

# QUANTIFYING THE USE OF STRUCTURE IN COGNITION

*David M. Rouse and Sheila S. Hemami*

Visual Communications Lab, School of Electrical and Computer Engineering,  
Cornell University, Ithaca, NY 14853

## ABSTRACT

Modern algorithms that process images to be viewed by humans analyze the images strictly as signals, where processing is typically limited to the pixel and frequency domains. The continuum of visual processing by the human visual system (HVS) from signal analysis to cognition indicates that the signal-processing based model of the HVS could be extended to include some higher-level, structural processing.

A preliminary experiment was conducted to study the relative importance of structural (higher-level) and signal-based (lower-level) representations of natural images in a cognitive task. The results from this experiment suggest that signal-based representations are only meaningful to human observers when the proportion of high frequency content, which conveys structural information, exceeds a seemingly fixed proportion. This work investigates the results of this experiment towards the development of a rudimentary measure of visual entropy.

## 1. INTRODUCTION

Modern algorithms that process images to be viewed by humans analyze the images strictly as signals, where processing is typically limited to the pixel and frequency domains. The human visual system (HVS) initially analyzes the images as signals and ends in cognition. Cognition is believed to occur gradually while the input is processed by the HVS, such that no single portion of the visual system transforms the internal representation to an abstract concept. Several models consistent with the signal analysis portion of the HVS decompose the input according to frequency, orientation, and contrast to suggest neural responses to primitive image components. Presently, no models exist that incorporate higher-level image characteristics to accompany the established signal analysis of images. The continuum of visual processing from signal analysis to cognition indicates that the signal-processing based model of the HVS could be extended to include some higher-level, structural processing.

When an image is distorted, the contrast of the distortion often is related to the distortions perceived by a human observer. The distorted image may contain at least two forms of distortion. Spatially uncorrelated distortions may be present, such as additive white noise, or spatially correlated distor-

tions, induced by quantizing the wavelet coefficients at all scales. Disrupting an image's structure alone has been shown to exhibit a perceived distortion that is four times greater than the distortion incurred by only adding white noise with the same distortion contrast [1]. The results from [1] suggest that a degradation in visual fidelity highly depends upon the observer's ability to process the image's visual structure. Furthermore, this indicates that the quality of an image's visual structure relates to its ability to convey visual information to the observer.

This paper describes an experiment to study the relative importance of structural (higher-level) and signal-based (lower-level) representations of natural images in a cognitive task. This paper has the following organization: Section 2 defines and discusses structural and signal-based representations of natural images. The methods and stimuli used in the experiments are described in Section 3. The results from the experiment are presented and discussed in Section 4. Conclusions are presented in Section 5.

## 2. STRUCTURAL AND SIGNAL-BASED REPRESENTATIONS

Given the evidence from the experiments in [1] that human observers are more sensitive to structural distortions than uncorrelated distortions in an image, it is of great interest to consider an observer's use of structure in a cognitive task. Toward this end, two representations of a natural image are formed: structural-based and signal-based.

Structural representations lack the fine details present in the original image while preserving the overall organization necessary to recognize the content. A simple line drawing of a natural image retains the visual structure of the original image. Ideally, line drawings represent the pertinent contours and, possibly, impart three-dimensional information to facilitate recognition of the scene [2]. A structurally-based image sequence evolves from a simple line drawing toward a far more elaborate line drawing. For example, such a sequence of a house would begin with an image depicting the house's coarse shape and gradually include more lines depicting the house's finer details.

Signal-based representations coincide with the initial processing of the human visual system, which decomposes the

image according to frequency, orientation, and contrast. A signal-based image sequence evolves from a blurry, distorted version to sharp, visually flawless version.

For either representation, subsequent images in a sequence reveal additional detail or information from the original image. By approximating the visual information by the entropy or bitrate of the corresponding image, the sequences also evolve according to increasing bitrate  $R$ .

### 3. METHODS

A preliminary experiment was conducted to determine the bitrate corresponding to an observer's recognition of the original image content when viewing either its signal-based or structural representation.

#### 3.1. Stimuli

Nine grayscale natural images of size  $512 \times 512$  pixels were cropped from original natural images. The content captured by the images was simple to facilitate analysis of responses by observers.

