https://doi.org/10.2352/ISSN.2470-1173.2020.10.IPAS-137

An expandable image database for evaluation of full-reference image visual quality metrics

Mykola Ponomarenko[1], Oleg Ieremeiev[2], Vladimir Lukin[2], Karen Egiazarian[1]
[1]Tampere University, FIN 33101, Tampere, Finland
[2]National Aerospace University, 61070, Kharkiv, Ukraine

Abstract

Traditional approach to collect mean opinion score (MOS) values for evaluation of full-reference image quality metrics has two serious drawbacks. The first drawback is a nonlinearity of MOS, only partially compensated by the use of rank order correlation coefficients in a further analysis. The second drawback are limitations on number of distortion types and distortion levels in image database imposed by a maximum allowed time to carry out an experiment. One of the largest of databases used for this purpose, TID2013, has almost reached these limitations, which makes an extension of TID2013 within the boundaries of this approach to be practically unfeasible. In this paper, a novel methodology to collect MOS values, with a possibility to infinitely increase a size of a database by adding new types of distortions, is proposed. For the proposed methodology, MOS values are collected for pairs of distortions, one of them being a signal dependent Gaussian noise. A technique of effective linearization and normalization of MOS is described. Extensive experiments for linearization of MOS values to extend TID2013 database are carried out.

Keywords: image visual quality assessment, full-reference image visual quality metrics, human visual perception

Introduction

Full-reference image visual quality metrics are widely used at different stages of digital image processing: verification of new image enhancement methods (e.g. image denoising), image quality monitoring in lossy image compression and digital watermarking, etc. Large test image databases with mean opinion score values are used for assessing the correspondence between image quality metrics and human visual perception [1]. One of the largest available databases is TID2013, which contains 25 reference (distortion free) images and 3000 distorted images (24 types of distortions and 5 levels for each distortion) [2].

To obtain MOS values for TID2013, about 1000 observers have been engaged. Each observer assessed visual quality of all distorted images for one reference image (from totally 120 images). An average time of one experiment was about 17 minutes, near to a maximum allowed time for an experiment (30 minutes [3]). Because of the time restriction, it is unfeasible to collect MOS values for a next generation of TID2013 with significantly larger number of distortion types, since in the case of a larger database (more distortion types or more levels of distortions), the time of one experiment will exceed an acceptable time limit. At the same time, it is possible to add to such a database more reference images, because experiments are performed for each reference image separately. In particular, the recently proposed KADID-10k database [4] includes the same number of distortions as TID2013, but 80 reference images instead 25 as in TID2013. However, to reach better representability of such image database for verification of visual quality metrics, one should extend number of different distortion types and levels of distortions, which is not possible for the methodology used in TID2013 and KADID-10k.

Another drawback of existing methodology is a nonlinearity of the obtained MOS (MOS values obtained in [2] are in a non-linear scale since the distributions of visual quality of distorted images are non-uniform). Specifically, the same value of MOS may correspond to slightly different levels of visual quality for different reference images. Due to this, one cannot truly rely on root mean square error (RMSE) for estimation of correspondence between full-reference metric and MOS. In practice, both Spearman and Kendall rank order correlation coefficients [5] are used in such an analysis, but their values highly depend on overall number of test images and the number of noise levels. Moreover, it is difficult to use these MOS values for linearization of a given full-reference metric. Therefore, to make the TID2013 database expandable, we have to solve abovementioned problems. We solve these problems in this paper by proposing the following approach.

For each reference image, a sequence of 20 calibration images distorted by signal dependent white Gaussian noise are created. Visual quality for the sequence of calibration images (SCI) has to vary from 100% for the first image (image is visually undistinguishable from the reference image) to 0% for the last image (image is practically invisible under the noise).

We utilize the following two peculiarities of human visual system (HVS):

•an increase of noise level is perceived as linear if a noise variance increases in a geometric progression;

•image regions with higher dissimilarity (e.g., noise-like textures and fine details) have higher noise masking ability [6].

