Early-Visual Features as Determinants of Perceived Texture Similarity

by

William T. Neveitt

Submitted to the Department of Electrical Engineering and Computer Science

January 22, 1996

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Electrical Science and Engineering and
Master of Engineering in Electrical Engineering and Computer Science.

Abstract

This thesis is about the perception of texture similarity and the procedures underlying its computation. A model of texture similarity was engineered using a collection of simulated early-visual operations. These operations include oriented, multi-scale linear filtering; rectifying non-linearity; divisive normalization; low-order stochastic modeling; recursive cascading; and whitening. The first aim of this thesis was to illuminate the gap between these low-level operations and high-level judgements of perceived texture similarity in humans. The second aim was to assess the merits and limitations of early-visual operations as a practical basis for image retrieval.

A taxonomy of early-visual models were used to retrieve natural textures from the Brodatz database. The fraction of correct textures found versus the total textures retrieved was used as an objective performance criterion. The most accurate early-visual model achieved a recall accuracy on par with the best known SAR [29] model, dramatically outperforming benchmark models by Haralick [15], Tamura [43], and Kabir [23]. In addition, early-visual texture models exhibited very favorable cost/performance characteristics: one model attained a recall accuracy of 80% using only 8 floating-point numbers to represent each texture in the database.

Previous work has shown that humans categorize natural textures along three orthogonal dimensions of periodicity, directionality, and complexity [39]. A pilot psychophysical experiment was performed to test how well early-visual features could predict human judgements of texture similarity within these characteristic dimensions. Three representative sets of periodic, directional, and complex textures were selected from the Brodatz database and used as stimuli for triadic ordinal comparisons. A computational model based on oriented, multi-scale features successfully predicted 88% of the judgements in which all subjects produced the same response.

The principal result of this thesis is an emprical demonstration that early-visual features offer a promising language for representing natural textures. Building from the lessons learned here, a milestone for future work is suggested. Because the models in this thesis use simple, global image statistics to describe a texture, they often lack the ability to discriminate spatial organization that is readily apparent to humans. In particular, the automatic classification of textures into periodic, directional, and complex groups is highlighted as an important future test for early-visual models. Two possible avenues towards this goal include exploring more sophisticated subband modeling techniques, and increasing the depth of the cascaded features.

Thesis Supervisor: Edward H. Adelson
Title: Professor of Brain and Cognitive Sciences

List of Figures

Chapter 1

Introduction

Texture decorates every corner of our visual experience, from the bark of a birch tree to a sparrow's feather, from a camel's hair overcoat to a dense wall of grafitti. Like most of our perceptions, texture is easy to experience but extraordinarily hard to define. Webster's *College Dictionary* offers us the following:

> **tex·ture** (teks'cher), n. **1.** The characteristic physical structure given to a material, an object, etc. by the size, shape, and arrangement of its parts: *soil of a sandy texture.* **2.** the characteristic structure of the threads, fibers, etc. that make up a textile fabric: *coarse texture.* **3.** essential or characteristic quality; essence. **4.** the visual and tactile quality of a work of art resulting from the way in which the materials are used.

Dozens of vision researchers have also tried their hands at defining texture [47], yet like Webster's attempt, these definitions seem either too broad to be useful, or too narrow to build upon. Although a satisfying definition of vision texture appears beyond our reach, we clearly recognize texture when we see it.

Understanding our perception of texture is central to visual science, graphic arts, and the engineering of artificial visual systems. As an instrument of theory, texture offers us a glimpse at the alphabet of early-vision [2, 3]. Through computer graphics, texture turns artificial stills into rich landscapes [16]. As an engineering tool, texture supports the retrieval of images from large archives [32] the coding of regions for

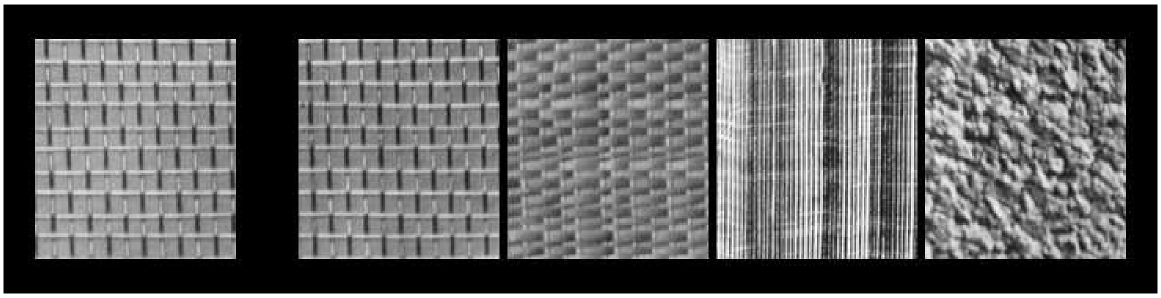

Figure 1.1: An example of a good texture ordering

digital video [27], and the estimation of surface properties from natural scenes [28]. These are just a few of the reasons why texture has been actively researched for over 30 years.

This thesis is about the perception of texture similarity and the procedures underlying its computation. Formally, this thesis casts the problem of texture similarity in the following terms: given a *database* of texture images and a single *key* texture selected from this database, sort the database with respect to the key in order of decreasing similarity. To be successful, a computational model of texture similarity ought to produce orderings that are well-aligned with the responses of human subjects. For example, consider the key and small database shown in Figure 1.1. The key texture on the left is woven aluminum wire. The four database textures on the right are, left to right, woven aluminum wire, straw matting, cheesecloth, and pressed cork. The ordering in Figure 1.1 is a good ordering of this database because, for example, most subjects agree that the two samples of woven aluminum wire look very similar, and therefore ought to appear next to each other in the ordering. By contrast, the ordering in Figure 1.2 is a poor ordering, in part because most subjects feel that woven aluminum wire looks more similar to straw matting than to pressed cork. Although searching for the unique, 'correct' ordering of a given key and database is probably nonsensical, some orderings will clearly be better than others.

Representation is the cornerstone of this thesis. Decades of visual science have yielded a principled view of the earliest visual operations in primates. This thesis engineers a computational model of texture similarity grounded in these operations, and evaluates its performance on a benchmark image retrieval task. Using a classic approach to the modeling of perceptual similarity, the resemblance of textures is computed using a geometric metaphor. Every texture image in the database is rep-

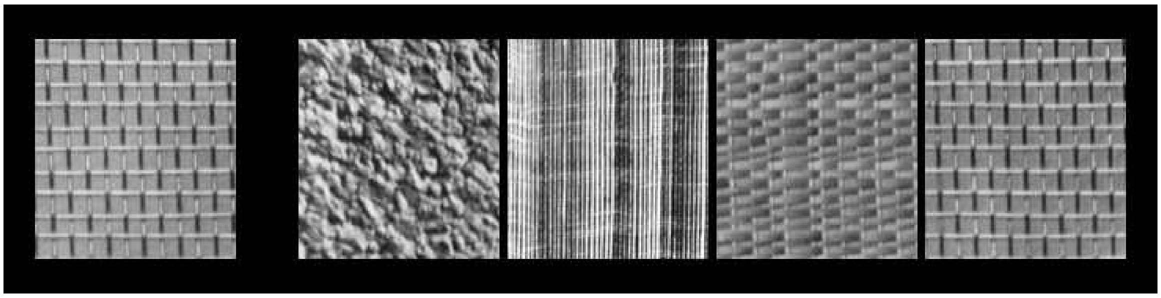

Figure 1.2: An example of a poor texture ordering

resented as a single point in a low-dimensional 'texture space' by processing it with a select set of early-visual operations. These operations include oriented, multi-scale linear filtering; a rectifying non-linearity; divisive normalization; recursive cascading; stochastic fitting; and whitening. An ordering of the database with respect to the key is produced by computing the distance in texture space between the key and every database texture. Model orderings are evaluated by comparison against objective benchmarks and the judgements of human subjects.

The thrust of this effort is twofold. By representing textures using early-visual features, this thesis aims to illuminate the gap between biologically motivated low-level image operations, and high-level judgements of perceived textural similarity. By working with an established database of natural textures, this thesis aims to contribute a practical metric for content-based image retrieval.

The remainder of this thesis is organized into four chapters. After briefly surveying our understanding of early-visual operations in Chapter 2, both a modeling strategy for these operations and a framework for the computation of textural similarity are specified in Chapter 3. Chapter 4 details experiments using this framework to both retrieve textures from the Brodatz texture database, and to predict the responses of human subjects. Finally, a summary of the principal results and directions for future research are outlined in Chapter 5.

Chapter 2

Texture and Early-Vision

Texture has been an important classical instrument for probing the earliest visual operations. This link between texture and early vision was forged over 30 years ago with the seminal work of Julesz [21] and Beck [1] on the effortless segregation of micro-patterns. Since then, a number of other experimental approaches, including both cognitive neuroscience and the psychology of visual search, have offered us glimpses at the early-visual machinery. This research reveals a remarkably direct relationship between human ability to rapidly discriminate textures and the receptive profiles of cells found in the striate cortex, inspiring the view that texture perception is grounded in simple processes. In this section, our current understanding of several early-visual operations and the supporting evidence is briefly reviewed.

2.1 Orientation and Scale

The existence of cortical cells tuned to particular orientations and spatial scales is widely accepted. In groundbreaking work, anatomical evidence for these cells in monkey cortex was established by stimulating the retina with spots or patterns of light and charting the responses of single cells [20]. The results suggest that cells in the primary visual cortex appear in at least three varieties: simple, complex, and hypercomplex. Simple cells are sensitive to lines and step edges, and show sharply localized 'on' and 'off' regions in their receptive fields. Complex-cells are also sensitive to abrupt changes in contrast, but have responses that are less sensitive to the absolute position of the stimulus. Hypercomplex cells are sensitive to two-dimensional image features, exhibiting flanked receptive fields with responses that first increase then decrease as a stimulus is lengthed or widened. Later anatomical investigations refined the spatial frequency selectivity and topology of these cells in the monkey cortex [9, 10].

In addition to neuroscientific data, the primacy of edges and blobs as features in early vision is supported by pre-attentive judgements. For example, the time it takes for a subject to search for a target defined purely by orientation is independent of the number of distractors [46]. In addition, orientation and size strongly determine the perceptual grouping of several classical micropatterns [1, 3]. One focus of this thesis will be to determine the stability of oriented, multi-scale features as anchors for the computation of textural similarity.

2.2 Non-Linearity

The responses of cortical cells are fundamentally non-linear. Because striate cells have a low maintained discharge rate, they cannot behave like truly linear operators, since positive and negative responses must be encoded in separate signals rather than through a single signal with a fixed bias. One popular approach views the simple-cell response as the output of a purely linear stage followed by a rectifying non-linearity, and views the complex-cell response as the pooled sum of simple-cell outputs [17].

The underlying linearity of simple-cells and the presence of a half-squared non-linearity have both been supported empirically. The linearity of simple-cells has been supported indirectly by comparing the discrepancies between simple-cell impulse and grating responses [40]. Assuming the existence of this underlying linear stage, modeling the non-linearity of simple-cells by half-wave rectifying and squaring the linear stage output has been supported by measuring the responses of simple-cells to compound [36] and counterphase [45] gratings.

Complex cells can be thought of as operators that pool the responses of simple-cells into an aggregate energy measure. Energy is computed by summing the half-squared outputs of four simple-cells whose underlying linear receptive fields are phase shifted by multiples of 90°. The energy model successfully accounts for a number of empirical observations, including the constant response of complex cells to drifting gratings [30].

Experimenters have also found pairs of cells with relative phase responses shifted by $90°$ [37]. In addition, the energy model successfully predicts the discriminability of certain micro-patterns used in classical texture segregation experiments [4]. One focus of this thesis will be to measure the effects of noise on magnitude and energy based non-linearities.

2.3 Normalization

There are two serious problems with the linear/energy model of striate cells presented so far. First, striate cell outputs saturate at high contrasts; ideal linear/energy cells may have arbitrarily large response amplitudes. Second, it has been observed that the responses of striate cells can be inhibited by the non-specific introduction of an additional stimulus; the linear/energy model offers no explicit mechanism accounting for the interactions between cell responses. These difficulties can be surmounted by augmenting the linear/energy model with simple feedback gain control [18]. In this approach, the energy response of each cell in a small population is normalized by dividing its response by the total pooled energy of all cells in the population. This model reflects both the saturation of striate cells at high contrasts, and non-specific supression phenomena [5].

The normalized gain control model need not discard overall stimulus contrast. It simply separates this information from measures of how contrast is distributed across orientations. Put differently, the gain-control model recodes complex-cell outputs without loss of information. One focus of this thesis will be to understand how this recoding affects the computation of texture similarity.

2.4 Two-Dimensional Image Features

Hypercomplex cells in the striate cortex have received far less attention than simple and complex cells. These cells appear to be sensitive to image terminations, with responses that first increase, then decrease as the length or width of a bar is steadily increased.

Unlike simple or complex cells, which are essentially sensitive to purely one dimensional variations, it is believed that hypercomplex cells encode two dimensional features of the stimulus. One popular view holds that hypercomplex cells are the neural basis for computing curvature [11], and in particular, are sensitive to corners and junctions. Such features are informative cues for recovering the ordinal relationships between occluding objects [13], and spatio-temporal variants of these cells may underly the perception of kinetic occlusion [31]. Visual search studies also support for the recovery of terminations and junctions at the pre-attentive level [46]. One focus of this thesis will be to recover two dimensional features by cascading one-dimensional operators, and to understand how much new information these hypercomplex features add to the computation of textural similarity.

2.5 Stochastic Description

The earliest models of texture perception were built around the intuition that textures are best described statistically. The seminal work of Beck and Julesz on micropattern segregation led to the Julesz conjecture: humans are only sensitive to local, second order distributions of the image intensity. This conjecture was refuted by counterexamples that showed subjects could both segregate textures with the same local second-order distributions, and could not segregate textures with different second-order distributions [22]. Although Julesz's conjecture and its variants have not survived empirical scrutiny, his view that low-order statistics of the image intensity are strong determinants of texture perception continues to guide the field. Unfortunately, work on the statistical character of texture perception in humans has largely been limited to segregation phenomena. One focus of this thesis will explore compact statistical features as a basis for representing and retrieving vision texture.

2.6 Whitening

The response profiles of striate cells often overlap significantly, leading to highly correlated output responses. This has lead researchers to postulate mechanisms by which the cortex might remove redundant information in low-level features. In a provocative paper [26], Linsker has evolved a locally connected, unsupervised neural network whose functional behavior matches that of a standard statistical technique known as whitening. Whitening, or principal components analysis (PCA), is a method for decorrelating a set of points in a vector space using a single rotation and scaling operation. Interestingly, Linsker was able to show that the receptive field of the input layer to this network closely approximates the classical center-surround receptive fields of cortical cells. One focus of this thesis will be to use whitening to gauge how many independent parameters are needed to represent a texture at a particular level of retrieval performance.

