq (m,n).$$

Definition 3 *A predictor of $y(m,n)$ is* **causal** *and of* **continuous support** *w.r.t. the order defined in Definition 1 if it depends on all and only the preceding samples.*

Definition 4 *The minimum-norm, causal, continuous support, linear predictor of $y(m,n)$ is the projection of $y(m,n)$ on the Hilbert space $\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n-1)}$. Denote the predictor as $\hat{y}(m,n)$, then*

$$\hat{y}(m,n) = \sum_{(0,0)\prec(k,l)} b_{(m,n)}(k,l)\, y(m-k,n-l). \tag{2.5}$$

2.2.3 Linear Predictor Based on Finite Support

In practice, only a finite number of samples are available. Therefore, it is necessary to consider a finite support.

Definition 5 *Define the 2-D* **discontinuous** *and* **finite half-plane support** *as*

$$S_{M,N} = \{(k,l) \mid k = 0,\, 1 \le l \le N\} \cup \{(k,l) \mid 1 \le k \le M, -N \le l \le N\}, \tag{2.6}$$

where M and N are positive integers.

Correspondingly, let

$$\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n);S_{M,N}} = \overline{Sp}\{y(m-k,n-l) \mid (k,l) \in \{S_{M,N} \cup \{(0,0)\}\}\}.$$

Definition 6 *The minimum-norm, causal, finite support, linear predictor of $y(m,n)$ is the projection of $y(m,n)$ on the Hilbert space $\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n-1);S_{M,N}}$. Denote the predictor as $\hat{y}_{S_{M,N}}(m,n)$, then*

$$\hat{y}_{S_{M,N}}(m,n) = \sum_{(k,l)\in S_{M,N}} b'_{(m,n)}(k,l)\, y(m-k,n-l). \tag{2.7}$$

It is shown in [31] that the prediction of $y(m,n)$ based on the continuous infinite half-plane support can be approximated by the prediction based on the discontinuous finite half-plane support.

Theorem 1

$$\lim_{M\to\infty}\lim_{N\to\infty} E\left[\hat{y}(m,n)-\hat{y}_{S_{M,N}}(m,n)\right]^2 = 0. \tag{2.8}$$

Therefore, although the 2-D Wold-like decomposition theory presented in the following sections is based on infinite 2-D support, it can be applied to random fields defined on finite 2-D discrete grids.

2.3 Decomposition of Non-homogeneous Random Fields

The definitions and theorems in this section are stated with respect to the total-order and NSHP support $\mathcal{S}$ $(o=(1,0))$.

Definition 7 *Let $\hat{y}(m,n)$ be the minimum-norm, causal, continuous support linear predictor of $y(m,n)$. Then the random field $\{u(m,n)=y(m,n)-\hat{y}(m,n)\}$ is called the* **innovation** *of the random field $\{y(m,n)\}$.*

Definition 8 *A random field $\{y(m,n)\}$ is* **regular** *if $E\left[y(m,n)-\hat{y}(m,n)\right]^2>0$ for at least one $(m,n)\in\mathcal{Z}^2$, i.e., its innovation field $\{u(m,n)\}$ does not vanish.*

Definition 9 *A random field $\{y(m,n)\}$ is* **deterministic** *if $E\left[(y(m,n)-\hat{y}(m,n))^2\right]=0$ for all $(m,n)\in\mathcal{Z}^2$, i.e., its innovation field $\{u(m,n)\}$ vanishes.*

Note that the deterministic field is a random field. It is deterministic only in the mean square sense.

Definition 10 *A regular field $\{y(m,n)\}$ is* **purely indeterministic** *if $\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n)}=\overset{o}{\mathcal{H}}{}^{u}{}_{(m,n)}$ for all $(m,n)\in\mathcal{Z}^2$, i.e., $\{u(m,n)\}$ spans the same Hilbert space spanned by $\{y(m,n)\}$.*

The following theorem is the basic theorem of the 2-D Wold-like decomposition of regular random fields. It is a generalization of Cramér's [20] 1-D Wold decomposition of non-stationary random processes by Francos *et al.* [31].

Theorem 2 *If $\{y(m,n)\}$ is a 2-D regular random field, then it can be represented uniquely by the following orthogonal decomposition:*

$$y(m,n)=v(m,n)+w(m,n), \tag{2.9}$$

where

$$w(m,n) = \sum_{(0,0)\preceq(k,l)} a_{(m,n)}(k,l)u(m-k,n-l) \tag{2.10}$$

and $E[v(m,n)] = E[u(m,n)] = 0$. *Field* $\{v(m,n)\}$ *is deterministic and field* $\{w(m,n)\}$ *is regular and purely indeterministic. The innovation field* $\{u(m,n)\}$ *is white, i.e.,* $E[u(m,n)u(s,t)] = 0$, *for all* $(m,n) \neq (s,t)$. *Fields* $\{v(m,n)\}$ *and* $\{u(s,t)\}$ *are orthogonal, i.e.,* $E[v(m,n)u(s,t)] = 0$, *for all* (m,n) *and* $(s,t) \in \mathcal{Z}^2$. *Thus fields* $\{v(m,n)\}$ *and* $\{w(s,t)\}$ *are also orthogonal. When* $E\left[u^2(m-k,n-l)\right] > 0$, *the coefficients* $a_{(m,n)}(k,l)$ *are given as*

$$a_{(m,n)}(k,l) = \frac{E\left[y(m,n)u(m-k,n-l)\right]}{E\left[u^2(m-k,n-l)\right]}. \tag{2.11}$$

When $E\left[u^2(m-k,n-l)\right] = 0$, $a_{(m,n)}(k,l)$ *are arbitrarily set to zero to accomplish the uniqueness of the sequence* $\{a_{(m,n)}(k,l)\}$.

It can be shown that field $\{u(m,n)\}$ is also the innovation field of $\{w(m,n)\}$ [31]. Since field $\{w(m,n)\}$ is purely indeterministic, $\overset{o}{\mathcal{H}}{}^{w}{}_{(m,n)} = \overset{o}{\mathcal{H}}{}^{u}{}_{(m,n)}$. Therefore, the purely indeterministic component $\{w(m,n)\}$ must exist for any regular field. If a random field $\{y(m,n)\}$ is regular and purely indeterministic, it can be represented completely by the white innovation driven moving average (MA) system

$$y(m,n) = \sum_{(0,0)\preceq(k,l)} a_{(m,n)}(k,l)u(m-k,n-l). \tag{2.12}$$

By Theorem 2, the subspace $\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n)}$ has a direct sum representation

$$\overset{o}{\mathcal{H}}{}^{y}{}_{(m,n)} = \overset{o}{\mathcal{H}}{}^{v}{}_{(m,n)} \oplus \overset{o}{\mathcal{H}}{}^{u}{}_{(m,n)}. \tag{2.13}$$

Theorem 2 gives the basic decomposition of a regular random field into its deterministic and purely indeterministic components. Shown next are some properties of the deterministic field $\{v(m,n)\}$.

Definition 11 *The* **remote past space** $\overset{o}{\mathcal{H}}{}^{y}{}_{(-\infty,-\infty)}$ *w.r.t. a specific total-order and NSHP definition is the intersection of all the Hilbert subspaces spanned by the samples of the regular field* $\{y(m,n)\}$, *i.e.,*

$$\overset{o}{\mathcal{H}}{}^{y}{}_{(-\infty,-\infty)} = \bigcap_{(m,n)\in\mathcal{Z}^2} \overset{o}{\mathcal{H}}{}^{y}{}_{(m,n)} \tag{2.14}$$

Let $\overset{o}{\mathcal{H}}{}^{v}{}_{(m,-\infty)} = \bigcap_{n=-\infty}^{\infty} \overset{o}{\mathcal{H}}{}^{v}{}_{(m,n)}$. Using Theorem 2, it can be shown that the deterministic field $\{v(m,n)\}$ has the property that $\overset{o}{\mathcal{H}}{}^{v}{}_{(m,n)} = \overset{o}{\mathcal{H}}{}^{v}{}_{(m,-\infty)}$ for all m [31]. Define $\overset{o}{\mathcal{H}}{}^{v}_{m} = \overline{Sp}\{v \mid v \in \overset{o}{\mathcal{H}}{}^{v}{}_{(m,-\infty)},\ v \perp \overset{o}{\mathcal{H}}{}^{v}{}_{(m-1,-\infty)}\}$. Then,

$$\overset{o}{\mathcal{H}}{}^{v}{}_{(m,-\infty)} = \overset{o}{\mathcal{H}}{}^{v}{}_{(m-1,-\infty)} \oplus \overset{o}{\mathcal{H}}{}^{v}_{m}.$$

By induction, the Hilbert space spanned by the deterministic field $\{v(m,n)\}$ can be written as the direct sum of the remote past space and the **row-to-row innovations** of the deterministic field w.r.t. the specific total-order and NSHP definition:

Theorem 3

$$\overset{o}{\mathcal{H}^v}_{(m,n)} = \overset{o}{\mathcal{H}^v}_{(m,-\infty)} = \overset{o}{\mathcal{H}^y}_{(-\infty,-\infty)} \oplus \bigoplus_{k=-\infty}^{m} \overset{o}{\mathcal{H}}{}^v_k. \tag{2.15}$$

Theorem 3 implies that the remote past space and the row-to-row innovations of the deterministic field $\{v(m,n)\}$ are mutually orthogonal.

Definition 12 *A 2-D deterministic random field $\{e(m,n)\}$ is* **evanescent** *w.r.t. a* specific *total-order and NSHP definition if it spans a Hilbert space that is identical to the one spanned by the row-to-row innovations of the deterministic random field at each coordinate (m,n).*

By Theorem 3, under each total-order and NSHP support $o \in \mathcal{O}$, at most one evanescent field can be resolved. This is the one that generates the row-to-row innovations aligned to the row orientation of o. The subspaces spanned by all other evanescent components $e_{o'}$, where $o' \in \mathcal{O}$ and $o' \neq o$, are contained in the remote past space $\overset{o}{\mathcal{H}^y}_{(-\infty,-\infty)}$. Therefore, to resolve all the evanescent components of the deterministic field, it is necessary to check the random field against every possible total-order definition in $\mathcal{O}$.

So far, the orthogonal decomposition of a 2-D regular random field has been studied without any assumption on the homogeneity of the field. In general, the regularity and determinism of a 2-D random field are total-order dependent. Therefore, the use of multiple total-order and RNSHP supports can result in a family of orthogonal decompositions of a 2-D random field. It will be shown in the next section that the decomposition of a regular and homogeneous random field is NSHP support invariant. The homogeneity of the random field also makes its spectral decomposition possible.

Summary:

By Theorem 2 and Theorem 3, a 2-D regular random field can be decomposed orthogonally into remote past, evanescent, and purely indeterministic random fields.

2.4 Decomposition of Homogeneous Random Fields

2.4.1 Homogeneous Random Fields

Definition 13 *A random field $\{y(m,n)\}$, $(m,n) \in \mathcal{Z}^2$, is* **homogeneous** *if*

$$E\left[y^2(m,n)\right] < \infty \tag{2.16}$$

and

$$r(k,l) = E\left[y(m+k,n+l)y(m,n)\right], \quad (k,l) \in \mathcal{Z}^2 \tag{2.17}$$

is independent of m and n.

If field $\{y(m,n)\}$ is homogeneous, its innovation field $\{u(m,n)\}$, as well as its deterministic field $\{v(m,n)\}$, its evanescent field $\{e(m,n)\}$, and its purely indeterministic field $\{w(m,n)\}$, are also homogeneous. Furthermore, the variance of field $\{u(m,n)\}$ is a constant for all $(m,n) \in \mathcal{Z}^2$. Denote this constant by σ^2. If field $\{y(m,n)\}$ is also regular, then σ^2 is strictly positive and the 2-D Wold decomposition of homogeneous regular random fields is unique by Theorem 2.

2.4.2 Spectral Decomposition

For a homogeneous random field, its spectral representation exists in the form of a Fourier-Stieltjes integral. In the following, all spectral functions are defined on the rectangular region

$$\mathcal{K} = \left[-\frac{1}{2}, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right].$$

Let $F_y(\xi,\eta)$ be the **spectral distribution function** of the homogeneous field $\{y(m,n)\}$. Then the corresponding **spectral density function** $f_y(\xi,\eta)$ is the 2-D Lebesgue derivative of $F_y(\xi,\eta)$:

$$f_y(\xi,\eta) = \frac{\partial^2 F_y(\xi,\eta)}{\partial\xi\,\partial\eta}. \tag{2.18}$$

The spectral representation of the field $\{y(m,n)\}$ is

$$y(m,n) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi j(m\xi+n\eta)} dZ(\xi,\eta), \tag{2.19}$$

where $j = \sqrt{-1}$, $Z(\xi,\eta)$ is a doubly orthogonal increment process,

$$E\left[dZ(\xi,\eta)dZ^*(\xi',\eta')\right] = 0, \qquad \xi \neq \xi', \eta \neq \eta', \tag{2.20}$$

and is related to $F_y(\xi,\eta)$ by

$$dF_y(\xi,\eta) = E\left[dZ(\xi,\eta)dZ^*(\xi,\eta)\right]. \tag{2.21}$$

The covariance function of $\{y(m,n)\}$ is

$$r_y(k,l) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi j(k\xi+l\eta)} dF_y(\xi,\eta). \tag{2.22}$$

Helson and Lowdenslager proved that a 2-D stationary (homogeneous) process can be orthogonally decomposed into three sub-processes: remote past, evanescent, and innovation [43]. The remote past and evanescent processes are deterministic and the innovation process is purely indeterministic. These are similar results as in Theorem 2, but for the homogeneous case. In the same paper, the decomposition of the spectral distribution function of a homogeneous regular random field is also given:

Theorem 4 *The spectral distribution function $F_y(\xi,\eta)$ of a homogeneous regular random field $\{y(m,n)\}$ can be uniquely represented as*

$$F_y(\xi,\eta) = F_v(\xi,\eta) + F_w(\xi,\eta), \tag{2.23}$$

where $F_v(\xi,\eta)$ and $F_w(\xi,\eta)$ are the spectral distribution functions of the deterministic and the purely indeterministic components of $\{y(m,n)\}$, respectively. Function $F_v(\xi,\eta) = F_y^s(\xi,\eta)$ is the singular part of $F_y(\xi,\eta)$ and function $F_w(\xi,\eta)$ is absolutely continuous. Thus, the spectral measure induced by $F_v(\xi,\eta)$ is singular w.r.t. the Lebesgue measure and is concentrated on a Borel set $\mathcal{L}$ with Lebesgue measure zero in $\mathcal{K}$. The derivative of $F_v(\xi,\eta)$ is zero except on the set $\mathcal{L}$. The spectral representations of the deterministic and the purely indeterministic component fields have the form

$$v(m,n) = \int_{\mathcal{L}} e^{2\pi j(m\xi+n\eta)} dZ(\xi,\eta). \tag{2.24}$$

and

$$w(m,n) = \int_{\mathcal{K}\setminus\mathcal{L}} e^{2\pi j(m\xi+n\eta)} dZ(\xi,\eta) \tag{2.25}$$

Therefore, the 2-D Wold-like decomposition of a homogeneous regular random field into deterministic and purely indeterministic components corresponds to: (1) in terms of spectral measures, the decomposition of the spectral measure of the random field into the sum of two mutually singular spectral measures (2.24) and (2.25) that are concentrated on the sets $\mathcal{L}$ and $\mathcal{K}\setminus\mathcal{L}$, respectively; (2) in terms of spectral distributions, the representation of $F_y(\xi,\eta)$ as the sum of its singular and absolutely continuous components (2.23). Clearly, the orthogonal decomposition of a homogeneous random field into deterministic and purely indeterministic components can be achieved by performing a spectral **Lebesgue decomposition** [82], which separates the singular and the absolutely continuous components of the spectral distribution of the random field.

2.4.3 Invariability

As mentioned at the end of last section, the set of multiple total-order and RNSHP support $\mathcal{O}$ gives rise to a corresponding family of orthogonal decompositions of a random field since the regularity and determinism of a non-homogeneous random fields are total-order dependent. In [42], Helson and Lowdenslager showed the following:

Theorem 5 *A 2-D homogeneous random field $\{y(m,n)\}$ is regular* **iff** *$f_y(\xi,\eta) > 0$ almost everywhere in $\mathcal{K}$ (in Lebesgue measure) and*

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \log f_y(\xi,\eta) d\xi d\eta > -\infty. \tag{2.26}$$

In the regular case, the variance of the innovation field $\{u(m,n)\}$ is given by

$$\sigma^2 = \exp \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \log f(\xi,\eta) d\xi d\eta. \tag{2.27}$$

The following theorem is based on Theorem 4, Theorem 5, and the fact that the results stated in these theorems are independent of the total-order and NSHP definition.

Theorem 6 *The regularity property of a homogeneous random field and the decomposition of a regular and homogeneous random field into deterministic and purely indeterministic components are NSHP support invariant. The resulting component fields from the decomposition are unique.*

Therefore, if a homogeneous random field is regular w.r.t. one total-order and NSHP definition, it is regular w.r.t. any other total-order and NSHP definition. The deterministic (purely indeterministic) component obtained w.r.t. one total-order and NSHP definition is identical to the deterministic (purely indeterministic) component obtained w.r.t. any other total-order and NSHP definition.

2.4.4 Deterministic Field

Further analysis of the deterministic field is pursued in this subsection. As shown previously, under the definition of the set of total-order and RNSHP support $\mathcal{O}$, multiple evanescent fields can be resolved from the deterministic random field.

Definition 14 *A random field $\{g(m,n)\}$ is called* **generalized evanescent** *if it can be represented as a linear (possibly infinite) combination of evanescent fields. Each of these evanescent fields generates row-to-row innovations w.r.t. a different total-order and RNSHP support.*

Definition 15 *A 2-D random field $\{p(m,n)\}$ is* **half-plane deterministic** *if it spans the Hilbert space $\mathcal{H}^y_{(-\infty,-\infty)} = \bigcap_{o\in\mathcal{O}} \overset{o}{\mathcal{H}}{}^y{}_{(-\infty,-\infty)}$, where $\overset{o}{\mathcal{H}}{}^y{}_{(-\infty,-\infty)}$ is the remote past space of the random field w.r.t. the total-order and NSHP definition o.*

Since the half-plane deterministic field spans the intersection of all Hilbert spaces spanned by the random field samples $y(m,n)$, $(m,n) \in \mathcal{Z}^2$, w.r.t. all total-orders and RNSHP definitions, it contains no innovations w.r.t. any total-order and RNSHP definitions.

From Theorem 3,

$$\overset{o}{\mathcal{H}}{}^v{}_{(\infty,\infty)} = \overset{o}{\mathcal{H}}{}^y{}_{(-\infty,-\infty)} \oplus \bigoplus_{k=-\infty}^{\infty} \overset{o}{\mathcal{H}}{}^v_k. \tag{2.28}$$

By the uniqueness and NSHP support invariance of the deterministic field and using (2.28),

$$\begin{aligned}
\mathcal{H}^v_{(\infty,\infty)} &= \overset{o}{\mathcal{H}}{}^v{}_{(\infty,\infty)} = \bigcap_{o\in\mathcal{O}} \overset{o}{\mathcal{H}}{}^v{}_{(\infty,\infty)} \\
&= \bigcap_{o\in\mathcal{O}} \left(\overset{o}{\mathcal{H}}{}^y{}_{(-\infty,-\infty)} \oplus \bigoplus_{k=-\infty}^{\infty} \overset{o}{\mathcal{H}}{}^v_k \right) \\
&= \mathcal{H}^y_{(-\infty,-\infty)} \oplus \bigoplus_{o\in\mathcal{O}} \bigoplus_{k=-\infty}^{\infty} \overset{o}{\mathcal{H}}{}^v_k.
\end{aligned} \tag{2.29}$$

This leads to the following theorem [28]:

Theorem 7 *The deterministic component $\{v(m,n)\}$ of a 2-D regular and homogeneous random field $\{y(m,n)\}$ can be decomposed uniquely as*

$$v(m,n) = p(m,n) + g(m,n), \tag{2.30}$$

where field $\{p(m,n)\}$ is half-plane deterministic and field $\{g(m,n)\}$ is generalized evanescent. Fields $\{p(m,n)\}$ and $\{g(m,n)\}$ are mutually orthogonal, i.e., $p(m,n) \perp g(s,t)$ for all $(m,n),(s,t) \in \mathcal{Z}^2$. Field $\{g(m,n)\}$ consists of a countable number of mutually orthogonal evanescent fields:

$$g(m,n) = \sum_{o \in \mathcal{O}} e_o(m,n), \tag{2.31}$$

where $e_o(m,n)$ is the evanescent field of $\{y(m,n)\}$ w.r.t. the total-order and RNSHP definition o.

The corresponding spectral decomposition of the deterministic random field is as follows [26].

Theorem 8 *Let $F_p(\xi,\eta)$ be the spectral distribution function of the half-plane deterministic component of a homogeneous regular random field $\{y(m,n)\}$ and $F_g(\xi,\eta)$ be the spectral distribution function of the generalized evanescent component of $\{y(m,n)\}$. The spectral distribution function $F_v(\xi,\eta)$ of the deterministic component of $\{y(m,n)\}$ can be uniquely represented as*

$$F_v(\xi,\eta) = F_p(\xi,\eta) + F_g(\xi,\eta) = F_p(\xi,\eta) + \sum_{o \in \mathcal{O}} F_{e_o}(\xi,\eta), \tag{2.32}$$

where $F_{e_o}(\xi,\eta)$ is the spectral distribution function of the evanescent field w.r.t. the total-order and RNSHP definition o. The spectral measures induced by the distribution functions in (2.32) are mutually singular.

From the definition of the evanescent field and Theorem 8, the spectral measure of the evanescent field w.r.t. the total-order and RNSHP definition o is a linear combination of spectral measures of the form

$$dF_{e_o}(\xi^o,\eta^o) = k(\xi^o)\, d\xi^o\, dF^s(\eta^o), \tag{2.33}$$

where $F^s(\eta^o)$ is a one-dimensional singular spectral distribution function and $k(\xi^o)$ is a one-dimensional spectral density function. Thus, the spectral distribution function of each evanescent field is absolutely continuous in one dimension and singular in the orthogonal dimension.

Summary:

By Theorem 2 and Theorem 7, a homogeneous regular random field $\{y(m,n)\}$ can be represented uniquely by the following orthogonal decomposition:

$$\begin{aligned} y(m,n) &= p(m,n) + g(m,n) + w(m,n) \\ &= p(m,n) + \sum_{o \in \mathcal{O}} e_o(m,n) + w(m,n). \end{aligned} \tag{2.34}$$

Correspondingly, the Hilbert space spanned by the homogeneous regular random variables $\{y(m,n)\}$, $(m,n) \in \mathcal{Z}^2$, can be decomposed into three mutually orthogonal subspaces: the subspace containing no innovations and spanned by the half-plane deterministic random field $\{p(m,n)\}$, the subspace representing the row-to-row innovations and spanned by the generalized evanescent random field $\{g(m,n)\}$, and the subspace containing the innovations of the 2-D field $\{y(m,n)\}$ and spanned by the purely indeterministic random field $\{w(m,n)\}$, which can be described by the white innovation driven MA term in (2.10).

By Theorem 4 and Theorem 8, the spectral distribution function of the homogeneous regular random field $\{y(m,n)\}$ can be uniquely represented as

$$\begin{aligned} F_y(\xi,\eta) &= F_p(\xi,\eta) + F_g(\xi,\eta) + F_w(\xi,\eta) \\ &= F_p(\xi,\eta) + \sum_{o\in\mathcal{O}} F_{e_o}(\xi,\eta) + F_w(\xi,\eta). \end{aligned} \tag{2.35}$$

The orthogonal decomposition of a homogeneous random field into deterministic and purely indeterministic components can be achieved by the spectral Lebesgue decomposition, which separates the singular and the absolutely continuous components of the spectral distribution of the field.

2.5 Approximations

Certain approximations to the 2-D Wold-like decomposition theory can be made for practical applications. (See [26] for more details.)

Since the spectral measure of the deterministic component of a homogeneous regular field is concentrated on a set with Lebesgue measure zero (Theorem 4), the derivative of the spectral distribution function of the deterministic component is zero almost everywhere in $\mathcal{K}$. In practice, the spectral density function of the deterministic component can be considered as being non-zero only on a countable set of points and curves in $\mathcal{K}$.

One frequently observed component of the half-plane deterministic field is the **harmonic random field** $\{h(m,n)\}$. In the "spectral density function", the harmonic field generates 2-D δ-functions supported by discrete points in $\mathcal{K}$. Hence, field $\{h(m,n)\}$ has the form of a countable sum:

$$h(m,n) = \sum_{p=1}^{P} \{A_p \cos 2\pi(m\xi_p + n\eta_p) + B_p \sin 2\pi(m\xi_p + n\eta_p)\}, \tag{2.36}$$

where (ξ_p, η_p) are the spatial frequencies of the p-th harmonic and the A_p's and B_p's are mutually orthogonal random variables with $E[A_p^2] = E[B_p^2] = \sigma_p^2$. The autocorrelation function of $\{h(m,n)\}$ is

$$r_h(m,n) = \sum_{p=1}^{P} \sigma_p^2 \cos 2\pi(m\xi_p + n\eta_p). \tag{2.37}$$

Similarly, the "spectral density function" of each evanescent component $f_{e_{(\alpha,\beta)}}(\xi,\eta)$ can be considered as containing 1-D δ-functions that are supported by lines of angle $\theta = tan^{-1}(\beta/\alpha)$ to the ξ axis in $\mathcal{K}$. Therefore, these "spectral density functions" are continuous along their support lines and singular in the orthogonal dimensions, *i.e.*,

$$f_{e_o}(\xi^o, \eta^o) = k(\xi^o) \sum_{i=1}^{I^o} \gamma_i^{o2} \left[\delta(\eta^o - \eta_i^o) + \delta(\eta^o + \eta_i^o)\right]. \tag{2.38}$$

Hence, the evanescent field $e_o(m^o, n^o)$ has the form

$$e_o(m^o, n^o) = s(m^o) \sum_{i=1}^{I^o} \left(C_i^o \cos 2\pi n^o \eta_i^o + D_i^o \sin 2\pi n^o \eta_i^o\right), \tag{2.39}$$

where $\{s(m^o)\}$ is a purely indeterministic 1-D process with spectral density $2k(\xi^o)$ and the C_i^o's and D_i^o's are mutually orthogonal random variables with $E[C_i^{o2}] = E[D_i^{o2}] = \gamma_i^{o2}$. The generalized evanescent field and its spectral density function have the form of a countable sum of representations in the form of (2.39) and (2.38) respectively.

As shown in (2.10), the purely indeterministic component of a regular random field has a white noise driven MA representation. In practice, one may want to explore the possibility of using other models. One such model is the 2-D autoregressive (AR) model:

$$w(m,n) = -\sum_{(0,0)\prec(k,l)} b(k,l)w(m-k, n-l) + u(m,n). \tag{2.40}$$

The validity of a MA to AR inversion can be determined by either examining the invertibility of the MA representation itself or testing whether the spectral density function $f_w(\xi, \eta)$ possesses certain properties. One sufficient condition for the existence of an AR representation is that $f_w(\xi, \eta)$ is strictly positive on and analytic in some neighborhood of the unit bicircle [24][94]. In practice, this condition is usually satisfied and an AR representation of the purely indeterministic field can be found [26].

In the following, the harmonic, the evanescent, and the purely indeterministic components of a random field are referred to as the **Wold components**. Shown in the next section, the spatial patterns of these Wold components appear to be visually repetitive, directional, and random.

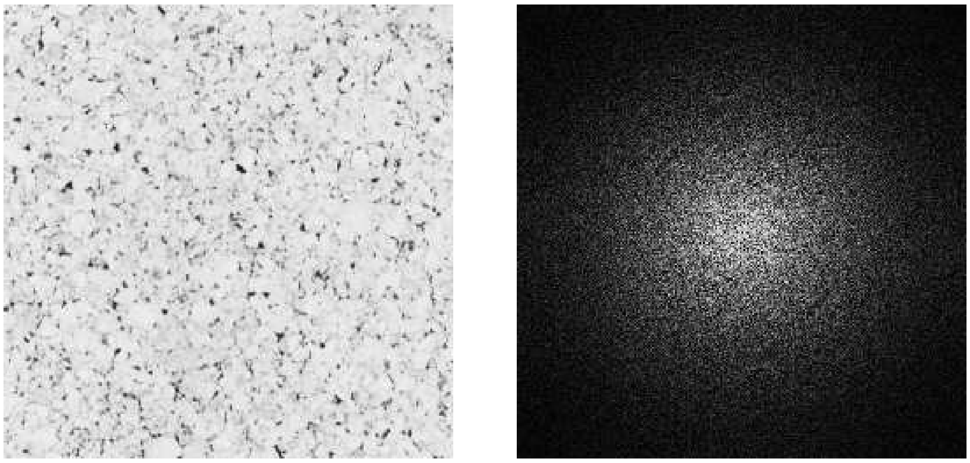
Summary:

A 2-D homogeneous regular random field can be represented as the sum of a harmonic component, a countable number of evanescent components, and a purely indeterministic component. These Wold components are mutually orthogonal.

The "spectral density function" of the harmonic (evanescent) component has the form of 2-D (1-D) δ-functions supported by points (lines) in the 2-D spectral domain. The spectral distribution function of the purely indeterministic component is absolutely continuous. In practice, an AR representation of the purely indeterministic component usually exists.

2.6 Examples of Wold Components

The Wold components of three natural textured images are shown here. Figure 2-3 shows the Brodatz texture Pressed Cork (D32). Since this pattern is mainly indeterministic, its Fourier magnitude image has no large peak. The sweater texture in Figure 2-4 has a prominent harmonic



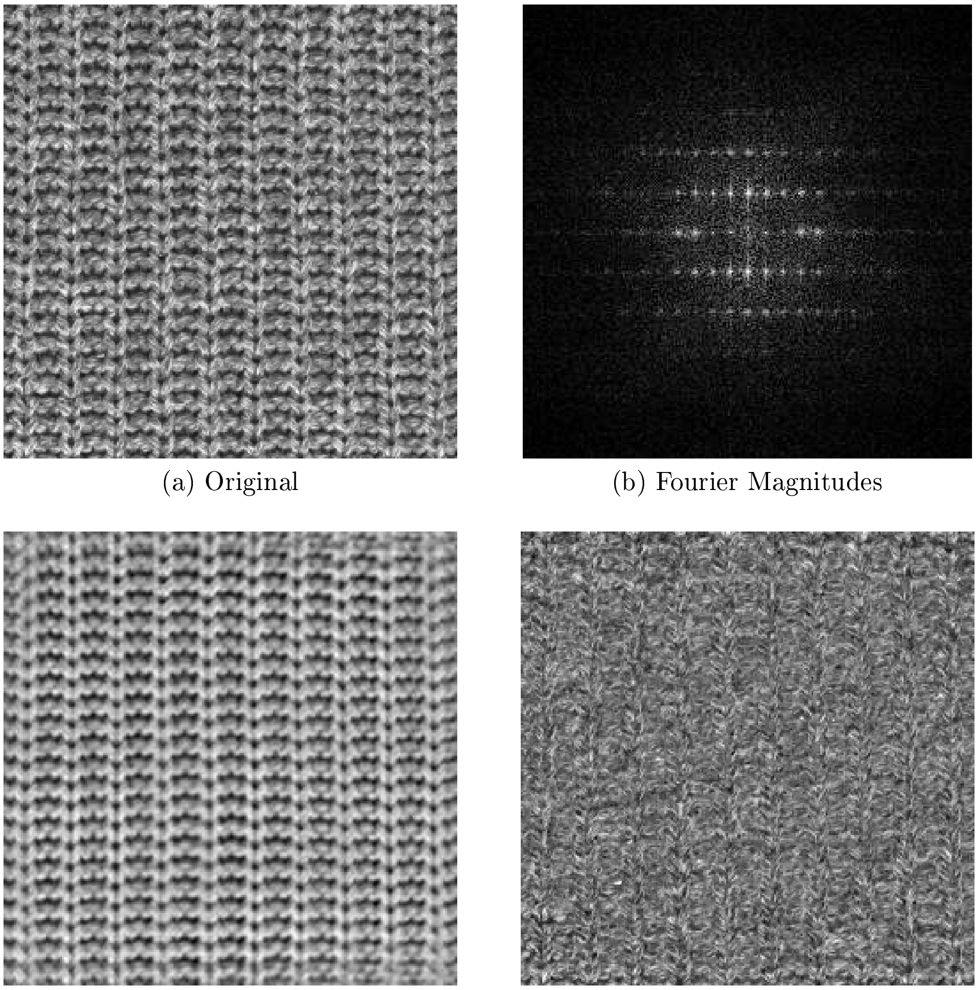
(a) Original (b) Fourier Magnitudes

Figure 2-3: Wold components of Brodatz texture D32: Pressed Cork. (a) Original. (b) Fourier magnitude image of (a). Since this pattern is mainly indeterministic, its Fourier magnitude image has no large peak.

---

component, which appears in its Fourier magnitude image as large peaks with point-like support. Figure 2-5 shows another Brodatz texture, Oriental Straw Cloth (D78). This pattern has strong evanescent components, which appear in its Fourier magnitude image as large values supported by line-like regions. Both Sweater and Oriental Straw Cloth patterns have certain amount of indeterministic component, which appears as the low value "smooth" background in their Fourier magnitude images. The extracted harmonic, evanescent, and indeterministic components of these two images are also shown.

Visually, the spatial images of the indeterministic components in the three examples are random looking, while the harmonic component in the Sweater pattern is very regular in both dimensions and the evanescent components in the Oriental Straw Cloth image appear to be directional.



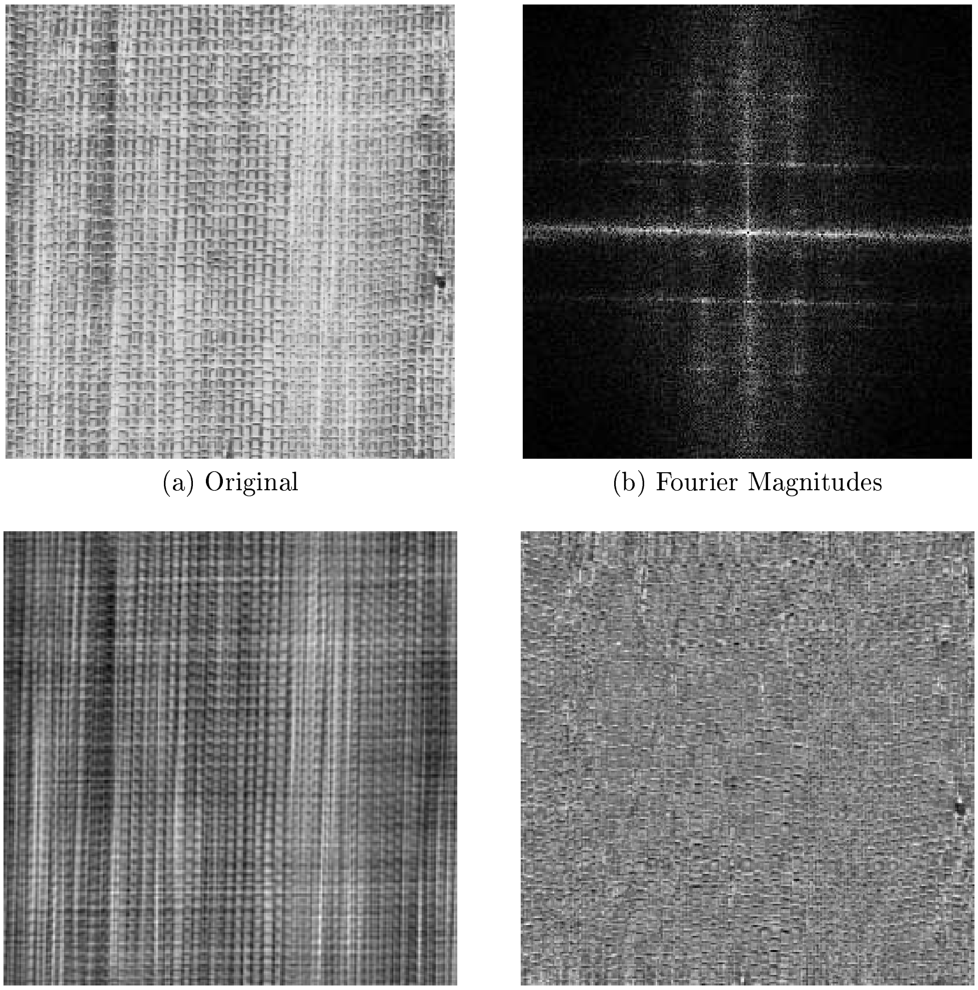
(a) Original

(b) Fourier Magnitudes

(c) Harmonic Component

(d) Indeterministic Component

Figure 2-4: Wold components of texture Sweater. (a) Original. (b) Fourier magnitude image of (a). (c) Harmonic component. (d) Indeterministic component. This pattern has a prominent harmonic component, which appears in its Fourier magnitude image as large peaks with point-like support.



(a) Original (b) Fourier Magnitudes

(c) Evanescent Component (d) Indeterministic Component

Figure 2-5: Wold components of Brodatz texture D78: Oriental Straw Cloth. (a) Original. (b) Fourier magnitude image of (a). (c) Evanescent component. (d) Indeterministic component. This pattern has a strong evanescent component, which appears in its Fourier magnitude image as large values supported by line-like regions.

Chapter 3

Spectral Decomposition

3.1 Introduction

The 2-D Wold-like decomposition theory presented in Chapter 2 provides the basis for two types of approach to the decomposition of a homogeneous random field. One approach is the direct parameter estimation from spatial data, and the other is the spectral decomposition based on the Lebesgue decomposition of the singular and continuous spectral components of the random field.

The effectiveness of the algorithms must be gauged in the context of their applications. For image coding, it is important that the estimated decomposition parameters provide enough information about the original image so that the reconstructed image resembles the original at least visually. For image similarity comparison, the discriminatory power of the extracted image features is more valuable than their reconstructive ability, and the form of the features should facilitate the distance computation in the feature space. In applications such as image database retrieval, automation and fast processing can be critical, especially when users introduce new images as query prototypes and the features need to be computed on the fly.

In the following, the existing 2-D Wold-based decomposition algorithms seen in the literature are first summarized and discussed. Then a new robust and computationally efficient spectral decomposition algorithm is presented.

3.2 Previous Work

Francos *et al.* proposed two 2-D Wold decomposition methods. One is a maximum-likelihood (ML) direct parameter estimation procedure and the other a periodogram thresholding scheme.

3.2.1 Direct Parameter Estimation

In [30][29], a conditional maximum-likelihood direct parameter estimation procedure was devised based on the assumption that the purely indeterministic component is a real-valued, Gaussian distributed, AR random field whose model is given by Equation (2.40) with $(k,l) \in \mathcal{S}_{M,N} \backslash \{(0,0)\}$.

The estimation problem is one of simultaneously estimating all the decomposition parameters from a finite number of samples taken from a single observation of the random field. Staring from the ML formulation, it was shown that the original ML problem can be transformed into a nonlinear problem of minimizing a new objective function over the deterministic component spectral support parameters, which include all the harmonic frequencies $\{\xi_p, \eta_p\}_{p=1}^{P}$, all the total orderings $(\alpha, \beta) \in \mathcal{O}$ that correspond to the evanescent components, and the frequencies $\{\eta_i^{(\alpha,\beta)}\}_{i=1}^{I^{(\alpha,\beta)}}$ for each (α, β). The amplitudes of the harmonic components and the AR parameters of the indeterministic components are then computed by solving a linear least-squares problem. The parameters of the modulating 1-D purely indeterministic processes of the evanescent components are estimated by solving a 1-D two-channel ARMA problem.

Difficult to solve analytically, the nonlinear minimization problem was dealt with numerically. To avoid an exhaustive search in the parameter space, a two-stage procedure was used. In the first stage, a parametric Fourier spectral estimation is conducted by fitting a high-order linear prediction model to the observed data and then computing the magnitude of the predictor transfer function inverse. Isolated peaks, as well as peaks that form continuous lines in the magnitude function, are identified as the candidates of the deterministic component spectral support parameters. In the second stage, a conjugate gradient procedure is used to refine the candidates. This procedure does not guarantee convergence to the global minimum unless the initial estimates are sufficiently close to the true values. It is reported in [30] that this iterative procedure can be computationally expensive, especially when the energy in the spectral peaks are not very high comparing to that in the neighboring Fourier frequencies. Unfortunately, this situation often arises in nature textures.

The effectiveness of the ML algorithm was demonstrated in [29] and [30] on two synthesized and six natural textures. The main advantage of this method is that it provides parametric descriptions of all Wold components in a random field. The main disadvantage is its computational cost. The entire procedure involves a high-order prediction model fitting, ARMA fittings, and gradient-based search.

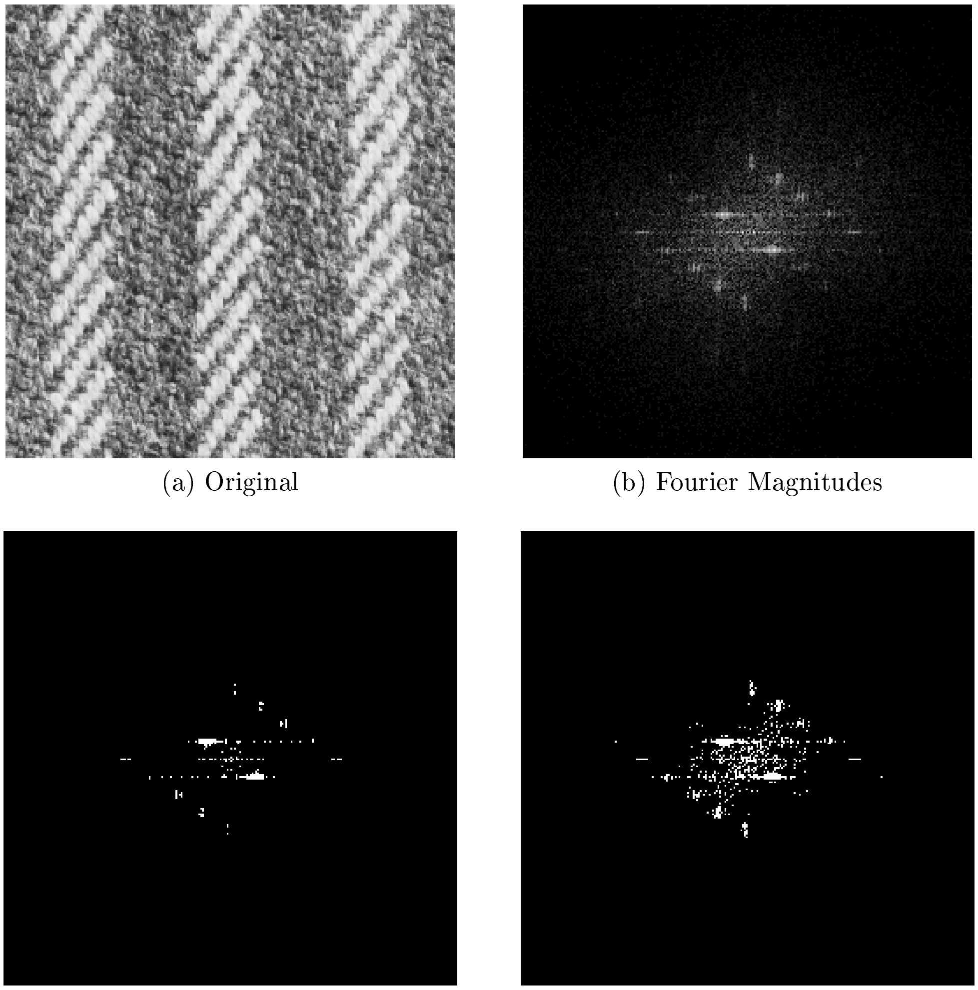
3.2.2 Periodogram Thresholding

The decomposition procedure proposed in [27] uses image periodogram thresholding to identify large Fourier spectral values as the harmonic and the evanescent components. The initial threshold is set to be the maximal value of the periodogram. Then the threshold is gradually lowered to qualify more frequency components as spectral peaks until "additional detected peaks are too wide to be considered as the contribution of harmonic components". The evanescent frequencies are determined by checking if the large spectral peaks are located in nearby frequencies along one dimension while fast decay of the periodogram values are observed in the orthogonal dimension.

After the removal of the deterministic component, the remaining purely indeterministic component is modeled by an AR model using a 2-D Levinson-type algorithm.

In the periodograms of natural texture images, the support region of each harmonic peak is usually not a point, but a small spread from the central frequency. There are essentially two issues in spectral Wold decomposition. One is to detect the spectral peaks; the other is to determine the peak support regions. The procedure in [27] resolves these two issues by using global thresholding of the periodogram values. However, there are cases in which this method fails. An example is shown in Figure 3-1. The pattern is Brodatz texture D11, Homespun Woolen Cloth, which has high frequency spectral peaks that are only locally large in value. Global spectral thresholding gives either poor segmentation of the peak support regions as in Figure 3-1 (c) or inaccurate peak identifications as in Figure 3-1 (d). In natural textures, this type of spectra abounds.

The main advantage of the periodogram thresholding approach is its computational simplicity. However, as shown above, this algorithm has serious limitations.



(a) Original

(b) Fourier Magnitudes

(c) Use High Threshold

(d) Use Low Threshold

Figure 3-1: Harmonic peak identification on Brodatz texture D11, Homespun Woolen Cloth, using global thresholding. (a) Original. (b) Fourier magnitude image of (a). (c) A high threshold results in poor segmentation of the support regions. Note that this threshold value is already not high enough since some low frequency random peaks are picked up. (d) A low threshold for better peak support gives inaccurate peak identification.

