ANALYZING THE ROLE OF VISUAL STRUCTURE IN THE RECOGNITION OF NATURAL IMAGE CONTENT WITH MULTI-SCALE SSIM

David M. Rouse and Sheila S. Hemami

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

ABSTRACT

The natural continuum of visual processing from signal analysis to cognition inspires the search for extensions to current models to include some higher-level structural processing. This paper continues the authors’ study regarding the role of visual structure in cognition, where two representations of a natural image are considered: signal-based and structurally based. Structural representations preserve the overall image organization necessary to recognize the image content and discard the finer details of objects such as textures. Signal-based representations decompose and process an image as a signal in terms of its frequency, orientation, and contrast. A recognition experiment was conducted, where grayscale cartoon-renderings replace the simple line drawings previously selected as structural representations. An analysis between the two types of image representations using the three components of the multi-scale structural similarity (MSSIM) image quality metric identifies signal structure as a crucial component in recognition.

Index Terms—Structural Modeling, Visual Cognition, Image Compression, Suprathreshold Image Quality Comparison

1. INTRODUCTION

When presented with a visual stimulus, the human visual system (HVS) initially analyzes the stimulus as a signal in the primary visual cortex (V1) and through additional, higher-level processing gradually interprets the original stimulus. Common image processing models, consistent with the lower-level signal analysis portion of the HVS, decompose the stimulus according to frequency, orientation, and contrast to suggest neural responses to primitive image components. However, no models currently extend the present signal analysis of images to incorporate higher-level image characteristics that result in cognition. The natural continuum of visual processing from signal analysis to cognition inspires the search for extensions to current models that include some higher-level, structural processing.

This paper extends the authors’ previous investigation regarding the role of visual structure in cognition, where two representations of a natural image are considered: signal-based and structurally based [1]. Grayscale cartoon-renderings replace the simple line drawings as structural representations, which affords a more comparative analysis with our signal-based representations. An analysis using the multi-scale structural similarity (MSSIM) image quality metric [2] identifies image structure as the most relevant feature for recognition among the two image representations considered.

This paper has the following organization: Section 2 reviews the MSSIM image quality metric. Definitions of the structural and signal-based representations tested are provided in Section 3. A discussion of the experimental methods implemented are described in Section 4. Section 5 presents the results from the recognition experiments and an analysis of those results. Extensions with a summary of the present paper appear in Section 6.

2. MSSIM IMAGE QUALITY METRIC

An analysis comparing the structural and signal-based representations is conducted using the multi-scale structural similarity (MSSIM) image quality metric [2]. MSSIM is an extension of the structural similarity (SSIM) image quality metric [3] and aims to quantify the perceptual quality of a distorted image $Y$ when compared to its reference image $X$. Let $x$ and $y$ denote patches from the same local window of the images $X$ and $Y$, respectively. MSSIM quantifies image quality as the product of three components: luminance ($l(x, y)$), contrast ($c(x, y)$), and structure $s(x, y)$. The SSIM index for the patches $x$ and $y$ is given as

$$SSIM(x, y) = \frac{l(x, y)^\alpha \times c(x, y)^\beta \times s(x, y)^\gamma}{(l(x, y)^\alpha + c(x, y)^\beta + s(x, y)^\gamma)^\delta}$$

where $\mu_x$ denotes the mean of $x$, $\sigma_x$ denotes the standard deviation of $x$, $\sigma_{xy}$ is the inner product of the the normalized images $x_{\mu_x}$ and $y_{\mu_y}$, and the $C_k$ are small positive constants. The exponents $\alpha$, $\beta$, and $\gamma$ allow adjustments to the components’ contribution to the overall SSIM index. The overall SSIM image quality index $SSIM(X, Y)$ for two images is computed by averaging the SSIM index values computed for small patches of the two images [3].

A problem with the structure component of SSIM arises when $s$ is zero for either $x$ or $y$. This might occur if the image is constant within the region over which $s$ is evaluated, and as defined above, $s(x, y)$ would be assigned one rather than zero. Since the overall SSIM index is the average of SSIM values for small patches, a value of one rather than zero for $s(x, y)$ inflates the quality index. To remedy this feature, we modify SSIM to force $s(x, y)$ to zero if either $\sigma_x$ or $\sigma_y$ is zero and set the constants $C_1$ to zero. Similarly, we force $l(x, y)$ to zero if $\mu$ is zero for either $x$ or $y$. These modifications revert SSIM to the universal quality index (UQI) introduced in [4] while avoiding stability issues when the denominator is zero [3].

