https://doi.org/10.2352/ISSN.2470-1173.2019.12.HVEI-214

An Improved Objective Metric to Predict Image Quality using Deep Neural Networks

Pinar Akyazi and Touradj Ebrahimi; Multimedia Signal Processing Group (MMSPG); Ecole Polytechnique Fédérale de Lausanne; CH 1015, Lausanne, Switzerland

Abstract

Objective quality assessment of compressed images is very useful in many applications. In this paper we present an objective quality metric that is better tuned to evaluate the quality of images distorted by compression artifacts. A deep convolutional neural networks is used to extract features from a reference image and its distorted version. Selected features have both spatial and spectral characteristics providing substantial information on perceived quality. These features are extracted from numerous randomly selected patches from images and overall image quality is computed as a weighted sum of patch scores, where weights are learned during training. The model parameters are initialized based on a previous work and further trained using content from a recent JPEG XL call for proposals. The proposed model is then analyzed on both the above JPEG XL test set and images distorted by compression algorithms in the TID2013 database. Test results indicate that the new model outperforms the initial model, as well as other state-of-the-art objective quality metrics.

Introduction

Continuous improvements in digital imaging and video keep motivating broadcasters and service providers to supply contents of superior visual quality to their viewers, despite larger storage and transmission requirements. Compression is used to reduce such resources. The principal trade-off in compression is between perceived quality of the compressed content and transmission rate. The goal of efficient state-of-the-art compression algorithms is to achieve lower rates while maintaining the quality. Compression related artifacts such as blur, blocking, ringing and change in contrast can distort a content and reduce its perceived quality. It is therefore important for content providers to be able to anticipate the degree of annoyance caused by such distortions, and to optimize their systems accordingly.

Quality assessment methods provide means of analyzing a content either subjectively or objectively. Subjective quality assessment methods employ human subjects and evaluate the quality of contents by collecting viewers ratings. While this is the most reliable form of measuring the quality of contents, the process is highly impractical in terms of time and labor costs. Objective quality assessment methods are more practical since they rely on mathematical models to evaluate the degradation and overall quality of contents. However, ensuring a good correlation between objective metrics and subjective ratings is a challenging problem. The general interest in image quality assessment (IQA) is to introduce an objective image quality metric (IQM) that is able to determine the quality of an image with high accuracy, i.e., with high correlation to the would-be perceived quality.

The simplest measure to assess the quality of distorted images in a full reference (FR) framework, i.e. when the reference image is available, is a difference-based metric called the peak signal to noise ratio (PSNR). PSNR and its derivatives do not consider models based on the human visual system (HVS) and therefore often result in low correlations with subjective quality ratings. [1]. Metrics such as structural similarity index (SSIM) [2], multi-scale structural similarity index (MS-SSIM) [3], feature similarity index (FSIM) [4] and visual information fidelity (VIF) [5] use models motivated by HVS and natural scenes statistics, resulting in better correlations with viewers' opinion.

Numerous machine learning based objective quality metrics have been reported in the literature. Weighted Average Deep Image Quality Measure for FR-IQA (WaDIQaM-FR) [6] is an end-to-end trained method that extracts features from reference and distorted image patches using convolutional filters. Recently, we have proposed a new model inspired by WaDIQaM-FR architecture that will be referred to as MTID2013 in the reminder of this paper [7]. Besides changes in the architecture and learning parameters when compared to those used in [6], MTID2013 incorporates both spatial and spectral information in its extracted features and consequently delivers more accurate results.

To the best of the authors' knowledge, none of the methods cited have been specially adjusted to compression artifacts, but trained using many different distortions in addition to compression, such as various types of noise, distortions due to quantization as well as transmission and sampling errors [8]. In this paper, we propose a FR-IQA based on deep convolutional neural network that is able to objectively predict the quality of distorted images suffering particularly from compression artifacts. The same architecture as in [7] is used and further trained with a new database that contains compression-based distortions only. This will make it more likely for the final model to correlates with subjective ratings when compared to the previous model, when evaluating the quality of compressed images.

