

## Statistics for Optimal Point Prediction in Natural Images

Wilson S. Geisler<sup>1</sup> and Jeffrey S. Perry<sup>1</sup>

<sup>1</sup>Center for Perceptual Systems  
1 University Station  
University of Texas at Austin  
Austin TX, 78712

## Abstract

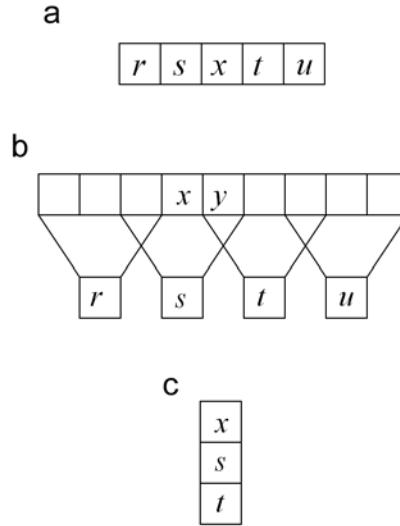
Sensory systems exploit the statistical regularities of natural signals, and thus a fundamental goal for understanding biological sensory systems, and creating artificial sensory systems, is to characterize the statistical structure of natural signals. Here we use a simple conditional-moments method to measure natural image statistics relevant for three fundamental visual tasks: (i) estimation of missing or occluded image points, (ii) estimation of a high resolution image from a low resolution image ('super resolution'), and (iii) estimation of a missing color channel. We use the conditional moments approach because it makes minimal invariance assumptions, can be applied to arbitrarily large sets of training data, and provides (given sufficient training data) the Bayes optimal estimators. The measurements reveal complex but systematic statistical regularities that can be exploited to substantially improve performance in the three tasks over what is possible with some standard image processing methods. Thus, it is likely that these statistics are exploited by the human visual system.

## Introduction

Natural images, like most other natural signals, are highly heterogeneous and variable; yet despite this variability, they contain a large number of statistical regularities. Visual systems have evolved to exploit these regularities so that the eye can accurately encode retinal images and the brain can correctly interpret them. Thus, characterizing the structure of natural images is critical for understanding visual encoding and decoding in biological vision systems and for applications in image processing and computer vision.

Here, we use a direct conditional moments approach to measure the joint statistics relevant for the three simple visual tasks illustrated in Figure 1. The first task is estimation of missing image points (Figure 1a). The symbols  $r$ ,  $s$ ,  $t$  and  $u$  represent observed values along a row of image pixels and  $x$  represents an unobserved value to be estimated. This kind of task arises when some image locations are occluded or a sensor element is missing. A different variant of the task arises when only some of the color values are absent at a location (e.g., the demosaicing task, Brainard et al. 2008; Li, Gunturk & Zhang 2008, and related tasks Zhang & Brainard 2004). The second task is estimation of a high resolution image from a low resolution image (Figure 1b). Again the symbols  $r$ ,  $s$ ,  $t$  and  $u$  represent observed low spatial resolution values obtained by locally averaging and then down sampling a higher resolution image;  $x$  and  $y$  represent unobserved values to be estimated. This kind of task arises naturally in interpreting (decoding) retinal responses in the periphery. For example, Figure 1b corresponds approximately to the situation around 1 degree eccentricity in the human retina where the sampling by midget (P) ganglion cells is about one half that in the fovea (one fourth the samples per unit area). In the computer vision literature the goal of this task is referred to as a 'super resolution' (Freeman, Jones & Pasztor 2002; Li & Adelson 2008; Glasner et al. 2009). The third task is estimation of a missing color channel given the other two. The symbols  $s$  and  $t$  represent the observed color values (e.g.,  $G$  and  $B$ ) at a pixel location, and  $x$  represents the unobserved value to be estimated (e.g.,  $R$ ). This is not a natural task, but it reveals the redundancy of color in natural scenes and sets an upper limit on how well a dichromat could estimate his missing color channel without knowledge of the spatial correlations in images or the objects in the scene. Furthermore, statistics measured for this task could be of use in other tasks (e.g., demosaicing). We find that

the image statistics for these tasks are complex but regular (smooth), are useful for performing the tasks, and make many testable predictions for visual coding and visual performance.



**Figure 1.** Three estimation tasks. **a.** Estimation of missing image points. **b.** Estimation of a high resolution image from a low resolution image. **c.** Estimation of a missing color channel. Letters x and y represent values to be estimated; other letters represent observed values.

The most studied regularities in natural images involve single and pair-wise statistics. For example, the intensities of randomly sampled points from a natural image have a probability distribution that is approximately Gaussian on a logarithmic intensity axis (Laughlin 1981; Ruderman et al. 1998), and the intensities of randomly sampled pairs of points have a covariance that declines smoothly as a function of the spatial separation between the points, in a fashion consistent with a Fourier amplitude spectrum that falls inversely with spatial frequency (Burton & Moorehead 1987; Field 1987; Ruderman & Bialek 1994). Many other statistical regularities have been discovered by measuring pair-wise statistics of image properties such as local orientation, color, spatial frequency, contrast, phase, and direction of motion (for reviews see Simoncelli & Olshausen 2001; Geisler 2008). Much can be learned by measuring such pair-wise statistics. For example, if the joint distribution of image features is Gaussian, then measuring the mean value of each feature and the covariance (a form of pair-wise statistic) of the values for all pairings of features is sufficient to completely specify the joint distribution.

However, the joint distributions of many image features are not Gaussian, and may not be completely described with pair-wise statistics (Field 1987; Daugman 1989; Buccigrossi & Simoncelli, 1999; Zhu and Mumford 1997), or it may be intractable to determine if they can be described by pair-wise statistics. A useful strategy for characterizing statistical structure of natural images is to fit image data with some general class of generative model such as linear models that assume statistically independent but non-Gaussian sources (Olshausen & Field 1997; Bell & Sejnowski 1997; van Hateren & van der Schaff 1998; Karklin & Lewicki 2008), or random Markov field models with filter kernels similar to those in the mammalian visual system (e.g., Portilla & Simoncelli 2000; for reviews see Simoncelli & Olshausen 2001, Zhu et al. 2009). Alternatively, more complex joint distributions could be approximated with Gaussian

mixtures (e.g., Maison & Vandendorpe 1998). While these approaches have yielded much insight into the structure of natural images and neural encoding, they involve assumptions that may miss higher dimensional statistical structure (Lee, Pedersen & Mumford, 2003).

In addition, weaker forms of assumed structure are sometimes implicit in normalization steps performed on each signal prior to the learning of their statistical structure. These normalization steps typically involve applying one of the following simple transformations to each training signal  $\mathbf{z}$  (e.g., image patch): subtract the mean of the signal ( $\mathbf{z}' = \mathbf{z} - \bar{\mathbf{z}}$ ), subtract the mean and then divide by the mean ( $\mathbf{z}' = (\mathbf{z} - \bar{\mathbf{z}}) / \|\mathbf{z} - \bar{\mathbf{z}}\|$ ), subtract the mean and scale to a vector length of 1.0 ( $\mathbf{z}' = (\mathbf{z} - \bar{\mathbf{z}}) / \|\mathbf{z} - \bar{\mathbf{z}}\|$ ), or subtract the mean and then multiply by a “whitening” matrix ( $\mathbf{z}'^T = \mathbf{W}(\mathbf{z} - \bar{\mathbf{z}})^T$ ). Usually, these normalization steps are applied to simplify the mathematics or reduce the dimensionality of the estimation problem. The assumption is that the normalization preserves the important statistical structure. For example, it is plausible that the statistical structure of image patches is largely independent of their mean luminance, if they are represented as contrast signals by subtracting and dividing by the mean. Similarly, standard image processing algorithms for some of the point prediction tasks considered here implicitly assume that subtracting off the mean luminance of the observed signal preserves the relevant statistical structure. Thus, in both cases the assumption is effectively that  $p(\mathbf{z}) = p(\mathbf{z}') p(\bar{\mathbf{z}})$ .

The most obvious strategy for characterizing statistical structure without making parametric assumptions is to directly estimate probability distributions by histogram binning. This strategy has proved useful for small numbers of dimensions (e.g., Geisler, Perry & Super 2001; Petrov & Zhaoping 2003; Thacik et al. 2010). A more sophisticated strategy uses clustering techniques to search for the subspace (manifold) where natural image signals are most concentrated (Lee et al. 2003). While this strategy has provided valuable insight (e.g., the manifold is concentrated around edge-like features), the subspaces found are still relatively high dimensional and not precisely characterized.

An alternative strategy emphasized here is based on directly measuring moments along single dimensions, conditional on the values along other dimensions. Measuring conditional moments is a well-known approach for characterizing probability distributions (e.g., see statistics textbooks such as Fisz 1967), and like directly estimating probability distributions, it is only practical for modest numbers of dimensions because of the amount of data required. Nonetheless, it has not been exploited for the current tasks and it has some unique advantages. First, univariate conditional distributions for local image properties are frequently unimodal and simple in shape, and thus the first few moments capture much of the shape information. Second, estimating conditional moments only requires keeping a single running sum for each moment, making it practical to use essentially arbitrarily large numbers of training signals and hence to measure higher dimensional statistics with higher precision. Third, it is relatively straightforward to specify Bayes optimal estimators from conditional moments. Of central relevance here, the conditional first moment (conditional mean) is the Bayes optimal estimator when the cost function is the mean square error (i.e., the so-called MMSE estimator). Fourth, conditional moments can be combined allowing the approach to be extended to higher numbers of dimensions than would otherwise be practical.

The proposed conditional moment method is related to other non-parametric methods such as example-based methods of texture synthesis, where individual pixels are assigned the median (or mean) value of the conditional probability distribution estimated for the given surrounding context pixels (De Bonet 1997; Efros & Leung 1999).

## Methods

### Training and test stimuli

The image set consisted of 1049 images of outdoor scenes collected in the Austin area with a calibrated Nikon D700 camera (see Ing et al. 2010 for the calibration procedure). The images contained no human made objects and the exposure was carefully set for each image so as to minimize pixel response saturation and thresholding (clipping at high and low intensities). Each 4284 x 2844 14-bit raw Bayer image was interpolated to 14 bit RGB using the AHD method (Hirakawa & Parks, 2005). The RGB values were then scaled using the multipliers

$m_R = 2.214791$ ,  $m_G = 1.0$ ,  $m_B = 1.193155$  (which are determined by the camera make and model) to obtain a 16 bit RGB image. In the case of gray scale, YPbPr color space conversion was used ( $\text{Gray} = 0.299R + 0.587G + 0.114B$ ). The dynamic range was increased by blacking out 2% of the darkest pixels and whiting out 1% of the brightest pixels. Blacking out pixels almost never occurred because in most images 2% of the pixels were already black. On the other hand, whiting out 1% of the brightest pixels has an effect because the image histograms have such long tails on the high end. For some measured statistics the images were then converted to L, M, S cone space. Finally, the images were converted to linear 8-bit gray scale or 24-bit color. The images were randomly shuffled and then divided into two groups: 700 training images that were used to generate the tables, and 349 test images. All pixels in the training images were used to estimate conditional moments, and thus the number of training samples was on the order of  $10^{10}$ .