The formed SCI are used to obtain linearized MOS for other distorted images of the same reference image, e.g. for new types of distortions added to an existing test image database, whose MOS shall be linearized.

This paper is organized as follows. In Section 2, psycho-visual experiments to analyze and ability of HVS to distinguish two levels of the noise are described. In Section 3, an algorithm of forming SCI is proposed. Section 4 describes experiments to obtain linearized MOS for TID2013, to add new types of distortions to the database and to merge two existing databases. Finally, Section 5, considers two examples, where obtained linearized MOS for TID2013 are applied for linearization of values of a full-reference image visual quality metric.

Experimental estimation of sensitivity of HVS to additive white Gaussian noise

Weber-Fechner law [7] states that the relationship between a stimulus and human perception is logarithmic. If a stimulus varies as a geometric progression, the corresponding human perception is altered in an arithmetic progression. Let us explore experimentally if this law works for perception of noise level (white Gaussian noise

in brightness color component in YCbCr color space) on images or not.

For psycho-visual experiments, we have created two noisy versions (noises with variances σ_1^2 and σ_2^2) of a homogeneous test image, demonstrated them to an observer asking to decide which image has a higher level of noise. This experiment has been repeated for many observers and the obtained percentage of true decisions (choices) allows us to estimate the ability of HVS to distinguish noises with variances σ_1^2 and σ_2^2.

We have carried out experiments setting σ_1^2 = {16, 50, 100, 200, 400}, and $\sigma_2^2 = \{1.0625\sigma_1^2, 1.125\sigma_1^2, 1.25\sigma_1^2, 1.5\sigma_1^2\}$ for all 20 possible combinations of σ_1^2 and σ_2^2.

Each observer during one experiment has been asked to make a decision on 60 pairs of images (3 times for each combination of σ_1^2 and σ_2^2). Every time the new homogeneous image was used to create a new pair of noisy images, in order to prevent learning effects. We asked some observers to repeat the experiments for several times after a reasonable long break in between. Observers chose the most convenient distance between their eyes and a monitor.

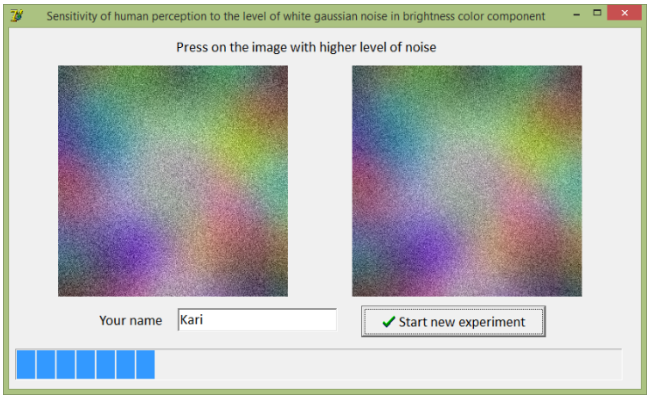


Figure 1. Main window of the designed software

Fig. 1 shows the main window of the designed software. A total number of experiments was 94. Four results we have rejected as abnormal. Thus, 270 elementary decisions for each combination of σ_1^2 and σ_2^2 have been collected. Next, we need to find a threshold T for the number of true decisions (D) which cannot be accidently obtained. If D is close to T, then people were unable to distinguish between noisy images with variances σ_1^2 and σ_2^2. The difference D-T corresponds to a confidence of observers in truly made decision, higher the difference – more confidence.

If the decisions would be done accidentally (in a random manner), then the number of true decisions shall follow the Gaussian distribution with the mean level μ=135 (270/2) and a standard deviation σ=8.22. Setting an upper bound T for randomly obtained true decisions as $\mu + 3\sigma$, for our experiments we obtain the threshold value T equal to 160.

