https://doi.org/10.2352/issn.2694-118X.2021.LIM-101

Automatic Noise Analysis on Still Life Chart

Salim Belkarfa, Ahmed Hakim Choukarah, Marcelin Tworski;DXOMARK Image labs; Boulogne-Billancourt, France

Abstract

In this paper, we tackle the issue of estimating the noise level of a camera, on its processed still images and as perceived by the user. Commonly, the characterization of the noise level of a camera is done using objective metrics determined on charts containing uniform patches at a given condition. These methods can lead to inadequate characterizations of the noise of a camera because cameras often incorporate denoising algorithms that are more efficient on uniform areas than on areas containing details. Therefore, in this paper, we propose a method to estimate the perceived noise level on natural areas of a still-life chart. Our method is based on a deep convolutional network trained with ground truth quality scores provided by expert annotators. Our experimental evaluation shows that our approach strongly matches human evaluations.

Introduction

Camera quality has been considerably improved in the last years to meet the ever-growing standards of the consumers. Image quality can be characterized through multiple attributes such as exposure, color, texture, and noise. In this work, we are focused on assessing the capability of a camera to control its level of noise. In addition, we aim to provide this assessment as a metric that correlates with human judgment.

To assess the quality of a camera, a common way is to capture for each camera the same chart in a controlled environment. A chart is designed to be reproducible and therefore allowing to fairly compare different cameras due to its consistent visual content.

Since noise in an image is a random granulation, it is not exactly reproducible from one image to another, but only statistically, so generally we aim at estimating its second central moment (i.e. its variance) to describe this random process. This metric is easier to estimate over uniform areas, this is why noise is commonly measured on charts with uniform patches.

One of the common metrics for assessing noise level is the signal-to-noise ratio (SNR). On a uniform area, this metric is the ratio between μ_{Image}, the average of the image values, and σ_{Image}, the standard deviation of the image values.

$$SNR = 20 \times \log_{10}(\frac{\mu_{Image}}{\sigma_{Image}})$$

However, the SNR only reflects the total amount of noise for a given signal level, it does not describe how the human observer actually perceives the noise. To tackle this issue, the visual noise metric has been proposed. This metric intends to measure noise as perceived by end-users. For example, noise that cannot be seen by the eye at a given viewing condition will not be included in the noise measurement.

The Visual Noise measurement is standardized by IEEE CPIQ P1858 (Camera Phone Image Quality) 2016 [1], this standard is an adaptation of ISO 15739 [2] proposal. To compute this metric, the used test target must be compliant with the ISO 14524 [3] opto-electronic conversion function (OECF) test chart. This test chart is represented in Figure 1.

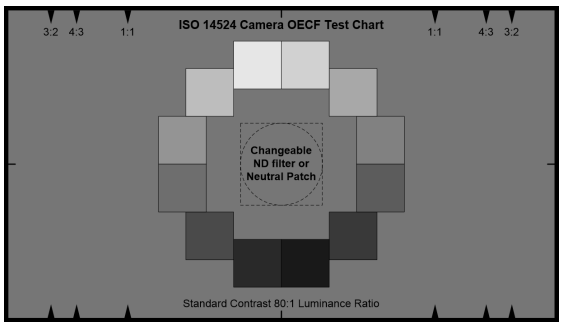

Figure 1. *Visual Noise compliant Chart*

However, nowadays, cameras integrate one or more denoising steps in their processing pipeline. The purpose of these steps is to reduce the noise in the image and restore the original signal. But it is challenging to reduce the noise in the high-frequency components while preserving the high-frequency content (such as edges and textured areas). While on the contrary, the low-frequency components have regular values, so noise is easily suppressed by averaging the pixels within a neighborhood. Thus, it is common to observe a different noise level between textured and uniform areas cf. Figure 2.

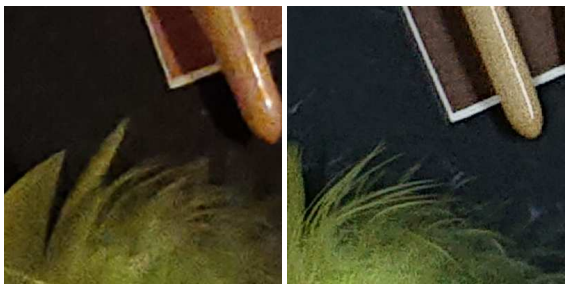

Figure 2. *Various level of noise between high frequencies and low frequencies components.*

As common measurements are not suitable for assessing noise in other than uniform areas, they cannot lead to adequate noise characterization of cameras with the behavior detailed above. To tackle this issue, we define two areas of interest well-suited for noise assessment in a still life chart (cf. Figure 3). We then propose a learning-based method using these specific areas of interest. The problem of assessing the perceived level of noise in these areas of interest can be formulated as a regression problem, so in order to solve this problem we suggest using a deep convolutional network. We train the network using annotations provided by image quality expert annotators, this annotation process allow obtaining a set of scores that will match with the perceptual user experience. We show that this learning-based approach strongly correlates with the perceptual ground truth and better predicts the perceived level of noise on natural scenes than standard approaches.