### Conditional moments

The conditional moments approach is very straight forward. From the definition of conditional probability, an arbitrary joint probability density function over  $n$  dimensions can be written as a product of single dimensional conditional probability density functions,

$$p(\mathbf{x}_n) = \prod_{i=1}^n p(x_i | \mathbf{x}_{i-1})$$

where  $\mathbf{x}_{i-1} = (x_1, \dots, x_{i-1})$  and  $\mathbf{x}_0 = \emptyset$ . Furthermore, the moments of a single dimensional probability distribution function defined over a finite interval (as in the case of image pixel values) are usually sufficient to uniquely specify the density function (e.g., Fisz 1967). Thus, the full joint probability density function can be characterized by measuring the single dimensional conditional moments,

$$E(x_i^k | \mathbf{x}_{i-1}) = \sum_{x_i} x_i^k p(x_i | \mathbf{x}_{i-1})$$

where  $k$  is the moment (e.g.,  $k=1$  gives the mean of the posterior probability distribution of  $x_i$  given  $\mathbf{x}_{i-1}$ ). As is standard, these moments can be estimated simply by summing the observed values of  $x_i^k$  for each unique (or quantized) value of the vector  $\mathbf{x}_{i-1}$  and then dividing by the number of times the vector  $\mathbf{x}_{i-1}$  was observed,

$$E(x_i^k | \mathbf{x}_{i-1}) \equiv \frac{\sum_{x_i \in \Omega(\mathbf{x}_{i-1})} x_i^k}{N(\mathbf{x}_{i-1})} \quad (1)$$

where  $\Omega(\mathbf{x}_{i-1})$  is the set of observed values of  $x_i$  for the given value of the vector  $\mathbf{x}_{i-1}$ , and  $N(\mathbf{x}_{i-1})$  is the number of those observed values.

We define the *MMSE estimation function* for  $x_i$  to be

$$f(\mathbf{x}_{i-1}) = E(x_i | \mathbf{x}_{i-1}) \quad (2)$$

which gives the Bayes optimal estimate of  $x_i$  when the goal is to minimize the mean squared error of the estimate from the true value (e.g., Bishop 2009). The reliability of these estimates is given by

$$\rho(\mathbf{x}_{i-1}) = 1/VAR(x_i | \mathbf{x}_{i-1}) \quad (3)$$

which is obtained from the first and second moments:  $VAR(x_i | \mathbf{x}_{i-1}) = E(x_i^2 | \mathbf{x}_{i-1}) - E(x_i | \mathbf{x}_{i-1})^2$ . The focus of most of this paper is on measuring and applying  $f(\mathbf{x}_{i-1})$  and  $\rho(\mathbf{x}_{i-1})$ .

Given the number of training images, we were limited to directly measuring conditional moments for probability distributions with five or fewer dimensions. However, it is possible to obtain estimates based on even higher numbers of dimensions by combining estimates from different sets of conditional variables. For example, let  $\hat{x}_1$  be the estimate of  $x$  given one set of conditional variables, let  $\hat{x}_2$  be the estimate of  $x$  given a different set of conditional variables, and let  $\hat{\rho}_1$  and  $\hat{\rho}_2$  be their estimated reliabilities. If the estimates are Gaussian distributed and uncorrelated, then the optimal estimate given the combined conditional variables is

$$\hat{x} = \hat{x}_1\gamma + \hat{x}_2(1-\gamma) \quad (4)$$

where  $\gamma = \hat{\rho}_1 / (\hat{\rho}_1 + \hat{\rho}_2)$  is the relative reliability of the two estimates (e.g., Yuille & Bultoff 1996; Oruc et al. 2003). The more general combination rules for the case where the estimates are correlated are known (e.g., see Oruc et al. 2003); however, there are not enough training

images to estimate these correlations. Nonetheless, applying equation (4) can increase the accuracy of the estimates.

A more general way to combine the estimates that takes into account relative reliability, and potentially other statistical structure, is to directly measure the MMSE function

$$g(\hat{x}_1, \hat{x}_2, \gamma) = E(x|\hat{x}_1, \hat{x}_2, \gamma) \quad (5)$$

If sufficient training data are available this function is guaranteed to perform at least as well as equation (4).

It is important to note that none of these methods for combining the estimates are guaranteed to capture all of the additional statistical structure contained in the conditional moments of the combined variables. Nonetheless, they do make it practical to extend the conditional moments approach beyond five dimensions.

## Results

### Statistics

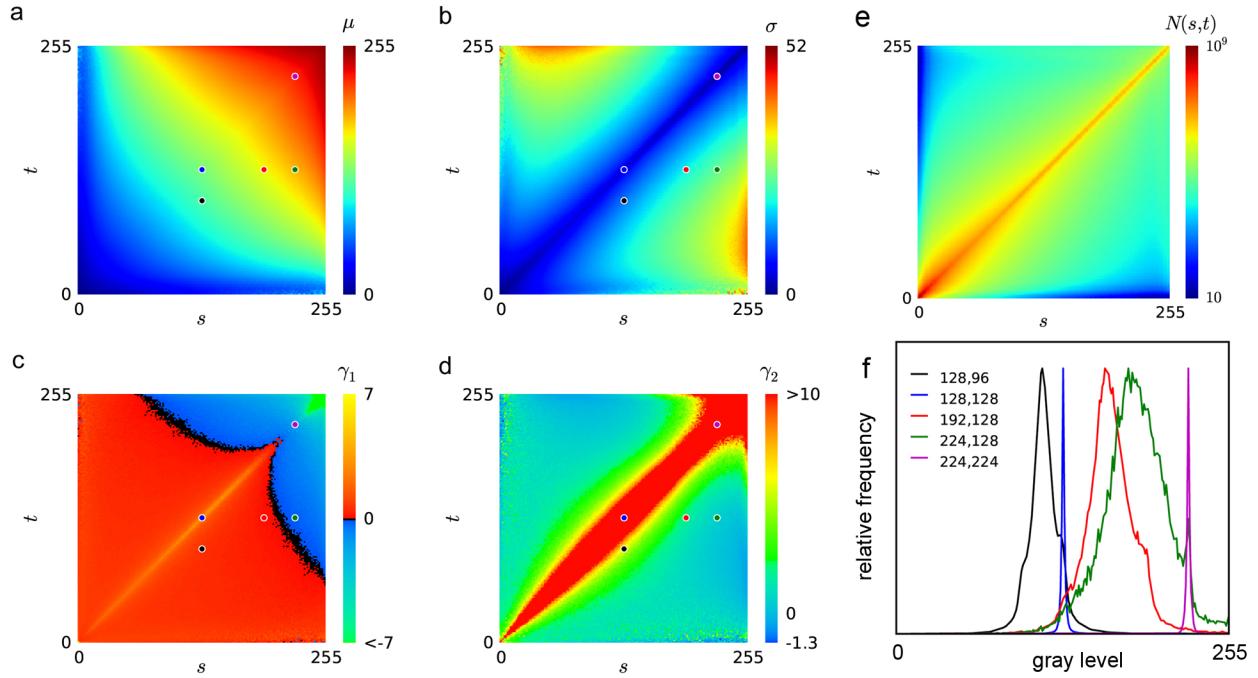
#### **Conditional moments for three neighboring image points**

To illustrate the conditional moments approach consider first the conditional moments of an image point  $x$  given the two flanking values  $s, t$  (see Figure 1a). The direct method of measuring the  $k^{\text{th}}$  moment,  $E(x^k | s, t)$ , is simply to compute the running sum of the values of  $x^k$  for all possible values  $s, t$ , using every pixel in the 700 training images (see eq. 1). The result is a 256 x 256 table for each moment, where each table entry is for a particular pair of flanking values. These tables were estimated separately for horizontal and vertical flanking points; however, the tables did not differ systematically and hence were combined. This property held for the other cases described below, and hence we show here only the combined tables.

Figures 2a-d shows tables of the first four moments after they have been normalized. Thus, the tables show the conditional mean, standard deviation, skew and kurtosis of the center point for each pair of values of the flanking points. These tables have not been smoothed. Figure 2e shows the number of training samples that went into the estimates for each pair of flanking values. As can be seen, these natural image statistics are complex but very regular. The only trivial property of the tables is the symmetry about the diagonal, which simply says that  $s$  and  $t$  can swap locations without affecting the statistics. Some sense of what these tables represent can be gained from Figure 2f, which plots normalized histograms of  $x$  for the five pairs of flanking values indicated by the dots in Figures 2a-d. These distributions are approximately unimodal and so are relatively well described by the first four moments.

The table of conditional first moments in Figure 2a gives the Bayes optimal MMSE estimate of  $x$  given a flanking pair of points. The table of conditional second moments can be used to combine separate MMSE estimates (e.g., estimates based on horizontal and vertical flanking points; see later). Maximum *a posteriori* (MAP) estimates (or estimates based on other cost functions)

should be possible using the four moments to estimate the parameters of a suitable family of unimodal density functions. However, the focus in the remainder of this paper is on MMSE estimates.

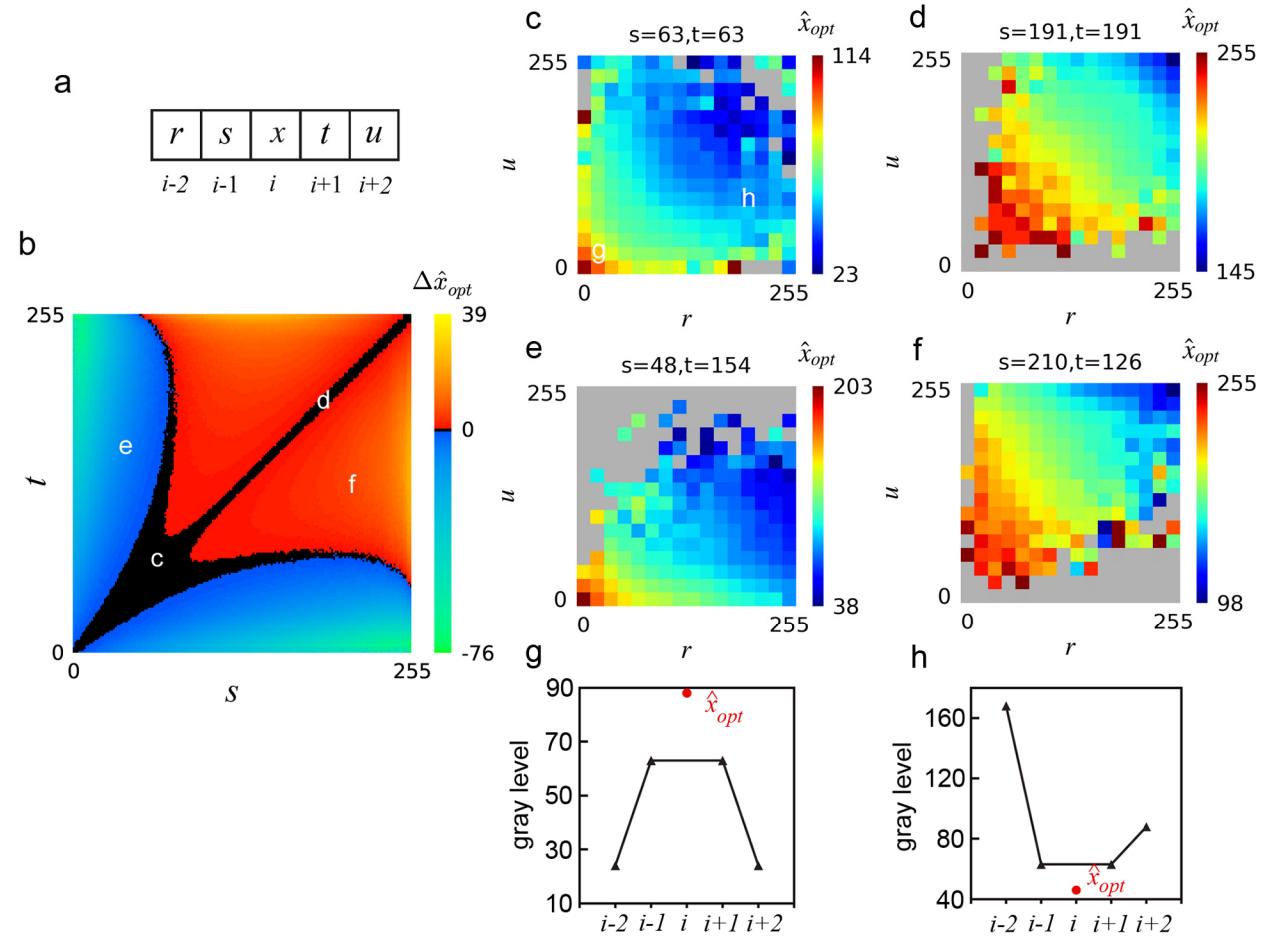


**Figure 2.** Conditional central moments for three neighboring image points. **a.** The 1<sup>st</sup> moments are the means:  $\mu = u_1$ . **b.** The 2<sup>nd</sup> moments are plotted as the standard deviation:  $\sigma = \sqrt{u_2}$ . **c.** The 3<sup>rd</sup> moments are plotted as the skewness:  $\gamma_1 = u_3/u_2^{3/2}$ . **d.** The 4<sup>th</sup> moments are plotted as the excess kurtosis:  $\gamma_2 = u_4/u_2^2 - 3$  (a Gaussian has an excess kurtosis of 0), where the central moments for each pair of values  $(s, t)$  were computed in the standard way:  $u_1 = n^{-1} \sum x$ ,  $u_2 = n^{-1} \sum (x - u_1)^2$ ,  $u_3 = n^{-1} \sum (x - u_1)^3$ ,  $u_4 = n^{-1} \sum (x - u_1)^4$  (see Kenny & Keeping 1951; Papoulis 1984). **e.** Number of training samples for each value of  $s, t$ . **f.** Conditional distributions for the five example pairs of  $(s, t)$  values indicated by the dots in **a-d**. The plots have not been smoothed.