The rest of the paper is structured as follows. The proposed framework is described in details in the next section. Experiments and results are reported after the framework description. The paper ends by providing an overview on the performance, followed by discussions as well as future directions.

Proposed Framework

The proposed framework is identical to the architecture introduced in [7]. Features extracted from both a reference image and its distorted version are concatenated into a single feature vector that is passed onto the fully connected layers for regression and an objective quality score is assigned to the output. We used a Siamese network to extract features from both the reference and distorted images as in [6], whereas the design of convolutional layers and the preferred building blocks have been changed.

Both color images and the wavelet decomposition of their grayscale versions up to three scales were used as input. 2-D wavelet decomposition is known to be effective in image processing tasks such as denoising, interpolation, sharpening and compression by providing information about both the spatial and the frequency components of an image in different scales. The compression related distortions in natural image databases affect distinct frequency areas in images differently. It is therefore important to analyze differences in both high frequency components and low frequency approximations of the reference and distorted images. Hence, the discrete wavelet transform of the reference and the distorted images were computed up to three scales using Daubechies wavelets and the resulting coefficients were used as features [7].

The architecture in VGGnets [9] was used to build convolutional layers and added residual connections resulting in a model that is easier to optimize and exhibits lower training error when the depth increases [10]. The input images are divided into N_p randomly selected patches. The dimensions of each patch were 128×128 pixels, thereby resulting in wavelet decompositions of size 64×64, 32×32 and 16×16. Images were normalized prior to network processing. The proposed VGGnet inspired residual convolutional layers are comprised of 8 to 10 weight layers with 3 to 4 shortcut connections for wavelet coefficient inputs and color patch inputs, respectively. The features are extracted using a series of 3x3 conv 32, 3x3 conv 32, 3x3 conv 64, 3x3 conv 64, 3x3 conv 128, 3x3 conv 128, 3x3 conv 256, 3x3 conv 256, with an addition of 3x3 conv 512, 3x3 conv 512 for the color input. The shortcut connections for residual architecture are established by 1x1 convolutional filters of size 64, 128 and 256, with an additional filter of size 512 for the color input. The downsampling is performed by using convolutional layers of stride 2 instead of pooling. At the end of each branch a 1x1 convolutional layer was used with 16 filters to reduce the output size. Further dimensional reduction was performed for the branches with input size greater than 16×16 by using max pooling. All max pooling layers have 2×2-size kernels. The output of each branch is then concatenated to form the final features vector, as shown in figure 1. The convolutional layers of same output size are activated through a leaky rectified linear unit (Leaky ReLU) where $\text{LeakyReLU}(x) = \max(0,x) + 0.01 \times \min(0,x)$. This activation function allows a small non-zero gradient when the unit is not active, thereby preventing all outputs from reducing to zero. Instead of random initial weights, those used in the model MTID2013 were used[7]. This model was trained and tested on the TID2013 image database [8] which contains 24 distortion types, of which only 2 are compression related.

The complete architecture of the proposed network is depicted in Figure 2. Following the feature extraction of both reference and distorted image patches, the distorted image features f_D are concatenated with the reference image features f_R. Moreover, we also add the difference vector $f_D - f_R$ as the accuracy of the model is reported to increase by using this configuration [6]. The features vectors are then passed through two fully connected layers for regression, FC 256 and FC 1. Between these layers the Leaky ReLU activation prior is used to dropout regularization with a ratio of 0.5 in order to prevent over fitting[11]. The features vector is separately fed into two fully connected layers to compute local patch weights. The architecture of this block is the same with the output regression layer, FC 256 and FC 1. Between these layers, ReLU activation prior to dropout with a ratio of 0.5 is used. Furthermore, a final ReLU activation is applied before weight computation, in order to ensure that weights are greater than or equal to zero. Afterwards a small constant $\varepsilon = 1\text{e}{-6}$ is added to the weights to prevent zero weights.

