

Higher order image pyramids: an early visual representation

Joshua Gluckman
Dept. of Computer Science
Polytechnic University
Brooklyn, NY 11201
jgluckma@poly.edu

Abstract

The scale invariant property of an ensemble of natural images is examined, which motivates a new early visual representation termed the higher order pyramid. The representation is a non-linear generalization of the Laplacian pyramid, tuned to the type of scale invariance exhibited by natural imagery as opposed to other scale invariant images such as $1/f$ correlated noise and the step edge. The transformation of an image to a higher order pyramid is simple to compute and straightforward to invert. Because the representation is invertible it is shown that the higher order pyramid can be truncated and quantized with little loss of visual quality. Images coded in this representation have much less redundancy than the raw image pixels and decorrelating transformations such as the Laplacian pyramid. This is demonstrated by showing statistical independence between pairs of coefficients. The representation is tuned to the ensemble redundancies, and hence the coefficients of the higher order pyramid are more efficient at capturing the variation within the ensemble which leads to improved matching results. This is demonstrated on several recognition tasks: object recognition with viewpoint changes, object recognition with scale changes, and face recognition with illumination changes.

Index Terms

image representations, image statistics, feature representation, multiscale, matching, recognition, early vision

I. INTRODUCTION

It has been argued for some time that the problem of early vision is to determine the most efficient method for representing images. Early formulations of this theory were given by Atteneave and Barlow who argued for redundancy reduction through the efficient coding hypothesis [1] [2]. The rationale is rooted in information theory, if later visual processes are to perform probabilistic inference, the input should be as statistically independent as possible. The implication for early vision is that images should be coded to match the expected input, the ensemble of natural images. Since, considerable advances have been made in understanding this ensemble by studying the statistics of natural images and arguably the most important feature to emerge is scale invariance (see [3] for a recent review). While multi-resolution representations, such as the Laplacian pyramid, the steerable pyramid, wavelets, and scale space, are motivated by scale invariance they are not necessarily tuned to the type of scale invariance found in images [4] [5] [6] [7]. In this paper, we describe a representation that is motivated by the higher

order scale invariant statistics of images.

The first statistical evidence of scale invariance was found by Field who discovered a $1/f$ power law in the amplitude spectra of natural images indicating that spatial correlations are scale invariant [8]. The spatial correlations of images can be removed by applying a “whitening” or derivative type filter. If the $1/f$ amplitude spectra was the only statistical regularity present in images, the distribution of coefficients of the filtered images would be Gaussian. However, the marginal statistics of the resulting coefficients are known to be highly non-Gaussian and sparsely distributed [8] [6]. Further evidence of scale invariance was found by Ruderman and by Zhu and Mumford who demonstrated that histograms of derivative filtered images are consistent across scale indicating scaling in the higher order statistics of images [9] [10]. Dependencies in the higher order statistics of decorrelated images have been documented by Simoncelli et al. and by Wegmann and Zetsche [11] [12]. The origin of scale invariance has been studied by creating random collage models that generate images with the above mentioned statistics [13] [14].

Although the presence of scale invariance in images is clear, no method exists for fully taking advantage of this property. One reason for this is the lack of a simple model that suggests how to represent the type of scale invariance found in images. One approach to develop a representation for images is to apply computational learning techniques. Examples of this approach include the sparse coding network of Olshausen and Field and the method of independent component analysis [15] [16]. Such methods learn a set of basis functions that are optimal for achieving independence, and there is a growing body of evidence that for natural images the optimal basis functions are oriented, localized and bandpass. Although these techniques are usually restricted to linear transformations, recently they have been extended to include non-linear hierarchical representations [17] [18]. Another approach is to assume a particular basis such as the steerable pyramid and learn a model of the dependencies of the coefficients in the transformed space. This approach has been taken by Simoncelli and colleagues who have suggested learning the parameters of a model of the joint statistics of various wavelet like transformations [19] [20] [21]. The primal sketch of Guo et al. is another representation motivated by image statistics [22]. In this approach, an image is described by two separate processes: a sparse code for the edges and a histogram based representation for the textured regions [22].

In this paper, we present a novel method for representing images which is motivated by the type of scale invariance found in the ensemble of images. This representation is a non-linear



Fig. 1. Scale invariance is studied by examining the statistics of images selected from the Van Hateren natural image database [25].

transformation based on the Laplacian pyramid of Burt and Adelson [4] and is referred to as the higher order image pyramid. Initial work on this representation is described in [23] [24]. In Section II the scale invariant statistics of natural images are studied. These studies reveal a particular form of scale invariance that motivates the use of higher order pyramids for representing images. In Section III the higher order pyramid is defined, and algorithms for building and inverting the pyramid are described. Section IV demonstrates that the higher order pyramid is a form of “higher order whitening” for images. A series of experiments are described in Section V that examine the properties of this representation. Because the transformation is invertible it is shown that the higher order pyramid can be quantized and truncated with little loss of visual quality. Once an image is coded the bits of the higher order pyramid exhibit far less redundancy than the raw image pixels and linear transformations such as the Laplacian pyramid. Redundancy reduction is shown by measuring the mutual information between pairs of coefficients in the pyramid. Tuning a representation to the ensemble redundancies implies the representation efficiently captures the variation within the ensemble which leads to improvements in recognition tasks. This is demonstrated by showing improved matching using the higher order pyramid on an object recognition task with varying viewpoint and scale changes and a face recognition task with varying illumination.

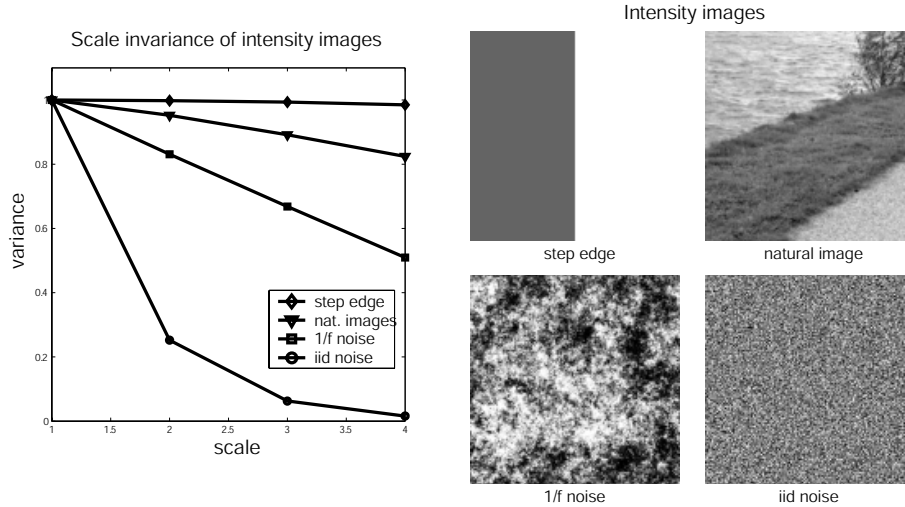


Fig. 2. The variance in intensity as a function of scale (the number of times the images are downsampled) for a pure step edge, a set of natural images, $1/f$ noise, and iid noise. The step edge, $1/f$ noise, and natural images all exhibit scale invariance indicated by a slow linear decay.