### Estimation of missing or occluded image points

Consider first the case of estimating  $x$  from just the two nearest flanking values  $(s, t)$ . The MMSE estimate function  $f(s, t)$  is given by the table in Figure 2a. If the joint probability distribution  $p(x, s, t)$  were Gaussian, then (given two obvious constraints that apply in our case: when  $s = t$  the optimal estimate is  $s$ , and symmetry about the diagonal) the optimal estimate would be the average of the two flanking values:  $f(s, t) = (s + t)/2$  (see Appendix). Similarly, if the joint probability distribution  $p(x, s, t)$  is Gaussian on log axes, then the optimal estimate would be the geometric average of the two flanking values:  $f(s, t) = \sqrt{st}$ . These are useful references against which to compare the measured table.

The plot in Figure 3b shows the difference between the measured table and the table predicted if the joint distribution were Gaussian (the prediction of linear regression). As can be seen, there are systematic but complex differences between the directly measured function and the Gaussian prediction. The black pixels show cases where the optimal estimate is equal to the average of the flanking pixel values. The warm-colored pixels indicate cases where the estimate is above the average of the flanking pixels and the cool-colored pixels where the estimate is below the average. Examining where differences from the Gaussian prediction occur in representative natural images suggest that they tend to be associated with well-defined contours.



**Figure 3.** Estimation of a missing or occluded point. **a.** In the case of 3<sup>rd</sup> order statistics, the aim is to predict  $x$  from  $(s, t)$ . In the case of 5<sup>th</sup> order statistics the aim is to predict  $x$  from  $(r, s, t, u)$ . **b.**

Optimal estimation of  $x$  given  $(s, t)$ . The color map shows the difference between the optimal estimate and average of  $s$  and  $t$ , which corresponds to the optimal estimate assuming a Gaussian joint distribution. **c-f.** Optimal estimation of  $x$  from  $(r, s, t, u)$ , for the specific values of  $(s, t)$  indicated by the white letters in **b**;  $c = (63, 63)$ ,  $d = (191, 191)$ ,  $e = (48, 154)$ ,  $f = (210, 126)$ . Each plot gives the optimal estimate as a function of the further flanking values  $r$  and  $u$ . Gray indicates no data. **g,h.**

Examples of specific point predictions for the two quantile bins labeled in **c**. The symbols show the observed values  $(r, s, t, u)$ , the red circle shows the optimal prediction.

Importantly, there are no obvious forms of invariance in Figure 3b (or 2a), except for the expected diagonal symmetry:  $f(s, t) = f(t, s)$ . In other words, the optimal estimates depend on the specific absolute values of  $s$  and  $t$ , not only on their relative values. The lack of obvious invariance in Figure 3b shows that less direct methods of measuring the image statistics could have missed important structure. For example, suppose each signal were normalized by subtracting the mean of  $s$  and  $t$  before measuring the conditional first moments, and then the mean was added back in to create a table like Figure 2a. If this procedure did not miss any information, then the difference values plotted in Figure 3b would be constant along every line parallel to the diagonal (see Appendix). In fact, most interpolation methods (e.g., linear, cubic, and spline methods) ignore the mean signal value and consider only the difference values. Similarly, if normalizing each signal by subtracting and dividing by the mean of  $s$  and  $t$  before measuring the conditional first moments did not lose information, then the difference values in Figure 3b would be constant along any line passing through the origin (see Appendix). In other words, these common pre-processing steps would have led us to miss some of the statistical structure shown in Figure 3b. Similar conclusions are reached if the log Gaussian prediction is used as the reference, suggesting that the complex structure and the lack of invariance are not due to the non-Gaussian shape of the marginal distribution of intensity values. As a further test, we quantized the 14-bit raw images into 8-bit images having an exactly Gaussian histogram, and then ran the above analyses. The resulting plots are qualitatively similar to those in Figure 3b.

Now consider the case of estimating  $x$  from all four neighboring values  $(r, s, t, u)$ . In this case the goal is to measure the MMSE estimation function

$$f(r, s, t, u) = E(x|r, s, t, u)$$

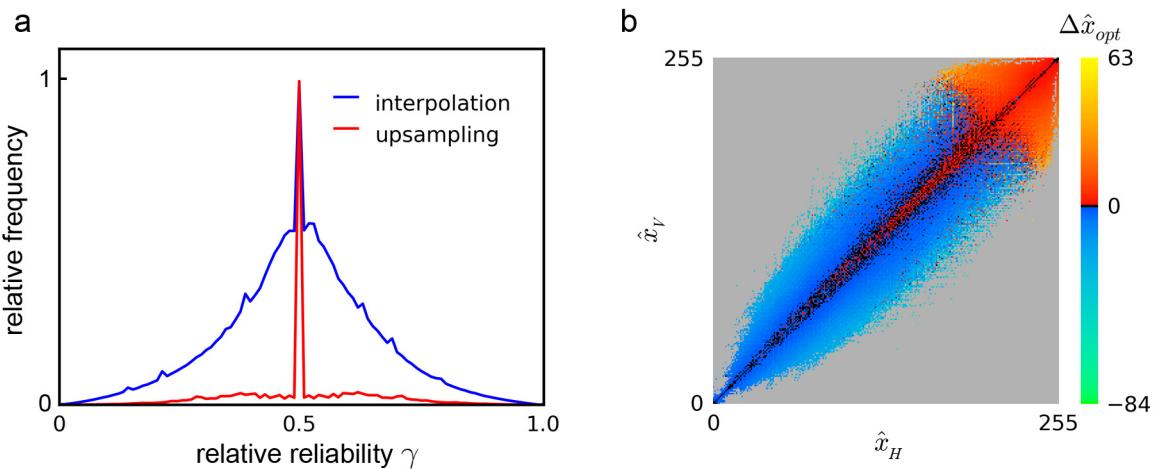
To facilitate measurement and visualization of this four-dimensional function it is useful to measure and plot separate two-dimensional tables for each pair of values  $(s, t)$ . In other words, we measure  $2^{16}$  tables of the form  $f_{s,t}(r, u)$ . This set of two-dimensional tables constitutes the full four-dimensional table.

The direct method of computing the mean of  $x$  for all possible flanking point values is not practical given the amount of image data we had available. One way to proceed is to lower the resolution of the table for the furthest flanking points. Reducing the resolution is justifiable for two reasons: (i) the furthest flanking points should generally have less predictive power, and (ii) reducing resolution is equivalent to assuming a smoothness constraint, which is plausible given the smoothness of the 3<sup>rd</sup> order statistics (plot in Figure 2a). We quantized the values of  $r$  and  $u$  into 16 bins each

$$f_{s,t}(r, u) \equiv f_{s,t}(q(r), q(u))$$

where  $q(\cdot)$  is a quantization function that maps the values of  $r$  and  $u$  into 16 integer values between 0 and 255.

The plots in Figure 3d-f are representative examples of tables for the different values of  $(s, t)$  corresponding to the white letters in Figure 3b. As in Figure 2a, the color indicates optimal estimate. (In the plot, gray indicates values of  $(r, s, t, u)$  that did not occur in the training set.) Again, the statistical structure is complex but relatively smooth and systematic. Consider the tables for values of  $(s, t)$  that are along the diagonal ( $s = t$ ). When the values of  $r$  and  $u$  are below  $s$  and  $t$  then  $(r, s, t, u)$  is consistent with a convex intensity profile and the estimate is above the average of  $s$  and  $t$  (see Figure 3g). On the other hand, when the values of  $r$  and  $u$  exceed  $s$  and  $t$  then  $(r, s, t, u)$  is consistent with a concave intensity profile and the estimate is below the average of  $s$  and  $t$  (see Figure 3h). Similar behavior is seen in the tables for values of  $(s, t)$  that are not along the diagonal, but the tables are less symmetric. These are intuitive patterns of behavior, but again there are no obvious forms of invariance except for the diagonal symmetry when  $s = t$ .



**Figure 4.** Reliability and recursive estimation. **a.** Relative frequency of the relative reliability of the two estimates based on values in the horizontal and vertical directions for the task of estimating the missing or occluded image point (blue curve) and estimating a higher resolution image point (red curve). **b.** Recursive estimation function for upsampling. The horizontal axis gives the optimal estimate of  $x$  given the observed values in the horizontal direction in the image, and the vertical axis gives the optimal estimate of  $x$  given the observed values in the vertical direction, for a relative reliability of 0.5. The color scale gives the estimate of  $x$  given the estimates in two directions minus the average of the two optimal estimates. Gray indicates no data.

As mentioned earlier, the estimation functions obtained by analyzing horizontal and vertical flanking points do not differ systematically, and thus in each case the horizontal and vertical data were combined to obtain a single estimation function, which can be applied in either direction. However, when this function is applied in both the horizontal and vertical directions, for the

same image point, somewhat different estimates are often obtained. In other words, there is additional information available in the two estimates.

It is possible to characterize some of this additional information by measuring the relative reliability of the two estimates, which determines how much weight each should be given in the combined estimate. The blue curve Figure 4a plots the distribution of relative reliability for missing-image point (interpolation) task when the two estimates are based on the horizontal and vertical values of  $(r, s, t, u)$ . There is considerable variation in the relative reliability, suggesting that this measure contains useful information. This large variation may be due to the structure of intensity gradients in natural images. As shown in Figure 2b, estimates based on image points that are near equal are more reliable than estimates based on image points that are less equal. But, image points that are relatively parallel to the local intensity gradient are closer to equal than image points that are relatively perpendicular to the local intensity gradient. In the performance demonstrations below, we combine horizontal and vertical estimates ( $\hat{x}_H$  and  $\hat{x}_V$ ) using relative reliability (see equation (4) in the Methods).

