NEAR-THRESHOLD PERCEPTUAL DISTORTION PREDICTION BASED ON OPTIMAL STRUCTURE CLASSIFICATION

Yucheng Liu and Jan P. Allebach

School of Electrical and Computer Engineering,
Purdue University, West Lafayette, IN 47906-2035, U.S.A.
{yc-liu11, allebach}@purdue.edu

ABSTRACT

Perceptual distortion prediction at near-threshold level has many applications in general image/video processing tasks. This paper presents a computational model to predict the near-threshold perceptual distortions based on optimal structure classification. This model accounts for contrast sensitivity, light adaptation, and various masking effects of the human visual system (HVS), and automatically adapts to local image structures by a soft classification scheme using a Gaussian Mixture Model (GMM). The proposed model is trained and verified on the public CSIQ local masking database. We demonstrate a superior prediction performance of the proposed model compared to previous research.

Index Terms— Perceptual distortion prediction, distortion estimation, visual masking, texture classification

1. INTRODUCTION

The perception of distortions present in natural images has been studied widely in multiple areas of image and video processing, such as image quality assessment [1], [2], image compression [3], and digital watermarking [4]. Today with the reduced cost of disk storage and Internet bandwidth, images and videos are stored and transmitted at very high quality where only visually imperceptible distortions are tolerable. Therefore, predicting the visibility of near-threshold distortion is a becoming an increasingly important concern.

In previous research, A number of full-reference image quality metrics have been proposed to address the problem of perceptual image similarity [1], [2], [5]. Although these methods were reported to perform excellently on public databases [6]-[8], they are not ideal for predicting the visibilities of near-threshold distortions [9]. A possible explanation is that these metrics did not directly model the primary cortex (V1) response of the HVS, which is believed to be critical for near-threshold human vision modeling [9].

The ability of the HVS to detect distortions in natural images is known to be affected by various factors: light adaptation [10], [11], contrast sensitivity [12], contrast masking [13] and entropy masking [9], [14]. The most widely studied HVS models are gain-control-based HVS models, such as Watson-Solomon's masking model and its variants [15], [16], and the contrast threshold elevation models [17]-[19]. However, it is not until recently that we have access to a large-scale masking database for natural images [20]; and most models had their parameters tuned on unnatural masks. Even with access to such a database, the parameter optimizations for these models are still difficult due to their divisive forms and complicated spatial and subband pooling strategy (e.g. Minkowski Pooling).

It is also found in prior work that the parameters of the HVS models should be optimized differently for different image structures, such as flat regions, edges, and textures, to achieve better performance [10], [11], [16]. However, the structure classification was done either manually or heuristically, rather than being data driven. And the parameters have never been optimized on a large-scale natural image dataset.

Recently, some machine learning frameworks have been proposed as data driven HVS models [3], [21]. However, such models can only operate on patches of the same size as the training patches, which limits the generalizability of the models. Moreover, these models never account for the spatial distribution of artifacts within the patches, which could be an important factor for distortion visibility.

In this paper, we propose a computational model based on optimal structure classification to predict the visibilities of near-threshold distortions in natural images. Specifically, to predict the perceptual distortion for each fovea-size patch in the image, we first divide the patch into smaller overlapping sub-blocks. Each sub-block is then softly classified into multiple structure categories using GMM posterior probabilities. Each category has a perceptual distortion prediction model optimized specifically for the structures within the category; and the distortion response of the sub-block is generated by combining the responses from multiple models weighted by the GMM posterior probabilities. Finally, we pool over the top-10% of the sub-block responses within the patch to obtain the distortion score for the patch.