II. SCALE INVARIANCE

Here we examine the scale invariant statistical properties of images using the Van Haterren natural image database [25]. This database is a large collection of images of both pastoral and city scenes. For our experiments we selected 40 images that represent a variety of scenes (some examples are shown in Fig. 1). Each image is resized and cropped to 1024×1024 pixels and converted to log intensities.

Scale invariance is examined by considering the behavior of each particular image as the scale is reduced and averaging the behavior over the ensemble of images. Once particular images are considered the range of intensities is fixed, hence the moments are finite. Then, scale invariance appears in the expected value of the moments of the distribution of intensity as scale is reduced which is

$$E\{|I(x/s)|^k\}, \quad (1)$$

where $I(x)$ is an image, s is the spatial scale, and the expectation E is taken over an ensemble of images. We assume the mean intensity is subtracted from each image so k is a central moment.

Fig. 2 shows the result of estimating (1) for $k = 2$ using the 40 images selected from the van Haterren database [25]. A scale change of $I(x)$ is implemented by smoothing and downsampling.

Each image is repeatedly blurred, downsampled and a variance ($k = 2$) is estimated. For each scale, the variances of all 40 images are averaged. The plot shows the reduction in variance as a function of scale and compares the behavior of images to that of a pure step edge, $1/f$ noise, and i.i.d. noise. All plots are normalized to 1 at the first scale. The step edge, the images, and $1/f$ noise all exhibit scale invariance because the variance falls linearly. The same result is found for the higher moments implying that

$$E\{|I(x/s)|^k\} = E\{|I(x/(2s))|^k\} + K. \quad (2)$$

The actual value of the constant K is unimportant. However, for images, K lies somewhere between that of a step edge and $1/f$ noise,

$$K_{1/f} < K_{\text{images}} < K_{\text{step}}. \quad (3)$$

What differentiates images is the *rate* at which the moments decay with scale. Note that any one image may not have scale invariance. This property is found when averaged over an ensemble of images.

Multi-resolution representations, such as the Laplacian and steerable pyramids, are motivated by the fact that images exhibit scale invariance [4] [5]. These representations code

$$L(x/s) = I(x/s) - I(x/(2s)), \quad (4)$$

which is the prediction error between scales. However, such representations do not explicitly differentiate between step edges, $1/f$ noise, and natural images. To describe the type of scale invariance that images exhibit, L must be examined.

Although L , the levels of a Laplacian pyramid, are no longer scale invariant according to Eq. (2), they still resemble images, which suggests a hidden scale invariance. Indeed, if a new image, the Laplacian magnitude, is defined

$$I^{(2)} = \log |L|, \quad (5)$$

which is independent of the sign, then scale invariance reappears. The log function is used to reduce the sensitivity of the statistics to the extreme values. The fact that redundancy exists in the magnitude is well known and has been documented by Simoncelli et al. and Ruderman [11] [5] [26] [9]. Here it is shown that the form of the redundancy is scale invariance. Next, the same experiment is repeated except using $I^{(2)}$ for each of the 40 images, the step

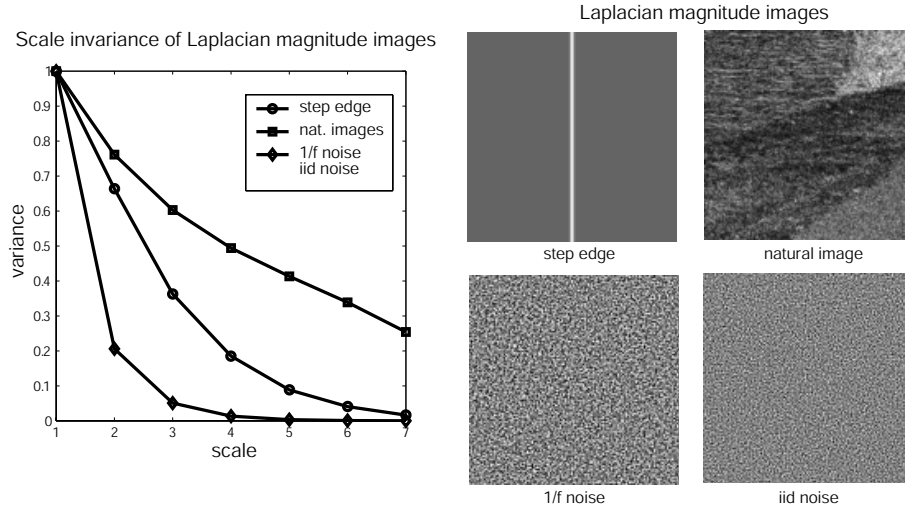


Fig. 3. The variance of the Laplacian magnitude as a function of scale for a pure step edge, a set of natural images, and $1/f$ noise. Only the natural images are scale invariant.

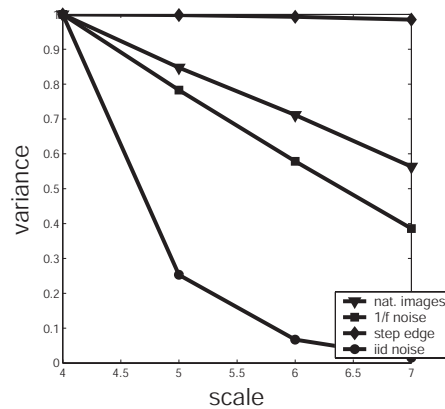


Fig. 4. The plot for the natural images from Fig. 3 is combined with the plot for the other images from Fig. 2. The rate of decay for the Laplacian magnitude of the natural images is similar to that of the intensity images.

edge, the $1/f$ noise, and the i.i.d. noise. Fig. 3 shows the result of the experiment. Now the behavior of images is very different. The Laplacian magnitude of a step edge is a line edge, and the Laplacian magnitude of $1/f$ noise is similar to i.i.d. noise neither of which have scale invariance. For images, the Laplacian magnitude appears scale invariant. Furthermore, if we start from a lower scale and include the data from Fig. 2, the same pattern emerges as in the first experiment (see Fig. 4). Thus, $I^{(2)}$ for natural images behaves like I meaning both (2) and (3) hold.

To describe the type of scale invariance exhibited by images, we define the following sequence

$$I^{(1)} = I(x/s) \quad (6)$$

$$I^{(n)} = \log |I^{(n-1)}(x/s) - I^{(n-1)}(x/(2s))|, \quad (7)$$

and assume that if (2) and (3) are true for $I^{(n)}$ they are true for $I^{(n+1)}$. While step edges and $1/f$ noise are described by the scale invariance of $I^{(1)}$, images are described by the scale invariance of $I^{(1)}, I^{(2)}, I^{(3)} \dots$. Just as the Laplacian pyramid is a representation that assumes $I^{(1)}$ is scale invariant, the higher order pyramid (described in the next section) is a representation that assumes $I^{(1)}, I^{(2)}, I^{(3)} \dots$ are scale invariant. The implication is that the higher order pyramid is tuned to the type of scale invariance of natural images and not that of other scale invariant images such as $1/f$ noise and the step edge.