3.3 A New Spectral Decomposition Algorithm

3.3.1 Overview

The objective here is to develop a robust practical algorithm to decompose the deterministic and the indeterministic components of a homogeneous regular random field. The new algorithm takes a spectral decomposition approach, which is based on the principle of Lebesgue decomposition. The focus is to detect and extract the spectral singularities, which appear in Fourier spectra as peaks supported by point-like and line-like regions. As mentioned before, two issues are essential: one is to locate the singularities, and the other is to determine their support regions.

The algorithm for extracting the deterministic components consists of three main parts: the harmonic peak detection, the evanescent line detection, and the peak support segmentation. After the spectral support of the deterministic component is determined, the component frequencies are separated from the rest, which comprises the indeterministic component. The spatial values of the deterministic and the indeterministic components can be obtained by taking the inverse Fourier transform of the corresponding frequency components. As illustrated by the example in Figure 3-1, both spectral peak detection and peak support determination should be local, as opposed to global, operations.

Notations

In the following, the spatial and frequency samples are indexed by (m, n) and (k, l) respectively, where m and k are the row indices and n and l the column indices. Unless specified otherwise, the samples are defined on the 2-D region

$$\mathcal{D} = \{(i, j) \mid 0 \le i \le N-1,\ 0 \le j \le N-1\}. \tag{3.1}$$

At times, vector notation $\mathbf{f}$ is used to denote frequency index (k, l).

3.3.2 Spectral Estimation

The first step of a spectral approach is to compute the spectrum of a random field. There exist a large variety of spectral estimation methods. In general, the ones providing better spectral estimates in terms of frequency resolution and estimation bias and consistency are computationally more expensive. To facilitate applications such as image database retrieval, the signal periodogram is used in this algorithm for its computational efficiency.

The basic periodograms can be computed as the squared magnitudes of the signal discrete Fourier transform (DFT). Given an image $y(m, n)$, $(m, n) \in \mathcal{D}$, its DFT and inverse DFT are defined as

$$Y(k,l) = \begin{cases} \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} y(m,n)\, e^{-j\frac{2\pi}{N}km} e^{-j\frac{2\pi}{N}ln}, & (k,l) \in \mathcal{D} \\ 0, & otherwise \end{cases} \tag{3.2}$$

and

$$y(m,n) = \begin{cases} \dfrac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} Y(k,l)\, e^{j\frac{2\pi}{N}mk} e^{j\frac{2\pi}{N}nl}, & (m,n) \in \mathcal{D} \\ 0, & otherwise \end{cases} \tag{3.3}$$

The periodogram estimate of the image spectrum is

$$\hat{P}_y(k,l) = \frac{1}{N^2}|Y(k,l)|^2, \qquad (k,l) \in \mathcal{D}. \tag{3.4}$$

It is well known that basic periodograms are "noisy". The variance of $\hat{P}_y(k,l)$ is often on the order of the true spectra $P_y(k,l)$, independent of the image size [59]. Typically, some type of smoothing technique is used in the estimation, as in the Bartlett estimator and Welch estimator [45]. However, in the spectral Wold decomposition, an image is decomposed via the inverse Fourier transform of its decomposed spectra. Hence, the spectral decomposition has to be based on the actual DFT values of the original image, so smoothing techniques can not be liberally applied. The noisiness of the periodogram poses a serious challenge to the decomposition algorithm.

Another issue related to the periodogram estimation is the image boundaries. Proper handling of the boundaries is especially important when the image has irregular boundaries or is not quite homogeneous. Circular Gaussian tapering is used in this algorithm when necessary. The Gaussian tapering window is

$$g_t(m,n) = e^{-\frac{(m-N/2)^2+(n-N/2)^2}{2\sigma^2}}, \qquad (m,n) \in \mathcal{D}, \tag{3.5}$$

where the standard deviation σ is 0.375, normalized by the image half-size $N/2$. Since a spatial multiplicative tapering corresponds to a spectral circular convolution, the periodogram of the tapered image is a low-pass filtered version of the one without tapering. The smoothing effect of Gaussian tapering is illustrated in Figure 3-2.

3.3.3 Spectral Harmonic Peak Detection

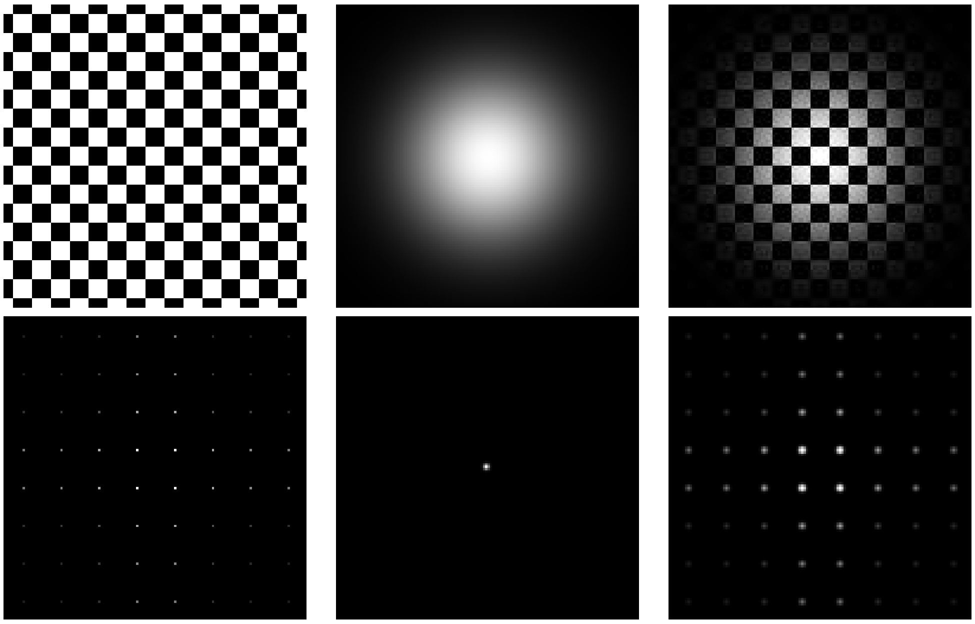
The harmonic peak detection is conducted on the 2-D Fourier magnitude image, which is the scaled square root of the periodogram. The basic idea is to first find the local maxima of the Fourier magnitudes. These local maxima provide the candidate locations of harmonic peaks. A local maximum qualifies to be a harmonic peak only when its frequency is either a fundamental or a harmonic.

Frequency Half-plane

Given an image of size $N \times N$, its DFT magnitude image has the same size, with the zero frequency (DC) at sample $(0,0)$. Translate the magnitude image by $(N/2, N/2)$ and wrap the image around at the edges so that the DC is at $\mathbf{f}_o = (N/2, N/2)$ of the frequency plane. (Note that the translation and wrapping are equivalent to swapping the quadrants.) When the image is real, the translated magnitude image is symmetric to $\mathbf{f}_o$. Define the **frequency half-plane** as

$$\mathcal{D}_h = \left\{(k,l) \mid 0 \le k \le \frac{N}{2} - 1,\ 0 \le l \le N-1\right\} \cup \left\{(k,l) \mid k = \frac{N}{2},\ 0 \le l \le \frac{N}{2}\right\}. \tag{3.6}$$

For a symmetric magnitude image, its frequency half-plane contains no symmetric frequency components. In the following discussion, all 2-D Fourier magnitudes and spectra are translated such that the DC components are at $\mathbf{f}_o$. Also, unless specified otherwise, only the frequency half-plane is under consideration.



(a) Original (b) Gaussian Window (c) Tapered

Figure 3-2: Smoothing effect of multiplicative Gaussian tapering. Top row: spatial images. Bottom row: corresponding Fourier magnitude images. (a) Original checkerboard pattern. (b) Gaussian tapering window of $\sigma = 0.375$ (σ is normalized by the image half-size $N/2$). (c) Tapered checkerboard pattern: the top image is the multiplication of the other two top images and the bottom image is the circular convolution of the other two bottom images.

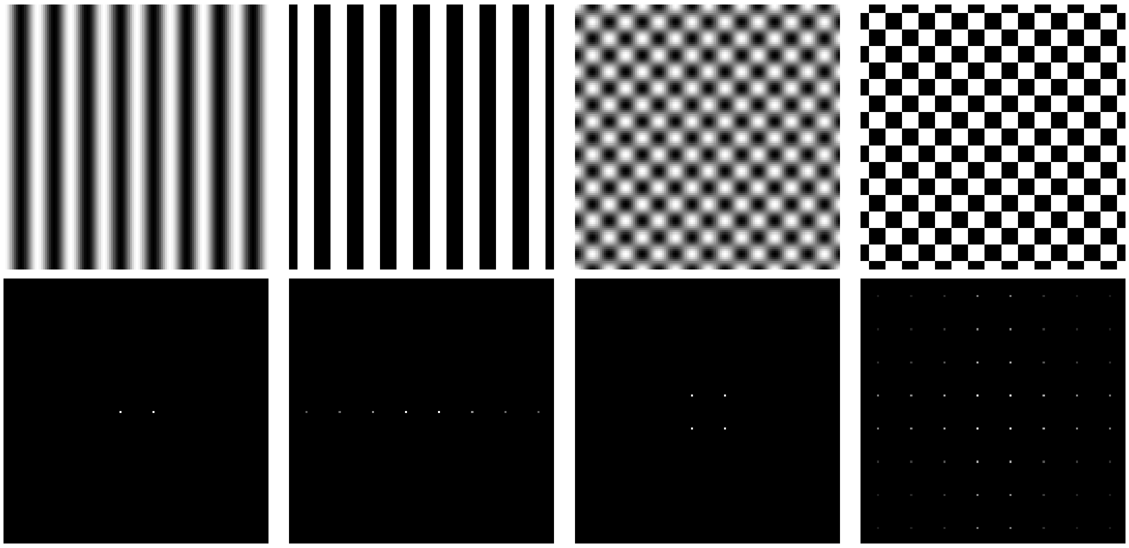
Local Maxima

Although harmonic peaks usually correspond to large magnitude values, as shown previously, some of them may be large only locally. For this reason, local maxima of the Fourier magnitudes are first detected as candidates of harmonic peaks.

Since the spacing between the harmonic peaks can be small, the estimation window for local maxima detection should not be large. However, the indeterministic "background" of the Fourier magnitude image is usually "lumpy". Under these conditions, many detected local maxima do not correspond to any harmonic peaks and are located rather randomly. Some intrinsic properties of the harmonic random field should be used to discriminate the local maxima for the true harmonic peaks.

Fundamental-harmonic Relationship

One important property of the harmonic random field is the fundamental-harmonic relationship among its spectral peaks. This is illustrated in Figure 3-3. The top row of the figure contains the



(a) (b) (c) (d)

Figure 3-3: Fundamental-harmonic relationship among harmonic peak frequencies. Top row: spatial images. Bottom row: corresponding Fourier magnitude images. (a) 1-D sine grating containing only one frequency component (considering half of the frequency plane). (b) Black and white stripes have one fundamental frequency, which is the same as the single frequency in (a), and three harmonic frequencies. (c) 2-D sine grating with two frequencies. (d) Checkerboard pattern, whose fundamentals are the same as the frequencies in (c), has a series of harmonics associated with its fundamentals.

---

spatial images and the bottom row the corresponding Fourier magnitudes. The 1-D sine grating pattern in Figure 3-3 (a) contains only one frequency component in $\mathcal{D}_h$. The black and white stripe pattern in (b) have one fundamental frequency, which is the same as the single frequency in (a), and three harmonic frequencies. Similar phenomenon can be observed in (c) and (d). While the 2-D sine grating pattern in (c) has only two frequency components, the checkerboard pattern, whose fundamentals are the same as the frequencies in (c), has a series of harmonics associated with its fundamentals.

Since the edges in a natural pattern usually do not have strictly sinusoidal profiles, one can expect to find harmonics associated with the fundamentals in the spectra of harmonic fields. Therefore, the fundamental-harmonic relationship can be used to identify the true harmonic peaks among the detected local maxima of the Fourier magnitude images.

For harmonic peak detection, the fundamental and the harmonic frequencies are defined as follows.

Fundamental

A Fundamental is a frequency that can be used to linearly express the frequencies of some other local maxima of the Fourier magnitudes.

Harmonic

A harmonic is a frequency that can be represented as a linear combination of some fundamentals.

Hence, a harmonic frequency $\mathbf{f}^h \in \mathcal{D}_h$ can be expressed as

$$\mathbf{f}^h = \mathbf{f}_o + \sum_{i=1}^{I_h} \alpha_i \left(\mathbf{f}_i^f - \mathbf{f}_o \right), \qquad \alpha_i \in \mathcal{Z}, \tag{3.7}$$

where $\mathbf{f}_i^f \in \mathcal{D}_h$ is the i-th contributing fundamental of the harmonic $\mathbf{f}^h$.

Effect of Frequency Sampling

The discrete Fourier spectrum of an image can be regarded as the sampling of the image continuous spectrum. In general, the sampling points do not fall right on to the very peaks of the continuous spectrum. Therefore, in the DFT plane, the discrete harmonics are in general not the exact multiples of their discrete fundamentals. Due to this sampling effect, certain amount of tolerance should be considered when examining the fundamental-harmonic relations among the local maxima. Furthermore, the fundamental frequency values should be refined to subsample precision since the accuracy of the fundamental frequencies is critical, especially when the harmonics are at the high multiples of the fundamentals. In the algorithm, the fundamental frequencies are refined by using the frequencies of their detected harmonics.

Implementation

To detect the harmonic peaks, the image is first zero-meaned and Gaussian tapered, and then its DFT magnitudes are computed. The local maxima of the magnitudes are found by searching a small neighborhood of each frequency sample, typically a 5×5 region. To save computation, frequencies whose magnitude values are below 5% of the entire magnitude range are not considered.

Next, the fundamental-harmonic relationship among the local maxima is examined. Starting from the lowest frequency to the highest, each local maxima is checked first for its harmonicity — if its frequency can be expressed as a linear combination of the existing fundamentals, and then for its fundamentality — if the multiples of its frequency, combined with the multiples of existing fundamentals, coincide with the frequency of another local maximum. To reduce the frequency sampling effect, a tolerance of two sample points in both row and column directions is used in the frequency matching.

When a new fundamental $\mathbf{f}^f$ is found, the algorithm detects the harmonics for which the new fundamental $\mathbf{f}^f$ is *solely* responsible and refines the fundamental frequency value after each harmonic is found. Denote the jth harmonic found as

$$\mathbf{f}_j^h = \beta_j \mathbf{f}_{j-1}^f + \Delta \mathbf{f}_j, \tag{3.8}$$

where $\mathbf{f}_{j-1}^f$ is the fundamental frequency value after the $(j-1)$-th refinement, $|\Delta \mathbf{f}_j| < |\mathbf{f}_{j-1}^f|$ and β_j

is a positive integer. Shown in Appendix A.1, the j-th fundamental refinement using $\mathbf{f}_j^h$ is

$$\mathbf{f}_j^f = \frac{\sum_{i=0}^{j} \mathbf{f}_i^h}{\sum_{i=0}^{j} \beta_i} = \mathbf{f}_{j-1}^f + \frac{\Delta \mathbf{f}_j}{\sum_{i=0}^{j} \beta_i}, \tag{3.9}$$

where $\beta_0 = 1$ and $\mathbf{f}_0^h = \mathbf{f}^f$.

All the local maxima whose frequencies are either fundamental or harmonic are qualified as the harmonic peaks.

Examples

Figure 3-4 shows two examples of harmonic peak detection. Shown in each example are the original image, the Fourier magnitude image, and the locations of detected harmonic peaks.

3.3.4 Spectral Evanescent Line Detection

The spectral evanescent line detection utilizes the Hough transformation [40] of the Fourier magnitude image. Consider a line in a conventional Cartesian X-Y coordinate system. Denote the line normal angle by ϕ and the line normal length by d. Then the line equation is

$$d = x \cos \phi + y \sin \phi. \tag{3.10}$$

A 2-D histogram is built for the normal angle ϕ and normal length d. The histogram bin size is half of a degree for ϕ and unity for d. For each pixel (x, y) of the magnitude image, the pixel value is added to every histogram cell whose index (ϕ, d) satisfies (3.10). Large local maxima of the histogram correspond to the prominent lines in the image while the bin indices of the maxima provide the line parameters. Note that for accurate evanescent line detection, the large Fourier magnitude values associated with the harmonic peak frequencies should be removed first since these frequencies leave strong signatures in the Hough transform histogram.

An example of the spectral line detection is shown in Figure 3-5. The original is the Brodatz texture D64: Handwoven Oriental Rattan. The Fourier magnitude image in Figure 3-5 (b) has six lines. Image (c) is the Hough transform of (b). Six large local maxima are detected in (c). Lines corresponding to the local maxima are shown in (d).

3.3.5 Determining Peak Support Regions

The support regions of the harmonic peaks and evanescent lines are determined by an iterative algorithm. At the beginning of each iteration, a 2-D Gaussian surface is fitted to the Fourier magnitude image, from which all the identified peak support frequencies are removed, to coarsely model the indeterministic component. Based on the local standard deviation (square root of the local variance) of the fitting residual, new support frequencies are identified. This process terminates automatically when the averaged local standard deviation at the near neighbor of the estimated support regions is comparable to that in the ambient areas.

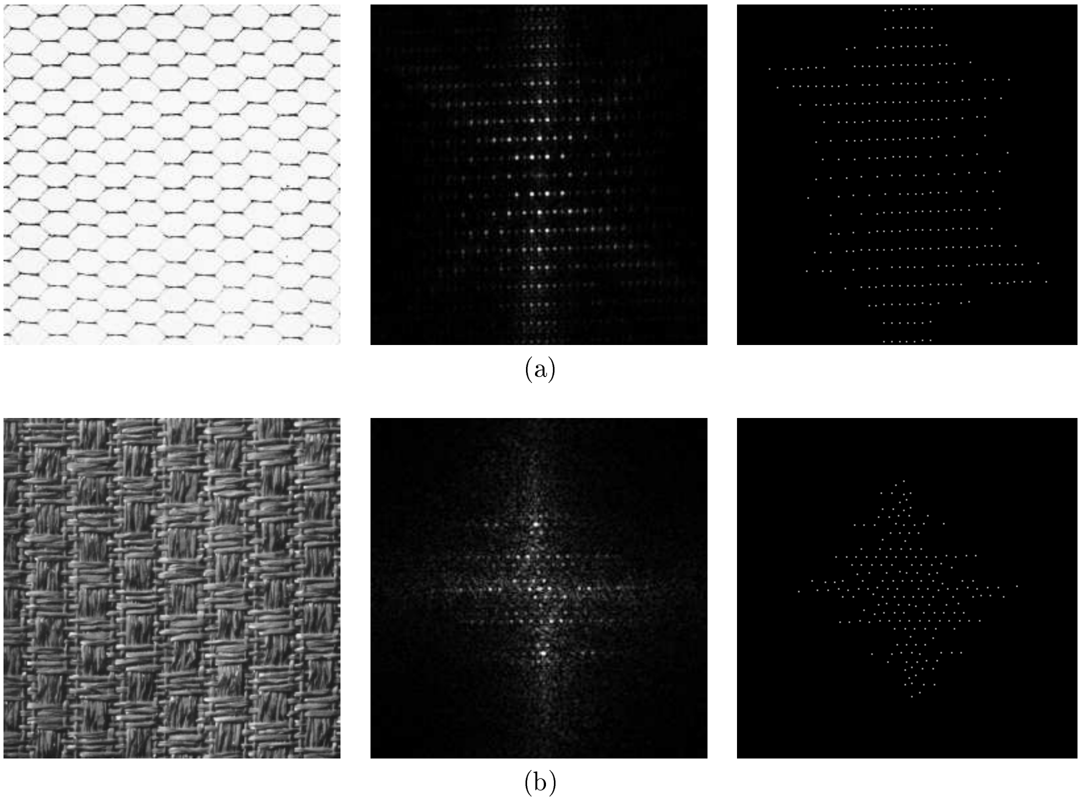


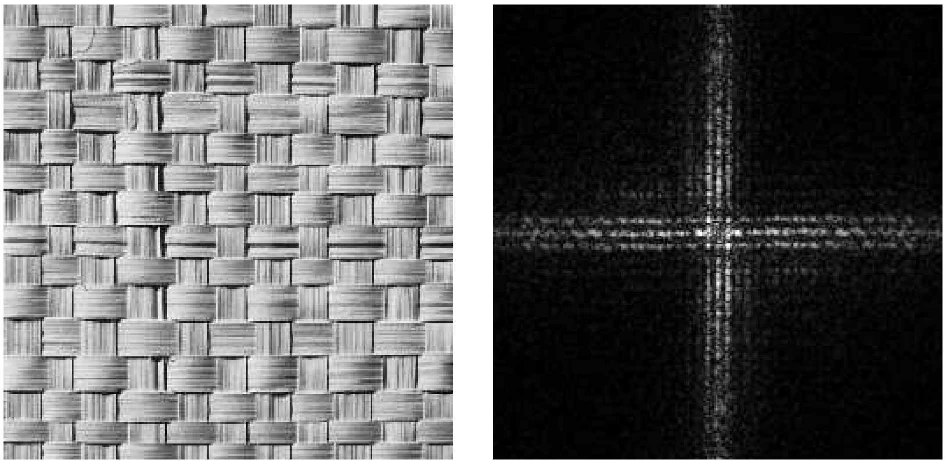
Figure 3-4: Examples of spectral harmonic peak detection. Left column: original image. Middle column: Fourier magnitudes. Right column: detected harmonic peak locations. (a) Brodatz texture D34: Netting. (b) Brodatz texture D82: Oriental Straw Cloth.

Initializing Mask

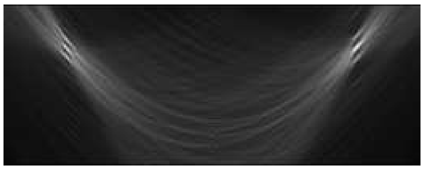
Before starting the iterative procedure, a mask image, which has the size of the Fourier magnitude image, is created and initialized to zero. The mask records the support regions estimated in each iteration. The frequency locations of the estimated harmonic peaks and evanescent lines are marked in the mask as the initial support regions. If a frequency on an evanescent line has a magnitude value less than 5% of the entire magnitude range, that frequency is not marked.

Gaussian Fitting

A 2-D Gaussian surface is used to coarsely model the indeterministic component in the Fourier magnitude image. The indeterministic component corresponds to the relatively smooth "background" of the Fourier spectra. The purpose of the Gaussian fitting is not to parameterize the indeterministic component, but to facilitate the determination of the harmonic peak and evanescent line support regions. This will become clear in the discussion that follows.



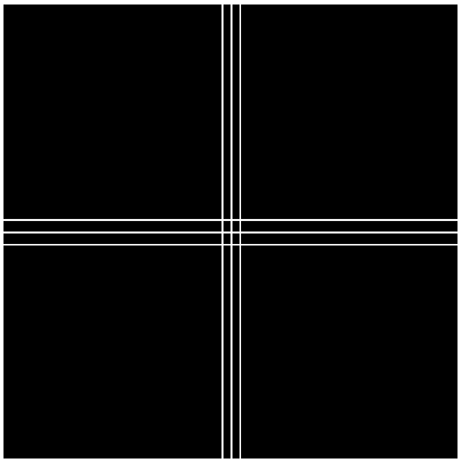
(a) Original (b) Fourier Magnitudes



(c) Hough Transform

d	ϕ	Hough Value
127.0	90.0	9220.0
128.0	0.0	7673.0
123.0	0.0	6648.0
133.0	0.0	6648.0
134.0	90.0	5951.0
120.0	90.0	5951.0

(d) Hough Local Maxima



(e) Detected Lines

Figure 3-5: Example of spectral evanescent line detection. (a) Brodatz texture D64: Handwoven Oriental Rattan. (b) Fourier magnitudes of (a). (c) Hough transform of the magnitude image. The abscissa is the normal angle ϕ, and the ordinate the normal length d. (d) List of large local maxima in (c). (e) Lines corresponding to the local maxima.

The equation of the Gaussian surface is

$$M_s\, g_s(\mathbf{f}) = M_s\, e^{-\frac{1}{2}(\mathbf{f}-\bar{\mathbf{f}})^T \boldsymbol{\Sigma}_{\mathbf{f}}^{-1}(\mathbf{f}-\bar{\mathbf{f}})}, \tag{3.11}$$

where T denotes the transpose and $\mathbf{f} \in \mathcal{D}$. Vector $\bar{\mathbf{f}}$, matrix $\boldsymbol{\Sigma}_{\mathbf{f}}$, and scaler M_s are the parameters to be estimated from the Fourier magnitude data. The least-squares estimation of $\bar{\mathbf{f}}$ and $\boldsymbol{\Sigma}_{\mathbf{f}}$ is a difficult nonlinear problem. An alternative is to regard the magnitude data as a histogram of Gaussian-distributed random frequency samples and estimate the parameters $\bar{\mathbf{f}}$ and $\boldsymbol{\Sigma}_{\mathbf{f}}$ by using a maximum likelihood estimator. The bin size of the magnitude "histogram" in this case is unity in both dimensions. Denote the integer part of the Fourier magnitude value $|Y(k,l)|$ by $N_{(k,l)}$. The histogram interpretation of the magnitude image is that there are $N_{(k,l)}$ random samples observed at frequency (k,l). The maximum likelihood estimate of the vector $\bar{\mathbf{f}}$ and matrix $\boldsymbol{\Sigma}_{\mathbf{f}}$ in (3.11) are (see Appendix A.2 for derivations)

$$\bar{\mathbf{f}} = (\bar{k}, \bar{l}), \qquad \boldsymbol{\Sigma}_{\mathbf{f}} = \begin{bmatrix} \sigma_{kk}^2 & \sigma_{kl}^2 \\ \sigma_{lk}^2 & \sigma_{ll}^2 \end{bmatrix}, \tag{3.12}$$

where

$$\bar{k} = \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} k\, N_{(k,l)} \tag{3.13}$$

$$\bar{l} = \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} l\, N_{(k,l)} \tag{3.14}$$

$$\sigma_{kk}^2 = \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} k^2\, N_{(k,l)} - \bar{k}^2 \tag{3.15}$$

$$\sigma_{kl}^2 = \sigma_{lk}^2 = \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} kl\, N_{(k,l)} - \bar{k}\,\bar{l} \tag{3.16}$$

$$\sigma_{ll}^2 = \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} l^2\, N_{(k,l)} - \bar{l}^2 \tag{3.17}$$

$$N_t = \sum_{(k,l)\in\mathcal{D}} N_{(k,l)}. \tag{3.18}$$

Since the Fourier magnitude image is symmetric with respect to $\mathbf{f}_o$, it is expected that $\bar{\mathbf{f}} \approx \mathbf{f}_o$.

After the parameters $\bar{\mathbf{f}}$ and $\boldsymbol{\Sigma}_{\mathbf{f}}$ are estimated, the magnitude M_s of the Gaussian surface can be obtained by solving a least-squares problem (see Appendix A.3 for derivations):

$$M_s = \arg\min_{M_s} \sum_{(k,l)\in\mathcal{D}} [\, |Y(k,l)| - M_s\, g_s(k,l)\,]^2 \tag{3.19}$$

$$= \frac{\sum_{(k,l)\in\mathcal{D}} g_s(k,l)\, |Y(k,l)|}{\sum_{(k,l)\in\mathcal{D}} g_s^2(k,l)} \tag{3.20}$$

Growing Support Regions

The support regions are grown in each iteration from the existing support areas that are marked in the mask. The Gaussian fitting residual, which is the difference between the magnitude image and its fitted Gaussian surface, is first computed. Then the local standard deviations of the residual image are estimated at each frequency using typically a 5×5 estimation window. The candidates of new support frequencies are found by detecting large positive residual values that are at least as large as the local standard deviation. These "outliers" become part of the peak support if they are adjacent to the existing support frequencies. The new support frequencies are then recorded in the mask and the values of these frequencies in the Fourier magnitude image are replaced by the corresponding values of the Gaussian surface.

Terminating Iterations

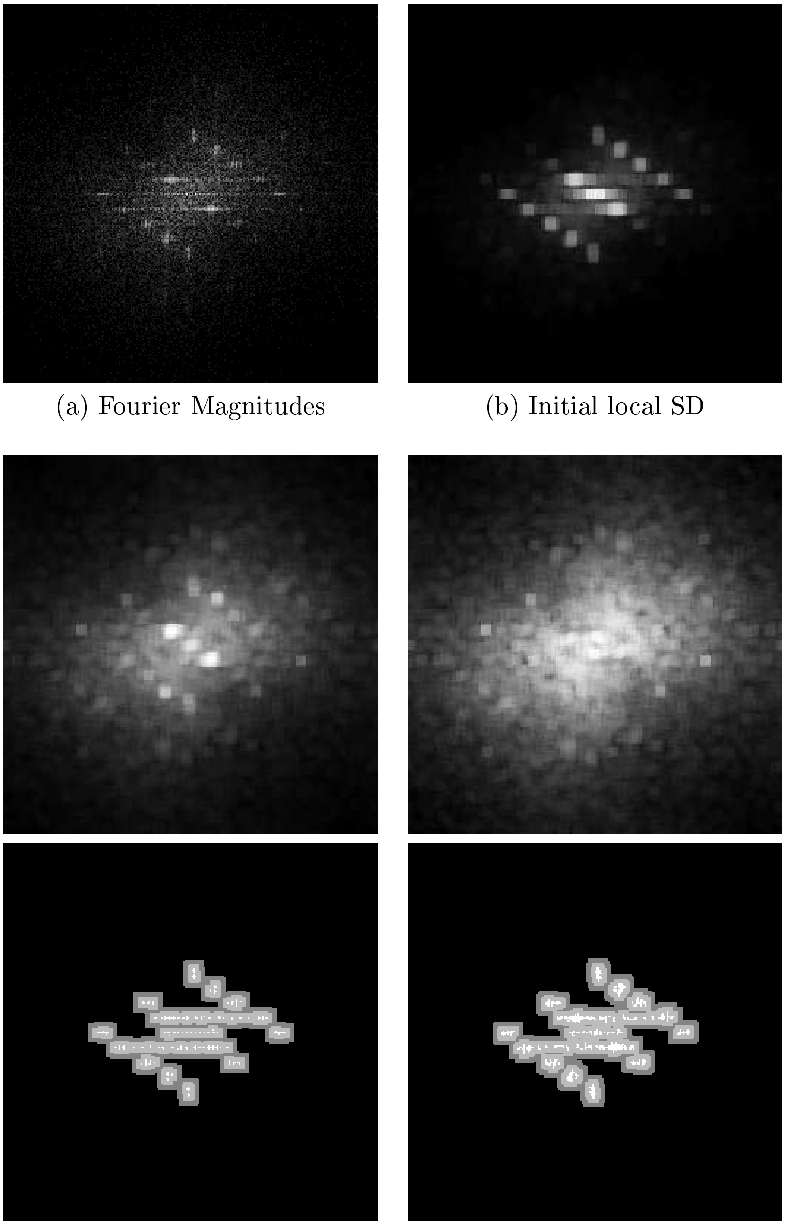
The iterative process should terminate when the large spectral peaks are removed and the remaining Fourier magnitude image becomes "smooth". However, this smoothness is only to the global scale; the Fourier magnitude image is generally very noisy. Hence, any applicable smoothness measure has to be statistical.

The local standard deviation (SD) is used in the algorithm to construct a smoothness measure. At the end of each iteration, the local SD of the Fourier magnitude image is computed. An example is given in Figure 3-6 to show how the local SD of the Fourier magnitude image becomes smoother over the iterations (image (b) and the middle row). At the beginning, the local SD values are large in areas near the spectral peaks. When more frequencies are extracted as peak support, the local SD values adjacent to the peak support regions become similar to those further away from the support regions. In the figure, the local SD images are individually scaled to the display range $[0, 255]$.

The Fourier magnitude smoothness measure used in the decomposition algorithm is constructed as follows. First, two kinds of regions surrounding the spectral peaks are found. One is the **adjacent** area, which includes all frequencies within distance d_{adj} from the estimated peak support frequencies. The other is the **ambient** area, which includes all frequencies within distance d_{amb} from the frequencies in the adjacent area. Examples of the adjacent and the ambient areas are shown in the bottom row of Figure 3-6, with $d_{adj} = d_{amb} = 3$. The **smoothness measure** is defined as the ratio between the averaged local SD in the adjacent areas, SD_{adj}, and the averaged local SD in the ambient areas, SD_{abm}. Figure 3-7 shows the local SD ratio SD_{adj}/SD_{abm} of 13 Brodatz textures in the first 7 iterations. For all 13 images, the local SD ratio tapers off after first few iterations. In the algorithm, the iterative processing terminates when the change of local SD ratio between iterations drops to below value 0.1.

3.3.6 Decomposition

The decomposition of a homogeneous random field is based on the decomposition of its spectral Wold components. When the peak support determination program is terminated, the mask contains the spectral frequencies of the deterministic component. Denote these frequencies as $\mathcal{D}_v$. The DFT



(a) Fourier Magnitudes (b) Initial local SD

(c) Iteration 1 (d) Iteration 4

Figure 3-6: Examples of Fourier magnitude local SD (5×5 estimation window) and the adjacent and the ambient areas (3 pixel width) for local SD ratio computation. (a) Fourier magnitudes of Brodatz texture D11 (the original is shown in Figure 3-1 (a)). (b) Initial magnitude local SD. (c)-(d) Shown for iteration 1 and 4: top image: magnitude local SD; bottom image: estimated peak support (white) and corresponding adjacent (light gray) and ambient areas (dark gray). The local SD images are individually scaled to the display range $[0, 255]$.

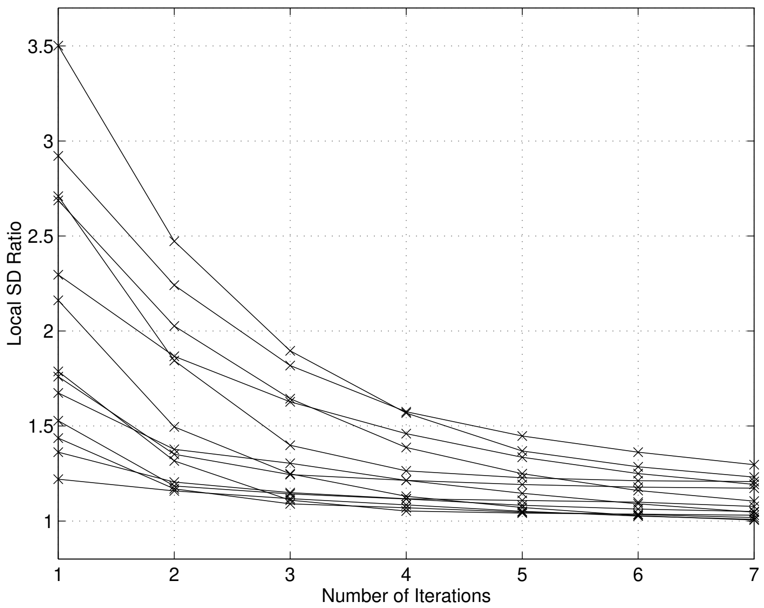


Figure 3-7: Fourier magnitude local SD ratio SD_{adj}/SD_{abm} of 13 Brodatz textures in the first 7 iterations. This ratio is used to measure the global smoothness of the Fourier magnitude images.

---

of the random field is then decomposed into the deterministic component

$$V(k,l) = \begin{cases} Y(k,l), & (k,l) \in \mathcal{D}_v \\ 0, & otherwise \end{cases} \tag{3.21}$$

and the indeterministic component

$$W(k,l) = \begin{cases} Y(k,l), & (k,l) \in \mathcal{D},\ (k,l) \notin \mathcal{D}_v \\ 0, & otherwise \end{cases} \tag{3.22}$$

The deterministic field $v(m,n)$ and the indeterministic field $w(m,n)$ are obtained by computing the inverse DFT of $V(k,l)$ and $W(k,l)$, respectively.

Note that the mask can also record whether an extracted peak frequency belongs to a harmonic peak or an evanescent line. Consequently, the deterministic field can be decomposed further into the harmonic and the generalized evanescent components by using a spectral decomposition procedure similar to the one presented above.

Iteration	D11	Sweater	D78
1	1.410867	1.579968	1.331581
2	1.183042	1.308202	1.215476
3	1.140345	1.199033	1.172281
4	1.121880	1.147715	1.145736

Table 3.1: Fourier magnitude local SD ratio of textures D11, Sweater, and D78 at the end of processing iteration 1 to 4.

3.3.7 Examples

Three decomposition examples are shown in Figures 3-8, 3-9, and 3-10. The images in the top rows are the original, the Fourier magnitudes of the original, and the detected harmonic peak central frequencies or evanescent lines. The Fourier magnitudes are computed after the original images are zero-meaned. Shown in row 2 to row 5 of each figure are the mask images, the Fourier magnitude images, the harmonic components, and the indeterministic components obtained at the end of iterations 1 to 4. The Fourier magnitude images are individually scaled to the display range $[0, 255]$.

The Fourier magnitude local SD ratio at the end of each iteration are shown in Table 3.1. For the three examples, the automatic termination of the iterative peak support estimation occurs after iteration 3, 4, and 3, respectively.

3.4 Discussion

Spectral peak support estimation is an important issue for a spectral decomposition algorithm since in practice the spectral singularities seldom appear as pure 1-D or 2-D impulses in estimated spectra. The algorithm presented here uses a non-parametric peak support estimation method, which is based on the local variance of image Fourier magnitudes. The main advantages of this algorithm are its robustness and computational efficiency. A parametric approach, such as fitting 2-D Gaussian models to spectral evanescent and harmonic peaks, can also be considered for robust peak support estimation. For fast computation of model parameter estimation, an efficient fitting algorithm, such as the one presented in Section 3.3.5, can be used.

3.5 Summary

In this chapter, a spectral 2-D Wold decomposition algorithm for homogeneous or near homogeneous random fields is presented. This algorithm relies on the fundamental-harmonic relationship among spectral peaks to identify the harmonic frequencies, and uses Hough transformation to detect spectral evanescent components. A local variance based procedure is developed to determine the spectral peak support. Comparing to the existing global thresholding scheme and maximum-likelihood parameter estimation, this algorithm is more robust and flexible for the large variety of natural textures, as well as computationally more efficient than the maximum-likelihood method.

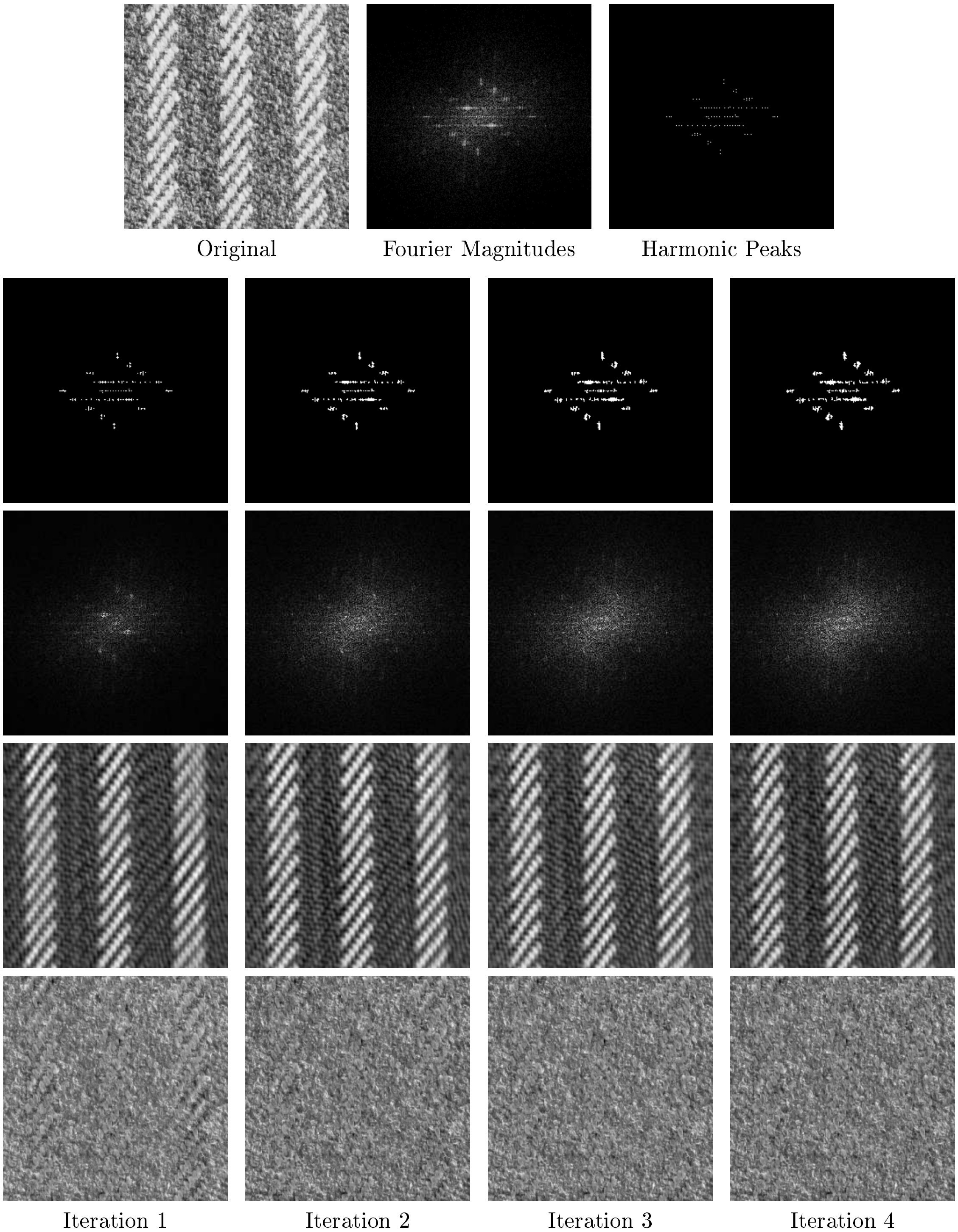


Figure 3-8: Decomposition of Brodatz texture D11: Homespun Woolen Cloth. Shown in row 2 to row 5 are the processing results at the end of iterations 1 to 4: mask images; Fourier magnitude images; Harmonic components; and indeterministic components.

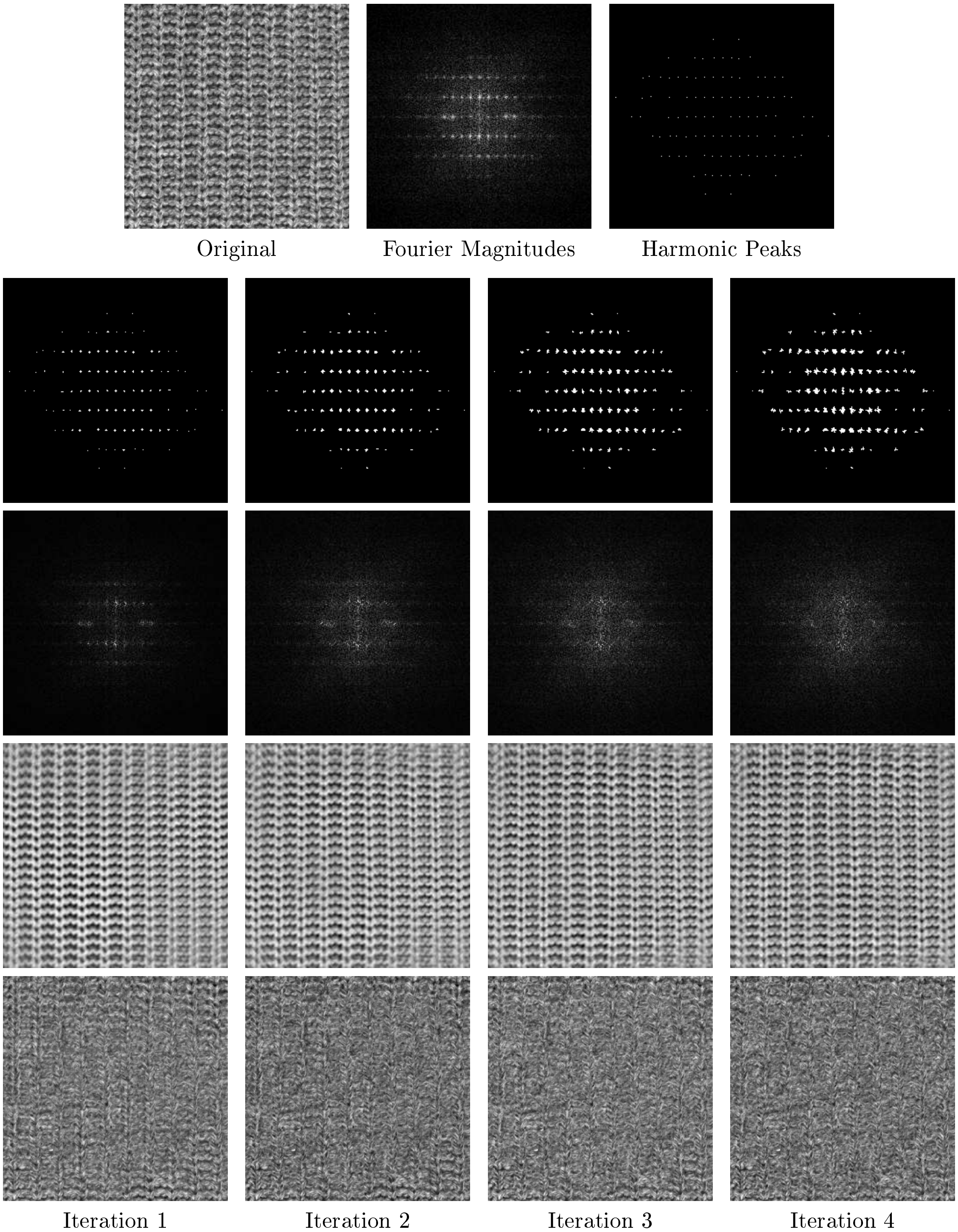


Figure 3-9: Decomposition of texture Sweater. Shown in row 2 to row 5 are the processing results at the end of iterations 1 to 4: mask images; Fourier magnitude images; Harmonic components; and indeterministic components.