This paper is organized as follows: Section 2 presents the details of the model. Section 3 provides the experimental re-

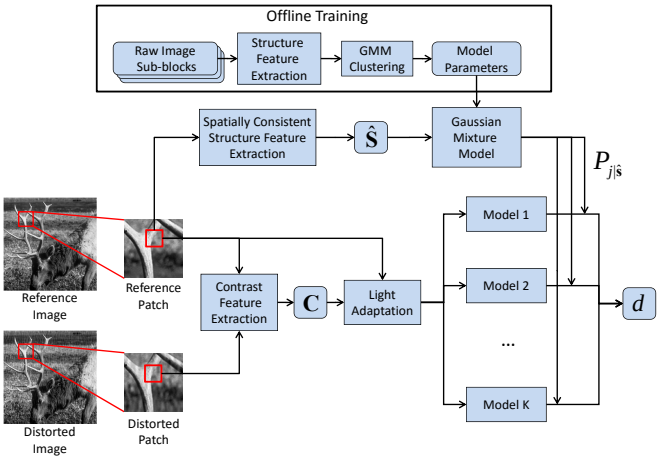

Fig. 1. Optimal classification-based JND model.

sults and analysis in comparison to other works; and Section 4 draws the conclusions.

2. OPTIMAL CLASSIFICATION-BASED PERCEPTUAL DISTORTION MODEL

2.1. Model Framework

The model framework adopted in this work is illustrated in Fig. 1. For each fovea-size patch extracted from the images, we further divide it into overlapping sub-blocks of size $M \times M$. The distortion response on each sub-block is computed independently. Major components in this framework include: 1) Structure feature extraction, 2) Contrast feature extraction, 3) Soft classification model, and 4) Distortion prediction model.

2.1.1. Structure Feature Extraction

The structure feature is used in our model to identify the type of a sub-block so that the corresponding optimal model can be applied more aggressively at a later stage.

We first perform a log-Gabor filter bank decomposition [22] on the reference image. Let $x_i(a,b)$ denote the subband coefficient at location (a,b) of the i-th log-Gabor filtered image, the I-dimension structure feature $\mathbf{s}(n) = [s_1(n), \ldots, s_I(n)]$ of the n-th sub-block is computed as

$$s_i(n) = \frac{1}{M^2} \sum_{(a,b)\in W_n} |\psi(x_i(a,b))|, i = 1, 2, \ldots, I, \qquad (1)$$

where I is the number of subbands, W_n denotes the n-th sub-block in the image patch. And $\psi(\cdot) = \tanh(\alpha t)$ is the bounded nonlinear function that maps the subband coefficient to $(0,1)$. Here, α is a model parameter.

When extracting the structure features, we shift the subband indices at each scale to prioritize the orientation with strongest subband energy to make the structure feature rotation invariant. In addition, we enforce spatial consistency on the structure features. Specifically, for each patch, we sample N sub-blocks on an evenly spaced grid, compute the structure features, and augment the structure features with the location coordinates of the sub-blocks: $\tilde{\mathbf{s}}(n) = [x(n), y(n), \mathbf{s}(n)]$. Then K-means clustering is applied to the augmented features $\{\tilde{\mathbf{s}}(n)\}_{n=1}^{N}$ to associate each sub-block with one of the centroid vectors $\mathbf{c}_k = [x_k^c, y_k^c, \mathbf{s}_k^c], k = 1, 2, \ldots, K_s$, where K_s is the number of classes in K-means clustering. The spatially consistent structure feature of a sub-block $\hat{\mathbf{s}}$ is taken to be the last I dimensions of its associated centroid vector, i.e. $\hat{\mathbf{s}}(n) = \mathbf{s}_{k_n}^c, n \in S_{k_n}$, where S_{k_n} is the cluster containing the n-th sub-block.

2.1.2. Contrast Feature Extraction

The contrast feature is the basic input to our model, which depicts the perceptual contrast difference of the distortion at different visual channels.