III. HIGHER ORDER PYRAMIDS

We begin with a description of the Laplacian pyramid on which higher order pyramids are based. The Laplacian pyramid is motivated by the scale invariance of $I^{(1)}$ and only codes the information at each scale that is not predicted by the lower scale. The Laplacian pyramid constructed from an image I is represented by \mathcal{L} an ordered set defined by

$$\mathcal{L}(I) = \{L_1, L_2, \dots, L_s, G\}, \quad (8)$$

where $L_i = I(x/2^{i-1}) - I(x/2^i)$ are levels of the pyramid. L_1 is the finest scale, L_s is the coarsest scale and G is the lowpass residual. Using \mathcal{L}_i to refer to the i^{th} member of the ordered set \mathcal{L} , it is straightforward to reconstruct I by

$$I = \sum_{i=1}^{s+1} \mathcal{L}_i, \quad (9)$$

where upsampling and convolution are implicit (for details see [4]).

A higher order pyramid is motivated by the scale invariance of $I^{(1)}, I^{(2)}, I^{(3)} \dots$ which suggests also representing these images via a pyramid. These pyramids in turn predict the magnitude of each level of \mathcal{L} . The only information not predicted is the set of sign bits $L_i/|L_i|$. However, the sign operation introduces aliasing and is not rotationally invariant, which is why scale invariance does not appear until the lower scales (see the abscissa axis of Fig. 3). Furthermore, the sign operation does not lead to a compact representation because each magnitude image is the same

size as L_i . Rather than coding the sign bits, we code $L_i(x)/M_i(x/2)$ where $M_i = |L_i|$, which is a form of contrast normalization. The operation $L_i(x)/M_i(x/2)$ is accomplished by dividing a level of the Laplacian pyramid by its magnitude after reducing the resolution. Using the operations S^\downarrow and S^\uparrow to denote downsampling and upsampling by a factor of 2 the residual magnitude is

$$M_i = S^\downarrow(|L_i|), \quad (10)$$

and the prediction error is

$$C_i = L_i \cdot S^\uparrow(M_i). \quad (11)$$

Filtering before and after sampling is assumed in order to prevent aliasing and ensure translation invariance. Applying these operations to each level of \mathcal{L} results in \mathcal{C} , a contrast normalized Laplacian pyramid, defined by the ordered set

$$\mathcal{C}(I) = \{C_1, C_2, \dots, C_s, G\}, \quad (12)$$

and \mathcal{R} , the residual magnitude, defined by

$$\mathcal{R}(I) = \{M_1, M_2, \dots, M_s\}. \quad (13)$$

Given $\mathcal{R}(I)$ and $\mathcal{C}(I)$ it is easy to invert the contrast normalization and reconstruct \mathcal{L} . The higher order pyramid is built by recursively constructing \mathcal{C} on each of the residual magnitude images in \mathcal{R} . Thus, the higher order pyramid differs from other contrast normalized image representations [9] [20] by explicitly representing the magnitudes via additional pyramids. The higher order pyramid is recursively defined by the ordered set

$$\mathcal{H}^{(1)}(I) = \{\mathcal{C}(I), \mathcal{R}(I)\} \quad (14)$$

$$\mathcal{H}^{(n)}(I) = \{\mathcal{C}(I), \{\mathcal{H}^{(n-1)}(\log M_1), \dots, \mathcal{H}^{(n-1)}(\log M_{s-1}), M_s\}\}. \quad (15)$$

Because this is a doubly recursive representation there are two notions of depth: s the number of times the scale is reduced when building the pyramids and n the number of times pyramids are built on the images $I^{(1)}, I^{(2)}, I^{(3)} \dots$. We refer to n as the order of the pyramid and s as the depth. A first order pyramid, $\mathcal{H}^{(1)}$, is a contrast normalized Laplacian pyramid along with $I^{(2)}$ the residual magnitude images. A second order pyramid, $\mathcal{H}^{(2)}$, constructs a first order pyramid on each of the residual magnitudes and keeps the $I^{(3)}$ residuals. For higher orders this process is repeated. \mathcal{H} is composed of three types of data: C the contrast normalized images, G the lowpass residuals, and M the residual magnitudes.

Algorithm 1 buildHigherOrderPyramid

Input: I the image, n the order, s the depth
Output: \mathcal{H} the higher order pyramid
 $\mathcal{L} \leftarrow \text{buildLaplacianPyramid}(I, s)$
for $i = 1 \dots s$ **do**
 $L \leftarrow \mathcal{L}_i$
 $M \leftarrow S^\downarrow(|L|)$ {downsample the magnitude}
 $\mathcal{C}_i \leftarrow L/S^\uparrow(M)$ {upsample and contrast normalize}
 if $(n > 1) \wedge (i < s)$ **then**
 $\mathcal{R}_i \leftarrow \text{buildHigherOrderPyramid}(\log(M + 1), n - 1, s - i)$
 else
 $\mathcal{R}_i \leftarrow M$ {the residual magnitude image}
 end if
end for
 $\mathcal{C}_{s+1} \leftarrow \mathcal{L}_{s+1}$ {the lowpass residual }
 $\mathcal{H} \leftarrow \{\mathcal{C}, \mathcal{R}\}$

Algorithm 1 recursively builds a higher order pyramid of an image for a given order and depth. Note that the recursive operation is applied to the $\log(M + 1)$ rather than $\log(M)$ to avoid difficulties with values near zero. In addition, care should be taken to avoid division by zero in the contrast normalization step. Algorithm 2 inverts the pyramid by recursively collapsing all of the pyramids, inverting the log, and multiplying back in the magnitudes. Like the Laplacian pyramid, there is no loss of information when inverting the higher order pyramid.

Fig. 5 demonstrates a higher order pyramid built up to order 3. Each time the order is increased the residual magnitudes decrease in scale by a factor of 2. It is straightforward to show that in the limit the total number of coefficients is twice the number of image pixels. Hence, the representation is twice overcomplete. The levels of the pyramids in Fig. 5 are displayed such that below any given level are the pyramids built from the residual magnitude of that level.

IV. HIGHER ORDER WHITENING

In this section we show that the higher order pyramid performs a form of “higher order whitening” for images. A random vector is whitened when the coefficients are decorrelated and the variance is equalized in all directions. For stationary signals this is achieved by applying a