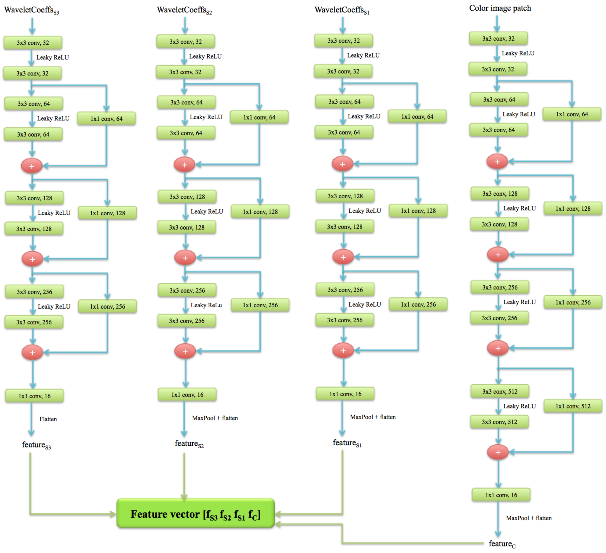


Figure 1. *Feature extractor composed of convolutional layers, as previously shown in Figure 2 as the CNN block. Inputs of the first three branches from left to right are the wavelet coefficients of the 128×128 image patch, where S3 corresponds to the coarsest and S1 corresponds to the finest scale. The rightmost branch is the color image patch branch. Features are extracted using a VGGnet inspired architecture involving shortcut connections and 1x1 convolution at the end for dimensional reduction. Max pooling is also applied when necessary. Features vectors of four branches are concatenated into a final features vector of the input image patch.*

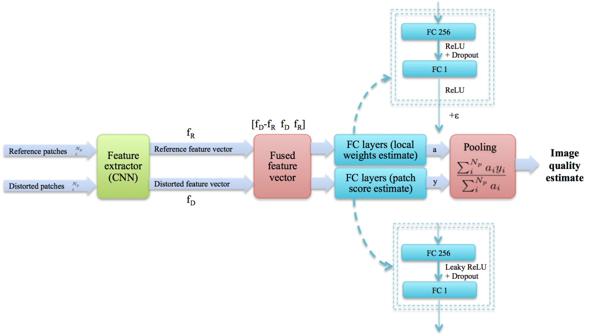


Figure 2. *The proposed framework for training and testing our model. Features are extracted from both reference and distorted image patches, using color information and wavelet decomposition. The reference and distorted features vectors are concatenated, also with a third difference vector. The final features vector is passed through parallel fully connected layers for local weight estimation and patch score estimation. Overall score of each image is computed as a linear combination of the weighted patch scores.*

For an input patch i, the computed weight is a_i such that $a_i = \max(0, a_i^*) + \varepsilon$ where a_i^* is the output prior to ReLU activation. The quality of patch i is computed in the parallel regression branch as y_i. The overall image quality is then computed as a linear combination of patch qualities and patch weights:

$$\hat{q} = \frac{\sum_i^{N_p} a_i y_i}{\sum_i^{N_p} a_i} \tag{1}$$

The mean squared error is used between the computed image quality and the ground truth, i.e., the mean opinion score (MOS) rating of the image as our loss function. The proposed network is trained iteratively by back propagation [12, 13] over a number of epochs using batch-wise optimization until the error is stabilized. In each batch, one image is used, from which $N_p = 128$ patches are extracted. Data augmentation is carried out by flipping each image from left to right and choosing additional $N_p = 128$ patches from each of the flipped images. As was done in [6, 7], patches are randomly sampled in every epoch to introduce as many different inputs as possible to the network during training. The ADAM method [14] is used for batch optimization with the recommended parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$ and a decaying learning rate starting from $\mathrm{lr} = 10^{-4}$ with a decay percentage of 10% every 5 epochs. The loss is computed on a separate validation set at the end of each epoch, where the validation set is defined at the beginning of the algorithm instead of choosing random patches at every epoch in order to ensure stability. The final model used for accuracy tests is the model with least validation error.