(a) First Region of interest Feather (b) Second Region of interest Woman

Figure 3. *Defined areas of interest to assess noise on natural areas.*

CSF Parameters defined by CPIQ P1858

	CSF_A	CSF_{C_1}	CSF_{C_2}
a_1	75	109.1413	7.0328
b_1	0.2	0.0004	0
c_1	0.8	3.4244	4.2582
a_2		96.5971	40.691
b_2		0.0037	0.1039
c_2		2.1677	1.6487
K	75	202.7384	40.691
S		0	7.0328

Related Work

In this section, we will review the existing works done on quality assessment of noise.

Visual Noise

Signal-to-noise ratio is often used as a metric to assess noise. However, SNR only reflects the total amount of noise for a given level of signal, it does not describe how the human observer actually perceives it. The level of noise can be critical for the image quality, as it can affect multiple of its aspects, from object visibility to face detection.

That is why the study of noise and in particular that of visual noise remains mandatory for the image quality assessment (IQA). Visual noise has been introduced to propose a metric that correlates more with human perception. The visual noise metric takes into account the spectral frequency content of luminance and chrominance noise by applying a contrast sensitivity function (CSF), a metric that integrated the noise power spectrum with properties of the human visual system. The computation of the visual noise described by CPIQ P1858 standard [1], based on the formulation made by ISO 15739 [2], requires the following steps:

- Conversion of the source image in a color opponent space AC_1C_2
- Filtering of the luminance and chrominance channels by respective CSFs
- Filtering of the channels by the display or print MTFs
- Application of a high pass filter to remove nonuniformities due to lens shading
- Conversion to CIELab color space and computation of variances of luminance and chrominance channels

The CSF used for the spatial filtering is defined as:

$$CSF_{luminance} = \frac{a_1 \times f^{c_1} \times \exp(-b_1 f)}{k}$$

$$CSF_{chrominance} = \frac{a_1 \times \exp(-b_1 f^{c_1}) + a_2 \times \exp(b_2 f^{c_2}) - S}{K}$$

where parameters are defined in Table 1.

The visual noise metric is then obtained by applying the log 10 base to the weighted sum of the L^*, a^*, b^* variances and L^*a^* covariance.

$$VN = \log_{10}(1 + 23\sigma^2(L^*) + 4.254\sigma^2(a^*) - 5.47\sigma^2(b^*) + 4.77\sigma^2(L^*a^*))$$

The previous formula weights the color noise for the b^* channel with a negative value, hence noise in the b^* channel leads to the decrease of the visual noise metric. Besides, a negative value on the b^* channel doesn't represent the human visual system property. For this reason, works are still under progress to improve the visual noise metric [4]. Moreover, the presence of the negative weights combined with the covariance, expressed by L^*a^*, can lead to negative values and the inability to estimate the visual noise metric for a given image.

Learning Based Methods

In opposition to the visual noise metric described in the previous section, learning-based methods require annotated datasets.

TID2008 [5] and its extension TID2013 [6]) are image quality datasets that give a Mean Opinion Score (MOS) for each distorted image. These distortions are artificially introduced and correspond mostly to compression or transmission scenarios. As these distortions are artificially introduced, they do not fully cover the ones introduced by real cameras. The LIVE in the wild [7] database contains 1162 authentically distorted images captured from many diverse mobile devices. Each image was viewed and rated online on a continuous quality scale by an average of 175 unique subjects with the goal of providing one MOS per each image, and not one score for each image quality attribute, such as the noise which is our interest study. Similarly, the KonIQ10k [8] dataset consists of samples from a larger public media database with unknown distortions. This dataset provides a ground truth for several image quality attributes, but does not consider the noise quality as one of them.

More recently, Yu et al. [9] collected a dataset of 12 853 natural photos from Flickr and annotated them according to image quality defects: exposition, white balance, color saturation, noise, haze, undesired blur, composition. They aimed to solve a multi-task learning problem and trained a multi-column deep convolutional neural network to simultaneously predict the severity of all the defects. While their approach showed promising results, we are tackling a different issue, that of noise estimation in specific areas only.

