AN ADAPTIVE LINEAR SYSTEM FRAMEWORK FOR IMAGE DISTORTION ANALYSIS

Zhou Wang and Eero P. Simoncelli

Lab for Computational Vision, New York University, New York, NY 10003
Email: zhouwang@ieee.org, eero.simoncelli@nyu.edu

ABSTRACT

We describe a framework for decomposing the distortion between two images into a linear combination of components. Unlike conventional linear bases such as those in Fourier or wavelet decompositions, a subset of the components in our representation are not fixed, but are adaptively computed from the input images. We show that this framework is a generalization of a number of existing image comparison approaches. As an example of a specific implementation, we select the components based on the structural similarity principle, separating the overall image distortions into non-structural distortions (those that do not change the structures of the objects in the scene) and the remaining structural distortions. We demonstrate that the resulting measure is effective in predicting image distortions as perceived by human observers.

1. INTRODUCTION

Signal analysis often involves decomposing a given signal into a linear combination of basic components. For discretely-sampled finite-length signals, we can write:

$$\mathbf{x} = \mathbf{L}\,\mathbf{c} = c_1\mathbf{l}_1 + c_2\mathbf{l}_2 + \cdots + c_M\mathbf{l}_M\,, \tag{1}$$

where $\mathbf{x}$ is an N-dimensional column vector representing the given signal, $\mathbf{L}$ is an $N \times M$ matrix whose columns $\{\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_M\}$ are the components, and $\mathbf{c}$ is an M-dimensional column vector containing a set of weighting coefficients, each associated with one component. Classical examples include Fourier and wavelet decompositions, in which the components are sinusoids and localized bandpass filters, respectively. Although such a linear decomposition paradigm has achieved great success in a wide variety of signal and image processing applications, the need for nonlinear signal representations is apparent in many problems. In particular, from the viewpoint of image statistics, no matter how the linear system $\mathbf{L}$ is designed, it seems impossible to achieve the desirable property of statistical independence between the coefficients in $\mathbf{c}$, because natural image signals are generally not linear combinations of independent sources (occlusion provides a clear counterexample).

Several methods have been proposed that depart from the basic linear system framework, while retaining some of its elegance and simplicity. One approach uses a large overcomplete "dictionary" of components [1], from which a subset are selected in an adaptive signal-dependent manner. The selection is optimized for some desired properties such as sparseness or independence. Other approaches incorporate an additional nonlinear stage after the linear system. For example, it has been shown that a divisive normalization procedure can substantially reduce the dependencies between the wavelet coefficients [2, 3].

Here we describe a different approach, in which the components (or a subset of the components) are not fixed, but are adaptively computed from the input signals. We find that this method is particularly effective in analyzing the distortions between two image signals. Specifically, we select the components based on the structural similarity principle [4]. The basic assumption is that the human visual system is designed to extract structural information from the visual scene, and it is thus sensible to separate the non-structural distortions (those that do not change the structures of the objects in the scene) and the remaining structural distortions.

This idea is demonstrated in Fig. 1, where the overall distortion signal (c) between the original image (a) and the distorted image (b) is separated into non-structural and structural distortions (d) and (e). The separation is implemented by decomposing (c) into a linear combination of two sets of components. The first set {(f), (g), ..., (h)} are considered "non-structural", in the sense that adding them into the original image (a) has little effect on the structural information of the image, as can be observed in image (l). But this is clearly not true of the second set {(i), (j), ..., (k)}, as can be observed in image (m). Fig. 1 also illustrates that the non-structural image distortion components are generally not fixed, but must be adaptively computed from the input image signals.