### ***Estimation of a high resolution image from a low resolution image***

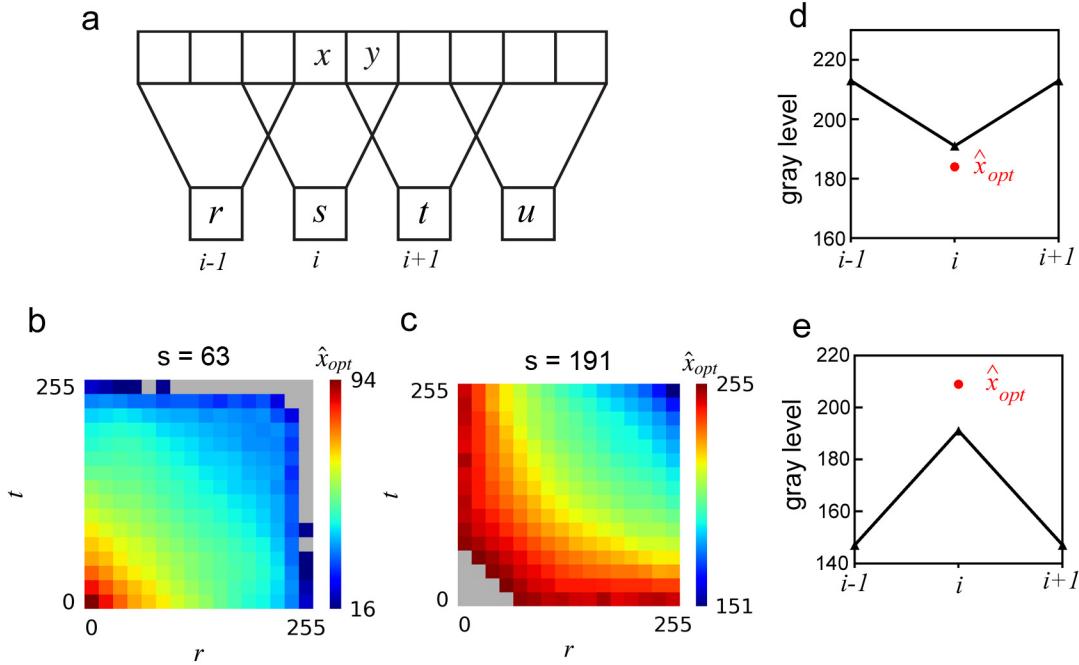
The task of upsampling from a low resolution to a high resolution image is illustrated in Figure 5a. Upsampling involves both interpolation and compensation for spatial filtering. In general, a low resolution image can be conceptualized as being obtained from a higher resolution image by the operation of low-pass spatial filtering followed by down-sampling. For example, a well-focused camera image (or a foveal cone image) can be regarded as a higher resolution image that has been filtered by a small amount of optical blur plus summation over the area of the sensor element, followed by down-sampling at the sensor element locations. As illustrated in Figure 5a, we consider here the case where the filtering operation has an odd kernel so that each low-resolution image point is centered on a high resolution image point. The case for an even kernel is similar, but differs in certain details (it is somewhat simpler).

As shown in Figure 5a, different statistics were used for estimating points like  $x$  that are aligned with low resolution points and those like  $y$  that are centered between low resolution points.

Following similar logic as before, the Bayes optimal estimate of  $x$  is given by  $E(x|r, s, t)$  and the optimal estimate of  $y$  is given by  $E(y|r, s, t, u)$ , and thus our aim is to measure the functions  $f_s(r, t)$  for the aligned points ( $x$ ) and the functions  $f_{s,t}(r, u)$  for the between points ( $y$ ). To measure these functions we first filtered all of the calibrated images with a 3x3 Gaussian kernel and then down-sampled them by a factor of two in each direction to obtain a set of low-resolution images together with their known higher-resolution “ground-truth” images. Measuring the functions required quantizing the more distant image points, and thus (as before) we measured  $f_s(q(r), q(t))$  and  $f_{s,t}(q(r), q(u))$ , where  $q(\cdot)$  is a quantization function.

Examples of the tables measured for the aligned points ( $x$ ) are shown in Figures 5b and 5c. The plots show the optimal estimate of  $x$ . As can be seen, if the values of  $r$  and  $t$  are greater than  $s$ , then the optimal estimate of  $x$  is less than  $s$  (see Figure 5d), and vice versa if the values of  $r$  and  $t$  are less than  $s$  (see Figure 5e). The tables measured for the between points ( $y$ ) are

similar to those in Figure 3c-f. The tables plotted in Figures 5 are intuitively reasonable, and reflect both the structure of the natural images and the effect of the blur kernel.



**Figure 5** Estimation of a high resolution image from a low resolution image. **a.** 4<sup>th</sup> order statistics are used to predict  $x$  from  $(r, s, t)$  and 5<sup>th</sup> order statistics are used to predict  $y$  from  $(r, s, t, u)$ . **b,c.** Optimal estimation of  $x$  given  $(r, s, t)$  for two values of  $s$ . The color indicates the optimal estimate of  $x$ . Gray indicates no data. **d,e.** Examples of specific point predictions for the two bins indicated in **c**. The black triangles are specific values  $(r, s, t)$ , the lines show the linear prediction, and the red circles are the optimal estimate.

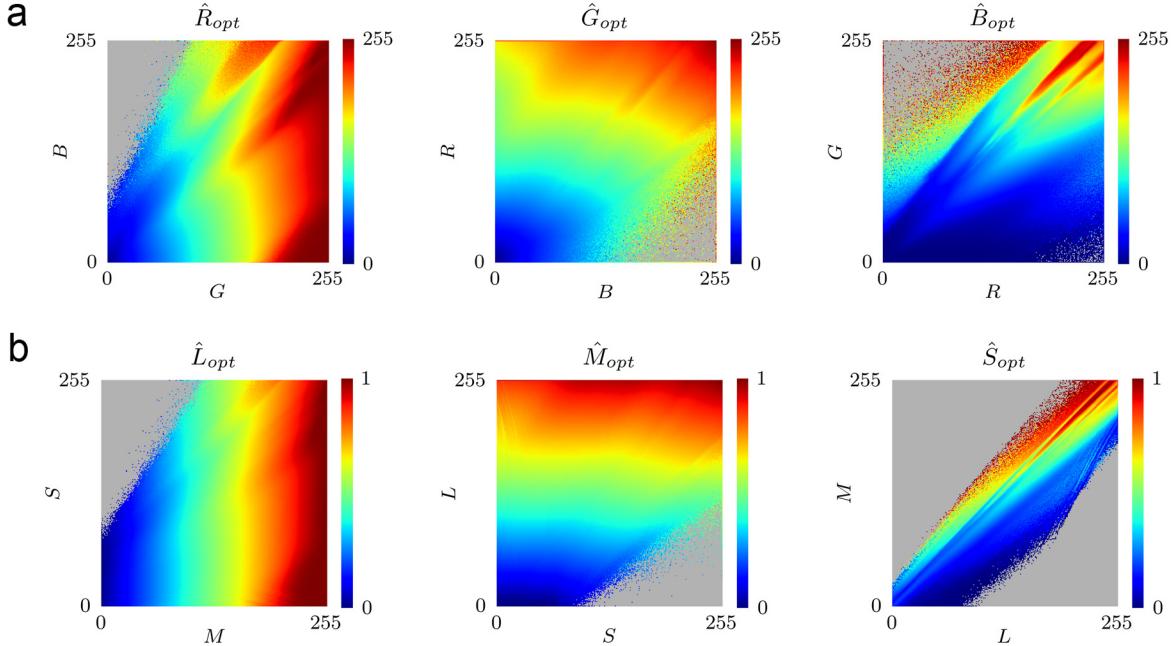
The relative reliability for the horizontal and vertical estimates of the aligned points ( $x$ ) is shown by the red curve in Figure 4a. The relative reliability is concentrated around 0.5 suggesting that this measure contains less useful information than in the case of the missing-image-points task. Indeed, we found that more useful information is contained in the recursive conditional first moments (see equation 5 in Methods). Figure 4b plots the estimate of  $x$  given  $\hat{x}_H$  and  $\hat{x}_V$  minus the average of the two estimates, for a relative reliability ( $\gamma$ ) of 0.5. Again the statistics are complex but regular. If the estimates nearly agree, then the best estimate is near the average of the two estimates. If the estimates do not agree, then the best estimate is higher than the average when the average is big, but lower than the average when the average is small.

### Estimation of a missing color channel

The task of estimating a missing color channel is illustrated in Figure 1c. Recall that in this task we ignore spatial information. We consider the task of estimating each color channel given the other two. We do this for two cases, one where the three channels are the  $R, G$ , and  $B$  sensors in the camera, and the other where the three channels are the  $L, M$ , and  $S$  cones in the human retina.

In other words, we measured the following statistics:  $E(R|G, B)$ ,  $E(G|R, B)$ ,  $E(B|R, G)$ ,  $E(L|M, S)$ ,  $E(M|L, S)$  and  $E(S|L, M)$ . Recall that these first conditional moments are the Bayes optimal estimates when the goal is to minimize the mean squared error.

Figure 6 plots the estimates for each class of sensor. Once again the statistics are complex but regular, and there are no obvious forms of invariance. In other words, the estimates of a color channel value depend on the specific absolute values of the other channels.

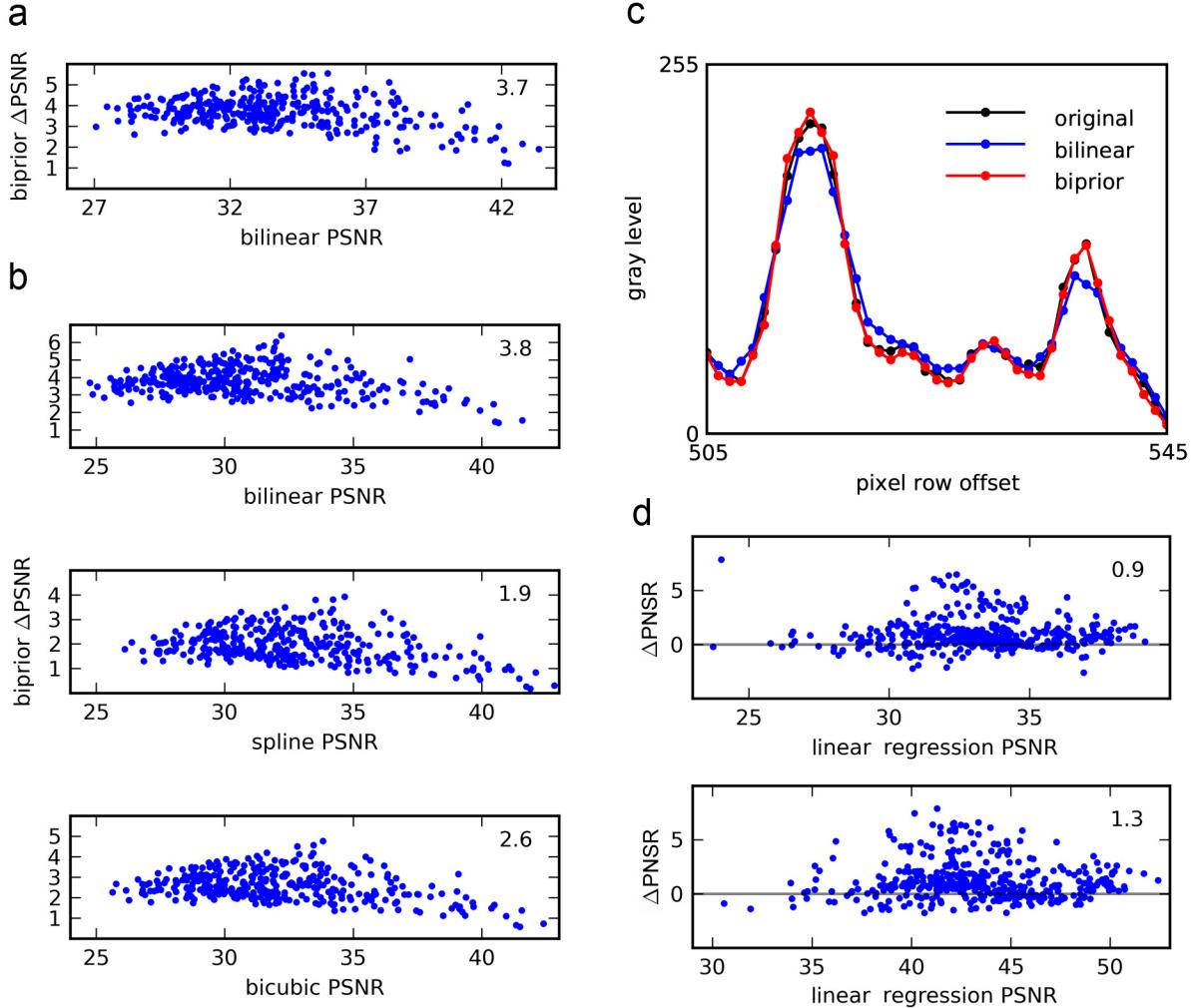


**Figure 6** Estimation of a missing color channel. Plots show the optimal estimate of the missing color value given the observed color values on the horizontal and vertical axes. **a.** Estimation of missing  $R$ ,  $G$  or  $B$  camera responses. **b.** Estimation of missing  $L$ ,  $M$  or  $S$  cone responses. Gray indicates no data.