2.7 Summary

Six early-visual operations and the challenges they present to this thesis have been reviewed in this section. Orientation and scale are basic primitives in the alphabet of early vision. But are orientation and scale stable anchors for computing the similarity of texture images? The essential non-linearity of cellular operators is successfully explained using an underlying linear model followed by half-squaring rectification. Yet how do these non-linearities respond in the presence of common noise sources? Normalization accounts for the saturation of cell responses at high contrasts and the effects of non-specific inhibition. How does this transformation affect the performance of classic similarity metrics? Two-dimensional image features like corners and terminations are coded by the responses of hypercomplex cells. Yet how much original information do hypercomplex features add to the representation of a texture? Low-order stochastic distributions are thought to characterize textures. Although compact, are these statistical features visually complete enough to be useful? Finally, whitening feature outputs can recreate the classical center-surround receptive profiles of cortical cells. How many uncorrelated numbers does it take to represent a database of textures at a given level of retrieval performance? Building from these basic operations and the issues they raise, a computational model for textural representation, comparison, and decision is described in the following chapter.

Chapter 3

Computational Modeling

The task of a texture similarity model is to order a texture database with respect to a key by decreasing similarity. A successful model will produce orderings well-aligned with the responses of human subjects. The principal hypothesis of this thesis is that a similarity model grounded in early-visual features will produce such orderings.

A rigorous framework for computing texture similarity must specify at least three things. First, it must describe how a texture is represented and specify algorithms that recover this representation from images. Second, it must describe how textures are compared once they have been represented, and state the assumptions that underly these comparisons. Finally, it must specify how textural judgements are produced from comparisons. This section specifies the representation, comparison metric, and decision model used in this thesis.

3.1 Representation

The representation used in this thesis is built squarely on the visual science reviewed in the previous chapter. The core analysis engine includes six basic operations: oriented, multi-scale linear-filtering; rectifying non-linearity; divisive normalization; cascading; stochastic fitting; and whitening. Selectively composing these operations produces a taxonomy of models with characteristic properties. Careful analysis of this taxonomy will be used to highlight the contributions of each operation to the computation of texture similarity.

3.1.1 Linear Filtering

As discussed in the previous chapter, the sensitivity of simple cortical cells to spatial orientations and scales is widely accepted and relatively well-understood. The underlying linearity of these simple cells is commonly modeled using a wavelet filter bank. This filter bank decomposes an image into a set of sub-band components by first applying a low-pass kernel and a set of oriented filters to the image stimulus, then downsampling and recursively processing the low-pass image.

As a model of early-visual processing, the attractive features of the wavelet architecture are its linearity, perceptual selectivity, and completeness. The linearity of the filter bank makes it a suitable model for the underlying stage of cortical cells. In addition, the selectivity of the filters to orientation and scale ensures that the sub-bands will represent key image features directly. For example, the recursive filtering and downsampling pyramid accurately models the observation that contrast sensitivity in the human visual system varies logarithmically with radial spatial frequency. Finally, the completeness of the pyramid representation implies that the original stimulus can be reconstructed from its sub-bands, a generic property thought to underly visual representations. In image processing, this recursive approach to image representation has become a powerful tool for image compression [41] and enhancement [42].

Several filter sets have been designed as models of early-visual processing [24, 6]. Key design criteria for these filter sets include negotiating the fundamental trade-off between spatial and spatial-frequency localization, efficient tiling of the spatial frequency plane, and closely approximating the detailed receptive profiles of cortical cells. A basic conclusion of research in this area is that there is no single 'optimal' filter set; the choice of filters is influenced to a large extent by the intended application.

One difficulty of most filter sets is that in order to effectively tile the spatial frequency plane, a large number of filters must be used. A clever solution to this dilemma generates a small set of basis filters from which a much larger set of filters can be synthesized on-demand [13]. This strategy is similar in spirit to pose-invariant object recognition by interpolation of view-based templates. Because of their computational efficiency, attractive localization properties, and self-invertibility, these 'steerable filters' are used as the initial linear filtering stage throughout this thesis.

Following the implementation described in [42], an invertible steerable pyramid is shown in Figure 3.1. This pyramid filters in scale and orientation using a polar-separable kernel design. In radial frequency, the image stimulus is split into high-pass, band-pass, and low-pass bands. In angular frequency, the radial band-pass image forms an anulus which is split into sub-bands using steerable filters. The low-pass band is subsampled by a fact of 2 in each dimension, and used as the input for a new pyramid level.

Steerability allows an arbitrarily oriented filter to be synthesized from a linear combination of basis filters. In this thesis, a texture stimulus is analyzed at each pyramid level by filters oriented at $0°$, $45°$, $90°$, and $135°$. Selecting horizontal and vertical orientations for analysis seems natural since the human visual system appears to be more sensitive to these directions [41]. Choosing $45°$ and $135°$ orientations allows circles and lines to be distinguished, since the mean projections of a circle and a line at $45°$ will look the same when projected solely onto horizontal and vertical orientations. In addition, using just these four projections of the stimulus keeps the representation compact, and has already demonstrated its utility by successfully accounting for a variety of classical texture segregation phenomena [4].

3.1.2 Energy Mechanism

There is strong support for the existence of an essential non-linearity following the underlying linear stage of simple and complex cortical cells. This thesis implements this non-linearity using a quadrature energy computation. Second-order Taylor approximations to the Hilbert Transforms of the even-steerable filters were designed in [12]. Like their even symmetric counterparts, these odd-symmetric filters are also steerable and self-inverting. Energy is computed by convolving the image stimulus $I[x,y]$ with both the even $E[x,y]$ and odd $O[x,y]$ filter sets, and invertibly transforming the outputs according to:

$$E[x,y] = \sqrt{(I[x,y] * E[x,y])^2 + (I[x,y] * O[x,y])^2} \tag{3.1}$$

$$\Phi[x,y] = atan\frac{I[x,y] * O[x,y]}{I[x,y] * E[x,y]} \tag{3.2}$$

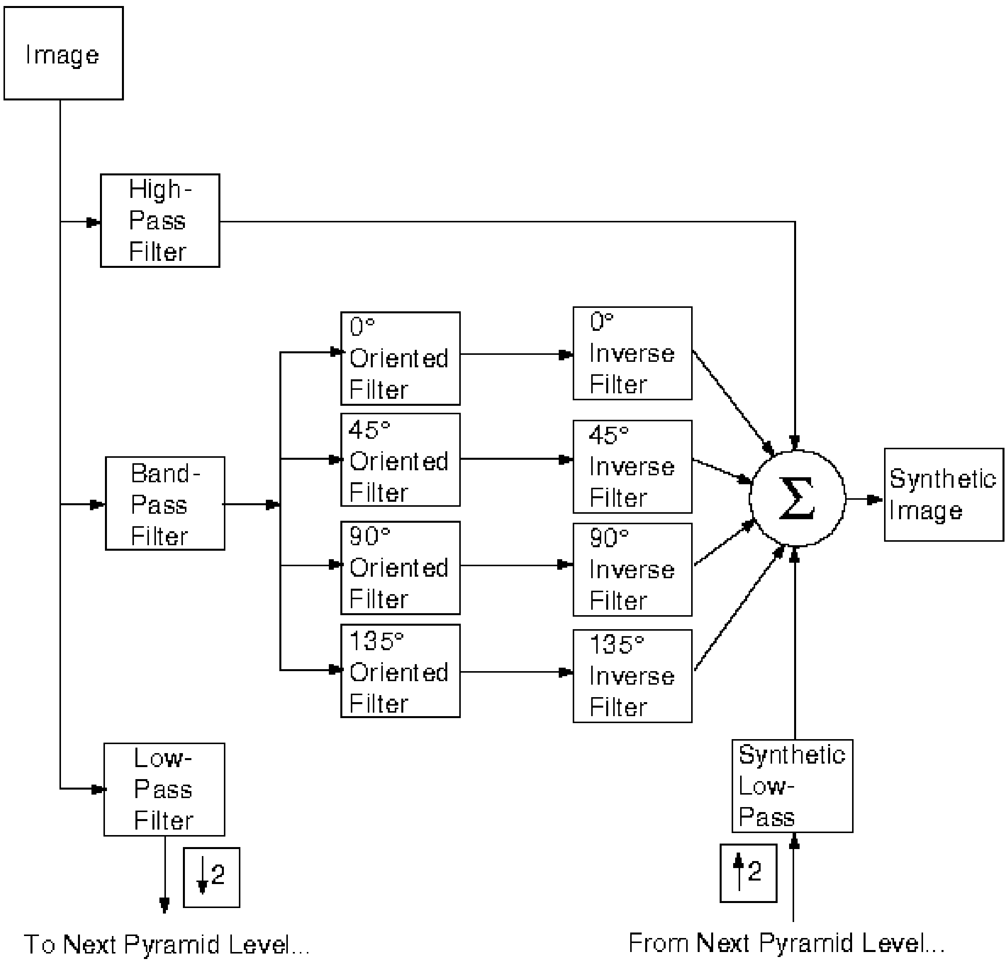

Figure 3.1: Block diagram of a self-inverting steerable pyramid

Consistent with the model of complex-cells presented in the previous chapter, the phase image Φ is ignored by later processing.

3.1.3 Normalization

Normalized gain-control successfully explains the saturation of cell outputs at high contrasts, and the effects of non-specific supression. Following the detailed analysis of [18], we implement normalization as a pooled, divisive computation. For each of the four oriented energy images ($O_{0°}[x,y]$, $O_{45°}[x,y]$, $O_{90°}[x,y]$, and $O_{135°}[x,y]$) at a particular scale in the steerable pyramid , every pixel is divided by the pooled energy at that pixel location in each of the four orientations, producing a normalized image $N_\theta[x,y]$, as shown in Figure 3.2.

$$N_\theta[x,y] = \frac{O_\theta[x,y]}{\sum_{\theta=0°}^{135°} O_\theta[x,y]} \tag{3.3}$$

Pixel locations with zero total energy in all four orientations are set to zero intensity in the normalized map.

The total summed energy at a particular pixel location, independent of orientation, is coded in a cumulative energy map:

$$C[x,y] = \sum_{\theta=0°}^{135°} O_\theta[x,y] \tag{3.4}$$

This ensures that normalizing codes the sub-band energy maps without loss of information, since the original energy profiles at all four orientations can be recovered from the normalized profiles of three orientations and the cumulative energy map:

$$O_\theta[x,y] = N_\theta[x,y] \cdot C[x,y], \quad \theta = 0°, 45°, 90° \tag{3.5}$$

$$O_{135°}[x,y] = C[x,y] - O_{0°}[x,y] - O_{45°}[x,y] - O_{90°}[x,y] \tag{3.6}$$

A similar scheme has been proposed as a perceptually-based distortion measure for image coding [44].

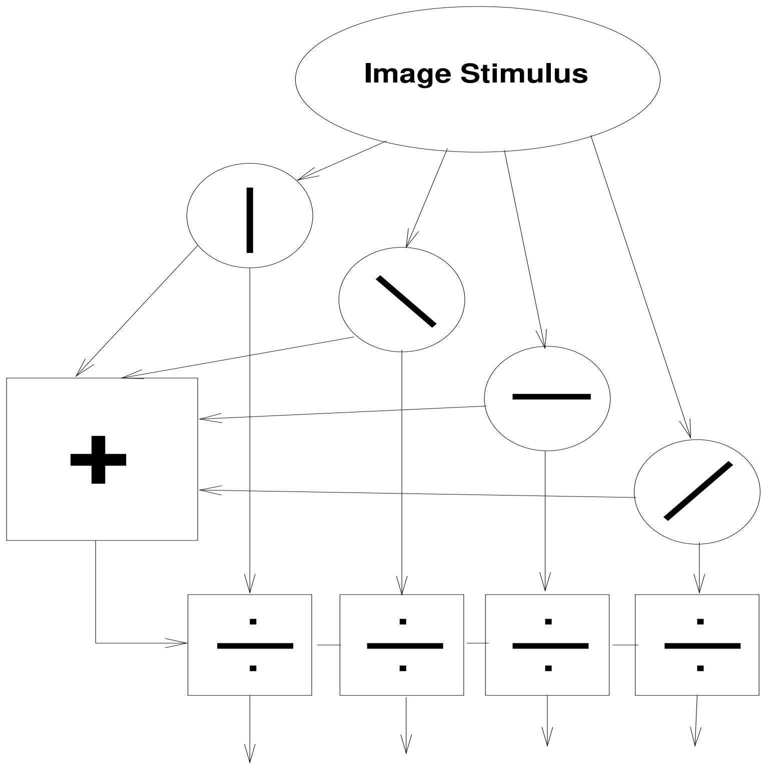

Figure 3.2: Block diagram of the normalization computation

3.1.4 Stochastic Description

Textures have been classically described using low-order statistics. This thesis uses a simple first-moment operator F to collapse sub-band energy images, $E[x,y]$, into real-valued features:

$$F < E[x,y] >= \frac{1}{x_{max} \cdot y_{max}} \sum_{x=0}^{x_max-1} \sum_{y=0}^{y_{max}-1} E[x,y] \tag{3.7}$$

Other statistical modeling techniques model the spatial variations in a texture by introducing relative spatial position as an explicit parameter in their models [29, 8, 33]. Using position as an explicit parameter, however, can dramatically increase the dimensionality of the model [38]. The novel approach of this thesis trades expressiveness for compactness by coding space implicitly using cascades of simple features. Whether or not such an approach can adequately capture the salient spatial structure we perceive in natural textures remains to be demonstrated.

3.1.5 Whitening

Whitening is a standard statistical technique used to decorrelate a vector of random variables. Given a set of N observation vectors of length L, arranged as the rows of a matrix V, view the columns of V as samples of a random process. Assume, without loss of generality, that the columns of V are zero mean. Whitening produces a new matrix V' whose columns are zero-mean, uncorrelated, and unit-variance:

$$K = \frac{1}{N} V'^T V' = I_{N \times N} \tag{3.8}$$

Here, $I_{N \times N}$ is the $N \times N$ identity matrix.

Whitening decorrelates the column space of V by the following procedure. First, the covariance matrix, K, of the matrix V is computed:

$$K = \frac{1}{N} V^T V \tag{3.9}$$

Next, using a familiar technique from linear algebra, compute two matrices E and Λ, where the columns of E are the orthonormal eigenvectors of K, and Λ is a diagonal matrix whose ith value is the associated eigenvalue of the ith column of E. E and Λ satisfy the relation:

$$KE = \Lambda E \tag{3.10}$$

Since K is a positive semi-definite matrix, the matrices E and Λ can always be found.

Next, an intermediate matrix T is produced according to:

$$T = VE \tag{3.11}$$

Because E is an orthonormal matrix, this operation corresponds to a simple rotation of the row space of V. Using the orthonormality of the columns of E, it is easily verified that the covariance of T is diagonal, and in particular, is equal to the matrix Λ:

$$T^T T = (VE)^T VE = E^T V^T VE = E^T \Lambda E = \Lambda \tag{3.12}$$

Note here that the ith element of Λ is the variance of the ith dimension of the matrix T.

Finally, to produce the unit-variance matrix V', the columns of T are simply scaled:

$$V' = \sqrt{\Lambda}^{-1} T \tag{3.13}$$

As demonstrated, whitening is equivalent to a single rotation and scaling operation on the columns of V. Because of the direct relationship between the eigenvalues of the matrix K and the variances along the dimensions of the intermediate matrix T, whitening can chart the distribution of uncorrelated information throughout the column space of V, allowing dimensions with little information content (i.e. very small variances) to be pruned away in a statistically optimal manner. For this reason, whitening has become a standard technique in data compression and recursive estimation.