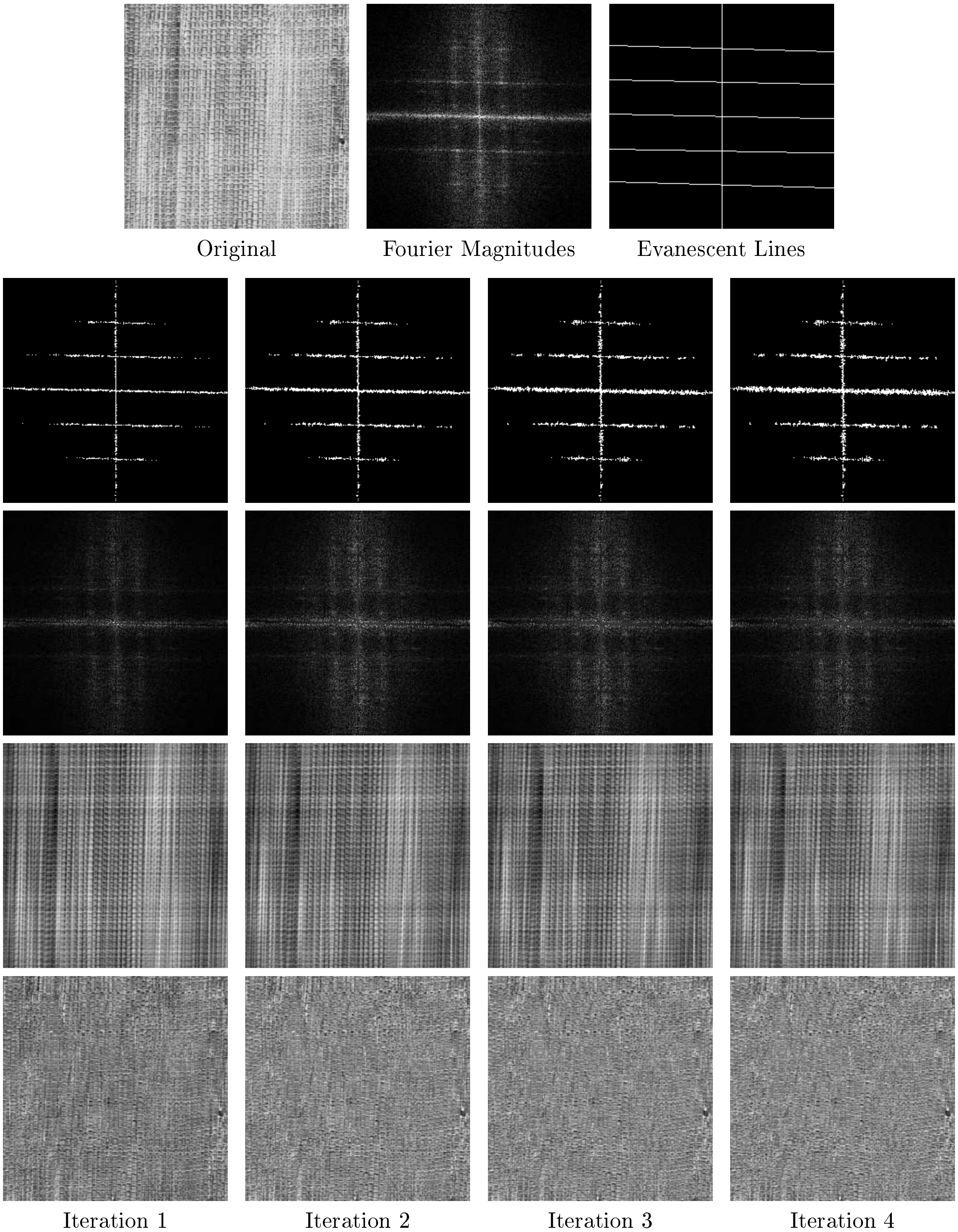


Figure 3-10: Decomposition of Brodatz texture D78: Oriental Straw Cloth. Shown in row 2 to row 5 are the processing results at the end of iterations 1 to 4: mask images; Fourier magnitude images; Harmonic components; and indeterministic components.

Chapter 4

Perceptual Properties

4.1 Introduction

In this chapter, the perceptual properties of Wold-based texture modeling are studied. The "perceptual property of a texture model" is the correspondence between the computational features provided by the model and the perceptual visual properties of the texture pattern.

Why is the perceptual property of a texture model important? This is perhaps best explained by using the example application of image database retrieval. (A retrieval experiment is presented in the next chapter.) In image retrieval, a computer system is expected to fetch back from the database the images that are *similar* to some user selected prototypes. Image retrieval involves similarity comparison. The underlying mechanism of a typical retrieval system is as follows. Each database image is represented by a set of pre-computed features in a feature space. In response to a query, distance measures are computed in the feature space to gauge the similarity between the database images and the prototypes. Images that are the most similar to the prototypes based on the particular similarity measures used are returned to the user. In this retrieval model, the construction of the features and the choice of the distance measures, usually closely related to each other, are crucial to the success of the system. One common criterion for evaluating the retrieval results is the *perceptual similarity* between the retrieved set and the prototype set, *i.e.,* whether the images are alike in their visual appearance. If the computational image features reflect the perceptual characteristics of the images, the image similarity measured by the computer algorithm can be expected to correspond well to the perceptual similarity.

In the following sections, previous work on the perceptual properties of computational texture models is first reviewed and then human and computer experiments conducted to study the perceptual properties of Wold-based texture modeling are presented.

4.2 Previous Work

4.2.1 Tamura *et al.*

From the descriptions seen in the literature and from observations of the Brodatz textures, Tamura *et al.* chose six visual textural properties — coarseness, contrast, directionality, line-likeness, regularity, and roughness — to model as computational texture features [87].

Human experiments were conducted to establish the ground truth ordering of 16 Brodatz texture samples based on each of the six textural properties. In the experiment, 48 human subjects, 28 male and 20 female, gave their pairwise judgment for all possible pairs of the 16 texture samples according to each of the six properties. The test data were processed to produce a one-dimensional ordering of the test images for each textural property. For example, for the coarseness property, the textures were ordered from coarse to fine.

To improvise computational features for the six textural properties, Tamura *et al.* tested and modified heuristic features proposed in the literature as well as composing new ones. For each feature, the 16 Brodatz samples were ordered based on the feature values computed on each texture. The final computational feature for each textural property was chosen as the one that provided the highest Spearman correlation coefficient (see Section 4.6.4) between the computer and the human orderings.

By the correlation between the human and the computer data, the computational features for coarseness, contrast, and directionality were considered to have achieved successful correspondences with the human data while the other three were not so successful. Strong correlations were observed between coarseness and contrast and between directionality and line-likeness. An attempt was made to measure texture similarities by using simple combinations of the six computational features, but the results did not correspond to the human data well.

4.2.2 Amadasun and King

Amadasun and King proposed five computational features corresponding to textural properties of coarseness, contrast, busyness, complexity, and texture strength [2]. No strong reasons were given why these properties were chosen. The computational features were composed heuristically from the absolute differences between the gray scale value of each pixel and the averaged gray scale value in a neighborhood surrounding the pixel.

The relations between the computational features and five textural properties were studied via a human texture ranking experiment. Ten Brodatz texture samples were ranked by 88 subjects, 48 male and 40 female, by each of the five properties. The computer rankings were based on the values of the features computed on each texture. The final human ranking for each feature was the order of the rank sum of each texture sample. The correspondences between the computational features and the textural properties were evaluated by computing the Spearman correlation coefficients between the computer and the human rankings for each property. The coefficient values ranged from 0.503 to 0.856, with the lowest for texture strength and the highest for coarseness.

The correlations among the feature rankings and among the property rankings were also computed. Strong correlations were observed between the coarseness and texture strength and between contrast and complexity for both the features and the properties. Combinations of the computational features were further tested in a similarity measurement experiment. The results were slightly better than that of the Tamura experiment.

4.2.3 Remarks

One common problem with the two studies above is that the textural properties were chosen largely based on the intuition and observation of the researchers. Although these features seem to be characteristic of natural textures, it is not clear what the relative importance of these features is and how well they span the perceptual space of human texture perception. As shown by the experimental data, strong correlations exist among some of the textural properties investigated.

Another common problem is that the computational features were improvised heuristically for the individual textural properties. It is not clear how well these separate features can be used together to represent a texture pattern.

Finally, while both studies used a large number of human subjects in the experiment, the concordance of the human data was not evaluated in any manner.

4.3 Experimental Design

4.3.1 Dimensions of Human Texture Perception

Rao and Lohse conducted a human study to identify the relevant dimensions of human texture perception [79]. In their experiment, twenty test subjects first rated 56 pictures from the Brodatz album on twelve 9-point Likert scales labeled by adjectives such as repetitive, directional, random, granular, uniform, regular, *etc.*. Then, the subjects were asked to sort the pictures into groups of similar items. The initial groupings were subsequently grouped again and again into higher-order clusters of similar groups, until all pictures were in a single group. The Likert scale data were analyzed by using classification and regression tree analysis, discriminant analysis, and principle component analysis, while the grouping data were analyzed by using hierarchical cluster analysis and non-parametric multidimensional scaling (MDS). Combining the analysis results of both scaling and grouping data, the top three dimensions of human texture perception were identified. These dimensions are shown in Figure 4-1. It was reported that the repetitiveness in texture, which is represented by the X axis, appears to be the most important feature used by humans in distinguishing textures. This property of repetitiveness is also significantly correlated with that of regularity, uniformity, and non-randomness.

4.3.2 Objectives

The purpose of this experiment is to investigate the perceptual properties of Wold-based texture modeling. The main premise of the experiment is that Wold texture modeling results in behavior similar to that of humans in discriminating textures along the most important dimension of human texture perception — repetitiveness vs. randomness.

4.3.3 General Procedure

The perceptual properties of Wold-based texture modeling are studied here in a texture ordering experiment. The experimental design is based on the result of Rao and Lohse's work. Since the top perceptual dimensions have already been identified by Rao and Lohse, the current experiment takes the form of texture ranking (ordering) instead of free sorting. The focus of this study is on the X axis of repetitiveness vs. randomness, the perceptually most salient dimension.

In the experiment, human subjects order a set of textured images along the chosen perceptual dimension. A computer program orders the same set of images using the Wold computational model. Then the images are ordered again based on the averaged human ranking scores to produce the final human ordering. The correlation between the final human ordering and the computer

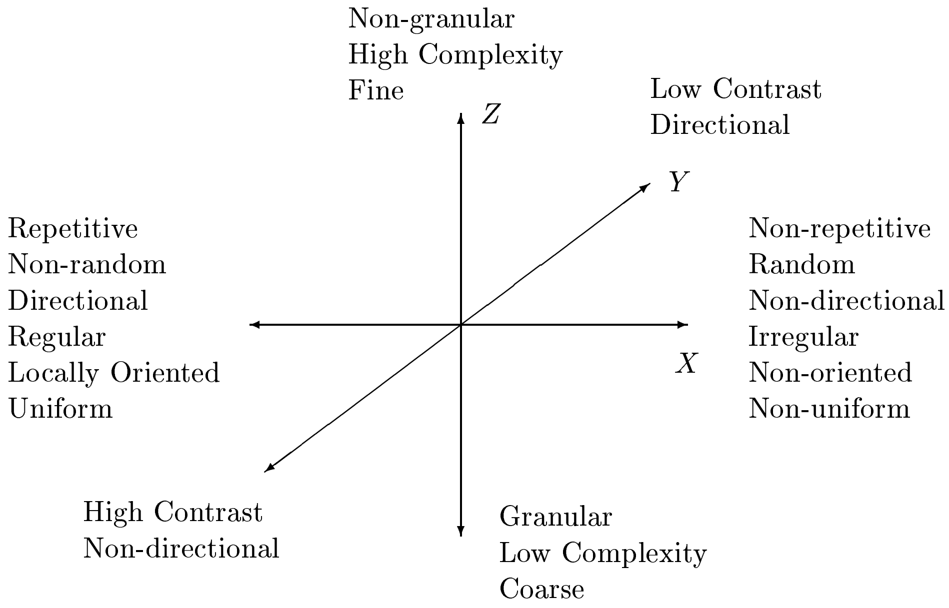


Figure 4-1: Top three dimensions of human texture perception identified by Rao and Lohse's study.

---

ordering is used to gauge how well the computational model captures the perceptual properties of the images along the axis.

4.3.4 Representing the Perceptual Axis

Different methods can be used to explain to the human subjects about the characteristics of the perceptual axis along which the test images are to be ordered. One method is to exemplify the two extremities of the axis by sample images. However, since the perceptual dimension is identified via multidimensional scaling and other statistical analysis of high dimensional test data, the visual properties it incorporates are better described in abstract terms. It is also difficult to select a small set of example images to convey the meaning of the axis accurately since the visual properties that human subjects derive from the examples can vary widely from person to person.

An alternative is to use the two sets of adjectives that are associated with the axis when it is identified. Using the adjectives also helps the subjects to focus on using the visual perceptual cues rather than the semantic categorical labels, such as the common names of the patterns, when comparing images. A potential problem is that the understanding of the English adjectives may vary among individual subjects. This problem is dealt with by averaging the ranking data across a large number of subjects.

4.3.5 Computational Periodicity Measure

For a computer algorithm to order the test images along the axis of repetitiveness vs. randomness, a quantitative measure is needed to gauge the amount of periodicity and randomness in each image. The orthogonal Wold components of an image have distinctive visual properties: the harmonic component appears to be regular and repetitive, the evanescent component looks directional, and the indeterministic component is random. Therefore, it is conceivable that the Wold components can be used to represent the perceptual properties of a texture pattern. The question is: what physical quantities of these components should be used to measure their perceptual strength?

It has been suggested that the human visual system contains simple mechanisms that measure the local energy present in the concentric and oriented receptive fields [9][8]. Computationally, this perception model has been implemented as a set of oriented linear filters followed by some rectifying nonlinearity that computes the local energy of the filter output [10]. By this model, the spatial properties of texture patterns are encoded in the local energy distribution of the filter output. Therefore, it is reasonable to use the total energy associated with a particular textural property to represent the perceptual strength of that property.

In the computer experiment, the **deterministic energy ratio** is used as the quantitative measure along the perceptual axis of repetitiveness vs. randomness. By Theorem 2, an image $y(m,n)$ can be decomposed into a deterministic component $v(m,n)$ and an indeterministic component $w(m,n)$ as

$$y(m,n) = v(m,n) + w(m,n),$$

where the deterministic component $v(m,n)$ includes both the harmonic and the evanescent Wold components. The **deterministic energy** of the image is the energy contained in component $v(m,n)$. Using Parseval's theorem [59], the deterministic energy E_v and the total energy E_y of the image can be computed as

$$E_v = \sum_{(m,n)\in\mathcal{D}} |v(m,n)|^2 = \frac{1}{N^2} \sum_{(k,l)\in\mathcal{D}} |V(k,l)|^2 \tag{4.1}$$

and

$$E_y = \sum_{(m,n)\in\mathcal{D}} |y(m,n)|^2 = \frac{1}{N^2} \sum_{(k,l)\in\mathcal{D}} |Y(k,l)|^2, \tag{4.2}$$

where $V(k,l)$ is defined by Equation (3.21). The deterministic energy ratio is then E_v/E_y.

4.4 Human Experiment

4.4.1 Method

Subjects

Thirty-two subjects, with an equal number of males and females, participated in the study. The subjects are MIT students and staff from various disciplines. Their ages range from 18 to 36.

Token	Material Name
D1	Woven Aluminum Wire
D9	Grass Lawn
D11	Homespun Woolen Cloth
D26	Ceramic-coated Brick Wall
D29	Beach Sand
D32	Pressed Cork
D34	Netting
D52	Oriental Straw Cloth
D55	Straw Matting
D57	Handmade Paper
D64	Handwoven Oriental Rattan
D78	Oriental Straw Cloth
D80	Oriental Straw Cloth
D82	Oriental Straw Cloth
D83	Woven Matting
D93	Fur
D94	Brick Wall
D101	Cane
D102	Cane
D110	Grassy Fiber

Table 4.1: Names of the Brodatz album pictures from which the test samples were made.

Materials

The test samples were the 20 Brodatz textures shown in Figure 4-2. The names of the Brodatz album pictures from which the test samples were made are listed in Table 4.1. These samples include all the relatively homogeneous patterns among the 56 Brodatz textures used in Rao and Lohse's study.[1] The physical test samples were made by pasting the cutouts of Brodatz album original glossy prints over 7cm by 7cm cardboards.[2] The samples were shown to the subjects under normal indoor lighting conditions.

The two sets of adjectives shown at the extremities of the X axis in Figure 4-1 were printed at the two ends of a 14cm by 175cm board. In the experiment, the test subjects were asked to order the texture samples along the board, between the two sets of adjectives.

[1]Recall that the Wold-based decomposition has the homogeneity assumption. In practice, certain inhomogeneities can be tolerated by the algorithms, but only to an extent.

[2]Samples of texture D1, D26, D64, and D94 were made of the laser printer printouts of the digitized originals in finer scale.

Procedure

At the beginning of an experiment, a page of written instructions was shown to the subject to explain the test apparatus and the task (see Appendix B). The texture samples were given to the subjects in a randomized pile, next to the test board. There was no time limit for completing the task.

The ordering score, from 1 to 20, was recorded for each sample image as the test data. In addition, the subject was asked to give a confidence measure of the ordering on Likert scale 1 (least confident) to 7 (most confident). The age and gender of the subjects were also recorded.

4.4.2 Experimental Data

The data collected in the experiment are shown in Table 4.2. The first column lists the subject numbers. The second column records the subjects' gender. The third column contains the confidence ratings each subject gave to his or her ordering. The confidence ratings are on Likert scale 1 (least confident) to 7 (most confident). The rest of the columns in the table are the ranking scores each of the 20 test image received from every subject.

4.5 Computer Experiment

4.5.1 Method

Input Data

The computer test images were the digitized version of the 20 texture samples used in the human experiment. The images were scanned from the Brodatz album by an HP ScanJet IICX scanner. The digital images are in 8-bit gray scale and cropped to 256 pixel by 256 pixel squares over the same image regions that are shown in the samples used in the human test.

Procedure

The test images were first zero-meaned and Gaussian tapered using the tapering window function shown in Figure 3-2 (b). Gaussian tapering eliminates any possible effect of the image boundary conditions on the spectral peak extraction and subsequently the energy computation. The images were decomposed into their deterministic and indeterministic components using the spectral decomposition method developed in Chapter 3. The final output of the program was the ratio E_v/E_y of each test image.

4.5.2 Experimental Data

The 20 test images are shown in Figure 4-2, and their Fourier magnitudes in Figure 4-3. The harmonic peak central frequencies and the evanescent lines detected from each test image are displayed in Figure 4-4 and Figure 4-5. Figure 4-6 shows the mask images, which contain the frequency locations of the deterministic component of each test sample. Shown in Figure 4-7 are the remaining Fourier magnitudes of the test images after the deterministic frequencies are extracted. In each of the images in Figure 4-7, the values at the deterministic frequencies are replaced by the corresponding values of the Gaussian surface that are fit to the magnitude image

Subj	Sex	Conf	D1	D9	D11	D26	D29	D32	D34	D52	D55	D57	D64	D78	D80	D82	D83	D93	D94	D101	D102	D110
01	M	5	1	20	11	3	19	18	2	4	9	17	7	15	14	12	10	13	8	6	5	16
02	M	6	4	18	13	7	19	17	1	5	6	16	9	12	14	11	8	15	10	3	2	20
03	M	6	4	19	11	6	17	20	1	8	9	18	7	14	13	12	10	16	5	2	3	15
04	M	6	1	19	11	9	18	14	6	4	2	15	3	13	16	5	12	17	10	8	7	20
05	M	6	1	16	8	5	18	17	2	10	9	19	7	13	15	12	11	14	6	3	4	20
06	F	6	2	18	12	6	16	17	1	11	5	20	8	14	13	7	9	15	10	3	4	19
07	M	6	1	19	5	10	15	18	2	11	8	16	13	12	14	7	6	17	9	4	3	20
08	F	5	1	19	3	14	16	17	4	10	2	15	7	11	12	9	8	18	13	5	6	20
09	M	5	2	19	6	8	18	15	1	12	3	16	10	14	7	13	11	17	9	4	5	20
10	F	4	6	16	1	8	19	20	11	12	10	18	9	13	14	3	2	15	7	5	4	17
11	M	6	1	19	6	5	17	18	4	3	2	16	8	12	11	9	7	15	10	13	14	20
12	M	6	6	19	5	4	15	17	3	9	10	16	8	11	14	12	13	18	7	2	1	20
13	F	6	7	19	11	6	16	18	5	4	1	17	2	13	14	8	12	15	3	10	9	20
14	M	5	3	19	13	4	16	18	1	2	5	17	8	12	14	10	11	15	9	6	7	20
15	M	5	1	18	13	4	16	20	7	11	3	19	10	14	12	9	2	15	8	5	6	17
16	F	3	1	16	10	12	18	19	13	9	2	17	6	3	14	7	8	15	11	5	4	20
17	M	5	3	19	6	4	17	18	7	8	12	16	9	14	13	11	10	15	5	1	2	20
18	M	6	3	17	10	6	18	20	4	13	9	19	7	14	12	8	11	15	5	2	1	16
19	F	5	10	19	6	11	17	18	1	2	5	16	4	7	13	3	14	15	12	8	9	20
20	F	5	1	17	13	5	18	20	2	11	7	19	9	8	14	12	10	15	6	3	4	16
21	M	5	11	18	13	10	16	20	4	3	2	17	6	1	14	5	12	15	9	7	8	19
22	M	4	5	16	7	13	18	19	3	9	4	17	10	14	8	11	6	15	12	1	2	20
23	F	4	6	16	9	7	18	19	3	4	5	20	10	14	13	11	12	15	8	1	2	17
24	F	6	3	18	6	10	17	19	2	1	9	16	7	8	15	4	5	14	11	13	12	20
25	F	6	2	19	8	10	16	17	1	7	3	15	4	14	13	5	6	18	9	12	11	20
26	M	5	1	17	14	9	20	18	13	12	6	19	4	7	15	5	10	11	8	2	3	16
27	F	6	1	19	8	11	17	16	2	5	3	18	12	10	15	4	9	14	13	7	6	20
28	F	5	4	19	14	9	16	17	3	8	5	18	6	7	11	13	12	15	10	2	1	20
29	F	6	4	16	10	3	18	19	11	5	8	17	6	12	14	9	13	15	7	2	1	20
30	F	6	2	16	6	11	19	18	1	5	7	17	8	13	14	10	9	15	12	4	3	20
31	F	5	2	16	12	1	18	17	7	8	5	19	6	13	14	10	9	15	11	4	3	20
32	F	6	1	16	2	11	19	17	3	7	9	18	10	14	13	8	6	15	12	5	4	20

Table 4.2: Human ranking data of 32 subjects on 20 test images along the perceptual dimension of repetitiveness vs. randomness. First column: subject numbers. Second column: subjects' gender. Third column: confidence ratings each subject gave to his or her ordering. These ratings are on a Likert scale 1 (least confident) to 7 (most confident). The rest of the columns are the ranking scores each test image received from every subject.

Name	E_v/E_y (%)	Rank
D1	92.49	3
D9	0.0	18
D11	72.25	10
D26	79.13	6
D29	0.0	18
D32	0.0	18
D34	88.57	4
D52	75.27	7
D55	84.43	5
D57	17.77	15
D64	74.60	8
D78	59.27	11
D80	35.41	14
D82	57.73	12
D83	73.60	9
D93	0.0	18
D94	43.65	13
D101	96.19	1
D102	94.51	2
D110	0.0	18

Table 4.3: Test sample deterministic energy ratio E_v/E_y. The ranking scores are obtained by ordering the images based on their deterministic energy ratios.

---

in the decomposition process. The magnitude images in Figure 4-7 are individually scaled to the display range of $[0, 255]$.

The energy ratio E_v/E_y of all test images are shown in Table 4.3. The ranking scores are obtained by ordering the images based on their deterministic energy ratios. Image D9, D29, D32, D93, and D110 are tied for ranks 16 to 20. These images are given the rank 18, which is the mean of the ranks for which the images are tied.

4.6 Data Analysis and Results

4.6.1 Overview

The data processing consists of a series of statistical tests. The goal is to decide if the human and the computer ranking data are significantly correlated. The Spearman rank correlation coefficient r_s and the Kendall rank correlation coefficient τ are used to assess the correlation between the ranking based on the averaged human ranks and the computer ranking.

To ensure that the ranking based on the averaged ranks is the best estimate of the "true" human

Rating Scale	1	2	3	4	5	6	7
Number of Subjects	0	0	1	3	12	16	0

Table 4.4: Summary of human ranking confidence data. Scale 1 is the least confident and 7 the most confident.

ranking, the overall agreement among the 32 sets of human rankings is ascertained by using the Kendall concordance coefficient W. The concordance within and between the rankings of the male and the female subject groups is also evaluated.

4.6.2 Confidence Ratings

The confidence ratings were collected to evaluate the validity of the rankings. As shown in Table 4.4, 28 out of 32 subjects were confident with their rankings, 3 were neutral, and 1 was not so confident. No subject was completely uncertain about his or her ranking. Therefore, all the ranking data are considered valid for analysis.

4.6.3 Final Rankings

For each test image, the human ranking scores are averaged across all 32 subjects. The final human ranking is generated by ordering the test images based on the averaged ranking scores. The averaged and the final human ranking scores are shown in Table 4.5.

The final human ranking and the computer ranking are shown together in Table 4.6 by the ascending ordering of the ranking scores. Increasing rank values correspond to moving along the perceptual axis from repetitive to random. In Figure 4-8 and Figure 4-9, from left to right and top to bottom, the test samples are displayed in the order of the final human ranking and the computer ranking respectively, from repetitive to random.

4.6.4 Spearman correlation coefficient r_s

Method

The Spearman rank correlation coefficient r_s measures the degree of association or correlation between two sets of ranking scores. The Spearman's r_s can be derived from the Pearson product-moment correlation coefficient r. Denote the two sets of ranking scores as X_j and Y_j, $j = 1, \cdots, N_s$. The Pearson's r is defined as

$$r = \frac{N_s(\sum X_j Y_j) - (\sum X_j)(\sum Y_j)}{\sqrt{[N_s \sum X_j^2 - (\sum X_j)^2][N_s \sum Y_j^2 - (\sum Y_j)^2]}}, \tag{4.3}$$

Sample Name	Average Rank	Final Rank
D1	3.15625	1
D9	17.84375	18
D11	8.84375	10
D26	7.56250	8
D29	17.34375	17
D32	17.96875	19
D34	4.09375	2
D52	7.28125	6
D55	5.78125	5
D57	17.28125	16
D64	7.50000	7
D78	11.43750	13
D80	13.18750	14
D82	8.59375	9
D83	9.18750	12
D93	15.21875	15
D94	8.90625	11
D101	4.93750	4
D102	4.87500	3
D110	19.00000	20

Table 4.5: Averaged and final human ranking scores of texture ordering.

where the sums are from $j = 1$ to N_s. When the sample values are ranks, $r = r_s$, and [55]

$$r_s = 1 - \frac{6 \sum_{j=1}^{N_s} (X_j - Y_j)^2}{N_s(N_s^2 - 1)}. \tag{4.4}$$

The range of r_s is $[-1, 1]$. The value -1 corresponds to complete disagreement between the two rankings, and the value 1 indicates complete agreement.

Testing the Significance of r_s

The significance of Spearman's r_s can be tested under the null hypothesis H_0 that the two rankings are not associated and the observed value of r_s differs from zero only by chance. When $N_s > 10$, the statistic

$$t = r_s \sqrt{\frac{N_s - 2}{1 - r_s^2}} \tag{4.5}$$

Human		Computer	
Rank	Name	Rank	Name
1	D1	1	D101
2	D34	2	D102
3	D102	3	D1
4	D101	4	D34
5	D55	5	D55
6	D52	6	D26
7	D64	7	D52
8	D26	8	D64
9	D82	9	D83
10	D11	10	D11
11	D94	11	D78
12	D83	12	D82
13	D78	13	D94
14	D80	14	D80
15	D93	15	D57
16	D57	18	D9
17	D29	18	D29
18	D9	18	D32
19	D32	18	D93
20	D110	18	D110

Table 4.6: Final human ranking and computer ranking scores. Increasing score values correspond to moving along the perceptual axis from repetitive to random.

---

has the Student's t-distribution with degrees of freedom $df = N_s - 2$ [55]. Thus, the probability p of observing under H_0 a value as large as r_s can be determined by first computing the corresponding t value and then finding the significance of that t in a t-distribution table.

Results

The Spearman correlation coefficient for the human and the computer ranking scores in Table 4.3 and Table 4.5 is

$$r_s = 0.9504$$

with $t = 12.96$. The probability for this r_s value to occur under H_0 is $p < .001$. Therefore, the hypothesis H_0 can be rejected with the probability of error less than 0.1%. The conclusion is that the human ranking and the computer ranking are significantly correlated.

4.6.5 Kendall correlation coefficient τ

Method

The Kendall correlation coefficient τ takes a different approach to assessing the correlation between two sets of ranking scores. Denote the two lists of N_s ranking scores as X and Y and put them side by side to form an N_s by 2 matrix. Now reorder the rows of the matrix so that the X column is in its natural order, *i.e.,* $1, \cdots, N_s$. Denote the ranks in the new Y column as Y_j, $j = 1, \cdots, N_s$. A credit score is calculated for each of the Y_j's as follows. For a particular rank Y_j, its value is compared to all the rank values Y_k, $k > j$. When $Y_k > Y_j$, Y_j earns a credit $+1$, otherwise, a credit -1. Next, the credit scores of each Y_j are summed together as the actual total credit S. The value of S is at its maximum $\frac{1}{2}N_s(N_s - 1)$ when the original X and Y rankings are identical, *i.e.,* the reordered Y column is also in its natural order. The Kendall's τ is defined as the ratio between the actual total credit S and the maximum total credit,

$$\tau = \frac{2\,S}{N_s(N_s - 1)}. \tag{4.6}$$

The range of τ is $[-1, 1]$. The value -1 corresponds to complete disagreement between the two rankings, and the value 1 indicates complete agreement.

Testing the Significance of τ

The significance of Kendall's τ can be tested under the null hypothesis H_0 that the two sets of ranks X and Y are unrelated. When $N_s > 10$, the distribution of τ under H_0 can be approximated by the normal distribution $N(\mu_\tau, \sigma_\tau)$ [55], where

$$\mu_\tau = 0, \qquad \sigma_\tau = \sqrt{\frac{2\,(2N_s + 5)}{9N_s(N_s - 1)}}.$$

That is,

$$z = \frac{\tau - \mu_\tau}{\sigma_\tau} = \tau\sqrt{\frac{9N_s(N_s - 1)}{2\,(2N_s + 5)}} \tag{4.7}$$

is approximately normally distributed with zero mean and unit variance. Thus, the probability p of observing under H_0 a value as large as τ can be determined by first computing the corresponding z value and then finding the significance of that z in a normal distribution table.

Results

The Kendall correlation coefficient for the human and the computer ranking scores in Tables 4.3 and 4.5 is

$$\tau = 0.7474$$

with $z = 4.61$. The probability for this τ value to occur under H_0 is $p < .001$. Again, the hypothesis H_0 can be rejected with the probability of error less than 0.1% and the conclusion is that the human ranking and the computer ranking are significantly correlated.

4.6.6 Comparison of r_s and τ

The values of r_s and τ are not identical when both are computed from the same ranking data. In fact, r_s and τ have different underlying scales, and numerically they are not directly comparable to each other. However, the two coefficients have the same power in detecting the existence of association in the rankings. That is, the distributions of r_s and τ are such that, with a given set of data, both will reject the null hypothesis at the same level of significance [84].

4.6.7 Concordance of All Human Data

Overview

In the above, the averaged human ranking scores are used to determine the final ranking of the test images. The final ranking scores are subsequently used to assess the correlation between the human and the computer ranking data. However, the validity of determining the final ranks based on the averaged ranking scores should be justified. This justification has two aspects. One is the inter-subject reliability of the human data, *i.e.*, whether there exist certain underlying criteria upon which the test subjects agree; the other is, assuming the reliability of the data, whether the final ranking determined by the averaged ranks are the best estimate of the "true" ranking according to the underlying criteria.

The Kendall concordance coefficient W is used to examine the inter-subject reliability of the data. The magnitude and significance of W provides evidence whether any underlying criterion exists in the ranking data. When the evidence is positive, it is shown by Kendall [55] that the best estimate of the "true" ranking is provided by the order of the averaged ranks in a least-squares sense.

Kendall Concordance Coefficient W

Given the rankings of K_s subjects on N_s entities, the Kendall concordance coefficient W is constructed as follows. First, all the ranking data are arranged into a K_s by N_s data matrix. The rows of the matrix are ordered by the subjects and the columns by the entities ranked. Thus, each column of the matrix contains the ranks given by all subjects to a particular entity. Denote the rank sum of each entity, which is the sum of the ranks in each column, by R_j, $j = 1, \cdots, N_s$. When all the subjects are in perfect agreement in their rankings, the rank sum R_j's take the value $K_s, 2K_s, 3K_s, \cdots, N_sK_s$, though not necessarily in that order. On the other hand, when there is no agreement among the subjects, the values of the R_j's will be approximately equal. Therefore, the degree of agreement among the subjects is reflected by the degree of variance among the N_s rank sums. This variance reaches its maximum when the perfect agreement occurs among the rankings. The Kendall concordance coefficient W is defined as the ratio between the actual variance V_R and the maximum variance of the rank sums. The value of W can be computed as [55]

$$W = \frac{12\,V_R}{K_s^2 N_s (N_s^2 - 1)}, \tag{4.8}$$

where

$$V_R = \sum_{j=1}^{N_s} \left(R_j - \frac{1}{N_s} \sum_{j=1}^{N_s} R_j \right)^2 . \tag{4.9}$$

The range of W is [0,1], and $W = 1$ when subjects are in perfect agreement.

The principle of the Kendall concordance coefficient W is further explained by its relationship with the Spearman correlation coefficient r_s. The concordance of K_s sets of rankings can be evaluated by computing the average value of the Spearman correlation coefficients between all possible pairs of the rankings. Denote the average value of the Spearman correlation coefficients as $r_{s_{av}}$. It can be shown that the Kendall concordance coefficient W has a linear relationship with value $r_{s_{av}}$ [55]:

$$r_{s_{av}} = \frac{K_s W - 1}{K_s - 1}. \tag{4.10}$$

Testing the Significance of W

The significance of Kendall concordance coefficient W can be tested under the null hypothesis H_0 that there is no agreement among the K_s sets of rankings. When $N_s > 7$, the distribution of statistic

$$\chi_r^2 = \frac{12\,V_R}{K_s N_s (N_s^2 + 1)} = K_s(N_s - 1)W \tag{4.11}$$

is approximately χ^2 with degrees of freedom $df = N_s - 1$ [55]. The probability p of observing under H_0 a value as large as W can be determined by first computing the corresponding χ_r^2 value and then finding the significance of that χ_r^2 in a χ^2 distribution table.

Results

The Kendall concordance coefficient for the 32 sets of human ranking data in Table 4.2 is

$$W = 0.7874$$

with $\chi_r^2 = 478.72$. The probability for this χ_r^2 value to occur under H_0 is $p < .001$. Therefore, the hypothesis H_0 can be rejected with the probability of error less than 0.1%. The conclusion is that the human ranking is reliable and there exist certain underlying criteria upon which the test subjects agree.

4.6.8 Concordance Within and Between Male and Female Groups

Further analysis has been conducted to examine the agreement in ranking within and between the male and female subjects. A straightforward procedure is to assess the concordance of the rankings among the male and female groups separately. If the magnitude and the significance level of the concordance coefficients suggest that there is strong agreement within each group, the agreement between the two groups can be assessed by computing the Spearman correlation coefficient for the rankings provided by the orders of the averaged ranks of each group.

For the male group, the Kendall concordance coefficient is $W = 0.8011$, with statistic $\chi_r^2 = 243.55$ and significance $p < .001$. For the female group, the coefficient is $W = 0.7872$, with $\chi_r^2 = 239.30$ and $p < .001$. Thus, both male and female groups exhibit strong within-group agreement in their rankings.

The averaged rankings of the male and female groups are shown in Table 4.7, together with the final ranking scores based on the averaged ranks. The Spearman correlation coefficient r_s for the

Sample Name	Male		Female	
	Average Rank	Final Rank	Average Rank	Final Rank
D1	3.0000	1	3.3125	1
D9	18.2500	19	17.4375	17
D11	9.5000	11	8.1875	9
D26	6.6875	6	8.4375	10
D29	17.3125	17	17.3750	16
D32	17.9375	18	18.0000	19
D34	3.8125	2	4.3750	2
D52	7.7500	7	6.8125	6
D55	6.1875	5	5.3750	4
D57	17.0625	16	17.5000	18
D64	7.8750	8	7.1250	7
D78	12.0000	13	10.8750	13
D80	12.8750	14	13.5000	14
D82	9.5000	12	7.6875	8
D83	9.3750	10	9.0000	11
D93	15.1875	15	15.2500	15
D94	8.1250	9	9.6875	12
D101	4.3125	3	5.5625	5
D102	4.5625	4	5.1875	3
D110	18.6875	20	19.3125	20

Table 4.7: Averaged and final ranking scores of male and female groups.

final rankings of the two groups is $r_s = 0.9519$, with statistic $t = 13.18$ and significance $p < .001$. Hence, the rankings of the male and the female subject groups are significantly correlated.

The conclusion of the concordance evaluation of the male and female ranking data is that the rankings exhibit strong agreement within each group and there is no significant difference between the rankings of the two groups.

Schucany and Frawley extended the concept of the L statistic introduced by Page [71] and proposed a rank test for two group concordance [83]. This test uses the $\mathcal{L}$ statistic, which is defined as the inner product of the two sets of rank sums of the two groups, to assess both the within-group and the between-group concordance. However, since the rank sum values are proportional to the averaged ranks, this method uses the same basic information in data as the procedure presented above does. Given the very significant results of the Kendall's concordance within each group and the Spearman's correlation between the groups, it is very unlikely the $\mathcal{L}$ test would conclude otherwise.

4.6.9 Concordance of Combined Human and Computer Data

The Kendall concordance coefficient for combined human and computer data is $W = 0.7902$, with statistic $\chi_r^2 = 495.44$ and significance $p < .001$. Notice that this W value is larger than that of the human data alone. Considering also the high correlation between the computer and the averaged human rankings, it can be seen that the behavior of the computer in the texture ordering experiment is indistinguishable from that of the human subjects.

4.7 Implications of the Experimental Results

The following conclusions can be drawn from the experimental results:

1. The highly significant correlation between the human and the computer texture ranking data suggests that the component energy resulting from the 2-D Wold decomposition of an image is a good computational measure for the most salient dimension of human texture perception, the dimension of repetitiveness vs. randomness.
2. The highly significant concordance of the human rankings indicates the following:
   (a) There exists a common interpretation to the semantic labels associated to the perceptual dimension.
   (b) These labels indeed correspond to certain underlying criteria, upon which the human subjects agree, for texture similarity measurement.

It should be emphasized that the Wold-based texture model is a computational image model, and not a model for visual texture perception. The purpose of investigating the perceptual properties of the Wold model is to provide assurance that the model behaves consistently with the texture perception and therefore can facilitate tasks such as image similarity comparison.

4.8 Discussion

4.8.1 Comparison to Previous Work

The current study differs in many respects from the existing work reviewed in Section 4.2. In accordance with the earlier remarks, the following observations can be made.

First, the textural properties for which the Wold-based texture modeling is examined are chosen based on an independent study of human texture perception. The experiment is carried out along the perceptually salient dimension. Therefore, there is little doubt that the observed correspondences between the textural properties and the computational features are important for texture modeling.

Second, Wold-based texture modeling has a solid theoretical foundation. The computational features are not composed purely heuristically. Consequently, the Wold model can provide computational descriptions that contain sufficient information for both texture representation and synthesis.[3]

[3]Texture synthesis examples can be found in [27][28].

Finally, the analysis of the experimental data is more rigorous in that both the significance of the ranking correlations and the concordance of the human data are tested.

4.8.2 Early Vision Models and Texture Modeling

Early Vision Models

The vast majority of human texture perception research has been carried out in the context of texture discrimination and segregation (See [7] for a review). The early work includes the texton theory by Julesz *et al.* [46][47][48] and the physical attribute based approach of Beck [5][3][4]. More recently, computationally explicit early vision models have been proposed [6][9][10][16] [41][64][91]. The fundamental elements of these models are the outputs of oriented linear filters (or spatial frequency channels), which are a crude model of the response properties of mammalian cortical cells. The explanatory and predictive power of these perception models has been demonstrated in various texture discrimination and segregation tasks. An example is the texture segregation model by Bergen and Landy [10]. The initial stage of the model consists of a set of linear filters in different orientations and spatial scales. These filters are followed by a rectifying nonlinearity, which computes the energy of the filter output. The basic image representation provided by the initial stage is then used by a correlation based decision making mechanism to achieve texture segregation. More recently, a model of two-stage linear filtering and decision making has been proposed by Sutter *et al.* [86].

Relations to Wold-based Texture Modeling

First and second directional derivatives have been used as the set of linear filters in early vision models [9][10]. However, the choice of the filters is not critical [10]. To gain some insights into the relations between Wold-based modeling and early vision models, the use of Gabor filters is considered here.

Gabor functions are essentially Gaussian functions modulated by complex sinusoids. In two dimensions, they take the form

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} e^{j2\pi\xi_0 x}, \tag{4.12}$$

where ξ_0 is the sinusoid frequency. A set of Gabor filters in various sizes, shapes, and orientations can be constructed by choosing the value of the parameters σ_x, σ_y, ξ_0 and applying appropriate amounts of spatial rotations. The Fourier transform of the Gabor function in (4.12) is

$$G(\xi,\eta) = e^{-2\pi^2\left[\sigma_x^2(\xi-\xi_0)^2+\sigma_y^2\eta^2\right]}, \tag{4.13}$$

which is a Gaussian function centered at the modulating frequency $(\xi_0, 0)$. The shape of this Gaussian function is determined by the values of σ_x and σ_y. Since a spatial rotation corresponds to a frequency rotation by the same angle, the location of the Gabor Gaussian function in the 2-D frequency plane is determined by the filter orientation and modulating frequency. Therefore, a set of Gabor filters can be built to cover the entire 2-D frequency range.

Now consider an input image with spectral harmonic peaks. The Gabor filters tuned to these peak frequencies will give large outputs while other filter outputs are relatively small. Therefore, the energy of the overall filter outputs concentrates on the periodic component of the image. Similarly, the filters should also respond strongly to image evanescent components. Hence, the behavior of the early vision model when using Gabor filters is consistent with the Wold-based modeling, which emphasizes the separate characterization of deterministic and indeterministic image components. This analysis also supports the use of deterministic energy ratio as a measure of pattern periodicity.

Early Vision Models and Texture Similarity

Although the existing early vision models can shed some light on the problem of judging texture similarity, the similarity comparison is a quite different visual task from texture segregation. Firstly, more complex usage of multiple texture properties may be required. This observation is also made by both Tamura [87] and Amadasun [2]. Secondly, when more than two patterns are under consideration, the task for the computational model is not only determining whether the patterns are different, but also how different. Therefore, after the relatively low-level and simple linear filtering stage, a more sophisticated process of information aggregation and quantization is most likely involved. Much research effort is needed to reach a better understanding of texture similarity comparison in humans.

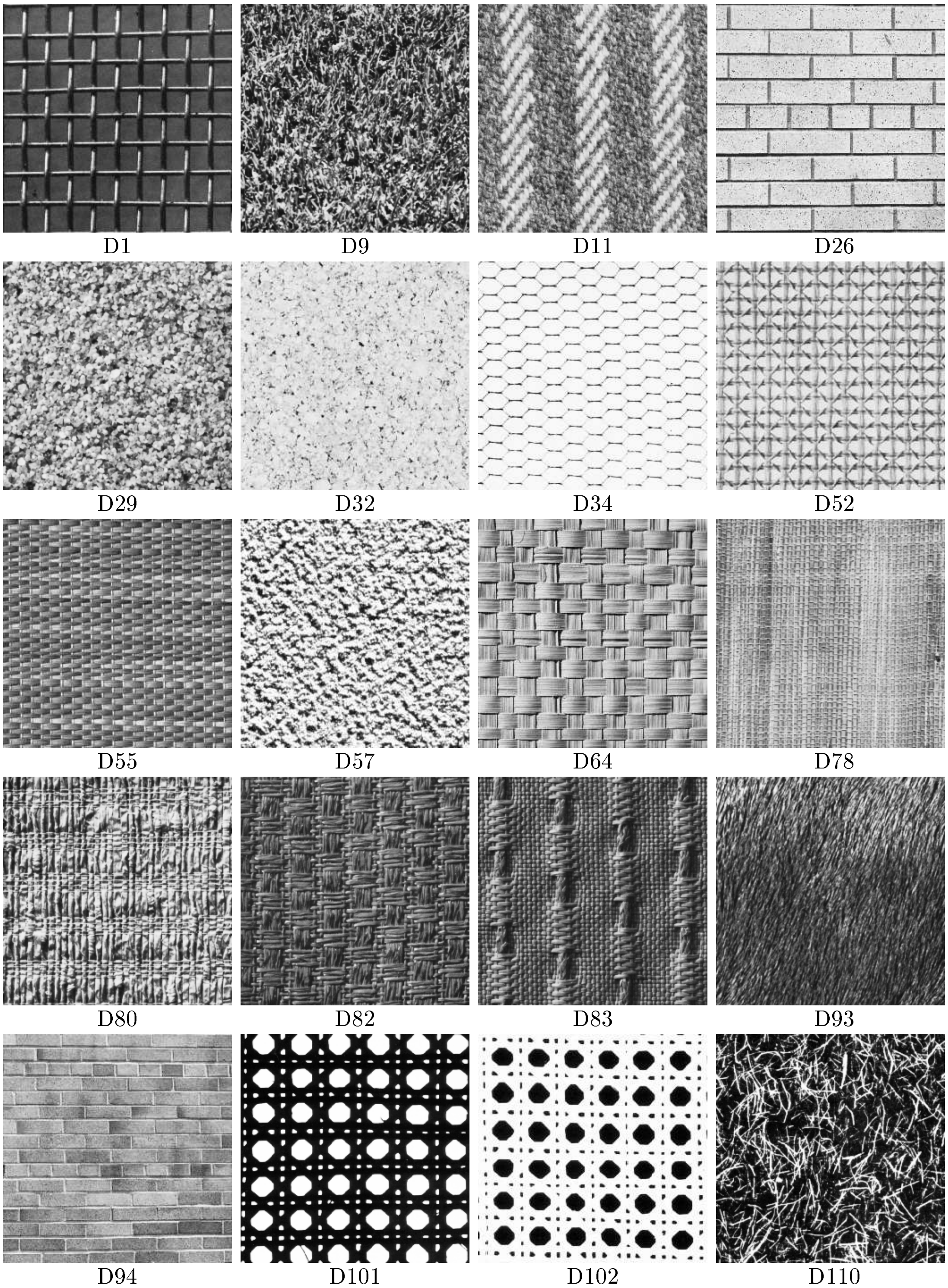


Figure 4-2: Brodatz texture test samples used in the perceptual study.

