https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-373

Inkjet Quality Ruler Experiments and Print Uniformity Predictor [1]

Yi Yang [a], Utpal Sarkar [b], Isabel Borrell [b], Jan P. Allebach [a]
[a]. **Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA.**
[b]. **HP Inc., Sant Cugat del Valles, SPAIN.**

Abstract

Macro-uniformity is an important factor in the overall quality of prints from inkjet printers. The International Committee for Information Technology Standards (INCITS) defined the macro-uniformity for prints, which includes several printing defects such as banding, streaks, mottle, etc. Although we can quantitatively analyze a certain kind of defect, it is difficult to assess the overall perceptual quality when multiple defects appear simultaneously in a print.

We used the Macro-uniformity quality rulers designed by INCITS W1.1 as experimental references, to conduct a psychophysical experiment for pooling perceptual assessments of our print samples from subjects. Then, calculated features can describe the severity of defects in a test sample; and we trained a predictive model using these data. The predictor can automatically predict the macro-uniformity score as judged by humans.

Our results show that the predictor can work accurately. The predicted scores are similar to the subjective visual scores (ground-truth). Also, we used 6-fold cross-validation to confirm the efficacy of our predictor.

Introduction

The $INCITS\ W1.1$ activity recognizes that printed image quality can be well described by a small set of attributes, including gloss uniformity, macro-uniformity, and micro-uniformity. Numerous recent papers show that among these attributes, macro-uniformity draws the most attention [1].

Macro-uniformity (ISO 19751 macro-uniformity) refers to the subjective impression of color uniformity across a large image area that is intended to have a uniform color. There are several kinds of print quality defects that influence the percept of macro-uniformity [1], [2]. They are:

- Banding: one-dimensional, periodic lightness and/or chromatic variations.
- Streaks: one-dimensional, isolated lightness and/or chromatic variations.
- Mottle: two-dimensional, random lightness and/or chromatic variations.
- Graininess: two-dimensional, fine-scale, random texture with a sand-like appearance.
- Mottle: two-dimensional, medium-scale, random lightness variations.
- Large area variation: two-dimensional, random lightness variations, the spatial region is larger than for mottle.
- Large-scale non-uniformity: one-dimensional, low-frequency lightness variations.

These defects are very important to print quality, yet it is difficult to evaluate the overall macro-uniformity when they occur simultaneously.

The $INCITS\ W1.1$ macro-uniformity team developed a method to measure overall macro-uniformity. They created macro-uniformity quality ruler samples by imposing increasing levels of non-uniformity in a synthetic defect pattern, which consists of a multitude of normally occurring defect types. Experiments were conducted to calibrate the quality ruler in terms of just noticeable differences (JND) [2].

Quality rulers labeled 3, 6, 9, 12, 15, 18, 21, 24, 27, and 30 can be generated by the macro-uniformity software. The smaller the label (JND) is, the better the print quality is. The first quality ruler (JND = 3) would appear nearly perfect, with only minor defects. The perceived defect level is approximately logarithmic with the amplitude of the defects, and all levels are visually equidistant according to the quality of the print samples.

With quality rulers, according to the ISO 20462-3 international standard, a psychophysical experiment for estimating printing quality can be designed [3]. The quality ruler method is superior to many other psychophysical methods, because it can assess a large range of printing quality levels with relatively few resources.

In this paper, we complete the following three tasks:

1. Compute a value for each defect in the macro-uniformity set, which can represent the severity of the defect.

2. Design and conduct a psychophysical experiment according to the ISO 20462-3 international standard, which includes the selection of test samples, the calibration of the printer and scanner, and the environmental preparation and detailed guidance during the experiment.

[1] Research supported by HP Inc., Barcelona, SPAIN

3. Analyze the data and built prediction models. We use Linear Regression and *SVM* to build prediction models which can predict the subjective assessment of the *JND* based on the objective defect values calculated for the test print.

Macro-Uniformity Quality Ruler

To print the samples for the image quality ruler, we used an *Epson Stylus Pro* 3880 Color Inkjet Printer[2], which is a photo quality inkjet printer, as recommended by the *INCITS W*1.1 macro-uniformity team. Printer calibration is required before generating and printing the quality ruler. We first used the W1.1 macro-uniformity software [4] to generate the printer calibration file. An image of the calibration page is shown in Figure 1.