Fig. 2 shows curves of dependence between the number of true decisions and the difference $\sigma_2^2 - \sigma_1^2$.

One can see that there are no obvious dependences. The difference 8 for σ_1^2=16 is well distinguishable for HVS (234 true decisions from 270). At the same time, the difference 25 for σ_1^2=400 is practically undistinguished (only 162 true decisions from 270).

Let us analyze now the dependences between the number of true decisions and relative variance σ_2^2/σ_1^2 (see Fig. 3).

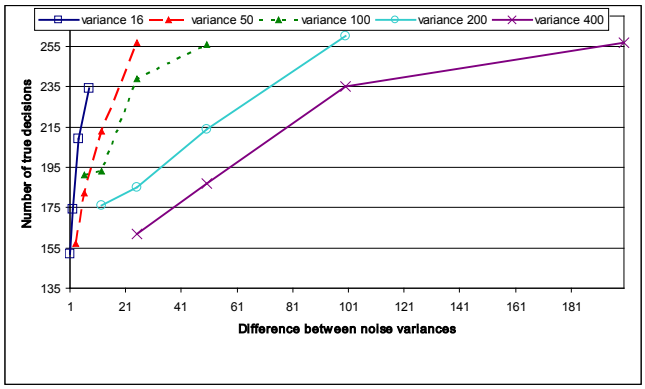


Figure 2. Dependence of the number of true decisions on the difference σ_2^2-σ_1^2

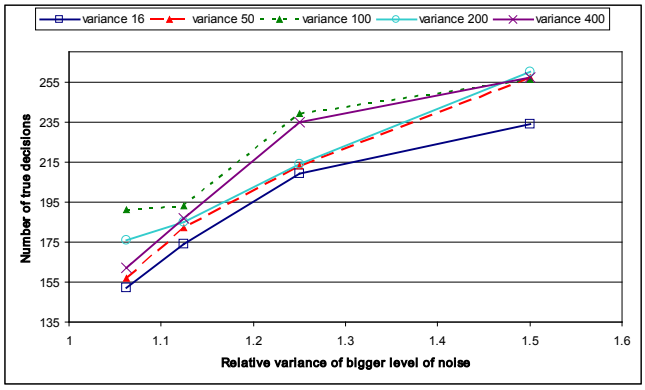


Figure 3. Dependence of the number of true decisions on σ_2^2/σ_1^2

Here, we can see a strong dependence (except two abnormal points for variances 100 and 200, and σ_2^2/σ_1^2=1.0625). Ability of HVS to distinguish noise levels as different is proportional to the value of σ_2^2/σ_1^2. Therefore, we have empirically demonstrated that the Weber–Fechner law works for perception of noise levels, too.

The presence of two abnormal points in Fig. 3 can be explained by learning or adaptation of observers during experiments to peculiarities of test images and to noise levels (mostly, to noise level which is in the middle of the range of variances (see Fig. 4)).

It is clearly seen that for σ_1^2=100 the largest number of true decisions takes place. It is possible also that HVS, on the average, is more trained to distinguish middle levels of contrasts, brightness, noise levels corresponding to σ_1^2=100. This will require an additional research.

The analysis of dependences in Fig. 3 shows, that HVS *distinguishes levels of noise well enough for σ_2^2/σ_1^2 exceeding 1.25*. For σ_2^2/σ_1^2=1.125, the number of true decisions is quite low (close to T). For σ_2^2/σ_1^2=1.0625, the number of true decisions is very low and is at the level of T.

Thus, if one neglects some roughness in the methodology of experiments, it is possible to make the following conclusions. The lower bound of distinguishable σ_2^2/σ_1^2 is 6%. For lower σ_2^2/σ_1^2, the difference is practically undistinguishable. In other words, *noise on an image region with variance σ_1^2 is able to effectively mask other noise with variance not larger than 1/16 σ_1^2*.