2. ADAPTIVE LINEAR SYSTEM FRAMEWORK

2.1. General framework

Given two signals $\mathbf{x}$ and $\mathbf{y}$, we decompose the distortion signal $\Delta\mathbf{x} = \mathbf{y} - \mathbf{x}$ using an adaptive linear system $\mathbf{L}(\mathbf{x}, \mathbf{y})$:

$$\Delta\mathbf{x} = \mathbf{L}(\mathbf{x}, \mathbf{y})\,\Delta\mathbf{c}\,. \tag{2}$$

In general, the distortion measure between $\mathbf{x}$ and $\mathbf{y}$ may then be defined as a function of the coefficients $\Delta\mathbf{c}$. Specifically, we define the distortion measure as the minimum of a weighted square energy function:

$$D(\mathbf{x}, \mathbf{y}) = \min_{\Delta\mathbf{c}} \|\mathbf{W}(\mathbf{x}, \mathbf{y})\Delta\mathbf{c}\|^2\,, \tag{3}$$

where $\|\cdot\|$ is the l_2 norm of a vector and $\mathbf{W}(\mathbf{x}, \mathbf{y})$ is a diagonal matrix, in which the *i*-th diagonal entry is the weight assigned to the *i*-th coefficient Δc_i. As with $\mathbf{L}(\mathbf{x}, \mathbf{y})$, these weights may also be adaptively computed from $\mathbf{x}$ and $\mathbf{y}$. For notational convenience, we replace $D(\mathbf{x}, \mathbf{y})$, $\mathbf{L}(\mathbf{x}, \mathbf{y})$ and $\mathbf{W}(\mathbf{x}, \mathbf{y})$ with D, $\mathbf{L}$ and $\mathbf{W}$, respectively, for the remainder of the paper.

Finding the optimal $\Delta\mathbf{c}$ is a least square optimization problem. When the representation is overcomplete, it can be solved using a Lagrange multiplier approach. Substituting the solution back into Eq. (3), we obtain

$$D = \|\mathbf{W}^{-1}\mathbf{L}^T(\mathbf{L}\mathbf{W}^{-2}\mathbf{L}^T)^{-1}\Delta\mathbf{x}\|^2\,. \tag{4}$$

To appear in *Proc. IEEE Inter. Conf. Image Proc., Genoa, Italy, Sept. 11-14, 2005*