A companion file for the calibration page contains a table of input values for the patches in the generated calibration test pattern. The first column in this file contains an arbitrary index. The second column is the row-index of the test patch. The third column is the column-index of the test patch. The fourth column is the input value to the printer, as a CIE Y value from 0 to 100, before mapping to an 8 *bit* value from 0 to 255.

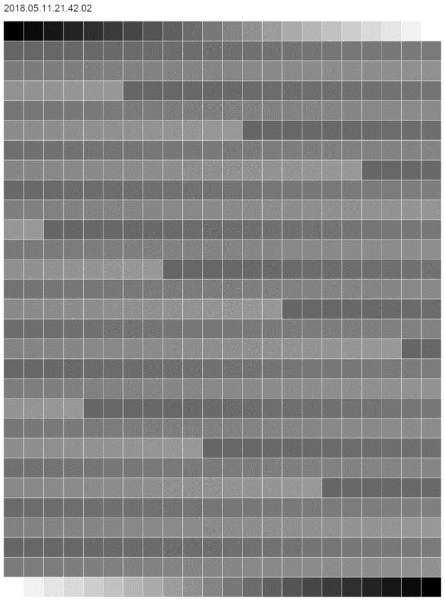

Figure 1. *The Calibration Pattern.*

This file needs to be printed using the same printer that is going to be used to print quality rulers. We then measured the CIE XYZ values of each patch on the printed file using an *X-Rite DTP*-70 [3] color measurement instrument. The measuring process can be repeated several times to minimize the effect of noise. The final step to complete printer calibration is to arrange all the patches' average CIE XYZ values into a specific format file and input it into the $W1.1$ macro-uniformity software.

The software automatically calibrates the CIE Y data according to the input file, and then generates quality ruler samples with specific defect scales.

Each quality ruler sample includes a 170 *mm* × 170 *mm* defect region and a test target surrounding the defect region. As shown in Figure 2, the quality rulers are created at fixed quality levels ranging from highest to lowest in steps that are 3 *JNDs* apart. Quality ruler samples labeled 3, 6, 9, 12, 15, 18, 21, 24, 27, and 30 can be generated by the software. But according to the severity of the defects in the test samples, the appropriate range should be chosen. (Quality rulers with *JND* from 3 to 21 were used in our work.)

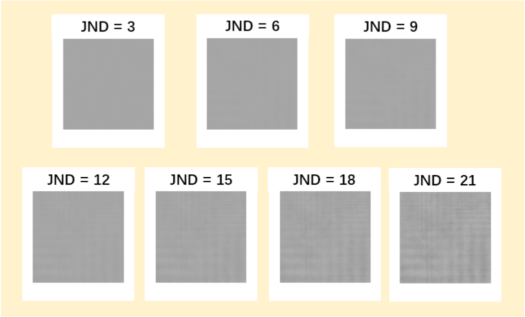

Figure 2. *Quality ruler samples labeled with JND = 3, 6, 9, 12, 15, 18, and 21. The smaller the label (JND) is, the better the print quality is. The first quality ruler (JND = 3) would appear nearly perfect.*

Compute Defect Features in the Macro-Uniformity Test Set

Our test samples were printed at four different tint levels with a prototype large-format printer using a page-wide array inkjet print bar. Print defect features such as banding, streaks, graininess, etc., could be seen in the print samples. The framework for computing the defect features includes scanning hard-copy test samples, removing halftoning pattern using a descreening processing and then computing values that represent the severity of defect features as described below.

First, scanner calibration needs to be performed before scanning the test samples. All test samples were scanned at 600 *dpi* resolution, as recommended by a previous study [5]. We used an *Epson Expression* 10000*XL* scanner [2]. The scanner calibration was performed as described in [6], [7]. A *Kodak Q*60 reflective target was scanned, and an *X-Rite DTP*-70 [3], was used to determine the CIE XYZ values of the patches. The gray patches were

[2] Epson America, Inc., Long Beach, CA.