MSSIM extends SSIM by computing the contrast and structure components at $K$ image scales, where the $k^{th}$ scale image corresponds to low-pass filtering and subsampling, by a factor of 2, the original image $(k - 1)$ times. The MSSIM index is given by

$$MSSIM(X, Y) = l_k(X, Y)^{\alpha K} \prod_{k=1}^{K} c_k(X, Y)^{\beta_k} s_k(X, Y)^{\gamma_k},$$

where

$$c_k(X, Y) = \left( \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right)^\alpha \left( \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right)^\beta \left( \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \right)^\gamma,$$

and

$$s_k(X, Y) = \left( \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right)^\alpha \left( \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right)^\beta \left( \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \right)^\gamma,$$

for $k = 1, \ldots, K$, where $\alpha, \beta, \gamma, \alpha_k, \beta_k,$ and $\gamma_k$ are the exponents for the $k^{th}$ scale.
objects serve as the components of structural representations.

A sequence of images for each representation is constructed to evolve from an unrecognizable version to a recognizable version of the original natural image. That is, subsequent images in a sequence contain additional detail or information from the original natural image.

3.1. Signal-based Representations

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

Signal-based representations were generated with the dynamic contrast-based quantization (DCQ) algorithm [5]. 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

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

where $\mu_{L_X}$ denotes the average luminance of the original image $X$, $\mu_{L_E}$ 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 [5]. The principle of global precedence contends that the HVS temporally processes a visual scene in a global-to-local order [6]. Thus, the DCQ algorithm discards subband coefficients in a fine-to-coarse order.

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.

3.2. Structural Representations

Structural representations possessing the properties described are consistent with signals restored via total variation (TV) regularization, an approach that has been applied to the problem of image denoising [7]. Let $g(t)$ be a continuous signal obtained by adding noise to a source signal for $t \in [a, b]$. TV regularization finds a restored signal $f$ from $g$ by solving an optimization problem of the form:

$$\min_f \int_a^b (f(t) - g(t))^2 + \lambda |d/dt f(t)| \, dt,$$

where the first term maintains the similarity between the $f$ and $g$, the second term penalizes deviations from smoothness, and $\lambda$ is a regularization parameter to control the amount of smoothing.

An alternative and equivalent approach to finding $f$ is via soft thresholding of undecimated Haar wavelet coefficients in all subbands except the coarsest LL subband [8]. The equivalence does not remain for 2-D signals, but sufficient visual similarities warrant...
this alternative for images [9]. Soft thresholding with thresholding parameter \( \tau \) is given by

\[
S_\tau(x) = \begin{cases} 
  x - \tau \text{sgn}(x) & |x| > \tau \\
  0 & |x| \leq \tau, 
\end{cases}
\]

where \( \text{sgn}(x) \) is the signum function. Given a thresholding parameter \( \tau \), thresholds are appropriately scaled for wavelet coefficients at each scale. Adjusting the parameter \( \tau \) varies the level of detail removed from the original image, where smaller values of \( \tau \) result in the removal of very few details. On the other hand, larger values of \( \tau \) induce more aggressive smoothing which may simultaneously compromise important structures, too.

4. METHODS

The experiment designed in the previous study [1] was conducted to determine the soft thresholding parameter \( \tau \) corresponding to an observer’s recognition of the original image content when viewing a sequence of structural representations. The sequence evolves from a low-detail cartoon, capturing the basic shapes of the image contents, toward a high-quality version of the original natural image. Image sequences corresponding to structural representations for each natural image were generated by soft thresholding undecimated Haar wavelet coefficients from a 5-level decomposition for various values of the parameter \( \tau \), which was logarithmically equally spaced from 2048 to 1. Two structural representations were tested, differing only with respect to the inclusion of the coarsest LL subband coefficients, to investigate how the absence of the information within the LL subband affects the recognition threshold. Twelve observers viewed one of the two structural representations for each natural image. For each image in a sequence, the observer was asked to provide a description of the image content. The next image in the sequence was shown upon submission of a description. Participants typically completed the experiment in about 30 minutes.