Algorithm 2 invertHigherOrderPyramid

Input: \mathcal{H} the higher order pyramid, n the order, s the depth

Output: I the image

$\mathcal{C} \leftarrow \mathcal{H}_1$

$\mathcal{R} \leftarrow \mathcal{H}_2$

for $i = 1 \dots s$ **do**

if $(n > 1) \wedge (i < s)$ **then**

$M \leftarrow \text{invertHigherOrderPyramid}(\mathcal{R}_i, n - 1, s - i)$

$M \leftarrow \exp(M) - 1;$

else

$M \leftarrow \mathcal{R}_i$ {the residual magnitude image}

end if

$\mathcal{C} \leftarrow \mathcal{C}_i$

$\mathcal{L}_i \leftarrow \mathcal{C} \cdot S^\dagger(M)$ {upsample and unnormalize}

end for

$\mathcal{L}_{s+1} \leftarrow \mathcal{C}_{s+1}$ {the lowpass residual}

$I \leftarrow \text{invertLaplacianPyramid}(\mathcal{L})$

filter whose frequency response is the inverse of the expected amplitude spectra of the random signal. Because images are scale invariant the amplitude spectra has a particular form. Let $\hat{I}(f)$ be the Fourier transform of an image $I(x)$, the amplitude spectra obey a power law,

$$E \left[|\hat{I}(f)| \right] \propto 1/f, \quad (16)$$

where E is the expectation operator taken over the ensemble of images [8]. In the frequency domain this implies that equal power is found in octave wide bands. In the spatial domain this implies that long range slowly decaying correlations exist. The expected correlations of an image can be removed by applying a linear operator w with a frequency response given by

$$|\hat{w}(f)| \propto f. \quad (17)$$

Then,

$$W(x) = w * I(x) \quad (18)$$

is a whitened image where $*$ is convolution with w , the whitening filter. Although there is no filter that satisfies Eq. (17), there are ways to approximate it. One approximation is to sample

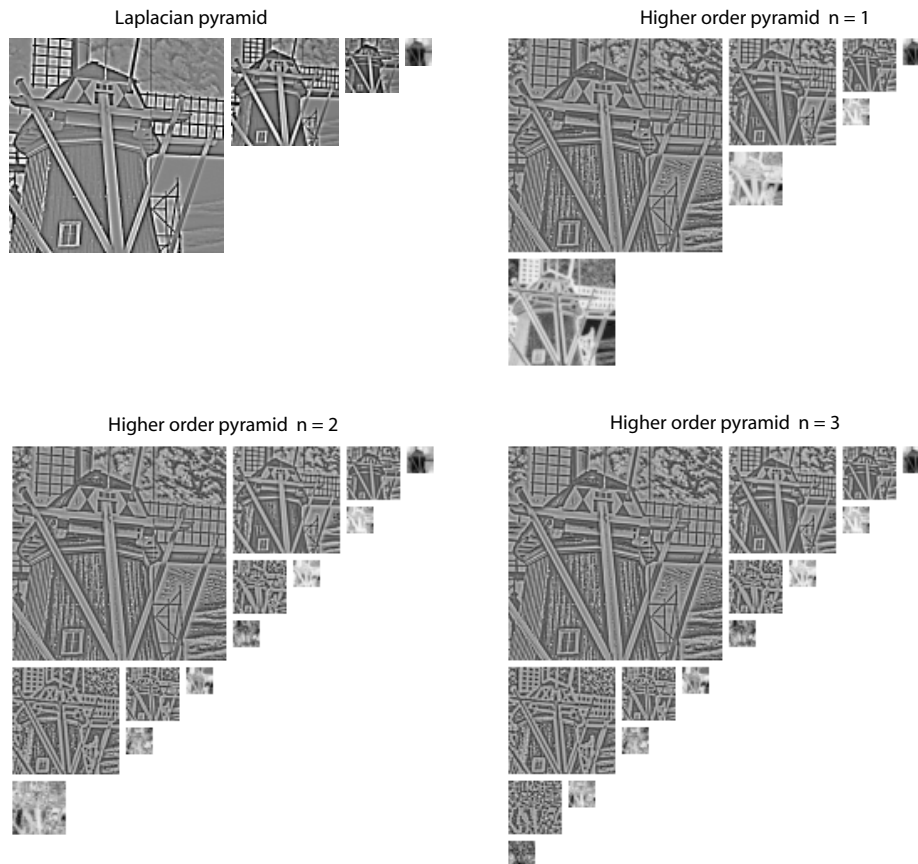


Fig. 5. A higher order pyramid is built from a Laplacian pyramid which in this example is three levels deep along with a lowpass residual. A first order pyramid, $n = 1$, is a contrast normalized pyramid with a residual magnitude shown below each level. A second order pyramid, $n = 2$, builds first order pyramids on the magnitudes that are larger than the lowpass residual. A third order pyramid repeats the process.

the function in the Fourier domain using frequency cutoffs and invert to construct w [27]. Atick and Redlich have shown that such a whitening filter corresponds well to the response properties of retinal ganglion cells [28]. Another way to approximate w is to properly weight the levels of a Laplacian pyramid. Assuming the levels of the pyramid are octave wide bandpass images, then according to Eq. (17)

$$W \approx L_1 + L_2/2 + L_3/4 + L_4/8 \dots \quad (19)$$

Thus, the scale invariant amplitude spectra of natural images indicates that the information in L_i should carry twice the weight as the information in L_{i+1} in order to whiten the data. However, for non-Gaussian signals such as images, whitening does not account for higher order statistical

dependencies. Recall that for natural images the sequence $I^{(1)}, I^{(2)}, I^{(3)} \dots$ is scale invariant which suggests that

$$E \left[|\hat{I}^{(n)}(f)| \right] \propto 1/f, n = 1, 2, \dots \quad (20)$$

Higher order whitening must weight the information in L_i twice that of L_{i+1} where the pair L_i and L_{i+1} comes from any of the $I^{(1)}, I^{(2)}, I^{(3)} \dots$

Next, we show that the higher order pyramid achieves this property when $n = s$, meaning that the order of the pyramid equals the depth. If $n = s$ the higher order pyramid is

$$\mathcal{H}^{(1)}(I) = \{\{C_1, G\}, \{M_1\}\} \quad (21)$$

$$\mathcal{H}^{(n)}(I) = \{\{C_1, \dots, C_n, G\}, \{\mathcal{H}^{(n-1)}(\log M_1), \dots, \mathcal{H}^{(1)}(\log M_{n-1}), M_n\}\}. \quad (22)$$

Let a_n be the total number of levels in an n^{th} order higher order pyramid not including the G 's and the M 's, the lowpass residuals and magnitude residuals. From Eq. (21-22) the number of levels is given by the following recurrence

$$a_0 = 1 \quad (23)$$

$$a_n = 2a_{n-1} + 1. \quad (24)$$

Consider an arbitrary pair of levels L_i and L_{i+1} from a Laplacian pyramid defined on any $I^{(n)}$. Let b_i and b_{i+1} be the number of levels in the higher order pyramid used to represent L_i and L_{i+1} . Each of L_i and L_{i+1} are represented in the higher order pyramid by C_i and C_{i+1} , the contrast normalized images, and $\mathcal{H}^{(n)}(\log M_i)$ and $\mathcal{H}^{(n-1)}(\log M_{i+1})$, the higher order pyramids built from the magnitude images. Then, the number of levels used to represent the Laplacian images is given by $b_i = 1 + a_n$ and $b_{i+1} = 1 + a_{n-1}$. From the recurrence in Eq. (24-24) we have,

$$b_i = 1 + a_n \quad (25)$$

$$= 1 + 2a_{n-1} + 1 \quad (26)$$

$$= 2(1 + a_{n-1}) \quad (27)$$

$$= 2b_{i+1}. \quad (28)$$

Thus, the number of levels used to represent L_i is twice the number used to represent L_{i+1} . Furthermore, this holds for any L_i and L_{i+1} defined on any of the images $I^{(1)}, I^{(2)}, I^{(3)} \dots$

which shows that the higher order pyramid is a form of higher order whitening for images. It is important to note that a_n refers to the number of levels not the number of coefficients. Thus, each level must be weighted the same regardless of the number of coefficients.