[3] X-Rite, Inc., Grand Rapids, MI.

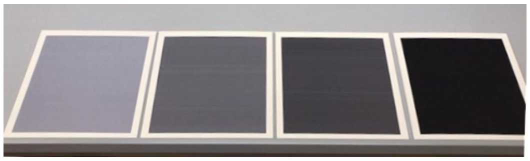

Figure 3. *Print samples with tint levels 30%, 50%, 70%, and 100% were printed with a prototype large-format printer using a page-wide array inkjet print bar. We selected 12 samples from levels 30%, 50% and 70%, and 6 samples from level 100%, resulting in a total of 42 test samples.*

used to determine the gray-balance curves for each of the R, G, and B scanner channels. Then, the 240 color patches were used to determine the elements of a 3×3 matrix used to transform from linear scanner RGB to CIE XYZ. Finally, we transform from CIE Y to CIE $L*$ to complete the transformation that is applied to the monochrome test pages.

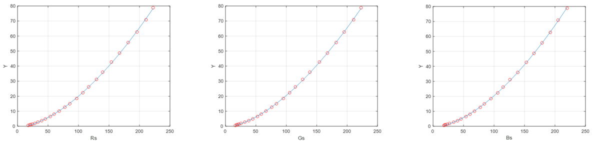

Figure 4. *Gray balancing curves for the R, G, B channels.*

The Gray balancing results are:

$$R_l = 99.4942(\frac{R_s}{255})^{1.6821} - 0.6268 \tag{1}$$

$$G_l = 98.5282(\frac{G_r}{255})^{1.6542} - 0.4967 \tag{2}$$

$$B_l = 102.6936(\frac{B_r}{255})^{1.7003} - 0.5815 \tag{3}$$

The transformation from Linear RGB to CIE XYZ is :

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.3628 & 0.3310 & 0.1875 \\ 0.1137 & 0.1407 & 0.1901 \\ 0.1819 & 0.1846 & 0.1197 \end{bmatrix} \begin{bmatrix} R_l \\ G_l \\ B_l \end{bmatrix}$$

A human vision model is then applied to the scanned samples to measure defects as perceived by a human subject. The contrast sensitivity function (CSF) we used in the human vision model was proposed by Mannos and Sakrison [8]. The viewing distance of the CSF is set to 15.7 *inches* (approximately 40 *cm*), which is also the viewing distance recommended by the $INCITS$ $W1.1$ working group for conducting psychophysical experiments using the quality ruler method.

Figure 5 and Figure 6 show a 70% tint test sample before and after filtering with our human vision model.

After applying the human vision model, we converted the samples from the RGB color space to the CIE $L^*a^*b^*$ space. Since our test samples are all printed in grayscale, we only use the lightness channel (L^* channel) to compute the defect features in the macro-uniformity test set.

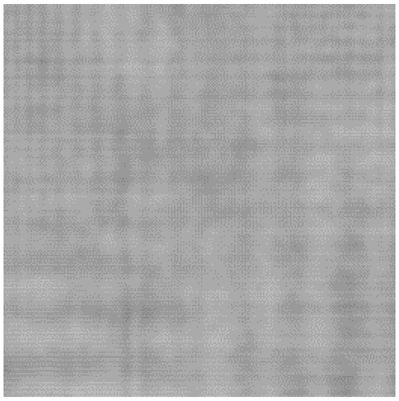

Figure 5. *The test image before HVS filtering.*

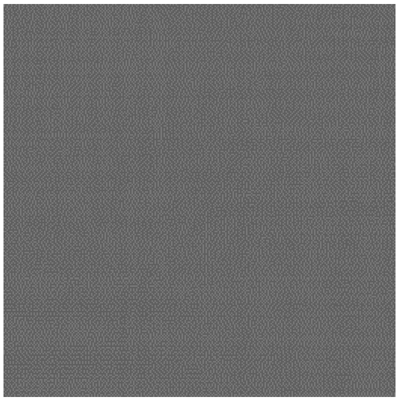

Figure 6. *The test image after HVS filtering.*

To compute values that can represent the severity of the defects in the macro-uniformity test set, spatial variations including one-dimensional, two-dimensional, periodic, aperiodic, localized, large-scale, and small-scale variations were considered. In the remainder of this section, we describe the attributes of each defect feature, and how it is computed. The methods we used to compute values for the defect features are mainly inspired by ISO image quality standards [3].