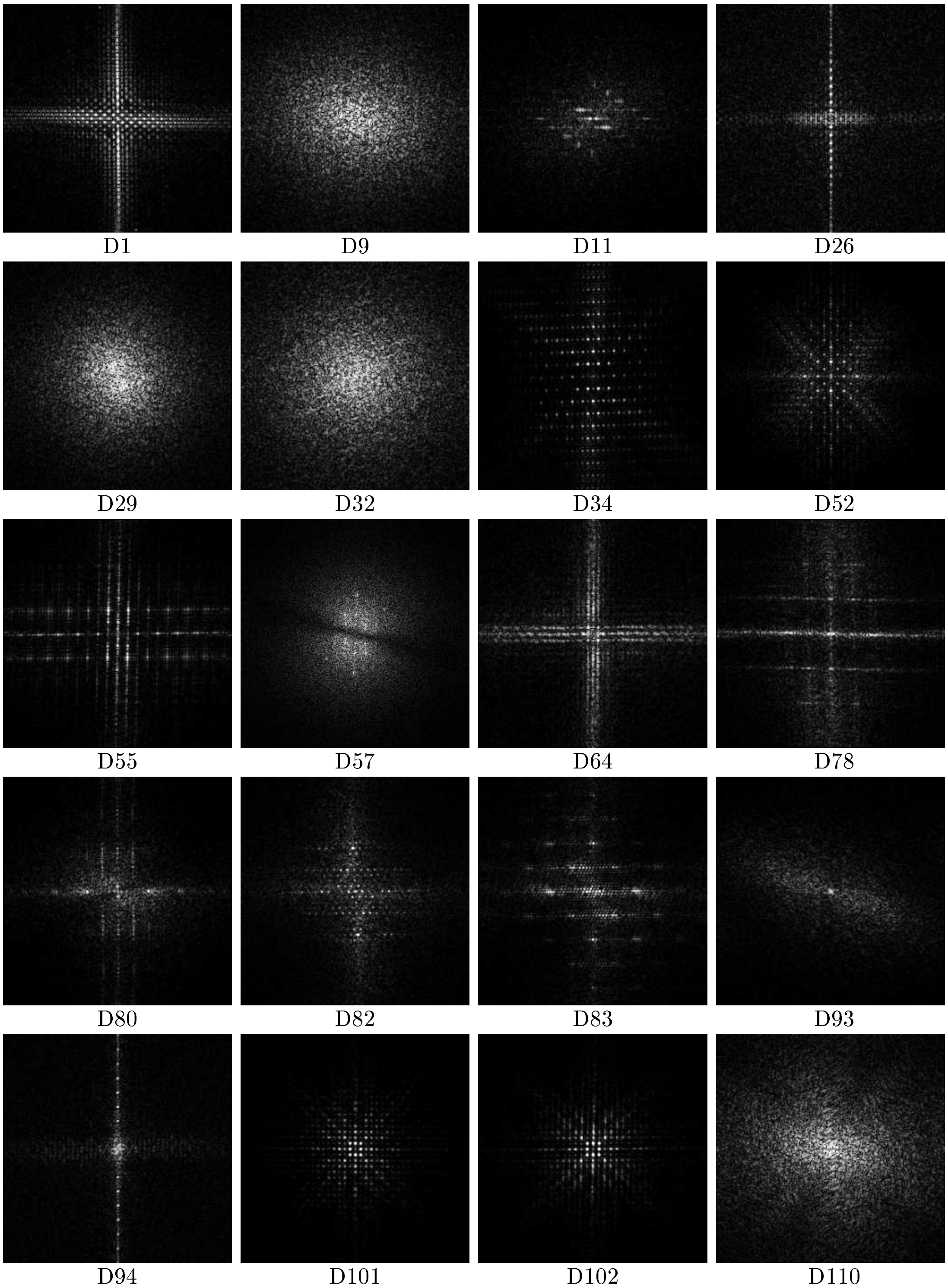


Figure 4-3: Test sample Fourier magnitudes.

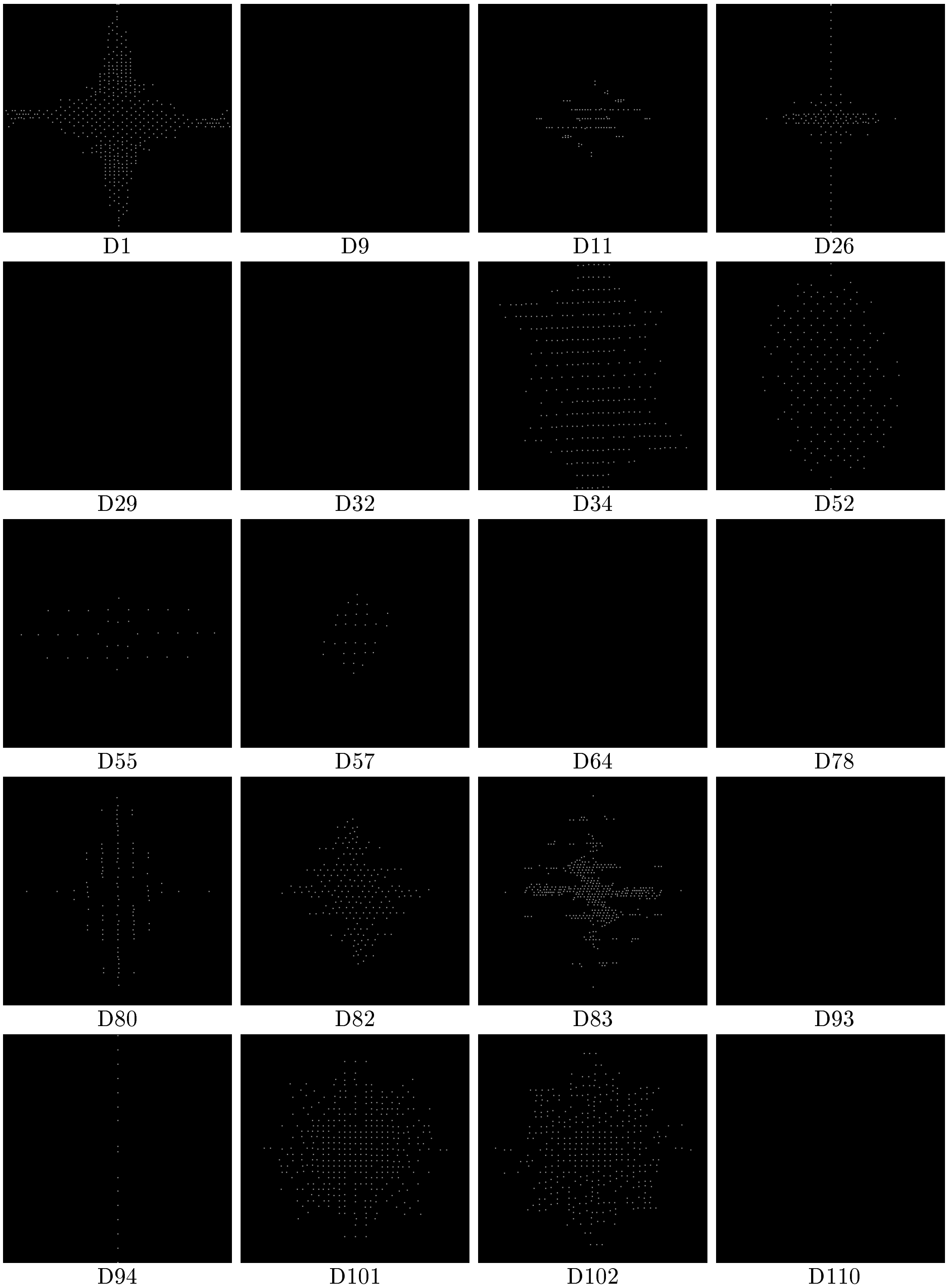


Figure 4-4: Test sample spectral harmonic peak frequencies.

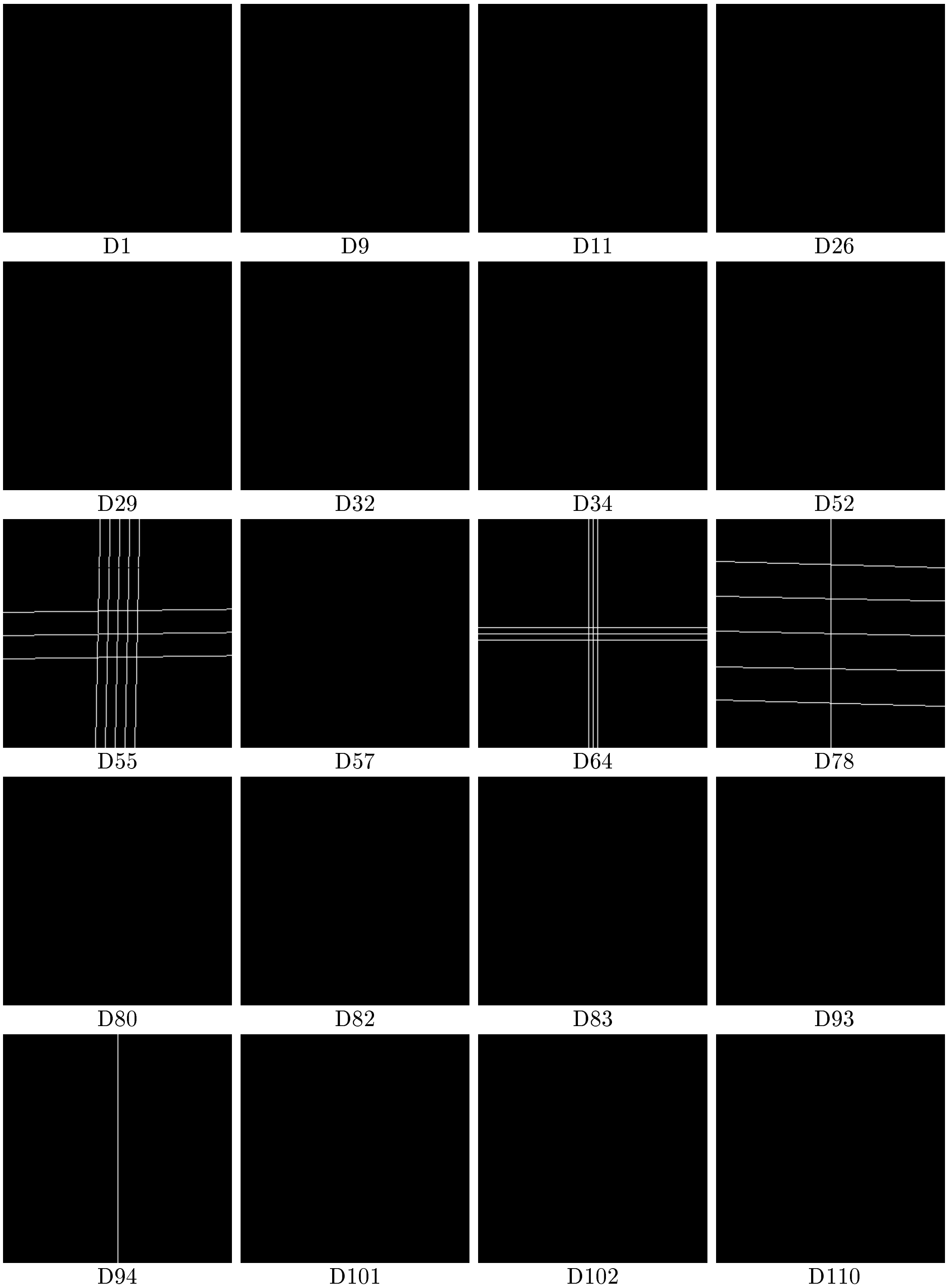


Figure 4-5: Test sample spectral evanescent lines.

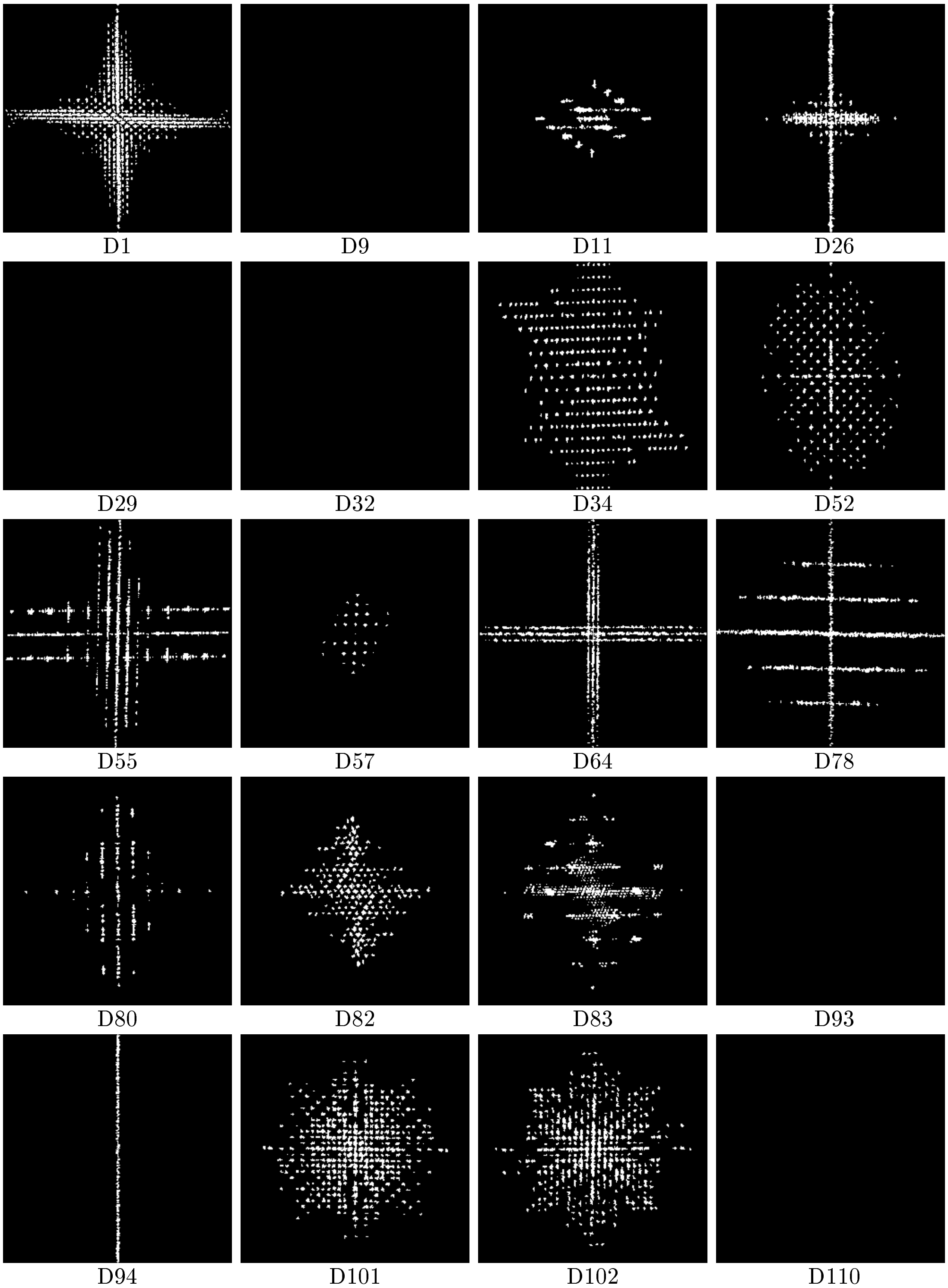


Figure 4-6: Test sample deterministic frequencies.

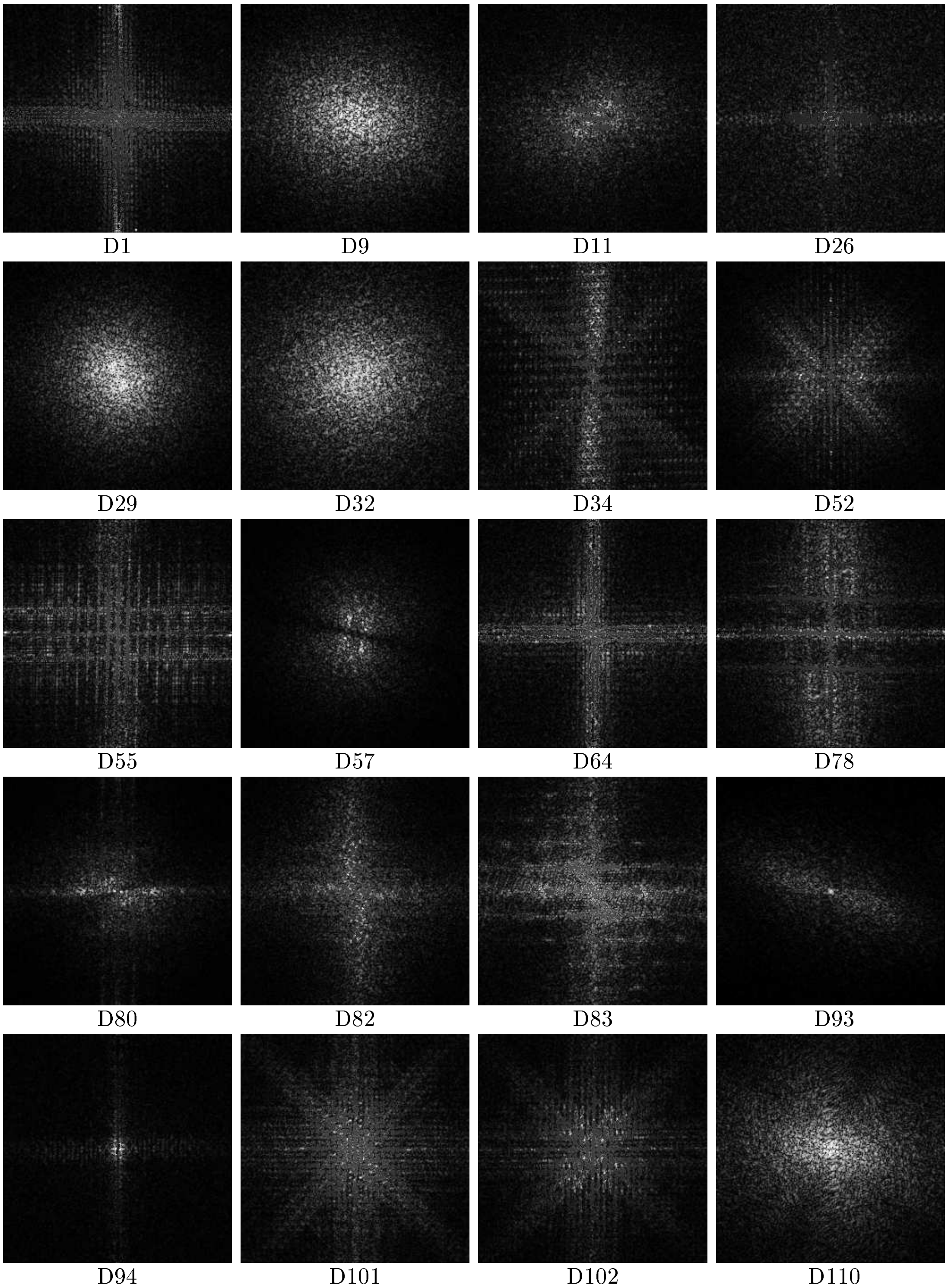


Figure 4-7: Test sample remaining Fourier magnitudes after the deterministic frequencies are extracted. The images are individually scaled to the display range of $[0, 255]$.

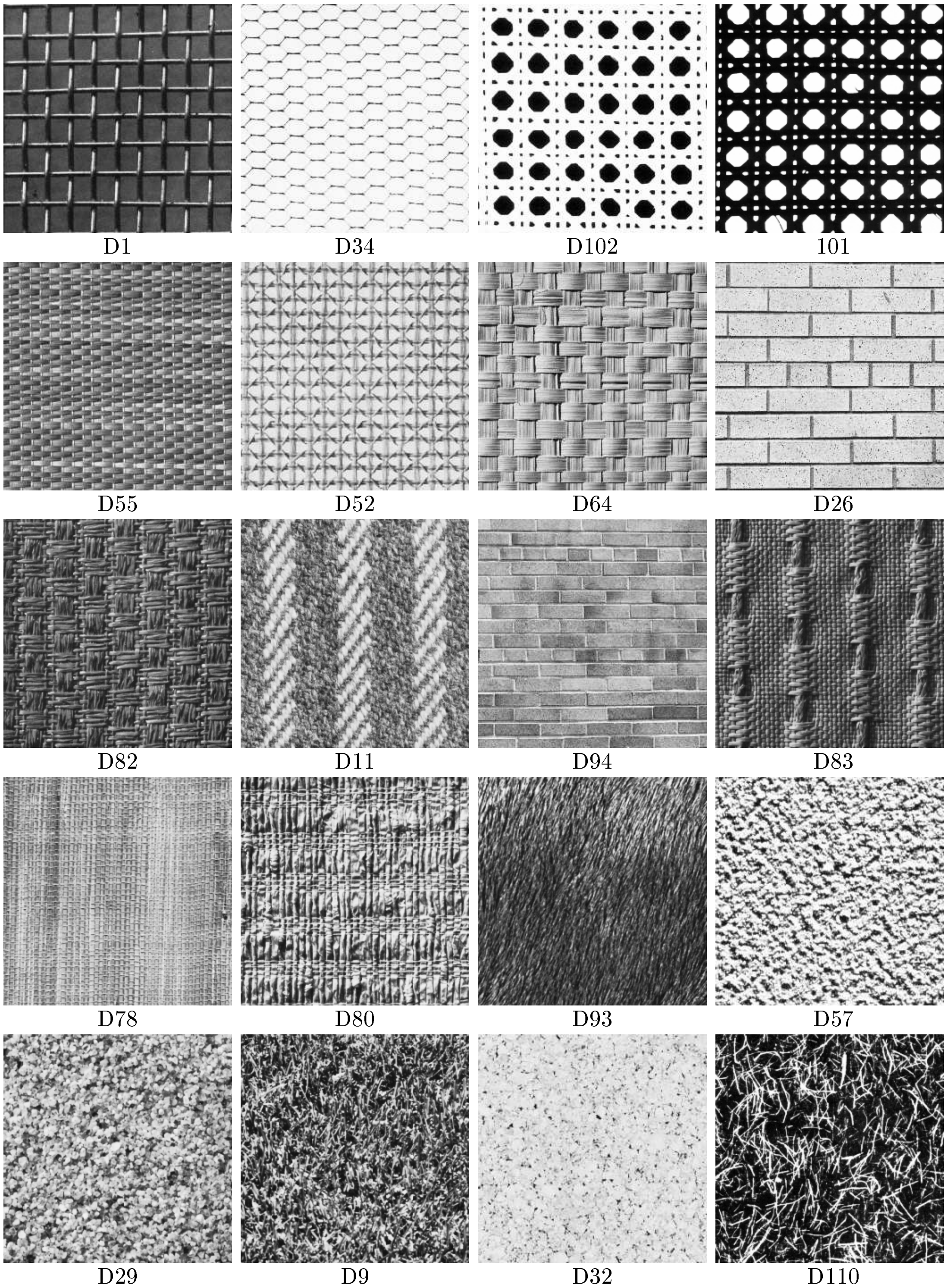


Figure 4-8: Test samples displayed in final human ranking order. From left to right, top to bottom, the images are from the most repetitive to the most random.

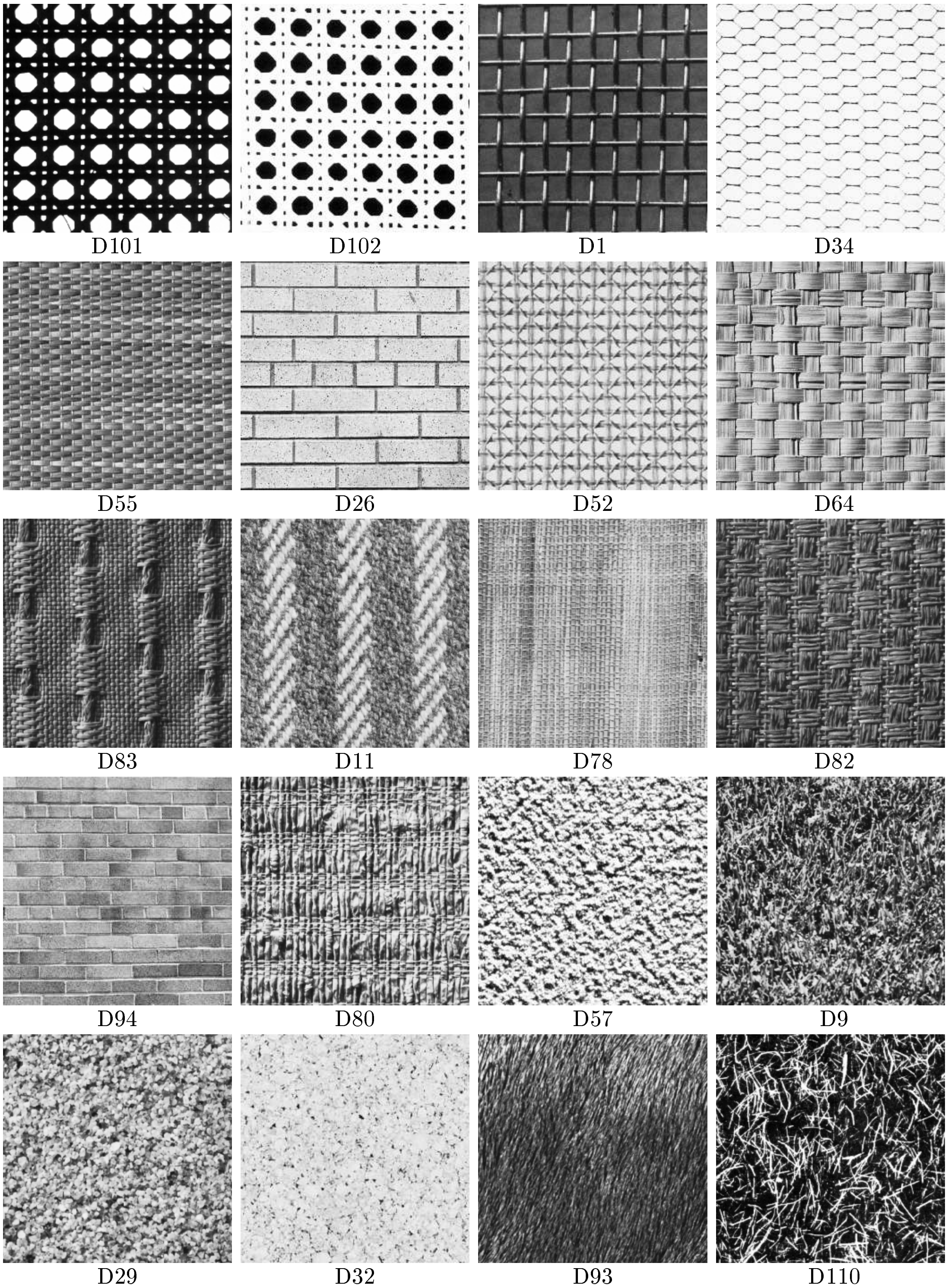


Figure 4-9: Test samples displayed in computer ranking order. From left to right, top to bottom, the images are from the most repetitive to the most random.

Chapter 5

Image Database Retrieval

5.1 Introduction

Current worldwide efforts of digitizing massive archives of image, film, and video have created an immediate demand for automated retrieval systems. Tools assisting search among texture-rich imagery have broad applications in, to name a few, video editing, medical image query, and commodity markets such as carpet, tile, and upholstery.

A retrieval system serves the purpose of saving human users the time and effort of browsing the entire database; hence, it is expected that the retrieved images resemble the visual properties of the prototype pattern provided by the human user. To build such a system, it is important that the computational features used for pattern comparison are faithful to those used by humans in comparing patterns. As shown previously, the perceptual properties of the image Wold components can be described as "periodicity", "directionality", and "randomness", agreeing closely with that of the top dimensions of human texture perception. Hence, perceptually salient features can be constructed based on the Wold theory.

Francos *et al.* [26][27][85] applied the 2-D Wold decomposition to spectral estimation and texture modeling. In their work, it is assumed that the images are homogeneous random fields and the model designs are based on the actual image decomposition. The proposed algorithms performed well on a few texture examples, but are not robust or computationally efficient enough to handle databases where image quantity is large and inhomogeneity abounds.

In this chapter, a Wold-based texture model is developed and shown to be robust in textured image database retrieval and natural scene representation applications. The emphasis of the model construction is on providing perceptually salient features for image recognition and similarity comparison. The model computational features, which preserve the perceptual properties of the Wold components, are extracted without decomposing each image. This model design eases the constraint on pattern homogeneity. The new texture model and the corresponding algorithms for image similarity comparison can tolerate a variety of pattern inhomogeneities, as well as transformations such as pattern rotation and scaling. The problem of aggregating different types of features for image similarity comparison is resolved by using a Bayesian probabilistic approach.

The effectiveness of the Wold model for natural texture modeling is demonstrated in image retrieval experiments in comparison to the performance of two other well-known pattern recognition methods, namely, the shift-invariant principal component analysis (SPCA) [75] and the multiresolution simultaneous autoregressive (MRSAR) [66] modeling. The Wold model appears to offer a perceptually more satisfying measure of pattern similarity while matching the best performance of these other methods by traditional pattern recognition criteria.

To illustrate how the Wold features can be used in natural scene representations, an image segmentation algorithm and experimental segmentation and representation results are also presented.

This chapter is organized as follows. Section 5.2 introduces the texture database used in the retrieval experiment. Section 5.3 contains a brief discussion of the image retrieval performance of several existing texture models. The Wold-based texture model is constructed in Section 5.4, and then applied to image database retrieval in Section 5.5. Section 5.6 demonstrates Wold texture modeling of natural scene images. A discussion of the strengths and weaknesses of the Wold-based model is in Section 5.7, followed by several conclusions.

5.2 Brodatz Texture Database

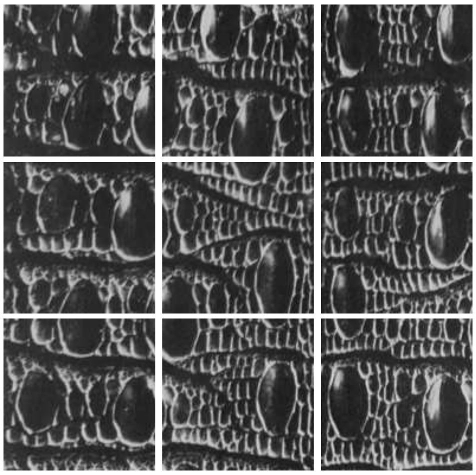
The "Brodatz texture database" contains 1008 natural texture patches, cropped from all 112 pictures in the Brodatz album [15]. Each Brodatz texture provides nine 128×128 non-overlapping sub-images in 8-bit gray levels. This collection contains a large variety of natural textures, including the many inhomogeneous ones which are not usually included in texture studies. Including the entire Brodatz collection in the database allows the potential of confusion and failure that exists when texture algorithms encounter non-texture regions in natural scenes. Examples of the database are shown in Figure 5-1.

5.3 Other Texture Models

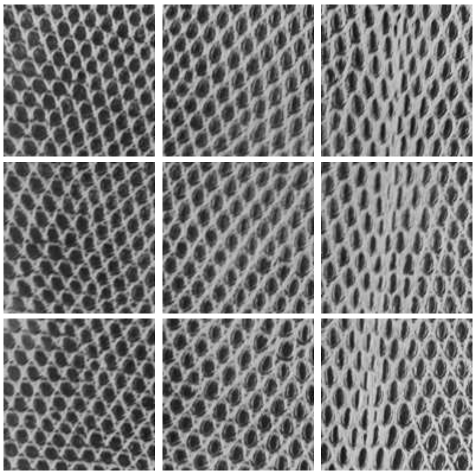
Using the benchmarking method reported in [76], the retrieval performance of several image models over the Brodatz database was evaluated by computing their recognition rate operating characteristics. The image classes are defined by the original Brodatz album pages. Using each image in the database once as a retrieval prototype, the average recognition rate is computed for different numbers of the top retrieved images. A 100% recognition rate is reached by a search when 8 matches are found within the top retrieved images considered. For example, if the first 15 retrieved images are considered and 4 matches are found for an image, then the recognition rate for that image at retrieved set size 15 is 50%. The models evaluated include the MRSAR, the SPCA, the tree-structured wavelet transform (TWT) [1] [17], and the three Tamura features of coarseness, contrast, and directionality [87] as used in [69]. Note that this evaluation method uses a traditional pattern recognition criterion, not necessarily agreeing with perceptual criteria.

The benchmarking results show that, when compared to the other three, the MRSAR model offers the best intra-class recognition rate (see Figure 5-13 in Section 5.7.3). Recently, a Gabor wavelet decomposition model was also applied to image retrieval and its performance benchmarked against the MRSAR model [65]. By the recognition rate operating characteristics, the retrieval performance of the Gabor and MRSAR methods are similar. Therefore, by this criterion, it would be reasonable to regard MRSAR as representative of the state-of-the-art texture modeling for image database retrieval. However, in many retrieval cases where structured image patterns are

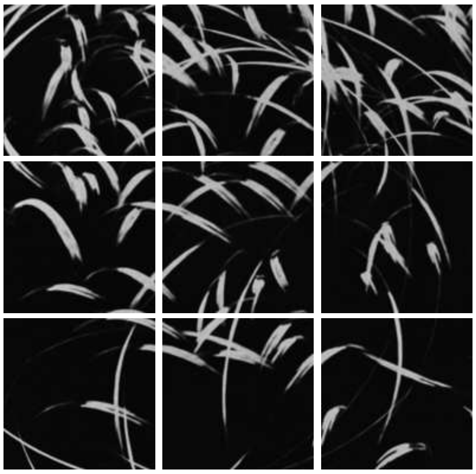
[1] The TWT method is sensitive to image sizes. Much smaller than the 512×512 used in [17], the 128×128 database image size had a negative impact on the TWT performance in the benchmarking.



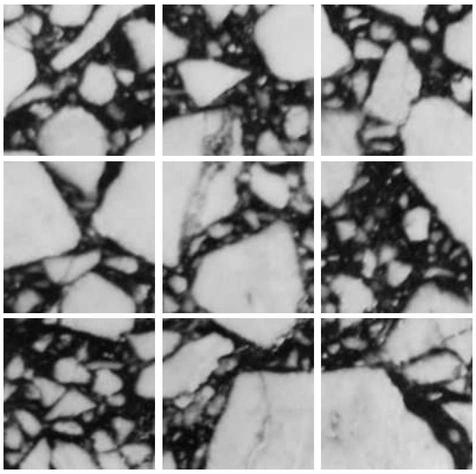
(a) D10: Crocodile Skin



(b) D36: Lizard Skin



(c) D45: Swinging Light



(d) D62: European marble

Figure 5-1: Example images of Brodatz texture database. Each original album picture contributes to the database nine 128 by 128 grey scale images.

---

involved, it is observed that the MRSAR model is incapable of distinguishing images with very little perceptual resemblance, showing its limitations in measuring perceptual similarity. Examples are shown in Section 5.5.3. This weakness of the MRSAR model is innate since the model only characterizes the interaction among neighboring image pixels, where neighbors are determined by the model order. As an autoregressive (AR) process, the MRSAR model is most appropriate for modeling random fields with continuous spectra (fine and purely random texture). When using an AR process to model an image with many spectral peaks (spatially periodic structures), it is often difficult to avoid both the information loss inherent in fitting with a low-order model and the extra computation and over-fitting with a higher-order model.

5.4 A Wold-based Texture Model

5.4.1 Model Construction

Perceptually, by Rao and Lohse's study [79], the existence of periodic structure is the strongest perceptual cue in texture discrimination. A careful examination of all Brodatz database images and their Fourier spectra reveals the following:

- Natural textures often contain multiple Wold components. Perceptually structured textures usually have dominant harmonic components which appear as structured spectral peaks. Conversely, when the harmonic components are significant, they usually dominate the perceptual pattern discrimination.
- Although certain local inhomogeneities (such as texture on an uneven surface or viewpoint distortion) spread out or change the frequencies of the spectral peaks slightly, the intrinsic structure of these peaks remains.
- Strong evanescent components correspond to eminent directionality in patterns; local inhomogeneities have only a minor effect on these components.

The distinct spectral signatures of some Brodatz database textures were shown earlier in Figure 1-1. Considering the observations above, the Wold-based texture model is designed to first conduct a "harmonicity test" on an image. This test provides a measure of the confidence that the image can be characterized as highly structured (or relatively unstructured). Based on this measure, either harmonic peak feature extraction or MRSAR fitting, or both, are deployed. The final Wold representation of the image contains the harmonic confidence measure and the corresponding harmonic peak features and MRSAR features.

The construction of the new model emphasizes the perceptually most salient harmonic information. It also incorporates the demonstrated robustness of the MRSAR model. The new model avoids the decomposition of images. The knowledge of harmonic and indeterministic components is combined probabilistically by using the harmonic confidence measure. Details of the model are explained in the following subsections.

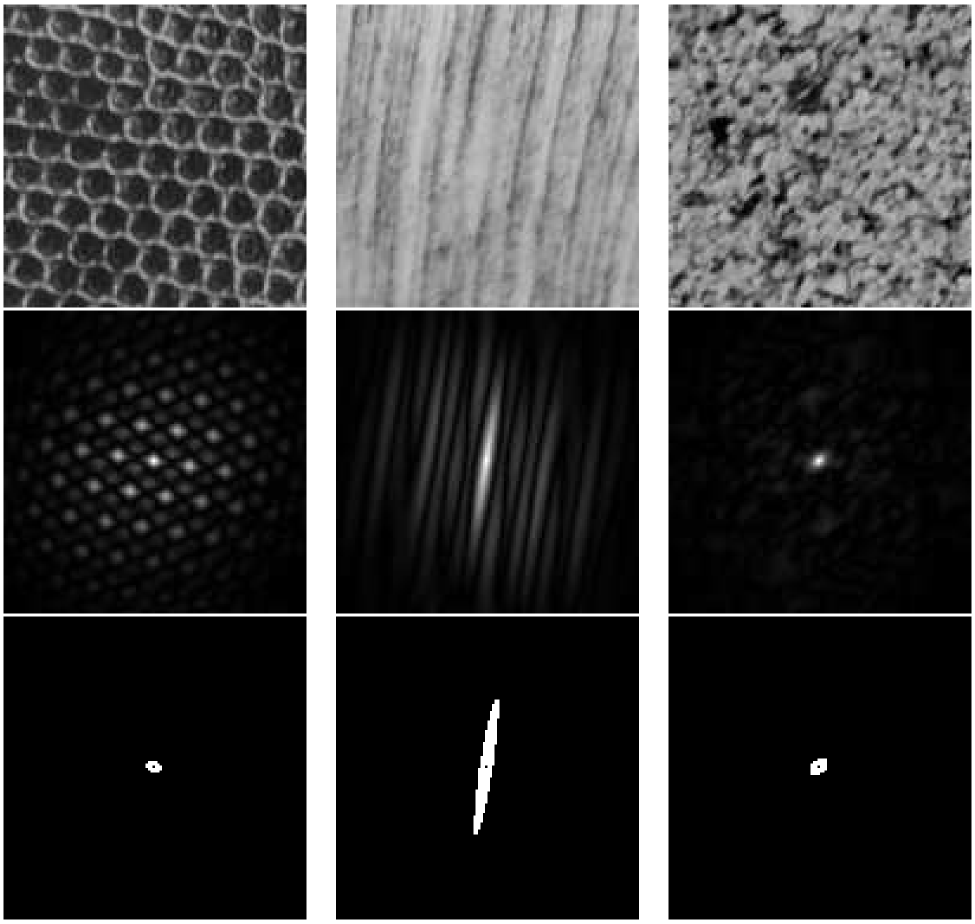
5.4.2 Harmonicity Test

Autocovariance Energy Ratio r_e

To determine the prominence of harmonic structures in a texture, the energy distribution of the image autocovariance function is examined. Given an $N \times N$ image $y(m,n)$, $(m,n) \in \mathcal{D}$ ($\mathcal{D}$ is defined by Eq. (3.1)), its autocovariance function $r_y(m,n)$ can be computed as the inverse Fourier transform of its Fourier magnitude squared:

$$r_y(m,n) = \begin{cases} \dfrac{1}{N^2} \displaystyle\sum_{k=0}^{N-1} \sum_{l=0}^{N-1} |Y(k,l)|^2 \, e^{j\frac{2\pi}{N}mk} e^{j\frac{2\pi}{N}nl}, & (m,n) \in \mathcal{D} \\ 0, & otherwise \end{cases} \tag{5.1}$$

where $Y(k,l)$ is defined in Equation (3.2)



(a) (b) (c)

Figure 5-2: Distinct autocovariance energy distribution of some Brodatz database textures. From top row to bottom: the originals; the absolute value of autocovariance functions; and the small displacement regions. (a) D3: Reptile skin: with periodic energy concentration in the entire displacement plane. (b) D69: Wood grain: with more small displacement energy. (c) D29: Beach sand: with most energy gathered in small displacement region.

As shown in the top two rows of Figure 5-2, the autocovariance energy of a highly structured texture is concentrated periodically throughout the 2-D displacement plane. In contrast, the autocovariance energy of a random-looking texture concentrates in a small displacement region. The ratio between the autocovariance "small displacement energy" (defined below) and its total energy (total sum of the absolute value of the function) can be used as an indication of the image harmonicity. (The autocovariance value at the zero displacement is ignored.) This ratio is denoted as r_e.

	ω_h	ω_r
mean	6.3579	43.6936
variance	10.0241	295.6701
prior	0.1270	0.8730

Table 5.1: Parameters of two Gaussian classes fitted to the autocovariance energy ratio data.

An image is first zero-meaned and Gaussian tapered using the tapering window function in Figure 3-2 (b). This window function is used in all image tapering in this chapter. The image autocovariance is computed as the inverse DFT of the image power spectrum. Starting from the zero displacement, a region is grown outwards continuously until the value of the autocovariance function is lower than a small portion of the function range (10% in the experiments). This region is regarded as the small displacement region. Examples are shown in the bottom row of Figure 5-2. The energy in this region is used as the "small displacement energy".

Harmonic Confidence Measure

The autocovariance energy ratio r_e is computed for each image in the Brodatz database. The histogram of these ratios has a bi-modal structure. Gaussian assumptions are made to model the energy ratio data using an expectation and maximization (EM) procedure. (See Appendix C.1.) Denote the resulting classes as ω_h (harmonic) and ω_r (random). The EM algorithm gives the means and variances of the Gaussian conditional probability density functions of r_e, denoted as $p(r_e|\omega_h)$ and $p(r_e|\omega_r)$, and the prior probabilities, denoted as $P(\omega_h)$ and $P(\omega_r)$. The estimated parameters are listed in Table 5.1. The joint probability density functions $p(r_e, \omega_h) = p(r_e|\omega_h)P(\omega_h)$ and $p(r_e, \omega_r) = p(r_e|\omega_r)P(\omega_r)$ are plotted in Figure 5-3, together with the energy ratio histogram.

Given the autocovariance energy ratio r_e of an image, the posterior probability of ω_h can be computed as

$$P(\omega_h|r_e) = \frac{p(r_e,\omega_h)}{p(r_e)} = \frac{p(r_e,\omega_h)}{p(r_e,\omega_h)+p(r_e,\omega_r)} = \frac{p(r_e|\omega_h)P(\omega_h)}{p(r_e|\omega_h)P(\omega_h)+p(r_e|\omega_r)P(\omega_r)}. \tag{5.2}$$

This probability is then used as the confidence measure of characterizing the image as highly structured. Consequently, the confidence of describing the image as relatively unstructured is

$$P(\omega_r|r_e) = 1 - P(\omega_h|r_e). \tag{5.3}$$

For a given image, the values of $P(\omega_h|r_e)$ and $P(\omega_r|r_e)$ determine what feature sets are computed. By the property of Gaussian functions, any value of r_e gives non-zero posterior probabilities. To save computation and storage, values of $P(\omega_h|r_e)$ and $P(\omega_r|r_e)$ smaller than 0.001 are considered insignificant and set to zero (for about 5% of Brodatz database images). Corresponding to the non-zero $P(\omega_h|r_e)$ and $P(\omega_r|r_e)$, the harmonic peak features and the MRSAR features are computed respectively.