To compute the contrast feature, we first convert the raw pixel values of the reference and distorted images to lightness values in the range $[0, 100]$ using the viewing parameters as given in [20]. We denote $L_r^*(a,b)$ and $L_d^*(a,b)$ as the lightness values of the reference and distorted images, respectively. Then a log-Gabor filter bank is applied to the lightness images. Let $x_i^{L_r^*}(a,b)$ and $x_i^{L_d^*}(a,b)$ denote the subband coefficients at location (a,b) of the i-th subbands of the reference and distorted patches, respectively. The $(J \times L)$-dimension contrast feature of the n-th sub-block $\mathbf{c}(n) = [c_{1,1}, \ldots, c_{1,L}, \ldots, c_{J,1}, \ldots, c_{J,L}]$ is given by

$$c_{j,l}(n) = \frac{1}{M^2} \sum_{(a,b)\in W_n} \left| \frac{x_j^{L_r^*}(a,b) - x_j^{L_d^*}(a,b)}{x_{n_l^j}^{L_r^*}(a,b) + x_{n_l^j}^{L_d^*}(a,b)} \right|^{\rho}, \qquad (2)$$

$$\text{for } j = 1, 2, \ldots, J \text{ and } l = 1, 2, \ldots, L,$$

where J is the number of subbands, L is the number of adjacent bands for each subband, ρ models the non-linear contrast response of HVS, and $N_j = \{n_1^j, n_2^j, \ldots, n_L^j\}$ is the index set of the adjacent bands of the j-th subband. We choose N_j to include the subbands within ± 45 degrees, and 1.7 octaves of frequency bandwidth of the j-th subband. The contrast feature captures the contrast masking effect due to the subband components in the mask that are close to those of the error signal, also known as the intra-band masking effect [10].

2.1.3. Soft Classification Model

The soft classification model categorizes the sub-blocks based on their structure features into K groups representing different image structures, such as flat regions, edges, textures and other mid-level structures. Each group has a prediction model optimized specifically for the structures within the group. For a new input sub-block, the model assign K posterior probabilities to it, indicating the likelihood that it belongs to each

group. The GMM posterior probability can be expressed as

$$P_{j|\hat{\mathbf{s}}} = \frac{\frac{\pi_j}{|\Sigma_j|^{1/2}} \exp\left(-(\hat{\mathbf{s}}-\boldsymbol{\mu}_j)^T \Sigma_j (\hat{\mathbf{s}}-\boldsymbol{\mu}_j)\right)}{\sum_{k=1}^{K} \frac{\pi_k}{|\Sigma_k|^{1/2}} \exp\left(-(\hat{\mathbf{s}}-\boldsymbol{\mu}_k)^T \Sigma_k (\hat{\mathbf{s}}-\boldsymbol{\mu}_k)\right)}, \quad \text{for } j = 1, 2, \ldots, K, \tag{3}$$

where π, $\boldsymbol{\mu}$ and Σ are model parameters obtained from the training stage, which will be elaborated later.

2.1.4. *Distortion Prediction Model*

Given the contrast feature and GMM posterior probability, the proposed perceptual distortion estimator for each sub-block can be expressed as follows

$$\begin{aligned} d(n) &= \sum_{i=1}^{K} P_{i|\hat{\mathbf{s}}(n)} \phi_i(n), \\ \text{in which } \phi_i(n) &= \phi\left(b_i + \mathbf{w}_i^T \left(g\left(\overline{L^*}(n)\right)\mathbf{c}(n)\right)\right), \end{aligned} \tag{4}$$

where $\mathbf{w}$ and b are model parameters, and ϕ is the sigmoid function $\phi(t) = \frac{1}{1+e^{-t}}$. The light adaptation function $g(\cdot)$ depends on the average lightness

$$\overline{L^*}(n) = \frac{1}{M^2} \sum_{(a,b)\in W_n} L_r^*(a,b), \tag{5}$$

of the reference sub-block. The individual model responses $\phi_i(n)$ can be interpreted as a linear combination of light adapted perceptual contrast differences in different subbands followed by a non-linear activation. Regarding the light adaptation function $g(\cdot)$, we evenly divide the range $[0, 100]$ into 10 bins and assign a representative adaptation factor for each bin $g_l, l = 0, 1, \ldots, 9$. The function $g(\cdot)$ can be expressed as

$$g(t) = g_l, \text{ where } l = \lfloor t/10 \rfloor. \tag{6}$$

The subband combining weight $\mathbf{w}$ automatically accounts for the contrast sensitivity. In addition, the inter-band masking effect is captured by choosing $\mathbf{w}$ adaptively according to the content of the sub-block. The distortion estimation is generated by combining the model responses based on the posterior probabilities of the sub-block belonging to the K groups.