3.1.6 Cascading

Unlike simple and complex cells, hypercomplex cells are sensitive to two-dimensional image features like corners and line terminations. This thesis models hypercomplex cells by cascading the operations discussed above in a tree structure, as shown in Figure 3.3. Each node of this tree is labeled with the operation it performs. After computing the oriented energy image at a particular orientation and scale, this energy image is used to seed a new steerable pyramid. Like the initial stimulus pyramid, successive sub-band images are expanded at all scales larger than the filter kernel. However, unlike the stimulus pyramid, cascaded pyramids are only analyzed at orientations that are equal or orthogonal to the orientations of their parents in the tree. This reflects the interpretation of cascaded bands as end-stopped and side-stopped operators that compute the mixed second-order derivatives of the stimulus energy.

Representing two-dimensional spatial variations using factored one-dimensional operators is, in general, not sufficient to express all variations in arbitrary images [25]. However, the fact that hypercomplex cells appear to share many of the same features of complex cells [19] lends some support to a recursively organized model. As a first approximation, this thesis adopts a cascaded model as a proxy for more sophisticated second-order computations.

3.1.7 Summary

To summarize, the representation process accepts a single image as input and produces a real-valued vector of features as output.

An initial linear filtering operator accepts an image stimulus as input and produces a set of even and odd filtered images at multiple scales as output.

An energy operation accepts an image and its Hilbert Transform as input, and produces a quadrature energy image as output.

A normalization operation converts a set of four oriented energy images at a particular scale into a set of three normalized energy images and a single cumulative energy image.

A simple first-moment operator accepts an image as input and produces a single floating-point number as output.

Cascading recursively applies the operations specified above in a manner that depends on the directional sensitivity of previous stages. For example, the vertical first-order energy band is cascaded with vertical and horizontal second-order operations, while the 45° first-order energy band is cascaded with 45° and 135° degree second-order operations. When performing a cascade on a map with no prior directionality (e.g. a cummulative energy map), both a 0°-90° cascade, and a 45°-135°

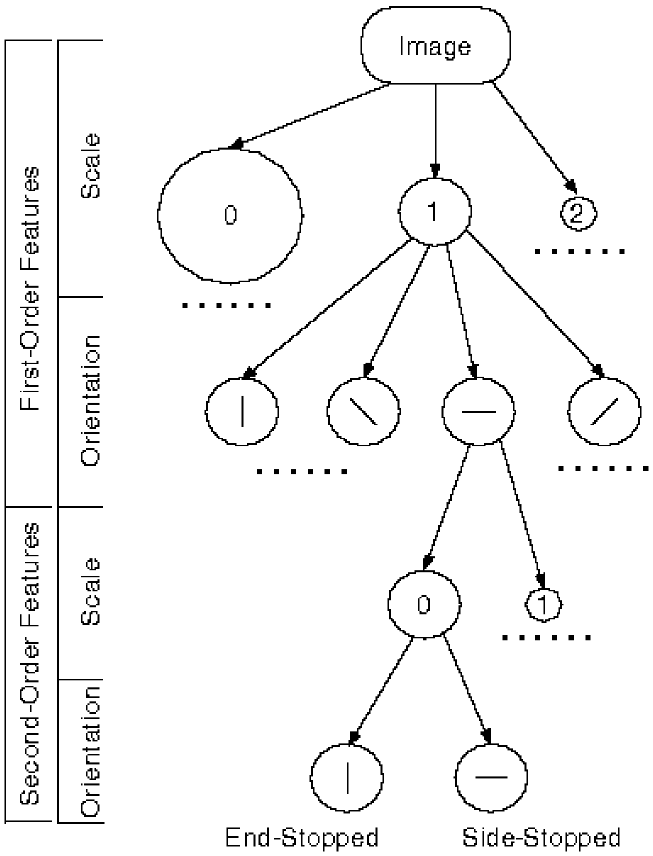

Figure 3.3: Block diagram of the cascading operation

Model Description	*Length*
Model 1: Orientation	4
Model 2: Normalized Orientation	4
Model 3: Multi-Scale Orientation	12
Model 4: Multi-Scale, Normalized Orientation	12
Model 5: End/Side Cascaded, Multi-Scale Orientation	60
Model 6: End/Side Cascaded, Normalized, Multi-Scale Orientation	72
Model 7: Model 6 with both Mean and Variance Based Sub-Band Features	144

Figure 3.4: Taxonomy of early-visual models.

cascade are applied.

Figure 3.4 presents a taxonomy of models based on different compositions of these basic operations. By comparing the performance of models in the taxonomy, this thesis will examine the impact of each operation on the computation of similarity.

3.2 Comparison

A texture's representation can be viewed geometrically as a single point in a multi-dimensional 'texture space'. Traditionally, a wide variety of psychological comparisons have been modeled using a function that computes the distances between pairs of points such a space. As a first approximation, this thesis adopts the geometric model of similarity, by using simple, euclidean distance to compute the dissimilarity between texture paris.

For a simple euclidean distance to be accurate, three assumptions concerning the comparison scheme used by humans must be true. These are the so called *metric assumptions*. Letting $\delta(a,b)$ represent a positive, real-valued measure of how dissimilar two textures a and b appear to a human subject, the metric assumptions may be stated as follows:

$$Minimality : \delta(a,a) = 0 \tag{3.14}$$

$$Symmetry : \delta(a, b) = \delta(b, a) \quad (3.15)$$

$$Triangle\ Inequality : \delta(a, c) > \delta(a, b) + \delta(b, c) \quad (3.16)$$

Although the metric distance model underlies the interpretation of a vast body of experimental psychology, persuasive arguments cast doubt on its validity. In a compelling paper [48], Tversky argues that human judgements of similarity often violate the metric assumptions, and proposes an alternative to metric similarity based on feature sets. Supporting the categorical nature of texture perception, recent work [39] has identified three salient dimensions along which humans perceive natural textures: periodicity, directionality, and randomness. A computational approach to texture similarity based on early classification along these dimensions is the subject of active research [34].

Although a categorical approach to texture similarity has much to recommend it, the link between these categories and the low-level features known to visual scientists has yet to be clarified. In particular, the hypothesis that these categories may be explainable through simple, early-visual processes has yet to be explored. As such, this thesis will focus primarily on the characterization of early-visual features as a substrate for texture perception, rather than on more sophisticated comparison procedures.

3.3 Decision

The decision stage of the similarity model accepts two real numbers representing the distances between two texture pairs, and produces a binary value which specifies which texture pair is most similar. An ordering of a texture database with respect to a key is produced by sorting the distances between all textures in the database and the key using this pair-wise decision operation.

One consequence of the similarity model as stated so far is that it is deterministic: given a database and key, the same ordering will always be produced. However, it seems wildly optimistic to believe that humans will behave this way when asked to render the same texture judgement more than once. It must be anticipated that, even in the same experiment, subjects will not produce consistent judgements due to factors beyond the understanding of contemporary cognitive science. Ideally, a complete model of texture similarity ought to account for the sources of noise in these judgements.

Unfortunately, a systematic approach to the study of noise in textural judgements has yet to be undertaken, although an interesting framework for modeling noise in the judgement process has been proposed in [14]. As a first approximation, this thesis will model the judgement process using a deterministic, noiseless comparator:

function: closer?(distance1 distance2) if (distance1 > distance2) T else F

Here, the result T indicates that the textures separated by distance1 are closer together than the textures separated by distance2.

3.4 Alternative Models

Four established benchmarks were used to rank the performance of the similarity model detailed in this chapter. Each of these alternatives was implemented by Tom Minka in the Photobook environment. It should be noted that although the implementations of these alternatives are representative of their use in the literature, strenuous efforts to optimize these models were not undertaken.

Haralick and his collaborators proposed one of the first computational feature sets for use in textural classification. Aligned with the then-popular view that texture perception was largely determined by local, second-order statistics of the image intensity, Haralick chose features that were extracted from grey-level co-occurance matrices. These features attained retrieval accuracies in excess of 80% on photomicrographs, aerial imagery, and satellite photos. A detailed specification of Haralick's methods can be found in [15]. The variant of Haralick's model used in this thesis represents each texture using 5 floating-point numbers: the energy, entropy, contrast, correlation, and inverse diference moment of the grey-level co-occurance matrix:

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i,j)^2 \tag{3.17}$$

$$f_2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j) log(P(i,j)) \tag{3.18}$$

$$f_3 = \sum_{n=0}^{N_g-1} n^2 \{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P(i,j)\}, \quad |i-j| = n \tag{3.19}$$

$$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{(ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{3.20}$$

$$f_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{P(i,j)}{1+(i-j)^2} \tag{3.21}$$

Here, $P(i,j)$ is the i,jth entry in the normalized co-occurance matrix. Following representation, a simple euclidean metric was used to order the Haralick texture space.

Tamura et.al. have studied a set of textural features related to human perception. Tamura manufactured operations to recover measures of directionality, coarseness, contrast, regularity, roughness, and 'line-likeness' from texture images, and used these features as predictors of human psychophysical performance. Tamura's coarseness, contrast, and directionality features were powerful predictors of human judgements on the rank orderings of 16 representative textures. However, simple linear combinations of Tamura's features were unsuccessfull at predicting nearest-neighbor similarity judgements. A thorough review of Tamura's work can be found in [43]. The version of Tamura's model used in this thesis represents each texture using 3 floating-point numbers: the coarseness, contrast, and directionality features of the texture.

Coarseness is computed using a three stage procedure that essentially averages the spatial scale at which neighboring image patches vary the most.

The contrast feature incorporates both raw grey-level range (i.e. the standard deviation of the image intensity σ), and the relative polarization of black and white image regions (i.e. the kuritosis, where μ_4 is the fourth-moment) according to:

$$\alpha = \frac{\mu_4}{\sigma^4} \tag{3.22}$$

$$F_{contrast} = \frac{\sigma}{\alpha^n} \tag{3.23}$$

Here n is a free-parameter, set to 1/4 empirically by Tamura.

The directionality feature is obtained by linearly filtering the texture with derivative filters, computing a histogram versus angle ϕ of all filtered intensities exceeding a threshold T, and estimating how strongly the orientation distribution is peaked using a simple statistic. This statistic essentially sums the second moment about every peak in the orientation histogram:

$$F_{dir} = 1 - rn_p \sum_{p}^{N_p} \sum_{\Phi} (\Phi - \Phi_p)^2 H_d(\Phi) \tag{3.24}$$

where Φ is the oriented angle modulo 180˙, n_p is the number of peaks in the orientation histogram, Φ_p is the position of the pth peak $H_d(\Phi)$, w_p is the width of the pth peak between valleys, and r is a normalizing coefficient.

For the model used in this thesis, Tamura's coarseness, contrast, and directionality features were whitened, and a euclidean metric was used to order the texture space.

Kabir and Picard have developed a texture classification model based on shift-invariant principal components [23]. In this scheme, a texture is represented in two steps. First, the DTFT magnitude of each texture image in the database is computed and used in all subsequent processing, eliminating sensitivity to linear phase. Next, 10% of the images in the database are selected at random and used to estimate a sample mean and covariance; this is done by converting each 128 x 128 texture image into a single, 128*128 long vector. Using the estimated sample covariance, a set of eigenvectors and eigenvalues were recovered. Euclidean distance was used to order the database. The distribution of perceptually relevant information in this space was probed by using comparing the accuracy of models using the largest 10, 20, 40, 60, 80, and 100 eigenvectors for recognition, and selecting the best performing model. This model represents each texture using 60 numbers per texture.

Mao and Jain have proposed a popular model of texture analysis based on a multi-scale, simultaneous auto-regressive model (SAR). SAR has demonstrated its value in both texture segmentation and classification [29]. SAR models texture variations using a regression across neighboring image positions:

$$g(s) = \mu + \sum_{r \in D} \theta(r) g(s+r) + \epsilon(s) \tag{3.25}$$

Where $g(s)$ is the grey-value intensity, μ is the mean image intensity, D is the mask neighborhood, the $\theta(r)$ are model parameters, and $\epsilon(s)$ is a zero-mean gaussian random variable with variance σ^2. The model parameters can be recovered using standard ML estimation. It should be noted that the SAR model given above can be made rotation invariant without neccessarily requiring an increase in the dimmensionality of the model. For the rotation-invariant model used in this thesis, a 4x4 image mask at 3 scales was used to produce 12 model parameters with an additional error-variance term at each scale, for a total of 15 numbers per 21x21 sub-window. Furthermore, a 15x15 covariance matrix of the model parameters was estimated using multiple sub-windows within a single image. The total number of parameters for the SAR model used in this thesis is therefore 15 mean and 225 covariance features. Gaussian divergence was used to order the texture space. Based on a longitudinal study of many popular texture models performed by Tom Minka, the implementation of the SAR model used in this thesis was found to be the best homogeneous retrieval model for the Brodatz database.

3.5 Summary

Texture similarity is computed using a three stage process of representation, comparison, and pair-wise decision to produce an ordering of a set of texture samples with respect to a key. The *representation* process accepts an image as input and produces a real-valued vector as output. The *comparison* process accepts two vectors as input and produces a real-valued measure of the distance between these vectors as output. The *decision* process accepts two distances as input and produces a binary ordinal judgement as output. This thesis focuses on the descriptive power of early-visual features for representation, while highlighting potential directions for extending the comparison and decision processes. As yardsticks to rank the performance of the model specified in this chapter, popular benchmarks due to Haralick, Tamura, Kabir, and Mao were used.

Chapter 4

Image Database Retrieval

The empirical problem studied in this thesis is the content-based retrieval of natural textures from a large image archive. Because image data constitutes the largest and fastest growing body of information in many on-line systems, devising accurate, scalable methods connecting users to the images they need is an important area of research. Within the content-based retrieval paradigm explored here, a user searches for a desired texture by selecting textures that look similar to it. If one has a computational model that successfully predicts the way humans judge texture similarity, this model could be used to navigate large databases, swiftly delivering the desired texture to the waiting user.

In this chapter, the Brodatz texture database is described, the retrieval task is formulated, and criteria for evaluation are established. The performance of the model taxonomy is examined and compared against established alternatives.

4.1 The Brodatz Database

Over 25 years ago, P. Brodatz composed an album of carefully photographed natural textures which he hoped would stimulate the imaginations of graphic artists, and advance our understanding of vision texture [7]. This album contains a rich variety of periodic, directional, complex, and inhomogeneous textures, and in recent years, has enjoyed popularity as a proving ground for computational texture research.