Experiments and Results

Datasets

The first model MTID2013 was trained on the TID2013 database. Here, the parameters of MTID2013 were used for initialization and training was further continued using the JPEG XL database [15].

The TID2013 database contains 25 different contents represented by 3000 distorted images of the same resolution, i.e. 512×384, in which for each reference image there are 24 types and 5 levels of distortions. A wide spectrum of distortion types have been included in this database, including additive Gaussian noise, Gaussian blur, high frequency noise, quantization noise and sparse sampling and reconstruction. JPEG and JPEG 2000 coding are the only compression related distortions within the TID2013 database. MOS values of the database lie in the range $[0,9]$ with 0 being the lowest quality score and 9 the highest. 15 reference images and their corresponding distorted versions were used for training, 5 left for validation and the remaining 5 for testing.

The JPEG XL database contains 7 different contents represented by 305 distorted images, where for each reference image there are 11 types and 3 to 4 levels of distortions. The resolution of images are typically varying between HD and UHD. All 11 types of distortions result from compression algorithms, i.e. use of different codecs, including JPEG [16], JPEG 2000 [17], HEVC/H.265 [18] and WebP [19] anchors along with seven proponents who submitted new compression algorithms for consideration. Subjective experiments have been conducted at Ecole Polytechnique Fédérale de Lausanne (EPFL) with the participation of 18 consenting subjects in a controlled environment [20], using the Double Stimulus Impairment Scale (DSIS) Variant I to compare the subjective quality of different coding schemes [21]. MOS values of the database lie in the range $[1,5]$ with 1 being the lowest quality score and 5 the highest. We have mapped these scores to the range of TID2013 scores, i.e. $[0,9]$, assuming a linear mapping. The JPEG XL dataset was separated into training, validation and test sets randomly, using 4, 1 and 2 images respectively.

Results on JPEG XL Database

The model was trained on the JPEG XL image database using a 5-fold cross validation. The total number of epochs was 140 for each fold, and the model with the lowest validation loss was selected as the final model. The averaged training and validation losses are shown in Figure 3. We can see that the validation error keeps decreasing at a very slow pace over the epochs, after a steep fall within the first five epochs. Table 1 presents the performance comparison of tested objective metrics in terms of the PLCC and SROCC values with respect to the MOS values of each image, averaged over all test images. We have used the same test images to evaluate the PSNR, MS-SSIM, FSIMc (c stands for color), MTID2013 and MJPEGXL metrics. Figure 4 depicts the performance improvement of MJPEGXL over MTID2013. The linear fitting shows that MJPEGXL scores and the MOS correlate highly whereas MTID2013 scores employ a flatter trend.

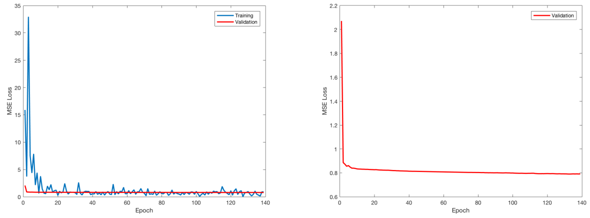


Figure 3. *Training and validation losses of the proposed model (left) and an enlarged plot of the validation loss (right) over 140 epochs.*

Performance comparison of objective quality metrics PSNR, MS-SSIM, FSIMc, MTID2013 and MJPEGXL in terms of PLCC and SROCC on the JPEG XL test set.