To the best of our knowledge, the most related work has been proposed by Tworski [10] et al.. They adopt a regression formulation and train a network to estimate the camera capacity to preserve texture using a common perceptual chart. In the next section, we will detail our deep regression framework for noise quality estimation and the method used to collect the datasets relevant to our noise assessment problem.

Method

In this section, we detail the proposed method for perceptual noise estimation on natural images. This task is a regression problem, in which we want to estimate for an image X of dimensions $Height \times Width \times 3$ its corresponding noise quality score Y, a scalar. To perform this, we use a learning-based method, meaning that we use the ground-truth noise quality of the given

image provided by expert annotators (cf. subsection Datasets).

Inspired by previous works [11, 10], we chose to rely on the very versatile ResNet-50 architecture. This network has already shown some excellent results in other related IQA tasks [11]. ResNet -short for Residual networks, is the neural network that won the imageNet [12] contest in 2015. The main addition of the ResNet architecture is to partially solve the vanishing gradient problem on extremely deep neural networks.

We have images with fixed size of $1000 \times 1000 \times 3$, ResNet50 can take an input of any dimensions but using large inputs usually leads to large memory consumption so often it is not an available option, e.g. a common input size for ResNet50 is $224 \times 224 \times 3$. As resizing the images to a lower resolution will affect the level of noise, we decide to take fixed size image crops input. During our investigation we observed better results when training the ResNet50 with a $448 \times 448 \times 3$ input size, so we decide to take crops of this size. We used the convolutional layers and average global pooling layer of the ResNet-50 model trained on ImageNet database and replaced the fully connected layer to fit our regression problem with a unique output. It is thus a layer with 2048 entries, that requires the training of 2049 additional parameters, with a single output to which we apply the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$ to obtain a continuous output ranging from 0 to 1.

At each epoch a crop is randomly selected, allowing the model to learn to estimate the perceived noise on variable zones and thus having a more robust estimation to field of view variations. As some crops may not be relevant for the evaluation, we choose to use Huber loss during training as this loss is less sensitive to outliers than the squared error loss.

At test time, we extract ten random crops and average their predictions to get the estimated noise score.

Datasets

Lighting conditions While having photos from different cameras is important for constructing our database, so are the lighting conditions that do affect heavily the level of noise. Our database therefore, contains multiple lighting conditions for each device and chart:

- 5 Lux Tungsten
- 20 Lux Tungsten
- 100 Lux Tungsten
- 300 Lux TL84
- 1000 Lux D65

Charts and devices As there is no well-established reference dataset for our problem, we collected annotated data using two different charts.

- **Still-Life**: First, we use the chart in Fig. 5. This dataset is referred to as *Still-Life*. This chart is specifically designed by DXOMARK to evaluate multiple IQA attributes and contains diversified content such as uniform zones, fine details, portraits, vivid colors for color rendering, as well as resolution lines and a low-quality Dead Leaves version. We extract 2 areas of interest represented in Figure 3, that we will note *Feather* and *Woman*. Images are acquired using 293 different smartphones and cameras from different brands commonly available in the consumer market. Thus this database consists of 1465 crops for each area of interest. In Fig. 4, we provide an example region captured with two different cameras in different lighting conditions. The left image corresponds to a low-quality device in low light conditions, while the other is obtained with higher quality. It illustrates the nature of distortions that appear in this dataset when using different lighting intensities.
- **Dead Leaves**: Second, we employ the Dead Leaves chart proposed in [13]. This chart depicts gray-scale circles with random radius and locations. This chart is compliant with ISO 14524 [3] and so allows to compute the visual noise on it. In all our experiments, we refer to this dataset as *Dead Leaves*. We use the same five lighting conditions and devices as for the *Still-Life* chart. Consequently, this database is made up of 1465 crops *Dead Leaves* shots.

(a) High Noise Image (b) Low Noise Image

Figure 4. *Different levels of Noise in the database*

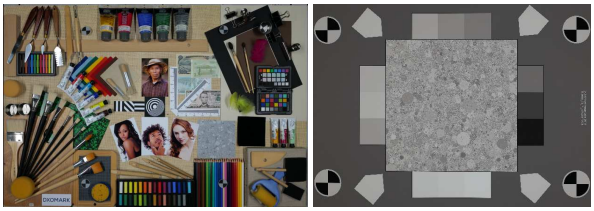

Figure 5. *Still-Life Chart used in our experiments. The* Still-Life *chart contains many diverse objects with varying colors and textures while the* Dead Leaves *chart depicts random gray-scale circles.*

Annotations In order to obtain more faithful results, we need to provide a reliable ground-truth annotation for each pair of device and lighting condition in our database. These annotations should correspond to a precise way of encapsulating the perceived visual noise quality. To obtain such quality annotations for each of our pairs, we first established the ground truth references by asking 20 human experts to rank the images according to the level of perceived noise. We then averaged the rankings by excluding the images rated with the highest and lowest positions within the obtained stack. In order to obtain continuous scores, we performed a linear rescaling of the ranks within the interval $[0, 1]$, where the best possible rank corresponds to a score of 1, while the worst to a score of 0. This reference set constitutes our noise quality ruler. For each image to be annotated, we ask an expert to correctly rank it by evaluating it with respect to the quality ruler (cf. Figure 6).