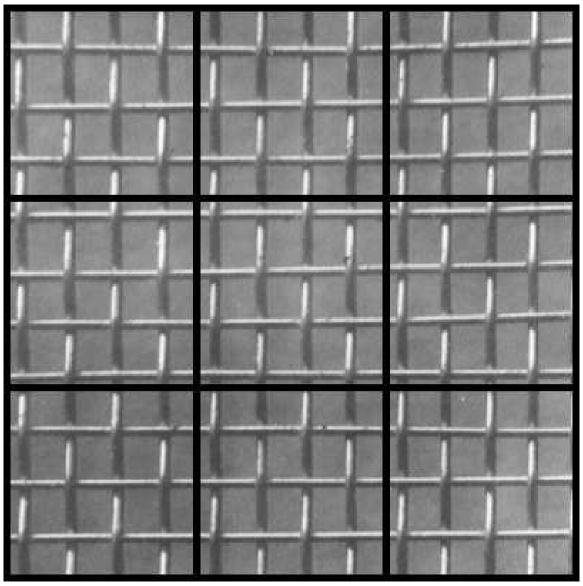

Figure 4.1: An example of a Brodatz texture class.

The complete digitized Brodatz database used in this thesis contains all 112 texture *classes* from the Brodatz album, 9 *samples* per class, for a total of 1008 images. An illustration of a typical Brodatz texture class and its samples is shown in Figure 4.1. All sample images are 128x128, 8-bit greyscale images, and are created by cropping non-overlapping regions from a large original [35]. In previous work, this database has been used for texture classification [35], segmentation, and coding [38].

The Photobook image database environment was used as an interface for the retrieval experiments [32]. In Photobook, users choose how similarity between textures will be computed by selecting a retrieval model. Retrieval begins by simply selecting a texture to be used as the key. Photobook orders all textures in the database with respect to the selected key in real-time using pre-processed representations, displaying the result in an application window.

4.2 The Task and Evaluation Strategy

The general retrieval task studied here may be formulated as follows:

Given:

- A texture database B. B is a set of images containing c texture classes, with s samples in each class. A class is a type of texture; samples are instances of a class.

- A single key texture T selected from B.

Produce:

- An ordering that maximizes both *objective* and *subjective* accuracy.

To be objectively accurate, the most similar textures to any key T ought to be all and only the members of T's class, since both the key and these members are instances of the same texture. To be subjectively accurate, the arrangement of texture classes appearing after the key's own class in the ordering ought to be well-aligned with human judgements of ordinal similarity. Objective and subjective performance of are examined in turn.

4.3 Objective Performance Criterion

Objective accuracy is quantified by plotting a receiver operating characteristic (ROC). The ROC is a graph of the average fraction of correct textures found versus the number of total textures retrieved. For an optimal retrieval model on the Brodatz database, the ROC would have the value 1.0 at 9 since for every texture T in the database, the first 9 textures in an ordering produced using T as a key would be T and the remaining 8 members of T's class. All Brodatz model ROC's have the value 1.0 at 1008, and are monotonically increasing, concave-down functions.

4.4 Analysis of Objective Performance

Each model in the early-visual taxonomy of Figure 4.2. is analyzed consecutively, from simplest to most complex.

Model Description	*Length*
Model 1: Orientation	4
Model 2: Normalized Orientation	4
Model 3: Multi-Scale Orientation	12
Model 4: Multi-Scale, Normalized Orientation	12
Model 5: End/Side Cascaded, Multi-Scale Orientation	60
Model 6: End/Side Cascaded, Normalized, Multi-Scale Orientation	72
Model 7: Model 6 with both Mean and Variance Based Features	144

Figure 4.2: Taxonomy of early-visual models.

4.4.1 Model 1: Simple Orientation

The ROC for model 1, based on 4 simple orientation features, is compared against the Haralick [15] and Tamura [43] models in Figure 4.3. The legend at the lower right of this graph labels the plots, from top to bottom in order of relative performance. Model 1 dramatically outperforms the Haralick model by 30% at 10, 28% at 50, and 13% at 100 textures retrieved. Model 1 also marginally outperforms the Tamura model by 5% at 10, 7% at 50, and 7% at 100 textures retrieved.

Retrieval rates averaged over all samples in a class were used to pinpoint the easiest and hardest classes for model 1 to represent. The easiest classes are shown in Figure 4.4. These textures are straw screening, netting, woven brass mesh, and french canvas. Not surprisingly, good classes are generally rigid, strongly oriented surfaces. Furthermore, examining the mean and variance of each oriented feature, averaged across all textures in the database, reveals a predominance of horizontal and vertical energy. This explains why a strongly oriented diagonal pattern like the woven brass mesh in Figure 4.4 performs so well; being comparatively rare, strongly diagonal textures are less likely to be confused with samples from other classes.

The poorest performing classes were grouped into five categories chosen to illustrate the sensitivity of model 1 to characteristic forms of intra-class noise. Left to right, these textures are a swinging light, tree bark, oriental glass fiber cloth, calfskin, and oriental straw cloth. Representative texture pairs from each of these categories are shown in Figure 4.5.

Slop texture classes are composed of samples that look so dramatically different, most humans would hesitate to call them textures, much less be able to relate them

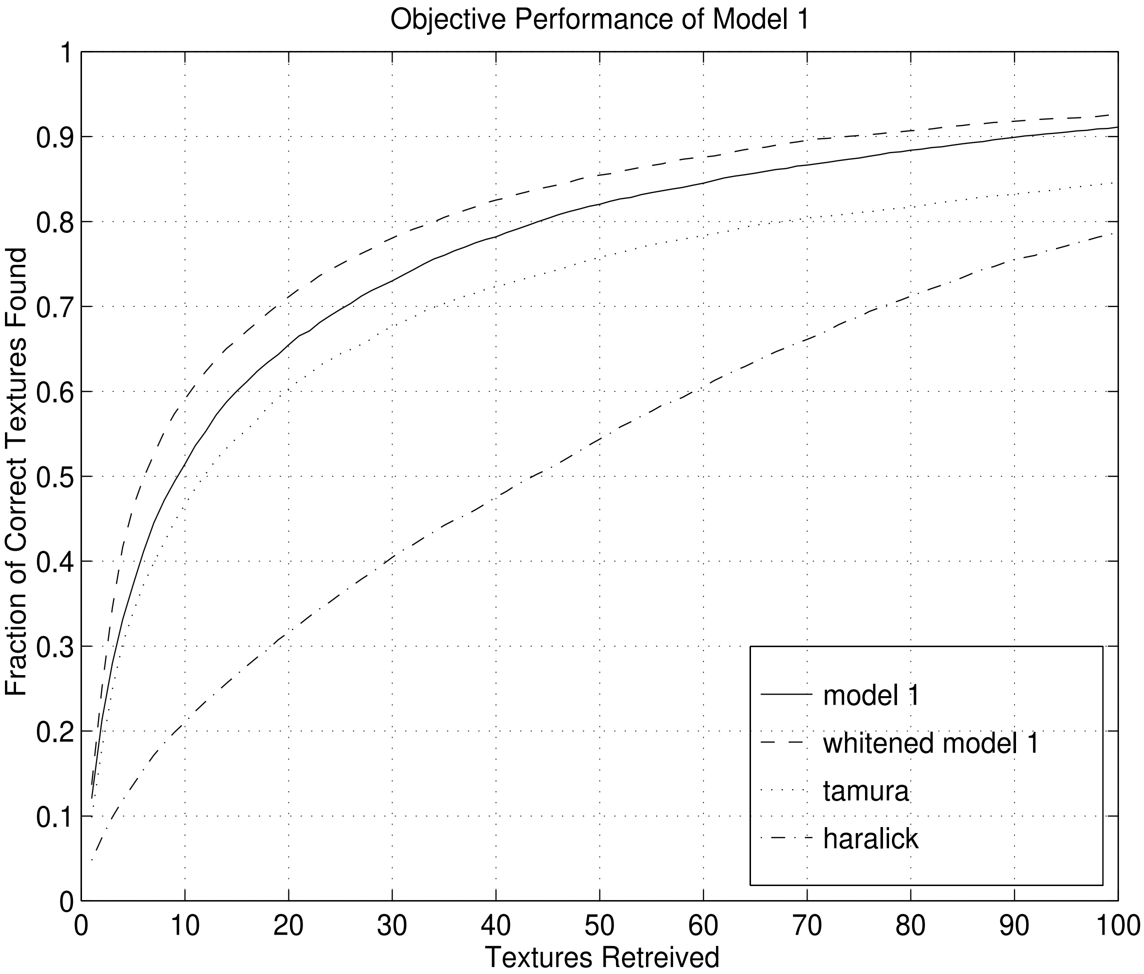

Figure 4.3: Objective Performance of Simple Oriented Features

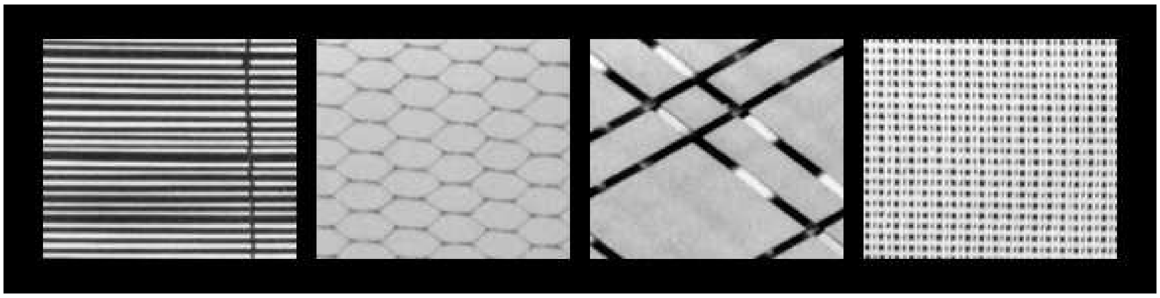

Figure 4.4: Simple oriented features perform perfectly on these texture classes.

as samples of a common texture class. For example, witness the two samples of the 'swinging light' class. One sample has virtually zero intensity, posing certain problems to the judgment that the first pair in Figure 4.5 come from the same class. This example illustrates an ongoing dilemma: the most subjective judgement in texture research is the judgement of whether or not a particular image is or is not a texture. Such images ought not to be included in the database.

Inhomogeneous textures, like the tree bark in Figure 4.5 are non-uniform surfaces photographed from a distance not much larger than the characteristic scale of the texture's surface height variations. Under strong illumination conditions, average oriented energy features vary considerably across the surface of these textures. When samples like the tree bark texture pair in Figure 4.5 are compared using a simple, euclidean metric, the result is noise, and the increased potential for confusion with samples from other classes.

The samples of Oriental glass fiber cloth in Figure 4.5 illustrate a texture class containing *specularities*. Specularities are directionally uniform surface inhomogeneities, and can be caused by gaps in annealed materials, bulges in non-rigid materials, or by the non-uniform density of embedded threads in fabrics. Specularities cause the overall magnitude of oriented features to change, while leaving the distribution of energy across each orientation direction relatively unchanged.

Rotational textures present a basic challenge to simple oriented features. Wholesale rotations of a texture cause major migrations in oriented energy between features, quickly separating members of the same class in a euclidean space. Similarity, local rotation noise, as illustrated by the calfskin textures in Figure 4.5, often appears in non-rigid, fur-like materials and textures with multiple surfaces.

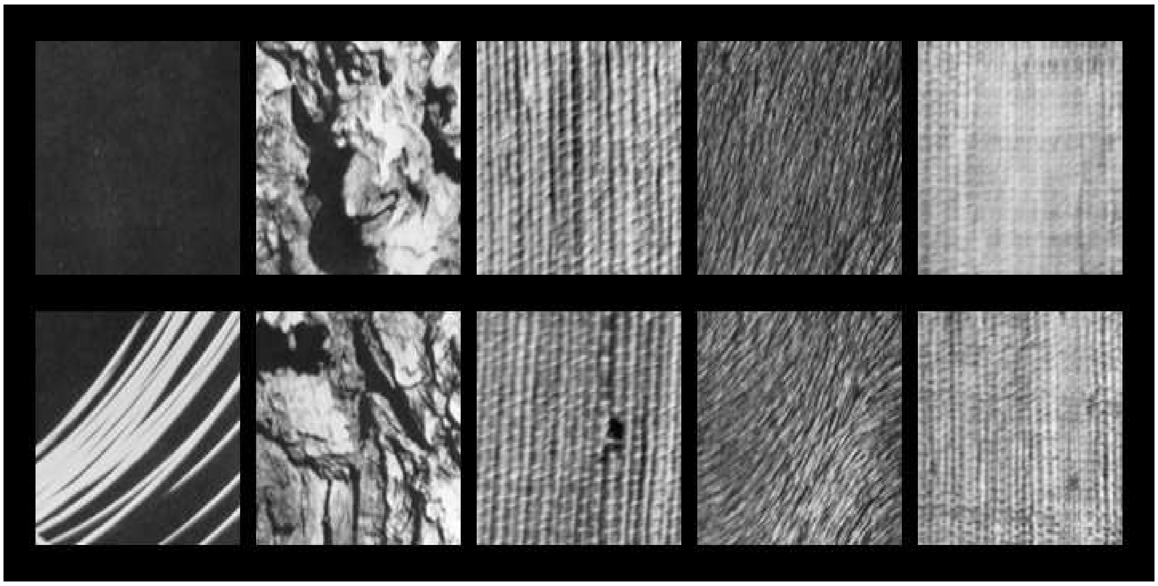

Figure 4.5: Simple oriented features perform poorly on these texture classes.

Finally, uneven surface weathering can reduce the overall magnitude of oriented energy in sample patches, as evidenced by the rightmost samples of oriental straw cloth in Figure 4.5.

Using the whitening procedure described in the previous chapter, the dimensions of model 1 were uncorrelated and distribution of variance along the uncorrelated dimensions were charted. Examining this distribution shows that 96% of the total variance in the space lies along the first two dimensions. This is understandable; because the power spectra of steerable filters overlap, adjacent orientation bands will be strongly correlated.

If so much of the total information in the space lies along the first two features, why not discard the low-variance features? If the variance along a particular dimension is very small, then on average, this direction will contribute almost nothing to the euclidean distance between texture pairs. However, there is reason to believe that the low-variance features in model 1 still contain useful information. Note that using only horizontal and vertical orientations would also decrease correlations between bands, yet it would be difficult to discriminate a 45° diagonal line from a circle using just these two features, since there would be equal average amounts of horizontal and vertical energy in each of these stimuli.

The prediction that the low-variance dimensions of model 1 contain useful information was tested by using whitening to equalize the variance along all four uncorrelated dimensions of model 1, and comparing the performance of the whitened and unwhitened models. As shown in Figure 4.3, whitened model 1 betters the unwhitened model by 8% at 10, 3% at 50, and 1% at 100 textures retrieved, a significant improvement

4.4.2 Model 2: Normalized Orientation

The performance of model 2 is plotted in Figure 4.6. Model 2 performs terribly, recalling less than 18% of the correct samples at 10 textures retrieved, a drop of over 41% from whitened model 1.

Why does this model perform so poorly? Examining the variances along the uncorrelated dimensions of model 2 provides the answer: virtually 100% of the variance is explained by a single dimension, the mean energy in the cummulative energy map. This makes perfect sense; by design, normalized features are limited to the range [0,1], yet the cumulative energy map has no such restriction on its range of output values. Examining the variance of unwhitened model 2 features averaged across the entire Brodatz database shows that the three orientation-normalized features have variances on the order of 3×10^{-3}, while the variance of the cummulative enegy map is on the order of 3×10^{2}, a difference of 5 orders of magnitude. The result is that the euclidean distance between texture pairs in the database under model 2 is determined almost completely by the single cumulative energy feature.

