Perceptual Image Enhancement

Lark Kwon Choi*
Department of Electrical and Computer Engineering,
The University of Texas at Austin,
UT Austin WNCG, 1616 Guadalupe St, UTA 7.518, Austin, TX, 78701, USA
larkkwonchoi@utexas.edu

Todd Goodall
Department of Electrical and Computer Engineering,
The University of Texas at Austin,
UT Austin WNCG, 1616 Guadalupe St, UTA 7.518, Austin, TX, 78701, USA
tgoodall@utexas.edu

Deepti Ghadiyaram
Department of Computer Sciences,
The University of Texas at Austin,
UT Austin WNCG, 1616 Guadalupe St, UTA 7.518, Austin, TX, 78701, USA
deepti@cs.utexas.edu

Alan C. Bovik
Department of Electrical and Computer Engineering,
The University of Texas at Austin,
UT Austin WNCG, 1616 Guadalupe St, UTA 7.518, Austin, TX, 78701, USA
bovik@ece.utexas.edu
Abstract

Enhancing image quality has become an important issue as the volume of digital images increases exponentially and the expectation of high-quality images grows insatiably. Digitized images commonly suffer from poor visual quality due to distortions, low contrast, the deficiency of lighting, defocusing, atmospheric influences such as fog, severe compression, and transmission errors. Hence, image enhancement is indispensable for better perception, interpretation, and subsequent analysis. Since humans are generally regarded as the final arbiter of the visual quality of the enhanced images, perceptual image enhancement has been of great interest. In this chapter, we visit actively evolving perceptual image enhancement research. The paper discusses the principles of the human visual system that have been used in perceptual image enhancement algorithms and then presents modern image enhancement models and applications on the perceptual aspects of human vision.

Keywords

Image enhancement, image quality, human visual system, natural scene statistics, denoising, contrast enhancement, face recognition, image defogging, mammogram, and adaptive image compression

1. Introduction

The advent of digital imaging and the high-speed Internet have opened up a new era in photography and image processing. Amateur users easily capture, store, edit, and share images, while professional users including researchers effectively examine details, identify objects, interpret findings, and present results. As the global volume of digital images increases exponentially from mobile devices to large screens, the expectation of high-quality images grows insatiably. Digitized images commonly suffer from poor image quality, particularly due to distortions (e.g., blocking, ringing, blur, Gaussian noise), lack of contrast (e.g., variant illumination), deficiency in focusing (e.g., motion effects), lighting (e.g., under or overexposure),
adverse atmospheric influences (e.g., fog), and transmission errors (e.g., packet loss). Hence, image enhancement is frequently applied to improve the overall visual quality of an image for better perception, interpretation, and subsequent analysis.

Since humans are generally regarded as the final arbiter of the visual quality of enhanced images, understanding how the human visual system (HVS) tunes to natural environments and how to achieve visual perception has been studied for decades among vision scientists and engineers. The HVS promptly and reliably recognizes objects in a blurred or noisy image and penetrates the variability in shape, color, and shading of objects under the drastically changing illuminant conditions. From retina to the primary visual cortex, the HVS efficiently codes the visual signal using multiscale/multiorientation subband decomposition and a divisive normalization process. In addition, the HVS imposes spatiotemporal visual masking effects. When visual signals are superimposed on similar orientations, spatiotemporal frequencies, motion, or color, the local high-frequency energy in an image can strikingly reduce the visibility of other signals such as noise.

Growing interest in the HVS has inspired developing perceptual image enhancement models whose successful performance has also shown that accounting for the HVS is an essential part of perceptual image enhancement. Perceptually optimized denoising, deblurring, and distortion-blind image repair models enhance the visual quality of distorted images based on the HVS and of natural scene statistics (NSS). The multiscale Retinex with color restoration model, called MSRCR, combines models of the color constancy property of human vision to provide both enhanced contrast and sharpness. More examples of image enhancement models that are based on perceptual image processing can be found such as contrast enhancement using Gaussian mixture models, fusion-based approaches, and tone mapping operations (TMO) to achieve high-dynamic range (HDR) using structural fidelity and statistical naturalness.
Perceptual image enhancement techniques vary depending on enhancement purposes and applications. Denoising and deblurring models alleviate visual distortions, while contrast and visibility enhancement algorithms allocate new luminance or chrominance values to emphasize edges based on global or local neighborhood. Tone mapping operations adjust the color appearance of HDR natural scenes for display mediums that have a limited dynamic range. These perceptual image enhancement techniques are used in a variety of applications from improvement of daily snapshots on mobile-devices to robust face recognition in security, visibility enhancement in bad weather scenes, diagnostic analysis of medical imaging such as mammogram, and adaptive image/video compression.

To introduce actively evolving perceptual image enhancement research, we discuss image enhancement models and applications based on the perceptual aspects of human vision. Section 2 summarizes principles of the HVS that have been used in perceptual image enhancement. Section 3 presents modern perceptual image enhancement models. Practical applications of perceptual image enhancement algorithms are described in Section 4. We conclude this chapter with future directions in Section 5.

2. Principles of Human Visual System

2.1. Spatial and Temporal Contrast Sensitivity Function

The HVS, having many interconnecting neurons with hidden functions, is often treated like a black box. Measuring spatial and temporal acuity is important for determining the amount of stimulus contrast necessary to elicit a criterion level of response from the neuron. The perceived contrast is a key for visual perception since the visual information in the HVS is represented in terms of contrast rather than the absolute level of light. The spatial contrast threshold indicates the amount of contrast for the retina to detect a spatial pattern, while the temporal contrast threshold is the required temporal contrast to sense a time-varying signal. Contrast sensitivity is the inverse of the contrast threshold.
The spatial contrast sensitivity function (CSF) depicted in Figure 1(a) describes how the retina responds to contrast as a function of frequency\[^1\]. Retinal cone cells are most sensitive to stimuli at 4-8 cycles per degree (cpd), while nothing is visible beyond approximately 60 cpd. Perceptual image enhancement methods mostly concentrate on signals in the visible range of spatial frequencies, discarding imperceptible signals based on the CSF. The retina is composed of a layer of ganglion cells, which can be modeled as having circularly symmetric center-surround excitatory and inhibitory, on and off receptive fields. Differences in light between the center and periphery of these receptive fields enable neurons to fire at different rates. In addition, contrast sensitivity is a function of viewing distance. When close to a viewing screen, individual pixels become visible to a viewer, while at normal viewing distance, individual pixels become indistinguishable. Quantifying the perceived contrast of an image is important to design perceptually enhanced images and displays.