IQM	PLCC	SROCC
PSNR	0.7433	0.7132
MS-SSIM	0.7410	**0.8413**
FSIMc	0.5809	0.7770
MTID2013	0.7358	0.7791
MJPEGXL	**0.7505**	0.7395

Results on TID2013 Database

We have also tested our final model on the TID2013 test set, which is the same test set used in [7]. Here we only include images that have been distorted by compression-related artifacts, i.e. JPEG and JPEG 2000 compressed images. We have a total of 50 test images from the TID2013 test set, associated with 5 different reference images. Instead of using $N_p = 128$ as we did for the high-resolution JPEG XL test images, we reduced N_p to 32, as was done during the training of the initial model. Table 2

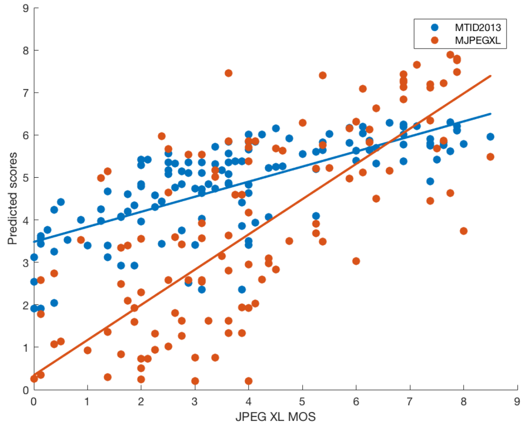


Figure 4. *Scatter plot of predicted scores vs. MOS on JPEG XL test set, with linear fitting for MTID2013 and MJPEGXL models.*

shows that the performance of our new model is superior to the initial model on the compressed images test set, which illustrates that our new model is better-suited for evaluating the quality of compressed images in both TID2013 and JPEG XL databases. This improvement is also depicted in Figure 5, where the linear fitting on the predicted scores using MJPEGXL correlates more with the underlying MOS when compared to MTID2013. An example image from TID2013 test set is included in Figure 6 with a maximum difference of 0.58 between the predicted scores and underlying MOS.

Performance comparison of objective quality metrics MTID2013 and MJPEGXL in terms of PLCC and SROCC on the images with compression artifacts in the TID2013 test set.

IQM	PLCC	SROCC
MTID2013	0.8725	0.8743
MJPEGXL	**0.8975**	**0.89545**

Conclusion

In this paper we have introduced a new objective metric for full reference image quality assessment. Our metric is trained using deep convolutional neural networks and tuned to predict the quality of images distorted by various compression artifacts. Our model parameters are initiated using a previous objective metric that is also learning-based, but trained on a dataset containing numerous types of distortions. We further trained our previous model on the JPEG XL dataset, which is composed of 7 contents and their compressed versions using 11 different codecs and 4 different bitrates. Results indicate that our new model is able to predict the quality of compressed images in the JPEG XL and TID2013 database test sets effectively, with higher correlation compared to our previous model and various state-of-the-art metrics.

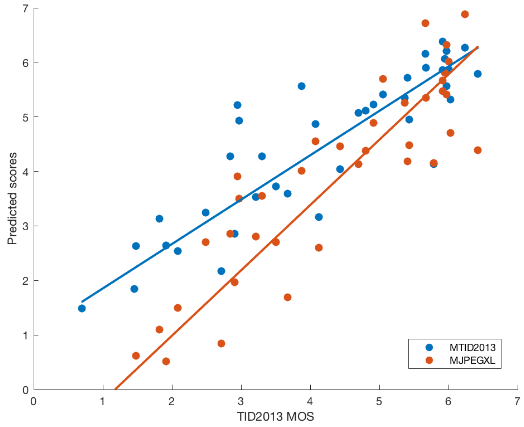


Figure 5. *Scatter plot of predicted scores vs. MOS on the images with compression artifacts in the TID2013 test set, with linear fitting for MTID2013 and MJPEGXL models.*