2.2. Stochastic Gradient Descent (SGD) Model Training

We evaluate our model on the public CSIQ local masking database [20]. The database provides 1080 image patches of size 85×85, the corresponding error signals, and contrast detection thresholds from six measurements. Let us use $\overline{C_i}$ and σ_i to denote, respectively, the average and standard deviation of the threshold measurements of the i-th image patch. Since our model is not designed to predict the thresholds directly, we use reference patches and error signals to generate reference-distorted training pairs with ground truth labels. Specifically, we generate training pairs with above-threshold distortion contrast $\overline{C_i} + \beta\sigma_i$ and label 1, below-threshold distortion contrast $\overline{C_i} - \beta\sigma_i$ and label 0, and threshold distortion contrast $\overline{C_i}$ and label 0.5. The parameter β is chosen to achieve a balance between model accuracy and model overfitting.

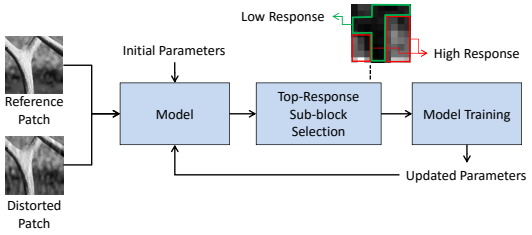

Fig. 2. Iterative model update and sample selection.

The top-response sub-blocks are sampled from the labelled reference-distorted pairs to train the model. The GMM classification model is trained using the EM algorithm independently of the labels and using sub-blocks from the reference patches only. The prediction model is trained using the labels and $L2$ cost by Stochastic Gradient Descent (SGD). Let f be the cost function. We have

$$\begin{aligned} \frac{\partial f}{\partial w_{ij}} &= \frac{1}{|B|} \sum_{n\in B} g(\overline{L^*}(n)) E_i(n) c_j(n), \\ \frac{\partial f}{\partial b_i} &= \frac{1}{|B|} \sum_{n\in B} E_i(n), \\ \frac{\partial f}{\partial g_l} &= \frac{1}{|B_l|} \sum_{n\in B_l} \sum_{j=1}^{D} F_j(n) c_j(n), \end{aligned} \tag{7}$$

in which E and F are defined as

$$\begin{aligned} E_i(n) &= \left(d(n) - t(n)\right) P_{i|\hat{\mathbf{s}}(n)} \phi_i(n)(1-\phi_i(n)), \\ F_j(n) &= \sum_{i=1}^{K} E_i(n) w_{ij}. \end{aligned} \tag{8}$$

For (7) and (8), $t(n) \in \{0, 0.5, 1\}$ is the ground truth label of the n-th sub-block, B is the sub-block batch for the current SGD update, and $B_l = \{n \in B | g(\overline{L^*}(n)) = g_l\}$.

Since our model takes a sub-block as input while the ground truth is labelled for each image patch, we iteratively extract training samples from each patch by choosing sub-blocks with the top-10% responses and assigning them the same labels as the patch. Specifically, after each epoch in SGD, we use the updated model to recompute the responses on the sub-blocks and re-select top-response training samples (sub-blocks) from each patch to use in the next epoch, as illustrated in Fig. 2. The initial model parameters are obtained by training on all sub-blocks. At the prediction stage, the response of a patch is estimated by averaging the top-10% sub-block responses in the patch. This practice is based on the prevalent assumption that the perceptual distortion of an image is determined by the worst regions.