## Performance

The measurements in Figures 2-6 reveal complex statistical structure in natural gray-scale and color images. These statistics are of interest in their own right, but an obvious question is what their implications might be for vision and image processing. Perhaps the most fundamental questions are: How useful would knowledge of these statistics be to a visual system? And, does the human visual system use this knowledge? The conditional first moments in Figures 2-6 give the Bayes optimal (MMSE) estimates for the specific tasks shown in Fig. 1. Thus, to address these questions we could compare optimal performance on these tasks with that of other simpler estimators, and we could also potentially compare ideal and human performance in the first two tasks (i.e. subjects could make estimates based a short row or column of natural image pixels). While this is doable and potentially useful, we show here predictions for the third task and for the two-dimensional versions of the first two tasks, where multiple estimates are combined using measured recursive conditional moments. Although we cannot guarantee optimal performance on the two dimensional tasks, they are closer to the normal image processing tasks carried out in

biological and artificial vision systems. All comparisons were carried out on a separate random sample of test images.



**Figure 7** Performance in three estimation tasks. **a.** Estimation of missing or occluded image points. Plotted is the difference in peak signal to noise ratio (PSNR) between the biprior and bilinear estimators for the 349 test images, as a function of the PSNR of the bilinear estimator. Each point represents a test image; points above zero indicate better performance for the biprior method. The average increase in PSNR across all test images is 3.7 db. **b.** Difference in PSNR for upsampling of the 349 test images as a function of the PSNR of the method indicated on the horizontal axis. **c.** Comparison of bilinear and biprior interpolation along a segment of scan line from a test image. (Note that every other interpolated scan line, like this one, contains no original image pixels.) **d.** Difference in PSNR between optimal estimator and linear estimator (linear regression) as a function of the PSNR for the linear estimator. The upper plot is for RGB images and the lower for LMS images. The values of PSNR are computed in the CIE  $I^*a^*b^*$  color space.

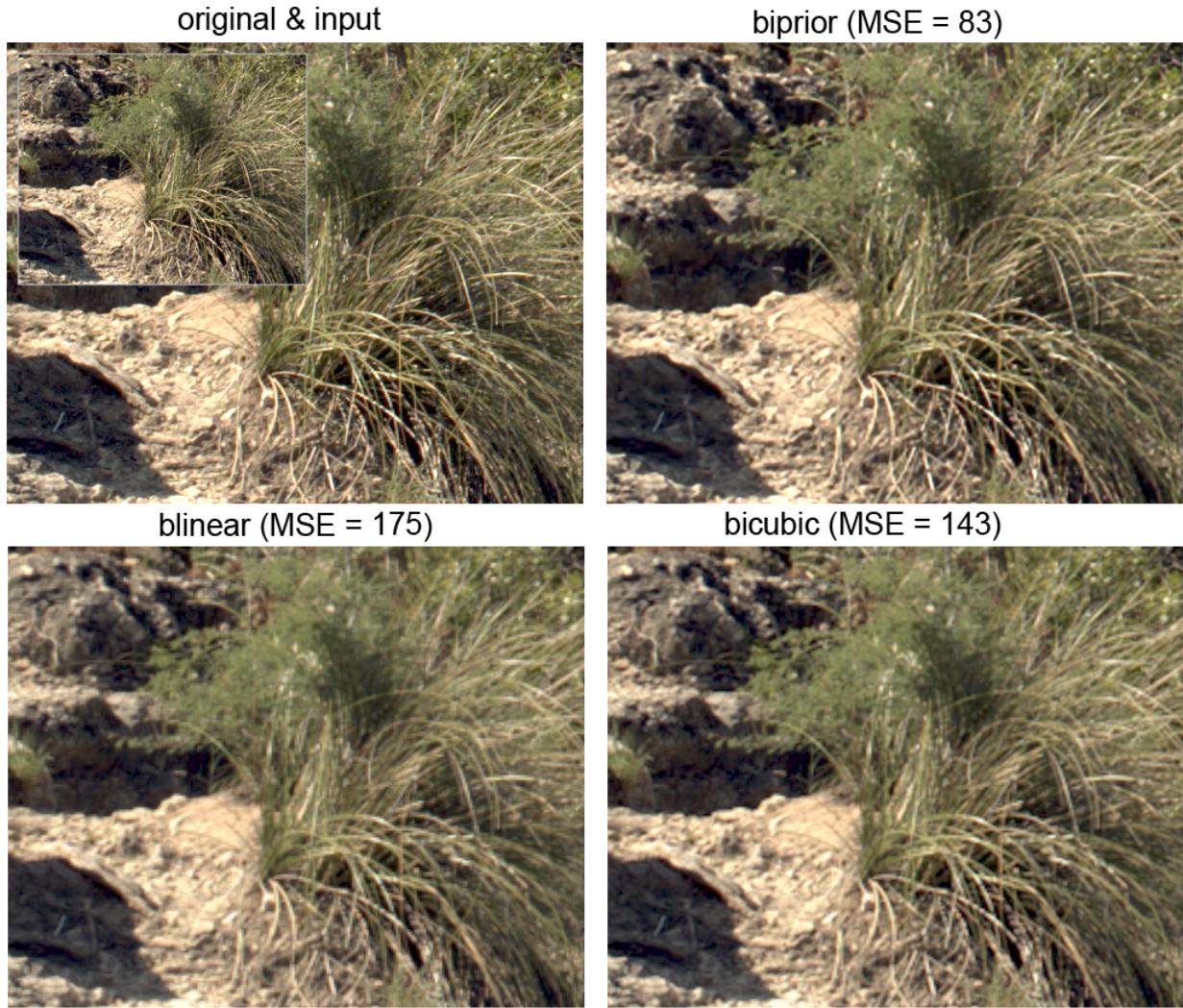
### **Estimation of missing or occluded image points**

For each pixel in each test image, we estimated its value by applying the MMSE estimation function  $f(r, s, t, u)$  in both directions and combining the two estimates using the measured relative reliability. This is what we call the *biprior* estimator. We also estimated the value of

each pixel using the standard bilinear estimator. Estimation accuracy is quantified as the peak signal-to-noise ratio,  $PSNR = 10 \log(255^2/MSE)$ , where  $MSE$  is the mean squared error between the estimated and the true values. Thus, a 50% decrease in  $MSE$  corresponds to a 3 db increase in  $PSNR$ . Figure 7a plots for each test image the difference in the  $PSNR$  of the two estimators as a function of the  $PSNR$  of the bilinear estimator. All blue points above zero indicate images for which the biprior method was more accurate. In general, the biprior method performs better. The increases in performance are substantial (3.7 db on average) and demonstrate the potential value of the natural image statistics in estimating missing or occluded points.

If the biprior estimate is based on only one direction then the increase in performance over bilinear is 1.0 db on average, and if the two estimates are simply averaged (given equal weight) then the average increase in performance over bilinear is 3.2 db. Thus, all components of the biprior estimator make a substantial contribution to performance. If only the two neighboring pixel values are used (i.e., applying  $f(s, t)$  in both directions and combining with relative reliability) the increase in performance over bilinear is 0.9 dB. This shows that the complex statistical structure shown in Figure 2 and 3a is useful, but it also shows that the additional flanking pixels make a substantial contribution to performance.

We also explored a number of related methods to determine if significant improvements in performance were possible. First, we applied exactly the same method for the diagonal directions rather than cardinal directions and found performance to be worse than even the simple bilinear estimator. Further, combining the diagonal estimates with the cardinal estimates did not improve the performance of the biprior estimator. This suggests that the pixels closest to the missing pixel carry most of the useful information. Next, we explored one and two dimensional kernels obtained with principal components analysis (PCA). The one dimensional kernels were obtained for the 6 pixels on the two sides of the missing pixel in the horizontal or vertical direction. The two-dimensional kernels were obtained for the 24 surrounding pixels in a 5x5 block. We then applied the current conditional-moments method for different choices of four PCA kernels. In other words, for each test stimulus we applied the four kernels and used the resulting four values as the context variables  $(r, s, t, u)$ . None of these performed well. Next, we explored one and two-dimensional kernels obtained using a recent method (accuracy maximization analysis) that finds kernels that are optimized for specific identification and estimation tasks (Geisler, et al. 2009). In the one dimensional case, we found that the AMA kernels were essentially equivalent to (covered approximately the same subspace as) the four kernels corresponding to the separate four pixels in the horizontal or vertical direction, and did not produce a significant improvement in performance. In the two dimensional case, we found that the four AMA kernels performed much better than the estimates based on only one direction, but did not perform better than the combined vertical and horizontal estimates. Finally, we tried combining the horizontal and vertical estimates with the full recursive table (equation 5) and found only a very slight increase (less than 1%) in performance over using equation (4). While none of these analyses is definitive, together they suggest that the biprior estimator makes near optimal use of the local image structure in natural images for the generalized missing image point task.