The temporal CSF provides a summary of the temporal response characteristics in the HVS\[^1\]. The highest contrast sensitivity is in the range of 8-15 Hz under sufficient lighting conditions, while visual stimuli over 50 Hz become undetectable. This temporal sensitivity varies with lighting conditions, so the temporal CSF curve shifts up or down depending on the integrated amount of light at the retina, as shown in Figure 1(b). For example, the HVS reduces the contrast sensitivity for higher frequency changes in low light. The lateral geniculate nucleus (LGN) decomposes the contrast signal from the retina into lagged and unlagged temporal responses, which can be modeled as temporally differenced outputs of gamma filters. The characteristics of the temporal CSF are also useful to develop perceptual video quality enhancement algorithms.
2.2. Multiscale and Multiorientation Decomposition

The natural environment is inherently multiscale and multiorientation. As a result, the HVS necessarily adjusts objects at various scales, orientations, and distance by efficiently coding the visual signal. The visual signals sensed by the eye are decomposed into bandpass channels as they pass through the visual cortex of the brain. The receptive fields of simple cells in the primary visual cortex (area V1) consist of adjacent excitatory and inhibitory regions, where they are elongated and oriented, and hence respond to stimuli in some orientations better than others. This orientation selectivity enables simple cells to detect the orientation of a stimulus. By using a wide range of orientations, lobe separations, and sizes, simple cells can be modeled as collectively providing a large number of a quadrature pair log-Gabor filter type responses; however, each simple cell is also sensitive to precise position of the visual stimuli within the receptive field. Complex cells are less dependent on stimulus position. They can be modeled as rectifying the responses of simple cells and further as normalizing them. The collective output of complex cells is often modeled using a filter-bank of log-Gabor filters tuned to various spatiotemporal orientations of

Figure 1. (a) Spatial contrast sensitivity function. (b) Temporal contrast sensitivity functions.
visual stimuli across scales. By using multiscale and multiorientation cell responses, the later stages of the HVS are able to efficiently encode both shape representations and object trajectories.

To mimic these multichannel characteristics of the HVS, several multiresolution image representations have been used for image analysis: The Gaussian and Laplacian pyramid, the wavelet decomposition, and Gabor filterbanks. The Gaussian pyramid provides a representation of the same image at multiple scales using lowpass filtering and decimation, while the Laplacian pyramid presents a set of detailed bandpass filtered images at different scales. The construction of a Gaussian pyramid involves continuous two-dimensional Gaussian lowpass filtering and subsampling (↓) by a factor of two. The Laplacian pyramid is then computed as the difference between the original image and the lowpass filtered image after upsampling (↑) by a factor of two. Figure 2(a) explains these processes. The wavelet decomposition represents an image in subbands having different scales and orientations. As shown in Figure 2(b), the LL (low, low) subband is a low resolution version of the original image, and the HL (high, low), LH (low, high), and HH (high, high) subbands, respectively, contain details with vertical, horizontal, and diagonal orientations. The Gabor filterbanks more closely approximates the response of the HVS, where the basis functions are Gaussian functions modulated by sinusoidal waves. As can be seen in Figure 2(c), Gabor filters have both frequency-selective and orientation-selective properties. Due to its optimally localized space-time frequency characteristics, Gabor filters are widely used in texture analysis, motion analysis, object recognition (e.g., fingerprint enhancement and iris recognition), etc. These natural and efficient multiscale/multiorientation decompositions are also extensively used for perceptual image enhancement models.
Figure 2. Multiresolution image representations: (a) Gaussian and Laplacian pyramid. (b) Wavelet decomposition. (c) Gabor filters and filtered images.
2.3. Visual Masking

Visual masking refers to the inability of the HVS to detect a stimulus, the target, when another stimulus, the mask, is superimposed on the target in a similar orientation, spatiotemporal frequency, motion, color, or other attributes.

2.3.1. Contrast (or Texture) Masking

Contrast masking indicates the variation of the detection thresholds of the target as a function of the contrast of the mask. The presence of local high-frequency energy in an image strikingly reduces the perceptual significance of the distortions [3]. For example, JPEG blocking artifacts are evident on the lady’s face and neck, but far less visible or eliminated in the flora regions as shown in Figure 3(a). Figure 3(b) is another example, where the image is uniformly distorted by the same level of additive Gaussian noise. Distortions are highly visible on the smooth regions (e.g., face), but nearly imperceptible on the more textured areas (e.g., hair and scarf).

Figure 3. Contrast masking of (a) JPEG artifacts and (b) additive white Gaussian noise.
Heeger’s model\textsuperscript{[4]} explains how contrast masking occurs in the HVS. As processed visual stimuli pass through the eyes to visual cortex, V1 neurons execute multiscale and multiorientation decompositions, computing local energy responses to the visual stimulus. These responses, in the form of contrast energy, are then normalized by local summations of neighboring responses. In this manner, neurons with separate spatiotemporal tunings respond relative to each other. This non-linearity is commonly called the divisive normalization. The computational divisive normalization transform (DNT) model is depicted in Figure 4.

![Diagram of Computational Divisive Normalization Model](image)

**Figure 4.** Illustration of the computational divisive normalization model.

### 2.3.2. Luminance Masking

The HVS may be modeled as performing a transform following processing by the retinal ganglion cells where the processed input is subtracted then divided by the local mean response. This operation accomplishes an adaptive gain control and entropy reduction for signal coding. This transform causes masking of luminance changes in proportion to the background luminance. The Weber-Fechner law states...
that the ratio of changes in luminance to the average luminance must be larger than a specific constant to be visible.