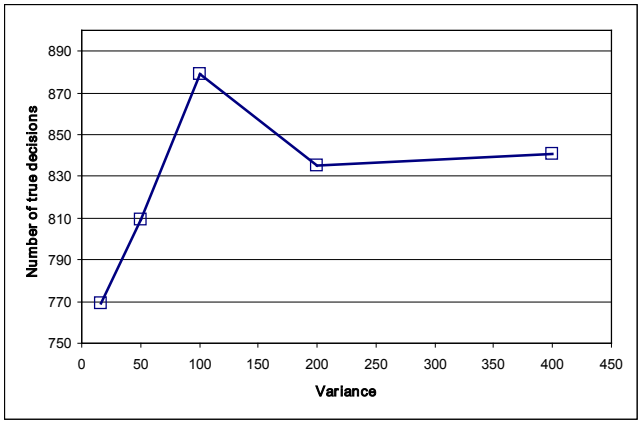


Figure 4. Number of true decisions for each variance σ_1^2

This conclusion is very important, for example, for the task of lossy image compression, because it allows, for a given image region, to estimate a level of losses that are visually unperceived (undetected) for HVS.

Forming distorted images with linearly decreased visual quality

Distortions which are visible in images with homogeneous regions, are often invisible in highly textured images. It depends on masking ability of image regions. For homogeneous regions, even noise with $\sigma^2=2$ is distinguishable. At the same time, noise with $\sigma^2=100$ can be invisible in images with contrast noise-like textures.

Note that regions with a larger level of local energy not always have greater masking ability. It was shown [6, 8, 9] that masking effect of image regions is proportional to dissimilarity (unpredictability) of these regions. This peculiarity of HVS is illustrated in Fig. 5.

Fig. 5, a shows a noise-free image. Dissimilarity map (RMSE of block matching procedure [6]) for this image is shown in Fig. 4, b. Image in Fig. 5, c contains the map of values of local standard deviations (for local variances calculated in 5x5 sliding window).

One can clearly see that unpredictability is higher for the right part of the image while local energies are much higher for the left part of the image.

According to these maps, we added to the image in Fig. 5 a signal dependent noise. Noise σ in image in Fig. 5, d for each pixel is directly proportional to RMSE in Fig. 5, b. Noise σ in image in Fig. 5, e for each pixel is directly proportional to standard deviations in Fig. 5, c. For both images, the overall variance is equal to 100.

One can see that in the image in Fig 5, d the noise is located mostly on stone textures and is low visible. For image in Fig 5, e, noise is mostly located on contrast but regular and well predictable textures in the left part of the image, and it is visually much higher. It is because HVS using neighbor textures is able to easily predict how the noise free textures should look.

Taking into accounting this peculiarity of HVS allows us to make the following very important conclusion. *Minimal visible level of noise for different image regions can be significantly different.* To estimate this level, one should be able to estimate masking ability (dissimilarity) separately for each image pixel.