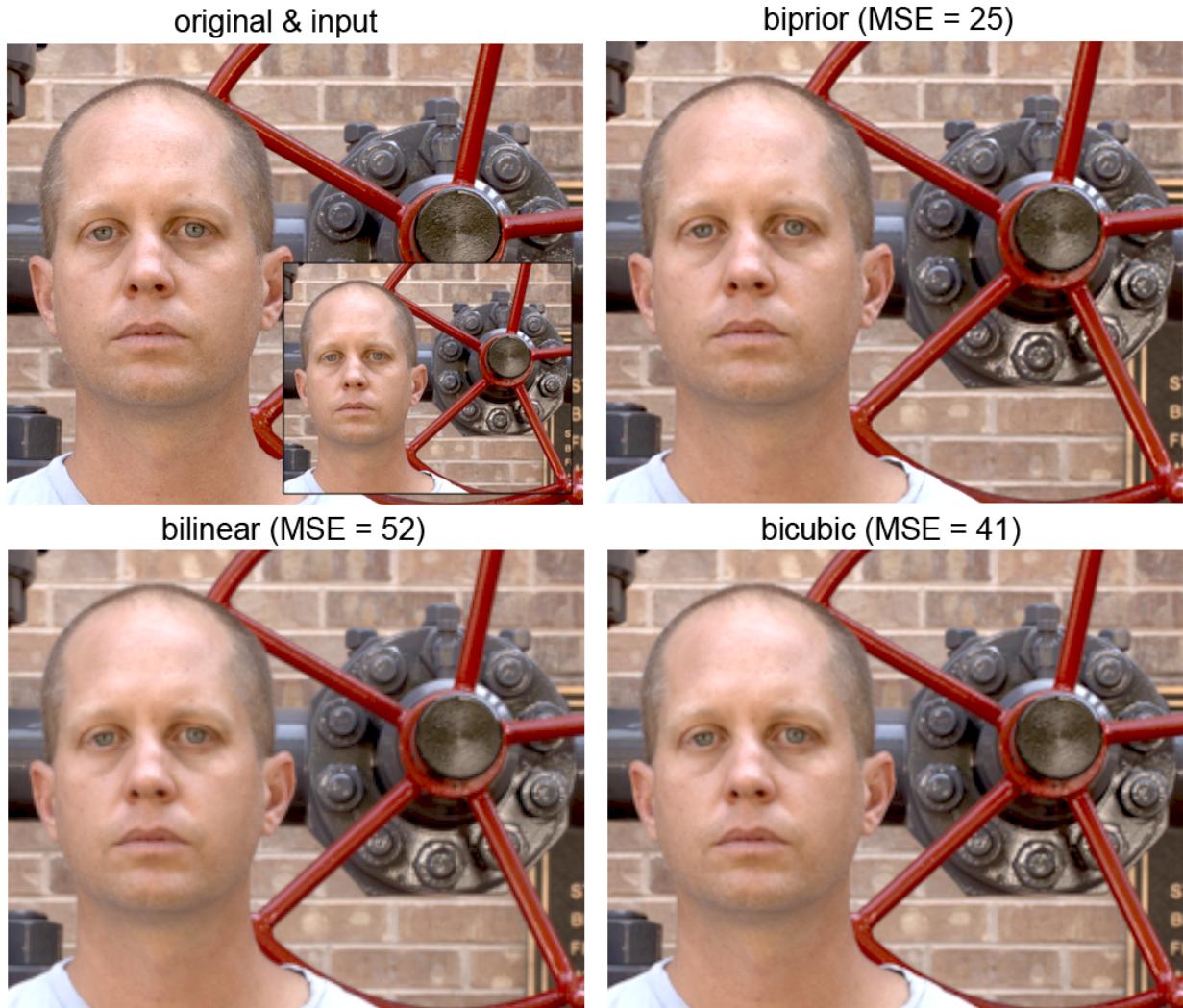


**Figure 8** Comparison of upsampling based on natural scene statistics (biprior method) with standard methods based on bilinear and bicubic interpolation. This is a representative example from the 349 test images; the reduction in mean squared error for biprior relative to bilinear is 53%, whereas the median reduction is 58%. The large image in the upper left panel is the original image and the inset is the down-sampled input image.

### ***Estimation of a high resolution image from a low resolution image***

For measuring the accuracy of the upsampling methods, the test images were blurred by a 3x3 Gaussian kernel, down sampled by a factor of two in each direction, and then upsampled using one of four methods: biprior, bilinear, spline, and bicubic upsampling. Bilinear upsampling uses linear interpolation of the two nearest neighbors, applied in both directions. Spline upsampling uses cubic interpolation of the four nearest neighbors (along a line) in both directions. Bicubic upsampling uses full cubic interpolation of the 16 (4 x 4) nearest neighbors. Biprior upsampling used tables for  $x$  points having 256 levels (8 bits) each for  $r$ ,  $s$  and  $t$ . The horizontal and vertical estimates for the  $x$  points were then combined using the recursive table (see equation 5 and Figure 4b). These estimated  $x$  points were then used as the four flanking points to estimate the  $y$  points. One final point in each 3x3 block (a  $z$  point not shown) was estimated by combining

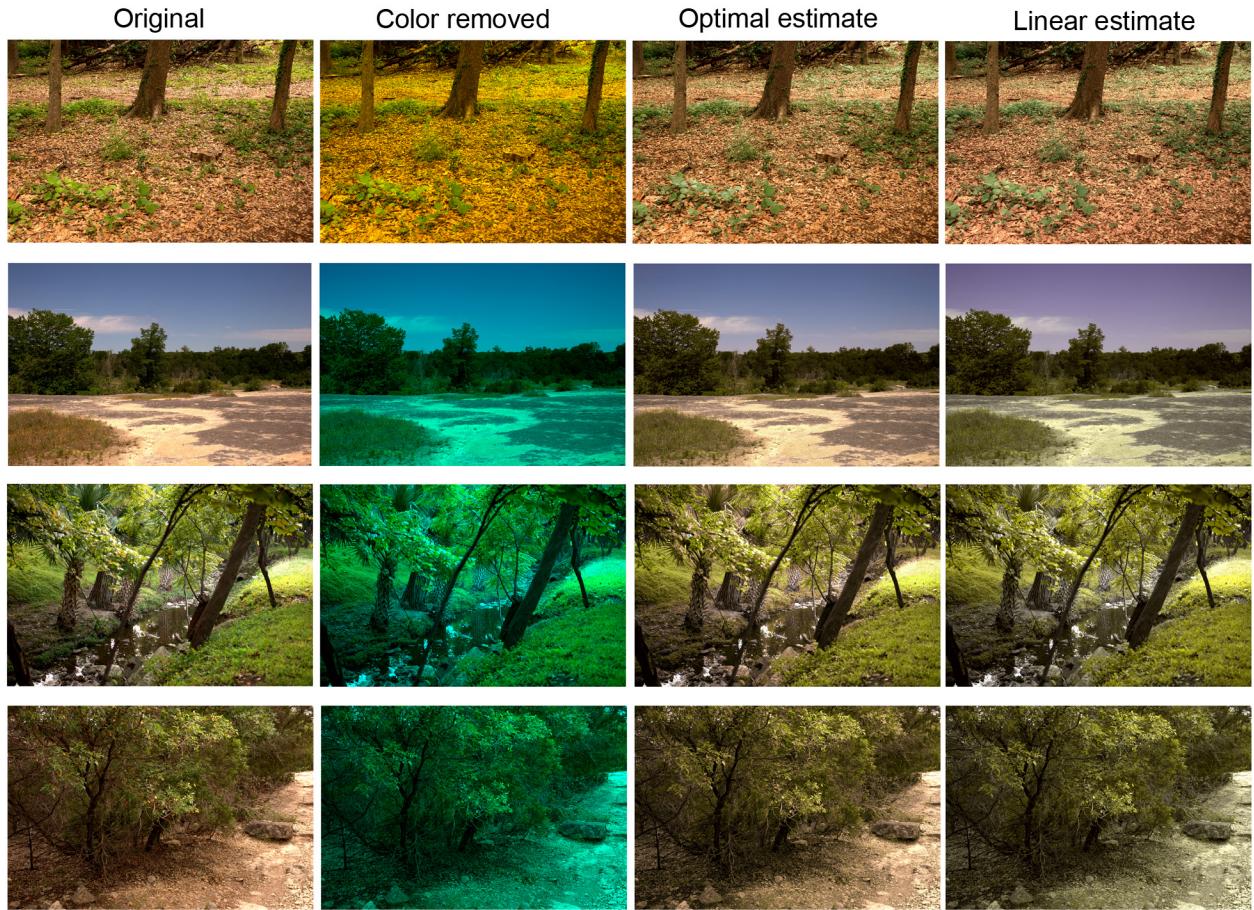
vertical and horizontal estimates from the flanking estimated  $y$  points, using the recursive table. The results are shown in Figure 7b. The biprior estimator performs substantially better than the standard methods on all the test images, demonstrating the potential value of the measured natural image statistics in decoding photoreceptor (or ganglion cell) responses. As in the case of estimating the missing point, all components of the biprior estimator make a substantial contribution to performance. However, unlike missing point estimation, using only the reliability to combine horizontal and vertical estimates (equation 4) did not work as well as using the recursive table, presumably because of the smaller variation in relative reliability (red curve in Figure 4a).



**Figure 9** Comparison of upsampling based on natural scene statistics (biprior method) with standard methods based on bilinear and bicubic interpolation. Images of faces and human-made objects were not in the training set of images. Thus, this example demonstrates the robustness of the statistics.

A closer look at the differences between bilinear and biprior estimates is shown in Figure 7c, which plots the bilinear (blue) and biprior (red) estimates against the original image points (black) for a typical segment of a scan line from a test image.

Figures 8 and 9 compare the original and the down-sampled input images with the upsampled images obtained using the biprior, bilinear, and bicubic methods. In addition to largely maintaining the original contrast, the biprior method does a better job of restoring detail. Similar behavior is seen with all the images we have inspected from the set of test images; indeed, the improvement in MSE for these images relative to bilinear (53% and 52%, respectively) is slightly less than the median improvement of 58%. The upsampling of these color images was obtained by applying the same estimator to each color channel. The relative improvement in MSE of the biprior estimator for these color images is only very slightly less (about 1%) than for the gray scale images on which the image statistics were learned. This demonstrates the robustness of the image statistics. Figure 9 also demonstrates the robustness of the image statistics in that the training set contained no human made objects or human faces.



**Figure 10.** Estimation of a missing color channel. In the first row the  $B$  channel is estimated. In the other rows the  $R$  channel is estimated. In row 3 the linear estimate is more accurate than the optimal estimate. Interestingly, the optimal estimate appears more “natural” in color, which was generally true.

### ***Estimation of a missing color channel***

To test the performance accuracy of the Bayes optimal estimators for missing color channels we separately removed each channel and then estimated the value of that channel at each pixel location in each test image. We quantified the estimation accuracy by the mean squared error

between the estimated and original color in the CIE L\*a\*b\* uniform color space. The average increase in PSNR of the optimal estimate over that of linear regression was as follows for each color channel: (R,G,B) = (1.2 db, 0.59 db, 0.93 db), (L,M,S) = (1.3 db, 0.91 db , 2.3 db). Note that linear regression incorporates all the pairwise correlations between the color channels (i.e., the full covariance matrix). Also note there are many locations in the color spaces of Figure 6 for which there is no data (the gray areas) and thus one might wonder if performance would be better if the spaces were filled in better. In fact this is not an issue. For the random set of test images, there were very few occasions when an empty location was encountered (less than one pixel per two images).

Estimation accuracy was not equally good for each color channel in an image. The percentage of images where each channel was easiest to estimate were as follow: (R,G,B) = (62%, 26%, 12%), (L,M,S) = (78%, 22%, 0%). Figure 7d plots the difference in *PSNR* between the optimal estimator and the linear estimator (linear regression) as a function of the *PSNR* for the linear estimator, for the channel that was easiest to estimate. (Usually, the channel that was easiest to estimate was the same for the optimal and linear estimators, but in a small percentage of cases the best channel was different.) The upper panel plots accuracy for a missing R, G, or B channel; the lower panel for a missing L, M, or S channel. Points above the horizontal line indicate images where the optimal estimate was more accurate than the linear-regression estimate.

Several randomly selected examples are shown in Figure 10. The first column shows the original image. The second column shows the image after a color channel has been removed (the channel values set to zero). The third column shows the result of applying the optimal estimator. Finally, the fourth column shows the result of the linear regression estimator. Although the color is not accurately reproduced, it is nonetheless surprising how natural the estimated images in the third column look, in the sense that one would be hard pressed to say that a color channel had been estimated without seeing the original. The estimates based on linear regression are also impressive, but they tend to look either too uniform or too purple/violet in color. For example, notice the lack of browns and tans in the bottom three rows, the purple tinted sky in the second row and the pink tinted leaves in the first row. Thus, it appears that the statistics in Figure 6 contain substantial color redundancy in natural outdoor images that is not captured by linear regression (e.g., Gaussian models).

## Discussion

A direct conditional moments approach was used to measure joint statistics of neighboring image points in calibrated gray-scale and color natural images. Specifically, optimal (MMSE) estimators were measured for the tasks of estimating missing image points, estimating a higher resolution image from a lower resolution image (super resolution), and estimating a missing color channel. In the different cases, the measured estimators are functions of two, three, or four neighboring image points. In general, we find that the estimation functions are complex but also relatively smooth and regular (see Figures 2-6). Importantly, the MMSE functions have no obvious forms of invariance, except for an expected diagonal symmetry. The lack of invariance implies that the various simplifying assumptions of invariance that are typical in less direct measurement methods probably miss important statistical structure in natural images. This conclusion is clearest for the invariance assumptions implicit in the preprocessing

transformations often applied to training signals (e.g., subtracting off the local mean or subtracting and scaling by the local mean), but it probably holds for a number of other invariance assumptions as well.