2.3.3. Temporal Masking and Silencing

Temporal masking in videos is analogous to spatial contrast masking in images, but is made more perceptually complex by the addition of motion. Models of temporal masking have been developed using heuristics for video compression. The commonly modeled case of sparse, global scene changes is one special case of temporal masking. A striking motion silencing illusion was recently introduced in which local salient temporal changes of objects in luminance, hue, size, and shape are masked in the presence of large, coherent motion [5]. Although the underlying mechanisms of silencing remain unknown, Choi and Bovik et al. [6] have developed a filter-based quantitative spatiotemporal flicker detector model that accurately predicts human judgments of silencing as a function of object velocity and object change rate. They suggest that the bandpass responses of cortical neurons may be implicated in whether changes in moving objects are perceived.

2.4. Modeling Statistical Regularities of Natural Images

The development of the HVS is strongly dependent on early visual stimulation and the statistics of the surrounding visual environment. The early stages of vision, from the retina to the primary cortex, are constrained to process local points of the visual field using groups of neurons representing localized responses. The output of these neurons is determined by natural statistical regularities. Hence, modeling the statistical regularities of natural images is a modern approach to model the HVS.

The primary goal of the HVS is to efficiently reduce redundancy while maximizing information of incoming visual stimuli using efficient statistical measures. Similar to the remarkably efficient coding of
visual signals into Gabor-like complex cell responses by cortical neurons in area V1, the first few local principal (and independent) components of natural images are also observed to maximally decorrelate the image content. The PCA filters yielding the highest response variance are found to be oriented edges or their derivatives, thus resembling the oriented receptive fields found in the cortex. This suggests that a decorrelation operation is accomplished by HVS to reduce the redundancy.

The highly structured statistical regularities exhibited by natural images are well modeled by spatial linear bandpass filters such as Difference of Gaussian (DoG), Laplacian of Gaussian (LoG), Gabor, or a predictive coding filter. DoG models the center-surround excitatory-inhibitory processes that occur at various stages of visual processing, while LoG measures the second derivative of an image after Gaussian smoothing. LoG filters are useful for finding edges and blobs. It has been observed that the empirical probability distributions of the responses of these filters can be reliably modeled using a Gaussian scale mixture (GSM) probability distribution. The GSM model can be used to capture both the marginal and the joint distributions of the wavelet coefficients across subbands helps to extract enhanced image by removing noisy signals. It has also been observed that the non-Gaussian long-tailed histograms of natural scenes imply the feasibility of ‘sparse coding’ by visual neurons. A large class of image enhancement algorithms is based on modeling of statistical regularities of natural images.

3. Perceptual Image Enhancement Models

3.1. Denoising, Deblurring, and Blind Repair of Natural Images

3.1.1. Perceptual Denoising

Visual signals are often degraded by noise during image acquisition and transmission, so the development of efficient denoising techniques is highly desirable to enhance perceived visual quality. The statistical regularities of natural images have been exploited by applying wavelet decomposition to the signal with a set of multiscale bandpass oriented filters. The observed marginal distributions of wavelet coefficients are
highly kurtotic and long-tailed. In addition to regularities in marginal distributions, it has been observed that higher order dependencies exist between neighboring subbands across scales and orientations. These observations were successfully incorporated by Portilla and Strela et al. \cite{7} in a model to denoise images corrupted by independent additive Gaussian noise of known variance.

When an image is decomposed into oriented subbands at multiple scales, the neighborhood includes coefficients from other subbands (i.e., at nearby scales and orientations) as well as from the same subband. The coefficients within each local neighborhood around a reference coefficient of a pyramid subband are modeled using a GSM \cite{7}: A random vector \( x \) is a GSM if it can be expressed as the product of two independent components

\[
x = \sqrt{z} u,
\]

where \( u \) is a zero-mean Gaussian vector with covariance matrix \( C_u \), and \( z \) is called a mixing multiplier.

Assuming that the original image is corrupted by independent additive white Gaussian noise, and then a noisy neighborhood coefficient vector \( y \) can be characterized as

\[
y = x + w = \sqrt{z} u + w,
\]

where \( w \) is a noise coefficient vector with covariance matrix \( C_w \).

The image denoising method 1) decomposes the image into pyramid subbands at different scales and orientations; 2) denoises each subband, except the lowpass residual band; and 3) inverts the pyramid transform, obtaining the denoised image. The group of neighboring coefficients constitutes a sliding window that moves across the wavelet subband. At each step, the reference center coefficient, \( x_c \), of the window is estimated (i.e., denoised). Therefore, a denoising method is to estimate \( x_c \) of a spatial location \( x \) from a noisy neighborhood \( y \). The Bayes least square estimator is used to compute the conditional mean as follows \cite{7},

\[
E[x_c | y] = \int_0^\infty E[x_c | y, z] p(z | y) \, dz,
\]
where $E\{x \mid y, z\} = zC_u(zC_u + C_w)^{-1}y$, from a local linear wiener estimate, and $C_u$ can be estimated from the observed noisy covariance matrix by $C_u = C_y - C_w$. The posterior density $p(z \mid y)$ can be predicted by Bayes’ rule: $p(z \mid y) \propto p(y \mid z)p_z(z)$, where $p_z(z) \propto 1 / z$. $p(y \mid z)$ can be calculated as a Gaussian function of zero-mean and covariance $C_y = zC_u + C_w$. These estimates in different subbands are used to reconstruct the denoised image. On a set of images corrupted by simulated additive white Gaussian noise of known variance, this model is shown to introduce fewer artifacts and to better preserve edges.