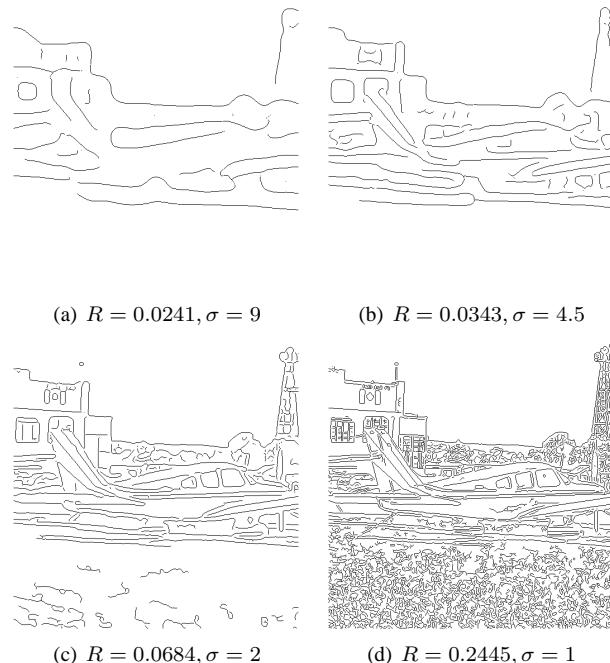
##### 3.1.1. Structurally-based Stimuli

Object structure is widely believed to rely on the perception of image details, such as sharp edges, which are conveyed by energy at high spatial frequencies [3, 4]. Edges, defined spatially by sudden intensity changes, may be identified by either the presence of an absolute maximum in the first derivative of an image or a zero-crossing in its second derivative [4]. Structural representations of natural images for this experiment were generated with the Canny edge detector [5]. This method filters the image with the derivative of a Gaussian specified for a particular  $\sigma > 0$  and applies thresholding to generate a binary image. The parameter  $\sigma$  in the Canny filter controls the suppression of high frequency energy before detecting edges. Decreasing  $\sigma$  preserves high frequency content, and the resulting structural representation will correspond to finer image details. The bitrate  $R$  of a structural representation was determined by compressing the binary image with a JBIG coder.

The image sets of structural representations for each natural image were created by varying  $\sigma$  in the Canny edge detector from 0.5 to 10 with increments of 0.5, where  $\sigma$  varies inversely with respect to the bitrate. Select images from a sequence of structural representations of the natural image *gray02* are shown in Figure 1.

##### 3.1.2. Signal-based Stimuli

From a signal analysis perspective, suppressing energy at specific spatial frequencies and orientations removes image content. The discrete wavelet transform (DWT) establishes a multiresolution representation by decomposing an image into subbands varying in spatial frequency and orientation. An



**Fig. 1.** Selected images from the sequence of structural-based representations of *gray02*. The bitrate  $R$  using the JBIG coder and the parameter  $\sigma$  for the Canny edge detector are included.

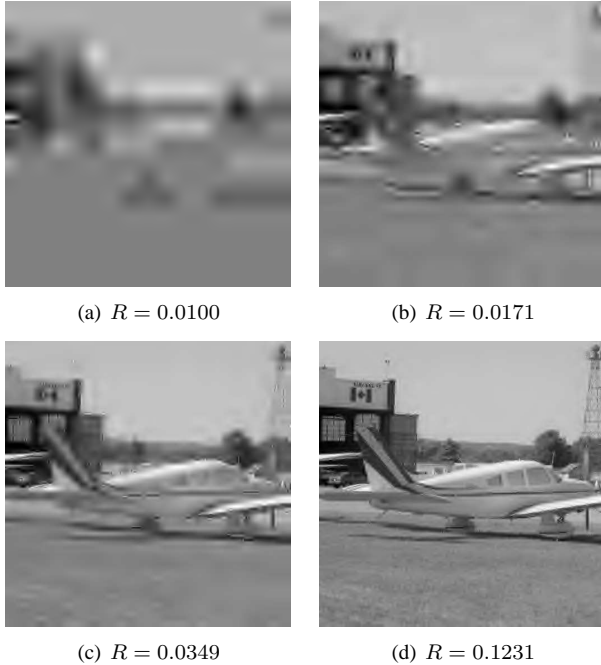
image  $I$  is then compressed by quantizing the coefficients in each subband.