V. EXPERIMENTS

Four experiments are performed. In order to ensure the image transformation is stable, the first experiment examines the effect of quantization and truncation of the coefficients of the higher order pyramid. The second experiment measures the degree to which pairs of coefficients are independent. The reason for this test is that the efficiency of a representation is reflected by the lack of redundancy or independence in the coefficients of the representation. The third and fourth experiments are recognition tests. A good representation should lead to improved matching. The rationale is that a less redundant representation is a better space in which to perform matching in the presence of uncertainty. In our experiments three types of uncertainty are considered: object recognition with viewpoint changes, object recognition with scale changes, and face recognition with illumination changes.

In all experiments the higher order pyramids are implemented using the algorithms described in Section III. The upsampling and downsampling is performed using 7-tap binomial filters, and image borders are handled via reflection. Higher order pyramids are compared to Laplacian pyramids and the image pixels which are two commonly used image representations. The Laplacian pyramids are implemented using the same binomial filters.

A. Truncation and quantization

For any order, the pyramid can be inverted without the residual magnitudes which is referred to as a truncated pyramid. This is done by replacing each of the residual magnitude images M with its mean $E[M]$ and inverting the pyramid. Although the result is a loss of information, the visual quality of the reconstructed images rapidly improves as n increases. Fig. 6 demonstrates examples of reconstructed images from truncated pyramids. For $n > 2$ it is very difficult to tell the difference between the original and reconstructed images.

The higher order pyramid is also stable with respect to quantization in that the root mean square error (RMSE) of the reconstructed images is small even when the coefficients are severely quantized. The effect of truncation and quantization is measured by computing the RMSE for the

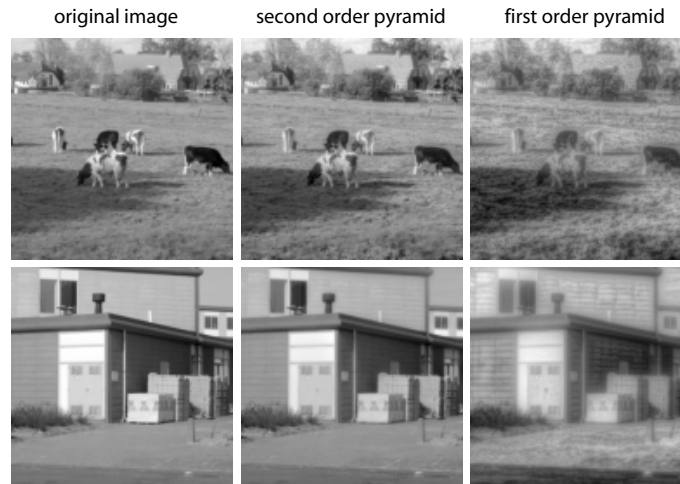


Fig. 6. The images are examples of inverting the higher order pyramid after truncating (removing the residual magnitudes) and quantizing the coefficients. The residual magnitude becomes significantly less important as the order increases. Furthermore, the transformation is stable with respect to quantization of the coefficients.

RMSE of reconstructed images				
	$q = 2$	$q = 4$	$q = 32$	$q = 128$
Truncated pyramid, $n=1$	17.4	14.3.	14.1	14.1
Truncated pyramid, $n=2$	11.1	4.2	3.4	3.3
Truncated pyramid, $n=3$	10.6	2.7	1.2	0.8
Truncated pyramid, $n=4$	10.6	2.6	1.0	0.5
Truncated pyramid, $n=5$	10.6	2.6	1.0	0.5

Fig. 7. The table shows the root mean square error (RMSE) estimated from a set of images after inverting an n^{th} order truncated pyramid quantized with q levels.

set of 40 images described in Section II. For each image the higher order pyramid is built, the coefficients are quantized and the residual magnitude is truncated. Then, the pyramid is inverted and the RMSE between the reconstructed images and the original images is estimated. An image resolution of 512×512 pixels is used and $s = 5$ so that the lowpass residual images are 16×16 pixels. The coefficients of the higher order pyramid are quantized to q levels using uniform quantization. The table in Fig. 7 reports the RMSE as a function of truncation and quantization. The input images are quantized between 0–255, thus only about 7 bits are needed per coefficient

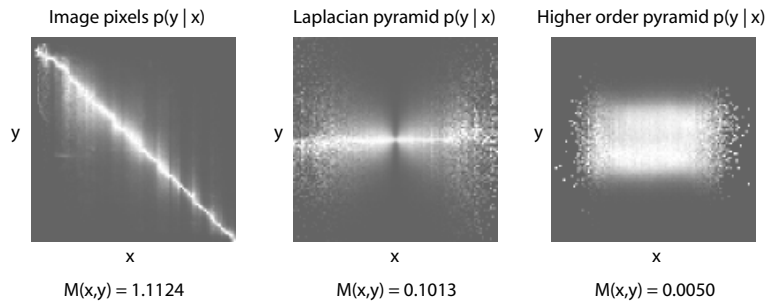


Fig. 8. Conditional distributions for x and y , pairs of coefficients at the same spatial offset. In the pixel domain the coefficients are correlated. In the Laplacian pyramid they are decorrelated but still dependent because the variance of y scales with x . In the higher order pyramid they are independent. Below each graph is M the measured mutual information between the coefficients for each representation.

in the pyramid to stay within the quantization error. These results show that the transformation is stable with respect to quantization and truncation. This suggests the representation could be used for image processing tasks such as compression and denoising.

B. Redundancy

To test for redundancy the joint distribution for a pair of coefficients is measured from the set of 40 natural images. For each image, statistics are gathered for all pairs of coefficients displaced by 5 pixel positions in the diagonal direction. This is done for the raw image pixels, for a level in the Laplacian pyramid, and for a level in the higher order pyramid. A distance of 5 is chosen to ensure the coefficients are not within the support of the convolution filter.