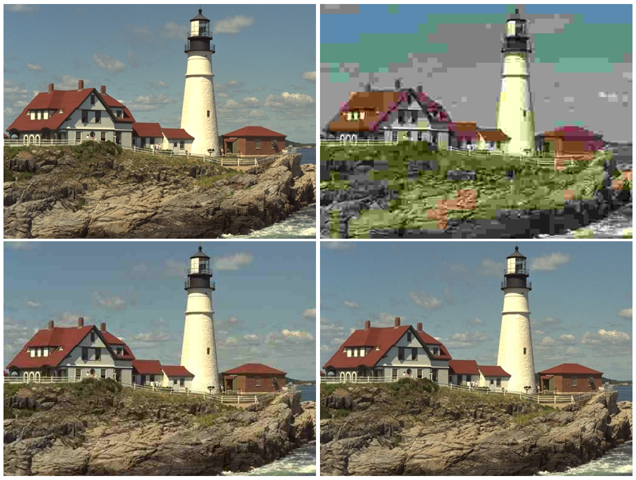


Figure 6. *An example reference image from TID2013 test set (top left) compressed using JPEG at distortion levels 5 (top right), 3 (bottom left) and 1 (bottom right). The scores predicted by MJPEGXL and underlying MOS are* $[1.50, 4.89, 6.32]$ *and* $[2.08, 4.92, 5.94]$*, respectively.*

In order to improve the accuracy of our model, it is important to find a good mapping between the initial training database and the JPEG XL database. Since the resolution of JPEG XL contents are roughly five times larger than the TID2013 database, the features extracted from the patches of same dimensions from the two databases are expected to differ. To address this issue, multiple resolutions of JPEG XL database can be incorporated into training, however scaling of the data could impair the contents and therefore cause the subjective ratings to become obsolete. We could also improve our results by assigning a more accurate mapping of the scores and distortion levels between the two databases, in comparison with linear mapping. Enhancement of the training set by including more contents and distortion levels would certainly increase the performance. To help the patch weighing in our model, we can also introduce robust saliency models. We plan to optimize our feature extraction scheme by using regression for aggregation of features extracted from the reference and distorted image patches, as well as extend our framework to image sequences to build an end-to-end model for predicting the quality of compressed video.

Acknowledgments

This paper reports a research performed under the framework of project Digital Eye: Deep Learning Video Quality Assessment Technology, funded by The Swiss Commission for Technology and Innovation (CTI) under the grant 27403.1 PFES-ES.

References

[1] W. Lin and C.-C. J. Kuo, “Perceptual visual quality metrics: A survey,” *Journal of Visual Communication and Image Representation* **22**(4), pp. 297–312, 2011.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing* **13**(4), pp. 600–612, 2004.

[3] Z. Wang, E. Simoncelli, A. Bovik, *et al.*, “Multi-scale structural similarity for image quality assessment,” in *ASILOMAR CONFERENCE ON SIGNALS SYSTEMS AND COMPUTERS*, **2**, pp. 1398–1402, Citeseer, 2003.

[4] L. Zhang, L. Zhang, X. Mou, D. Zhang, *et al.*, “Fsim: a feature similarity index for image quality assessment,” *IEEE transactions on Image Processing* **20**(8), pp. 2378–2386, 2011.

[5] H. R. Sheikh and A. C. Bovik, “A visual information fidelity approach to video quality assessment,” in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 23–25, 2005.

[6] S. Bosse, D. Maniry, K. R. Mller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing* **27**, pp. 206–219, Jan 2018.

[7] P. Akyazi and T. Ebrahimi, “A new objective metric to predict image quality using deep neural networks,” in *Applications of Digital Image Processing XLI*, **10752**, p. 107521Q, International Society for Optics and Photonics, 2018.

[8] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, *et al.*, “Image database tid2013: Peculiarities, results and perspectives,” *Signal Processing: Image Communication* **30**, pp. 57–77, 2015.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556* , 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research* **15**(1), pp. 1929–1958, 2014.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE* **86**(11), pp. 2278–2324, 1998.

[13] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, “Efficient backprop,” in *Neural networks: Tricks of the trade*, pp. 9–48, Springer, 2012.

[14] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980* , 2014.

[15] ISO/IEC JTC 1/SC 29/WG1 N80024, “Jpeg xl: Additional information to the final call for proposals,” *80th JPEG meeting, Berlin, Germany* , 7-13 July 2018.