To generate the signal-based representations in this experiment, the dynamic contrast-based quantization (DCQ) algorithm [6] is utilized. The DCQ algorithm dynamically computes quantization step-sizes for each subband such that the induced distortions  $E$  exhibit a specified RMS contrast. The RMS contrast is defined as

$$C_{rms} = \frac{1}{\mu_{L_I}} \sqrt{\frac{1}{M} \sum_{i=1}^M (L_E[i] - \mu_{L_E})^2}, \quad (1)$$

where  $\mu_{L_I}$  denotes the average luminance of the original image  $I$ ,  $\mu_{L_E}$  the average luminance of the distortions  $E$ ,  $L_E[i]$  the luminance of the  $i^{th}$  pixel of  $E$ , and  $M$  is the total number of pixels.

Given a specified bitrate  $R$ , a measure of visual distortion is calculated, which the DCQ algorithm utilizes to generate subband quantization step-sizes. In addition, the DCQ strategy incorporates the property of global precedence [6]. The principle of global precedence contends that the HVS temporally processes a visual scene in a global-to-local order [7]. Thus, the DCQ algorithm discards subband coefficients in a fine-to-coarse order. Furthermore, by directly specifying the visual distortion to one, DCQ generates a very compressed yet visually lossless image. A visually lossless image is visually indistinguishable from the original.



**Fig. 2.** Selected images from the sequence of signal-based representations of *gray02* with the corresponding bitrate  $R$  using the JPEG-2000 coder supplied with DCQ step-sizes.

Signal-based representations were produced by quantizing wavelet subbands obtained by transforming a natural image using  $N = 5$  decomposition levels and the 9/7 biorthogonal DWT filters. The bitrate  $R$  of a signal-based representation was computed by compressing the image using a JPEG-2000 coder supplied with DCQ step-sizes for each subband. An iterative bisection search was performed to determine the step-sizes for a specific bitrate.

The series of images in each set of signal-based representations corresponded to a set of encoding rates, which were logarithmically equally spaced between 0.0112 and 0.3020. The choice of extremely low rates guaranteed images such that the recognition is questionable. Select images from a sequence of signal-based representations of the natural image *gray02* are shown in Figure 2. By virtue of degrading the original natural image by quantizing the wavelet coefficients with the DCQ algorithm, an image reconstructed at a very low rate, though unrecognizable, is noted to exhibit the spatial organization of the original image.

### 3.2. Procedure

For each of the nine images, sequences of signal-based and structural representations were generated as described. Observers viewed image sequences corresponding to one of the two representations for several of the nine natural images. Not every observer was available to view a representation for each of the natural images. For each image in a sequence, the ob-

server was asked to provide a description of the image content. The next image in the sequence was shown upon submission of a description; a time limit was not imposed.

### 3.3. Participants

Twenty-six observers with normal or corrected-to-normal acuity participated in this preliminary experiment. Each series of representations was viewed by at least 9 observers and at most by 14. On average the structural representations and signal-based representations were viewed by 12.1 and 11.8 observers, respectively.

## 4. RESULTS AND DISCUSSION

An observer's point of recognition was identified when the description contained both adequate and accurate information to briefly describe the image content. For all nine images, all observers identified the image content before viewing the signal-based representation with the largest bitrate. However, several observers did not recognize the image content for three of the structural representation sequences.

The average bitrate  $m_R$  corresponding to the initial recognition was noted for both the structural and signal-based representations, and the standard deviation  $s_R$  of the initial recognition bitrates were computed for each representation. Columns two through five of Table 1 summarize these statistics. Larger standard deviations reflected a difficulty in recognizing the content. For both representations, the three largest standard deviations corresponded to the same three original natural images, which are emphasized by an asterisk beside the image tag. In addition, the structural representations of these three images were not recognized by several observers who viewed the corresponding sequence. These findings suggest that a difficulty in content recognition for one representation predicts a similar difficulty for the other representation.

A noted shortcoming for the sequences of structural representations using the Canny edge detector by varying  $\sigma$  is apparent when examining Figure 1. Without prior knowledge of the image content, the representations in Figures 1(a) and 1(b) do not provide adequate information to facilitate recognition of the content. While varying  $\sigma$  captures edges a different scales, larger values of  $\sigma$  also smooth the contours indicating the edges as observed in Figure 1. Specifying  $\sigma = 1$  captures the desired structural content in addition to other undesirable content (e.g. the texture of the grass beneath the plane). Varying the thresholds when  $\sigma = 1$  can control the amount of undesired content that appears in the structural representation. Additional experiments will need to be performed to analyze the impact of varying the thresholds of the Canny edge detector for a fixed  $\sigma$ .