The joint distributions are displayed as conditional distributions in Fig. 8. The pairs of coefficients are denoted by x and y where intensity in the figure represents the likelihood of y conditioned on the value of x . The Laplacian pyramid has the familiar bowtie shape documented by Simoncelli [11]. This is indicative of the fact that the magnitudes are correlated. Meanwhile, the coefficients of the higher order pyramid are clearly independent because the likelihood of x is the same for all values of y . This is also reflected by the mutual information shown below each plot. Similar results are found for different spatial offsets both within and between different levels of the higher order pyramid. Independence has also been demonstrated after applying contrast normalization [20]. However, because the higher order pyramid is invertible we can be sure that the independence does not arise from destroying information. It is important to note

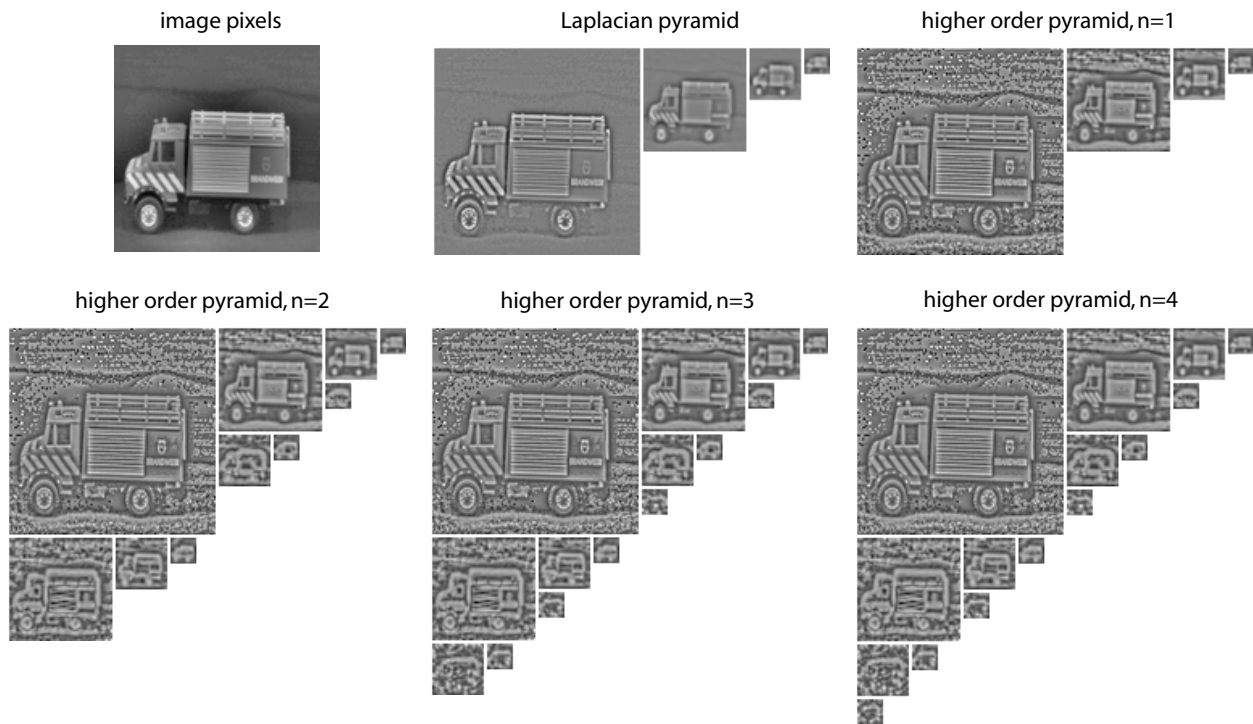


Fig. 10. The following representations are compared: higher order pyramids ($n = 1, \dots, 4$), the Laplacian pyramid and the image pixels. An example of each representation for the truck image is shown.

For the recognition experiments, the training set consists of one image of each of the 1000 objects. A test image is compared to each image in the training set and recognition is determined by the nearest neighbor. The training and test images are compared by building a higher order pyramid for each, and computing the distance

$$\sum_i \frac{|\mathbf{x}_i - \mathbf{y}_i|^2}{N_i}, \quad (29)$$

where \mathbf{x}_i and \mathbf{y}_i are vectors of the coefficients at corresponding levels, N_i is the number of coefficients in the level, and the summation is performed over all levels except the residuals, the G 's and M 's. This is just the L_2 norm defined on the coefficients where each level is weighted the same. Recall that to achieve higher order whitening each level must be weighted the same. In the experiments, the order of the pyramid is varied from $n = 1$ to $n = 4$ but the depth is set to $s = 4$ so that higher order whitening corresponds to $n = s = 4$. For comparison each object is also represented by a Laplacian pyramid and the image pixels. Like the higher order pyramid, the Laplacian pyramid is built with $s = 4$ and the coefficients are weighted by the size of the

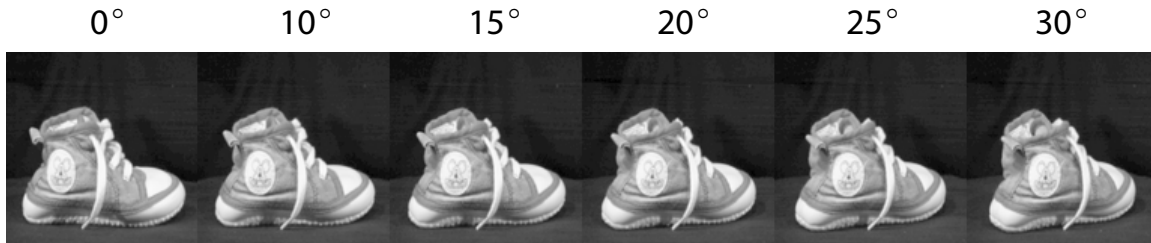


Fig. 11. Examples of test images with pose changes between 10 – 30 degrees. Recognition is performed using nearest neighbor matching between the test image and the training set of 1000 object images.

Object recognition error rate (%) vs. pose change (degrees)					
	10°	15°	20°	25°	30°
Image pixels	2.8	9.0	19.2	30.8	40.7
Laplacian pyramid	0.2	2.8	6.9	15.5	23.9
Higher order pyramid, n=1	0.4	1.6	4.5	10.1	16.2
Higher order pyramid, n=2	0.3	0.8	2.3	5.1	9.7
Higher order pyramid, n=3	0.0	0.6	1.4	3.8	8.2
Higher order pyramid, n=4	0.0	0.5	1.3	4.1	8.0

Fig. 12. Object recognition error rates for each representation as a function of pose change.

level. Laplacian pyramids are compared in the same way as in Eq. (29), except the L_1 norm is used. In the case of the Laplacian pyramid the L_1 norm is preferred because the distribution of coefficient values is sparse. For the image pixels representation the lowpass data is removed, which is $I(x) - I(x/2^{s+1})$ for $s = 4$. Comparisons are made using Eq. (29), which in this case is just the sum of squared distances. Fig. 10 shows an example of each representation for one of the object images. For efficiency the representations are pre-computed for each training image and stored as a single vector of coefficients. For the pyramids the coefficients are pre-weighted by the size of the level. Because these images do not exhibit scale invariance at small resolutions, the data at or below a resolution of 8×8 pixels is never used for any of the representations.

In the first experiment, the test images consist of pose changes in range 10 – 30 degrees. Examples of these pose changes are in Fig. 11 for one of the objects. Each pose change is

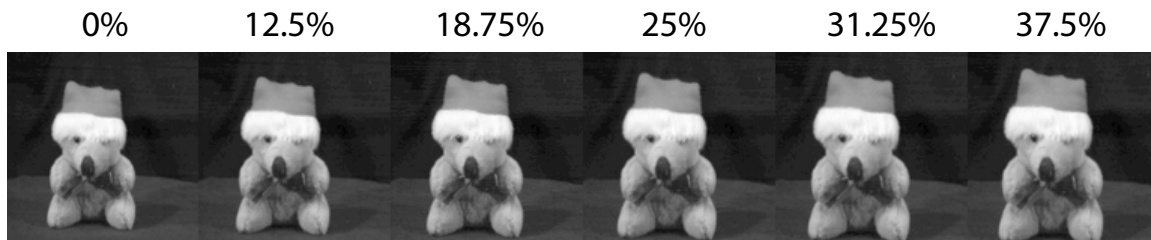


Fig. 13. Examples of test images with scale changes between 12.5 – 37.5%.