Another model in the wavelet domain but designed for video denoising was proposed [8]. Since there exist strong correlations between wavelet coefficients across adjacent video frames, grouping wavelet coefficients along temporal directions based on a GSM framework reinforces the statistical regularities of the coefficients and enhances the performance of Bayesian signal denoising algorithm. A GSM was used to capture the local correlations between the spatial and temporal wavelet coefficients, and such correlations were strengthened with a noise-robust motion compensation process, resulting in competitive performance of denoising [8]. The keys of the spatiotemporal GSM (ST-GSM) denoising method were to include temporal neighborhoods in the wavelet coefficient vector and to correctly estimate global motion in the presence of noise. Figure 5 shows the visual effects of the denoising algorithm for the Suzie sequence. ST-GSM shows quite effective performance at suppressing background noise while maintaining the edge and texture details, where the structural similarity (SSIM) maps of the corresponding frames were used to verify the denoising enhancement.
Figure 5. Denoising results of Frame 100 in *Suzie* sequence corrupted with noise standard deviation $\sigma = 20$. (a) Original image. (b) Noisy image. (c) ST-GSM denoised sequence. (d) Corresponding SSIM quality maps: brighter indicates larger SSIM value, high image quality. Images are generated using an algorithm [8].

3.1.2. *Perceptual Deblurring*

One of the most common artifacts in digital photo and videography is blurring, which could be caused by a number of factors - camera shake, long exposure, poor focus etc. This artifact can be modeled as a blur kernel convolved with image intensities, and thus recovering an unblurred image signal from a single blurred signal translates to signal deconvolution. The heavy-tailed property of the empirical distributions of natural images has been successfully used for a number of image enhancement models including deconvolution. The NSS of derivative distributions of unblurred images are exploited by Fergus and Singh et al. [9]. The authors adopted a Bayesian approach using the aforementioned scene statistics as a prior to find a blur kernel, then reconstructed the deblurred image using the estimated kernel via a standard deconvolution method.
The algorithm proposed in [9] takes a blurred input image $B$, and few more manual inputs from the user: 1) a rectangular patch, rich in edge structure, within the blurred image, 2) the maximum size of the blur kernel, and 3) an initial $3 \times 3$ blur kernel. The blurry image $B$ was assumed to have been generated by convolution of a blur kernel $K$ with a latent sharp image $L$ plus sensor noise $N$ at each pixel:

$$B = K \otimes L + N,$$  \hspace{1cm} (4)

where $\otimes$ denotes discrete image convolution (with non-periodic boundary conditions). In the first step, the blur kernel $K$ is inferred using multiscale approach. Specifically, let $P$ denote the blurred patch selected by the user. The blur kernel $K$ and the latent patch image $L_p$ are estimated by varying image resolution in a coarse-to-fine manner in order to avoid local minima. This estimation problem is posed as a search for the max-marginal blurring kernel, by finding the highest probability, guided by a prior on the statistics of $L$. In the second step, using this estimated kernel, a standard deconvolution (e.g., Richardson-Lucy) algorithm is applied to estimate the unblurred image $L$. \cite{9}

Figure 6 shows that the algorithm significantly sharpens a blurred image. The result suggests that applying natural image priors and advanced statistical techniques greatly assists in recovering an unblurred to blurred signal from a blurry image. Although ringing artifacts occur near saturated regions and regions of significant object motion, the algorithm achieves superior performance when applied on a large number of real images with varying degrees of blur.
Figure 6. Top: A scene with a blur. The patch selected by the user is indicated by the gray rectangle. Bottom: The deblurred image. Close-up images of the selected patch are shown in the right. Images are generated using an algorithm [9].

3.1.3. Perceptually Optimized Blind Repair of Natural Images

A distortion-blind, perceptually optimized GEneral-purpose No-reference Image Improver (GENII) enhances images corrupted by any of JPEG compression, JPEG2000 compression, additive white noise, or Gaussian blur distortions [10]. The GENII framework operates based on NSS features extracted from a blind image quality assessment (IQA) algorithm in the wavelet or spatial domains. Since natural images possess highly reliable statistical regularities that are measurably modified by the presence of distortions,
NSS features can be used to identify unknown distortions by type and severity. Once a distortion type has been identified, the extracted features are utilized to predict the perceptual quality of the image using a quality assessment process. If the predicted quality falls below a certain threshold, a trained regression module such as support vector regression (SVR) is queried to estimate perceptually optimized parameters of that distortion. GENII then proceeds to invoke an appropriate image repair algorithm such as deringing, deblocking, denoising, or deblurring models based on the predicted distortion type. The repaired intermediate image is then passed back into the loop in order to evaluate visual quality and identify its distortion type. This loop continues until the obtained intermediate image has the highest possible quality or until a specific finite number of repair iterations have been performed. An illustration of the GENII framework including a distorted image and the corresponding repaired image is shown in Figure 7.

Figure 7. An illustration of the GENII framework. NSS features are used to predict the distortion type, the severity, the visual quality, and the distortion parameters that serve as inputs to a possible non-blind repair algorithm. The intermediate result is fed back to the system until the best possible quality is achieved. Images are provided by the authors of [10].
As a working exemplar model of the GENII framework, GENII-1 is aligned with distortion-specific image repair algorithms. For example, the block matching 3D (BM3D) algorithm and the trilateral filter are used for denoising and deringing, respectively, but other repair algorithms can be adopted. GENII-1 was tested to evaluate the performance on 4000 distorted images. GENII-1 performs quite well predicting and ameliorating the distortions present in the image, even when the image has multiple distortions.

3.2. Perceptual Image Contrast Enhancement

3.2.1. Multiscale Retinex with Color Restoration

Retinex is a model of the lightness and color perception of human vision. As the word “Retinex” implies, this model relates to the neurophysiological functions of individual neurons in the primate retina, lateral geniculate nucleus, and cerebral cortex of the brain as a form of a center-surround spatial structure \[1^{11}\].