Given that model 2 is essentially a one dimensional model based on overall oriented energy, an illustration of the best and worst performing classes is shown in Figure 4.7. Good classes, shown on the left, included straw screening and an abstract illusion of woven wire. Bad classes, shown on the right, included crocodile skin and tree bark. Not surprisingly, the best performing classes are either very strongly or very weakly oriented, while the worst classes are inhomogeneous.

Clearly, whitening model 2 ought to improve performance, since the distribution of oriented energy by direction will become a factor in the computation of distance between texture pairs. As shown in Figure 4.6, whitened model 2 outperforms the unwhitened model 2 by more than 46% at 10 textures retrieved, and also betters

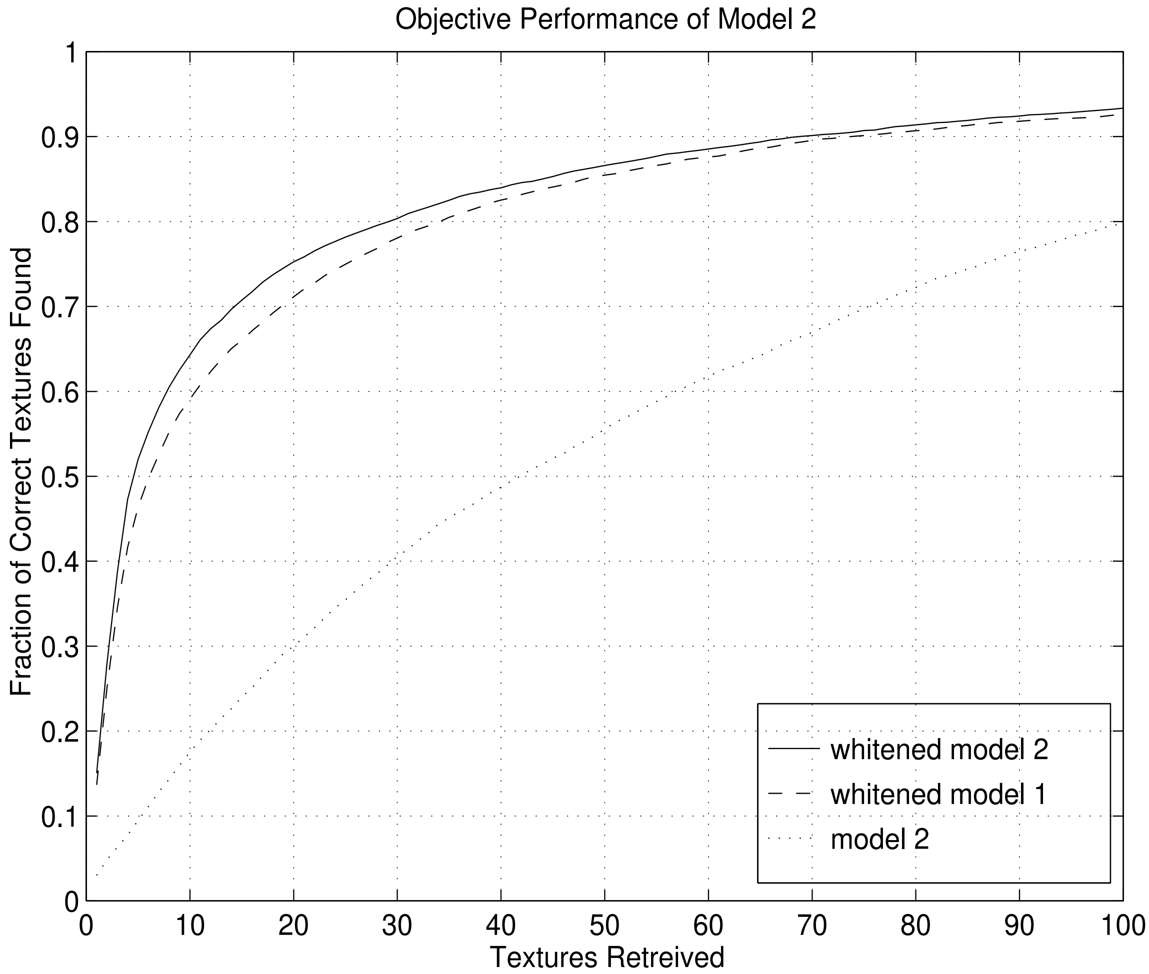

Figure 4.6: Objective Performance of Normalized Orientation Model

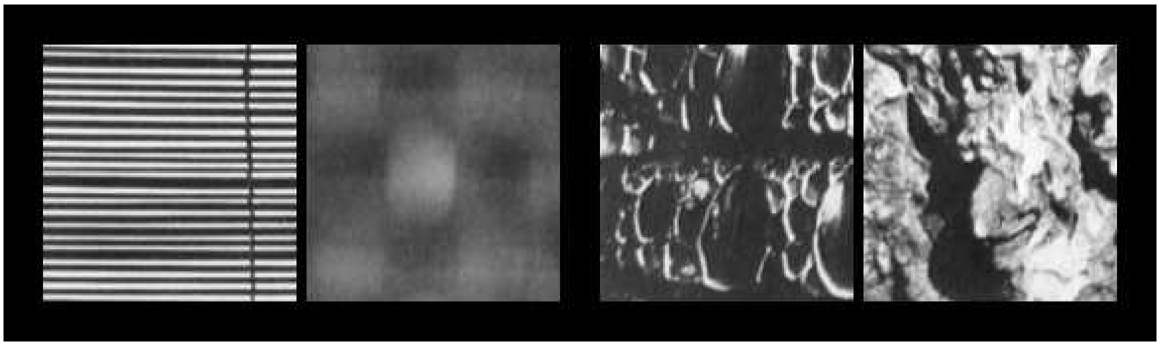

Figure 4.7: Good and bad texture classes under normalized features.

model 1 by 5% at 10 textures retrieved.

Investigation of the most improved classes between model 1 and whitened model 2 suggests that normalization provides enhanced robustness to the sources of surface noise that account for the worst classes in model 1; performance on the tree bark, oriental glass fiber cloth, calfskin, and oriental straw cloth classes shown in Figure 4.5 improves dramatically.

4.4.3 Model 3: Multi-scale Orientation

The performance of model 3 is shown in Figure 4.8. Model 3 betters the eigenvector model by about 6% at 10 textures retrieved. Model 3 also outperforms whitened model 2 by about 10% at 10, 5% at 50, and 2% at 100 textures retrieved, a significant improvement. Investigating the relative performance of model 3 versus whitened model 2 highlights scale as a useful discriminator. Consider the whitened model 2 texture confusion shown in Figure 4.9. Model 2 judges the large brick to be more similar to the small brick key than another sample of small brick, a clear error. Under model 3, this judgement is corrected through the addition of scale information. Further, the most improved textures between whitened model 2 and model 3, shown in Figure 4.10, were typically textures that had a very uniform size.

Variance analysis of model 3 shows that 97% of the total variance lies along the first 6 uncorrelated dimensions, with a gradual falloff in variance. Like orientation, strong correlations across scale are to be expected since the power spectra of successive low-pass kernels overlap considerably.

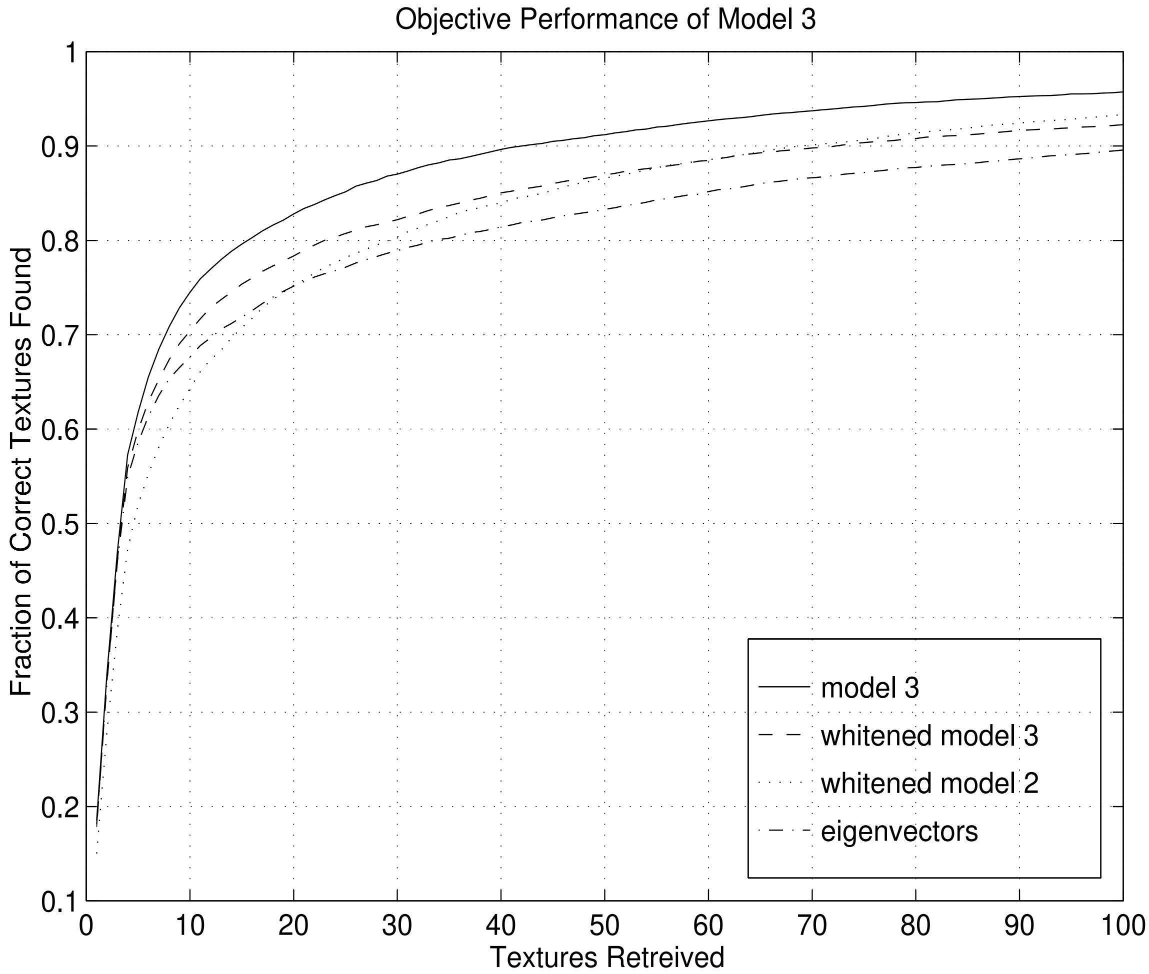

Figure 4.8: Objective Performance of Multi-Scale Model

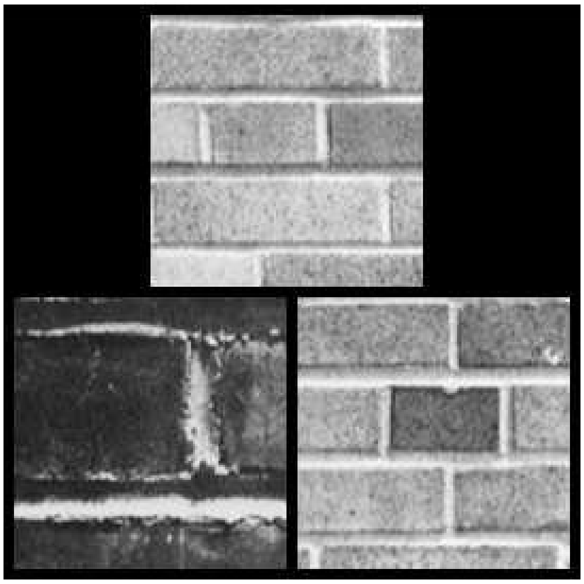

Figure 4.9: A triad confused by whitened model 2, corrected by model 3.

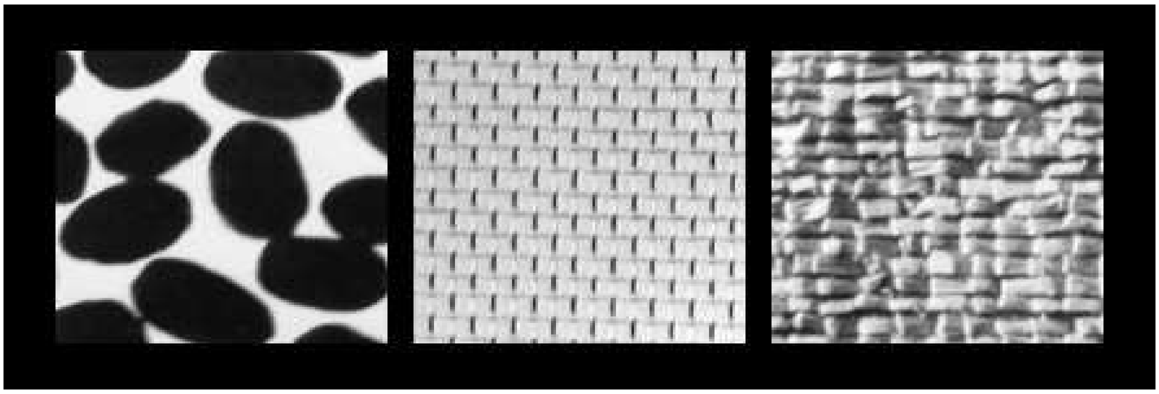

Figure 4.10: Most improved classes between whitened model 2 and model 3.

Whitening model 3 results in a marginal 4% loss in performance at 10 textures retrieved. One hypothesis for the performance decrease between model 3 and its whitened counterpart is that, unlike model 2, dimensions with low variance in model 3 provide little useful information, acting like noise sources in the computation of distance. This hypothesis was tested by incrementally removing the lowest-variance dimensions of whitened model 3 and measuring the relative performance change. The results are shown in Figure 4.11. Successive removal of the 1, 3, and 5 lowest-variance dimensions of the whitened model resulted in a steady performance increase. This is an attractive property, allowing the representation of textures to be compressed while simultaneously improving overall retrieval performance.

4.4.4 Model 4: Normalized, Multi-scale Orientation

The performance of model 4 is shown in Figure 4.12. Like unwhitened model 2, unwhitened model 4 performs poorly due to cummulative energy features that dominate the distance computation. However, by adding scale, model 4 betters model 2 by more than 37%. Whitened model 4 performs slightly worse than model 3, falling 2% at 10 textures retrieved.

4.4.5 Model 5: End/Side Cascade, Multi-scale Orientation

As shown in Figure 4.13, addition of cascaded features in model 5 results in a modest 3However, examining the eigenvalue distribution of model 5 shows that 97% of the variance can be explained by the first 7 of 60 dimensions. This poor distribution of information content underscores the problems raised by strong correlations between features, and suggests that a model of inhibition across cascades will be a necessary step to producing a scalable architecture.