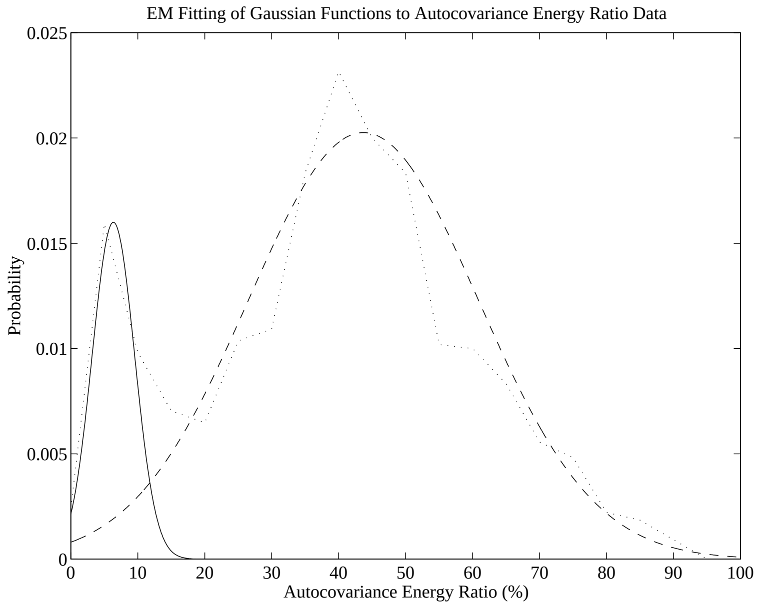


Figure 5-3: EM fitting of Gaussian density functions to image autocovariance energy ratio data. Shown are the joint probability density functions $p(r_e, \omega_h)$ (solid line) and $p(r_e, \omega_r)$ (dash line), together with the energy ratio histogram (dot line). The histogram has bin size 5% and is scaled down by a factor of 5400.

5.4.3 Features for Harmonic Structures

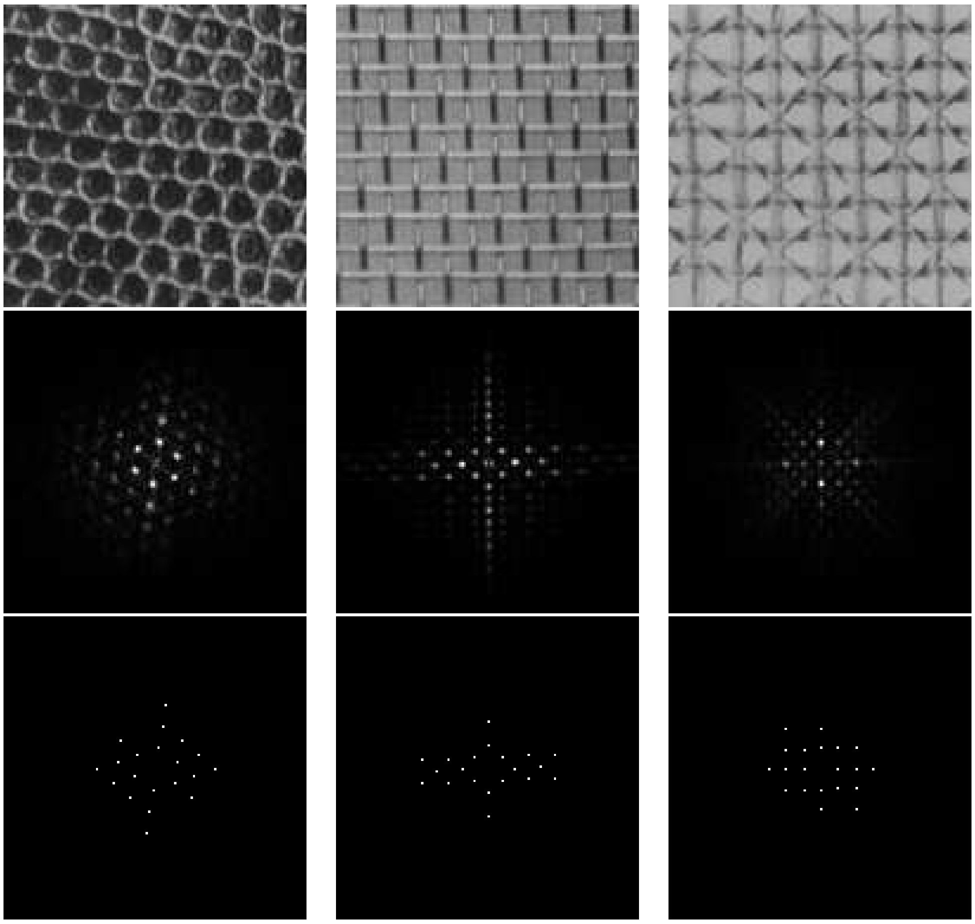
Feature Extraction

The Wold feature set characterizing the harmonic structure of an image consists of the frequencies and the magnitudes of the spectral harmonic peaks. To extract the feature set, the image is first zero-meaned and Gaussian tapered. Then the spectral harmonic peaks are detected using the method presented in Section 3.3.3. Examples of harmonic Wold features are shown in Figure 5-4. Note that it is usually not necessary to use all detected harmonic peaks for the feature sets. In this work, only the ten largest ones are kept for each image.

Feature Invariance

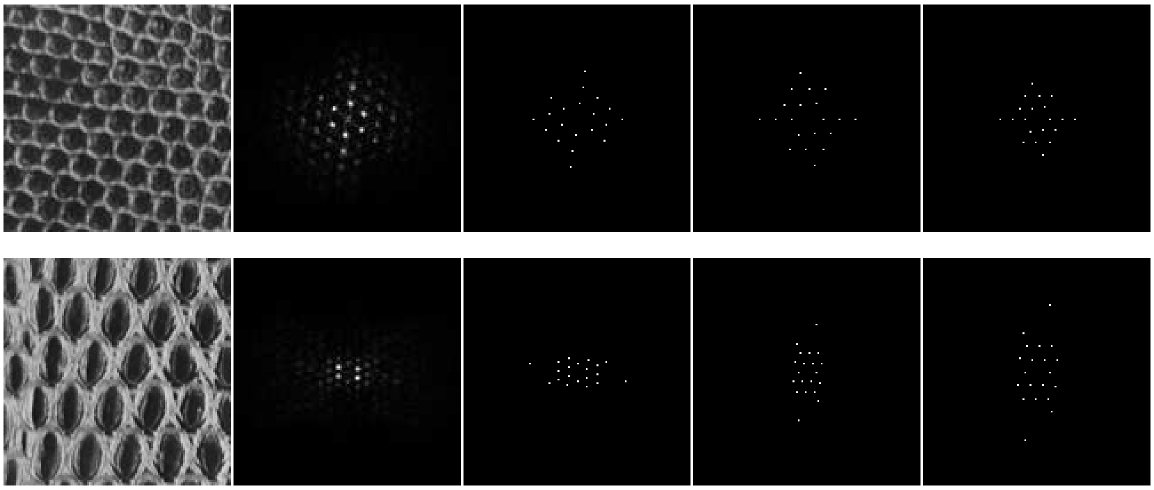
The harmonic Wold features inherit from the Fourier spectral magnitude the property of spatial shift-invariance, a property that is usually important when comparing images. It is often desirable for a retrieval system to also provide users options such as pattern comparison with respect to relative rotation and scaling.

Since the spatial relationship of the harmonic peaks in a Wold feature set does not vary under rotation, effects of relative rotation among textures may be reduced by rotating the peaks to align



(a) (b) (c)

Figure 5-4: Harmonic features of three Brodatz database textures. Each pattern contains two fundamental frequencies. From the top row to the bottom: originals; DFT magnitudes; and harmonic peak feature frequencies. (a) D3: Reptile skin. (b) D14: Woven aluminum wire. (c) D52: Oriental straw cloth.



Original Fourier Mag. Peak Feature Rotated Rot. & Scaled

Figure 5-5: Harmonic peak feature rotation, and scale invariance. Top row: D3, Reptile skin. Bottom row: D35, Lizard Skin. Although the original patterns are in different scale and have relative rotation, their harmonic peak features allow rotation and scale invariant image similarity comparison.

---

the main orientation of the texture to a chosen direction (horizontal in this work). The main orientation of a texture is defined here as the direction of the lowest fundamental frequency in the feature set. Note that this direction may not correspond to the *perceptually* most salient orientation in the image, but this does not matter for the purposes of comparing images after aligning their orientations. Aligning the peaks using the frequency with the most energy (not necessarily the lowest fundamental frequency) is not as useful since the energy distribution can be influenced by many non-pattern attributes, such as local lighting and contrast. Since each feature set typically consists of a small number of peaks, its rotation involves minimal computation compared to a rotation in the spatial domain.

In a similar manner, the harmonic Wold features can be made scale invariant by scaling the 2-D frequency values of each peak by a factor that puts the lowest fundamental frequency at a chosen distance from the zero frequency.

Examples of the rotated and scaled harmonic peak features are shown in Figure 5-5. Although the original patterns are in different scale and have relative rotation, their harmonic peak features allow rotation and scale invariant image similarity comparison.

Harmonic Peak Matching

In the image retrieval applications considered here, the user selects a **prototype image** and the retrieval algorithm searches through the database **test images** for the ones that are similar to the prototype. The comparison of the texture harmonic structures is carried out by matching the Wold feature sets. Denote the peak feature magnitude values of a prototype and a test image by $m_p(s)$ and $m_t(r)$ respectively, where $s = (s_1, s_2), r = (r_1, r_2) \in \mathcal{D}_h$. As defined in Equation (3.6), region $\mathcal{D}_h$ is half of the discrete frequency plane. The harmonic pattern similarity between the two images is measured as:

$$M_{pt} = \sum_{s \in \mathcal{D}_h} m_p(s) \sum_{r \in \mathcal{D}_h} w_p(r-s) \frac{m_p(s) m_t(r)}{[m_p(s) + m_t(r)]^2}, \tag{5.4}$$

where $w_p(\cdot)$ is a point spread weighting function, implemented here as a 5×5 (size found empirically) Gaussian mask with unity at the center and standard deviation $\sigma = \sqrt{5}$. This function enables peak matching within a small neighborhood of the prototype peaks. This not only compensates for the frequency sampling effects of the DFT operation, but also tolerates small frequency shifts of the harmonic peaks caused by inhomogeneities in the data. The function of the ratio term is to weigh the difference of the peak magnitudes since quantity

$$\frac{m_p(s)}{m_p(s) + m_t(r)} \cdot \frac{m_t(r)}{m_p(s) + m_t(r)}$$

reaches its maximum when $m_p(s) = m_t(r)$. Note that the larger the value M_{pt}, the more similar the two harmonic patterns.

5.4.4 Features for Relatively Unstructured Textures

Feature Extraction

The indeterministic component of a texture can be modeled by an AR process (Section 2.5). Various AR implementations have been applied to texture modeling. In this work, the second-order symmetric MRSAR model of Mao and Jain [66] is used.

The least-squares estimation (LSE) method is used to estimate the MRSAR model parameters. Other methods, such as the maximum likelihood (ML) estimation [53] and the 2-D Levinson type algorithm [67], can also be used. It has been shown that under the experimental circumstances similar to this work, the LSE and the ML estimates offer very similar performance [56]. The 2-D Levinson algorithm is especially useful when the model order determination is involved in the parameter estimation. Since the MRSAR modeling in this work targets the relatively unstructured patterns in an image, a fixed second-order model is chosen and the LSE estimation is used for its computational simplicity.

For every other pixel of an image, four SAR coefficients and the standard deviation of the SAR fitting error are estimated at each of the second, third, and fourth resolution levels. These parameters are then concatenated to form fifteen-parameter feature vectors. The mean and the covariance matrix of these feature vectors comprise the MRSAR feature set for the image. Details of the LSE estimation procedure can be found in Appendix C.2.

Image Comparison

Two relatively unstructured images are compared by examining the Mahalanobis distance of their MRSAR feature vectors. Let $\mathbf{f}_p$ and $\mathbf{f}_t$ be the feature vectors of the prototype and a test image respectively. Let $\mathbf{K}_p$ be the covariance matrix of the prototype feature $\mathbf{f}_p$. The Mahalanobis distance between the two images is

$$d_{pt} = (\mathbf{f}_t - \mathbf{f}_p)^T \mathbf{K}_p^{-1} (\mathbf{f}_t - \mathbf{f}_p). \tag{5.5}$$

Note that the smaller the Mahalanobis distance, the more similar the two images.

In Section 5.5.3, the results of image retrieval based solely on the MRSAR features are compared to the performance of the Wold-based model.

5.4.5 Detecting Evanescent Components

Since the spectral signatures of evanescent components are straight lines, an algorithm using the gray-scale Hough transform was developed to detect evanescent components in the frequency domain. After computing the Hough transform of the image DFT magnitudes, the histogram of line slope angles is built. The variance of this histogram and the variance of the Fourier energy along lines corresponding to the sharp peaks in the histogram are found to be discriminative features for evanescent detection. This algorithm accurately identifies the images from the Brodatz pictures D49, D105, and D106 as highly evanescent. Perceptually, these images indeed have distinctively strong directional properties.

The fact that the Brodatz database contains few strongly evanescent samples makes it impossible to statistically determine how the evanescent information should be incorporated into the modeling procedure. In this work, the evanescent database images are modeled by MRSAR processes.

5.4.6 Measuring Similarity of Textures

Using the Wold features of textures, the image similarities can be measured by either the harmonic peak matching or the MRSAR feature Mahalanobis distances. However, since the harmonic and the MRSAR features are of different types, it is an open question how the two measures should be best combined so that the resulting measure reflects the overall similarity of textures. In the context of image retrieval, the following probabilistic joint measure for image similarity is devised.

Given a prototype image, the system generates two image orderings by using the harmonic peak and the MRSAR features respectively. In each ordering, the entire database is sorted by the descending order of the image similarity to the prototype. When multiple images have the same similarity measure value to the prototype, they share the same ordering rank. For an arbitrary test image, its ordering ranks in the two orderings are typically different. Denote its rank in the harmonic ordering by O_h and the one in the MRSAR ordering by O_r. As discussed in Section 5.4.2, the posterior probabilities $P(\omega_h|r_e)$ and $P(\omega_r|r_e)$ can be used as a confidence measure of characterizing the *prototype* texture as highly structured or relatively unstructured. More specifically, these probabilities indicate the degree of belief in the two orderings. Hence, the joint rank of the test image is computed as

$$O_{joint} = O_h P(\omega_h|r_e) + O_r P(\omega_r|r_e).$$

The final similarity ordering of the database is formed by sorting images in the ascending order of their joint rank values.

As an additional benefit, it is found that, with this similarity measure, the system is less sensitive to the choices of threshold parameters (such as the 10% for the small displacement energy calculation in Section 5.4.2), while giving improved overall retrieval performance.

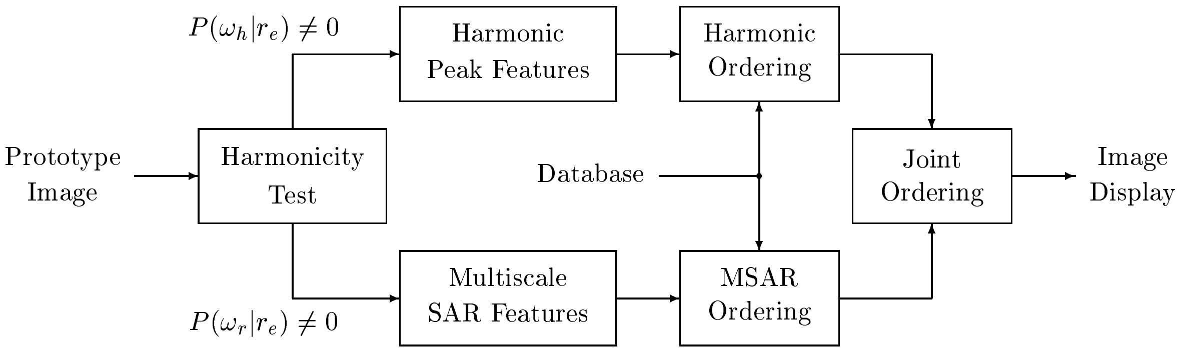


Figure 5-6: Flow-chart of image retrieval system based on the Wold texture model.

5.5 Textured Image Database Retrieval

5.5.1 Image Retrieval System

The flow-chart of the image database retrieval system is shown in Figure 5-6. This system consists of four stages. The first stage is the harmonicity test. Given a prototype image, its autocovariance energy ratio is computed to obtain the posterior probabilities $P(\omega_h|r_e)$ and $P(\omega_r|r_e)$. Probability values smaller than 0.001 are set to zero. In the second stage, corresponding to the non-zero posterior probabilities, the harmonic peak features and the MRSAR features are estimated respectively. The harmonic peaks in the feature set are rotated to align their main orientation to horizontal and scaled such that the distance between the lowest fundamental frequency and zero frequency is 10. The third stage provides database image orderings where the entire database is sorted by the descending order of the image similarity to the prototype. In each ordering, the similarities are measured by either the harmonic peak matching or the MRSAR feature Mahalanobis distances. In the final stage, different orderings are combined using the method described in Section 5.4.6 to obtain the final joint ordering.

The retrieval experiments are carried out on the Brodatz texture database using the Photobook test environment described in [72]. Parameters used to compute the posterior probabilities for a prototype image in harmonicity testing can be found in Table 5.1. The Gaussian weighting function for harmonic peak matching is shown in Figure 5-7.

Each harmonic peak feature set contains the 2-D frequencies and magnitudes of ten harmonic peaks, yielding twenty integers and ten floating-point numbers per image. A MRSAR feature set includes the 15-parameter feature vector and the 15×15 feature covariance matrix (120 distinct numbers due to symmetry), for a total of 135 floating-point numbers per image. For a 128×128 image, feature computation takes typically 0.18 second for the harmonic peaks and 38 seconds for the MRSAR features on an HP9000/735 workstation.

0.4	0.6	0.7	0.6	0.4
0.6	0.8	0.9	0.8	0.6
0.7	0.9	1.0	0.9	0.7
0.6	0.8	0.9	0.8	0.6
0.4	0.6	0.7	0.6	0.4

Figure 5-7: The non-zero region of the Gaussian weighting function $w_p(\cdot)$ for harmonic peak matching ($\sigma = \sqrt{5}$).

5.5.2 Retrieval Performance Criteria

When evaluating the image retrieval results in the experiments, two performance criteria are considered. One is *quantitative*: the nine samples from each original Brodatz texture form a class and a perfect "traditional pattern recognition performance" implies that the class members of the prototype image appear as the first eight retrieved images. The other criterion is *qualitative* and more difficult to evaluate: the retrieved images should be in the order of their perceptual similarity to the prototype image. In fact, the latter criterion is subject to cognitive and other influences, and there may not exist a unique "correct" ordering upon which all people agree. The claim that Wold-based modeling provides perceptually sensible features rely on the results of the human study reported in Chapter 4, and on the observations in the image retrieval experiment.

5.5.3 Image Retrieval Examples

In Figures 5-8 and 5-9, two examples of Wold-based image retrieval are shown together with the results given by the SPCA model and the MRSAR model. The two latter models are described and benchmarked in [76]. In each display, the upper left image is the user selected prototype image. In raster-scan order after the prototype, the retrieved images are shown by descending similarity to the prototype[2]. With pre-computation of the features, all three methods search the database in interactive-time (the search is faster than loading the images for display).

Figure 5-8 demonstrates the superior qualitative and quantitative performance of the Wold model. Here, the prototype image is straw cloth. In (a) and (b), both the SPCA and the MRSAR methods fail to find other straw cloth pictures as the most similar; they each retrieve images perceptually very different from the prototype. In (c), the Wold model provides both "intra-class" accuracy and "inter-class" similarity. It finds all eight other straw cloth patterns in the database

[2]The drawback of sequential display is that images having the same order number appear as different in their ordering.

and fills the display with other highly structured textures.

In Figure 5-9, the experiment is repeated on a prototype image of reptile skin. The results in (a) and (b) show that the SPCA and the MRSAR methods confuse the periodic reptile skin patterns and the random-looking paper fiber or cork patterns. In (c), the Wold method retrieves large number of reptile skin images up front, exhibiting its robustness to rotation, scale, and other image local inhomogeneities.

In both examples, the Wold-based method uses largely the harmonic information in the textures ($r_e = 6.36\%$ and 5.52%, $P(\omega_h|r_e) = 0.893$ and 0.900). This is consistent with the fact that both prototype images contain prominent periodic structures.

5.6 Natural Scene Representation

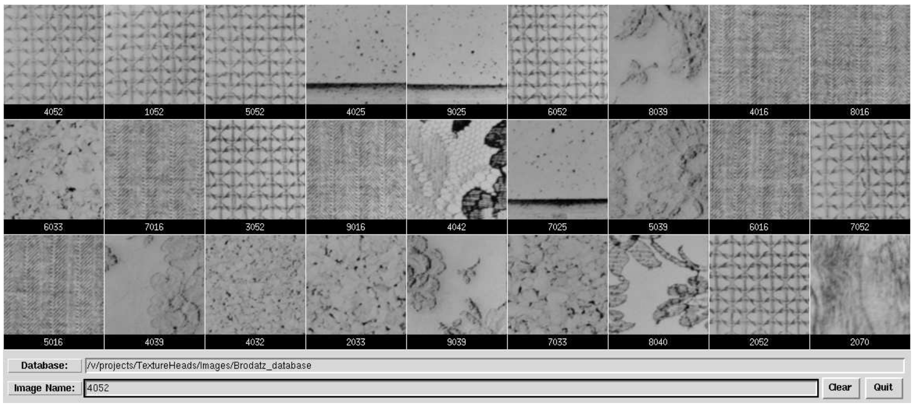
This section demonstrates how to generate descriptions for textured regions of natural scenes in terms of Wold features. A scene image is first segmented by using its MRSAR features and a K-means-based clustering algorithm. The Wold features are then extracted for the segmented image regions.

5.6.1 Textured Region Segmentation

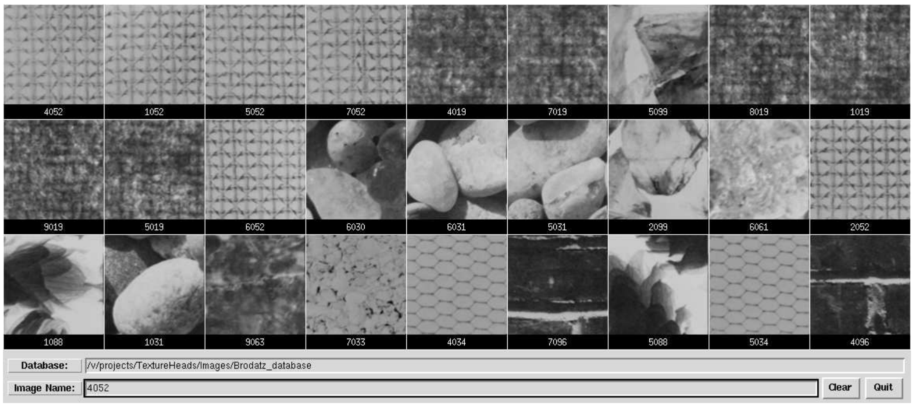
Numerous image segmentation methods have been proposed for various tasks [39][32][81]. While the common practice is to partition the entire image, the focus here is to detect and segment sizable and relatively homogeneous regions in a scene. Note that precision of region boundaries is not a primary concern in representing natural scene contents for retrieval; it is more important to extract features that provide a basis for subsequent content identification.

An unsupervised segmentation algorithm is developed to find reasonably homogeneous image regions. The algorithm is robust to slight inhomogeneities due to perspective viewpoint and uneven textured surfaces. Smooth regions (small variations in pixel values) are first detected by thresholding the local variances at each pixel in a 9×9 neighborhood. These regions are useful for retrieval requests such as "find pictures with a patch of sky at upper left". The main segmentation algorithm is a K-means-based clustering of image pixels in MRSAR feature space. Pixels in smooth regions are excluded from this procedure since the LSE estimates of their MRSAR coefficients are unreliable due to the under-determined linear equations.

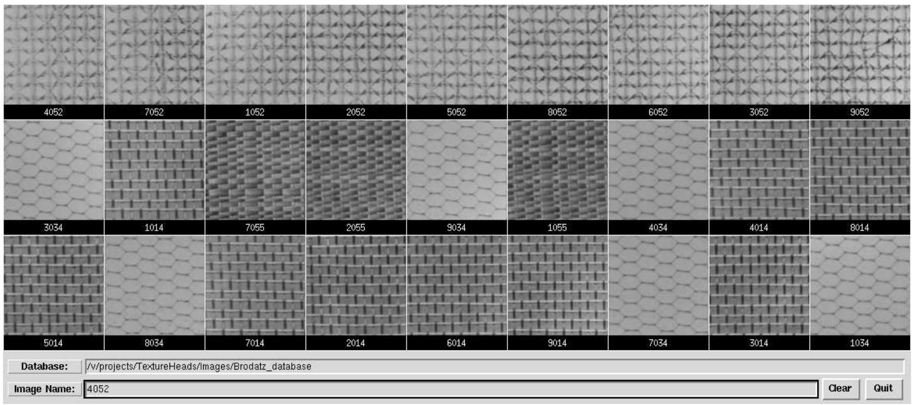
The pixel MRSAR features are computed in the same manner as described in Section 5.4.4. To initialize the clustering algorithm, the image is tessellated into rectangular regions (64×64 squares on 256×384 8-bit gray scale images in the experiments below). In a typical iteration, the Mahalanobis distances of each pixel to every cluster are computed and the pixel is re-assigned to the nearest cluster. Small clusters (less than 4000 pixels) are eliminated and their members re-assigned. Clusters are merged when their mutual Mahalanobis distance is small. The program terminates after a given number of iterations or when no pixel changes its cluster membership in an iteration. One morphological closing [80] operation is applied to the segmentation output to smooth the boundaries. The circular structuring elements used in the two examples below have diameters 15 and 30 pixels respectively.



(a)

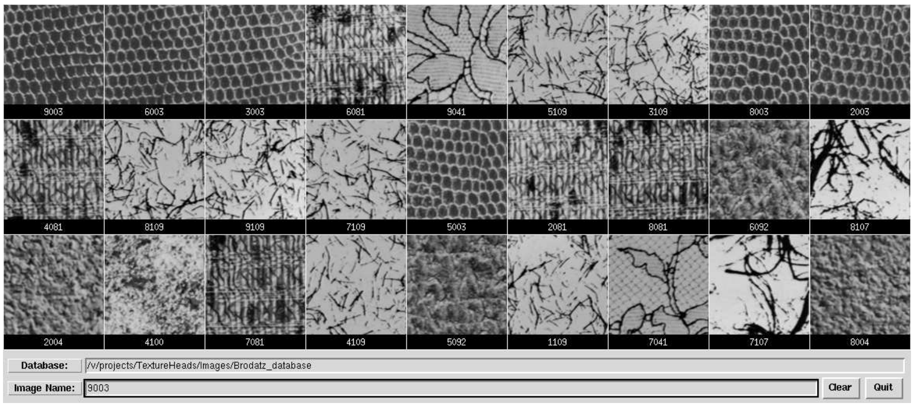


(b)

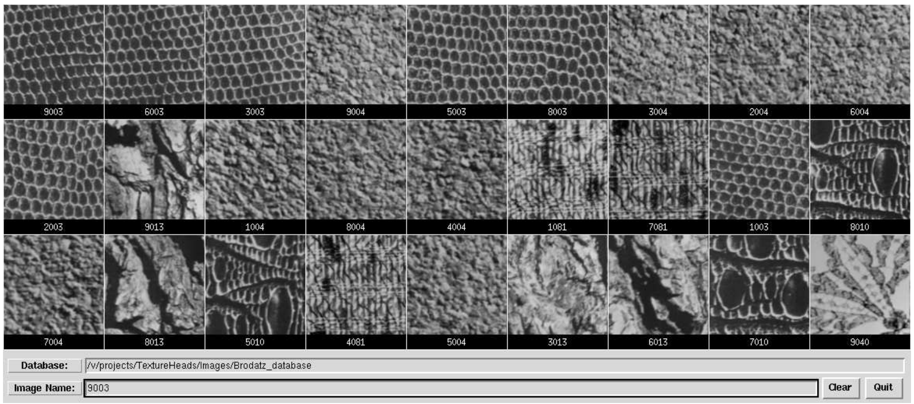


(c)

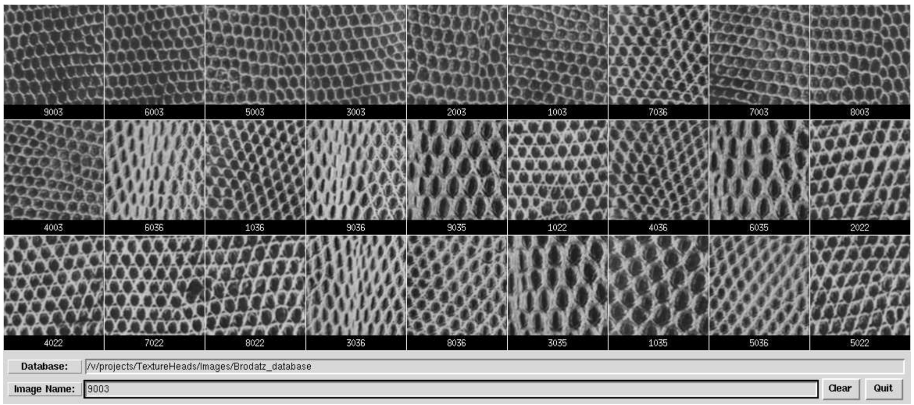
Figure 5-8: Image retrieval of the straw cloth pattern comparing three methods: (a) SPCA, (b) MRSAR, and (c) Wold. In each picture, the images are raster-scan ordered by their similarities to the image in upper left.



(a)



(b)



(c)

Figure 5-9: Image retrieval of the reptile skin pattern comparing three methods: (a) SPCA, (b) MRSAR, and (c) Wold. In each picture, the images are raster-scan ordered by their similarities to the image in upper left.

5.6.2 Natural Scene Representation Examples

Figures 5-10 and 5-11 show two examples of textured region segmentation and representation in natural scenes. In both figures, the K-means-based segmentation results are shown with smooth regions marked in black. The number of iterations used in clustering are 15 and 30 respectively for the two images.

The example shown in Figure 5-10 illustrates the segmentation and representation of a city scene. The segmented building is shown in (c). The autocovariance energy ratio of the building is $r_e = 11.88\%$ ($P(\omega_h|r_e) = 0.489$), hence the region should be represented by both the harmonic peak and the MRSAR features. The DFT magnitudes and the harmonic peak features of the building patch are shown in (d) and (e). In computing the DFT, a 128×128 Gaussian window ($\sigma = 24$) is applied to the center of the building. The harmonic peak extraction in this example shows the robustness of the algorithm to inhomogeneity due to perspective, even though no explicit perspective coordinate transform was included.

Note that not only does the segmentation find the building patch in the image, but also the Wold representation of the patch indicates the presence of a "highly structured region." For recognition and retrieval, this description rules out large categories of content such as "grass." If a user were browsing for city scenes, the algorithm could skip over images without any highly-structured regions.

Figure 5-11 shows a national park scene and its segmentation. Both the cliff and rock patches have no harmonic structures ($r_e > 45\%$) and hence are modeled by their MRSAR features. In addition, the cliff has a strong evanescent component which can be detected by the method described in Section 5.4.5.

5.7 Discussion

5.7.1 Image Inhomogeneity

The effectiveness of the Wold-based model presented above depends on the properties of the estimated image spectra. On the spectra estimated by the simple method (windowed periodogram) used here, the proposed image modeling and comparison system is surprisingly insensitive to small surface inhomogeneities and viewpoint changes. However, the performance of the model will be compromised when the inhomogeneities alter image spectra substantially.

One example is given in Figure 5-12. If shown to a human, the two lace pictures could be judged similar. Nevertheless, one of the lace patterns has prominent spectral harmonic peaks and the other does not. The reason is that the netting pattern in (c) is not homogeneous enough to form strong peaks in its spectra, nor does the netting cover enough area of the image to reinforce the weak periodicity that is present. Instead, the high contrast flowers in (c) overwhelm the harmonic component in spectra. However, a human viewer seems to "homogenize" the netting, and considers the two lace pictures to be similar.

5.7.2 Perspective Transformation

Image spatial perspective transformation is a special kind of image inhomogeneity. For a structured pattern, the perspective transformation can cause the deformation of pattern spectral harmonic peaks. An example is the building patch and its spectrum shown in Figure 5-10. Although in this case the peak detection algorithm was able to extract the peak features in spite of the peak

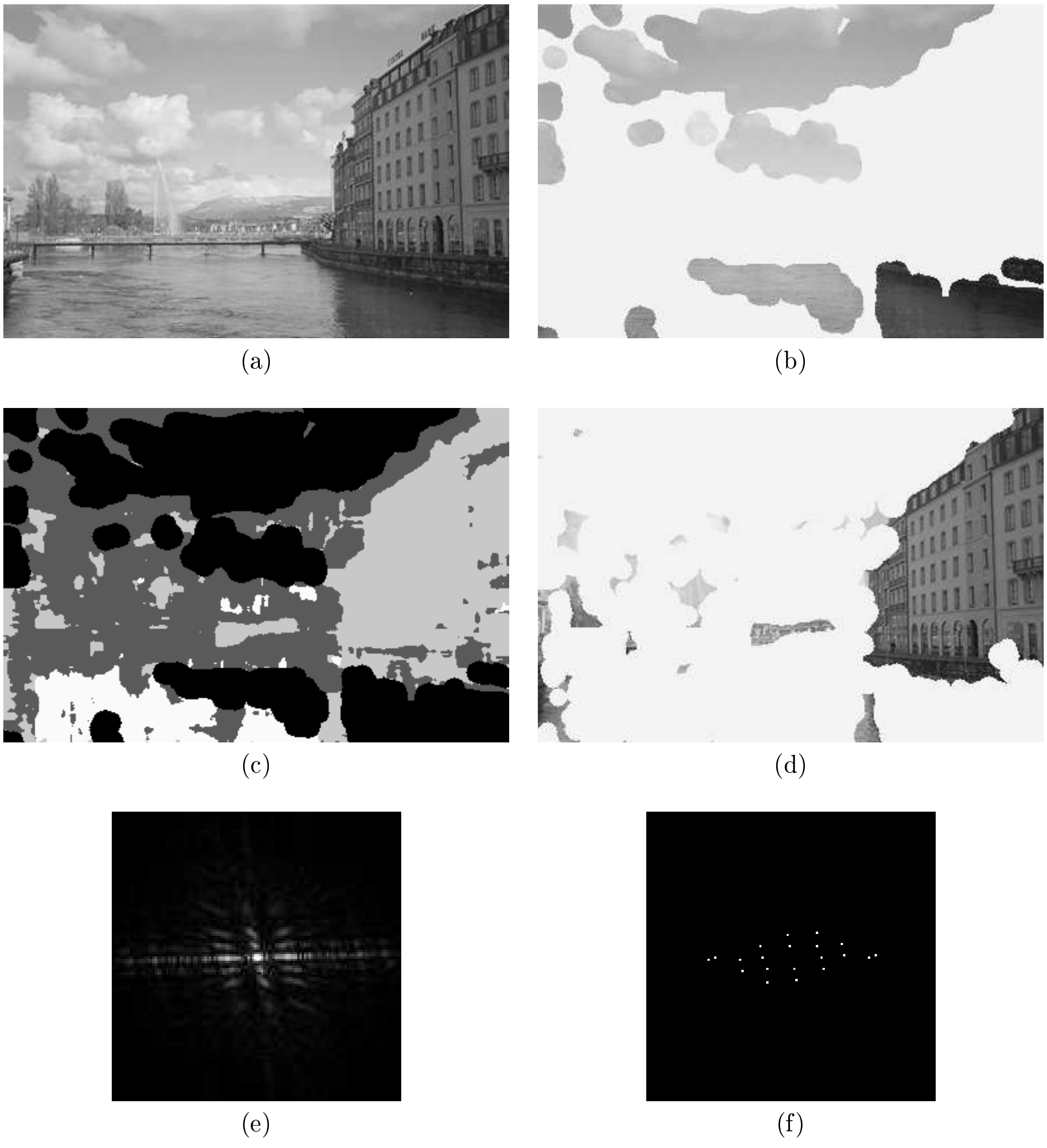


Figure 5-10: Segmentation of city scene. (a) Original; (b) Smooth regions; (c) Segmentation result with smooth regions in black; (d) Segmented building; (e) DFT magnitudes of building; (f) Extracted harmonic peaks.

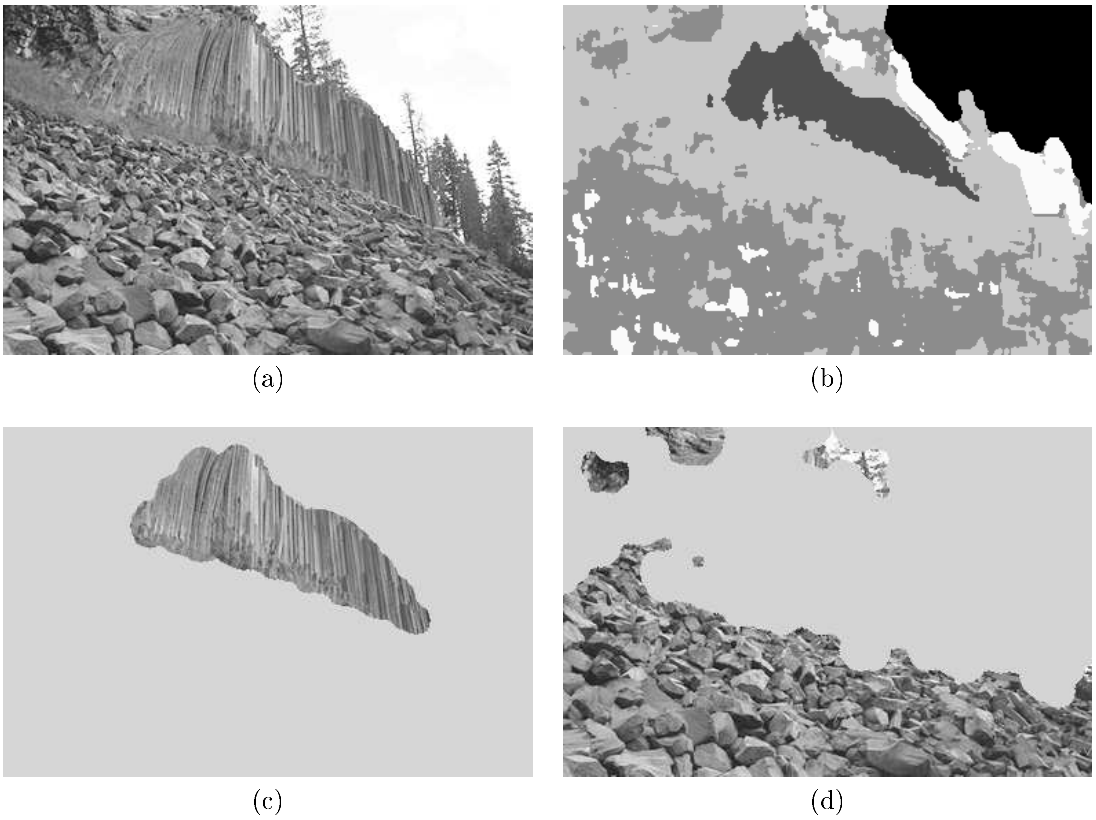


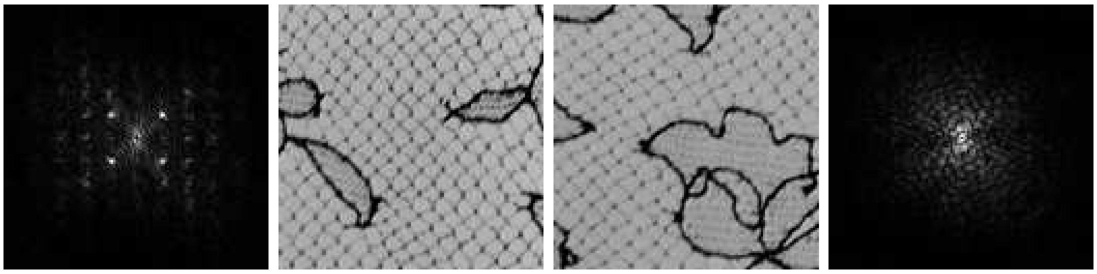
Figure 5-11: Segmentation of national park scene. (a) Original; (b) Segmentation result with smooth regions in black; (c) Segmented cliff; (d) segmented rocks.

---

spreading, the locations of these peaks in the frequency plane are different from where they would be if the building were frontal. Since the perspective effect is common in natural scene images, it is important to make the image features perspective invariant.

Various estimation methods have been proposed to recover the planar surface normal from surface texture information [35][58][63]. These methods first approximate the perspective transformation locally with an affine one and then use either spatial or spectral texture features to estimate the surface orientation. Among other computational concerns, the methods need at least two sizable texture patches for estimation. In natural scene images, very often the textured regions are not large enough to deploy these methods.

Aiming at a solution that will allow planar surface normal recovery from a single harmonic texture patch, a spectral approach is formulated as a decomposition of image perspective transformation into affine and chirp transformations. The perspective deformation of harmonic peaks can be expressed as a convolution of the spectral peaks with a shift-variant frequency kernel. The current results of this on-going research are described in Appendix D.



(a) (b) (c) (d)

Figure 5-12: Examples of perceptually similar textures which exhibit distinct spectral signatures. (b) and (c): two patches of Brodatz database texture D41: Lace. (a) and (d): DFT magnitudes of (b) and (c) respectively.

5.7.3 Performance Over the Brodatz Database

The quantitative measure of the retrieval performance is obtained using the benchmarking method described in Section 5.3. In Figure 5-13, the average recognition rate characteristic of the Wold-based method over the Brodatz database is plotted against those of four other methods: MRSAR, SPCA, TWT, and the Tamura features. Figure 5-13 shows that, by the traditional pattern recognition criterion, the effectiveness of the Wold model matches that of the MRSAR and is much better than those of the SPCA, the TWT, and the Tamura features.[3] It should be stressed that the perceptual advantage of the Wold model, which is demonstrated by some examples in the last section, is not well captured by this traditional quantitative evaluation.

Comparing the recognition rate averaged within each Brodatz class, the MRSAR method performs better on 10 of the 112 classes at neighbor size 8. Examples are D38 (Water), D41 (Lace), D80 (Straw cloth), and D84 (Raffia). In all ten classes, some class members have relatively homogeneous and structured patterns with prominent spectral peaks while the others do not. When the prototype image is a structured patch from such a class, the Wold model, which uses the harmonic peak information, may consider some other structured database images as more similar to the prototype than some of the patches in the prototype class. For these classes, the MRSAR model captures the average local spatial interaction and outperforms the Wold model by up to 18%. However, it is arguable if humans would agree with the original Brodatz grouping for some of these classes.

Although the Brodatz collection contains a large variety of natural textures, it is still a limited set. For instance, the Wold model, which represents both the harmonic structure and the randomness in a pattern, should outperform the MRSAR model on textures with mixed-spectra. However, most of the highly structured Brodatz textures have uniform backgrounds and simple local features. On these images, the MRSAR model, which is incapable of representing large scale

[3]The Wold performance curve in Figure 5-13 is slightly different from the one reported in [60]. This is due to the minor variations in the model implementation.

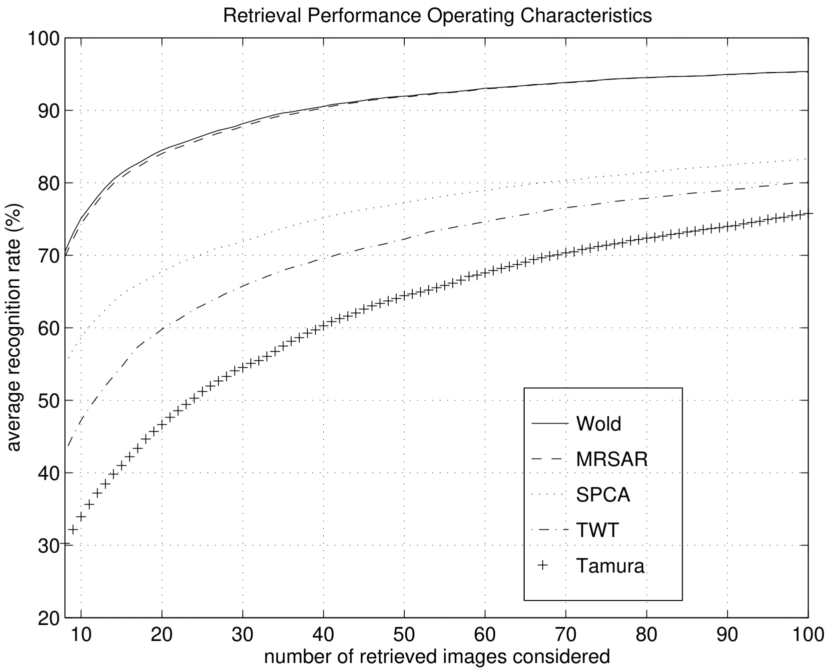


Figure 5-13: Retrieval performance operating characteristics — average recognition rates over the entire Brodatz database, considering from 8 up to 100 top retrieved images. Methods shown (from top curve to bottom): Wold-based model, multiresolution simultaneous autoregressive (MRSAR) model, shift-invariant principal component analysis (SPCA), tree-structured wavelet transform (TWT), and Tamura features.

---

spatial structures, performs no worse than the Wold model and can even achieve 100% recognition (eg.: D14, D20, D34, and D47). Although the Wold model does better in cases such as D3 (93.1% vs. 54.2%) and D52 (98.6% vs. 58.3%), its strength is not shown strongly against the MRSAR model given the limited variety of the database images.