The Retinex algorithm uses multiscale analysis to decompose the image into different frequency bands, and then enhances its desired global and local frequencies. Retinex image processing achieves sharpening, local contrast enhancement, color constancy, and natural rendition of high dynamic range (HDR) lightness and color. The Retinex concept evolved from single scale Retinex (SSR) into a full scale image enhancement algorithm, called the multiscale Retinex with color restoration (MSRCR) \[1^{11}\]. MSRCR improves the visibility of color and details in shadows by using color constancy. Here, the constancy refers to the resilience of perceived color and lightness to spatial and spectral illumination variations. MSRCR is widely used from everyday images to lunar orbiter images, medical imaging such as digital X-ray, mammography, CT scans, and MRI by enhancing contrast and sharpness, while reducing the bandwidth requirement (e.g., telemedicine applications).

MSRCR algorithm consists of multiscale Retinex (MSR) and color restoration as follows \[1^{11}\],

\[
R_{\text{MSRCR}}(x, y) = \alpha_{MSR}(x, y) \sum_{r=1}^{K} W_r \{\log I_r(x, y) - \log[ F_r(x, y) \otimes I_r(x, y)]\}.
\] (5)
\[ \alpha_i(x, y) = \log \left( \frac{N I_i(x, y)}{\sum_{i=1}^{N} I_i(x, y)} \right), \]  
(6) \[ F_k(x, y) = \kappa \exp[-(x^2 + y^2)/\sigma_k^2], \]  
(7) \[ \kappa = \frac{1}{\sum_{x} \sum_{y} F(x, y)}, \]  
(8)

where index \( i \) is the \( i \text{th} \) spectral band in an image \( I \), \((x_1, x_2)\) is the pixel location, and \( \otimes \) is the convolution operator. \( \alpha_i(x_1, x_2) \) is the color restoration coefficient in the \( i \text{th} \) spectral band, \( N \) is the number of spectral band (Typically, \( N = 1 \) for grayscale images, and \( N = 3 \) for RGB color images), \( F_k \) is the \( k \text{th} \) Gaussian surround function, while \( K \) is the number of scales. \( \sigma_k \) is the standard deviation of the Gaussian kernels.

MSRCR performs a log operation on the ratio of each pixel in each spectral band to both spatial and spectral averages. This suppression of spatial and spectral lighting variations mimics human perception of color and lightness.

MSRCR has been tested on a large number of images and has consistently proven to be better than conventional contrast enhancement methods. Results show that MSRCR automatically improves images with diverse degrees of visual deficit. For example, to produce more realistic enhanced images, MSRCR slightly sharpens the dimed original image or removes the effects of the shadows for the moderately underexposed images. Further, MSRCR reduces the underexposure and completely eliminate the effect of dark shadows by enhancing the details in the dark zones as shown in Figure 8. MSRCR image processing is patented (e.g., USPTO #5,991,456, # 6,843,125, #6,842,543, CAN #PCT/US1997/007996, etc.) and available for licensing to commercial purposes.
Figure 8. Top: Original images. Bottom: Output images using the MSRCR algorithm. Images are generated using a demo version of TruView® PhotoFlair.

3.2.2. Histogram Equalization Approach using Gaussian Mixture Model

Adaptive image equalization using the Gaussian mixture model (GMM) is another contrast enhancement method based on the HVS [12]. For decades, histogram modification techniques are generally regarded as the most straightforward and intuitive contrast enhancement approaches due to their simplicity. Global histogram equalization, local histogram equalization, brightness-preserving histogram equalization with maximum entropy, the flattest histogram specification with accurate brightness preservation, histogram modification framework, and contrast enhancement based on genetic algorithm are representative examples of contrast enhancement methods using histogram modification. However, the aforementioned techniques cause problems when the histogram has spikes, often resulting in a hashed, noisy appearance of the output image or undesirable checkerboard effects. Further, they often suffer from empirical parameter settings. To solve these problems and to improve the visual quality of input images, the automatic image equalization and contrast enhancement algorithm using the GMM is proposed [12].
Celik and Tjahjadi [12] explained that human eyes are more sensitive to widely scattered fluctuations than small variations around dense data. Hence, in order to increase the contrast while retaining image details, dense data with low standard deviation should be spread, whereas scattered data with high standard deviation should be compacted. To achieve this, the GMM based algorithm\cite{12} partitions the histogram distribution of an input image into a mixture of different Gaussian components as shown in Figure 9(b). The significant intersection points are selected from all the possible intersections between the Gaussian components. Then, the contrast enhanced image is generated by transforming the pixel’s gray levels in each input interval to the appropriate output gray-level interval according to the dominant Gaussian component and the cumulative distribution function of the input interval. Since the HVS is more sensitive to sudden changes in widely scattered data and less sensitive to smooth changes in densely scattered data, the Gaussian components with small variances are weighted with smaller values than the Gaussian components with larger variances, and vice versa\cite{12}. Further, the gray-level distributions are used to weight the components in the mapping of the input interval to the output interval as shown in Figure 9(c). For color images, the luminance component is processed for contrast enhancement, while the chrominance components are preserved in $L^*a^*b^*$ color space.

Performance evaluation using absolute minimum mean brightness, discrete entropy, and edge-based contrast on 300 gray and color images of the Berkeley image database shows that the proposed algorithm outperforms other histogram modification algorithms without any parameter tuning, even under diverse illumination conditions\cite{12}. The proposed algorithm attains drastic dynamic range compression while preserving fine details without yielding artifacts such as halos, gradient reversals, or loss of local contrast.
3.2.3. Image Fusion

Global contrast enhancement is limited in its ability to improve local details, while local enhancement approaches can suffer from block discontinuities (e.g., checkerboard effects), noise amplification, and unnatural image modifications. A fusion-based contrast enhancement solves these problems and balances the effects of local and global contrast \cite{13}.