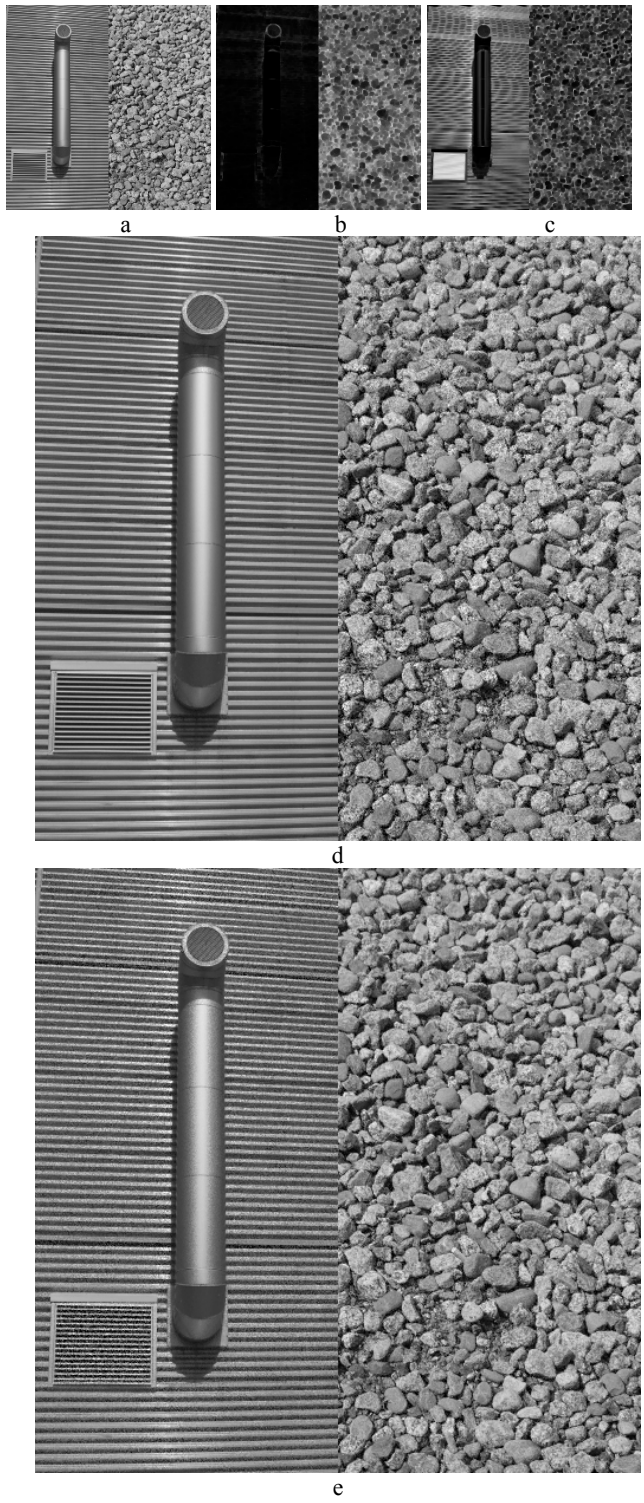


Figure 5. Illustration of masking ability of different regions

Calculation of masking ability map of image pixels

Let us calculate the map of pixels masking ability separately for each color component of a given image in YCbCr color space.

As the first step, for a given image color component A, the dissimilarity map D is calculated (in a sliding window 5x5, search

zone 21x21, size of area excluded from searching 3x3) [8]). A fast Matlab code for calculation of the map is available at http://ponomarenko.info/flt.htm.

Masking ability M(i,j) of a pixel with indexes i,j is calculated as

$$M(i,j) = \frac{\min\{D(i-2:i+2, j-2:j+2)\}}{4} + 3, \quad (1)$$

where 1/4 is the difference in variances providing confident perception of higher level of the noise (see conclusions in Section 2), 3 is a variance of the noise which is clearly visible in image homogeneous regions.

An example of calculated map M is given in Fig. 5, b (square root of values). HVS.

Sequence of calibration images with a linear decrease of visual quality

Let us (for a given reference image) create a SCI consisting of 20 distorted images with a linear decrease of visual quality from 100% to 0%.

The first image in the SCI is distorted by an additive noise with σ=0.3 (such a noise is invisible). The 20th image is distorted by an additive noise with σ=500 (image distorted by such a high level of the noise will be impossible to perceive). Remaining 18 images are distorted by signal dependent noise, where σ for each pixel is calculated as:

$$\sigma(i,j) = \sqrt{M(i,j)K(i,j)^{L-2}},$$
$$K(i,j) = \sqrt[18]{\frac{500^2}{M(i,j)}}, \quad (2)$$

where i,j are pixel indexes, L is an index of an image in the SCI (2..19).

Here, the coefficient K(i,j) defines a geometric progression for noise variances for the pixel and provides linearity of decreasing of visual quality. For homogeneous regions, K is equal to 1.77. For texture regions, K decreases depending on dissimilarity value of the region.