3. RESULTS AND ANALYSIS

In our experiment, we randomly partition the 1080 image patches in the CSIQ local masking database into five subsets,

Table 1. Optimal model parameters

Symbol	Description	Value
	General	
M	Sub-block size	16
Δ	Sub-block sampling stride	8
K	Number of structure classes	10
	Structure Features	
α	Subband coefficient scaler	0.25
K_s	Number of local clusters	3
I	Number of subbands[1]	6×4
	Contrast Features	
ρ	Contrast feature exponent	0.9
J	Number of subbands[1]	6×4
L	Number of adjacent subbands	3

and conduct a five-fold cross-validation to verify the performance of the proposed model. Each time, we use three subsets for training, one subset for validation, and the remaining subset for testing. The training stage learns the optimal parameters $\mathbf{w}$, b, and g in (4). Other parameters are tuned by minimizing the error on the validation set, as given in Table 1.

The performance of the model is measured by RMSE, Pearson linear correlation coefficient (CC), and Spearman rank-order correlation (SROCC) of the threshold predictions on the test set. In the proposed model, we deem that the distortion is present at threshold level when the model gives a response of 0.5. For comparison, we implemented 1) Watson-Soloman model [15], 2) our previously proposed adjacent band inhibition model [18], 3) SSIM [1], and 4) a Support Vector Machine (SVM) visibility threshold regression using the 14 normalized features suggested in [20] and a RBF kernel. For a fair comparison, the parameters of all the models are optimized on the validation set before the models are evaluated on the test set. The average performance of the models on the five test sets are given in Table. 2. We also list in the table the results from a recent work using an optimal gain-control model and a CNN model [21] for comparison, although their experiment was conducted under slightly different conditions.

In Fig. 3 we illustrate the classification map of an example image, where each sub-block is color-coded according to the structure class it belongs to with highest probability. The map roughly segments the image into four types of structures: smooth (cyan) as over the sky, edges (green) as around the birds and flowers, texture (blue) as in the body of the cactus, and mid-level structures (orange) as in the flower clusters. The prediction framework automatically applies a different model for each type with corresponding optimal parameters.

To see the distribution of errors among different types of patches, we use average lightness L^* and average texture intensity T [23] to classify the 1080 patches into four groups: dark ($L^* < 10$), flat ($L^* \geq 10$ and $T < 0.3$), mid-texture ($L^* \geq 10$ and $0.3 \leq T < 0.6$), and high-texture ($L^* \geq 10$ and $T \geq 0.6$). The RMSE within each group is shown in Table 3. It appears that the detection thresholds on dark regions are less predictable. By inspection of the ground truth data, we found that very similar dark patches with little structures in the patches can result in totally different thresholds, which explained the errors in the model prediction.

[1] I and J are represented as $N_f \times N_\theta$, where N_f and N_θ are the number of frequency bands and the number of orientation channels, respectively.

Table 2. Performance comparison of models over the test set

Model	RMSE (dB)	CC	SROCC
Five-fold cross-validation			
Watson-Soloman [15]	6.65	0.74	0.76
Liu-Allebach [18]	4.98	0.82	0.84
SSIM [1]	7.16	0.52	0.48
SVM with features from [20]	4.56	0.85	0.86
Proposed model	**4.42**	**0.86**	**0.88**
Training set 70%, validation set 15%, test set 15%			
optimal gain-control [21]	5.24	N/A	N/A
CNN model [21]	5.55	N/A	N/A

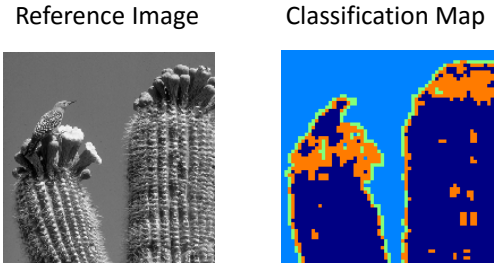

Fig. 3. Optimal classification map with $K = 4$.