Specific conditions were prepared to make the comparison as reliable as possible, we use a 24" full hd monitor with a pixel pitch of 0.27 millimeters, while the distance between the analyst and the screen is fixed to 40 centimeters. Note that the images used for annotation were provided with no down-sampling. However, for low resolution images, bicubic resizing is applied to match their size to the highest in the image stack. Each position among the set of references is assigned a score between 0 and 1. In the *Still-Life* chart, we have considered two different

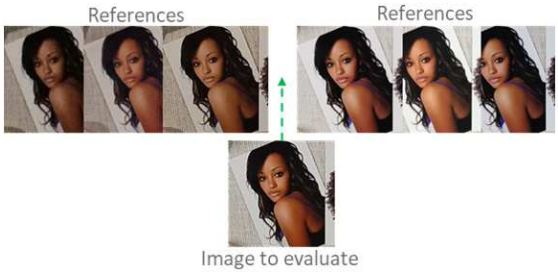

Figure 6. *Perceptual noise evaluation by quality ruler*

regions of interest to study as seen in Figure 3. In the case of the Dead Leaves charts, since the charts are unnatural images, human perceptual annotation is quite complex due to multiple reasons, but it is mainly the presence of different types of patches at different intensities that make the annotation task quite hard for the annotator. Therefore, we chose to transfer the annotations obtained on the *Still-Life* chart to the *Dead Leaves* one, rather than re-annotating the images. This assumes that our annotations are device based: the quality of a given image depends mostly on the device itself. The *Still-Life* chart contains diverse scenes similar to what real images would contain. Evaluating devices according to their performance on this card allows us to obtain a subjective device evaluation in a setting more similar to real-life scenarios.

Metrics

A straightforward way to assess our results could consist in computing the correlation between the predictions and the annotation. However, the underlying assumption that the predictions of each method correlate linearly with our annotations is not always correct and might bias our evaluation. Thus we decided to use two distinct metrics based on the correlation of the rank-order. First, the Spearman Rank-Order Correlation Coefficient (*SROCC*) defined as the linear correlation coefficient of the ranks of predictions and annotations. We also note the Kendall Rank-Order Correlation Coefficient (*KROCC*) defined by the difference between concordant and discordant pairs divided by the number of possible pairs. This second metric allows us to check the similarity of the ranking. For both metrics a value of 1 means that the observation of the predictions and annotations are identical.

For all visual charts, the dataset is split into training and test sets as follows. First, among the devices we use in our experiments several are produced by the same brand. So, to avoid bias between training and test, we impose no brand-overlap between training and test sets. To do that, we create 6 distinct manufacturer families chosen at random to balance the set of images from each family of devices. Then for each family, we proceed for a training on the rest of our set excluding it, and using that said family of devices as a test set. Thus, for each performed training and test there are approximately 1221 images in the training database and 244 images in the test database.

Comparison to state of the art

In this section, we compare the performance of our approach to existing methods. We compare the measurements performed on the *DeadLeaves* chart and predictions on the *Still-Life* chart on the whole database (293 devices). We chose to benchmark our method to three different formula of the visual noise metric:

- The formula standardized by CPIQ [1] (VN_{CPIQ})
- The formula in discussion in ISO15739 and lastly proposed [4] (VN_{ISO})
- The formula used by DXOMARK [14] ($VN_{DXOMARK}$)

As the visual noise metric provides one metric for each patch, we consider for each formula the one interpolated for $CIE-L^* = 50$. Besides this, the visual noise takes into account the sensitivity of the human eye to different spatial frequencies under various viewing conditions. Hence the measurement is always dependent on the size of the image (i.e. print or on-screen) and the viewing distance. The effect of the viewing conditions is to stretch the CSF along the frequency axis. To evaluate the ability of the visual noise measure to assess the noise level in our dataset, we use two different conditions:

- Viewing Condition Print: a commonly used viewing condition of a print of 120 centimeters height viewed at 100 centimeters
- Viewing Condition Display: a viewing condition as the one used during the annotation process, involving a display viewed at 40 centimeters with a pixel pitch of 0.27 millimeters

Moreover, our method on the *Still-Life* chart, gives predictions on two areas of interest for each image: *Woman* and *Feather*. We will therefore evaluate the predictions of *Woman* and *Feather* compared to the ground truth of their respective areas as well as the average of the two predictions compared to the average of the annotations. Quantitative results are reported in Table 2

Performance on the devices database.