A fusion-based contrast enhancement involves extracting the most informative regions from the input images and filtering the local areas of input images with weighted sum to achieve the contrast enhanced output. Since the HVS is more sensitive to local contrast changes such as edges, and since multiscale decomposition provides convenient localization of contrast changes, multiresolution-based image fusion approaches are widely used in practice. Specifically, Laplacian multiscale refinement achieves a seamless, halo-free fused image. Several input images and corresponding normalized weight maps are decomposed via the Gaussian and the Laplacian pyramid, respectively, and then they are blended into a fused pyramid as follows \cite{13},

\[
F_i = \sum_{k=1}^{N} G_i \{\overline{W}_k \} L_i \{I_k \},
\]

\[
\overline{W}_k = W_k / \sum_{k=1}^{N} W_k,
\]
where $k$ is the index of input images $I$, $N$ is the number of input images, $W_k$ is a weight map of the input image at index $k$, $\overline{W}_k$ is a normalized weight map to ensure that the sum of $W_k$ to unity, and $l$ is the number of pyramid levels. $l$ can be various but is recommended to be over 6 in order to eliminate halo effects. $G_l\{\cdot\}$ and $L_l\{\cdot\}$ represent the Gaussian and the Laplacian decomposition at pyramid levels $l$, respectively.

Operations are performed successively for each level, in a bottom-up manner. Finally, a contrast enhanced image $J$ is achieved by the Laplacian pyramid reconstruction as follows $^{[13]}$,

$$ J = \sum_l F_l \uparrow^n, $$

where $\uparrow^n$ is the upsampling operator with factor $n = 2^{l-1}$.

The challenging part of the fusion-based approaches is the design of the weight maps. The weight maps vary depending on applications, but are largely based on visual quality features or salience dictated by the particular vision task. For example, the scalar weight map,

$$ W_k = (W_k^{chr})^\alpha (W_k^{sat})^\beta (W_k^{con})^\gamma (W_k^{sal})^\delta, $$

where $W_k^{chr}$, $W_k^{sat}$, $W_k^{con}$, and $W_k^{sal}$ denote chrominance, saturation, contrast, and saliency respectively, while exponent $\alpha$, $\beta$, $\gamma$, and $\delta$ represent impact portions between weighting features. Results over a wide range of test images show that the fusion-based approach balances local and global contrasts without block discontinuities and artifacts amplifications, while preserving the original image appearance. Fusion-based contrast enhancement techniques are widely used to improve image focus, image exposure, HDR photography, and image segmentation. Figure 10 shows some example results of the fusion-based contrast enhancement.
Figure 10. (a), (c) Original images. (b), (d) Output images of the fusion-based contrast enhancement. Images are generated using an algorithm [13].

3.3. Tone Mapping and High-Dynamic Range Image Enhancement

The dynamic range of the natural environment is extremely large. The HVS efficiently adapts colors, luminance, and tones in the real world into the best appearance, while standard acquisition and display devices suffer from visualization of the real world due to their limited low dynamic ranges (LDR). To effectively visualize the natural environment, various tone-mapping operators (TMOs) and HDR image enhancement techniques have been suggested. In spite of significant progress on developing TMOs, further improvement of HDR image enhancement are challenging due to the difficulty of perceptual IQA for HDR images [14]. Recently, a perceptually relevant tone mapped image quality index (TMQI) [14] and a new optimized TMO [15] based on TMQI were introduced, which have shown to correlate well with human judgments on a wide range of tone-mapped images.
The authors of [14] explain that a good tone mapped image should achieve a high performance both in multiscale structural fidelity and in statistical naturalness. Since the HVS extracts structural information from the visual scene, multiscale signal fidelity is a good predictor of perceptual quality. In addition, the deviation from the NSS of an input image can be used as an indicator of image naturalness in terms of perceptual quality [14]. TMQI not only provides an overall quality score of the tone-mapped image, but also creates multiscale quality maps that reflect the structural fidelity variations across scale and space. Let $X$ and $Y$ be the HDR and the tone-mapped LDR image, respectively. The TMQI computation is expressed as follows [14],

$$
\text{TMQI}(X,Y) = a[S(X,Y)]^\alpha + (1-a)[N(Y)]^\beta,
$$

(13)

$$
S(X,Y) = \frac{1}{M} \sum_{i=1}^{M} S_{\text{local}}(x_i, y_i),
$$

(14)

$$
S_{\text{local}}(x, y) = \frac{2\sigma_x \sigma_y + C_1}{\sigma_x^2 + \sigma_y^2 + C_1} \times \frac{2\sigma_x \sigma_y + C_2}{\sigma_x \sigma_y + C_2},
$$

(15)

$$
\sigma = \frac{1}{\sqrt{2\pi} \theta} \int_{-\infty}^{\infty} \left[ -\frac{(t - \tau_\sigma)^2}{2\theta^2} \right] dt,
$$

(16)

$$
N(Y) = \frac{1}{K} P_m P_d,
$$

(17)

where $S$ and $N$ denote the structural fidelity and statistical naturalness, respectively. The parameters $\alpha, \beta$ determine the sensitivities of the two factors, and $0 \leq a \leq 1$ adjusts the relative importance between them. $x$ and $y$ denote two local image patches extracted from $X$ and $Y$. $S_{\text{local}}$ means the structural similarity (SSIM)-motivated local structural fidelity, while $\sigma_x, \sigma_y,$ and $\sigma_{xy}$ indicate the local standard deviations and cross correlation between the two corresponding patches. $C_1$ and $C_2$ are stability constants. $\tau_\sigma$ is a contrast threshold, and $\theta = \tau_\sigma / 3$. The statistical naturalness $N$ is modeled as the product of two density functions of image brightness and contrast. $K$ is a normalization factor given by $K = \max\{P_m, P_d\}$, where $P_m, P_d$ are a Gaussian density function and a Beta density function, respectively.
A new TMO approach \cite{15} based on TMQI is an iterative algorithm that searches for the best solution in the image space. Specifically, the algorithm adopts a gradient ascent method to improve the structural fidelity $S$, then solves a constrained optimization problem to enhance the statistical naturalness $N$. The two-step iteration alternately continues until convergence to an optimized TMQI. Performance evaluation is executed on a database of 14 HDR images including various contents such as humans, landscapes, architectures, as well as indoor and night scene. Results show that both the structural fidelity and naturalness enhancement steps significantly improve the overall quality of the tone-mapped image as can be seen in Figure 11.