To begin exploring the implications of the measured statistics, we compared the performance of the Bayesian MMSE estimators with standard methods. We find that against objective ground truth, the MMSE estimators based on natural image statistics substantially outperform these standard methods, suggesting that the measured statistics contain much useful information. Thus, our results have potentially important implications for understanding visual systems and for image processing. Specifically, given the demonstrated usefulness of the statistics, it seems likely that these statistics have been incorporated, at least partially, into biological vision systems, either through evolution or through learning over the lifespan. Furthermore, because the measured statistics are so rich and regular, they make many predictions that should be testable in psychophysical experiments.

For example, the task of estimating missing image points could be translated into psychophysical experiments using a paradigm similar to that of Kersten (1987), where observers are asked to estimate the gray scale value at the missing image points. The natural image statistics measured in Figures 2 and 3 make many strong predictions. Whenever the values of  $r$  and  $u$  are below  $s$  and  $t$  in  $r, s, t, u$  space (see Figure 3), observers are predicted to estimate gray levels above the average of  $s$  and  $t$ , and the opposite when  $r$  and  $u$  are above  $s$  and  $t$ . Further, the uncertainty in the observers' estimates (e.g., psychometric function slopes) should be greater in those parts of the space where the standard deviations of the conditional moment distributions are greater (see Figure 2b).

Similarly, the task of estimating a higher resolution image from a lower resolution image could be translated into psychophysical experiments where observers are asked to compare altered foveal image patches with unaltered patches presented in the periphery. Specifically, if the visual system exploits the natural image statistics in Figures 4b and 5b, then a foveal patch that corresponds to the optimal upsampled estimate from the half-resolution version of the patch should be the best perceptual match to the (unaltered) patch presented at the appropriate eccentricity. Such predictions could be tested by varying the foveal patch around the optimal estimate.

The task of estimating the missing color channel is probably not so directly translatable into psychophysical experiments. However, the joint prior probability distribution represented by the conditional moments in Figure 6, and the other moments not shown, could play a role in understanding how photoreceptor responses are decoded. For example, Brainard, Hofer and Williams (2008) have shown that estimates of color names from the stimulation of single cones (Hofer, Singer & Williams 2005) is moderately predictable using a Bayesian ideal estimator that takes into account the spatial arrangement of the subjects'  $L$ ,  $M$  and  $S$  cones, as well as the prior probability distribution over natural images. They assumed that the prior is Gaussian and separable over space and color. The present statistics show that the prior over color is not Gaussian and that color estimation is more accurate if the joint statistics are taken into account. Thus, the present results may lead to somewhat different predictions for estimation experiments like those of Hofer et al. (2005).

The optimal estimators for the present tasks depend on both the statistical properties of the images in the training set as well as the spatial filtering and noise characteristics associated with the imaging system capturing the images. This suggests that there may be many different scientific and practical applications for the conditional moments approach described here. For example, the statistics of medical images, microscopic images, satellite images and diagnostic images in manufacturing and safety inspection are likely to be different in detail from the natural images considered here. Measurement of conditional moments for these other classes of image could potentially lead to a better understanding of their statistical structure and to substantially improved resolution, interpolation, and/or noise suppression. In general, the conditional moments approach may be applicable and useful whenever there is a sufficiently large set of training images. The approach could also be applied to natural signals other than natural images.

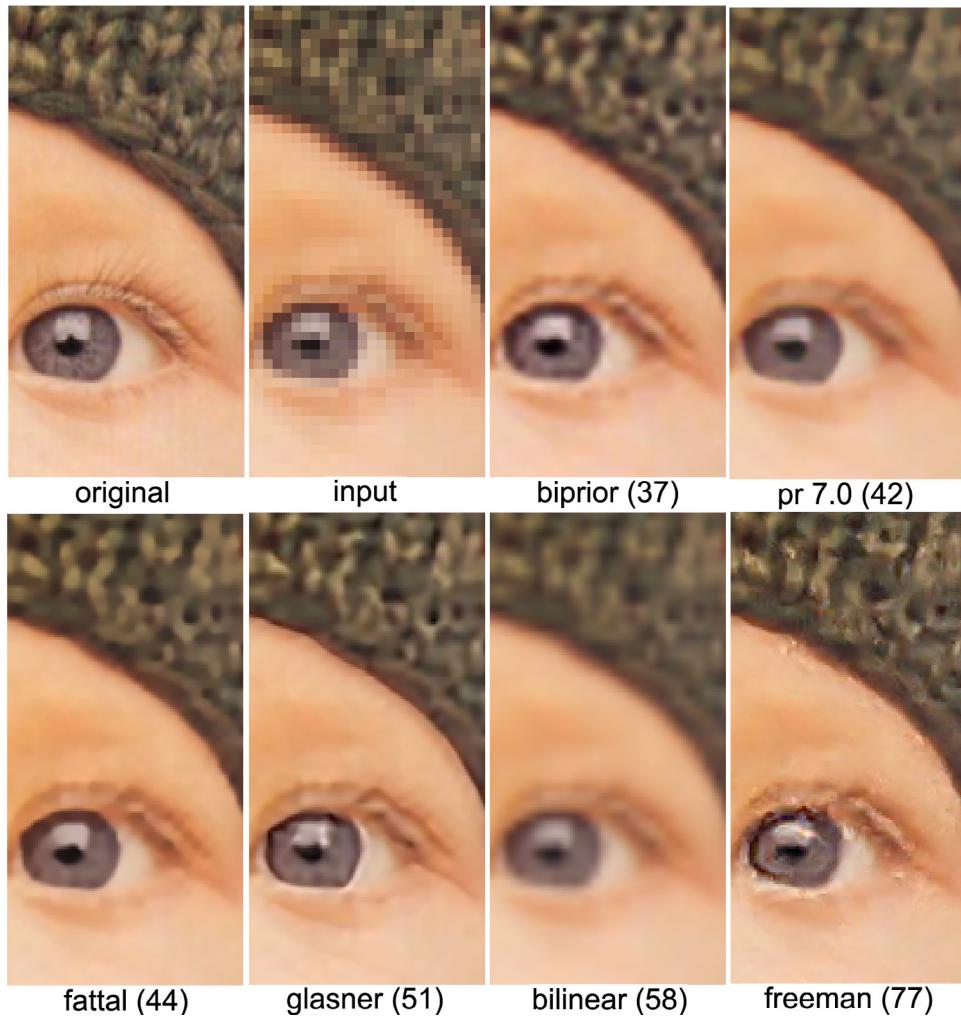
An advantage of the simple conditional moments approach is that it can be applied easily to arbitrarily large training sets, allowing relatively precise measurements of joint statistics. The number of training signals here was on the order of  $10^{10}$  and much larger training sets are possible. However, measuring the conditional moments is not feasible once the signal dimensionality becomes sufficiently high (although the dimensionality can be extended somewhat by measuring recursive conditional moments that take into account reliability; see equations 4 and 5 in the Methods). In high dimensions, parametric models are the most viable approach. Nonetheless, even when the dimensionality is high, conditional-moments measured in lower dimensions may be useful for constraining parameters and testing the assumed forms of invariance in the parametric models. An important two-dimensional example of this was the discovery that the unsigned magnitudes of nearby orthogonal wavelet values are positively correlated, even when their signed values are uncorrelated (Buccigrossi & Simoncelli 1999).

The conditional moments approach works very well for the estimation tasks described here because these tasks are inherently low dimensional. The missing color task is exactly three dimensional and hence there is sufficient data to accurately determine the full MMSE estimators. For the two-dimensional missing point task we demonstrated that the relevant information is highly concentrated in the neighboring four pixels in the horizontal and vertical direction. Our analyses are not definitive, but nonetheless suggest that substantially better performance cannot be obtained by moving to higher dimensions. Similar conclusions are likely for the two-dimensional upsampling (super resolution) task. However, for tasks that are inherently higher dimensional (e.g., that involve integrating information over larger spatial areas) the conditional moments approach will be less useful.

With respect to practical application it is worth noting that the estimators described here consist of fixed tables and hence can be applied by a simple lookup (albeit with substantial memory requirements for the tables). (Of course, parametric models in similarly low dimensional spaces could also be tabled.) The estimation functions also appear to be relatively smooth so it may be possible to find simple formulas that provide close approximation.

Could the conditional moments approach be useful for practical image processing tasks? Our goal here is to characterize image statistics rather than develop image processing algorithms. Nonetheless, we have made some preliminary comparisons with existing algorithms. Figure 11

shows a comparison with various recent super resolution (upsampling) algorithms (Fattal 2007; Freeman et al. 2002; Glasner et al. 2009; and a commercial algorithm, Perfect Resize 7.0), rank ordered by the mean squared error from the original image (in parentheses). The upsampling is by a factor of 4 and only a part of the image is shown, in order to make the differences between the algorithms more visible. Although the biprior algorithm has the lowest MSE, the smoothness of the pupil boundary is better captured by some of the other algorithms. On the other hand, the texture of the hat and the inside of the iris is better captured by the biprior algorithm. Also, some of the other algorithms (Fattal, Glasner, Freeman) are computationally intensive; the Freeman algorithm is reported to take over 30 min to produce a 0.25 MB (512 x 512) image, the Glasner algorithm 90 sec, and the Fattal algorithm 6 sec. On the other hand, the biprior algorithm produces a 12 MB (4000 x 3000) image in about a 0.1 sec. Thus, there may be upsampling applications where the conditional moments approach will have practical value.



**Figure 10.** Comparison of different upsampling (super resolution) algorithms. The original image was 512 x 512 and input was 128 x 128. The original and input images were provided by R. Fattal (<http://www.cs.huji.ac.il/~raananf/projects/upsampling/results.html>) and the upsampled images for the Fattal, Glasner and Freeman algorithms were provided by M. Irani (<http://www.wisdom.weizmann.ac.il/~vision/SingleImageSR.html>). Some appreciation for the sources of some of the distortions seen in the upsampled images can be obtained by inspecting the input image.

To get some idea of whether the conditional moments approach might be useful for other image processing tasks we put together a few un-optimized test algorithms. Their performance suggests that the conditional moments approach is promising, or at least may provide useful insight in developing practical algorithms. A first pass at interpolation of Bayer images (demosaicing digital camera images) gave mean squared error (MSE) performance better than the AHD algorithm (e.g., see Hirakawa & Parks 2005), a common current standard, and comparable to more recent algorithms (Li et al. 2008); also see Mairal et al. (2008). A first pass at denoising (for additive Gaussian noise) gave MSE performance equal to adaptive Weiner filtering (MatLab wiener2; Lim 1990). A first pass at lossless compression gave compression ratios in between those of PNG and JPEG-LS, the best current standards (for references and discussion see Santa-Cruz et al. 2000). Another potential application is in color reproduction. To enhance color reproduction, video displays have begun using a 4<sup>th</sup> color channel (e.g., the Sharp Quattron). As in our third task, the conditional-moments method could be used to find the Bayes optimal estimate of the 4<sup>th</sup> (yellow) channel values given the RGB values, and hence potentially improve the quality (and accuracy) of standard RGB imagery displayed on a 4-channel device.

Another future direction is to investigate the underlying causes of the complex statistical regularities reported here. For the first two tasks, inspections of the image locations where deviations from the linear prediction are greatest suggest (not surprisingly) that the causes are at least in part related to the properties of well-defined edges (see Figure 7c), but much remains to be done.