Table 3. Prediction RMSE of different types of patches

	Dark	Flat	Mid-texture	High-Texture
RMSE (dB)	10.83	2.83	4.01	3.54

4. CONCLUSION

This paper presents a computational model to predict the near-threshold perceptual distortions based on optimal structure classification. This model accounts for contrast sensitivity, light adaptation, and various masking effects of the human visual system (HVS), and automatically adapts to local image structures by a soft classification scheme using a Gaussian Mixture Model (GMM). The proposed model is trained and verified on the public CSIQ local masking database. We demonstrate a superior prediction performance of the proposed model compared to previous research.

5. REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to

structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[2] D. M Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, pp. 2284–98, 2007.

[3] Md Mushfiqul Alam, Tuan D. Nguyen, Martin T. Hagan, and Damon M. Chandler, "A perceptual quantization strategy for HEVC based on a convolutional neural network trained on natural images," *Proc. SPIE*, vol. 9599, 2015.

[4] M. Masry, D. M. Chandler, and S. S. Hemami, "Digital watermarking using local contrast-based texture masking," *Conf. Rec. 37th Asilomar Conf. Signals, Syst. and Comput.*, vol. 2, pp. 1590–1594, 2003.

[5] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[6] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[7] E. C. Larson and D.M. Chandler, "Categorical subjective image quality csiq database," http://vision.okstate.edu/csiq/, 2009.

[8] N. Ponomarenko, A. Zelensky V. Lukin, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances Modern Radioelectronics*, vol. 10, pp. 30–45, 2009.

[9] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *ISRN Signal Process.*, vol. 2013, pp. 53, 2013.

[10] Yuting Jia, Weisi Lin, and Ashraf Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits and Syst. for Video Technol.*, vol. 16, no. 7, pp. 820–829, 2006.

[11] X. H. Zhang, W. S. Lin, and Ping Xue, "Improved estimation for just-noticeable visual distortion," *Signal Process.*, vol. 85, no. 4, pp. 795–808, 2005.

[12] S. J. Daly and B. E. Rogowitz, "Visible differences predictor: an algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, no. 2, pp. 2–15, 1992.

[13] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Am.*, vol. 70, pp. 1458–71, 1980.

[14] A. B. Watson, R. Borthwick, M. Taylor, B. E. Rogowitz, and T. N. Pappas, "Image quality and entropy masking," *Proc. SPIE*, vol. 3016, pp. 2–12, 1997.

[15] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A*, vol. 14, pp. 2379–91, 1997.

[16] D. M Chandler, M. D. Gaubatz, and S. S. Hemami, "A patch-based structural masking model with an application to compression," *EURASIP J. Image and Video Process.*, p. 22, 2009.

[17] Marcus J. Nadenau, Julien Reichel, and Murat Kunt, "Performance comparison of masking models based on a new psychovisual test method with natural scenery stimuli," *Signal Process.: Image Commun.*, vol. 17, no. 10, pp. 807–823, 2002.

[18] Yucheng Liu and Jan P. Allebach, "A computational texture masking model for natural images based on adjacent visual channel inhibition," *Proc. SPIE*, vol. 9016, pp. 11, 2014.

[19] Yucheng Liu and Jan P. Allebach, "A patch-based cross masking model for natural images with detail loss and additive defects," *Proc. SPIE*, vol. 9394, pp. 14, 2015.

[20] Md Mushfiqul Alam, Kedarnath P. Vilankar, David J. Field, and Damon M. Chandler, "Local masking in natural images: A database and analysis," *J. Vision*, vol. 14, no. 8, pp. 1–38, 2014.

[21] M.M. Alam, P. Patil, M.T. Hagan, and D.M. Chandler, "A computational model for predicting local distortion visibility via convolutional neural network trained on natural scenes," *IEEE Int. Conf. Image Process. (ICIP)*, pp. 3967 – 3971, 2015.

[22] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, pp. 2379–2397, 1987.

[23] Ruth Bergman, Hila Nachlieli, and Gitit Ruckenstein, "Detection of textured areas in natural images using an indicator based on component counts," *J. Electron. Imaging*, vol. 17, no. 4, pp. 13, 2008.