![Figure 11. Top: Initial images created by Gamma corrections. Bottom: Output images of the tone mapped and HDR image enhancement. Images are provided by the authors of [15].](image)

4. Perceptual Image Enhancement Applications

4.1. Robust Face Recognition under Variable Illumination Conditions

Face recognition plays a central role in a variety of applications such as security and surveillance systems. Various face recognition methods including eigenface, fisherface, independent component analysis, and
Bayesian face recognition have been suggested in the past decades. However, due to the difficulties in handling the lighting conditions in practice, recognizing human faces in variable illuminations, especially in the side lighting effects, is challenging. To achieve robust face recognition under varying illuminations, an adaptive region-based image enhancement method is proposed [16].

The first step of the algorithm of [16] is segmenting an image into differently lit-regions according to its local homogeneous illumination conditions. The region segmentation is based on an edge map of the face image obtained using two-level wavelet decomposition. Adaptive region-based histogram equalization (ARHE) is then regionally applied to enhance the contrast and to minimize variations under different lighting conditions based on the spatial frequency responses of the HVS. The variations of illumination generally lie in the low frequency band. Next, since the high-frequency components become more substantial in recognition of details with poor illuminations, the algorithm enlarges the high-frequency coefficients at different scales so as to enhance the edges, which is called EdgeE. At the end, an enhanced image is reconstructed using the inverse discrete wavelet transform with the modified coefficients. Finally, the reconstructed enhanced image is used for face recognition. Figure 12 shows the examples of face images under various illumination conditions and corresponding enhanced images using an algorithm [16].

The experimental results using the ARHE and EdgeE on a variety of face recognition databases (e.g., Yale B, Extended Yale B, and Carnegie Mellon University Pose, Illumination, and Expression) show that the method significantly improves the face recognition rate when the face images include illumination variations. The simplicity and generality of the algorithm enable the approach to be applied directly to any single image without any lighting assumption, without any prior information about the 3D face geometry, and without any data fitting steps [16].
4.2. Perceptual Image Defogging and Scene Visibility Enhancement

The perception of outdoor natural scenes is important for successfully conducting visual activities such as object detection, recognition, and navigation. In bad weather, the absorption or scattering of light by atmospheric particles such as fog, haze, and mist can seriously degrade visibility. As a result, objects in images captured under such conditions suffer from low contrast, faint color, and shifted luminance. Since degraded visibility can cause serious operator misjudgments in vehicles guided by camera images and can induce erroneous sensing in surveillance systems, image defogging designed to enhance scene visibility of foggy images is important [17].
Recently a referenceless, perceptual image defogging and visibility enhancement algorithm was introduced [18] based on a space domain NSS model and multiscale “fog aware” statistical features [17]. Here, ‘referenceless’ means that the perceptual defogging algorithm does not require multiple foggy images, different degrees of polarization, pre-detected salient objects in a foggy scene, side geographical information obtained from an onboard camera, a depth dependent transmission map, training on human-rated judgments, or content-dependent assumptions such as smoothness of airlight layers, smoothness of a depth map, or substantial variations of color and luminance in a foggy scene [18]. Fog aware statistical features include the variance of local mean subtracted contrast normalized (MSCN) coefficients, the variance of the vertical product of MSCN coefficients, the sharpness, the coefficient of variance of sharpness, the Michelson contrast (or contrast energy), the image entropy, a dark channel prior, the color saturation in HSV color space, and the colorfulness.

The defogging and visibility enhancer utilizes statistical regularities observed on foggy and fog-free images to predict perceptual fog density of the input foggy image, then extract the most visible information from three processed images: one white balanced and two contrast enhanced images. Chrominance, saturation, saliency, perceptual fog density, fog aware luminance, and contrast weight maps are applied to effectively filter the most visible areas of the three preprocessed images and to smoothly blend them via Laplacian multiscale refinement [18].

Evaluation on a variety of foggy images in subjective and objective comparisons shows that the perceptual image defogging algorithm produces more naturalistic, clear contrast of edges and vivid color restoration after defogging. Figure 13 shows the examples of foggy images and their corresponding defogged, visibility enhanced images.
Figure 13. Top: Original foggy images. Bottom: defogged, visibility enhanced images using the perceptual image defogging algorithm. Images are provided by the authors of [18].

4.3. Nonlinear Unsharp Masking for Mammogram Enhancement

Digital mammography is the most common technique to detect breast cancer since it yields better image quality, uses a lower X-ray dose, and offers faster diagnosis with confident interpretation [19]. Mammogram enhancement algorithms generally focus on the contrast improvement of specific regions and details in mammograms, and then apply a threshold in order to separately visualize them from their surroundings. Nonlinear unsharp masking (NLUM) method combining the non-linear filtering and unsharp masking (UM) technique is one of the powerful mammogram enhancement algorithms [19].

A traditional UM technique is to subtract a lowpass filtered image from the original image. The UM improves visual quality by adding a scaled high-frequency part of the signal to its original; however, it can simultaneously amplify noise and over-enhances steep edges. NLUM enhances the fine details of mammographic images by integrating the nonlinear filtering with UM. The original mammogram $I(m,n)$ is filtered by a nonlinear filter, then normalized and combined with the original mammogram using the arithmetic addition and multiplication in order to obtain an enhanced mammogram. Mathematically, the enhanced mammogram is expressed by [19]
\[ E(m,n) = A_1 I(m,n) + A_2 \frac{F(m,n)}{|F|_{\text{max}}} I(m,n), \]  

(18)

where \( A_1 \) and \( A_2 \) are scaling factors, and \( |F|_{\text{max}} \) is the maximum absolute value of the nonlinear filtered mammogram, \( F(m,n) = w_0 I_0 + w_1 I_1 + w_2 I_2 \), where \( w_0, w_1, \) and \( w_2 \) are weight coefficients working as highpass filters, and \( I_0, I_1, I_2 \) are exponentially center-weighted mean filter of the original mammogram \( I \), respectively. The second-derivative-like measure of enhancement (SDME) is applied to optimize the NLUM parameters, obtaining the best enhancement result. Specifically, the NLUM parameters are decided where the SDME curve reaches the local maxima by changing the scaling factor and exponential gain coefficients \(^{[19]}\).