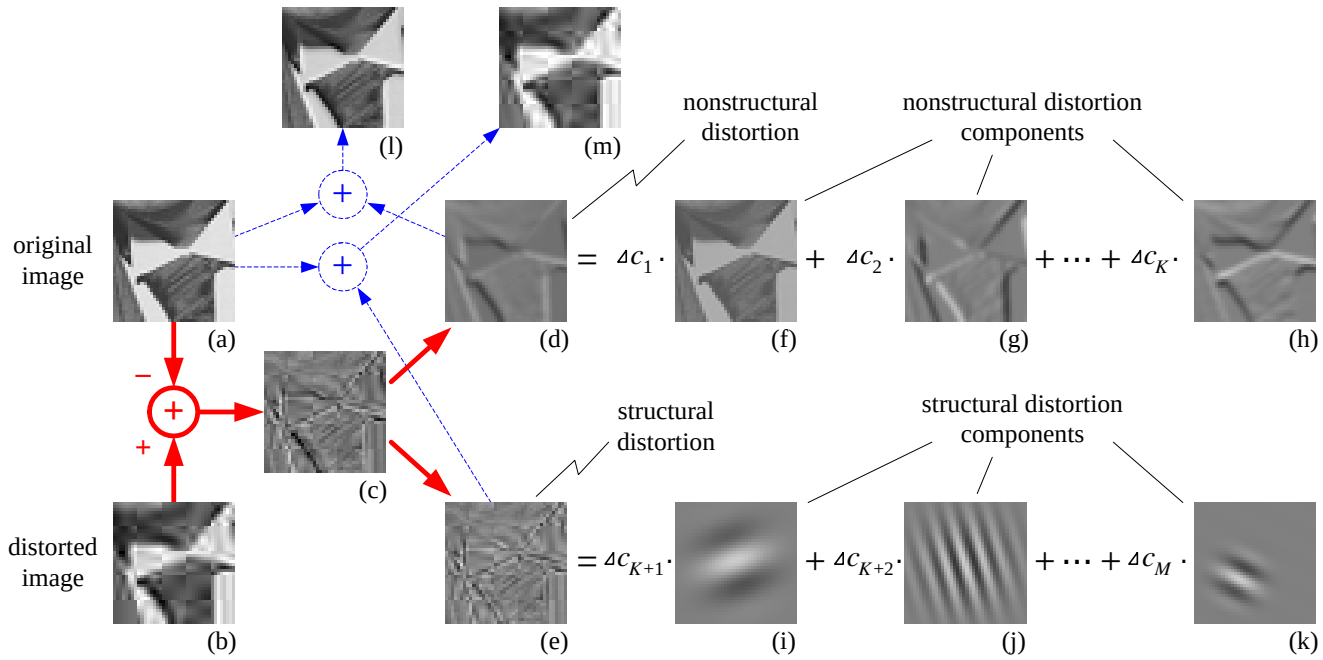

Fig. 1. Separation of structural and non-structural distortions using an adaptive linear system.

This distortion measure is general, and relies only on the invertability of matrices $\mathbf{W}$ and $\mathbf{L}\mathbf{W}^{-2}\mathbf{L}^T$. In practice, however, such computation can be very expensive because it requires inverting an $M \times M$ matrix $\mathbf{L}\mathbf{W}^{-2}\mathbf{L}^T$. Even worse, when the distortion measure is applied locally in an image (as in our implementation, see Section 3), such matrix inversion needs to be computed at every spatial location because both $\mathbf{L}$ and $\mathbf{W}$ may vary across the image.

To simplify the computation, we divide $\mathbf{L}$ into two parts: $\mathbf{L} = [\mathbf{A}\ \mathbf{B}]$, where $\mathbf{B}$ contains a complete generic (non-adaptive) basis for the space of all images ($\mathbf{B}$ is $N \times N$ and full rank) and $\mathbf{A}$ contains $M - N$ adaptive components. Correspondingly, the coefficient vector $\Delta\mathbf{c}$ and the weighting matrix $\mathbf{W}$ are divided into $\Delta\mathbf{c}_A$ and $\Delta\mathbf{c}_B$, and $\mathbf{W}_A$ and $\mathbf{W}_B$, respectively. Thus we can write

$$\Delta\mathbf{x} = \mathbf{L}\,\Delta\mathbf{c} = \mathbf{A}\Delta\mathbf{c}_A + \mathbf{B}\Delta\mathbf{c}_B\,, \tag{5}$$

$$\|\mathbf{W}\Delta\mathbf{c}\|^2 = \|\mathbf{W}_A\Delta\mathbf{c}_A\|^2 + \|\mathbf{W}_B\Delta\mathbf{c}_B\|^2\,. \tag{6}$$

From Eq. (5), we have $\Delta\mathbf{c}_B = \mathbf{B}^{-1}(\Delta\mathbf{x} - \mathbf{A}\Delta\mathbf{c}_A)$. Substituting this into Eq. (6), setting the partial derivative of $\|\mathbf{W}\Delta\mathbf{c}\|^2$ with respect to $\Delta\mathbf{c}_A$ to zero, and solving for $\Delta\mathbf{c}_A$, we obtain the least square solution (the convex formulation of Eq. (6) ensures a unique minimum):

$$\Delta\hat{\mathbf{c}}_A = (\mathbf{W}_A^2 + \mathbf{A}^T\mathbf{G}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{G}\Delta\mathbf{x}\,, \tag{7}$$

where $\mathbf{G} = \mathbf{B}^{-T}\mathbf{W}_B^2\,\mathbf{B}^{-1}$. Finally, we can write the overall distortion measure as:

$$D = \|\mathbf{W}_A\Delta\hat{\mathbf{c}}_A\|^2 + \|\mathbf{W}_B\,\mathbf{B}^{-1}(\Delta\mathbf{x} - \mathbf{A}\Delta\hat{\mathbf{c}}_A)\|^2\,. \tag{8}$$

Compared with Eq. (4), the advantage of this solution is that we only need to adaptively invert an $(M - N) \times (M - N)$ matrix (for the calculation of $\Delta\hat{\mathbf{c}}_A$ in Eq. (7)). Since $\mathbf{B}$ is non-adaptive, $\mathbf{B}^{-1}$ and $\mathbf{G}$ only need to be computed once.