Examples of SCI built for different reference images can be found in http://ponomarenko.info/mmsp2019.html. It is clearly seen that for any reference image we have the same scale of visual quality.

Linearization of MOS of TID2013

For each reference image of TID2013 a corresponding SCI was created. Each observer in correspondence with the methodology described in [2] carried out pair-wise sorting of SCI images with 120 distorted images of this reference image. The main window of the software is shown in Fig. 6.

Each experiment takes, on the average, 5 minutes. Results of experiments for 427 observers are obtained. It is approximately 17 observers for each reference image. We have used the same methodology of calculation of MOS as in [2], rejecting 1% of results as abnormal.

Finally, for each of 25 reference images, there are 20 MOS values for SCI images in the scale of MOS of TID2013. It allows (by a simple linear interpolation) to rearrange MOS of all 120 distorted images of this reference image into a linear scale of visual quality of SCI images (0...10, where 0 is a minimal possible visual quality, 10 is a maximal possible visual quality).

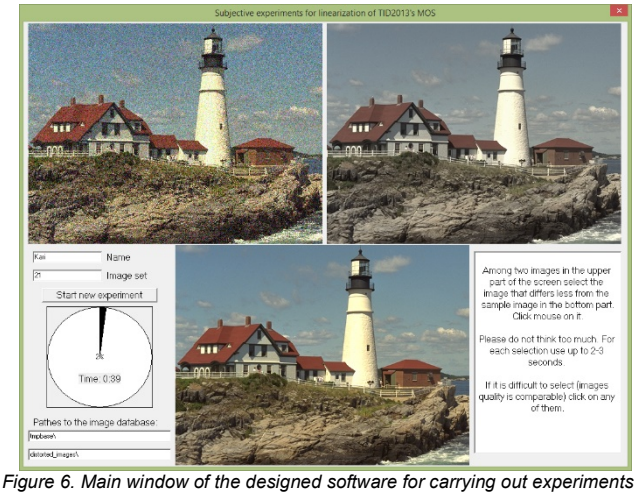


Figure 6. Main window of the designed software for carrying out experiments for linearization of MOS for TID2013

Fig. 7 shows histograms of MOS values of TID2013 before and after linearization. As it is seen, distribution of values has changed significantly.

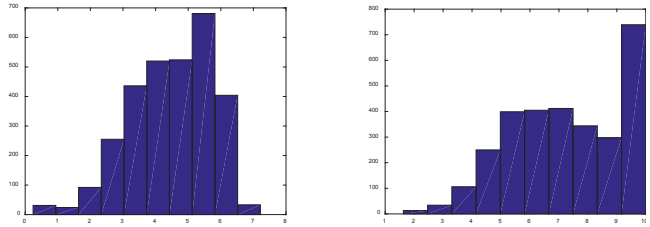


Figure 7. Histograms of MOS of TID2013 before and after the proposed linearization

The linearized MOS values of TID2013 will be available at *http://ponomarenko.info/tid2013.htm*.

Note that linearized MOS (due to linking to the fixed linear scale of visual quality) allows to add new reference images and new types of distortion to the test image database by small portions, step by step. Thus, the base becomes easily expandable.

To add to the database, for example, a reference image *B* and *N* distorted images, the following experiment should be carried out. We have to create SCI for the image *B*. According to [2], MOS of these *N*+20 images will be obtained. For *N*<20 one experiment will take not more than 6-7 minutes. Finally, MOS of *N* distorted images will be converted into linear scale using the obtained MOS of SCI images. After this, the reference image and *N* distorted images with linearized MOS can be added to the database.

Similar methodology may be used for a merger of existing databases into one larger database, which is able to provide more reliable metrics' verification.

Let us note that linearized MOS allows to calculate RMSE between linearized metrics and human perception (for TID2013 only rank order correlation coefficients can be used). It is possible also to calculate weighted RMSE giving larger weights to distortions important for a practical task.