Object recognition error rate (%) vs. scale change (percent increase)					
	12.5%	18.75%	25.00%	31.25%	37.5%
Image pixels	11.0	14.3	37.7	58.8	73.5
Laplacian pyramid	0.1	4.7	18.5	36.5	54.6
Higher order pyramid, $n=1$	0.2	9.5	30.3	51.7	68.8
Higher order pyramid, $n=2$	0.0	0.6	7.7	27.0	50.6
Higher order pyramid, $n=3$	0.0	0.0	1.7	11.5	30.0
Higher order pyramid, $n=4$	0.0	0.0	1.1	8.2	24.3

Fig. 14. Object recognition error rates for each representation as a function of scale change.

tested for all 1000 objects. The error rates are shown in Fig. 12 as a function of pose. The higher order pyramid with $n = 1$ is representative of contrast normalization. The superior performance for higher n demonstrates the importance of representing the magnitude.

In the second experiment, the test images consist of scale changes. Examples are shown in Fig. 13. Again, each scale change is tested for all 1000 objects, and the error rates are reported in Fig. 14. As before, the higher order pyramids with $n \geq 2$ demonstrate considerable improvements in recognition rates. It is important to note that no searching over scale or position is performed.

D. Face recognition with illumination changes

Using the Yale illumination face database and experimental setup reported in [30], we test the ability of higher order pyramids to perform face recognition in the presence of illumination changes. The database contains 10 faces each under 64 different illuminants. They are grouped into 5 subsets according to the severity of the direction of illumination. Examples are shown in

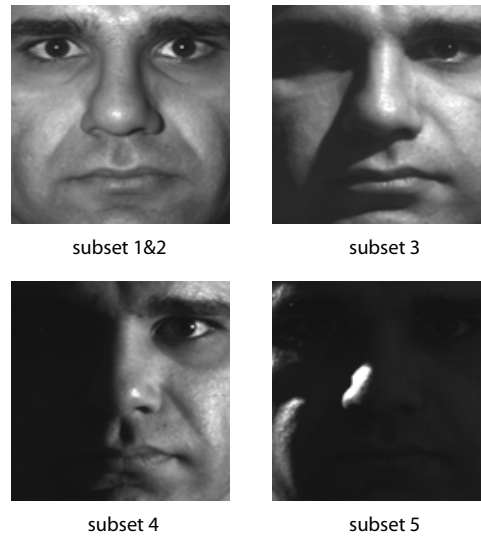


Fig. 15. Example test images with illumination changes. The images are grouped according to the severity of the illuminant direction.

Fig. 15. Each face is cropped and resampled to a resolution of 128×128 pixels and converted to log intensities.

The same method for recognition as discussed in the previous section is used. The training set consists of one frontal illumination face image per person. Each of the remaining 630 test images is matched to each image in the training set and the nearest neighbor is selected. Again, for matching, the L_2 norm is computed for the image pixels and the higher order pyramids, and the L_1 norm is used for the Laplacian pyramid. Fig. 16 shows an example of each representation computed for one of the face images. As before, the pyramids are constructed with $s = 4$ and treated as a single vector of weighted coefficients so that each level contributes equally.

The results are reported in Fig. 17 and demonstrate a clear advantage for the higher order pyramids as compared to the Laplacian pyramid and the image pixels. The superior performance for $n \geq 2$ is further evidence that it is important to represent the magnitude, in a less redundant space, rather than just applying contrast normalization.

For further comparison the results from Chen et al. are included in Fig. 18 [30]. It should be pointed out that all of the other methods, with the exception of Gradient Angle, use all of the images in Subset 1 and 2 for training. Furthermore, these methods are specifically designed to deal with variation in illumination. Subset 5, the one with the most severe illumination directions,

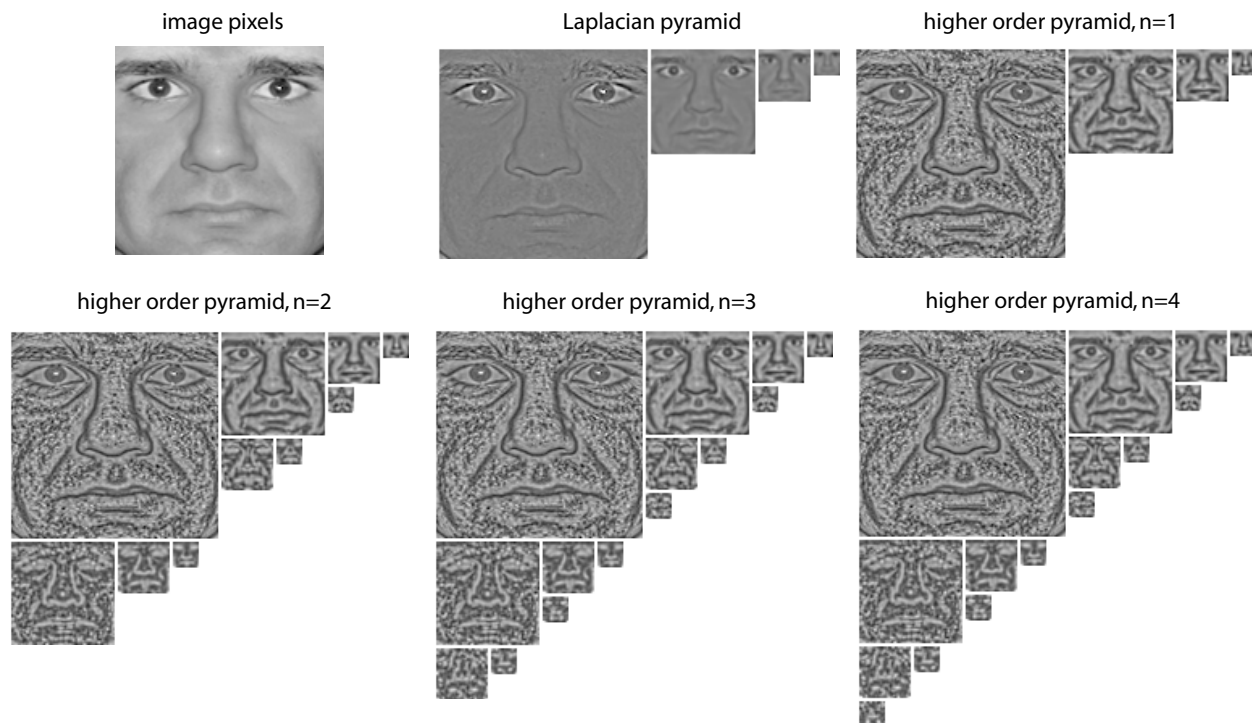


Fig. 16. An example of each representation for one of the face images.

is not used in any of the other experiments. The Gradient Angle method learns, from a database of images, an illumination insensitive weighting of the magnitude of the gradient. Because the magnitude is sensitive to illumination the Gradient Angle method is a form of contrast normalization thus bears a resemblance to the higher order pyramids.