2.2. Relationship to existing methods

A number of existing image comparison methods may be described within the context of this general framework, each associated with a particular choice of $\mathbf{L}$ and $\mathbf{W}$.

Mean squared error (*MSE*) is the most widely used image distortion measure. It is non-adaptive and corresponds to the simple case that $\mathbf{L} = \mathbf{W} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix.

Space/frequency weighting: The key idea of this approach is that the visual error sensitivity of different components are different and therefore should be given different weights. Usually the weights do not change with the input image signal, so the method is still non-adaptive. Depending on the linear transform, the weights (given by $\mathbf{W}$) may be space-variant (when $\mathbf{L} = \mathbf{I}$) [5], frequency-variant (when $\mathbf{L}$ is of Fourier-type) [6], or jointly space/frequency-variant (when $\mathbf{L}$ is of wavelet-type) [7]. This type of method has been widely used in the design of transform-based image coders such as JPEG and JPEG2000.

Transform domain masking: Masking refers to the psychophysical/physiological phenomenon that the visual sensitivity of one image component is surpressed by the other components that are close in space, frequency and orientation. Therefore, the weight given to one component should be adjusted according to the strength of the other components (e.g., [8–10]). We consider this method to be partially adaptive because although $\mathbf{W}$ is adaptively adjusted with the input signal, the component matrix $\mathbf{L}$ is not.

The tangent distance [11] has been successfully used in pattern recognition, especially digit recognition applications. The idea is to define a set of operations that are invariant for pattern recognition tasks, e.g., small spatial translation and rotation. A set of tangent vectors are computed by taking derivatives of the image with respect to these invariant variables. Let $\mathbf{A} = [\mathbf{d}_1(\mathbf{x})\ \mathbf{d}_2(\mathbf{x})\ \cdots\ -\mathbf{d}_1(\mathbf{y})\ -\mathbf{d}_2(\mathbf{y})\ \cdots]$, where $\{\mathbf{d}_1(\mathbf{x}), \mathbf{d}_2(\mathbf{x}), \cdots\}$ and $\{\mathbf{d}_1(\mathbf{y}), \mathbf{d}_2(\mathbf{y}), \cdots\}$ are the tangent vectors computed from the two im-

ages, respectively. Also, let $\mathbf{B} = \mathbf{I}$, $\mathbf{W}_A = \mathbf{0}$ and $\mathbf{W}_B = \mathbf{I}$. The solution of Eq. (7) becomes $\Delta\hat{\mathbf{c}}_A = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\Delta\mathbf{x}$ and we obtain the distortion measure $D = \|(\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T)\Delta\mathbf{x}\|^2$. Careful comparison shows that the square root of D is indeed the tangent distance measure given in [11]. Although the weighting matrix $\mathbf{W}$ is fixed here, the linear system $\mathbf{L}$ is adapted to the input signal, so this method is categorized as an adaptive approach.

Differential methods for optical flow estimation [12] have a broad application in motion/velocity estimation and image registration. The formulations of many of these algorithms are quite similar to the tangent distance measure. The differences are 1) only spatio-temporal derivatives (those directly related to spatial translation) are used; 2) the estimated coefficients $\Delta\hat{\mathbf{c}}_A$, and not the distortion measure D, are of primary interest; and 3) the measurement is localized to allow calculation of an optical flow field, as opposed to the tangent distance measure (which is global).

The structural similarity index (*SSIM*) [4] is an adaptive image quality measure that provides a convenient way to separate out the non-structural luminance and contrast distortions. It cannot be written exactly in the adaptive linear system framework because it adopts a *polar*, rather than *rectilinear*, coordinate system. In the limit of small distortions, the two coordinate systems are indistinguishable, but it is not clear which provides a better prediction for the perception of large distortions. Nevertheless, we find that the adaptive linear system framework is more general and more flexible, easily allowing inclusion of additional non-structural distortion components.