The visually lossless bitrate  $R_{VL}$  for an image provides a reference bitrate where recognition is unquestionable. For structural representations, the definition of a visually lossless image is incompatible, since a reference line drawing is not

**Table 1.** Statistical and Data Analysis Corresponding to the Initial Recognition According to Natural Image and Representation

Natural Image Tag	Structural		Signal-Based					
	$m_R$	$s_R$	$m_R$	$s_R$	$m_R/R_{VL}$	$R_{VL}$	$m_{R_{noLL}}/m_R$	$m_{R_{noLL}}$
gray06*	0.2119	0.1172	0.0729	0.0241	0.0456	1.5999	0.9064	0.0661
gray22*	0.2093	0.0840	0.1166	0.0599	0.0641	1.8181	0.9422	0.1098
gray07*	0.2747	0.0814	0.0510	0.0268	0.0164	3.1091	0.8866	0.0452
gray02	0.0838	0.0539	0.0348	0.0115	0.0188	1.8479	0.8546	0.0298
gray11	0.0625	0.0529	0.0341	0.0086	0.0203	1.6824	0.8247	0.0281
gray10	0.0914	0.0482	0.0255	0.0095	0.0124	2.0599	0.8094	0.0206
gray12	0.0738	0.0408	0.0301	0.0086	0.0185	1.6307	0.7978	0.0240
gray13	0.0778	0.0387	0.0380	0.0123	0.0164	2.3203	0.8405	0.0319
gray25	0.0855	0.0142	0.0379	0.0076	0.0243	1.5593	0.8474	0.0321

readily available. Furthermore, it is more appropriate to consider a structural representation that captures the visual information available in the original natural image. Such an image would be information lossless in a visual sense. On the other hand, the DCQ algorithm allows the generation of a visually lossless image for the signal-based representations. Normalizing  $m_R$  for signal-based representations by  $R_{VL}$  specifies the average recognition bitrate as a percentage of the visually lossless bitrate. The normalized  $m_R$  and  $R_{VL}$  are listed in the sixth and seventh columns, respectively, of Table 1. The average normalized recognition bitrate  $m_R/R_{VL}$  for all nine images is 0.0263. This rudimentary relationship may provide a coarse recognition threshold.

The structural importance of the LL band was considered in the signal-based representation. Reconstructing the image sequences using only the LL band generated images lacking sufficient detail for recognition at all bitrates. Considering the observers' viewing distance, this would correspond to frequencies below about 1.15 cpd. The content in the LL band may provide depth information to the observer. When reconstructing the image sequences using all bands except the LL band, the images appeared to convey structure similar to the experiment images. Here energy at spatial frequencies above about 1.15 cpd is preserved. These observations indicate that information valuable for recognition resides in the higher frequency content. The eighth and ninth columns of Table 1 list the percentage of the average recognition rate  $m_R$  when excluding the LL band and the average recognition rate when excluding the LL band  $m_{R_{noLL}}$ , respectively. Excluding the images difficult to recognize, indicated by an asterisk, the proportions span over the small interval of 0.79 to 0.85. The values in column eight of Table 1 support the assertion that more information which is used by observers for recognition is available in the higher frequency content. Note further that the proportion of information is nearly constant with an average of 0.8566 among the nine images.

Signal-based representations were noted to contain very little, if any, energy at higher spatial frequencies at  $m_R$ . Since the Canny filter also suppresses high frequencies, the highest frequency in the signal-based representation may relate to the

highest frequency in the structural representation. Additional investigation and analysis could validate this claim.

## 5. CONCLUSIONS

The results from this preliminary experiment suggest that signal-based representations are only meaningful to human observers when the proportion of high-frequency content, which conveys structural information, exceeds a seemingly fixed proportion. Further investigation will extend this analysis to the structural-based recognition data, and we will attempt to cross-validate the recognition rates for the two representations in an effort to unify the results into a rudimentary measure of visual entropy.

## 6. REFERENCES

- [1] D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," in *Proc. SPIE Human Vision and Electronic Imaging XI*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, Eds., San Jose, CA, Jan. 2006.
- [2] D. Marr, *Vision*. W. H. Freeman and Company, 1982.
- [3] R. L. De Valois and K. K. De Valois, *Spatial Vision*. Oxford University Press, 1990.
- [4] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B*, vol. 207, no. 1167, pp. 187–217, Feb. 1980.
- [5] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.
- [6] D. M. Chandler and S. S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 397–410, Apr. 2005.
- [7] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive Psychology*, vol. 9, pp. 353–383, 1977.