5. RESULTS AND ANALYSIS

5.1. Average Recognition Thresholds

An observer’s point of recognition was identified when the description contained both adequate and accurate information to briefly describe the image content. Comparisons of the average recognition thresholds for the new structural representations with our signal-based representations [1] revealed similar trends among the nine images tested. The average recognition thresholds from the experiments for the two structural representations are summarized in Figure 3, while the results from our previous experiment [1] for signal-based representations generated with the distortion-contrast quantization (DCQ) algorithm [5] are provided for comparison in Figure 4. The recognition thresholds are presented with respect to the logarithm of the parameter varied for the corresponding representation. For the structural representation results, the ordinate axis has been reversed such that larger values of \( \tau \) correspond with lower values of bitrate for the signal-based representations. Between the two representations the images gray06, gray07, and gray22 have higher recognition thresholds than the remaining six images tested.

Results from psychological studies emphasize the value of higher spatial frequency content of visual information in recognition tasks [10, 11, 12]. Among the two structural representations tested, there is no significant difference in the recognition threshold when the LL subband is discarded, which corroborates the authors’ previous hypothesis that discarding LL subband coefficients would not significantly alter the recognition thresholds [1].

5.2. MSSIM Analysis of Recognition Threshold Images

Among the three components of MSSIM, the structure component, which quantifies signal structure in terms of the inner product of \( x \) and \( y \), accounts best for the correspondence between the two image representations at their recognition threshold. This correspondence reinforces the premise that human observers rely on structural features to recognize image content. A representative example illustrating this conclusion is provided in Figure 5 for the natural image airplane. In Figure 5, the MSSIM index value along with its contrast and structure components\(^1\) are shown when comparing the average recognition threshold of the signal-based representation to each image in the sequence of structural representations. The average recognition threshold for the structural representations is de-

\(^{1}\)The luminance component was approximately 1 for every comparison, so it has been omitted from the figure.
5.3. Recognition Thresholds and MSSIM Components

The preceding MSSIM analysis identifies a relationship between signal structure and recognition, and it follows to consider the capacity in which MSSIM predicts recognition thresholds for an original image. Denote $X$ as an original image and $Y$ as the recognition threshold image for a particular representation. Table 1 summarizes the average and sample standard deviation of the contrast and structure components of MSSIM for the recognition thresholds for both representations of the nine natural images tested. The average value for the structure component differs the least between the two representations. Furthermore, the images $gray06$ and $gray22$ account for the larger standard deviation in $s(X, Y)$ and have larger values ($\approx 0.8$) for $s(X, Y)$ than the other seven natural images. Removing these two images yields averages $0.61$ and $0.59$ with standard deviations $0.05$ and $0.04$ for the signal-based and structural representations, respectively.

The larger structure values for the images $gray06$ and $gray22$ result from the presence of more non-zero wavelet coefficients at finer scales relative to the other images at their recognition thresholds. A closer investigation of the signal-based representation’s wavelet coefficients will disclose their impact on image structure. For example, a side-by-side visual inspection of each representation’s recognition threshold revealed that discarding the wavelet coefficients in the finest 2 or 3 scales in a 5-level DWT decomposition produced negligible visual distortions without disrupting the perceived visual structure. In fact, discarding the finer scale coefficients in the signal-based representations removed unnatural visual artifacts. Models cognizant of a natural image’s structural features will yield algorithms that avoid such egregious and distracting artifacts at very low bitrates.

6. SUMMARY AND FUTURE WORK

Cartoon-renderings as structural representations facilitate a comparative analysis with our signal-based representations of natural images. The analysis with the structure component of MSSIM illustrates that image structure is a crucial component when interpreting an image. The discrepancy between the peak of the MSSIM structure component and the recognition thresholds demonstrates that signal structure may not be identical to the perceived structure. Understanding this disparity would reveal techniques toward developing a perceptual image quality metric between images contaminated with different suprathreshold distortions.

7. REFERENCES