Whitening model 5 results in a dramatic performance fall off, down 14% from the unwhitened model at 10 textures, shown in Figure 4.13; this is consistent with the interpretation of low-variance dimensions as noisy features. Compressing the space using only the 7 largest-variance dimensions brings performance back to par with the unwhitened model 5.

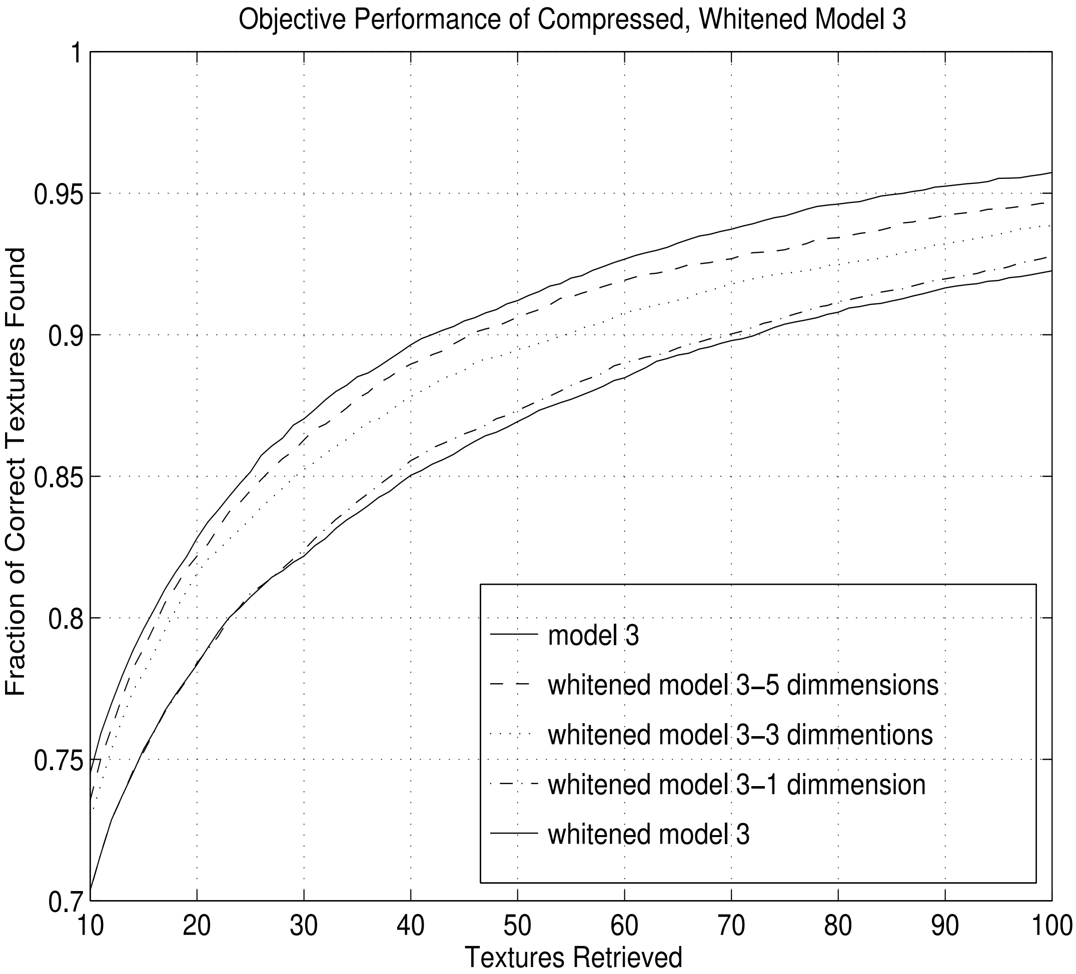

Figure 4.11: Objective Performance of Whitened, Compressed Multi-Scale Model

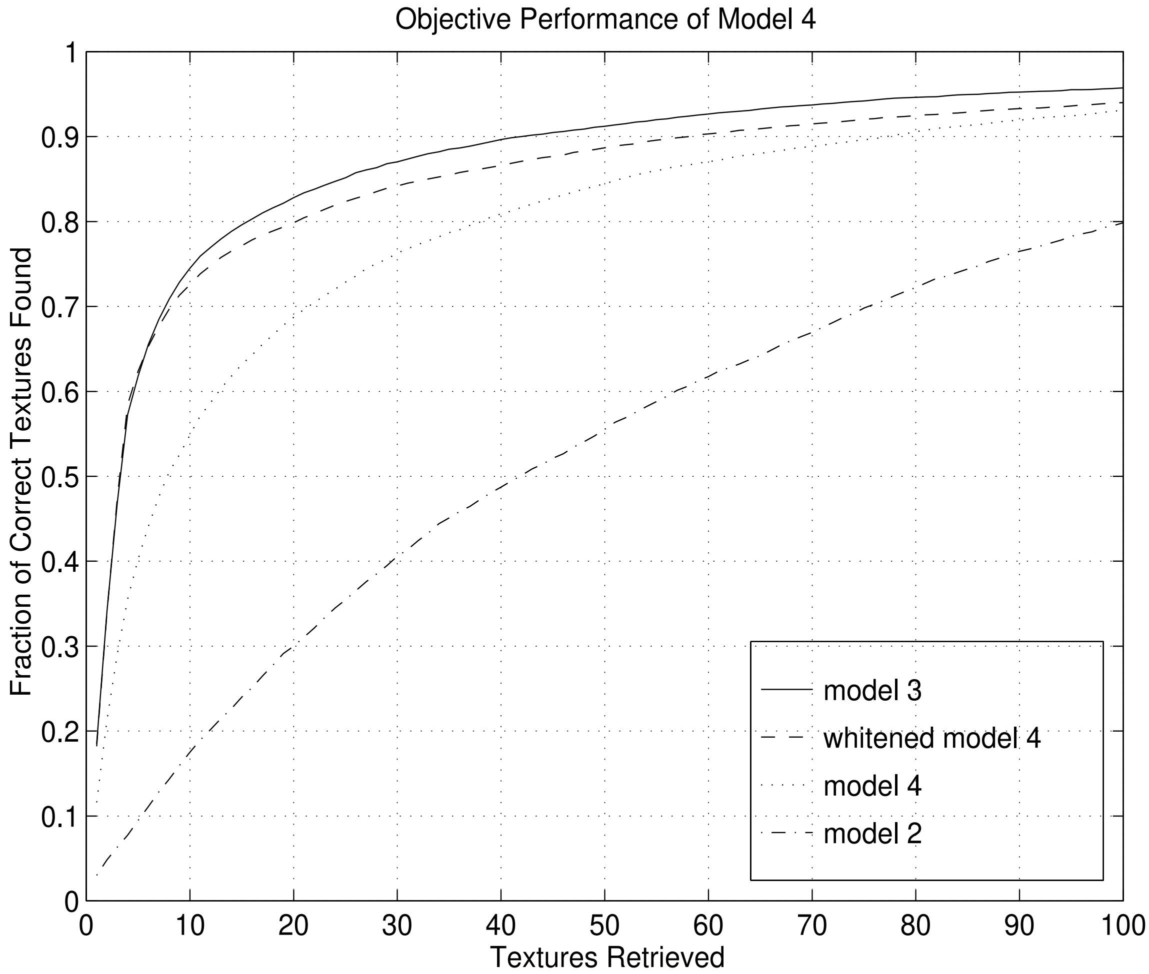

Figure 4.12: Objective Performance of Normalized, Multi-Scale Model

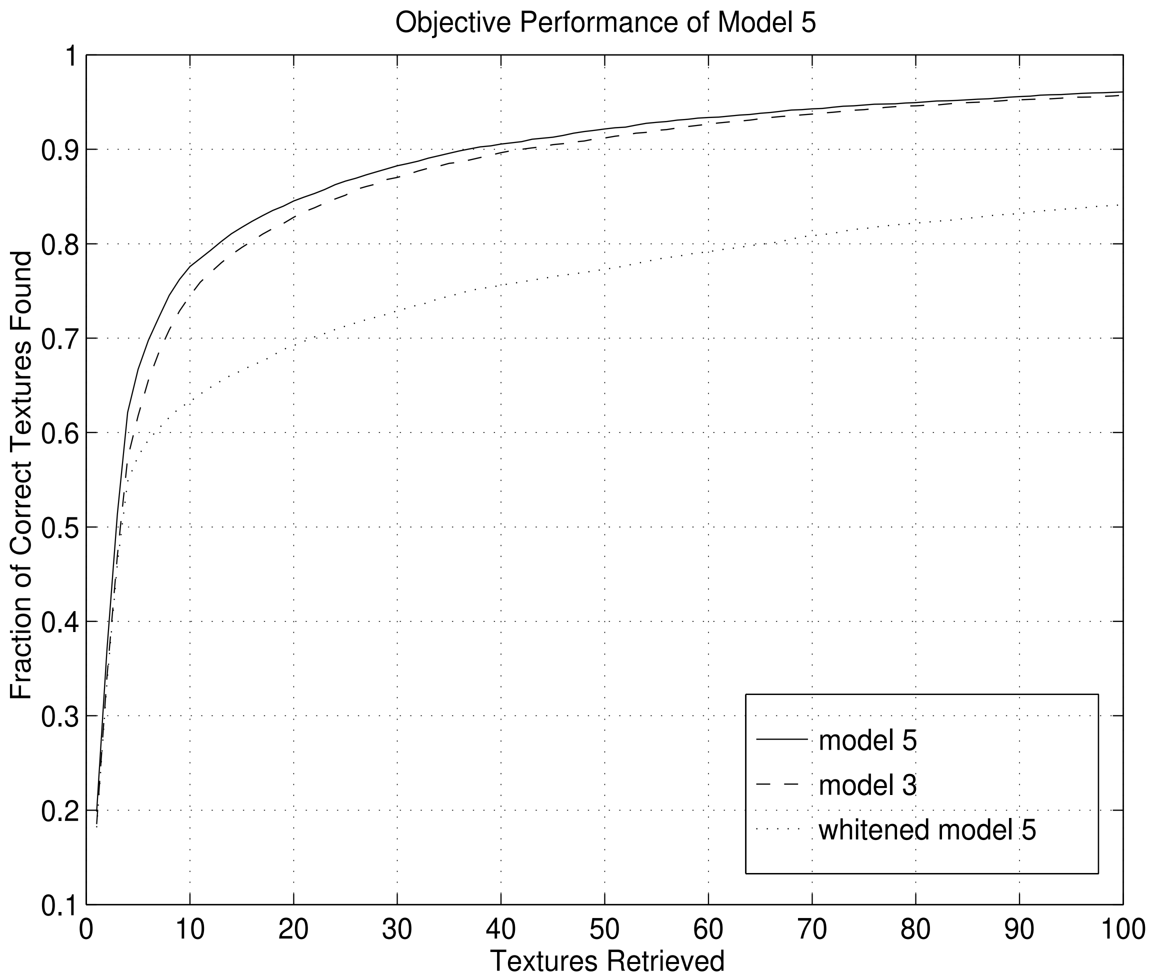

Figure 4.13: Objective Performance of End/Side Cascaded, Multi-Scale Model

4.4.6 Model 6: End/Side Cascade, Normalized Multi-scale Orientation

Model 6 performs terribly (Figure 4.14), due to the same strong correlation problems that plauge model 5; the variances along the uncorrelated dimensions of model 6 range over 17 orders of magnitude.

Using the successive compression technique discussed previously, the 49, 52, 55, 58, 61, 64, and 67 lowest-variance dimensions of whitened model 6 were successively discarded and relative performance changes were measured. The result is shown in Figure 4.15; performance increases steadily as low-variance dimensions are removed from the whitened space, peaking at 64 dimensions removed. The best model achieves an objective recall accuracy of 80% at 10 textures, within 8using just 8 numbers to represent each texture.

4.4.7 Model 7: Adding Variance Features

The models reviewed so far trade spatial expressiveness for compactness by using a simple mean feature to describe the sub-band energy images of a texture. As shown, these models can be both compact and highly accurate. Yet could one do better by using a slightly more complex description of the energy sub-bands?

The features of Model 7 are extracted using almost the same procedure as Model 6; the only difference is that rather than describing each texture as a single point in a multi-dimmensional space using sub-band means, Model 7 describes a texture as a multi-dimmensional gaussian distribution using both the mean and variance of each sub-band. For simplicity, this distribution is modeled using a diagonal covariance matrix. A single texture is therefore represented under Model 7 using 72 mean and 72 diagonal variance features.

The performance of Model 7 is shown in Figure 4.16, plotted against the best-known (SAR-divergence) model, and the best compressed model 6. As shown, model 7 performs on-par with SAR (down only about 2% at 10 textures retrieved) while using a representation almost half its size.

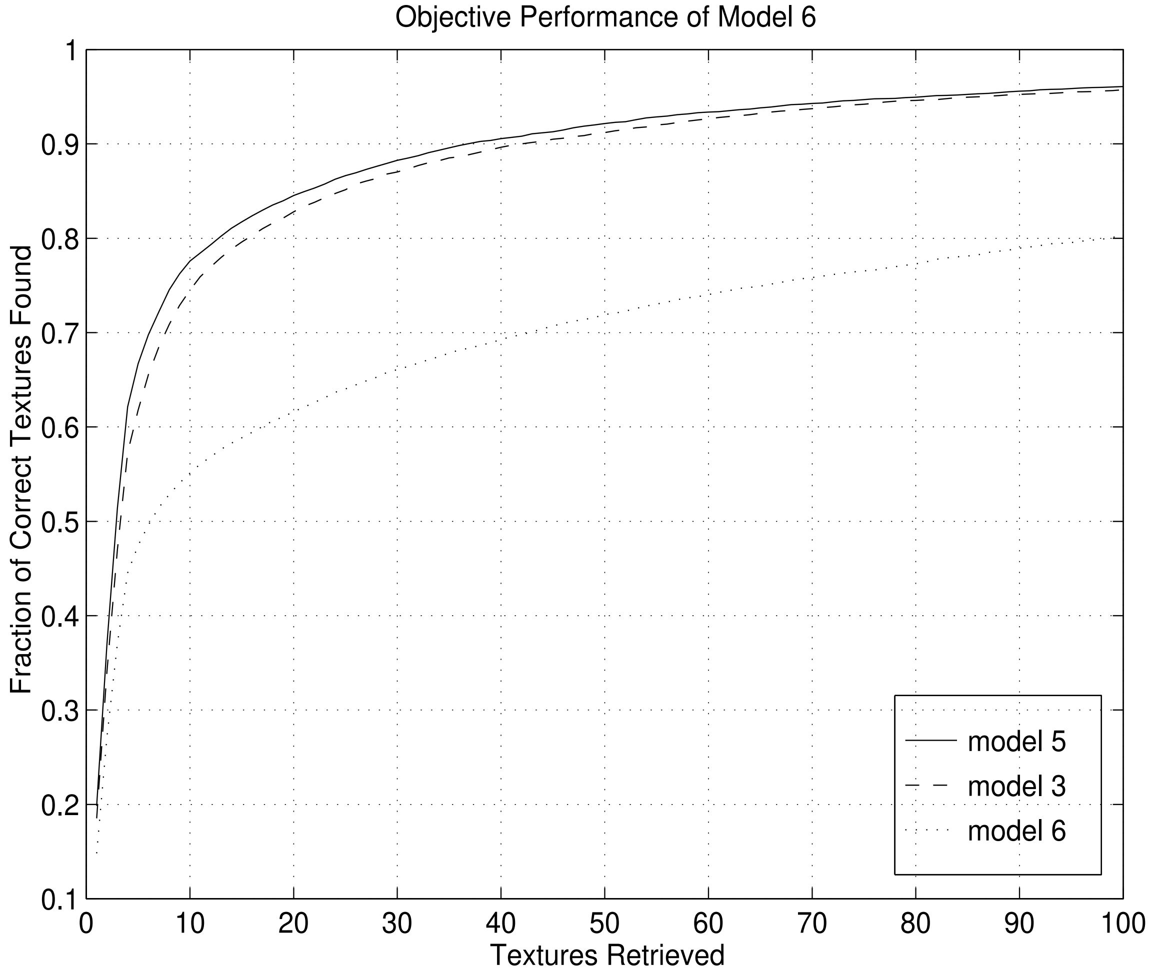

Figure 4.14: Objective Performance of End/Side Cascaded, Normalized Multi-Scale Model