The fundamental weakness of this performance evaluation is the lack of a meaningful benchmarking method for perceptual similarity in image retrieval. The current database classes are defined by the image origin in the Brodatz album, instead of the visual similarities. This is especially problematic for inhomogeneous images, where members of different classes can be perceptually more similar than images from the same original Brodatz picture. Examples are the sub-images of D36 (Figure 5-1 (b) Lizard skin) and D3 (top image of Figure 5-2 (a) Reptile skin). However, regrouping the images by their perceptual categories is not as easy as it appears. For instance, it is difficult for a human to judge pattern similarity without being biased by the image semantic content. One example is the five Brodatz brick wall patterns which differ in scale and surface properties.

5.8 Summary

In this chapter, a texture model based on the 2-D Wold random field decomposition theory is developed and evaluated for image retrieval in the Brodatz texture database.

The structure of the Wold-based model reflects the correspondence between the perceptual properties of the Wold components and the properties of human texture perception. It emphasizes the perceptually most salient harmonic structures in a texture while using the robust statistical models to represent the relatively unstructured patterns. The model avoids the actual decomposition of images and is designed to tolerate a variety of inhomogeneities in natural data, including pattern scaling and rotation. The robustness of the model makes it suitable for use in large collections of natural patterns.

The Wold-based model provides a new approach in modeling textures with mixed-spectra. Since the model uses harmonic peak extraction and MRSAR modeling to target different parts of the spectra, it is able to avoid a common problem found in statistical modeling: the information loss inherent in fitting highly structured textures with a low-order model, or the extra computation and over-fitting with a higher-order model.

For model perspective invariance, a linear system characterization of image perspective transformation and its decomposition into affine and chirp transformations are presented. The relation between geometric and spectral descriptions of perspective transformation is formulated to form a basis for future algorithms to infer image perspective parameters from a single sample of harmonic texture data.

Based on the new Wold texture model, an image retrieval algorithm is developed. Different types of image features are aggregated for similarity comparison by using a Bayesian probabilistic approach. Compared to other texture models, the Wold model appears to offer perceptually more satisfying results in the image retrieval experiments while matching or surpassing the best performance in recognition by traditional quantitative criteria.

A K-means-based image segmentation method is presented to demonstrate how Wold-based modeling can be used to characterize textured regions in natural scenes. The Wold feature sets constructed for these regions can be used subsequently in image content description.

Chapter 6

Temporal Texture Modeling

6.1 Introduction

Periodicity is common in the natural world. It is also a salient cue in human perception. Information regarding the nature of a periodic phenomenon, such as its location, strength, and frequency, is important for the understanding of the environment. Techniques for periodicity detection and characterization can assist in many applications requiring object and activity recognition and representation.

Although surface patterns may come to mind first, periodicity often involves both space and time, such as cyclic motion. The main body of work on periodic motion is model-based (eg. [1][44]). More recently there is work on motion recognition directly using low-level features of motion information (eg. [77][78][14]). However, to date, there has not been a method which uses low-level features to detect and systematically characterize periodicity in space and time. This work [61] attempts to tackle this problem by using *periodicity templates* to incorporate the location, strength, and other characteristic information of a periodic phenomenon. The templates are useful in applications such as periodic motion representation and action recognition. The template generating procedure also provides a tool for detecting and segmenting regions of periodicity. The proposed method is Fourier spectral based and uses 1-D Fourier transforms; therefore, it is also computationally efficient.

6.2 The Approach

The term *temporal texture* is defined in [78] as "motion patterns of indeterminate spatial and temporal extent". This definition emphasizes the randomness of a temporal event. In this work, "temporal texture" is used to refer to any approximately homogeneous spatiotemporal phenomenon. Hence, periodic temporal activities such as walking, swimming, or playing a violin are considered as a type of temporal texture.

The one-dimensional signals along the temporal dimension of a three-dimensional data volume formed by an image sequence can be considered as stochastic processes. When assuming stationarity, a stochastic signal can be decomposed into deterministic (periodic) and indeterministic (random) components. This is the classic 1-D Wold decomposition of stochastic processes [96]:

Theorem 9 *A zero-mean, regular, and stationary stochastic process $\{y(n)\}$, $n \in \mathcal{Z}$, can be represented uniquely by*

$$y(n) = v(n) + w(n) \tag{6.1}$$

where

$$w(n) = \sum_{k=0}^{\infty} a(k)u(n-k) \tag{6.2}$$

and $E[v(n)] = E[u(n)] = 0$. The coefficient $a(0) = 1$ and $\sum_{k=0}^{\infty} a(k)^2 < \infty$. Process $\{v(n)\}$ is deterministic and process $\{w(n)\}$ is regular and purely indeterministic. The innovation process $\{u(n)\}$ is white, i.e., $E[u(n)u(k)] = 0$ for all $n \neq k$. Processes $\{v(n)\}$ and $\{u(n)\}$ are orthogonal, i.e., $E[v(n)u(k)] = 0$ for all $n, k \in \mathcal{Z}$. Thus processes $\{v(n)\}$ and $\{w(n)\}$ are also orthogonal.

Different from the 2-D case, evanescent components do not exist in 1-D signals. Therefore, the deterministic component of a 1-D signal can be approximated solely by the harmonic component, which corresponds to the repetitive structure in the signal. As discussed previously, the repetitive structure in a signal contributes to the Fourier spectral harmonic peaks, and the random behavior to the smooth part of the spectra. Shown in Chapter 4, the deterministic energy ratio of a 2-D signal is a good measure of signal periodicity. Referring to the energy contained in the spectral harmonic peaks of a signal as the harmonic energy, the ratio between the harmonic energy and the total energy of a 1-D signal along the temporal dimension is used in this work to detect and segment spatiotemporal periodicity.

The approach described above assumes that the spatiotemporal periodicity is observable along lines parallel to the temporal (T) axis. In other words, the action needs to be tracked just like humans fix their eyes on a walking person. The problem of object tracking is conceptualized here as foveating. Typically, optical flow based techniques are used for tracking. However, flow based methods are usually susceptible to noise. A non-flow-based procedure — foveating by frame alignment — is developed here for object stabilization.

In this chapter, examples of walking people are used to illustrate the techniques. However, it should be stressed that the goal of this research is not to detect and segment a moving object, but to detect and characterize in a three-dimensional data volume those regions that exhibit periodicity. The algorithm is not expected to segment out the walking person. Instead, regions of legs and arms and the outline of the bouncing head and shoulder should be identified.

6.3 Related Work

The work of Polana and Nelson on periodic motion detection [78] is perhaps the most relevant to the approach presented in this chapter. In their work, reference curves, which are lines parallel to the trajectory of the motion flow centroid, are extracted and their power spectra computed. The periodicity measure p_f of each reference curve is defined as the normalized difference between the sum of the spectral energy at the highest amplitude frequency and its multiples and the sum of the energy at the frequencies half way between. Besides the value of the periodicity measure itself, there is no checking on the signal harmonicity along the curve, which is a weakness of the method. The periodicity measure for an entire sequence is the maximum of p_f averaged among pixels whose highest power spectrum values appear on the same frequency. The final periodicity measure is used to distinguish periodic and non-periodic motion by thresholding.

In [77], flow based algorithms are used to transform an image sequence so that the object in consideration is stabilized at the center of the image frame. Then flow magnitudes in tessellated frame areas of periodic motion were used as feature vectors for motion classification. It will be shown later in the chapter that flow based methods are very sensitive to noise.

The proposed approach differs from the work discussed above in the following ways: 1) the harmonic relationship among spectral peaks is explicitly verified; 2) a more accurate measure of periodicity in the form of harmonic energy ratio is proposed; 3) multiple fundamentals can be extracted along a temporal line; 4) the values of fundamental frequencies are used in processing to help distinguish periodicity of different activities; 5) regions of periodicity are actually segmented; and 6) optical flow based methods are not used here, so the proposed algorithm is robust in the presence of noise.

6.4 Method

6.4.1 Overview

The algorithm for periodicity detection and segmentation consists of two stages: 1) preprocessing: foveating by frame alignment; 2) simultaneous detection and segmentation of regions of periodicity.

Foveating (or tracking) is by itself an important research area. In this work, two types of image sequences are considered:

I. Area of interest (typically a moving object) is as a whole stationary to the camera, but the background can be moving;

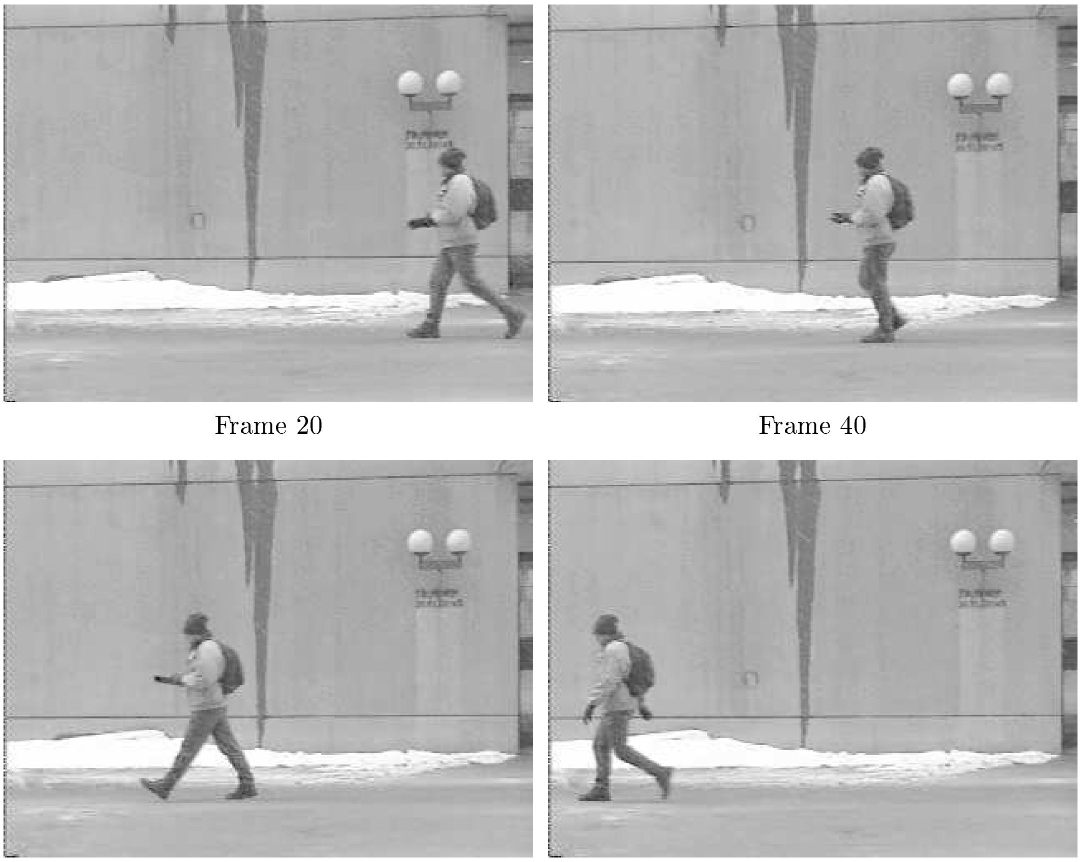
II. There is very little ego-motion involved, permitting minor shake and drift as in the hand-held situation, and approximately each moving object is as a whole moving frontoparallel to the camera along a straight line and at a constant speed.

In practice, a large number of image sequences containing periodicity can be categorized into one of these two types.

When watching a sequence of a person walking across the image plane, as shown in Figure 6-1, a notion of repetitiveness is experienced. However, if the frames are examined individually, there are no re-occurring scenes. The reason why periodicity is perceived in the sequence is due to one's ability to focus the visual attention on the moving object, or, *foveating*. The effect of foveating can be accomplished computationally by frame alignment. Obviously, foveating is not necessary for sequence type I, but in fact is a process of transforming type II sequences into type I.

In the second stage, 1-D Fourier transforms are performed along the temporal dimension of the aligned frames. The spectral harmonic peaks are detected and used to compute the harmonic energy. A periodicity template of frame size is then generated by using the extracted fundamental frequencies and the harmonic energy ratio at each frame pixel location. The original sequence is then masked for regions of periodicity.

In the following, the term *data cube* refers to the three dimensional (X: horizontal; Y: vertical; and T: temporal) data volume formed by stacking all the frames in a sequence, one in front the other. The XT and the YT slices of the data cube reveal the temporal behavior usually hidden from the viewer. Figure 6-2 shows the head and ankle level XT slices of the Walker sequence. The head leaves a more or less straight track (non-periodic) in (a) while the walking ankles in (b) make



Frame 20 Frame 40

Frame 60 Frame 80

Figure 6-1: Four frames in the 97 frame sequence Walker. Frame size is 320 by 240. The goal is not to segment out the walking person, but to detect and characterize regions that are involved in periodic motion, such as the legs, the arms, and the outline of the bouncing head and shoulder.

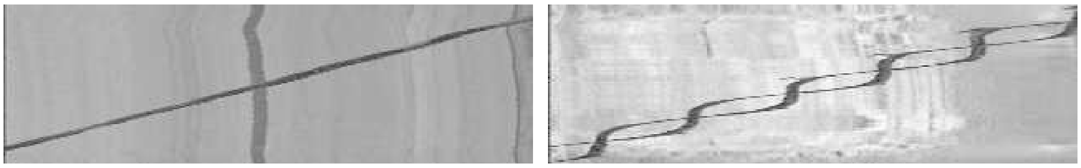
---

a crisscross pattern (periodic). Note that the periodicity in (b) is difficult to characterize since it is along neither the X nor the Y dimension, but the diagonal line. In this sense, foveating, or frame alignment, is a process of transforming data into a form in which periodicity can be easily detected and measured.

Throughout this section, the Walker sequence will be used to illustrate the technical points. More complex examples are given in Section 6.5.

6.4.2 Preprocessing: Foveating by Frame Alignment

To align a sequence to a particular moving object, the trajectory of the object needs to be detected first. To be demonstrated in Section 6.5, optical flow based methods are subject to the noise



(a) (b)

Figure 6-2: Head and ankle level XT slices of the Walker sequence. (a) Head leaves a straight track (non-periodic). (b) Walking ankles make a crisscross pattern (periodic). Note that the periodicity in (b) is difficult to characterize since it is along neither X nor Y dimensions, but the diagonal line.

---

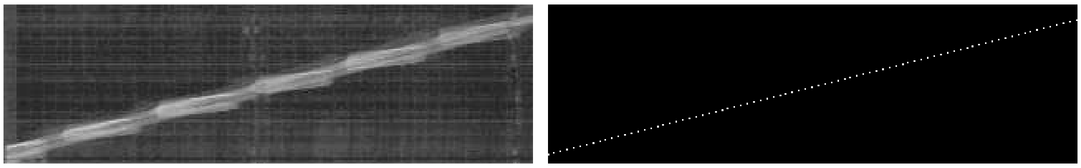
sensitivity inherent in pairwise frame differencing. A filtering method similar to the one in [70] is used here to find the trajectories.

A 1-D median filter is first applied to the data volume along the temporal dimension T to exclude moving objects and result in a sequence containing mostly the background. Filter length of 11 was used in the Walker sequence. Subtracting the background from the original sequence, a *difference sequence* containing mainly the moving objects is obtained. Since the object trajectories in consideration are approximately linear, the 2-D representations of the trajectories can be obtained by simply computing the average of the XT or the YT slices of the difference cube. This is equivalent to collapsing the difference cube top down to the XT plane or sideways to the YT plane. Next, lines in the 2-D trajectory images are detected via the Hough transform. The detected lines give the X or the Y positions of the moving objects in each frame. These position values are used as *alignment indices*. The averaged XT image of the Walker difference sequence and the line found by a Hough transform based method are shown in Figure 6-3. Note that multiple object trajectories can be detected simultaneously using this procedure. An example of three walking persons is shown in Section 6.5.

Using the alignment indices of an object, each frame in the image sequence can be repositioned to center the object to any specified position in the XY plane. After alignment, the object should appear as moving in place in the sequence. For instance, after aligning the Walker sequence in the X dimension (alignment in the Y dimension is not necessary since the overall Y position of the person does not change much), the position of the walker's torso remains in place, but the surroundings move to the right. In effect, this is equivalent to focusing the visual attention on an object when viewing a sequence in which the object's position changes frame by frame. This process is referred to as *foveating by frame alignment*. The aligned sequences are passed on to the second stage of the algorithm.

6.4.3 Finding Regions of Periodicity

The input to this stage is the aligned sequences. To save computation and storage, an aligned sequence can be cropped to limit processing to the area of interest. It will become clear later that the cropping does not effect the detection of periodicity. The location and size of the cropping



(a) (b)

Figure 6-3: (a) Averaged XT image of the Walker sequence after background removal. (b) Line found in (a) by using the Hough transform method. Each dot marks the object location in a frame. (Each horizontal line in the picture represents a frame.)

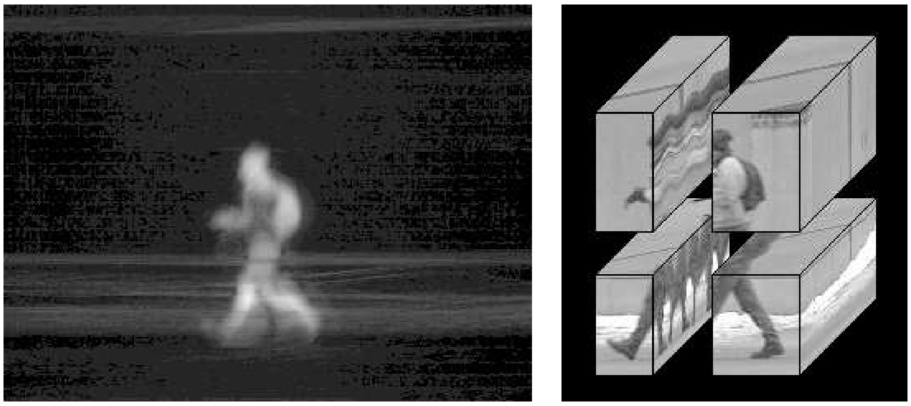
---

window can be determined by first aligning the difference sequence using the estimated alignment indices and then computing the average XY image. Figure 6-4 shows the averaged aligned XY image and the aligned and cropped Walker sequence with splits near the center of the frames to show the inside of the data cube.

Facing the data cube, a line can be drawn from a pixel (x_0, y_0) in the first frame all the way through the cube to arrive at pixel (x_0, y_0) in the last frame. Clearly, this line contains pixel (x_0, y_0) of all frames. This line, named as the **temporal line** at (x_0, y_0), is of essential importance to the discussion. If the frame size is N_x by N_y, then there are $N_x N_y$ temporal lines in the data cube.

In the aligned sequence, the object of interest should be moving in place. Apparently, if the object is moving cyclically in any manner, the periodicity will be reflected in some of the temporal lines. Figure 6-5 (a1) and (b1) show the head level and ankle level XT slices of 64 frames (Frame 17 to 80) of the data cube in Figure 6-4 (b). Every column in the images is a temporal line. These images are in fact the aligned and cropped version of the two XT slices in Figure 6-2. Each column in Figure 6-5 (a2) and (b2) is the 1-D power spectra of the corresponding column in (a1) and (b1). Note that the power spectrum values are normalized among all temporal lines in the data cube. Figure 6-5 (c1) and (c2) give the details of the signal along the white vertical lines in (b1) and (b2). While the head level slice in (a1) shows no harmonicity, the periodicity of the moving ankles in (b1) is reflected by the spectral harmonic peaks, shown clearly in (c2). Referring to the energy contained in the spectral harmonic peaks as the **temporal harmonic energy**, the ratio between the harmonic energy and the total energy of the signal along a temporal line, the **temporal harmonic energy ratio**, is used as a measure of temporal periodicity at the corresponding pixel location.

The 2-D spectral harmonic peak detection algorithm described in Section 3.3.3 is adapted here for use on 1-D signals. Given the signal along a temporal line, it is first zero-meaned and Gaussian tapered, then its power spectral values are computed using a fast Fourier transform. To locate the harmonic peaks, local maxima of the spectrum values (excluding values below 10% of the value range in the data cube) are found by searching a size 7 neighborhood of each frequency sample. These local maxima provide candidate locations of harmonic peaks. A local maximum marks the location of a harmonic peak only when its spectral frequency is either a fundamental or a harmonic. The definitions of the fundamental and the harmonic frequencies are similar to the ones given in



(a) (b)

Figure 6-4: (a) Averaged XY image of the aligned Walker sequence after background removal. The area of interest is clearly shown. (b) Aligned and cropped Walker sequence with splits near the center of the frames to show the inside of the data cube.

Section 3.3.3. A tolerance of one sample point is used in the frequency matching. Note that multiple fundamental frequencies can exist on a temporal line.

Due to the nature of the natural temporal signal and the windowing effect in spectra computation, a harmonic peak usually does not appear as a single impulse. Therefore the procedure so far only provides the central frequency location of each peak. The support of a peak is determined by growing outward from the central frequency location along the frequency axis until the spectrum value is below certain small value (5% of the spectrum value range in this work). After the harmonic peaks are identified, it is straightforward to compute the harmonic energy ratio associated with each fundamental frequency and its harmonics on the temporal line.

One may argue that the technique discussed above will fail when a temporal line contains only a sinusoidal signal that produces a single spectral peak. Theoretically, this is correct. However, this situation is highly unlikely to ever occur in image sequences of natural scenes and objects. The explanation to this traces back to the formation of the signal on a temporal line. Since a temporal line corresponds to a particular pixel in the image plane, having a pattern moving across the pixel is equivalent to having the pixel scanning across the pattern. When will this scan create a sinusoidal signal? The answer is only when the pattern has a sinusoidal profile. (An example is to translate horizontally a vertical sine grating pattern frontoparallel to the camera at a constant speed. See also Figure 3-3.) However, natural edges, patterns, and surfaces hardly ever have such a profile. Therefore, it is safe to say that higher harmonics will usually accompany the fundamentals in the Fourier spectra of the signals along the temporal lines.

Applying the peak detection procedure to all temporal lines in a data cube, the **periodicity template** of the aligned sequence is built by registering the fundamental frequencies and the corresponding values of temporal harmonic energy ratio at each pixel location in an data structure array of frame size. At places where no periodicity is involved, the corresponding values in the

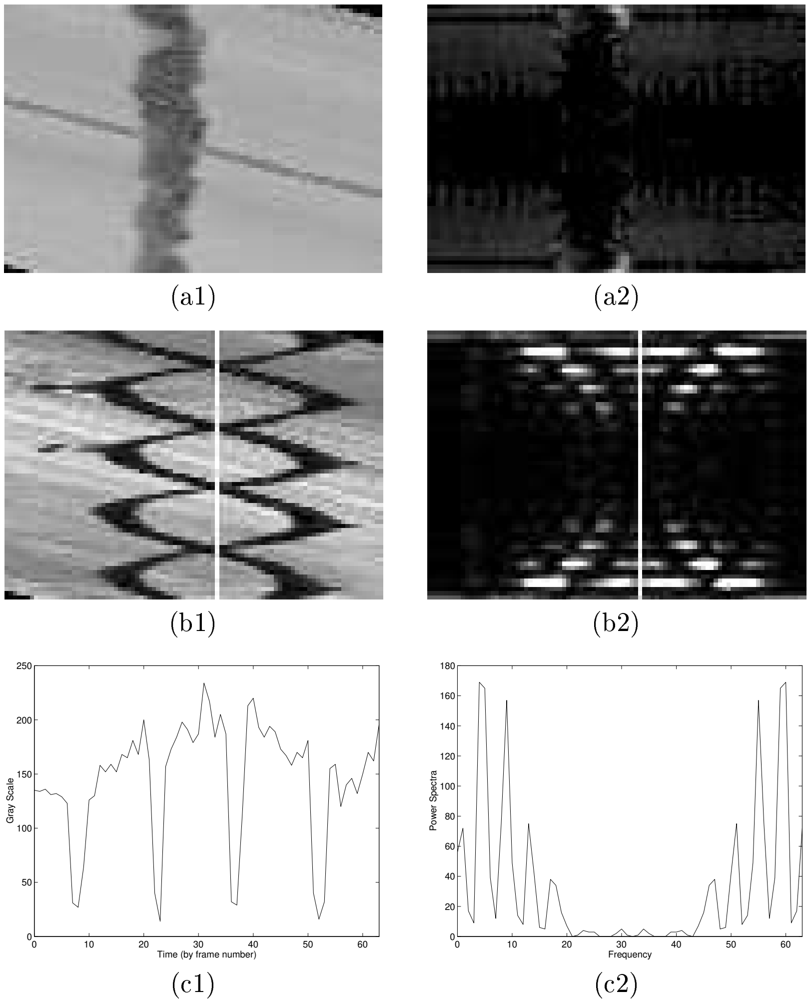
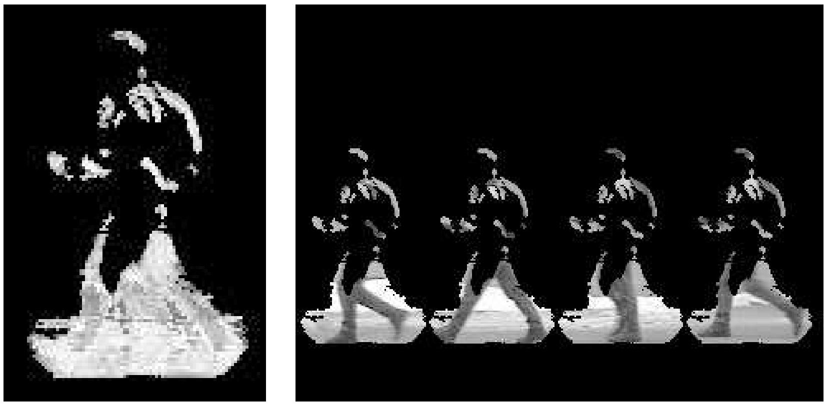


Figure 6-5: Temporal lines and their normalized power spectra. (a1) and (b1) show the head level and ankle level XT slices that are the aligned and cropped version of the two XT slices in Figure 6-2. Every column in the images is a temporal line. Each column in (a2) and (b2) is the 1-D power spectra of the corresponding column in (a1) and (b1). (c1) and (c2) give the details of the signal along the white vertical lines in (b1) and (b2). While the head level slice in (a1) shows no harmonicity, the periodicity of the moving ankles in (b1) is reflected by the spectral harmonic peaks, shown clearly in (c2).



(a) (b)

Figure 6-6: (a) Temporal harmonic energy ratio values of the aligned Walker sequence of Figure 6-4 (b). High value indicates more harmonic energy at the location. As expected, the brightest region is the wedge shape created by the walking legs. The head, the shoulder, and the outline of the backpack are shown because the walker bounces. The hands appear at the front of the body since in most parts of the sequence the walker was fixing his gloves and moving his hands in a rather periodic manner. Note that the moving background and parts of the walker do not appear in the template since there is no periodicity present in those areas. (b) Using the periodicity template and the alignment indices to mask the original sequence. The four frames in Figure 6-1 are masked and stacked together into one frame.

---

template data structure remain zero. Under circumstances such as a noisy background, some speckle noise may appear in the template. Simple morphological closing and opening operations can be applied to the template to remove the speckles.

Figure 6-6 (a) shows the temporal harmonic energy ratio values of the Walker sequence after one closing and one opening operation with a 3 pixel diameter circular structuring element. The larger the energy ratio value, the more harmonic energy at the location. As expected, the brightest region is the wedge shape created by the walking legs. The head, the shoulder, and the outline of the backpack are shown because the walker bounces. The hands appear at the front of the body since in most parts of the sequence the walker was fixing his gloves and moving his hands in a rather periodic manner. Note that the moving background and parts of the walker do not appear in the template since there is no periodicity present in those areas.

Since the non-periodic activities of the background do not light up in the templates, it is clear that the sequence cropping applied earlier in the second stage does not effect the processing results, but only increases the computational efficiency.

Using the alignment indices generated at the first stage, the periodicity template can be applied to the original sequence to mask the regions of periodicity in each frame. The masked frames corresponding to the ones in Figure 6-1 are stacked together and shown in Figure 6-6 (b).

6.5 Examples

In addition to the Walker sequence, four more example sequences are used here to demonstrate the effectiveness of the proposed algorithm: Trio, Dog, Wheels, and Jumping Jack. The Walker and Trio sequences were recorded by a hand-held consumer-grade camcorder during a snow storm. Camera drift and the influence of breathing of the cameraman are visible in the sequences. The Dog and Wheels sequences were taken by the same camera, but set on a tripod. The Jumping Jack sequence was recorded by a fixed Betacam camera in an indoor setting. Except for the Jumping Jack, none of the subjects in the sequences was aware of the filming; hence the activities are natural and exhibit natural irregularities. All original sequences have 320 by 240 frame size.

These examples are used to demonstrate 1) the effectiveness of the new algorithm in finding and characterizing periodicity in various settings; 2) the robustness of the algorithm under noisy conditions; and 3) the noise sensitivity of optical flow based estimation methods, which have been used for trajectory detection in many existing works, and avoided by the method proposed here.

6.5.1 Trio

Trio is a 156 frame sequence of three people walking and passing each other. Frames 40, 61, and 88 of the sequence are shown in the left column of Figure 6-7.

As in the Walker example, a temporal median filter is used to extract the background. After the background is largely removed from the sequence, the averaged XT image is computed. The lines in the XT image are then detected via Hough transform. Figure 6-8 shows the averaged XT image and the lines detected from the image. The detected lines provide the alignment indices of each objects. Note that the alignment indices of all three objects are estimated simultaneously by the proposed algorithm.

Next, as in the Walker example, the original sequence is aligned and cropped for each of the three moving individuals. All aligned sequences contain 64 frames. Then the aligned sequences go though the process of power spectrum estimation and harmonic peak detection. Finally, the temporal harmonic energy ratio at each pixel location is computed to generate the periodicity templates. Figure 6-9 shows example frames of the aligned sequences and the harmonic energy ratio values of the corresponding periodicity templates. Again, the goal here is not to segment out the walking person, but to detect and characterize regions of periodicity, such as the legs, the arms, the outline of the bouncing heads and shoulders, and even the dangling straps of the backpack. Finally, the templates are used to mask the original sequence. Examples of masked sequence are shown in the right column of Figure 6-7.

Notice that, besides the center person, there is a second or even a third person passing through in all three aligned sequences. However, their appearance has no effect on the result of periodicity detection. This is because, to any individual temporal line, these passersby are one-time event and do not contribute to the temporal harmonic energy of the temporal line. The Trio example demonstrates that the proposed algorithm is well suited for the detection of multiple periodicities, even under the circumstances of temporary object occlusion.

6.5.2 Dog

Dog is a 104 frame sequence, in which a person walks with two dogs in front of a picket fence. Figure 6-10 (a) shows the 13th frame of the 64-frame aligned sequence. Images (b1) and (b2) show

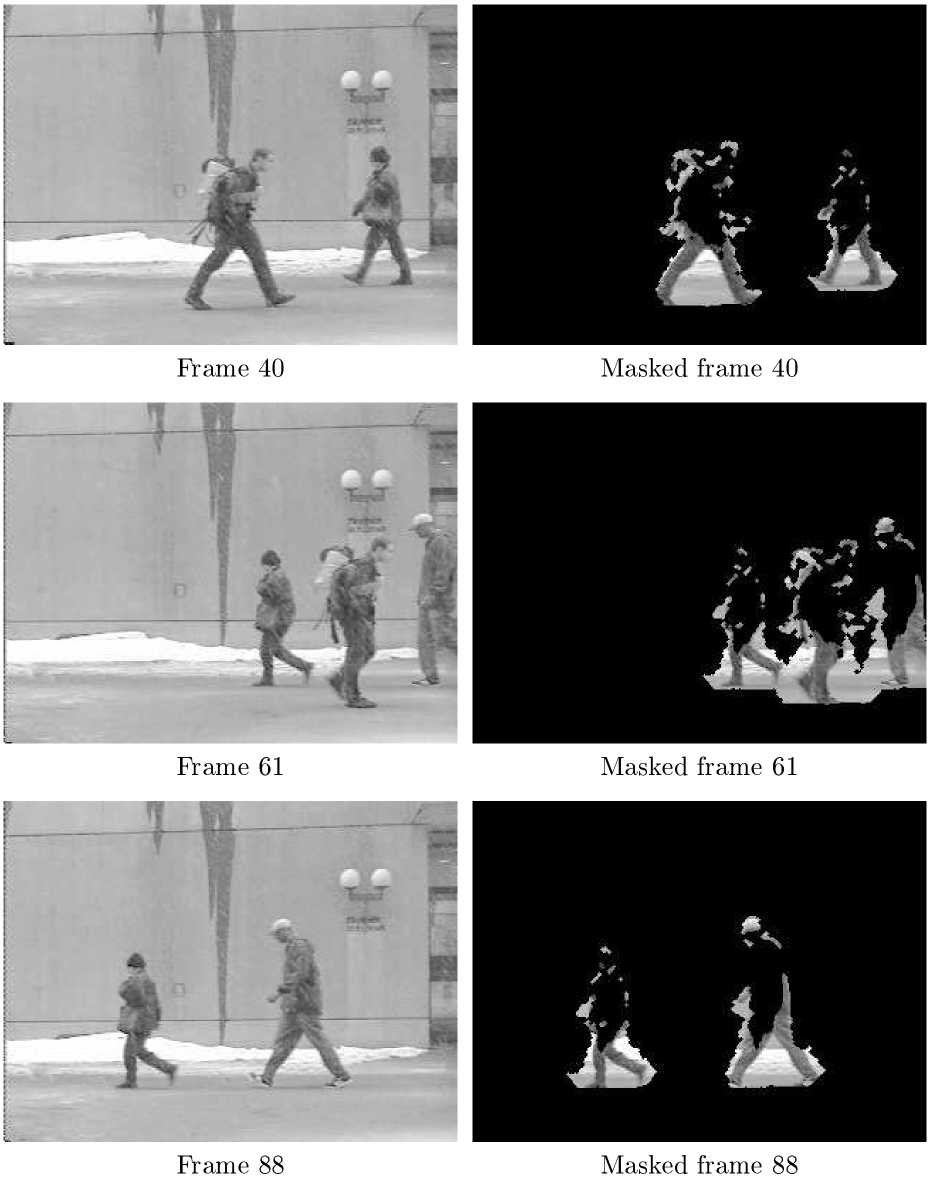
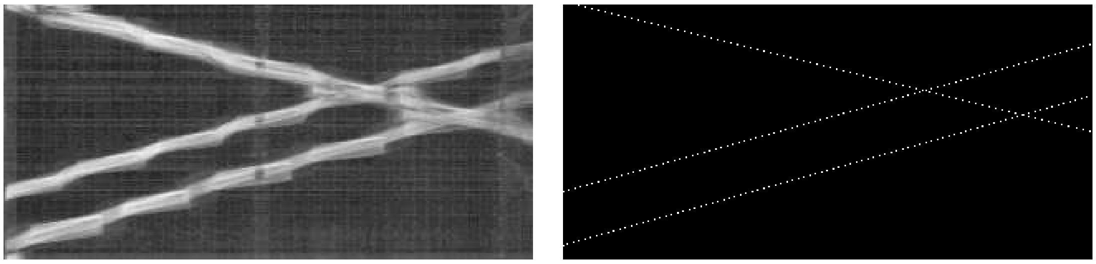


Figure 6-7: Three pairs of original and masked frames of sequence Trio. Left column: originals, where three people are walking and passing each other. Right row: frames in the left column are masked by the periodicity templates of three individuals. Again, the goal here is not to segment out the walking person, but to detect and characterize regions of periodicity, such as the legs, the arms, the outline of the bouncing heads and shoulders, and even the dangling straps of the backpack.



(a) (b)

Figure 6-8: (a) Averaged XT image of the Trio sequence after background removal. (b) Lines found in (a) by using the Hough transform method.

---

the first and the second fundamental frequencies in the periodicity template, while (c1) and (c2) contain the corresponding harmonic energy ratios. Note that there are double fundamentals at many pixel locations.

The complication here is the picket fence. In the original sequence, the fence is part of the fixed background, exhibiting only spatial periodicity. However, when the sequence is aligned to the person, the fence starts to move in the background, leaving periodic signature on many temporal lines. As shown in (b1) and (c1), the fence area lights up in the periodicity template.

Figure 6-10 (b3) shows the fundamentals with value near 0.875π, which is the temporal frequency of the fence in the aligned sequence. The frequency values can be used to extract the fence. Figure 6-10 (c3) shows the harmonic energy ratios in the template after the fence frequency is taken out. Figure 6-10 (d) shows frame 46 of the original Dog sequence. Images (e) is (d) masked to show the fence region, and image (f) is (d) masked to show other regions of periodicity.

6.5.3 Wheels

The examples shown so far all involve walking. However, the algorithm is not limited to periodicity caused by human activities, but works in general for any periodic spatiotemporal phenomenon.

Wheels is a 64 frame sequence of a car passing by a building. Near the top of the building, two spinning wheels are connected by a figure 8 belt. Periodicity occurs at the car hub caps and the spinning wheels. One side of the figure 8 belt is patterned and appears periodic. Every region with periodicity is expected to be captured : the hub caps, the wheels, and one side of the belt. As shown in Figure 6-11, the algorithm accomplishes just that.

6.5.4 Jumping Jack

In the Jumping Jack sequence, there is no translatory motion involved and most part of the background is very smooth. This sequence and the noisy versions of it are used to demonstrate the robustness of the new algorithm under noisy conditions and also to show the sensitivity of the optical flow based motion estimation to noise. There are three different kinds of input given to

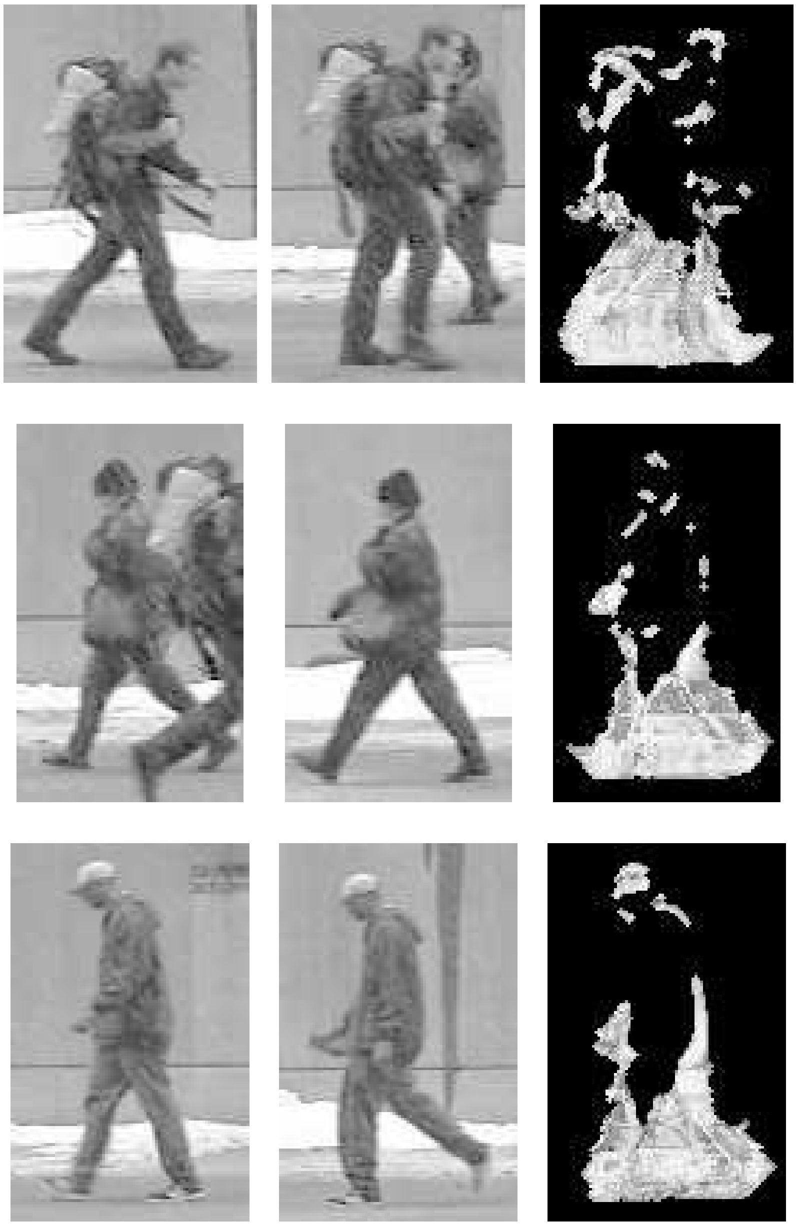
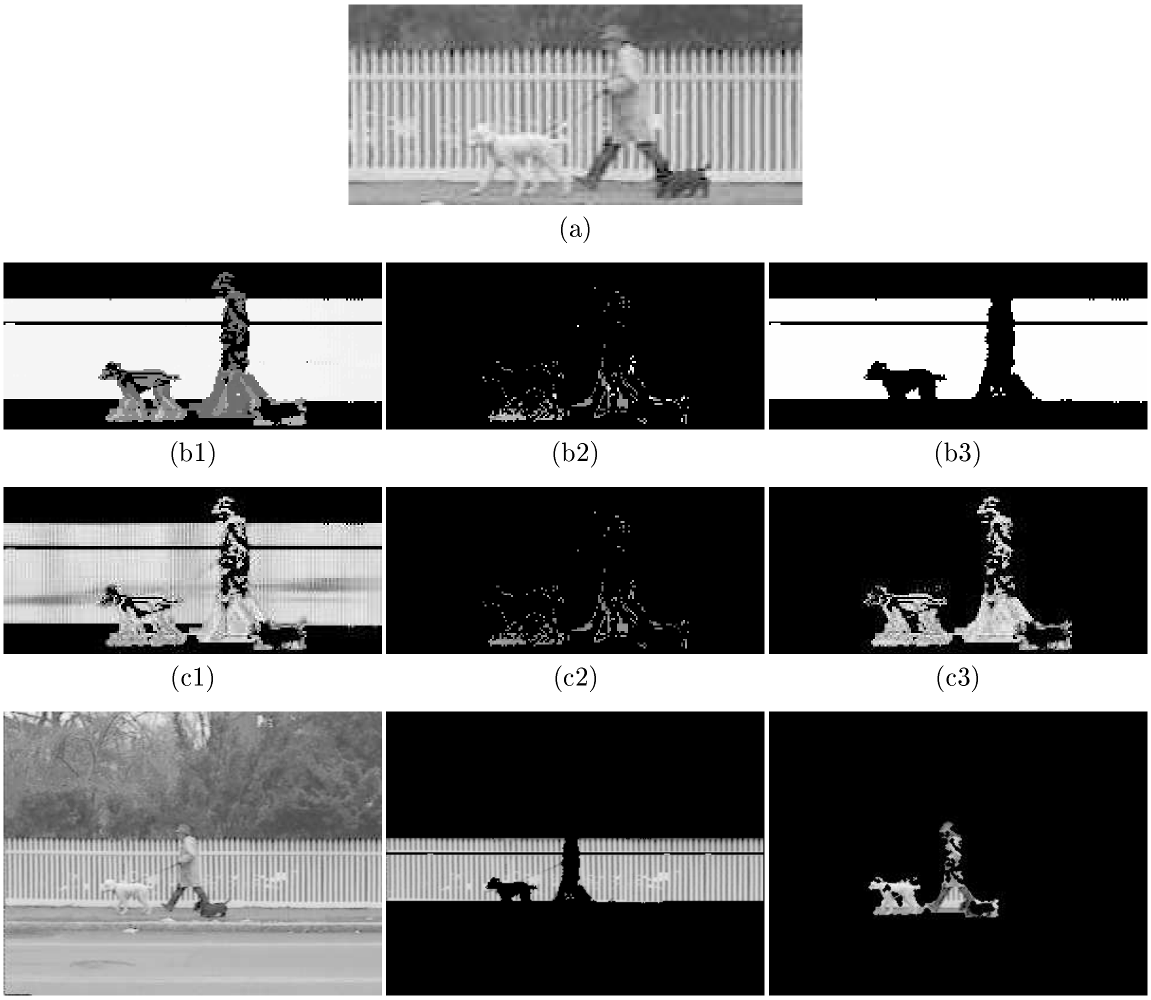


Figure 6-9: Example frames of the aligned sequences and the harmonic energy ratio values of the corresponding periodicity templates for each individual of the Trio sequence. Two left columns: example frames. Right column: harmonic energy ratios.



(a) (b1) (b2) (b3) (c1) (c2) (c3) (d) (e) (f)

Figure 6-10: The Dog sequence. (a) Frame 13 of the 64-frame aligned sequence. (b1) and (b2): first and second fundamental frequencies in periodicity template. (b3) Fundamentals with value near the temporal frequency of the fence. The frequency values can be used to extract the fence. (c1) and (c2): harmonic energy ratios in the template, corresponding to the frequencies in (b1) and (b2). (c3) Harmonic energy ratios in the template after the fence frequency is taken out. (d) Frame 46 of the original sequence. (e) Frame in (d) masked to show fence region. (f) Frame in (d) masked to show other regions of periodicity.