### Acknowledgements

We thank Alan Bovik, Daniel Kersten and Jonathan Pillow for helpful comments. We also thank the reviewers for many helpful suggestions. Both authors conceived and designed the study; JSP did the data analysis; WSG wrote the manuscript. Supported by NIH grant EY11747. The methods described in this paper are also described in a provisional patent application 22084-P062V1.

### References

- Bell A. J. & Sejnowski T. J. (1997) The “independent components” of natural scenes are edge filters. *Vision Res.* 37:3327–38.
- Bishop, C.M. (2006) *Pattern recognition and machine learning*. Springer, New York NY.
- Brainard, D. H., Williams, D. R. & Hofer H. (2008) Trichromatic reconstruction from the interleaved cone mosaic: Bayesian model and the color appearance of small spots. *Journal of Vision*, 8(5):15, 1–23.
- Buccigrossi R. W. & Simoncelli E. P. 1999. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Proc.* 8(12):1688–701.
- Burton G. J. & Moorehead I. R. (1987) Color and spatial structure in natural scenes. *Appl. Opt.* 26:157–70
- Daugman J. G. (1989) Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Trans. Biomed. Eng.* 36(1):107–14.
- De Bonet, J.S. (1997) Multiresolution sampling procedure for analysis and synthesis of texture images. *Computer Graphics*. ACM SIGGRAPH.

- Efron, A.A. & Leung, T.K. (1999) Texture synthesis by non-parametric sampling. *IEEE Conference on Computer Vision*. Corfu, Greece.
- Fattal, R. (2007) Image upsampling via imposed edges statistics. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*.
- Freeman, W.T., Thouis R. Jones, T.R. & Pasztor, E.C. (2002) Example-Based Super-Resolution, *IEEE Computer Graphics and Applications*, 22, 56-65.
- Geisler, W. S. (2008) Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59, 167-192.
- Geisler, W.S., Perry, J.S., Super, B.J., & Gallogly, D.P. (2001) Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41, 711-724.
- Glasner, D., Bagon, S. & Irani, M. (2009) Super-resolution from a single image. *International Conference and Computer Vision (ICCV)*.
- Hirakawa, K. & Parks, T. W. (2005) Adaptive homogeneity-directed demosaicing algorithm. *IEEE Transactions on Image Processing*, 14, 360-369.
- Hofer, H, Singer, B. & Williams, D. R. (2005) Different sensations from cones with the same photopigment. *Journal of Vision*, 5(5), 444-454.
- Ing A. D., Wilson A. J. & Geisler, W. S. (2010) Region grouping in natural foliage images: Image statistics and human performance. *Journal of Vision*. 10(4):10, 1-19.
- Karklin, Y. & Lewicki, M. S. (2008) Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457, 83-87.
- Kenny, J. F. & Keeping, E. S. (1951) *Mathematical Statistics, Pt. 2, 2<sup>nd</sup> Ed.* Princeton NJ: Van Nostrand.
- Kersten, D. Predictability and redundancy of natural images. (1987) *J. Opt. Soc. Am. A*, 4, 2395-2400.
- Laughlin, S. B. (1981) A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C* 36:910-12
- Lee, A. B., Pedersen, K. S. & Mumford, D. (2003) The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision* 54(1/2/3), 83–103.
- Li, X., Gunturk, B. & Zhang, L. (2008) Image demosaicing: a systematic survey. *Visual Communications and Image Processing* 2008, Proceedings of the SPIE, Volume 6822, pp. 68221J-68221J-15. San Jose, CA, USA
- Li, Y. & Adelson E. (2008) Imaging mapping using local and global statistics. *Proceedings of SPIE-IS&T Electronic Imaging*, 6806, 680614.
- Lim, Jae S., (1990) *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ, Prentice Hall.
- Mairal, J., Elad, M., & Sapiro, G. (2008) Sparse representation for color image restoration. *IEEE Transactions on Image Processing*.
- Maison B., & Vandendorpe, L. (1998) Nonlinear Stochastic Image Modeling by Means of Multidimensional Finite Mixture Distributions, *SPIE Proc. Nonlinear Image Processing IX*, San Jose CA, USA.
- Olshausen, B. A. & Field, D. J. (1997) Sparse coding with an overcomplete basis set: a strategy by V1? *Vis. Res.* 37:3311–25
- Oruç, I., Maloney, L. T., & Landy, M. S. (2003), Weighted linear cue combination with possibly correlated error, *Vision Research*, 43, 2451-2468.
- Papoulis, A. (1984) *Probability, Random Variables and Stochastic Processes*, 2<sup>nd</sup> Ed. New York: McGraw-Hall.

- Petrov, Y. & Zhaoping, L. (2003) Local correlations, information redundancy, and sufficient pixel depth in natural images. *J. Opt. Soc. Am. A* 20, 56-66.
- Portilla, J. & Simoncelli, E. P. (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40(1), 49–71.
- Ruderman, D. L. & Bialek, W. (1994) Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.* 73, 814–17.
- Ruderman, D. L., Cronin, T. W. & Chiao, C. (1998) Statistics of cone responses to natural images: implications for visual coding. *J. Opt. Soc. Am. A* 15, 2036–2045.
- Simoncelli, E. P. & Olshausen, B. A. (2001) Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24:1193–216.
- Santa-Cruz, D., Ebrahimi, T., Askelof, J., Larsson, M. & Christopoulos (2000) JPEG 2000 still image coding verus other standards. Proceedings of SPIE 4115, 446-454.
- Tkačik, G., Prentice, J., Victor, J. D. & Balasubramanian, V. (2010) Local statistics in natural scenes predict the saliency of synthetic textures. *Proc. Natl. Acad. Sci. USA* 107, 18149–18154.
- van Hateren, J. H. & van der Schaaf, A. (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. London B Biol. Sci.* 265:359–66.
- Yang, J., Wright, J., Ma, Y. & Huang, T. (2008) Image super-resolution as sparse representation of raw image patches. *IEEE Computer Vision and Pattern Recognition*, vol. 1, pp. 1-8.
- Yuille, A. L., & Bulthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference*. Cambridge: Cambridge University Press.
- Zhang, X. & Brainard, D. H. (2004). Estimation of saturated pixel values in digital color imaging. *Journal of the Optical Society of America A*, 21, 2301-2310.
- Zhu, S.C. & Mumford, D. (1997) Prior learning and Gibbs reaction-diffusion. *IEEE Trans. Pattern Anal. Machine Intell.* 19 (11), 1236–1250.
- Zhu S. C., Shi K. & Si, Z. (2010) Learning explicit and implicit visual manifolds by information projection. *Pattern Recognition Letters*, 31, 667-685.

## Appendix

### 1. Gaussian and log Gaussian models for optimal interpolation between two points

Here we derive the optimal estimate of the central point in optimal interpolation under the assumption that the underlying distributions are Gaussian and log Gaussian. A conditional Gaussian distribution  $p(\mathbf{x}_a | \mathbf{x}_b)$  has mean value which is a linear function of  $\mathbf{x}_b$  and a covariance that is independent of the value of  $\mathbf{x}_b$ . Therefore, in the case of interpolation between two points the optimal estimate is given by

$$\hat{x}_{opt} = w_1 s + w_2 t + w_0$$

However, there are two empirical constraints that also hold in our case. First, when  $s = t$  we always find (in agreement with intuition) that

$$\hat{x}_{opt} = s = t$$

which implies that  $w_0 = 0$ . Second, we observe (in agreement with intuition) symmetry about the diagonal (see Figure 1b), which requires that

$$w_1 = w_2$$

which implies that  $w_1 = w_2 = \frac{1}{2}$  and hence  $\hat{x}_{opt} = \frac{s+t}{2}$

The same relations hold for a log Gaussian distribution, except  $\log \hat{x}_{opt} = \frac{\log s + \log t}{2}$  or  $\hat{x}_{opt} = \sqrt{st}$  (the geometric average).

## 2. Implications of standard transformations of training signals

Two of the standard preprocessing steps in measuring natural signal statistics are: (a) subtract off the mean of each training signal to obtain a collection of zero mean signals or (b) subtract off the mean and then divide by the mean to obtain a collection of normalized zero-mean signals (e.g., signed contrast signals). Even when these steps are not carried out explicitly, they are frequently an implicit component of statistical analyses. These steps are intuitively sensible because they remove seemingly irrelevant information when the goal is to characterize the statistics of signal shape. However, they correspond to invariance assumptions that can lead to missed information. To see this, consider the case of measuring the expected value function for the central pixel given the two neighboring pixel values:  $f(s, t) = E(x|s, t)$  (see Figure 3a and 3b).

### 2.1 Transforming training signals to zero mean

Let the  $(s_i, x_i, t_i)$  be an arbitrary original training signal and let  $m_i$  be the local mean  $(s_i + t_i)/2$  that is subtracted off. The transformed values of  $s_i$ ,  $x_i$ , and  $t_i$  are given by

$$s'_i = s_i - m_i, \quad x'_i = x_i - m_i, \quad t'_i = t_i - m_i$$

In this case, for all possible learning algorithms

$$\text{if } (s'_i, t'_i) = (s'_j, t'_j) \text{ then } E(x'|s'_i, t'_i) = E(x'|s'_j, t'_j)$$

What is the set of  $(s_j, t_j)$  for which  $(s'_i, t'_i) = (s'_j, t'_j)$ ? For any given  $(s_i, t_i)$ , the  $(s_j, t_j)$  must satisfy the following two equations:

$$s_j - m_j = s_i - m_i, \text{ and } t_j - m_j = t_i - m_i$$

Hence,

$$s_j - t_j = s_i - t_i$$

This implies that subtracting off the mean is a valid preprocessing step if and only if

$$E(x'|s + \alpha, t + \alpha) = E(x'|s, t) \forall \alpha$$

If this property were true in natural images, then the plot of  $\hat{\Delta x}_{opt}(s, t)$  in Figure 3b would have a constant value along any line parallel to the central diagonal, which does not hold except for the central diagonal itself. In other words, applying the preprocessing step of subtracting the mean would miss much of the important statistical structure shown in Figure 3b. The same result holds if the subtracted mean is  $(s_t + x_i + t_i)/3$ .

## 2.2 Transforming training signals to a normalized zero mean

Now suppose the transformed training stimuli are given by

$$s'_i = \frac{s_i - m_i}{m_i}, \quad x'_i = \frac{x_i - m_i}{m_i}, \quad \text{and} \quad t'_i = \frac{t_i - m_i}{m_i}$$

Again, for all possible learning algorithms

$$\text{if } (s'_i, t'_i) = (s'_j, t'_j) \text{ then } E(x'|s'_i, t'_i) = E(x'|s'_j, t'_j)$$

What is the set of  $(s_j, t_j)$  for which  $(s'_i, t'_i) = (s'_j, t'_j)$ ? For any given  $(s_i, t_i)$  the  $(s_j, t_j)$  must satisfy the following two equations:

$$\frac{s_j - m_j}{m_j} = \frac{s_i - m_i}{m_i}, \quad \text{and} \quad \frac{t_j - m_j}{m_j} = \frac{t_i - m_i}{m_i}$$

Hence,

$$s_j = \frac{m_j}{m_i} s_i, \quad \text{and} \quad t_j = \frac{m_j}{m_i} t_i$$

This implies that subtracting off and then dividing by the mean is a valid preprocessing step if and only if

$$E(x'|\alpha s, \alpha t) = E(x'|s, t) \forall \alpha$$

If this property were true in natural images, then the plot of  $\hat{\Delta x}_{opt}(s, t)$  in Figure 3b would have a constant value along any line passing through the origin, which does not hold except for the central diagonal itself. Thus, applying this preprocessing step would also miss much of the important statistical structure. A similar result holds if the subtracted mean is  $(s_t + x_i + t_i)/3$ .