Method	Viewing Condition	SROCC	KROCC
VN_{CPIQ}	Print	-0.640	-0.460
VN_{CPIQ}	Display	-0.620	-0.445
VN_{ISO}	Print	-0.585	-0.416
VN_{ISO}	Display	-0.576	-0.408
$VN_{DXOMARK}$	Print	-0.646	-0.464
$VN_{DXOMARK}$	Display	-0.654	-0.470
Ours Woman		0.883	0.717
Ours Feather		0.862	0.689
Ours Average		**0.904**	**0.734**

First, we observe that our method strongly matches with the provided annotations, and that it also outperforms other benchmark methods. These results must also be weighted, as the predictions were made on the same chart as the annotations (i.e. the Still-Life chart), while the visual noise metrics were established on the Dead Leaves chart. The results of the visual noise metrics show that the concerns raised in Introduction are valid: measuring the noise on uniformly gray patches does not sufficiently allow us to assume the perceived level of noise of the camera on a natural image.

Conclusion

In this paper, we propose an efficient learning-based method to assess the perceived level of noise of a camera. Compared to traditional methods, our approach can be used in images with natural contents. The experimental results show that our predictions strongly match that of the user experience. These promising results show the great potential of deep learning for image quality assessment. Future work will focus on improving the proposed method and will consist in building a system able to evaluate the noise more exhaustively, namely by characterizing its chromaticity as well as its frequency.

References

[1] IEEE standard for camera phone image quality.

[2] Photography – Electronic Still Picture Imaging – Noise Measurements. Standard, International Organization for Standardization, Geneva, CH, 2013.

[3] Photography – Electronic Still Picture Cameras – Methods For Measuring Opto-Electronic Conversion Functions (OECFs). Standard, International Organization for Standardization, Geneva, CH, (2009.

[4] Dietmar Wueller, Akira Matsui, and Naoya Katoh. Visual noise revision for ISO 15739. *Electronic Imaging*, 2019(10):315–1–315–7, January 2019.

[5] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 01 2009.

[6] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Color image database tid2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing (EUVIP)*, pages 106–111, 2013.

[7] D. Ghadiyaram and A.C. Bovik. Live in the wild image quality challenge database. `http://live.ece.utexas.edu/research/ChallengeDB/`.

[8] Vlad Hosu, Hanhe Lin, Tamás Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:1–1, 01 2020.

[9] Ning Yu, Xiaohui Shen, Zhe Lin, Radomir Mech, and Connelly Barnes. Learning to detect multiple photographic defects. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1387–1396, 2018.

[10] Marcelin Tworski, Stéphane Lathuilière, Salim Belkarfa, Attilio Fiandrotti, and Marco Cagnazzo. Dr2s : Deep regression with region selection for camera quality evaluation, 2020.

[11] Chen-Hsiu Huang and Ja-Ling Wu. Multi-task deep cnn model for no-reference image quality assessment on smartphone camera photos, 2020.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[13] F. Cao, F. Guichard, and Hervé Hornung. Measuring texture sharpness of a digital camera. In Brian G. Rodricks and Sabine E. Süsstrunk, editors, *Digital Photography V*. SPIE, January 2009.

[14] Gabriele Facciolo, Gabriel Pacianotto, Martin Renaudin, Clement Viard, and Frédéric Guichard. Quantitative measurement of contrast, texture, color, and noise for digital photography of high dynamic range scenes. *Electronic Imaging*, 2018(12):170–1–170–10, January 2018.

Author Biography

Salim Belkarfa *received his Master's degree in engineering from Telecom Bretagne (2013). He joined DXOMARK Image Labs in 2015 to continue developing state of the art camera quality assessment technology.* ***Ahmed Hakim Choukarah*** *is a master student at the École Normale Supérieure de Paris.* ***Marcelin Tworski*** *received his Master's degree from Institut polytechnique de Grenoble (2018), he is currently pursuing his PhD in Automatic analysis of image quality criteria on natural scenes using deep neural networks at Telecom Paris.*

AUTHOR INDEX

Society for Imaging Science and Technology
7003 Kilworth Lane
Springfield, VA 22151 USA
703/642-9090; 703/642-9094 (fax)