Graininess refers to the image fluctuation in both the horizontal and vertical directions of the image. We use the root mean square fluctuation ($RMSF$) of L^* value to measure Graininess.

$$G = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(L*_{ij} - \bar{L}*)^2} \tag{4}$$

where M and N are the number of image samples in the vertical and horizontal directions, and $\bar{L}*$ is the global mean of $L*$ in the sample.

Mottle refers to random lightness variations in both the horizontal and vertical directions of the image. To compute its value, an averaging window of 2×2 mm^2 (47×47 pixels at the scanning resolution of 600 dpi) is convolved with the sample. Then, the standard deviation of the resulting array is computed as the Mottle.

Large Area Variation also refers to two-dimensional random lightness variations, but the spatial region is larger than for Mottle. We use an averaging window of 4×4 mm^2 (95×95 pixels at the scanning resolution of 600 dpi), and convolve it with the sample. Then the difference between the largest and smallest array entries was computed as the Large Area Variation.

Banding and Streaks represent high-frequency lightness variations in different directions. So their computational methods are the same except for the processing direction. First, the $1D$ (one-dimensional) projection in both the horizontal and vertical directions is performed. Then the average of the signal is subtracted from the signal to exclude the DC component. After that, its DFT (discrete Fourier transform) is computed; and the strengths of signal peaks are measured. The Banding and Streaks features are obtained by integrating the energy of the peaks whose frequency is higher than 10 $cycles/inch$.

Large-scale Non-uniformity represents two-dimensional low-frequency lightness variations. First, we perform $1D$ projection in the horizontal and vertical directions. Then, we smooth the projections with an averaging window of length 3 mm (80 pixels at the scanning resolution of 600 dpi). After that, we use a piecewise linear spline fit to iteratively add knots until the maximum error between two adjacent knots is less than $0.5\Delta E$ units. The Large-scale Non-uniformity is obtained by computing the mean absolute slope of those line segments for which $\frac{\Delta L*}{\Delta d} > 0.5$, where d is the distance in $inch$.

Psychophysical Experiment

The psychophysical experiments were conducted under controlled viewing conditions in a laboratory at Purdue University dedicated for this purpose. The viewing booth used was the $Graphiclite\ CVX2$ [4]. For viewing, the quality ruler samples and test samples were placed in frames. The frames were fabricated by $MatShop$ [5]. For the ruler samples, the frames were 9×9 $inch^2$ in size with a 6×6 $inch^2$ opening. The frame border was sufficiently large to hide the test target around the border. For the test samples, the frames were 9×9 $inch^2$ in size with an 8×8 $inch^2$ opening. These frames can be seen in Figure 3. For both the ruler samples and the test samples, the frame color was cream with a white color core, which was chosen to be as neutral as possible.

A total of 26 subjects participated in the entire experiment. 14 of these subjects came from our research group; and most of them had an image processing background. The other 12 subjects were from non-engineering departments at Purdue, and did not have an image processing background. We conducted a pilot experiment before the official experiment to double-check that the procedure worked well. The data from the pilot experiment was not used to train or evaluate the predictors.

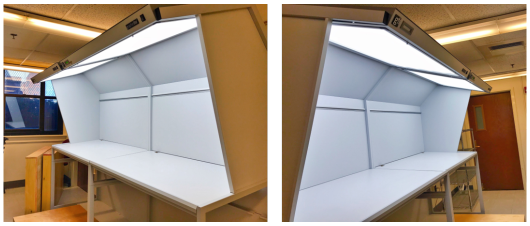

Figure 7. Two views of the lab environment and viewing booth. In the far left of the left image, some of the wooden platforms used to adjust viewing height can be seen standing on end.