Results tested on a variety of 322 mammograms from the mini-MIAS database show that NLUM achieves excellent performance for improving the local contrast of specific regions and fine details such as microcalcifications. The abnormal regions are more recognizable in the enhanced mammograms. It was also observed that NLUM helps doctors to diagnose breast cancers successfully. For detailed examples, we refer to \([19]\).

### 4.4. Adaptive Perceptual Image and Video Compression

Most existing approaches in image and video compression have focused on developing reliable methods to minimize mathematically tractable and easy to measure distortion metrics. Some of these non-perceptual techniques sometimes fail to correlate well with the perceived quality of the reconstructed images, especially at lower bit rates \(^{[20]}\). As a new effort to overcome these limitations, certain properties of the HVS have been used in the context of signal compression in order to minimize perceptually redundant information while preserving visual quality.

Höntsch and Karam \([20]\) proposed a discrete cosine transform (DCT)-based, locally adaptive, and perceptual image coder based on the contrast sensitivity and visual masking effects of the HVS. This
coder locally and adaptively discriminates between signal components based on their perceptual relevance. Specifically, the coder attempts to conceal coding distortions using the computed local spatial detection thresholds thereby discarding the signal components which are imperceptible to the human receiver.

The coder proposed in [20] decomposes a given image into several subbands that vary in frequency and orientation. The subbands at varying local frequencies and orientations are used to adaptively compute local distortion sensitivity profiles in the form of distortion thresholds (i.e., perceptual weights). These thresholds are then used to adaptively control the quantization and dequantization stages of the coding system by adapting the quantization reconstruction levels to the varying characteristics of the visual data. The nature of this approach is that the process depends on image content, and the thresholds are derived locally. These thresholds are needed both at the encoder and at the decoder in order to reconstruct the coded visual data. To avoid the need to transmit certain amount of side information, locally available masking characteristics of the visual input were estimated only at the decoder [20].

The authors compared the bit rates (in bits per pixel) obtained using their locally adaptive perceptual image compression (APIC) with a non-locally adaptive Safranek-Johnston method (called a perceptual image coder, PIC) for perceptually lossless compression. We reproduced the results presented by the authors in their work in Table 1 and Figure 14. As shown in Table 1 and Figure 14, APIC successfully removes a large number of perceptually redundant information in images and results in superior performance when compared to existing non-locally adaptive schemes.
Table 1. Bit rates for the PIC and the APIC for (a) perceptually-lossless compression and (b) near transparent compression with a step size multiplier of 2.0.

<table>
<thead>
<tr>
<th>Image</th>
<th>Bit rate (bits per pixel)</th>
<th>Compression gain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PIC</td>
<td>APIC</td>
</tr>
<tr>
<td>Baboon</td>
<td>1.57</td>
<td>1.24</td>
</tr>
<tr>
<td>Actor</td>
<td>0.99</td>
<td>0.55</td>
</tr>
<tr>
<td>Lighthouse</td>
<td>0.92</td>
<td>0.61</td>
</tr>
<tr>
<td>Lena</td>
<td>0.79</td>
<td>0.54</td>
</tr>
<tr>
<td>Boat</td>
<td>0.90</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image</th>
<th>Bit rate (bits per pixel)</th>
<th>Compression gain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PIC</td>
<td>APIC</td>
</tr>
<tr>
<td>Baboon</td>
<td>1.05</td>
<td>0.85</td>
</tr>
<tr>
<td>Actor</td>
<td>0.71</td>
<td>0.37</td>
</tr>
<tr>
<td>Lighthouse</td>
<td>0.67</td>
<td>0.41</td>
</tr>
<tr>
<td>Lena</td>
<td>0.54</td>
<td>0.35</td>
</tr>
<tr>
<td>Boat</td>
<td>0.64</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Figure 14. Results of coding *Actor* and *Lenna* with a step size multiplier of 2.0. (a), (d) original images at 8 bits per pixel (bpp). (b), (e) PIC at 0.706 and 0.635 bpp. (c), (f) APIC at 0.372 and 0.351 bpp, respectively. The images are generated using an algorithm [20].

5. Conclusion

We have discussed actively evolving perceptual image enhancement models and applications based on the principles of the HVS. Since the HVS has remarkable abilities to analyze visual stimuli,
understanding how the HVS tunes to natural environments and achieves stable, clear visual perception leads to the development of successful perceptual image enhancement methods. Among numerous properties of the HVS, contrast sensitivity function, multiscale and multiorientation decomposition, visual masking, and modeling of the NSS have played crucial roles in the development of algorithms that alleviate visual distortions, that increase contrast and visibility, and that adjust HDR natural scenes to display mediums. Perceptually inspired image enhancement models have been extensively applied to practical applications for better perception, interpretation, and subsequent analysis such as robust face recognition, image defogging, mammogram diagnosis, and image/video compression. In summary, understanding and modeling the HVS is fundamental to developing modern image enhancement models, and perceptual image enhancement is a promising direction for future research.

Although we introduced representative perceptual image enhancement models and applications, many emerging image enhancement topics remain in an early stage. Enhancement of color quality is an important area, yet there are not well-accepted perceptual quality predictions of color images. Progress remains to be made on color perception including color constancy and opponency. Enhancement of stereoscopic 3D images is also a motivating topic. The role of depth and visual comfort are worth exploring for 3D image enhancement. As mobile cameras become very widely used, enhancement of pictures acquired using mobile devices such as image stabilization, color balance, and noise reduction under a low light condition are pressing areas for future work on perceptual image enhancement.
References


[Online]. Available: https://ece.uwaterloo.ca/~z70wang/research/stgsm/


[Online]. Available: [https://ece.uwaterloo.ca/~z70wang/research/tmqi/](https://ece.uwaterloo.ca/~z70wang/research/tmqi/)