3. STRUCTURAL IMAGE DISTORTION ANALYSIS

To decompose the image distortions as demonstrated in Fig. 1, we first need to define the non-structural and structural components. From the perspective of image formation, gentle distortions caused by variations of lighting conditions, spatial movement, or pointwise monotonic intensity changes caused by image acquisition and display devices should not change the perceived structure of the objects in the scene. The structural similarity principle posits that the human visual system discounts these distortions, and is primarily sensitive to the remaining distortions that change the structural information in the scene. Specifically, we define the following non-structural distortion components:

- Luminance change: $\mathbf{a}_1 = \mathbf{1}/\sqrt{N}$, where $\mathbf{1}$ denotes a column vector with all entries set to 1.
- Contrast change: $\mathbf{a}_2 = (\mathbf{x} - \mu_x\mathbf{1})\,/\|\mathbf{x} - \mu_x\mathbf{1}\|$, where μ_x is the mean of $\mathbf{x}$.
- Gamma distortion: This is a type of pointwise nonlinear intensity distortion (modelled as $f(x) = x^\gamma$) commonly used in describing image acquisition and display devices. Expanding in a Taylor series with respect to γ around $\gamma = 1$, and dropping the higher-order terms, we have $f(x) \approx x + x\log(x)(\gamma - 1)$. As such, we define $\mathbf{a}_3 = \mathbf{x}^*/\|\mathbf{x}^*\|$, where $\mathbf{x}^*$ is a vector whose i-th entry is given by $x_i^* = x_i\log(x_i)$.
- Horizontal translation: Writing a Taylor expansion of $\mathbf{x}$ with respect to horizontal translation and dropping higher-order terms, the distortion is approximated linear in the first-order horizontal derivative, $d_h(\mathbf{x})$. Thus, we define $\mathbf{a}_4 = d_h(\mathbf{x})/\|d_h(\mathbf{x})\|$.
- Vertical translation: Similar to horizontal translation, we define $\mathbf{a}_5 = d_v(\mathbf{x})/\|d_v(\mathbf{x})\|$, where $d_v(\mathbf{x})$ is the partial derivative of $\mathbf{x}$ with respect to the vertical spatial position.

We can now define a matrix of non-structural components, $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \mathbf{a}_3\ \mathbf{a}_4\ \mathbf{a}_5]$, with corresponding weights given by:

$$w_{a_1} = W_0 + \frac{|\mu_x - \mu_y|}{\sqrt{\mu_x^2 + \mu_y^2}}, \quad w_{a_2} = W_0 + \frac{|\sigma_x - \sigma_y|}{\sqrt{\sigma_x^2 + \sigma_y^2}},$$
$$w_{a_3} = w_{a_4} = w_{a_5} = W_0, \tag{9}$$

where σ_x and σ_y are the standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively, and W_0 is employed as a baseline (minimum) weight for all components. We set $W_0 = 0.1$ in our current implementation. Note that the weights given to the luminance and contrast components are adaptive, in order to enhance penalties for large distortions. For example, the weight given to the contrast component has minimum value W_0 when the two images have the same standard deviations, but grows when the ratio between them (σ_x/σ_y or σ_y/σ_x) is large. For structural distortion components, we apply a frequency decomposition method so that the variations of visual error sensitivity with respect to spatial frequency can be taken into account. In particular, for every local 8×8 image patch, we let $\mathbf{B} = \mathbf{F}^{-1}$, where $\mathbf{F}$ is the discrete cosine transform (DCT) matrix. We then define the diagonal entries of $\mathbf{W}_B$ to be inverse proportional to the JPEG DCT quantization table.

For a given pair of original and distorted images, we apply our distortion measure locally. At each spatial location, we extract the corresponding local 8×8 windows from the two images, and calculate the distortion measure. This results in a distortion map, and the total distortion is computed by averaging the distortion map.