Before each subject's experiment, we tested their visual acuity and color vision, and adjusted their viewing distance based on their height by having them stand on an appropriate number of stacked wooden platforms. We showed all quality ruler samples as well as the test samples, and briefly explained the experimental process.

During the experiment, the subjects walked along the viewing booth, and slid the test sample in front of them, comparing it with the hard-copy quality ruler samples, until finding a suitable location based on overall visual uniformity (The test sample's location meets the condition that each ruler sample farther to the right is lower in quality than the test sample, and each ruler sample farther to the left is higher in quality). Since the reference stimuli are labeled 3, 6, 9, etc., if the test sample's location fell in between two adjacent ruler images, as it often did, the subject then selected an intermediate integer value from the ruler scale. For example, if the location was between the ruler prints "12" and "15", but was closer to "12", the value of the test sample was assigned to be "13".

Data Analysis

Before using the subjects' assessments, we pre-processed the data and removed the outliers. In our work, we define a subject's data to be an outlier if it meets both of the following conditions:

1. Weak consistency.
2. Large average absolute deviation.

We conducted the consistency test as part of the experimental process. We selected 4 samples from different tint levels and repeatedly added these 4 samples into the waiting list in the early, middle, and end stages of each participant's experiment. So each subject assessed the uniformity of these four samples 3 times throughout the experiment.

We performed the same procedure for each subject to collect data to analyze their score consistency. For a given test sample,

[4] GTI Graphic Technology, Inc., Newburgh, NY.
[5] MatShop, Victoria, BC, Canada.

if the difference between the highest and lowest scores given by a subject was greater than 6 *JND*, the subject's score was considered to be weakly consistent. Thus, the weak consistency condition is defined as:

$$X^i_{max} - X^i_{min} \geq 6 \quad (5)$$

where X^i_{max} and X^i_{min} represent the highest and lowest scores given to the same print sample i by the subject at different stages of the experiment.

The second condition for the subject's data to be an outlier is that it exhibits a large average absolute deviation, which is:

$$\frac{1}{N}\sum_{i=1}^{N}|S_i - \overline{S_i}| > 3 \quad (6)$$

where N denotes the total number of subjects, S_i denotes the score assigned by the subject to the print sample i, and $\overline{S_i}$ denotes the average score assigned by all subjects to the print sample i. If the subject failed both these tests, then all their scores were eliminated from further analysis of the dataset. Of the 26 subjects, a total of 3 subjects were eliminated as outliers.

Prediction Models

At this point, we had collected 26 participants' perceptual assessments (*JND*) of 42 samples. After removing the outliers as described in the previous section, we obtained the average perceptual assessments of all samples, which was the ground-truth used to build the prediction model. For each sample, the seven features described previously, which represent the severity of printing defects in the macro-uniformity set were computed. This data was also the input for building the prediction model. Our goal was to find the relation between defects and the overall perceptual uniformity. Using this correspondence, we can predict the overall *JND* score for prints in the future.

In this paper, we used linear regression (*LR*) and support vector regression (*SVR*) to build predictors. Linear regression is a simple algorithm that models the relationship between a scalar response and one or more explanatory variables, and then predicts future data response based on that relationship. The support-vector machine (*SVM*) is a supervised learning model used for classification and regression analysis. The learning strategy of *SVM* is to identify the hyperplane which maximizes the margin, so the task is transformed into a convex quadratic programming problem.

In our work, we used the Least-squares support-vector machines ($LS-SVM$) to solve the regression problem. It has the same principles as the *SVM* for classification, with only a few minor differences. For the kernel function K, typical choices are the linear kernel, polynomial kernel, and Gaussian radial basis (*RBF*) kernel. We use the *RBF* kernel represented by the following formula.

$$K(x,x_i) = exp(\frac{-||x-x_i||^2}{\sigma^2}) \quad (7)$$

We used Matlab and Python to implement the *SVR* model. For Matlab, we used the Least Squares Support Vector Machines (*LS-SVM*) [9]; and the *Scikit-Learn* [10] package was used for the Python version *SVR* model.