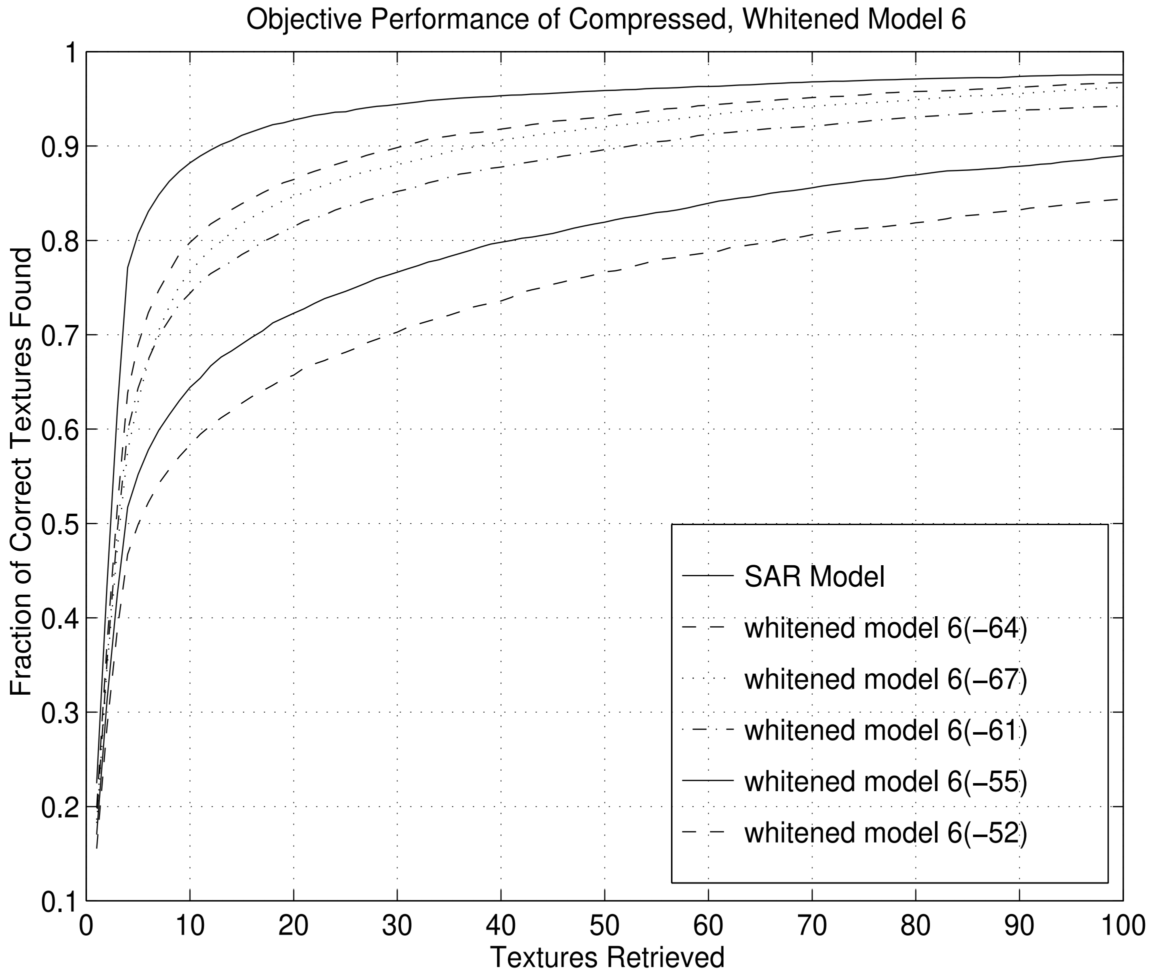

Figure 4.15: Objective Performance of Compressed, Whitened Model 6

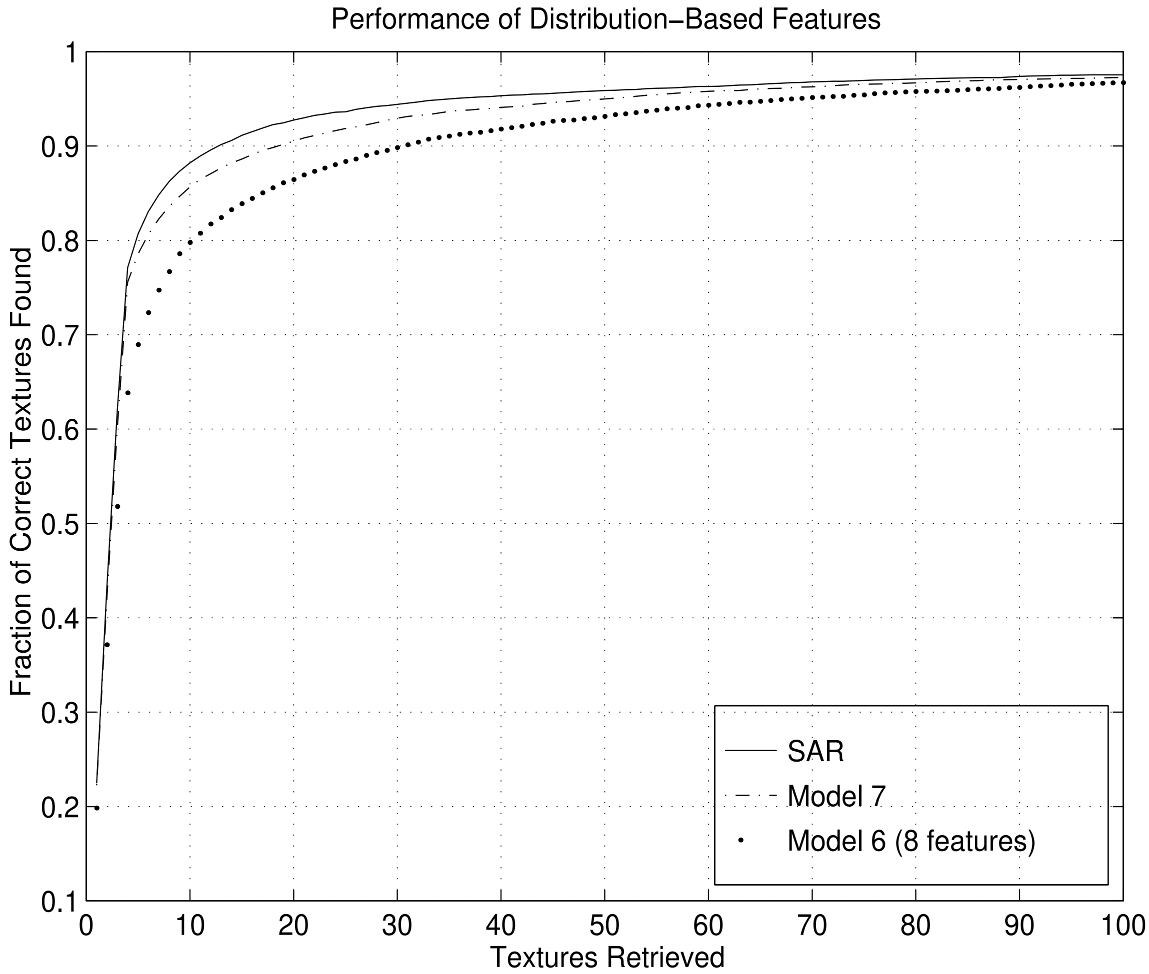

Figure 4.16: Objective Performance of Distribution-Based Features

4.4.8 Cost-Performance Summary

Figure 4.17 illustrates the cost-performance characteristics of the models examined in this thesis. The cost axis is the log total number of floating-point numbers used to represent a single texture. The performance axis is the log average fractional error at 10 textures retrieved. As summarized by this graph, models based on early-visual features demonstrate accuracies and cost/performance characteristics comparable to, or better than the best known texture analysis models.

4.5 Subjective Performance Criterion

Subjective accuracy was evaluated by comparing the orderings produced by a representative model against those of human subjects. Rao [39] demonstrates that humans categorize Brodatz textures along the three orthogonal dimensions of periodicity, directionality, and complexity. Working from this result, a simple psychophysical experiment was designed to gauge how humans judge texture similarity within these three dimensions.

Texture triads like the one shown in Figure 4.18 were presented to subjects on a computer screen for 2 seconds, then removed. Subjects were then asked to decide which texture, the one on the left or the one on the right, was more similar to the texture in the center. Subjects responded by entering a (1) to indicate left, and a (0) to indicate right.

As stimuli for this experiment, three sets of 6 six textures were chosen from the Brodatz database (Figure 4.19). Working from Rao's findings, the textures in these three sets were chosen to correspond to representative periodic, directional, and complex textures. In Figure 4.19, the firt row are the periodic stimuli, the second row are the directional stimuli, and the third row are the complex stimuli. The mean and variance of the image intensity for all textures were equalized prior to presentation. Three experimental runs were conducted using all 60 possible symmetric comparisons in each of the three sets. Five subjects were used, with each subject completing all three runs.

In addition to the 5 human subjects, Model 3 was used to produce synthetic similarity judgements for the same three sets of textures.

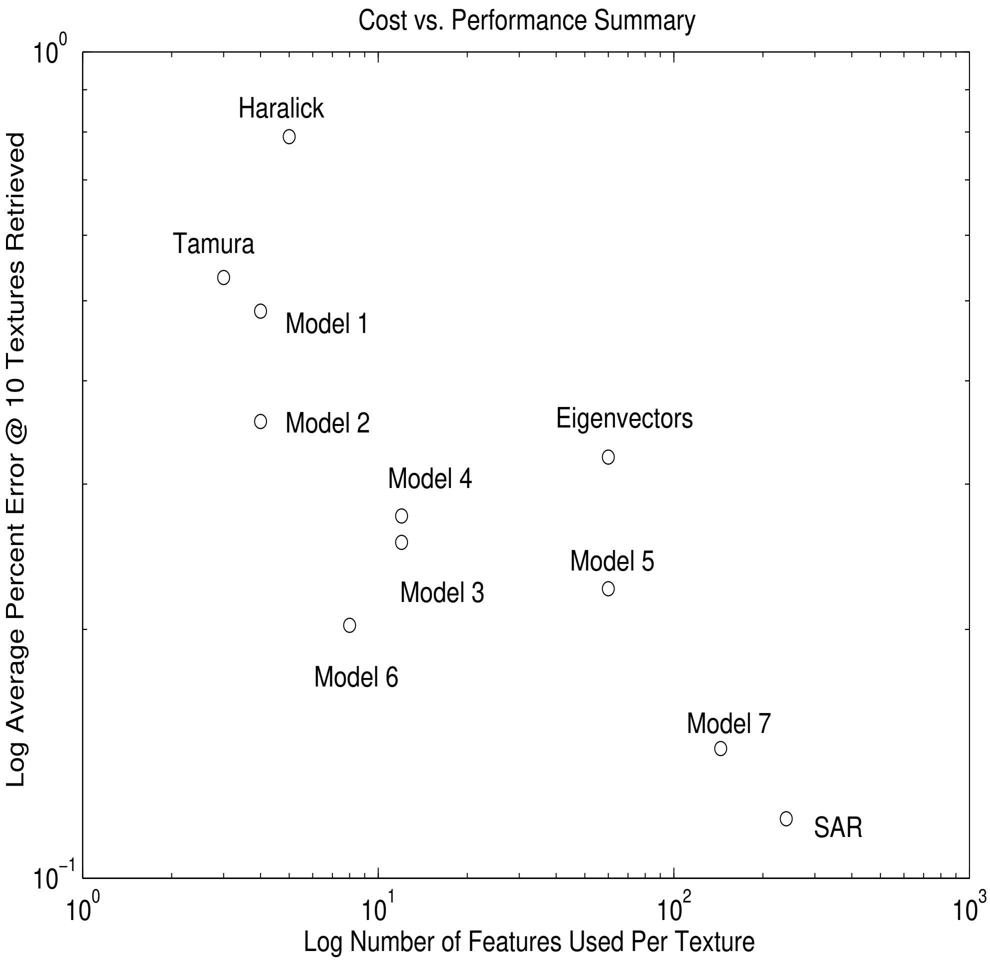

Figure 4.17: Cost/Performance Summary of the Models Examined in this Thesis

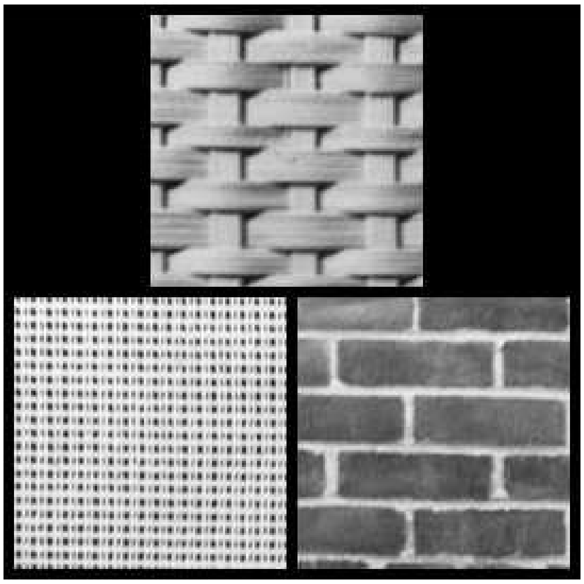

Figure 4.18: A sample texture stimulus used in the subjective experiments

4.6 Analysis of Subjective Performance

The judgements rendered by the 5 human subjects were sorted according to variance. Judgements on which all subjects agreed were isolated and used as a basis for evaluating the performance of the synthetic model.

In the periodic, directional, and complex data sets, all 5 subjects agreed on 15 of 60, 17 of 60, and 26 of 60 total judgements respectively.

The judgements produced by the synthetic model matched those of human subjects 13 of 15 times in the periodic set, 15 of 17 times in the directional data set, and 23 of 26 times in the random data set, for a cummulative predictive accuracy of approximately 88%.