We use the examples shown in Fig. 2 to demonstrate the performance of the proposed distortion measure. In Fig. 2, the original image is altered with a variety of distortion types. We have adjusted the degree of each distortion, so that all distorted images have very similar MSE with respect to the original image (although note that images (i) and (j) have slightly higher MSEs). By contrast, the perceived image distortion is drastically different: the quality of images (b)-(e) is quite poor, but images (f)-(j) do not appear to have significant loss of image quality. Three methods, in addition to MSE, are used to provide an objective evaluation of perceived image distortion/quality. The weighted MSE (WMSE) measure provides a direct comparison with respect to the proposed method when the adaptive part is removed. In particular, it uses $\mathbf{L} = \mathbf{F}^{-1}$ and the diagonal entries of $\mathbf{W}$ are set to be inverse proportional to the JPEG DCT quantization table. Using Eq. (4) (and given the fact that $\mathbf{F}$ is orthogonal: $\mathbf{F}^{-1} = \mathbf{F}^T$), we have WMSE = $\|\mathbf{WF}\Delta\mathbf{x}\|^2$. In Fig. 2, we see that WMSE does not show a significant improvement over MSE. The SSIM index is adaptive and provides a much more consistent image quality prediction for images (b)-(h), but it assumes all images are perfectly aligned and thus is too sensitive to the spatially-shifted images in (i) and (j). The proposed measure, D, is clearly the closest to human perception of the distortions in these example images.

4. CONCLUSION

We have formulated an adaptive linear system framework that generalizes many image comparison algorithms. The combination of this framework with the structural similarity principle gives rise to a new image distortion measure that correlates well with perceived image distortion. One attractive feature of the framework is that it provides a flexible nonlinear solution without sacrificing the

(a) MSE = 0, WMSE = 0 SSIM = 1, D = 0 | (b) MSE=212, WMSE=508 SSIM=0.675, D=296 | (c) MSE=210, WMSE=580 SSIM=0.713, D=248 | (d) MSE=212, WMSE=600 SSIM=0.760, D=131 | (e) MSE=212, WMSE=204 SSIM=0.802, D=166

(f) MSE=212, WMSE=595 SSIM=0.950, D=4.47 | (g) MSE=212, WMSE=542 SSIM=0.971, D=3.90 | (h) MSE=212, WMSE=542 SSIM=0.979, D=3.41 | (i) MSE=224, WMSE=534 SSIM=0.816, D=2.34 | (j) MSE=219, WMSE=500 SSIM=0.822, D=1.96

Fig. 2. Comparison of image distortion measures. (a) original image; (b) JPEG compression; (c) JPEG2000 compression; (d) blurring; (e) salt-pepper noise; (f) contrast reduction; (g) gamma distortion ($\gamma > 1$); (h) gamma distortion ($\gamma < 1$); (i) horizontal shift; (j) vertical shift.

elegance and tractability of representing signals with linear combinations of simple components. In particular, the nonlinear aspect of the process lies in the adaptive computation of the components and their associated weights, which is implemented as a preprocessing stage. The remaining process (signal decomposition and distortion measure) can be computed using standard linear algebra techniques. Of course the adherence to linearity also limits the framework to describing only those non-structural distortion types that lie within linear subspaces. This assumption may not hold when the distortion is highly nonlinear in the image space, such as large geometric distortions or changes in viewpoint.

Future work will include validation of the measure against human subjective data, both with standard image databases and with a set of images specifically optimized to compare this measure against other measures [13]. Finally, we believe that the measure may be extended by applying it to coefficients of a multi-scale decomposition, and by including other nonstructural distortions.

5. REFERENCES

[1] S. G. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[2] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, pp. 819–825, August 2001.

[3] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Nonlinear image representation for efficient perceptual coding," *IEEE Trans Image Processing*, 2005. Accepted for publication.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[5] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, pp. 129–132, Mar. 2002.

[6] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, vol. 4, pp. 525–536, 1974.

[7] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, pp. 1164–1175, Aug. 1997.

[8] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. ICASSP*, pp. 1945–1948, May 1989.

[9] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. SPIE*, vol. 1913, pp. 202–216, 1993.

[10] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 982–986, 1994.

[11] P. Y. Simard, Y. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition – tangent distance and tangent propagation," *International Journal of Imaging Systems and Technology*, vol. 11, no. 3, 2000.

[12] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys*, vol. 27, pp. 433–467, Sept. 1995.

[13] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Human Vision and Electronic Imaging IX, Proc. SPIE*, vol. 5292, pp. 99–108, Jan. 2004.