In the process of building the model, we used k-fold cross-validation. Cross-validation is a verifying method used to evaluate how the results of a statistical analysis will generalize to an independent data set. It is mainly used for assessing the ability of a predictive model to predict new data. Also, it can help to find problems such as overfitting or selection bias [11].

In k-fold cross-validation, the original sample is randomly divided into k equal-sized subsets. Among the k subsets, one is used as testing data, and the remaining k-1 subsets are used as training data. Then, the cross-validation process is repeated k times, and each of the k subsets is used as the testing data only once. The k results can then be averaged to produce the model estimation. The advantage of this method is that all observations are used for training and testing, and each observation is only used once for testing.

For our data, setting $k = 6$ will result in 6 folds. We randomly shuffle the samples into 6 folds indicated by d_0 to d_5 so that each set is equal in size, which is with 7 samples in each fold. Then, we train on $d_0,...,d_4$ and verify on d_5, then train on $d_1,...,d_5$ and verify on d_0, ... This process is repeated 6 times. The average prediction is the model estimation.

Results

To evaluate our regression models, we use mean absolute error (*MAE*) and mean squared error (*MSE*). The standard deviation of *MAE* is a measure of the robustness of the predictions.

$$MAE = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n} \quad (8)$$

$$MSE = \frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n} \quad (9)$$

where p_i represents predicted data and a_i represents actual data. For the predictor built by *LR*, we have:

$MAE = 0.9100$ *JND* $\quad$ $Std.Dev.$ of $MAE = 0.6340$
$MSE = 1.2399$ *JND*

For the predictor build by *SVR*, we have:

$MAE = 0.8305$ *JND* $\quad$ $Std.Dev.$ of $MAE = 0.5469$
$MSE = 0.9463$ *JND*

It is a quite encouraging that the *MAE* between the predicted scores and the subjects' scores for both models is less than 1. As mentioned earlier, the interval between two adjacent quality rulers used as a reference is $JND = 3$ so for an accurate model, its *MAE* should be less than 3 *JND*. Also, the standard deviation shows that the error distribution is stable.

We were particularly interested in some samples with a relatively large absolute error. We found that most of these samples were from the 100% tint level. We feel that this is understandable, because the 100% tint level means the sample is totally black. It was more difficult for subjects to notice the defects; and thus these samples were given a relatively better quality evaluation than was predicted.

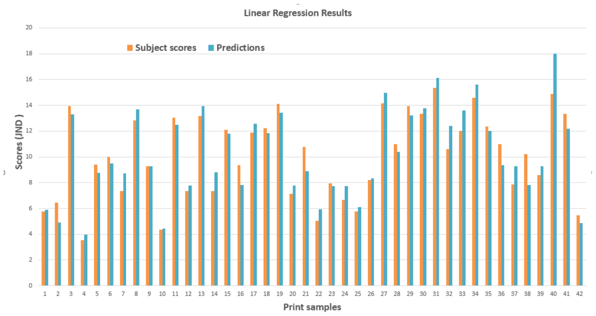

Figure 8. *The results of the Linear Regression Predictor. The abscissa is the sample ID and the ordinate is the JND score. The orange bars indicate the predicted score for each sample. The blue bars indicate the subjects' mean score for each sample.*

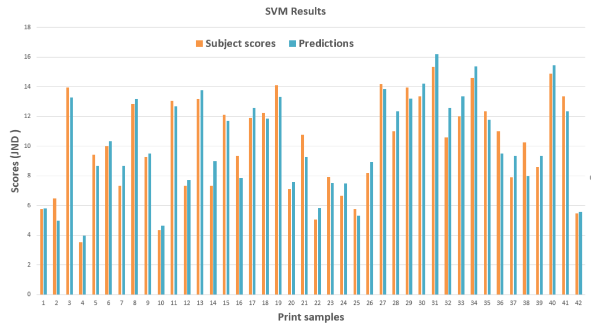

Figure 9. *The results of the SVR Predictor. The abscissa is the sample ID and the ordinate is the JND score. The orange bars indicate the predicted score for each sample. The blue bars indicate the subjects' mean score for each sample.*

Conclusion

In this paper, we designed and conducted a psychophysical experiment to collect the perceptual assessment of print macro-uniformity. The experiment worked well, and showed that the quality ruler method proposed by the *INCITS W*1.1 macro-uniformity team can provide a reliable method to assess macro-uniformity of print samples. Also we developed models for predicting the overall macro-uniformity as judged by humans. We confirmed the efficacy of the predictors using 6-fold cross-validation. Also, the model evaluation metrics *MAE*, *MSE* and standard deviation of *MAE* indicated that the models can perform accurate prediction.