VI. DISCUSSION

The scale invariant properties of natural images leads to the higher order pyramid as an early visual representation, which is simple to compute and straightforward to invert. The representation is stable with respect to quantization of the coefficients and truncation of the residual magnitudes. Images coded in the space of higher order pyramids exhibit far less redundancy than the raw image pixels and decorrelating transformations such as the Laplacian pyramid. This is demonstrated by showing the independence between pairs of coefficients and improved matching in several recognition experiments. The recognition experiments are only intended to demonstrate the potential for redundancy reduction as a goal for the representation of visual

Face recognition error rate (%) vs. Illumination				
	Subset 1&2	Subset 3	Subset 4	Subset 5
Image pixels	0.0	2.5	12.1	19.5
Laplacian pyramid	0.0	0.0	7.9	7.4
Higher order pyramid, n=1	0.0	0.0	7.9	12.6
Higher order pyramid, n=2	0.0	0.0	0.0	0.5
Higher order pyramid, n=3	0.0	0.0	0.7	0.5
Higher order pyramid, n=4	0.0	0.0	1.4	1.0

Fig. 17. Face recognition error rates as a function of the illumination directions.

Face recognition error rate (%) vs. Illumination				
Method	Subset 1&2	Subset 3	Subset 4	Subset 5
Eigenfaces	0.0	16.7	69.3	–
Linear subspace	0.0	1.7	12.9	–
Cones-attached	0.0	0.8	9.3	–
Gradient Angle	0.0	0.0	1.4	–
Cones-cast	0.0	0.0	0.0	–
Higher order pyramid, n=2	0.0	0.0	0.0	0.5

Fig. 18. For comparison, face recognition error rates reported in [30].

patterns and the higher order pyramid as a step in this direction. The real benefit is the further processing of the higher order pyramid and the incorporation of learning techniques.

It is interesting to note that the representation is not oriented. Because a Laplacian pyramid is used to construct the higher order pyramid, the basis functions are all circularly symmetric Gaussian functions albeit combined in a non-linear way. This might seem strange given the wide spread use of oriented filters such as Gabor functions and steerable derivatives throughout computer vision. In addition, oriented receptive fields are commonplace in computational models

of human vision. Furthermore, it is well known that the optimal linear basis in which to represent natural images is made up of oriented basis functions [8] [16] [15]. If we assume that scale invariance as defined in this paper is a natural description of images, then we must question the use of oriented basis functions. We briefly speculate on some possible answers: (1) The model proposed here does not fully capture the scale invariance of images; (2) Other statistical properties of images that are not implied by scale invariance are more important; (3) Oriented basis functions are simply the best way to represent scale invariance when restricted to a linear framework. However, in a non-linear framework they may no longer be needed.

ACKNOWLEDGMENT

Parts of this paper first appeared in the European Conference on Computer Vision 2006. This work was supported by a NSF ITR award no. IIS-0219078.

REFERENCES

- [1] F. Attneave, "Some informational aspects of visual perception," *Psych. Rev.*, vol. 61, pp. 183–193, 1954.
- [2] H. Barlow, *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, pp. 217–234, MIT Press, 1961.
- [3] S.C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 691–712, June 2003.
- [4] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. on Communications.*, vol. 31, no. 4, pp. 532–540, 1983.
- [5] E. Simoncelli and W. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Int'l Conf. on Image Processing*, 1995.
- [6] S.G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1989.
- [7] J. Koenderink, "The structure of images," *Biological Cybernetics*, vol. 50, pp. 363–370, 1984.
- [8] D.J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. of the Opt. Soc. of Am. A*, vol. 4, no. 12, pp. 2379–, 1987.
- [9] D. Ruderman, "The statistics of natural images," *Network*, vol. 5, no. 4, pp. 477–500, 1993.
- [10] S.C. Zhu and D. Mumford, "Prior learning and gibbs reactionn-diffusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 1236–, 1997.
- [11] E. Simoncelli and E. Adelson, "Noise removal via bayesian wavelet coding," in *Int'l Conf. on Image Processing*, 1996, pp. 379–382.
- [12] B. Wegmann and C. Zetsche, "Statistical dependence between orientation filter outputs used in a human vision based image code," *Proc. of Visual Comm. and Image Processing*, vol. 1360, 1990.
- [13] A. Lee and D. Mumford, "Occlusion models for natural images: a statistical study of scale invariant dead leaves model," *International Journal of Computer Vision*, vol. 41, no. (1,2), 2001.

- [14] D.L. Ruderman, “The origins of scaling in natural images,” *Vision Research*, vol. 37, 1997.
- [15] A.J. Bell and T.J. Sejnowski, “The independent components of natural scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–38, 1997.
- [16] B.A. Olshausen and D.J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, pp. 607–609, 1996.
- [17] Y. Karklin and M. Lewicki, “Learning higher-order structures in natural images,” *Network: Comp. in Neural Syst.*, vol. 14, pp. 483–499, 2003.
- [18] P. Hoyer and A. Hyvarinen, “A multi-layer sparse coding network learns contour coding from natural images,” *Vision Research*, vol. 42, no. 12, pp. 1593–1605, 2002.
- [19] J. Portilla, V. Strela, M.J. Wainwright, and E.P. Simoncelli, “Image denoising using gaussian scale mixtures in the wavelet domain,” in *IEEE Trans. on Image Processing*, 2003.
- [20] O. Schwartz and E. Simoncelli, “Natural signal statistics and sensory gain control,” *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, August 2001.
- [21] M. Wainwright and E. Simoncelli, “Scale mixtures of gaussians and the statistics of natural images,” in *Advances in Neural Information Processing Systems*, 2000.
- [22] C. Guo, S. Zhu, and Y. Wu, “Towards a mathematical theory of primal sketch and sketchability,” in *Proceedings of the Ninth Int’l Conf. on Computer Vision*, 2003.
- [23] J. Gluckman, “Higher order image pyramids,” in *Proc. of the European Conference on Computer Vision*, 2006.
- [24] J. Gluckman, “Higher order image whitening,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.
- [25] J.A. van Hateren, “Independent component filters of natural images compared with simple cells,” *Proc. of the R. Stat. Soc. of London B*, vol. 265, pp. 359–366, 1998.
- [26] R.W. Buccigrossi and E.P. Simoncelli, “Image compression via joint statistical characterization in the wavelet domain,” *IEEE Trans. on Image Processing*, vol. 8, no. 12, pp. 1688–, 1999.
- [27] B. Olshausen and D. Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1?,” *Vision Research*, vol. 37, pp. 3311–3325, 1997.
- [28] J.J. Atick and A.N. Redlich, “What does the retina know about natural scenes?,” *Neural Computation*, vol. 4, pp. 196–210, 1992.
- [29] J.M. Geusebroek, G.J. Burghouts, and A.W.M. Smeulders, “The amsterdam library of object images,” *Int’l J. Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.
- [30] H. Chen, P. Belhumeur, and D. Jacobs, “In search of illumination invariants,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.