Linearization of metrics values

Let us consider two examples of how metrics values can be linearized using linearized MOS of TID2013.

Fig. 8 shows the results of fitting the values of widely used SSIM metric [10] to a linearized MOS of TID2013 using "cftools" of Matlab.

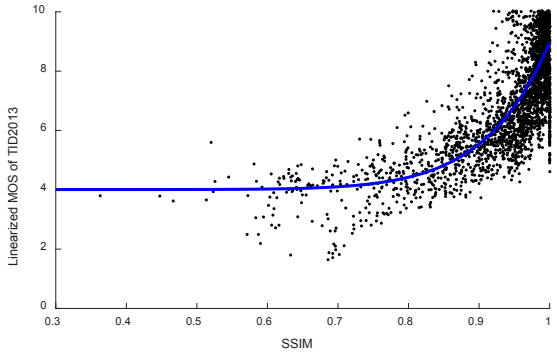


Figure 8. The results of fitting SSIM metric to linearized MOS of TID2013

Linearized SSIM can be calculated as

$$SSIM_l = 4.947\ SSIM^{11.11} + 3.977, \quad (3)$$

where 4.947, 11.11 and 3.977 are coefficients obtained as a result of the fitting.

It is interesting that RMSE between linearized MOS of TID2013 and $SSIM_l$ is equal to 1.11. Taking into accounting that the difference in visual quality between SCI images is well distinguished, and the difference corresponds to 0.5 (approximately) in the scale of linearized MOS, RMSE equal to 1.11 means the possibility of an error in estimation of visual quality by SSIM on up to 4 grades of visual quality distinguished for human perception.

Fig. 9 shows the results of fitting the values of PSNR-HMA metric [11] to a linearized MOS of TID2013.

Linearized PSNR-HMA can be calculated as

$$PSNR\text{-}HMA_l = 0.2264\ PSNR\text{-}HMA - 0.4754, \quad (4)$$

where coefficients 0.2264 and 0.4754 are obtained as a result of the fitting. Note, that the equation (4) is linear. This proves that mean square error in a logarithmic scale (the base for calculation of PSNR-like metrics) is almost linear for human perception.

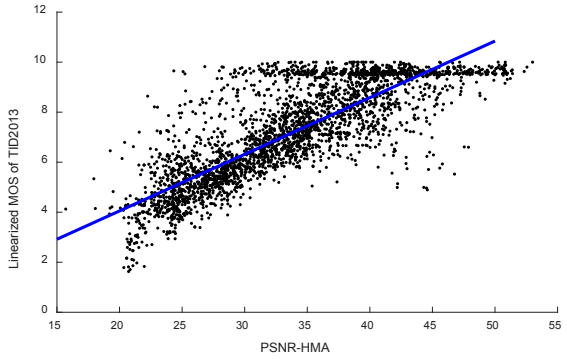


Figure 9. Fitting PSNR-HMA metric to linearized MOS of TID2013

RMSE between linearized MOS of TID2013 and $PSNR\text{-}HMA_l$ is equal to 1.15.

Conclusions

In the paper, we explored important peculiarities of HVS, in particular, its ability to distinguish difference in levels of white additive Gaussian noise in images. An effective algorithm of forming SCI providing sequence of images with linear changing of visual quality was proposed. Experiments by linearization of MOS for TID2013 test image database have been carried out. It was demonstrated that obtained results allow to add new types of distortions to the database as well as to linearize metric values.

Acknowledgement

This work is supported by the FlexISP project.

References

[1] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms, IEEE Transactions on Image Processing, 15 (2006) 3440-3451.

[2] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C.Jay Kuo, "Image database TID2013: Peculiarities, results and perspectives", Signal Processing: Image Communication, Vol. 30, pp. 55-77, 2015.

[3] ITU, Methodology for the subjective assessment of the quality of television pictures, in: Recommendation BT.500-11, Geneva, Switzerland, 2002.