Acknowledgments

This work builds on earlier work with the image quality ruler reported in [6], [7] below. We wish to thank *Rene Rasmussen* for providing the software to generate and print the image quality ruler samples, as well as the software to calibrate the printer used to print the image quality ruler samples. We also wish the thank *Chin-Ning Chen* for calibrating the scanner that was used to scan the test samples.

References

[1] René D. Rasmussen, Frans Gaykema, Yee S. Ng, Kevin D. Donohue, William C. Kress, and Susan Zoltner. "W1.1 Macro Uniformity", *Proceedings of SPIE*. Vol. 7242, 2009.

[2] René D. Rasmussen, William C. Kress, Yee S. Ng, Marguerite Doyle, Kevin D. Donohue, Kate Johnson, and Susan Zoltner. "INCITS W1. 1 macro-uniformity", *Image Quality and System Performance*. Vol. 5294, 2003.

[3] Brian W. Keelan and Hitoshi Urabe. "ISO 20462: a psychophysical image quality measurement standard", *Image Quality and System Performance*. Vol. 5294, 2003.

[4] René D. Rasmussen. *W1.1 macro-uniformity software 0.2.0.*

[5] Jim Grice and Jan P. Allebach. "The print quality toolkit: An integrated print quality assessment tool", *Journal of Imaging Science and Technology*. Vol. 43, 1999.

[6] Weibao Wang, Gary Overall, Travis Riggs, Rebecca Silveston-Keith, Julie Whitney, George Chiu, and Jan P. Allebach. " Figure of Merit for Macrouniformity Based on Image Quality Ruler Evaluation and Machine Learning Framework", *Image Quality and System Performance*. Vol. 8653, 2013.

[7] Weibao Wang. "A Study on Image Quality Evaluation In Image Capture and Production Process", Ph.D. Dissertation, Purdue University, West Lafayette, IN. *ProQuest Dissertations and Theses*, 2016.

[8] James Mannos and David Sakrison. "The effects of a visual fidelity criterion of the encoding of images", *IEEE Transactions on Information Theory*, Vol.20, 1974.

[9] https://www.esat.kuleuven.be/sista/lssvmlab/

[10] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

[11] Ron Koha, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *Appears in the International Joint Conference on Artificial Intelligence.* 1995.

[12] Xing Liu, Gary Overall, Travis Riggs, Rebecca Silveston-Keith, Julie Whitney, George Chiu, Jan P. Allebach. "Wavelet-Based Figure of Merit for Macrouniformity", *Image Quality and System Performance X*. Vol. 8653, 2013.

[13] https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html.

Author Biography

Yi Yang received her M.S in Geomatics Engineering from Chinese Academy of Sciences in 2016. She is currently working on a Ph.D. in Electrical and Computer Engineering at Purdue University. Her primary area of research has been image processing, computer vision and machine learning.

Utpal Sarkar received an M.Sc. in Mathematics from the University of Utrecht, the Netherlands in 1997, and an M.Sc. in Theoretical Physics *from the University of Barcelona, Spain, in 2018. He works as a software engineer at the Large Format Division of HP in Barcelona, on image processing and printing pipelines for large format, 3D, and textile printers.*

Isabel Borrell received her M.S in Mechanical Engineering from the Escola Tècnica Superior d'Enginyers Industrials de Barcelona. She joined Hewlett-Packard in 2000 and since then she has worked in the development of the writing systems of a variety of large-format inkjet printers. Most recently she has contributed to the HP PageWide XL printer platform.

Jan P. Allebach is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for Imaging Science and Technology (IS&T), and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award and the IS&T/OSA Edwin Land Medal, and is a member of the National Academy of Engineering.