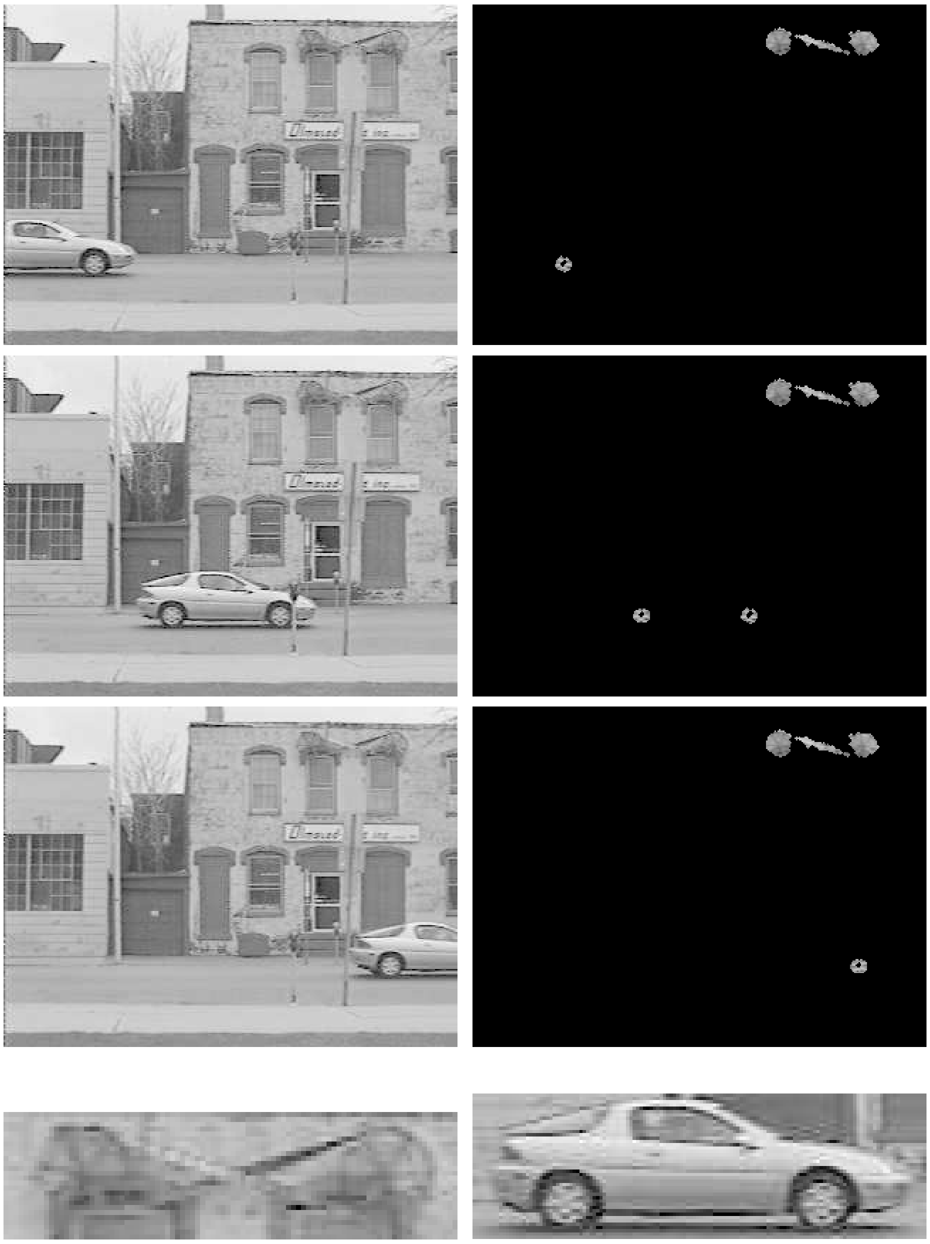
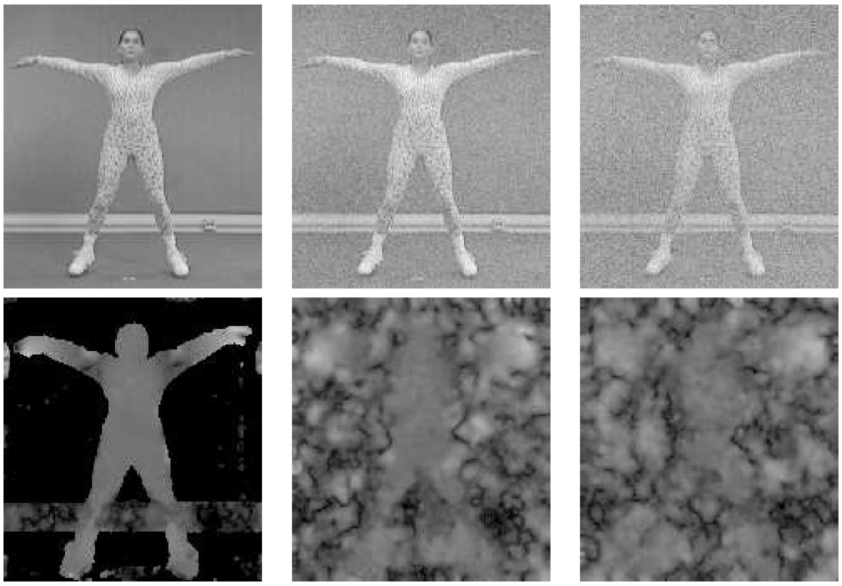


Figure 6-11: Sequence Wheels contains a car passing by a building. Near the top of the building, two spinning wheels are connected by a figure 8 belt. Periodicity occurs at the car hub caps and the spinning wheels. One side of the figure 8 belt is patterned and appears periodic. Only the hub caps, the wheels, and one side of the belt should be captured. As shown in the first three rows, the algorithm accomplishes just that. The original and the masked frames shown are frames 6, 24, and 45. Details of the spinning wheels and the car are shown in the bottom row.



(a) (b) (c)

Figure 6-12: Optical flow estimates of the Jumping Jack sequences: (a) from original sequence; (b) from sequence corrupted by AGWN with variance 100; (c) from sequence corrupted by AGWN with variance 400. Top row: frame 61 of the Jumping Jack sequences. Bottom row: corresponding optical flow magnitudes. Under noisy conditions, the flow-based algorithm is mostly ineffective.

---

the algorithm: the original sequence and the sequences corrupted by additive Gaussian white noise (AGWN) with variance 100 and 400. The length of the sequences used in power spectrum estimation is increased to 128 due to the cycle of the jumping motion. The input sequences have frame size 155 by 170.

Most related work uses flow based methods to extract spatiotemporal surfaces or curves to locate the moving objects in a sequence. However, as demonstrated in Figure 6-12, the noise sensitivity of the flow based method can be a drawback in real applications. The optical flow magnitudes are estimated here by using the hierarchical least-squares algorithm [92] which is based on a gradient approach described in [11][62]. This algorithm is representative of the existing optical flow estimation techniques.

In Figure 6-12, frame 61 of the original and the two white-noise corrupted Jumping Jack sequences are shown in the top row, and the corresponding optical flow magnitudes in the bottom row. When given a clean input such as the original Jumping Jack sequence, the flow magnitudes can be used to segment out the moving object. However, under the noisy conditions, the algorithm is mostly ineffective.

Figure 6-13 shows the processing results produced by the proposed algorithm. Again, frame 61 of the original and the two white-noise corrupted Jumping Jack sequences are shown in the top row. The second row in Figure 6-13 is the 57th TY (not YT!) image of each sequence. It shows

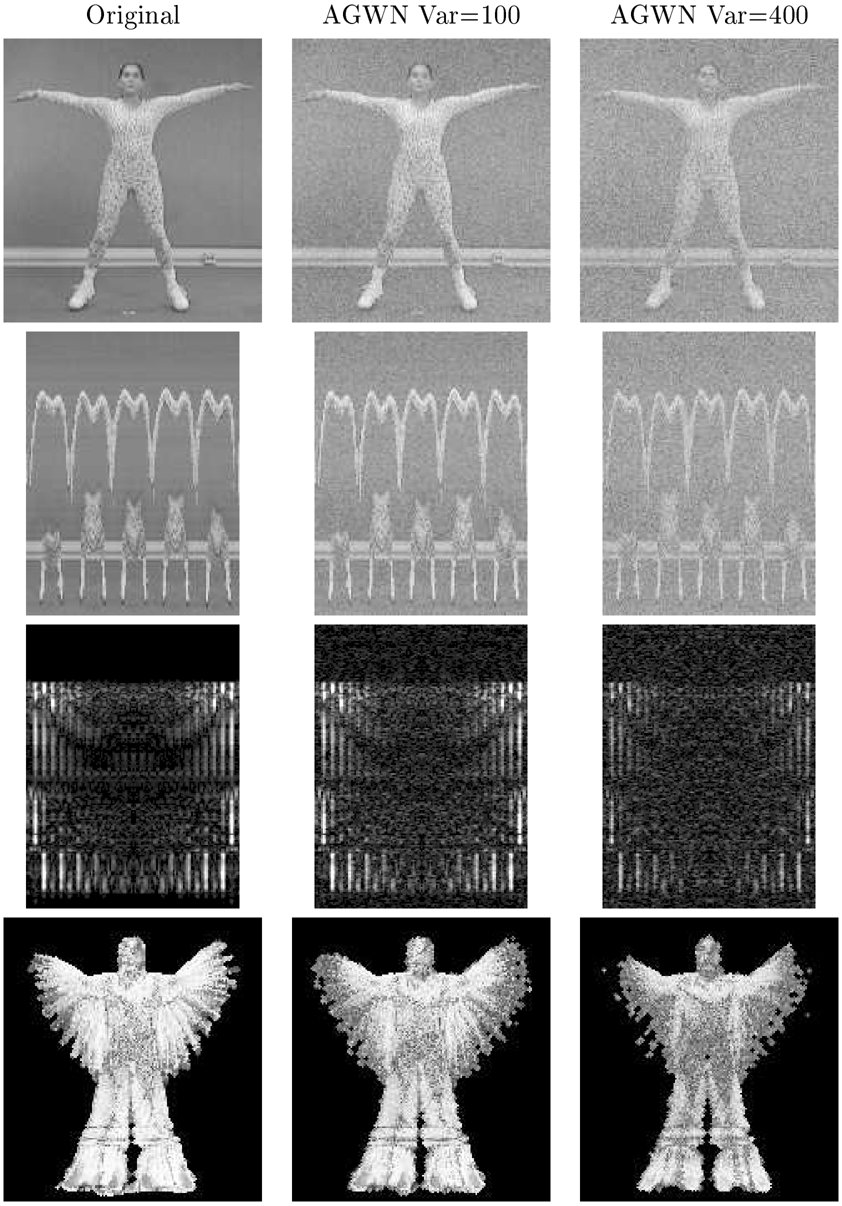
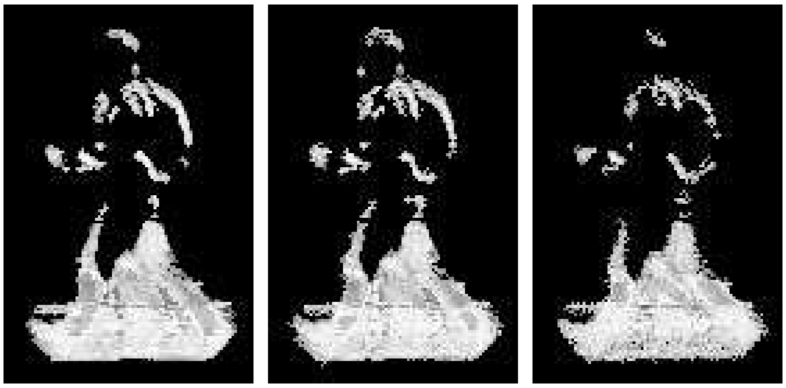


Figure 6-13: Processing results of the Jumping Jack sequences. Left column: from original sequence. Middle column: from sequence corrupted by AGWN with variance 100. Right column: from sequence corrupted by AGWN with variance 400. Row 1: frame 61 of the Jumping Jack sequences. Row 2: TY slice 57, showing the tracks left by the right hand and leg; each row of these images is a temporal line. Row 3: temporal line power spectra of TY slice 57. Row 4: harmonic energy ratios of periodicity templates. The proposed algorithm is robust in the presence of noise. As shown in Row 4, although the noise causes some degradation in the arm regions, other areas of the templates are well preserved.



(a) (b) (c)

Figure 6-14: Harmonic energy ratios of the periodicity templates of the Walker sequence: (a) from original sequence; (b) with AGWN of variance 100; (c) with AGWN of variance 400. The proposed algorithm is robust in the presence of noise.

---

the tracks left by the right hand and leg. The rows in these images are temporal lines and the corresponding power spectra are shown in the third row of the figure. The harmonic energy ratio values of the periodicity templates can be found in the last row of the figure. Although the noise causes some degradation in the arm regions, other areas of the templates are well preserved.

The reason why the proposed algorithm is robust in the presence of certain amounts of white noise is that white noise only contributes to the relatively smooth indeterministic component of the power spectrum. As long as the noise energy is not so high that it overwhelms the spectral harmonic peaks, the algorithm works.

6.5.5 Walker

The detection results of the Walker sequence are mostly shown in Section 6.4. Presented here are the results from noisy inputs. Additive Gaussian white noise of variance 100 and 400 was used to corrupt the original sequence. The length of the aligned sequences for power spectrum estimation is 64. The resulting harmonic energy ratios of the periodicity templates in Figure 6-14 show clearly that, unlike optical-flow based methods, the proposed algorithm is robust in the presence of noise.

6.6 Discussion

6.6.1 Algorithm

Compared to the one used in [78], the periodicity measure proposed here in the form of the temporal harmonic energy ratio is a more accurate and more reliable measure of signal periodicity. It not only can indicate the presence of periodic activities, but also gives a quantitative measure of how much energy at each pixel location is contributed by the periodicity.

The fundamental frequencies of the temporal signals are extracted and registered to the periodicity templates. Using this information, areas involved in periodic activities with different cycles can be distinguished easily. This is demonstrated in the example of a person walking in front of a picket fence.

The proposed algorithm can also be considered as a periodicity "filter". At first, all moving objects are targets for foveating, but then only the ones exhibiting periodic behavior remain. Given an input sequence of a street with cars and pedestrians, the algorithm will find the moving legs of the pedestrians and filter out the cars and other non-periodic activities. Since periodicity is a salient feature to human visual perception, the proposed algorithm provides a model of low-level periodicity perception, even though it may not work exactly like the human visual system.

Existing related work often uses flow based methods to extract the trajectories of moving objects. Since flow based methods can be susceptible to noise, foveating by frame alignment is used here instead to focus on individual objects. This approach not only improves the performance of the algorithm in the presence of noise, but also is efficient in that it generates alignment parameters of all moving objects simultaneously.

The method introduced here is computationally efficient. The most machine intensive part of the algorithm is the 1-D fast Fourier transform used in power spectrum computation. However, when the activity cycle is reasonably short, such as walking in normal speed, a sequence length of 64 frames suffices. Cropping of aligned sequences also helps to speed up the processing.

In the current work, a few assumptions are made on the data. The steady background condition in the translatory moving object case is mainly for the background subtraction. The algorithm actually tolerates small camera movement quite well. When an object is not moving in a translatory manner with respect to the camera, its trajectory will not be linear in the data cube and a scheme more sophisticated than the Hough line detection will have to be used for the frame alignment. If the object is not moving frontoparallel to the camera, the perspective effect will change the size of the object in the sequence. However, this change should not be significant during the period of 64 frames when the distance between the camera and the object is sufficiently large. In practical situations, this is often the case.

6.6.2 Applications

Among other possible applications, the proposed algorithm can be applied to motion classification and recognition. In [14], the shape of the active region in a sequence was used for activity recognition. In [77], the sum of the flow magnitudes in tessellated frame areas of periodic motion were used as feature vectors for motion classification. The periodicity templates produced by the algorithm introduced here can provide not only distinct shapes of regions of periodic motion, such as the wedge for the walking motion and the snow angle for the jumping jack, but also accurate pixel-level description of a periodic action in the form of temporal harmonic energy ratio and motion fundamental frequencies.

The characterization of periodicity is also important to video database related applications. The presence, position, strength, and frequency information of periodic activities can be used for video representation and retrieval.

In general, periodicity is a salient attention-getting feature. The proposed algorithm can be used in numerous surveillance applications for detecting ambulatory activity without having to do full-person recognition.

6.7 Summary

A new algorithm for finding periodicity in space and time is presented. The algorithm consists of two main parts: 1) foveating, or frame alignment, which transforms data into a form in which periodicity can be easily detected and measured; 2) Fourier spectral harmonic peak detection and energy computation to identify regions of periodicity and measure its strength. This method allows the detection, segmentation, and characterization of spatiotemporal periodicity to be accomplished simultaneously, and is computationally efficient. The effectiveness of the technique and its robustness to noise over optical-flow based methods are demonstrated using real-world video examples.

The use of periodicity templates for characterizing spatiotemporal periodicity is proposed. The templates contain information such as the fundamental frequencies and the temporal harmonic energy ratio at each frame pixel location. The periodicity templates and the template generating algorithm are useful tools for applications such as action recognition, video databases, and video surveillance.

Chapter 7

Conclusions

This work has established Wold-based texture modeling as an important method for a wide range of applications that benefit from efficient and effective characterization of textural information. This was achieved by bridging the gap between the Wold random process decomposition theory and practical texture modeling. Applications demonstrated in this work include image and video analysis, representation, understanding, and similarity comparison.

- A new spectral 2-D Wold decomposition algorithm for homogeneous or near homogeneous random fields was presented. This algorithm detects the Fourier spectral harmonic and evanescent frequencies of a textured image and decomposes the image by extracting these frequency components from the image spectrum. The harmonic frequencies are identified by using the fundamental-harmonic relationship among spectral peaks, while the evanescent frequencies are detected via Hough transformation. Compared to the prior Wold decomposition methods, this fully automated algorithm is more robust and flexible for a large variety of natural textures, and is also computationally efficient.

- A psychophysical study was conducted to investigate the perceptual properties of Wold-based texture modeling. A highly significant correlation was found between the human and computer texture ranking data, suggesting that the component energy resulting from the 2-D Wold decomposition of an image is a good computational measure for the most salient dimension of human texture perception, the dimension of repetitiveness vs. randomness. The highly significant concordance of the human rankings also verifies that the top perceptual dimension indeed corresponds to certain underlying criteria, upon which the human subjects agree, for texture similarity measurement.

- A Wold-based shift, rotation, and scale invariant texture model was developed and presented. The structure of the model reflects the correspondence between the perceptual properties of the Wold components and the properties of human texture perception. The model features are extracted without explicitly performing a Wold decomposition. By modeling the structured and relatively unstructured texture components separately, the model overcomes the common deficiency of purely statistical models in characterizing structured patterns. This model is designed to tolerate a variety of inhomogeneities in natural data, making it suitable for use in large collections of natural patterns.

- For model perspective invariance, a linear system characterization of image perspective transformation and its decomposition into affine and chirp transformations were presented. The relation between geometric and spectral descriptions of perspective transformation was formulated to form a basis for future algorithms to infer image perspective parameters from a single sample of harmonic texture data.

- Based on the new texture model, an image retrieval algorithm was developed for textured image databases. Different types of image features were aggregated for similarity comparison by using a Bayesian probabilistic approach. Compared to other well-known models, the Wold model appears to offer perceptually more satisfying results in the image retrieval experiments while matching or surpassing the best recognition performance of state-of-the-art texture models.

- A K-means-based image segmentation method was presented to demonstrate the use of Wold-based modeling in characterizing textured regions in natural scene images. The Wold feature sets constructed for these regions can be used subsequently in image content description.

- Based on the principle of 1-D Wold decomposition, a new algorithm was developed to model temporal textures for image sequence analysis. The algorithm first performs foveating via frame alignment and then identifies spatiotemporal regions exhibiting periodicity. This computationally efficient method allows the detection, segmentation, and characterization of periodic motion to be accomplished simultaneously. Compared to commonly used flow-based techniques, this method is more robust in the presence of noise. The effectiveness of the algorithm has been demonstrated on a variety of complex natural scene videos with multiple motions.

- The use of periodicity templates was proposed for characterizing periodicity in space and time. The information carried by the templates not only indicates the presence and position of signal periodicity, but also gives an accurate quantitative measure of how much energy at each frame pixel location is contributed by the periodicity. The periodicity templates and the template-generating algorithm provide useful tools for detecting and representing periodicity in applications such as action recognition, video database retrieval, and video surveillance.

Chapter 8

Future Research Suggestions

8.1 Model Perceptual Property

The perceptual property of the Wold models can be further studied in an image similarity comparison experiment using human subjects. Instead of ordering a set of image samples along an abstract perceptual dimension, the samples can be ordered by their similarities to a prototype image. The correlation between the human and the computer ordering data can indicate how well the Wold models correspond to the human perception of image similarity.

8.2 Retrieval Algorithm

When the homogeneity condition of the data permits, the image retrieval algorithm presented in Chapter 3 can be improved by using the actual decomposition of database images. Image decomposition should improve at least two aspects of the system. First, the MRSAR can have a better fit to the indeterministic component in data and therefore be more effective in comparing image similarity. Second, a better feature aggregation can be achieved by using the image deterministic energy ratio instead of the autocovariance energy ratio. It has been demonstrated in Chapter 4 that the deterministic energy ratio is a good measure of image harmonicity.

8.3 Model Performance Evaluation

In Chapter 3, the performance of the retrieval algorithms was evaluated by using the averaged recognition rate criterion. However, the image classes used in the computation are defined by the origin of the images in the Brodatz album, not by visual similarities. Since the album contains many inhomogeneous images, the class definitions are not always appropriate. In addition, the texture variety in the album is limited. Establishing the "ground truth" of the image perceptual classes for a larger and more diverse texture collection by using human subjects should notably improve the quality of the performance evaluation.

8.4 Temporal Textures

In Chapter 6, the periodicity templates and the template-generating algorithm are proposed as general tools for detecting and charactering periodicity in the spatiotemporal domain. As discussed in Section 6.6.2, there are many potential applications of these tools. Such applications include real-time action recognition, video database retrieval, automated video surveillance, just to name a few.

Appendix A

Derivations in Spectral Decomposition

A.1 Fundamental Frequency Refinement

Equation (3.9) can be derived as follows. From

$$\mathbf{f}^f_{j-1} = \frac{\sum\limits_{i=0}^{j-1} \mathbf{f}^h_i}{\sum\limits_{i=0}^{j-1} \beta_i} \qquad\qquad \mathbf{f}^h_j = \beta_j \mathbf{f}^f_{j-1} + \Delta\mathbf{f}_j,$$

it follows

$$\begin{aligned}
\sum_{i=0}^{j} \mathbf{f}^h_i &= \sum_{i=0}^{j-1} \mathbf{f}^h_i + \mathbf{f}^h_j \\
&= \mathbf{f}^f_{j-1} \left(\sum_{i=0}^{j-1} \beta_i \right) + \mathbf{f}^h_j \\
&= \mathbf{f}^f_{j-1} \left(\sum_{i=0}^{j} \beta_i \right) - \beta_j \mathbf{f}^f_{j-1} + \mathbf{f}^h_j \\
&= \mathbf{f}^f_{j-1} \left(\sum_{i=0}^{j} \beta_i \right) + \Delta\mathbf{f}_j.
\end{aligned}$$

Therefore,

$$\mathbf{f}^f_j = \frac{\sum\limits_{i=0}^{j} \mathbf{f}^h_i}{\sum\limits_{i=0}^{j} \beta_i} = \mathbf{f}^f_{j-1} + \frac{\Delta\mathbf{f}_j}{\sum\limits_{i=0}^{j} \beta_i}.$$

A.2 Gaussian Surface $\bar{\mathbf{f}}$ and $\boldsymbol{\Sigma}_{\mathbf{f}}$ Estimation

For N independent, identically distributed (i.i.d.) Gaussian samples $\mathbf{x}_i$, $i = 1, \ldots, N$, the maximum likelihood estimates of the mean $\bar{\mathbf{x}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ are [23] [1]

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \qquad \boldsymbol{\Sigma}_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

Given that $N_{(k,l)}$ i.i.d. 2-D Gaussian random samples $\mathbf{f} = (k,l)$ are observed at each $(k,l) \in \mathcal{D}$, the maximum likelihood estimates of the mean $\bar{\mathbf{f}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{f}}$ are

$$\bar{\mathbf{f}} = \frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\mathbf{f}\,N_f \tag{A.1}$$

and

$$\boldsymbol{\Sigma}_{\mathbf{f}} = \frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}(\mathbf{f} - \bar{\mathbf{f}})(\mathbf{f} - \bar{\mathbf{f}})^T N_f, \tag{A.2}$$

where $\mathbf{f} = (k,l) = [\ k \quad l\ ]^T$ and $N_t = \sum_{(k,l)\in\mathcal{D}} N_{(k,l)}$. From (A.1),

$$\bar{\mathbf{f}} = \left(\frac{1}{N_t}\sum_{(k,l)\in\mathcal{D}} kN_{(k,l)},\ \frac{1}{N_t}\sum_{(k,l)\in\mathcal{D}} lN_{(k,l)}\right) = (\bar{k}, \bar{l}).$$

From (A.2),

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{f}} &= \frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\left(\mathbf{f}\mathbf{f}^T - \mathbf{f}\bar{\mathbf{f}}^T - \bar{\mathbf{f}}\mathbf{f}^T + \bar{\mathbf{f}}\bar{\mathbf{f}}^T\right)N_f \\
&= \frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\mathbf{f}\mathbf{f}^T N_f - \left(\frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\mathbf{f}N_f\right)\bar{\mathbf{f}}^T - \bar{\mathbf{f}}\left(\frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\mathbf{f}N_f\right)^T + \left(\frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}N_f\right)\bar{\mathbf{f}}\bar{\mathbf{f}}^T \\
&= \frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\mathbf{f}\mathbf{f}^T N_f - \bar{\mathbf{f}}\bar{\mathbf{f}}^T - \bar{\mathbf{f}}\bar{\mathbf{f}}^T + \bar{\mathbf{f}}\bar{\mathbf{f}}^T \\
&= \frac{1}{N_t}\sum_{\mathbf{f}\in\mathcal{D}}\mathbf{f}\mathbf{f}^T N_f - \bar{\mathbf{f}}\bar{\mathbf{f}}^T
\end{aligned}$$

[1] The maximum likelihood estimate of the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ is biased. An unbiased estimate for $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the sample covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

However, when N is large, which is the case here, the two estimates are virtually identical.

$$= \begin{bmatrix} \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} k^2 N_{(k,l)} - \bar{k}^2 & \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} kl\, N_{(k,l)} - \bar{k}\,\bar{l} \\ \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} kl\, N_{(k,l)} - \bar{k}\,\bar{l} & \frac{1}{N_t} \sum_{(k,l)\in\mathcal{D}} l^2 N_{(k,l)} - \bar{l}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{kk}^2 & \sigma_{kl}^2 \\ \sigma_{lk}^2 & \sigma_{ll}^2 \end{bmatrix}.$$

A.3 Gaussian Surface Magnitude Estimation

After parameters $\bar{\mathbf{f}}$ and $\mathbf{\Sigma_f}$ are estimated, the squared Gaussian surface fitting error is

$$\mathcal{E}^2 = \sum_{(k,l)\in\mathcal{D}} \left[\, |Y(k,l)| - M_s\, g_s(k,l)\, \right]^2,$$

with a single variable M_s. Taking the derivative of $\mathcal{E}^2$ w.r.t. M_s and setting it to zero,

$$\frac{d\,\mathcal{E}^2}{dM_s} = -2 \sum_{(k,l)\in\mathcal{D}} g_s(k,l)\left[\, |Y(k,l)| - M_s\, g_s(k,l)\,\right] = 0$$

$$\sum_{(k,l)\in\mathcal{D}} g_s(k,l)\, |Y(k,l)| = M_s \sum_{(k,l)\in\mathcal{D}} g_s^2(k,l)$$

$$M_s = \frac{\sum_{(k,l)\in\mathcal{D}} g_s(k,l)\, |Y(k,l)|}{\sum_{(k,l)\in\mathcal{D}} g_s^2(k,l)}.$$

The last equation is (3.19).

Appendix B

Instruction Sheet Used in Human Experiment

Instructions

On the table in front of you, you will find twenty texture images and two sets of adjectives at either end of a board.

Your task is, according to the visual properties of the textures, to **order the twenty test images in one line** between the two sets of adjectives.

The left-most image in your ordering should be the most repetitive, non-random, directional, regular, locally oriented, and uniform; the right-most one should be the most non-repetitive, random, non-directional, irregular, non-oriented, and non-uniform; and the images you place toward the center may contain elements of both sets of properties.

Please ignore the differences in image brightness, and pattern element size.

Please try not to order the images by their common names, such as grass, or bricks; use only their visual properties.

There is **no time limit** for completing the task.

While you are waiting for your turn, please **do not** observe other subjects' ordering.

Thank you.

Appendix C

Estimation Algorithms for Image Database Retrieval

C.1 EM Procedure for Harmonic Confidence Measure

The autocovariance energy ratios of the Brodatz database images have a bi-modal distribution, which is modeled here as two Gaussian distributions. Denote the energy ratios as $\{x_n\}$, $n = 1, \cdots, N$. The conditional density functions of the Gaussian classes, ω_m, $m = 1, 2$, are

$$p(x_n|\omega_m) = \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(x_n-\mu_m)^2}{2\sigma_m^2}}.$$

The initial values of the means μ_m, the variances σ_m, and the prior probabilities $P(\omega_m)$, $m = 1, 2$, are estimated via the K-means clustering of $\{x_n\}$.

The main steps of the EM procedure are as follows [13]:

E-step:

$$P(\omega_m|x_n) = \frac{p(x_n|\omega_m)P(\omega_m)}{\sum_{j=1}^{2} p(x_n|\omega_j)P(\omega_j)} = \frac{\frac{1}{\sigma_m} e^{-\frac{(x_n-\mu_m)^2}{2\sigma_m^2}} P(\omega_m)}{\sum_{j=1}^{2} \frac{1}{\sigma_j} e^{-\frac{(x_n-\mu_j)^2}{2\sigma_j^2}} P(\omega_j)}$$

M-step:

$$\mu_m = \frac{\sum_{n=1}^{N} x_n \, P(\omega_m|x_n)}{\sum_{n=1}^{N} P(\omega_m|x_n)}$$

$$\sigma_m^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_m)^2 \, P(\omega_m|x_n)}{\sum_{n=1}^{N} P(\omega_m|x_n)}$$

$-1,-1$	$-1,0$	$-1,1$
$0,-1$	$\times$	$0,1$
$1,-1$	$1,0$	$1,1$

Figure C-1: Second-order SAR model neighborhood $\mathcal{D}_s$. The center is the current site $s = (s_1, s_2)$.

$$P(\omega_m) = \frac{\sum_{n=1}^{N} P(\omega_m|x_n)}{\sum_{j=1}^{2}\sum_{n=1}^{N} P(\omega_m|x_n)} = \frac{\sum_{n=1}^{N} P(\omega_m|x_n)}{N}$$

The procedure terminates if the absolute change of the log-likelihood

$$\sum_{n=1}^{N} \ln \left\{ \sum_{m=1}^{2} p(x_n|\omega_m)P(\omega_m) \right\}$$

after an EM iteration is lower than a threshold value (0.001 is used in this work).

C.2 Least-squares Estimation of SAR Model Parameters

C.2.1 Estimating SAR Model Parameters

Simultaneous autoregressive models characterize the interaction among neighboring image pixels as a random field linear prediction problem. Given a zero-mean random field $\{y(s)\}$, $s = (s_1, s_2) \in \mathcal{D}$ (region $\mathcal{D}$ is defined by (3.1)), a second-order symmetric SAR model can be expressed as

$$y(s) = \sum_{r \in \mathcal{D}_s} \theta(r)\, y(s + r) + \varepsilon(s), \tag{C.1}$$

where region $\mathcal{D}_s$, as shown in Figure C-1, is the 8-pixel neighborhood of sample site s, parameters $\theta(r) = \theta(-r)$, $r \in \mathcal{D}_s$, are the SAR coefficients, and prediction errors $\varepsilon(s)$, $s \in \mathcal{D}$, are independent and zero-mean random variables with variance σ_s^2.

At any given sample location, the SAR model coefficients can be estimated by using the least-squares estimation (LSE) method within an estimation window $\mathcal{D}_w \in \mathcal{D}$, which typically puts the given sample at the center. The window size is usually determined empirically. In this work, 21×21 windows are used.

Now, consider a particular estimation window $\mathcal{D}_w$. Let $y_1(s), \cdots, y_p(s)$ denote the sample values in neighborhood $\mathcal{D}_s$ of sample site $s \in \mathcal{D}_w$, and $\theta_1, \cdots, \theta_p$ the corresponding SAR coefficients ($p = 8$ for second-order SAR models). Then

$$\varepsilon(s) = y(s) - \sum_{i=1}^{p} \theta_i y_i(s), \qquad s \in \mathcal{D}_w$$

is the prediction error for $y(s)$. Assuming that there are P samples in $\mathcal{D}_w$, let

$$\mathcal{E} = \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(P) \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(P) \end{bmatrix}, \quad \mathbf{\Theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}, \quad \mathbf{y}(s) = \begin{bmatrix} y_1(s) \\ y_2(s) \\ \vdots \\ y_p(s) \end{bmatrix},$$

and

$$Y = \begin{bmatrix} \mathbf{y}(1) & \mathbf{y}(2) & \cdots & \mathbf{y}(P) \end{bmatrix} = \begin{bmatrix} y_1(1) & y_1(2) & \cdots & y_1(P) \\ y_2(1) & y_2(2) & \cdots & y_2(P) \\ \vdots & \vdots & \cdots & \vdots \\ y_p(1) & y_p(2) & \cdots & y_p(P) \end{bmatrix}.$$

Then

$$\mathcal{E} = \mathbf{y} - Y^T \mathbf{\Theta},$$

and the sum of squared errors is

$$\mathcal{E}^T \mathcal{E} = \sum_{s \in \mathcal{D}_w} \varepsilon(s)^2 = \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T Y^T \mathbf{\Theta} + \mathbf{\Theta}^T Y Y^T \mathbf{\Theta}.$$

The sum of squared errors can be minimized by solving equation

$$\frac{d}{d\mathbf{\Theta}}(\mathcal{E}^T \mathcal{E}) = 0.$$

This results in the normal equation

$$Y Y^T \mathbf{\Theta} = Y \mathbf{y},$$

or,

$$\sum_{s \in \mathcal{D}_w} \mathbf{y}(s) \mathbf{y}^T(s) \mathbf{\Theta} = \sum_{s \in \mathcal{D}_w} \mathbf{y}(s) y(s).$$

The SAR coefficient estimates $\hat{\theta}_1, \cdots, \hat{\theta}_p$ are obtained by solving this set of linear equations.

The variance σ_s^2 of the prediction error $\varepsilon(s)$ is estimated by using the estimated SAR model coefficients:

$$\hat{\sigma}_s^2 = \frac{1}{P} \sum_{s \in \mathcal{D}_w} \left[y(s) - \hat{\mathbf{\Theta}}^T \mathbf{y}(s) \right]^2,$$

where $\hat{\mathbf{\Theta}}^T = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_2 & \cdots & \hat{\theta}_p \end{bmatrix}$.

d_3			d_3			d_3
	d_2		d_2		d_2	
		d_1	d_1	d_1		
d_3	d_2	d_1	$\bullet$	d_1	d_2	d_3
		d_1	d_1	d_1		
	d_2		d_2		d_2	
d_3			d_3			d_3

Figure C-2: Second-order MRSAR model neighborhood definitions. Pixel locations labeled as d_l, $l = 1, 2, 3$, are the neighbors of the center pixel at the l-th resolution level.

C.2.2 Multiresolution SAR Modeling

Multiresolution image modeling usually involves filtering and subsampling. However, subsampling sharply reduces the number of available pixels at each scale level. Small number of pixels can be problematic for estimating SAR parameters via the LSE method. An alternative is to vary the SAR neighborhood definitions. Shown in Figure C-2, the second-order MRSAR neighborhood at the l-th resolution level, $\mathcal{D}_{s_l}$, is composed of sample sites d_l.

At each resolution level, four SAR coefficients and the prediction error standard deviation are estimated for every other pixel of an image. For a particular pixel, parameters from different resolution levels are concatenated into one feature vector. The average of these vectors and their covariance matrix form the MRSAR feature set for the image.

Appendix D

Decomposition of Perspective Transformation

D.1 Perspective Projection

As illustrated in Figure D-1, three reference frames are used in analyzing the perspective projection of a three-dimensional (3-D) scene onto a two-dimensional image plane in a pinhole camera: the camera frame X-Y-Z, the image frame x-y, and the surface frame s-t-n.

The origin O of the camera frame is at the projection center (the pinhole), and the Z axis coincides with the camera optical axis. The origin o of the image frame is the intersection of the Z axis and the image plane, with distance d from O. Axes x and y are parallel to the axes X and Y, respectively.

For a locally planar surface in the 3-D scene, the surface coordinate frame s-t-n can be constructed by using a surface point $M = (X_m, Y_m, Z_m)$ as the origin and making the n axis parallel to the surface normal. Denote the 3-D surface as $Z(X, Y)$, and

$$p = \frac{\partial Z(X,Y)}{\partial X}, \qquad q = \frac{\partial Z(X,Y)}{\partial Y}.$$

The surface normal can be expressed as

$$\hat{n} = \frac{1}{r}\,(-p, -q,\, 1) \tag{D.1}$$

with $r = \sqrt{p^2 + q^2 + 1}$. The directions of the s and t axes are chosen such that, when the s-t-n frame is rotated around the unit vector

$$\hat{\omega} = \left(\frac{-q}{\sqrt{p^2+q^2}}, \frac{p}{\sqrt{p^2+q^2}}, 0 \right)$$

to a position where the n axis is parallel to the Z axis, the s and t axes are parallel to the X and Y axes, respectively. The image of the surface point M is $m = (x_m, y_m)$.

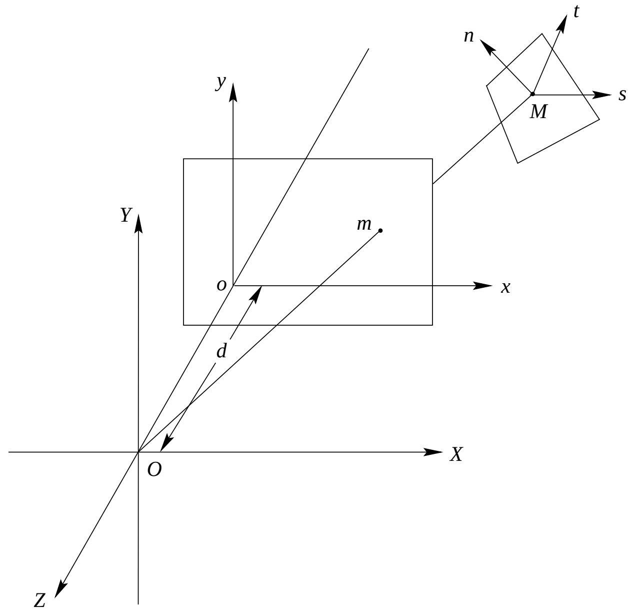


Figure D-1: Reference frames for the perspective projection in a pinhole camera. Camera frame X-Y-Z has origin O at the projection center and axis Z coinciding with the optical axis. Image frame x-y has origin o at the intersection of the Z axis and the image plane, with distance d from O. Axes x and y are parallel to the axes X and Y, respectively. Surface frame s-t-n has origin at surface point $M = (X_m, Y_m, Z_m)$ and axis n is parallel to the surface normal.

For a point in the 3-D scene, its surface coordinates (s, t, n) and its camera coordinates (X, Y, Z) are related by a 4×4 homogeneous transformation matrix T:

$$\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = T \begin{bmatrix} s \\ t \\ n \\ 1 \end{bmatrix} \tag{D.2}$$

where

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t_{21} & t_{22} & t_{23} & t_{24} \\ t_{31} & t_{32} & t_{33} & t_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \frac{1}{r} \begin{bmatrix} \dfrac{p^2 + rq^2}{p^2 + q^2} & \dfrac{pq(1-r)}{p^2 + q^2} & -p & rX_m \\ \dfrac{pq(1-r)}{p^2 + q^2} & \dfrac{rp^2 + q^2}{p^2 + q^2} & -q & rY_m \\ p & q & 1 & rZ_m \\ 0 & 0 & 0 & r \end{bmatrix}. \tag{D.3}$$

A point on the locally planar surface has coordinates $(s, t, 0)$. Applying the transformation matrix T, the corresponding coordinates of the point in the camera frame are

$$\begin{aligned} X &= t_{11}s + t_{12}t + X_m \\ Y &= t_{21}s + t_{22}t + Y_m \\ Z &= t_{31}s + t_{32}t + Z_m \end{aligned} \tag{D.4}$$

Under perspective, the point is projected to the image plane at

$$\begin{aligned} x &= \frac{d}{Z}X = d\,\frac{t_{11}s + t_{12}t + X_m}{t_{31}s + t_{32}t + Z_m} \\ y &= \frac{d}{Z}Y = d\,\frac{t_{21}s + t_{22}t + Y_m}{t_{31}s + t_{32}t + Z_m} \end{aligned} \tag{D.5}$$

Rewrite (D.5) using vector and matrix notations,

$$\mathbf{x} = \frac{\mathbf{A}\mathbf{s} + \mathbf{b}}{\mathbf{c}^T\mathbf{s} + 1} = \begin{bmatrix} \dfrac{a_{11}s + a_{12}t + b_1}{c_1 s + c_2 t + 1} \\ \dfrac{a_{21}s + a_{22}t + b_2}{c_1 s + c_2 t + 1} \end{bmatrix} \tag{D.6}$$

where $\mathbf{s} = [\ s \quad t\ ]^T$, $\mathbf{x} = [\ x \quad y\ ]^T$, and

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \frac{d}{Z_m}\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \tag{D.7}$$

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{d}{Z_m}\begin{bmatrix} X_m \\ Y_m \end{bmatrix} = \begin{bmatrix} x_m \\ y_m \end{bmatrix} \tag{D.8}$$

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \frac{1}{Z_m}\begin{bmatrix} t_{31} \\ t_{32} \end{bmatrix} \tag{D.9}$$

Physically, matrix $\mathbf{A}$ represents image rotation, scaling, and skew, vector $\mathbf{b}$ translation, and vector $\mathbf{c}$ chirping.

D.2 Coordinate Transformations

D.2.1 Perspective Transformation

Denote the planar pattern on the 3-D surface as $f(s,t)$, and the perspective image of this pattern taken by the pinhole camera as $f_p(x,y)$. By Equation (D.6), a surface point (s,t) corresponds to a unique point (x,y) in the image. Therefore, $f_p(x,y) = f(s,t)$ with respect to (D.6).

Inverting Equation (D.6) yields (see Section D.5.1 for derivations)

$$\begin{aligned}
\mathbf{s} &= (\mathbf{A} - \mathbf{x}\mathbf{c}^T)^{-1}(\mathbf{x} - \mathbf{b}) \\
&= \frac{(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{A}^{-1}\mathbf{x} + (\mathbf{A}^{-1}\mathbf{b})(\mathbf{c}^T\mathbf{A}^{-1})\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}}{1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}} \\
&= \frac{\mathbf{A}^p\mathbf{x} + \mathbf{b}^p}{\mathbf{c}^{pT}\mathbf{x} + 1}
\end{aligned} \tag{D.10}$$

where

$$\begin{aligned}
\mathbf{A}^p &= \begin{bmatrix} a^p_{11} & a^p_{12} \\ a^p_{21} & a^p_{22} \end{bmatrix} = (1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{A}^{-1} + (\mathbf{A}^{-1}\mathbf{b})(\mathbf{c}^T\mathbf{A}^{-1}) \\
&= \frac{1}{|\mathbf{A}|}\begin{bmatrix} a_{22} - c_2 b_2 & -a_{12} + c_2 b_1 \\ -a_{21} + c_1 b_2 & a_{11} - c_1 b_1 \end{bmatrix}
\end{aligned} \tag{D.11}$$

$$\mathbf{b}^p = \begin{bmatrix} b^p_1 \\ b^p_2 \end{bmatrix} = -\mathbf{A}^{-1}\mathbf{b} = -\frac{1}{|\mathbf{A}|}\begin{bmatrix} a_{22}b_1 - a_{12}b_2 \\ -a_{21}b_1 + a_{11}b_2 \end{bmatrix} \tag{D.12}$$

$$\mathbf{c}^p = \begin{bmatrix} c^p_1 \\ c^p_2 \end{bmatrix} = -\mathbf{A}^{-T}\mathbf{c} = -\frac{1}{|\mathbf{A}|}\begin{bmatrix} a_{22}c_1 - a_{21}c_2 \\ -a_{12}c_1 + a_{11}c_2 \end{bmatrix} \tag{D.13}$$

and $|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$ is the determinant of $\mathbf{A}$.