[4] H. Lin, V. Hosu, D. Saupe, KADID-10k: A Large-scale Artificially Distorted IQA Database, accepted to Eleventh International Conference on Quality of Multimedia Experience, 2019, 3p.

[5] M.G. Kendall, The advanced theory of statistics. Vol. 1, Charles Griffin & Company limited, London, UK, 1945.

[6] O.I. Ieremeiev, N.N. Ponomarenko, V.V. Lukin, J.T. Astola, K.O. Egiazarian, "Masking effect of non-predictable energy of image regions", Telecommications and Radio Engineering, 76 (8), pp. 685-708, 2017.

[7] D.M. Mackay, "Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' laws", Science, pp. 1213–1216, 1963.

[8] K. Egiazarian, M. Ponomarenko, V. Lukin, O. Ieremeiev. "Statistical evaluation of visual quality metrics for image denoising", International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5 p, 2018.

[9] M. Ponomarenko, K. Egiazarian, V. Lukin, V. Abramova, "Structural Similarity Index with Predictability of Image Blocks", International Conference on Mathematical Methods in Electromagnetic Theory (MMET), pp. 115-118, 2018.

[10] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, 13(4), pp. 600–612, 2004.

[11] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, M. Carli, Modified image visual quality metrics for contrast change and mean shift accounting, in: Int. Conf. The Experience of Designing and Application of CAD Systems in Microelectronics, 2011, pp. 305-311.

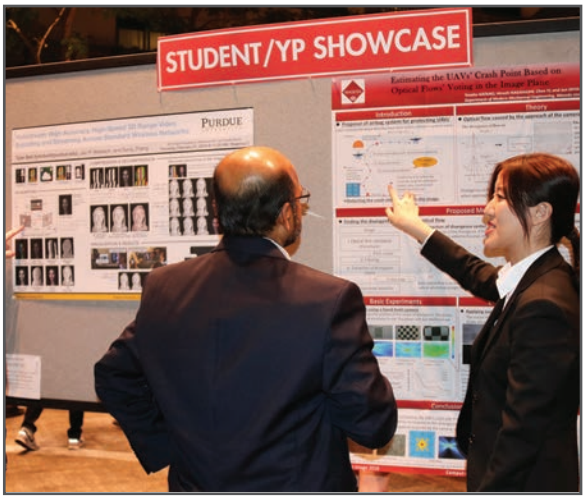
JOIN US AT THE NEXT EI!

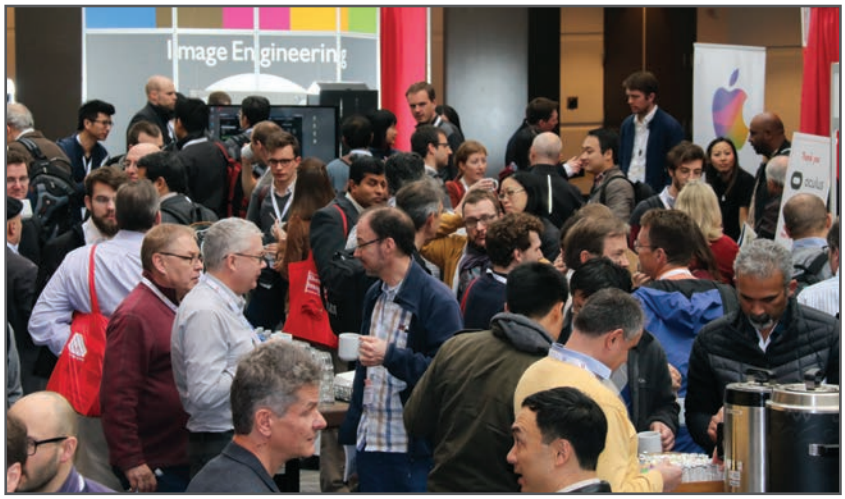
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org