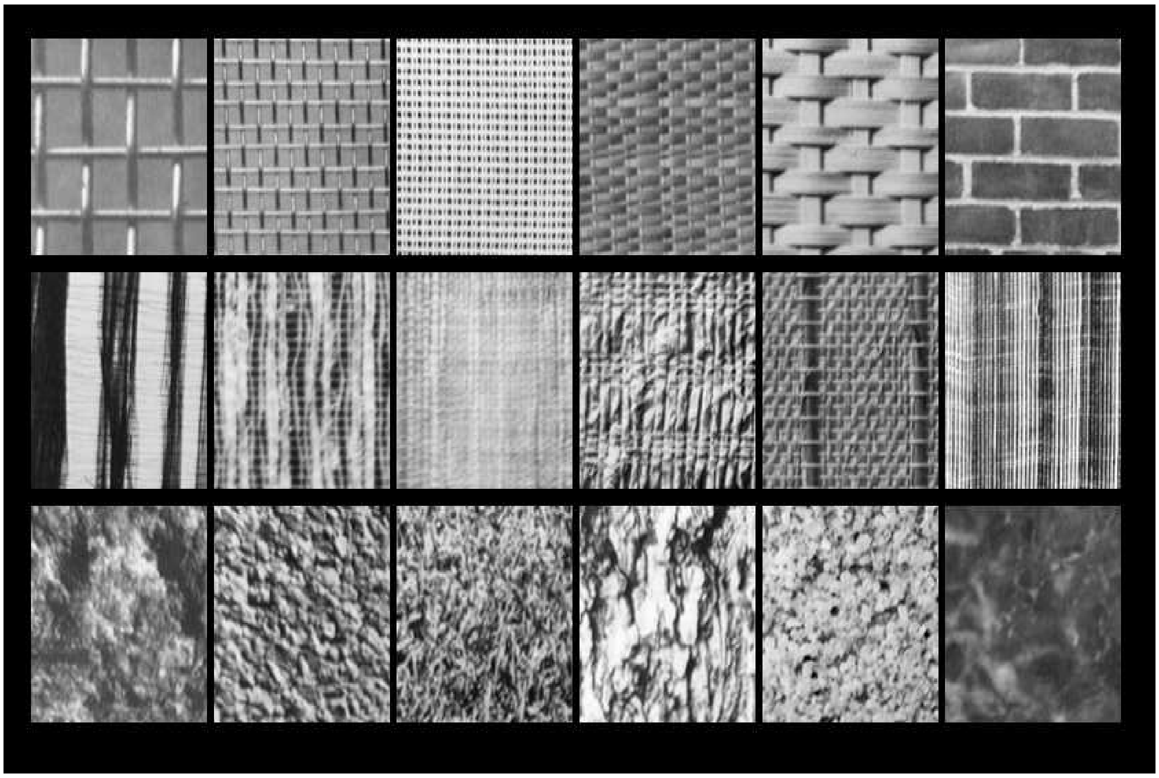

Figure 4.19: Stimuli used in the subjective experiments.

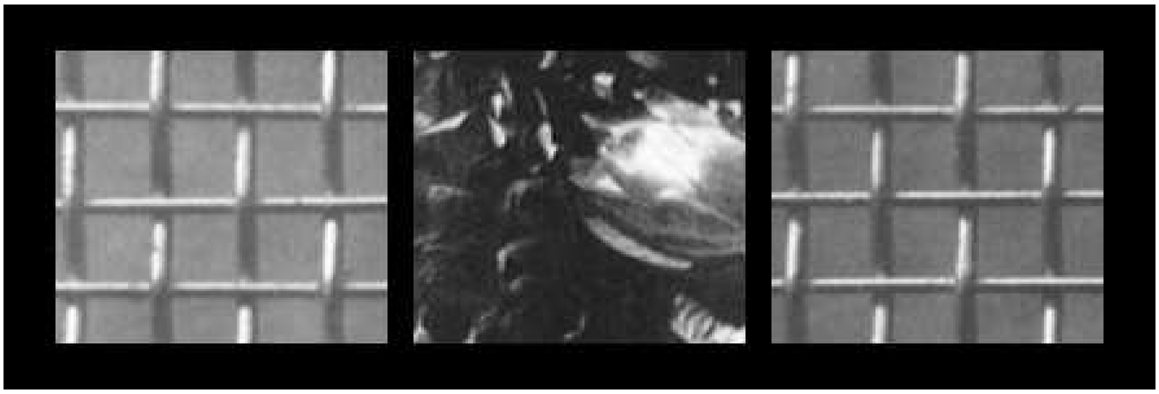

Figure 4.20: Example of a subjective texture confusion due to poor spatial modeling.

4.7 Limitations and Future Research

Despite the favorable objective and subjective achievements of the early-visual models, interactively browsing the Brodatz Database highlights difficult problems.

Consider for example,the judgement shown in Figure 4.20. This example is representative of the many subjective confusions that typically occured in free image retrieval using the early-visual models. Most humans feel that the two samples of aluminum wire are very similar because they are both periodic looking textures. However, by using features averaged across position, model 6 judges the aluminum wire key to be more similar to feathers than to the other aluminum wire sample. Clearly, the early-visual models are failing to capture spatial qualities that humans consider to be 'obvious'.

As discussed earlier, many approaches to modeling the spatial character of textures, including the SAR model, use relative position as an explicit parameter. One advantage of this approach is that by representing position explicitly, these models provide an expressive basis for spatial comparisons. SAR, for example, does not commit the error shown in Fig.20 (although it too makes many subjective judgements that humans would consider to be quite poor). One disadvantage of this approach is that typically, many parameters are used to characterize a particular texture.

This thesis has pursued an alternative approach to the representation of space. Instead of using position as an explicit parameter, this thesis attempted to use cascades of filtering operations to split a texture into sub-band components, next using a simple first-moment operation to describe these sub-bands. By using the spatial constraints implicit in the underlying filters and cascading order, a texture could be represented with very few parameters.

Unfortunately, although simple early-visual features appear to be both strong predictors of perceived texture similarity within Rao's dimensions and stable anchors for texture retrieval, confusions like those of Fig.20 suggest that a more direct test of spatial discrimination is needed. In particular, the accurate classification of Brodatz textures into periodic, directional, or complex groups based on Rao's findings would make a challenging litmus test for future early-visual models.

Chapter 5

Conclusions

This thesis has tested the expressiveness of early-visual features as determinants of perceived texture similarity. The first aim of this thesis was to illuminate the gap between these low-level visual primitives, and high-level judgements. The second aim was to assess the practical merits and limitations of an image retrieval model grounded in early-visual operations.

A taxonomy of early-visual models were used to retrieve natural textures from the Brodatz database. Objective and subjective criteria were used to evaluate the performance of models in this taxonomy.

The average fraction of correct samples found versus the total number of textures retrieved was used as a criterion for objective performance.

Simple orientation features performed surprisingly well, recalling on average over 50% of the correct samples within the first 10 textures retrieved, while using only 4 numbers to represent each texture. Analyzing poorly performing classes highlighted characteristic sources of intra-class noise that fundamentally limit the performance of oriented features, including variations in the direction of oriented energy due to local rotations in non-rigid materials, and changes in the overall magnitude of oriented energy due to strongly illuminated, inhomogeneous surfaces.

Normalized features empirically outperform simple orientation features when whitened, providing improved robustness to common sources of intra-class noise.

Multi-scale features broadly improve retrieval performance, capturing distinctions between homogeneous and inhomogeneous materials that simple orientation features are too weak to represent.

The amount of original information added by cascaded features scales very poorly with the total length of the feature vector, adding virtually no new degrees of freedom to the representation due to strong feature correlations. These correlations are caused in part by the overlapping spectra of the underlying filters. Whitening the 72 features generated by a cascaded model reveals that 8 dimensions of the whitened space explain virtually 100% of the total variance.

A useful empirical relationship is discovered between dimensions with low variance and the recall performance of whitened, early-visual models; systematically discarding dimensions with low variance after whitening simultaneously compresses the space and monotonically improves overall accuracy.

The most accurate early-visual outperforms well-known models by Tamura [43] and [15], and achieves objective performance within 2% of the best known SAR model. In addition, the early-visual models achieve very favorable cost/performance characteristics: the leanest model achieves an accuracy of 80% at 10 textures retrieved using only 8 numbers to represent each texture in the database.

A simple psychophysical experiment was used to gauge how well simple early-visual features predicted human judgements of texture similarity within three textural dimensions. Three sets of representative periodic, directional, and complex textures were used as stimuli for triadic ordinal comparisons. 5 subjects were used, with each subject completing all three texture sets. All 5 subjects used produced identical judgements on 15 of 60 periodic texture comparisons, 17 of 60 directional texture comparisons, and 26 of 60 random texture comparisons. A computational model based on oriented, multi-scale features was able to successfully predict 88% of these consensus judgements.

The limitations of the models examined here suggest a milestone for future work: the classification of natural textures into periodic, directional, or complex groups is emphasized as a litmus test for the spatial expressiveness of filtering cascades and low-order statistical features.

Bibliography

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

[2] J. Bergen. Theories of visual texture perception. In D. Regan, editor, *Spatial Vision (Vol. 10 of Vision and Visual Disfunction)*, pages 114–134. Macmillan, New York, 1991.

[3] J. Bergen and E. Adelson. Early-vision and texture perception. *Nature*, 333(6171):363–364, 1988.

[4] J. Bergen and M. Landy. Computational modeling of visual texture segregation. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing*, pages 253–273. MIT Press, Cambridge, 1991.

[5] A. Bonds. Role of inhibition in the specification of orientation selectivity in the cat striate cortex. *Visual Neuroscience*, 2:41–55, 1989.

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

[7] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover, New York, 1966.

[8] G. Cross and A. Jain. Markov random field texture models. *IEEE Transactions on Pattern Matching and Machine Intelligence*, 5:25–39, 1983.

[9] R. DeValois, D. Albrecht, and L. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22:545–559, 1982.

[10] R. DeValois, E. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22:531–544, 1982.

[11] A. Dobbins, S. Zucker, and M. Cynader. End-stopping and curvature. *Vision Research*, 29:1371–1387, 1989.

[12] W. Freeman. The design and use of steerable filters. *IEEE Transactions on Pattern Matching and Machine Intelligence*, 13(9), 1991.

[13] W. Freeman. *Steeable Filters and Local Analysis of Image Structure*. PhD thesis, MIT EECS, 1992.

[14] D. Getty, J. Swets, and D. Green. On the prediction of confusion matrices from similarity judgements. *Perception and Psychophysics*, 26(1):1–19, 1979.

[15] R. Haralick. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6), 1973.

[16] P. Heckbert. Survey of texture mapping. *IEEE Computer Graphics and Applications*, 6(11):56–67, 1986.

[17] D. Heeger. Half-squaring in responses of cat striate cells. *Visusal Neuroscience*, 9:427–443, 1992.

[18] D. Heeger. Normalization of cell responses in cat striate cortex. *Visusal Neuroscience*, 9:181–198, 1992.

[19] F. Heitger, L. Rosenthaler, R. Von der Heydt, E. Peterhans, and O. Kubler. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research*, 32:963–981, 1992.

[20] D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195:215–243, 1968.

[21] B. Julesz. Visual pattern discrimination. *IEEE Transactions on Information Theory*, 8(2):84–92, 1962.

[22] B. Julesz, H. Frish, E. Gilbert, and L. Shepp. Inability of humans to discriminate between visual textures that agree in second-order statistics - revisited. *Perception*, 2:391–405, 1973.

[23] T. Kabir and R. Picard. Finding similar patterns in large image databases. In *IEEE ICASSP Vol. V*, pages 161–164, 1993.

[24] K. Knutsson and G. Granlund. Texture analysis using two-dimmensional quadrature filters. *IEEE Workshop on Pattern Analysis and Image Database Management*, pages 206–213, 1983.

[25] J. Koenderink and A.J. van Dorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–376, 1987.

[26] R. Linsker. Self-organization in a perceptual network. *Computer Magazine*, 21:105–117, 1988.

[27] S. Mallat and J. Froment. Second generation compact image coding with wavelets. In C. Chui, editor, *Wavelets - A Tutorial in Theory and Applications*, pages 655–678. Academic Press, 1992.

[28] S. Mann and S. Haykin. The chirplet transform: A new signal analysis technique based on affine relationships in the time-frequency plane. Technical Report 250, MIT Media Laboratory Perceptual Computing, 1994.

[29] J. Mao and A. Jain. Texture classification and segmentation using multiresolution simultaneous auto-regressive models. *Pattern Recognition*, 25(2):173–188, 1992.

[30] J. Movshon, I. Thompson, and D. Tolhurst. Receptive field organization of complex-cells in the cat striate cortex. *Journal of Physiology*, 283:79–99, 1978.

[31] S. Niyogi and E. Adelson. Kinetic occlusion. Technical report, MIT Media Laboratory Perceptual Computing, 1994.

[32] A. Pentland, R. Picard, and S. Scarloff. Photobook: Tools for content-based manipulation of image databases. Technical Report 255, MIT Media Laboratory Perceptual Computing, 1994.

[33] R. Picard. *Texture Modeling: Temperature Effects on Markov/Gibbs Random Fields*. PhD thesis, MIT EECS, 1991.

[34] R. Picard and F. Liu. A new wold ordering for image similarity. Technical Report 237, MIT Media Laboratory Perceptual Computing, 1994.

[35] R. Picard and T. Minka. Vision texture for annotation. Technical Report 302, MIT Media Laboratory Perceptual Computing, 1994.

[36] D. Pollen, J. Gaska, and L. Jacobson. Responses of simple and complex cells to compound gratings. *Vision Research*, 28:25–39, 1988.

[37] D. Pollen and S. Ronner. Phase relationships between adjacent simple cells in the visual cortex. *Science*, 212:1409–1411, 1981.

[38] K. Popat and R. Picard. Novel clusterbased probability model for texture synthesis, classification, and compression. Technical Report 234, MIT Media Laboratory Perceptual Computing, 1993.

[39] R. Rao and G. Lohse. Towards a texture naming system: Identifying the relevant dimmensions of texture. In *IEEE Conference on Visualization*, 1993.

[40] R. Shapley, R. Reid, and R. Soodak. Spatiotemporal receptive fields and direction selectivity. In M. Landy and J. Movshon, editors, *Computational Models of Visual Processing*, pages 109–118. MIT Press, Cambridge, 1991.

[41] E. Simoncelli. Orthogonal sub-band image transforms. Master's thesis, MIT EECS, 1988.

[42] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger. Shiftable multi-scale transforms. Technical Report 161, MIT Media Laboratory Perceptual Computing, 1991.

[43] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):460–473, 1978.

[44] P. Teo and D. Heeger. Perceptual image distortion. In *IEEE ICIP vol. 2*, pages 982–986, 1994.

[45] D. Tolhurst and A. Dean. Evaluation of linear model of directional sensitivity in simple cells of the cat's striate cortex. *Visual Neuroscience*, 6:421–428, 1991.

[46] A. Triesman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[47] M. Tuceryan and A. Jain. Texture analysis. In L. Pau and P.Wang, editors, *The Handbook of Pattern Recognition and Computer Vision*, chapter 12. World Scientific Publishing, 1992.

[48] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.