Consequently,

$$f_p(x, y) = f\left(\frac{a^p_{11}x + a^p_{12}y + b^p_1}{c^p_1 x + c^p_2 y + 1}, \frac{a^p_{21}x + a^p_{22}y + b^p_2}{c^p_1 x + c^p_2 y + 1}\right) \tag{D.14}$$

D.2.2 Affine Transformation

The coordinate transformation between image $f(s, t)$ and its affine image $f_a(u, v)$ is

$$\mathbf{u} = \begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{A}\mathbf{s} + \mathbf{b} = \begin{bmatrix} a_{11}s + a_{12}t + b_1 \\ a_{21}s + a_{22}t + b_2 \end{bmatrix} \tag{D.15}$$

and

$$\mathbf{s} = \mathbf{A}^{-1}(\mathbf{u} - \mathbf{b}) = \mathbf{A}^a\mathbf{u} + \mathbf{b}^a, \tag{D.16}$$

where

$$\mathbf{A}^a = \begin{bmatrix} a^a_{11} & a^a_{12} \\ a^a_{21} & a^a_{22} \end{bmatrix} = \mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{D.17}$$

$$\mathbf{b}^a = \begin{bmatrix} b^a_1 \\ b^a_2 \end{bmatrix} = -\mathbf{A}^{-1}\mathbf{b} = \mathbf{b}^p \tag{D.18}$$

Therefore,

$$f_a(u,v) = f(a_{11}^a u + a_{12}^a v + b_1^a,\ a_{21}^a u + a_{22}^a v + b_2^a) \tag{D.19}$$

D.2.3 Chirp Transformation

The chirp transformation is defined as the coordinate transformation between image $f_a(u,v)$ and $f_p(x,y)$. Equations (D.6), (D.15), and (D.16) can be used to obtain (see Section D.5.1 for derivations)

$$\mathbf{x} = \frac{\mathbf{u}}{\mathbf{c}^T\mathbf{A}^{-1}\mathbf{u} + (1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})} = \frac{\mathbf{u}}{-\mathbf{c}^{pT}\mathbf{u} + a^c} \tag{D.20}$$

and

$$\mathbf{u} = (\mathbf{I} - \mathbf{x}\mathbf{c}^T\mathbf{A}^{-1})^{-1}\mathbf{x}(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b}) = \frac{(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\,\mathbf{x}}{1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}} = \frac{\mathbf{A}^c\,\mathbf{x}}{\mathbf{c}^{cT}\mathbf{x} + 1} \tag{D.21}$$

where $\mathbf{I}$ is a 2×2 identity matrix and

$$a^c = 1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b} = 1 - \frac{1}{|\mathbf{A}|}(a_{22}c_1 - a_{21}c_2)b_1 - \frac{1}{|\mathbf{A}|}(-a_{12}c_1 + a_{11}c_2)b_2 \tag{D.22}$$

$$\mathbf{A}^c = (1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\,\mathbf{I} = a^c\,\mathbf{I} \tag{D.23}$$

$$\mathbf{c}^c = \begin{bmatrix} c_1^c \\ c_2^c \end{bmatrix} = -\mathbf{A}^{-T}\mathbf{c} = \mathbf{c}^p \tag{D.24}$$

Therefore,

$$f_p(x,y) = f_a\left(\frac{a^c x}{c_1^c x + c_2^c y + 1}, \frac{a^c y}{c_1^c x + c_2^c y + 1}\right) \tag{D.25}$$

D.2.4 Relationship Among Three Coordinate Transformations

The basic parameters for the perspective, affine, and chirp coordinate transformations are

$$\begin{aligned}
\mathbf{A}^a &= \mathbf{A}^{-1} \\
\mathbf{A}^c &= a^c\,\mathbf{I} = (1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{I} \\
\mathbf{A}^p &= a^c\mathbf{A}^a + \mathbf{b}^a\mathbf{c}^{cT} \\
\mathbf{b}^a &= \mathbf{b}^p = -\mathbf{A}^{-1}\mathbf{b} \\
\mathbf{c}^c &= \mathbf{c}^p = -\mathbf{A}^{-T}\mathbf{c}
\end{aligned} \tag{D.26}$$

The relationship of the transformations is summarized in Table D.1.

D.3 Decomposition of Perspective Transformation

D.3.1 Coordinate Transformations as Linear System Operations

The three coordinate transformations in discussion can be characterized as three linear system operations (see Section D.5.2 for proof of linearity). Consequently, expression (D.14), (D.19), and (D.25) are the output of the corresponding systems.

	Perspective	Affine	Chirp
$\mathbf{A}$	$\mathbf{A}^p = a^c\mathbf{A}^a + \mathbf{b}^a\mathbf{c}^{cT}$	$\mathbf{A}^a = \mathbf{A}^{-1}$	$\mathbf{A}^c = a^c\,\mathbf{I} = (1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{I}$
$\mathbf{b}$	$\mathbf{b}^p = \mathbf{b}^a = -\mathbf{A}^{-1}\mathbf{b}$	$\mathbf{b}^a = \mathbf{b}^p = -\mathbf{A}^{-1}\mathbf{b}$	—
$\mathbf{c}$	$\mathbf{c}^p = \mathbf{c}^c = -\mathbf{A}^{-T}\mathbf{c}$	—	$\mathbf{c}^c = \mathbf{c}^p = -\mathbf{A}^{-T}\mathbf{c}$
Coordinate Transform	$\mathbf{x} = \dfrac{\mathbf{A}\mathbf{s}+\mathbf{b}}{\mathbf{c}^T\mathbf{s}+1}$ $\mathbf{s} = \dfrac{\mathbf{A}^p\mathbf{x}+\mathbf{b}^p}{\mathbf{c}^{pT}\mathbf{x}+1}$	$\mathbf{u} = \mathbf{A}\mathbf{s}+\mathbf{b}$ $\mathbf{s} = \mathbf{A}^a\mathbf{u}+\mathbf{b}^a$	$\mathbf{x} = \dfrac{\mathbf{u}}{-\mathbf{c}^{cT}\mathbf{u}+a^c}$ $\mathbf{u} = \dfrac{\mathbf{A}^c\,\mathbf{x}}{\mathbf{c}^{cT}\mathbf{x}+1}$

Table D.1: Parameters of perspective, affine, and chirp transformations.

Given an input $f(x, y)$ to a general linear system, the output can be written as

$$g(x,y) = \iint_{-\infty}^{+\infty} h(x,y;s,t)\, f(s,t)\, ds\, dt \tag{D.27}$$

where $h(x, y; s, t)$ is the system impulse response. Hence, Equations (D.14), (D.19), and (D.25) can be written as

$$\begin{aligned} f_p(x,y) &= \iint_{-\infty}^{+\infty} h_p(x,y;s,t)\, f(s,t)\, ds\, dt && \text{(D.28)} \\ f_a(u,v) &= \iint_{-\infty}^{+\infty} h_a(u,v;s,t)\, f(s,t)\, ds\, dt && \text{(D.29)} \\ f_p(x,y) &= \iint_{-\infty}^{+\infty} h_c(x,y;u,v)\, f_a(u,v)\, du\, dv && \text{(D.30)} \end{aligned}$$

where

$$\begin{aligned} h_p(x,y;s,t) &= \delta\left(\frac{a^p_{11}x + a^p_{12}y + b^p_1}{c^p_1 x + c^p_2 y + 1} - s,\ \frac{a^p_{21}x + a^p_{22}y + b^p_2}{c^p_1 x + c^p_2 y + 1} - t\right) && \text{(D.31)} \\ h_a(u,v;s,t) &= \delta\left(a^a_{11}u + a^a_{12}v + b^a_1 - s,\ a^a_{21}u + a^a_{22}v + b^a_2 - t\right) && \text{(D.32)} \\ h_c(x,y;u,v) &= \delta\left(\frac{a^c x}{c^c_1 x + c^c_2 y + 1} - u,\ \frac{a^c y}{c^c_1 x + c^c_2 y + 1} - v\right) && \text{(D.33)} \end{aligned}$$

and the $\delta(\cdot)$'s are two-dimensional Dirac delta functions.

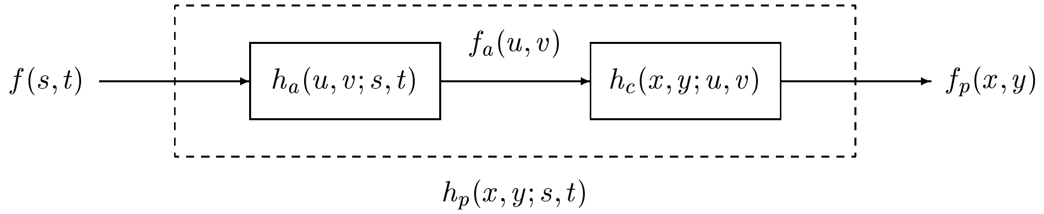


Figure D-2: Decomposition of image perspective transformations.

D.3.2 Decomposition of Perspective Transformation

A perspective transformation can be decomposed into an affine and a chirp transformation. This decomposition is shown as three linear systems in Figure D-2. The three system impulse responses are related as follows (see Section D.5.3 for proof):

$$h_p(x,y;s,t) = \iint_{-\infty}^{+\infty} h_c(x,y;u,v)\, h_a(u,v;s,t)\, du\, dv \tag{D.34}$$

D.3.3 Example of Perspective Decomposition

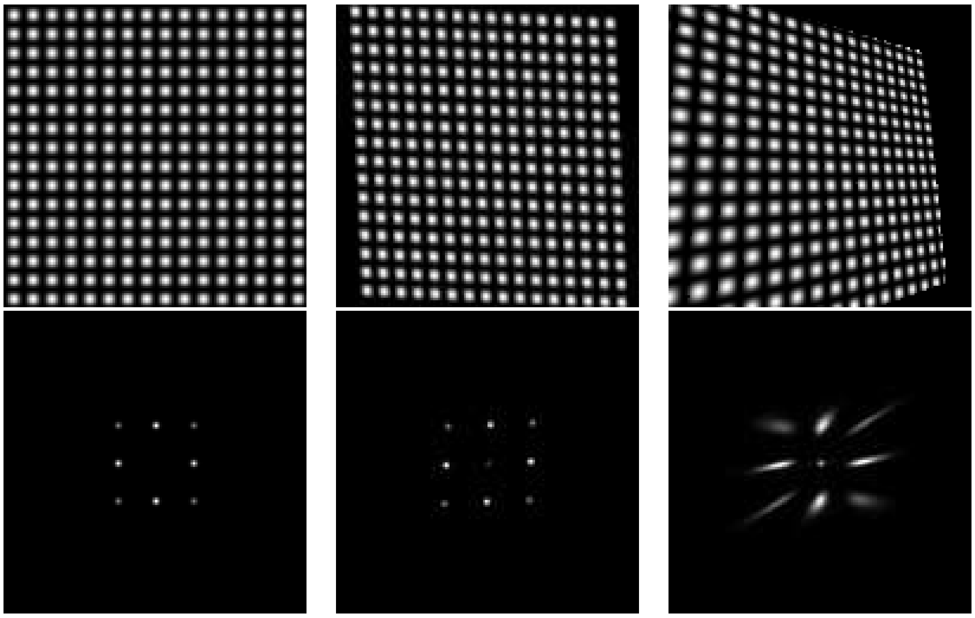
Figure D-3 shows an example of perspective decomposition. The top row pictures are the spatial images $f(s,t)$, $f_a(u,v)$, and $f_p(x,y)$, while the bottom row contains the corresponding Fourier magnitude images $|F(\sigma,\tau)|$, $|F_a(\mu,\nu)|$, and $|F_p(\xi,\eta)|$. When computing the Fourier magnitudes, the spatial images are first zero-meaned and tapered by the Gaussian window function shown in Figure 3-2 (b).

The spatial image $f(s,t)$ in Figure D-3 (a) is a 2-D sinusoidal grating pattern:

$$f(s,t) = [1 + sin(\omega_s s + \phi_s)][1 + sin(\omega_t t + \phi_t)],$$

where the spatial frequencies $\omega_s = \omega_t = 16$ (normalized by the image size) and phases $\phi_s = \phi_t = -90°$. The parameters used in the image transformations are $p = 0.5$, $q = 0.2$, and $d = Z_p = 100$. The perspective image in (c) can be obtained by applying either a perspective transformation on the original image in (a) or a chirp transformation on the affine image in (b). The centers of the top row images are on the Z axis of the camera frame.

Shown in the Fourier magnitude images in Figure D-3 (b) and (c), spatial affine and chirp transformations have distinct spectral signatures. While the affine transformation gives rise to the frequency shifts of the spectral harmonic peaks, the chirp transformation is responsible to the peak deformation. The relationship among the Fourier transforms of the original, the affine, and the perspective images is examined in the next section.



(a) Original (b) Affine (c) Perspective (Chirp on Affine)

Figure D-3: Example of perspective decomposition. Top row: spatial images $f(s,t)$, $f_a(u,v)$, and $f_p(x,y)$. Bottom row: corresponding Fourier magnitudes $F(\sigma,\tau)$, $F_a(\mu,\nu)$, and $F_p(\xi,\eta)$. Shown in (b) and (c), spatial affine and chirp transformations leave distinct spectral signatures.

D.4 Relationship in Frequency Domain

Define the Fourier transforms of image $f(s,t)$, $f_a(u,v)$, and $f_p(x,y)$ as

$$F(\sigma,\tau) = \iint\limits_{-\infty}^{+\infty} f(s,t)\, e^{-j2\pi(\sigma s+\tau t)}\, ds\, dt \tag{D.35}$$

$$F_a(\mu,\nu) = \iint\limits_{-\infty}^{+\infty} f_a(u,v)\, e^{-j2\pi(\mu u+\nu v)} du\, dv \tag{D.36}$$

$$F_p(\xi,\eta) = \iint\limits_{-\infty}^{+\infty} f_p(x,y)\, e^{-j2\pi(\xi x+\eta y)} dx\, dy \tag{D.37}$$

The focus here is on deriving the expression of $F_a(\mu,\nu)$ in terms of $F(\sigma,\tau)$, and $F_p(\xi,\eta)$ in terms of $F_a(\mu,\nu)$. With these expressions, it is straightforward to obtain the relationship between the perspective pair $F_p(\xi,\eta)$ and $F(\sigma,\tau)$.

D.4.1 Affine Transformation

For the affine pair $f(s,t)$ and $f_a(u,v)$, it can be shown that their Fourier transforms $F(\sigma,\tau)$ and $F_a(\mu,\nu)$ also have an affine relationship (see Section D.5.4 for derivations):

$$F_a(\mu,\nu) = |\mathbf{A}|\, e^{-j2\pi(b_1\mu+b_2\nu)} F\left(a_{11}\mu + a_{21}\nu, a_{12}\mu + a_{22}\nu\right) \tag{D.38}$$

By Equation (D.38), the Fourier transform of the affine image is a rotated, skewed, and phase-shifted version of that of the original image. The frequency coordinate transformations are:

$$\begin{bmatrix} \sigma \\ \tau \end{bmatrix} = \mathbf{A}^T \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \qquad \begin{bmatrix} \mu \\ \nu \end{bmatrix} = \mathbf{A}^{-T} \begin{bmatrix} \sigma \\ \tau \end{bmatrix}.$$

The effect of affine transformation is demonstrated in Figure D-3 (b). Although some of the spectral harmonic peaks are shifted, the shape of the peaks is preserved.

D.4.2 Chirp Transformation

Unlike the affine case, the chirp transformation does not preserve image homogeneity. As a consequence, different areas in an image usually have different frequency content, even though the original pattern is homogeneous. In practice, a tapering window can be used to isolate the area of interest in the perspective image.

Centering a tapering window $g(x,y)$ at location (x_m, y_m) in the perspective image $f_p(x,y)$, the Fourier transform of the image is

$$\begin{aligned} F_p(\xi,\eta) &= \iint_{-\infty}^{+\infty} f_p(x,y)\, g(x-x_m, y-y_m)\, e^{-j2\pi(\xi x+\eta y)} dx\, dy \\ &= \iint_{-\infty}^{+\infty} F_a(\alpha,\beta)\, H_c(\xi,\eta;\alpha,\beta)\, d\alpha\, d\beta \end{aligned} \tag{D.39}$$

where

$$H_c(\xi,\eta;\alpha,\beta) = \iint_{-\infty}^{+\infty} g(x-x_m, y-y_m)\, e^{-j2\pi\left(\xi x - \frac{a^c \alpha x}{c_1^c x + c_2^c y + 1} + \eta y - \frac{a^c \beta y}{c_1^c x + c_2^c y + 1}\right)} dx\, dy \tag{D.40}$$

From Equation (D.40) and Figure D-3, it can be seen that the integration kernel $H_c(\xi,\eta;\alpha,\beta)$, which is responsible for the spectral harmonic peak deformation, is a shift-variant function. If the kernel H_c can be evaluated and expressed in a form that allows numerical modeling, it is conceivable that the perspective parameters p and q, which define the surface normal, can be recovered from the shape of the peak deformation. The kernel evaluation is a difficult problem, and is currently studied in a joint effort with researchers from the Mathematics Departments of MIT and the Harvard University.

D.5 Derivations and Proofs

D.5.1 Coordinate Transforms

First, the derivation of (D.21). From $\mathbf{I} - \mathbf{x}\mathbf{c}^T\mathbf{A}^{-1} = (\mathbf{A} - \mathbf{x}\mathbf{c}^T)\mathbf{A}^{-1}$,

$$\begin{aligned}
|\mathbf{I} - \mathbf{x}\mathbf{c}^T\mathbf{A}^{-1}| &= |\mathbf{A} - \mathbf{x}\mathbf{c}^T|\,\frac{1}{|\mathbf{A}|} \\
&= \frac{1}{|\mathbf{A}|}\left[(a_{11}a_{22} - a_{12}a_{21}) - (a_{22}c_1 - a_{21}c_2)x - (-a_{12}c_1 + a_{11}c_2)y\right] \\
&= 1 - \frac{1}{|\mathbf{A}|}\left[\begin{array}{cc} a_{22}c_1 - a_{21}c_2 & -a_{12}c_1 + a_{11}c_2 \end{array}\right]\mathbf{x} \\
&= 1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}
\end{aligned}$$

and

$$\begin{aligned}
(\mathbf{I} - \mathbf{x}\mathbf{c}^T\mathbf{A}^{-1})^{-1}\mathbf{x} &= (\mathbf{I} - \mathbf{x}\,\mathbf{c}^{pT})^{-1}\mathbf{x} \\
&= \left(\mathbf{I} - \left[\begin{array}{c} x \\ y \end{array}\right]\left[\begin{array}{cc} c_1^p & c_2^p \end{array}\right]\right)^{-1}\mathbf{x} = \left[\begin{array}{cc} 1 - c_1^p x & -c_2^p x \\ -c_1^p y & 1 - c_2^p y \end{array}\right]^{-1}\mathbf{x} \\
&= \frac{1}{|\mathbf{I} - \mathbf{x}\mathbf{c}^T\mathbf{A}^{-1}|}\left[\begin{array}{cc} 1 - c_2^p y & c_2^p x \\ c_1^p y & 1 - c_1^p x \end{array}\right]\left[\begin{array}{c} x \\ y \end{array}\right] \\
&= \frac{\mathbf{x}}{|\mathbf{I} - \mathbf{x}\mathbf{c}^T\mathbf{A}^{-1}|} = \frac{\mathbf{x}}{1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}}
\end{aligned}$$

Hence, Equation (D.21).

Equation (D.10) can be obtained by bringing (D.21) into (D.16):

$$\begin{aligned}
\mathbf{s} &= \mathbf{A}^{-1}\mathbf{u} - \mathbf{A}^{-1}\mathbf{b} = \frac{(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{A}^{-1}\mathbf{x}}{1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}} - \mathbf{A}^{-1}\mathbf{b} \\
&= \frac{(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{A}^{-1}\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x})}{1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}} \\
&= \frac{(1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{b})\mathbf{A}^{-1}\mathbf{x} + (\mathbf{A}^{-1}\mathbf{b})(\mathbf{c}^T\mathbf{A}^{-1})\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}}{1 - \mathbf{c}^T\mathbf{A}^{-1}\mathbf{x}}
\end{aligned}$$

D.5.2 System Linearity

Proof:

Denote the affine system operator as $\mathcal{L}$. Given two input signals $f_1(s,t)$ and $f_2(s,t)$, by Equation (D.19), the output signals of the system are

$$\begin{aligned}
f_{a_1}(x,y) &= \mathcal{L}\{f_1(s,t)\} = f_1\left(a_{11}^a x + a_{12}^a y + b_1^a,\ a_{21}^a x + a_{22}^a y + b_2^a\right) \\
f_{a_2}(x,y) &= \mathcal{L}\{f_2(s,t)\} = f_2\left(a_{11}^a x + a_{12}^a y + b_1^a,\ a_{21}^a x + a_{22}^a y + b_2^a\right)
\end{aligned}$$

For input $\alpha_1 f_1(s,t) + \alpha_2 f_2(s,t)$, where α_1 and α_2 are arbitrary constants, the output is

$$\begin{aligned}
&\mathcal{L}\{\alpha_1 f_1(s,t) + \alpha_2 f_2(s,t)\} \\
&\quad = \alpha_1 f_1\left(a_{11}^a x + a_{12}^a y + b_1^a,\, a_{21}^a x + a_{22}^a y + b_2^a\right) + \alpha_2 f_2\left(a_{11}^a x + a_{12}^a y + b_1^a,\, a_{21}^a x + a_{22}^a y + b_2^a\right) \\
&\quad = \alpha_1 \mathcal{L}\{f_1(s,t)\} + \alpha_2 \mathcal{L}\{f_2(s,t)\}
\end{aligned}$$

Therefore, the system is linear. The linearity of the chirp and perspective systems can be shown in a similar manner.

■

D.5.3 Relationship of Three Impulse Responses

Proof:

$$\begin{aligned}
&\iint_{-\infty}^{+\infty} h_c(x,y;u,v)\, h_a(u,v;s,t)\, du\, dv \\
&\quad = \iint_{-\infty}^{+\infty} \delta\left(\frac{a^c x}{c_1^c x + c_2^c y + 1} - u, \frac{a^c y}{c_1^c x + c_2^c y + 1} - v\right) \times \\
&\qquad\qquad \delta\left(a_{11}^a u + a_{12}^a v + b_1^a - s,\, a_{21}^a u + a_{22}^a v + b_2^a - t\right) du\, dv \\
&\quad = \delta\left(a_{11}^a u + a_{12}^a v + b_1^a - s,\, a_{21}^a u + a_{22}^a v + b_2^a - t\right) \Big|_{\substack{u = \frac{a^c x}{c_1^c x + c_2^c y + 1} \\ v = \frac{a^c y}{c_1^c x + c_2^c y + 1}}} \\
&\quad = \delta(\alpha, \beta)
\end{aligned}$$

$$\begin{aligned}
\begin{bmatrix} \alpha \\ \beta \end{bmatrix} &= \mathbf{A}^a \mathbf{u} + \mathbf{b}^a - \mathbf{s} \Big|_{\mathbf{u} = \frac{a^c \mathbf{x}}{\mathbf{c}^{cT}\mathbf{x} + 1}} = \mathbf{A}^{-1}\left(\frac{1 - \mathbf{c}^T \mathbf{A}^{-1}\mathbf{b}}{1 - \mathbf{c}^T \mathbf{A}^{-1}\mathbf{x}}\right)\mathbf{x} - \mathbf{A}^{-1}\mathbf{b} - \mathbf{s} \\
&= \frac{(1 - \mathbf{c}^T \mathbf{A}^{-1}\mathbf{b})\mathbf{A}^{-1}\mathbf{x} + (\mathbf{A}^{-1}\mathbf{b})(\mathbf{c}^T \mathbf{A}^{-1})\mathbf{x} - \mathbf{A}^{-1}\mathbf{b}}{1 - \mathbf{c}^T \mathbf{A}^{-1}\mathbf{x}} - \mathbf{s} \\
&= \frac{\mathbf{A}^p \mathbf{x} + \mathbf{b}^p}{\mathbf{c}^{pT}\mathbf{x} + 1} - \mathbf{s}
\end{aligned}$$

Hence,

$$\delta(\alpha, \beta) = \delta\left(\frac{a_{11}^p x + a_{12}^p y + b_1^p}{c_1^p x + c_2^p y + 1} - s,\, \frac{a_{21}^p x + a_{22}^p y + b_2^p}{c_1^p x + c_2^p y + 1} - t\right) = h_p(x,y;s,t)$$

■

D.5.4 Fourier Transform of Affine Images

Using (D.19),

$$
\begin{aligned}
F_a(\mu,\nu) &= \iint_{-\infty}^{+\infty} f_a(u,v)\, e^{-j2\pi(\mu u+\nu v)} du\, dv \\
&= \iint_{-\infty}^{+\infty} e^{-j2\pi(\mu u+\nu v)} \left[\iint_{-\infty}^{+\infty} f(s,t)\, \delta\left(a_{11}^a u + a_{12}^a v + b_1^a - s,\; a_{21}^a u + a_{22}^a v + b_2^a - t\right) ds\, dt \right] du\, dv \\
&= \iint_{-\infty}^{+\infty} e^{-j2\pi(\mu u+\nu v)} \times \\
&\qquad \left[\iint_{-\infty}^{+\infty} f(s,t) \left(\iint_{-\infty}^{+\infty} e^{-j2\pi\left\{\alpha\left[s-\left(a_{11}^a u+a_{12}^a v+b_1^a\right)\right]+\beta\left[t-\left(a_{21}^a u+a_{22}^a v+b_2^a\right)\right]\right\}} d\alpha\, d\beta \right) ds\, dt \right] du\, dv \\
&= \iint_{-\infty}^{+\infty} \left[\iint_{-\infty}^{+\infty} f(s,t)\, e^{-j2\pi(\alpha s+\beta t)} ds\, dt \;\times \right. \\
&\qquad \left. \iint_{-\infty}^{+\infty} e^{-j2\pi\left[\mu u-\alpha\left(a_{11}^a u+a_{12}^a v+b_1^a\right)+\nu v-\beta\left(a_{21}^a u+a_{22}^a v+b_2^a\right)\right]} du\, dv \right] d\alpha\, d\beta \\
&= \iint_{-\infty}^{+\infty} F(\alpha,\beta)\, e^{j2\pi\left(b_1^a\alpha+b_2^a\beta\right)} \left[\iint_{-\infty}^{+\infty} e^{-j2\pi\left[\left(\mu-a_{11}^a\alpha-a_{21}^a\beta\right)u+\left(\nu-a_{12}^a\alpha-a_{22}^a\beta\right)v\right]} du\, dv \right] d\alpha\, d\beta \\
&= \iint_{-\infty}^{+\infty} F(\alpha,\beta)\, e^{j2\pi\left(b_1^a\alpha+b_2^a\beta\right)}\, \delta\left(\mu - a_{11}^a\alpha - a_{21}^a\beta, \nu - a_{12}^a\alpha - a_{22}^a\beta\right) d\alpha\, d\beta
\end{aligned}
$$

Substituting variables as $\alpha' = a_{11}^a\alpha + a_{21}^a\beta$ and $\beta' = a_{12}^a\alpha + a_{22}^a\beta$,

$$
\begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} = \mathbf{A}^{aT} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \qquad\qquad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \mathbf{A}^{a-T} \begin{bmatrix} \alpha' \\ \beta' \end{bmatrix}
$$

The Jacobian of this variable substitution is

$$
J = \begin{vmatrix} \dfrac{\partial\alpha'}{\partial\alpha} & \dfrac{\partial\alpha'}{\partial\beta} \\ \dfrac{\partial\beta'}{\partial\alpha} & \dfrac{\partial\beta'}{\partial\beta} \end{vmatrix}^{-1} = \begin{vmatrix} a_{11}^a & a_{21}^a \\ a_{12}^a & a_{22}^a \end{vmatrix}^{-1} = \frac{1}{|\mathbf{A}^a|}.
$$

Therefore,

$$F_a(\mu,\nu) = \frac{1}{|\mathbf{A}^a|}\iint_{-\infty}^{+\infty} F\left(\frac{a^a_{22}}{|\mathbf{A}^a|}\alpha' - \frac{a^a_{21}}{|\mathbf{A}^a|}\beta',\, -\frac{a^a_{12}}{|\mathbf{A}^a|}\alpha' + \frac{a^a_{11}}{|\mathbf{A}^a|}\beta'\right) e^{j2\pi(u_0\alpha'+v_0\beta')} \times \delta(\mu-\alpha',\nu-\beta')\, d\alpha'\, d\beta'$$

$$= \frac{1}{|\mathbf{A}^a|} e^{j2\pi(u_0\mu+v_0\nu)} F\left(\frac{a^a_{22}}{|\mathbf{A}^a|}\mu - \frac{a^a_{21}}{|\mathbf{A}^a|}\nu,\, -\frac{a^a_{12}}{|\mathbf{A}^a|}\mu + \frac{a^a_{11}}{|\mathbf{A}^a|}\nu\right)$$

where

$$u_0 = \frac{1}{|\mathbf{A}^a|}(a^a_{22}b^a_1 - a^a_{12}b^a_2), \qquad v_0 = \frac{1}{|\mathbf{A}^a|}(-a^a_{21}b^a_1 + a^a_{11}b^a_2).$$

Using (D.26),

$$\frac{1}{|\mathbf{A}^a|} = |\mathbf{A}|$$

$$\begin{bmatrix} u_0 \\ v_0 \end{bmatrix} = \mathbf{A}^{a^{-1}}\mathbf{b}^a = \mathbf{A}\mathbf{b}^a = \mathbf{A}(-\mathbf{A}^{-1}\mathbf{b}) = -\mathbf{b}$$

$$\frac{1}{|\mathbf{A}^a|}\begin{bmatrix} a^a_{22} & -a^a_{21} \\ -a^a_{12} & a^a_{11} \end{bmatrix}\begin{bmatrix} \mu \\ \nu \end{bmatrix} = \mathbf{A}^{a^{-T}}\begin{bmatrix} \mu \\ \nu \end{bmatrix} = \mathbf{A}^T\begin{bmatrix} \mu \\ \nu \end{bmatrix}.$$

Hence, Equation (D.38).

D.5.5 Fourier Transform of Chirp Images

Using (D.25),

$$\begin{aligned} F_p(\xi,\eta) &= \iint_{-\infty}^{+\infty} f_p(x,y)\, g(x-x_m, y-y_m)\, e^{-j2\pi(\xi x+\eta y)} dx\, dy \\ &= \iint_{-\infty}^{+\infty} g(x-x_m, y-y_m)\, e^{-j2\pi(\xi x+\eta y)} \times \\ &\qquad \left[\iint_{-\infty}^{+\infty} f_a(u,v)\, \delta\left(\frac{a^c x}{c^c_1 x + c^c_2 y + 1} - u,\, \frac{a^c y}{c^c_1 x + c^c_2 y + 1} - v\right) du\, dv\right] dx\, dy \\ &= \iint_{-\infty}^{+\infty} g(x-x_m, y-y_m)\, e^{-j2\pi(\xi x+\eta y)} \times \\ &\qquad \left[\iint_{-\infty}^{+\infty} f_a(u,v) \left(\iint_{-\infty}^{+\infty} e^{-j2\pi\left[\alpha\left(u - \frac{a^c x}{c^c_1 x + c^c_2 y + 1}\right) + \beta\left(v - \frac{a^c y}{c^c_1 x + c^c_2 y + 1}\right)\right]} d\alpha\, d\beta\right) du\, dv\right] dx\, dy \end{aligned}$$

$$= \iint\limits_{-\infty}^{+\infty} \left[\iint\limits_{-\infty}^{+\infty} f_a(u,v)\, e^{-j2\pi(\alpha u+\beta v)} du\, dv \;\times \right.$$
$$\left. \iint\limits_{-\infty}^{+\infty} g(x-x_m, y-y_m)\, e^{-j2\pi\left(\xi x - \frac{a^c \alpha x}{c_1^c x + c_2^c y + 1} + \eta y - \frac{a^c \beta y}{c_1^c x + c_2^c y + 1}\right)} dx\, dy \right] d\alpha\, d\beta$$

$$= \iint\limits_{-\infty}^{+\infty} F_a(\alpha,\beta) \left[\iint\limits_{-\infty}^{+\infty} g(x-x_m, y-y_m)\, e^{-j2\pi\left(\xi x - \frac{a^c \alpha x}{c_1^c x + c_2^c y + 1} + \eta y - \frac{a^c \beta y}{c_1^c x + c_2^c y + 1}\right)} dx\, dy \right] d\alpha\, d\beta$$

Bibliography

[1] M. Allmen and C.R. Dyer. Cyclic motion detection using spatiotemporal surface and curves. In *Proc. Int. Conf. on Pattern Recognition*, pages 365–370, 1990.

[2] M. Amadasun. Textural features corresponding to textural properties. *IEEE T. Sys., Man, and Cyber.*, 19(5):1264–1274, Sep./Oct. 1989.

[3] J. Beck. Effect of orientation and of shape similarity on perceptual grouping. *Perception & Psychophysics*, 1:300–302, 1966.

[4] J. Beck. Perceptual grouping produced by changes in orientation and shape. *Science*, 154:538–540, 1966.

[5] J. Beck. Perceptual grouping produced by like figures. *Perception & Psychophysics*, 2:491–495, 1967.

[6] J. Beck, A. Sutter, and R. Ivry. Spatial frequency channels and perceptual grouping in texture segregation. *Computer Vision, Graph., and Image Proc.*, 37:299–325, 1987.

[7] J. R. Bergen. Theories of visual texture perception. In D.M. Regan, editor, *Spatial Vision*, Vision and Visual Dysfunction, vol. 10, pages 114–134. CRC Press, 1991.

[8] J. R. Bergen and E. H. Adelson. Visual texture segmentation based on energy measures. *J. Opt. Soc. of Amer. A*, 3(13), 1986.

[9] J. R. Bergen and E. H. Adelson. Early vision and texture perception. *Nature*, 333:363–364, 1988.

[10] J. R. Bergen and M. S. Landy. Computational modeling of visual texture segregation. In M. S. Landy and J. A. Movshon, editors, *Computational Models of Visual Processing*, pages 253–271, Cambridge, MA, 1991. MIT Press.

[11] J. Bergen *et al.*. Hierarchial model-based motion estimation. In *Proc. Europ. Conf. on Computer Vision*, pages 237–252, 1992.

[12] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc., Ser. B*, 36:192–236, 1974.

[13] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[14] A.F. Bobick and J.W. Davis. Real-time recognition of activity using temporal templates. In *Proc. Third IEEE Workshop on Appl. of Computer Vision*, pages 39–42, Sarasota, FL, December 1996.

[15] P. Brodatz. *Textures: A Photographic Album for Artists and Designers.* Dover, New York, 1966.

[16] T.M. Caelli. On discriminating visual textures and images. *Perception & Psychophysics*, 31(2):149–159, 1982.

[17] T. Chang and C.-C. J. Kuo. Texture analysis and classification with tree-structured wavelet transform. *IEEE T. Image Processing*, 2(4):429–441, Oct. 1993.

[18] T.P. Chiang. On the line extrapolation of a continuous homogeneous random field. *Theory Prob. Appl.*, 2:58–89, 1957.

[19] T.P. Chiang. The prediction theory of stationary random fields. III. Fourfold Wold decompositions. *J. of Multivariate Anal.*, 37:46–65, 1991.

[20] H. Cramér. On some classes of nonstationary stochastic processes. In *Proc. Fourth Berkeley Symp. on Mathematics, Statistics, and Probability*, pages 57–77. University of California Press, 1961.

[21] G. R. Cross and A. K. Jain. Markov random field texture models. *IEEE T. Pat. Analy. and Machine Intel.*, PAMI-5(1):25–39, 1983.

[22] H. Derin and P. A. Kelly. Discrete-index Markov-type random processes. *Proceedings of IEEE*, pages 1485–1510, Oct. 1989.

[23] R. Duda and I. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, Inc., 1973.

[24] M. P. Ekstrom and J. W. Woods. Two-dimensional spectral factorization with applications in recursive digital filtering. *IEEE T. Acous., Speech, and Signal Proc.*, ASSP-24(2):115–128, April 1976.

[25] I. M. Elfadel and R. W. Picard. Gibbs random fields, co-occurrences, and texture modeling. *IEEE T. Pat. Analy. and Machine Intel.*, 16(1):24–37, Jan. 1994.

[26] J. M. Francos. Orthogonal decompositions of 2-D random fields and their applications in 2-D spectral estimation. In N. K. Bose and C. R. Rao, editors, *Signal Processing and Its Applications*, Handbook of Statistics, Vol. 10, pages 207–227. North Holland, 1993.

[27] J. M. Francos, A. Zvi Meiri, and B. Porat. A unified texture model based on a 2-D Wold-like decomposition. *IEEE T. Signal Processing*, pages 2665–2678, August 1993.

[28] J. M. Francos, A.Z. Meiri, and B. Porat. A Wold-like decomposition of two-dimensional discrete homogeneous random fields. *The Annals of Applied Probability*, 5(1):248–260, Feb. 1995.

[29] J. M. Francos, A. Narasimhan, and J. W. Woods. Maximum likelihood parameter estimation of textures using a Wold-decomposition based model. *IEEE T. Image Processing*, pages 1655–1666, December 1995.

[30] J. M. Francos, A. Narasimhan, and J. W. Woods. Maximum likelihood parameter estimation of discrete homogeneous random fields with mixed spectral distributions. *IEEE T. Signal Processing*, 44(5):1242–1255, May 1996.

[31] J. M. Francos, B. Porat, and A. Zvi Meiri. Orthogonal decompositions of 2-D nonhomogeneous discrete random fields. *Mathematics of Control, Signals, and Systems*, 8(4):375–389, December 1995.

[32] K. S. Fu and J. K. Mui. A survey on image segmentation. *Pattern Recognition*, 13:3–16, 1981.

[33] M. Galloway. Texture analysis using gray level run lengths. *Comput. Graphics Image Processing*, 4:172–199, 1974.

[34] L. Garand and J. A. Weinman. A structural-stochastic model for the analysis and synthesis of cloud images. *J. of Climate and Appl. Meteorology*, 25:1052–1068, 1986.

[35] J. Gårding. Surface orientation and curvature from differential texture distortion. In *Proc. Int. Conf. on Computer Vision*, pages 733–739, Boston, MA, June 1995.

[36] L. V. Gool, P. Dewaaele, and A. Oosterlinck. Survey: texture analysis anno 1983. *Computer Vision, Graph., and Image Proc.*, 29:336–357, 1985.

[37] R. Haralick. Statistical and structural approaches to texture. *Proceedings of IEEE*, 67:786–804, May 1979.

[38] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE T. Sys., Man, and Cyber.*, SMC-3(6):610–621, 1973.

[39] R. M. Haralick and L. G. Shapiro. Image segmentation techniques. *Computer Vision, Graph., and Image Proc.*, 29:100–132, 1985.

[40] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*, volume 1. Addison-Wesley Publishing Company, Inc., 1992.

[41] L.O. Harvey, Jr. and M.J. Gervais. Visual texture perception and Fourier analysis. *Perception & Psychophysics*, 24(6), 1987.

[42] H. Helson and D. Lowdenslager. Prediction theory and Fourier series in several variables. *Acta Math*, 99:165–202, 1958.

[43] H. Helson and D. Lowdenslager. Prediction theory and Fourier series in several variables. II. *Acta Math*, 106:175–213, 1962.

[44] D.D. Hoffman and B.E. Flinchbuagh. The interpretation of biological motion. *Biological Cybernatics.*, pages 195–204, 1982.

[45] S.L. Marple Jr. *Digital Spectral Analysis.* Prentice-Hall, Inc., 1987.

[46] B. Julesz. Visual pattern discrimination. *Proceedings of IRE*, 8(2):84–92, 1962.

[47] B. Julesz. Textons, the elements of texture perception and their interactions. *Nature*, 290:91–97, 1981.

[48] B. Julesz and J. R. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *The Bell System Technical Journal*, 62(6):1619–1645, July-August 1983.

[49] H. Kaizer. A quantification of textures on aerial photographs. Technical Report No. 121, AD 69484, Boston Univ. Research Labs, Boston, MA, 1955.

[50] G. Kallianpur and V. Mandrekar. Nondeterministic random fields and Wold and Halmos decompositions for commuting isometries. In *Prediction Theory and Harmonic Analysis*, volume The Pesi Masani Volume, pages 165–190, Amsterdam, 1983. North-Holland.

[51] G. Kallianpur, A.G. Miamee, and H. Niemi. On the prediction theory of two-parameter stationary random fields. *J. of Multivariate Anal.*, 32:120–149, 1990.

[52] R. L. Kashyap. Analysis and synthesis of image patterns by spatial interaction models. In L.N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, pages 149–186. North-Holland, 1981.

[53] R. L. Kashyap and R. Chellappa. Estimation and choice of neighbors in spatial-interaction models of images. *IEEE T. Information Theory*, IT-29(1):60–72, 1983.

[54] J.M. Keller and S. Chen. Texture description and segmentation through fractal geometry. *Computer Vision, Graph., and Image Proc.*, 45:150–166, 1989.

[55] M. Kendall and J.D. Gibbons. *Rank Correlation Methods.* Oxford Univ. Press, fifth edition, 1990.

[56] A. Khotanzad and J.Y. Chen. Unsupervised segmentation of textured images by edge detection in multidimensional features. *IEEE T. Pat. Analy. and Machine Intel.*, 11(4):414–421, 1989.

[57] H. Korezlioglu and P. Loubaton. Spectral factorization of wide sense stationary processes on $\mathcal{Z}^2$. *J. of Multivariate Anal.*, 19:24–47, 1986.

[58] J. Krumm and S.A. Shafer. Texture segmentation and shape in the same image. In *Proc. Int. Conf. on Computer Vision*, pages 121–127, Boston, MA, June 1995.

[59] J. Lim. *Two-dimensional Signal and Image Processing.* Prentice Hall, 1990.

[60] F. Liu and R. W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE T. Pat. Analy. and Machine Intel.*, 18(7):722–733, July 1996.

[61] F. Liu and R. W. Picard. Finding periodicity in space and time. In *Proc. Int. Conf. on Computer Vision*, Bombay, India, January 1998. To appear.

[62] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. Image Understanding Workshop*, pages 121–131, 1981.

[63] J. Malik. A differential method for computing local shape-from-texture for planar and curved surfaces. In *Proc. Int. Conf. on Comp. Vis. and Pat. Rec.*, pages 267–273, New York, NY, June 1993.

[64] J. Malik and P. Perona. Preattentive texture discrimination with early visual mechanisms. *J. Opt. Soc. of Amer. A*, 7:923–932, 1990.

[65] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE T. Pat. Analy. and Machine Intel.*, 18(8):837–842, 1996.

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

[67] T.L. Marzetta. Two-dimensional linear prediction: autocorrelation arrays, minimum-phase prediction error filters, and reflection coefficient arrays. *IEEE T. Acous., Speech, and Signal Proc.*, ASSP-28(6):725–733, December 1980.

[68] T. Matsuyama, S-I Miura, and M. Nagao. Structural analysis of natural textures by Fourier transformation. *Computer Vision, Graph., and Image Proc.*, 24:347–362, 1983.

[69] W. Niblack *et al.*. The QBIC project: querying images by content using color, texture, and shape. In W. Niblack, editor, *Storage and Retrieval for Image and Video Databases*, pages 173–181, San Jose, CA, Feb. 1993. SPIE.

[70] S. A. Niyogi and E. H. Adelson. Analyzing gait with spatiotemporal surfaces. In *IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 64–69, Austin, Texas, Nov. 11-12 1994.

[71] E.B. Page. Ordered hypothesis for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, 58:216–230, 1963.

[72] A. Pentland, R. Picard, and S. Sclaroff. Photobook: tools for content-based manipulation of image databases. *Int. Journal of Computer Vision*, 18(3):233–254, June 1996.

[73] A. P. Pentland. Fractal-based description of natural scenes. *IEEE T. Pat. Analy. and Machine Intel.*, PAMI-6(6):661–674, 1984.

[74] R. W. Picard. Structured patterns from random fields. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, pages 1011–1015, Pacific Grove, CA, Oct 1992.

[75] R. W. Picard and T. Kabir. Finding similar patterns in large image databases. In *Proc. Int. Conf. on Acous., Speech, and Signal Proc.*, pages V–161–V–164, Minneapolis, MN, 1993.

[76] R. W. Picard, T. Kabir, and F. Liu. Real-time recognition with the entire Brodatz texture database. In *Proc. Int. Conf. on Comp. Vis. and Pat. Rec.*, pages 638–639, New York, June 1993.

[77] R. Polana and R. Nelson. Low level recognition of human motion. In *Proc. IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 77–82, Austin, TX, Nov. 11-12 1994.

[78] R. Polana and R. C. Nelson. Detecting activities. In *Proc. Int. Conf. on Comp. Vis. and Pat. Rec.*, pages 2–7, New York, NY, June 1993.

[79] A. R. Rao and G. L. Lohse. Towards a texture naming system: identifying relevant dimensions of texture. *Vision Research*, 36(11):1649–1669, 1996.

[80] C. H. Richardson and R. W. Schafer. The symbolic manipulation and analysis of morphological algorithms. In Alan V. Oppenheim and S. Hamid Nawab, editors, *Symbolic and knowledge-based signal processing*, pages 142–172. Prentice Hall, 1992.

[81] A. Rosenfeld and L. S. Davis. Image segmentation and image models. *Proceedings of IEEE*, 67(5):764–772, 1979.

[82] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1987.

[83] W.R. Schucany and W.H. Frawley. A rank test for two group concordance. *Psychometrika*, 38(2):249–258, 1973.

[84] S. Siegel. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, 1956.

[85] R. Sriram, J. M. Francos, and W. A. Pearlman. Texture coding using a Wold decomposition model. In *Proc. Int. Conf. on Pattern Recognition*, volume III, pages 35–39, Jerusalem, Israel, Oct. 1994.

[86] A. Sutter, G. Sperling, and C. Chubb. Measuring the spatial frequency selectivity of second-order texture mechanisms. *Vision Research*, 35(7):915–924, 1995.

[87] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE T. Sys., Man, and Cyber.*, SMC-8(6):460–473, 1978.

[88] M. Tuceryan and A. K. Jain. Texture analysis. In C. H. Chen, L. F. Pau, and P. S. P. Wang, editors, *The Handbook of Pattern Recognition and Computer Vision*, pages 235–276. World Scientific Pub. Co, 1993.

[89] F. M. Vilnrotter, R. Nevatia, and K. E. Price. Structural analysis of natural textures. *IEEE T. Pat. Analy. and Machine Intel.*, PAMI-8(1):76–89, Jan. 1986.

[90] P. Volet and M. Kunt. Synthesis of natural structured textures. In I. T. Young *et al.*, editor, *Signal Processing III: Theories and Applications*, pages 913–916. North-Holland, 1986.

[91] H. Voorhees and T. Poggio. Computing texture boundaries from images. *Nature*, 333:364–367, 1988.

[92] J.Y.A. Wang. *Layered Image Representation: Identification of Coherent Components in Image Sequences.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, September 1996.

[93] H. Wechsler. Texture analysis – a survey. *Signal Processing*, 2(3):271–282, 1980.

[94] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

[95] C.P. Winder. *Markovian Analysis of Texture: Serial and Parallel Paradigms in Low-level Vision*. PhD thesis, Oxford Univ. Computing Laboratory, St. Hugh's College, Oxford Univ., UK, 1992.

[96] H. Wold. *A Study in the Analysis of Stationary Time Series*. Stockholm, Almqvist & Wiksell, 1954.

[97] J. W. Woods. Two-dimensional discrete Markovian fields. *IEEE T. Information Theory*, IT-18(2):232–240, 1972.