[16] JPEG XT reference software, v1.53. Available at http://jpeg.org/jpegxt/software.html.

[17] Kakadu v7.10.2. Available at http://www.kakadusoftware.com.

[18] “HEVC reference software HM 16.18+SCM-8.7.” Available at https://hevc.hhi.fraunhofer.de/svn/svn__HEVCSoftware/tags/HM-16.18+SCM-8.7/.

[19] WebP v1.0.0-rc2. Available at https://developers.google.com/speed/webp/download.

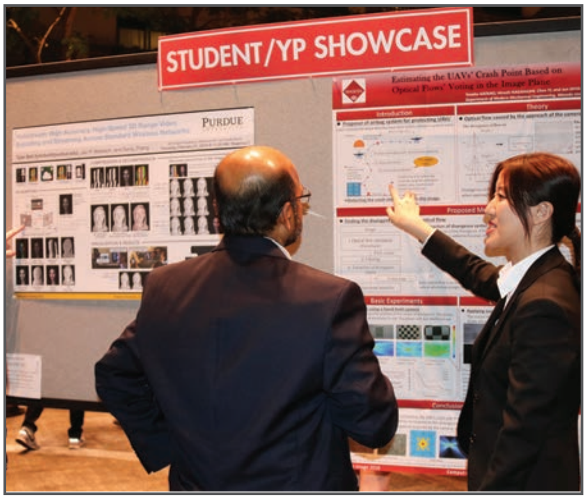
[20] ITU-R BT.2022, “General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays,” *International Telecommunication Union* , August 2012.

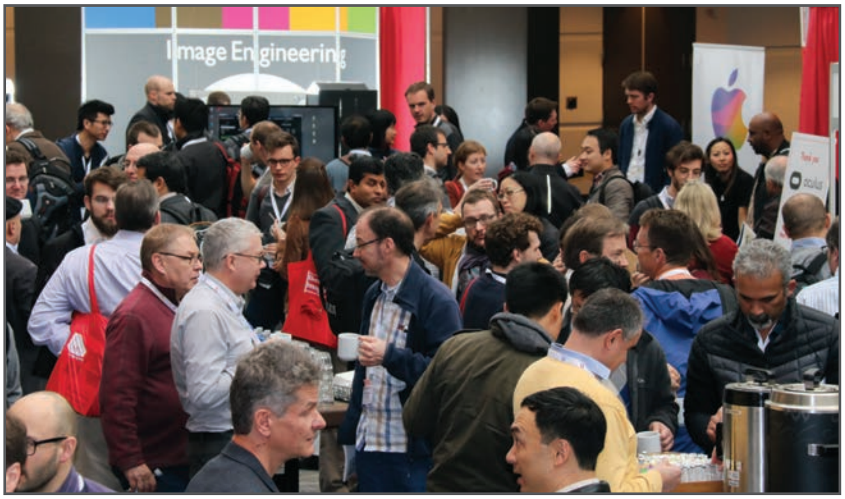
[21] ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures,” *International Telecommunication Union* , January 2012.

Author Biography

Pinar Akyazi received her B.S. and M.Sc. in Electrical and Electronics Engineering from Bogazici University, Turkey, in 2010 and 2013, respectively. Since 2018 she has been working as a PhD student in Multimedia Signal Processing Group of EPFL. Her current research is focused on learning-based image quality assessment and learning-based image compression.

Touradj Ebrahimi is currently Professor at EPFL heading its Multimedia Signal Processing Group. He is also the Convener of JPEG standardization Committee. His research interests include still, moving, and 3D image processing and coding, visual information security (rights protection, watermarking, authentication, data integrity, steganography), new media, and human computer interfaces (smart vision, brain computer interface). He is the author or the co-author of more than 200 research publications, and holds 14 patents.





- AR is a fundamental and transformative interface and platform.
- 20+ years from my initial working demo to today's AR displays
- Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

My first working AR prototype

Today's platforms

Optimis… this is << 20

Ubiquitous consumer AR