https://doi.org/10.2352/CIC.2022.30.1.24

Analysis of Individual Quality Scores of Different Image Distortions

Olga Cherepkova, Seyed Ali Amirshahi, Marius Pedersen; Norwegian University of Science and Technology Gjøvik, Norway.

Abstract

In this paper, we study individual quality scores given by different observers for various image distortions (saturation, contrast, and color quantization) at different levels. We created a database that contains a total of 232 images, derived from 21 pristine images, three distortions, and five levels. The database was rated by 31 participants collected through an online platform. The study shows that observers have distinguishable patterns with respect to different distortions. Using quadratic regression models, we visualized the behavior patterns of different groups of observers. The database and the individual scores collected are publicly available and can be further used for quality assessment research.

Introduction

Image quality assessment is widely used in various fields of research. As the requirements for quality assessment are increasing, pressure on the reliability of the development methods is also growing. Generally, to assess image quality, two different methods are used: subjective methods, which involve the participation of observers, and objective methods (metrics), which aim to predict subjective scores and provide a fast and simple solution to avoid time-consuming subjective studies. Although the first image quality metrics were mainly based on simple techniques such as PSNR and the SSIM [16], in recent years there has been a shift towards more complex approaches, which nowadays are based on deep learning algorithms [2, 3, 4, 5].

Research has shown [8] that there are differences between different observers in image quality assessment. Due to these differences, a potential way to increase the precision of an metric for a particular observer lies in the personalization of image quality assessment methods. However, due to the complexity of the prediction of individual scores, most metrics are developed to predict the Mean Opinion Score (MOS). Such an approach can lead to loss of information and, in some cases, not enough to reflect observers' preferences. Furthermore, different people are sensitive to different types of distortion, and while for some distortions there may be a high degree of agreement between observers, other distortions could lead to a greater deviation in opinions [8].

In this work, we address the differences in personal preferences in judging the quality of images affected by certain distortions. Our hypothesis considers the possibility of having different groups of people agreeing on level of degradation from some types of distortions while disagreeing on others. Our objective is to find groups of observers with similar preferences in quality judgment and to understand their preference patterns. This will allow us to find image quality metrics which are the best fit for each group allowing us to predict the preferences of the observers for each group as closely as possible.

Background

Multiple research works mention individual differences in image quality assessment tasks. Sun et al. [14] reported differences in the variability of objective scores for different databases, which shows how distortion influences the reaction of observers. Ponomarenko et al. [13] also pointed out differences in variability between distortions in the TID2013 dataset. The influence of specific image attributes has also been investigated. Bringier et al. [6] and Calabria and Fairchild [7] inspected the preferences of the images with respect to contrast. They found a certain optimal point of contrast after which observers react negatively to further increase in contrast. Calabria and Fairchild [7] in their work also found interrelations between different image attributes, such as contrast, sharpness, chroma, and lightness. This shows the influence of different attributes on the perception of human quality. In addition, they found an optimal point for the sharpness level, which had a similar influence as the contrast. Speaking of sharpness, Del Pin and Amirshahi [1] found that observers are influenced differently by sharpness of the image, resulting in differences in individual ratings. Furthermore, some works report the influence of content on individual observers. Virtanen et al. [15] reported the difference in objective scores for different scenes in the CID2013 database, highlighting the connection between image content and image quality assessment. Ninassi et al. [12] specifically mentioned the deviance in observers' opinions of JPEG2000 distorted images, where observers were more critical in their judgment of images with larger homogeneous areas rather than busy images.

Recently in 2022, Cherepkova et al. [8] reported variabilities in the observers' scores for 21 different distortions. By conducting a quantitative analysis and interview, they found distortions with the largest disagreement between the observers. The research has shown that saturation, contrast, change in the sharpness level, introduction of quantization or noise artifacts, and lens distortion cause large standard deviation in observers' scores. They have shown that the standard way of image assessment using MOS does not always reflect the preferences of real observers. In particular, the bimodal distribution of observers who disagree on a particular image would be averaged by MOS. They have also analyzed the quality assessment process with relation to image content and found that observers generally disagree not only on a particular distortion, but on a combination of a distortion and an image.

Experiment

Dataset preparation

In this work, we used 21 reference images of different content, level of detail, and visual attributes. We applied five different distortions of three types, which previously have been reported to

Figure 1. *Original images in the dataset.*

have the largest deviation in observers' ratings [8]: change in contrast, change in saturation, and color quantization. The distortion generator from the Kadid10K database [10] was used. Each distortion is applied at two different levels, which also have been chosen with regard to the highest variability in the ratings. Saturation increase corresponds to distortion 8 (levels 1 and 2) in Kadid10 database and saturation decrease to distortion 7 (levels 1 and 3), contrast increase to distortion 25 (levels 3 and 2) and contrast decrease to distortion 25 (levels 4 and 5), and color quantization to distortion 6 (levels 2 and 3). Thus we have 10 distorted images for each reference image, making a total of 232 images. Images are presented in Figure 1. The images were cropped to 800x800 pixels to avoid rescaling.

Collecting individual scores

To collect observer scores, we used the Amazon Mechanical Turk crowd-sourcing platform. Only master workers with an approval rate greater than 95% were allowed to participate. Although online data collection allows to gather more data in less time, limitations of an uncontrolled environment, such as different lightning and viewing conditions, displays can affect observers judgment. The influence of different viewing conditions was not included in the analysis process of this work due to Mechanical Turk limitations. A category judgment experiment was chosen to rate the quality of the images, due to its simplicity for observers and convenient analysis. We selected a neutral gray background (128, 128, 128) for the representation of the images. Observers had to judge the image using a scale from one to five, where five corresponded to the best image quality and one to the worst. The instructions were the following: "Here you will see different images. Please rate the quality of the images using the scale from 1 to 5, where 1 corresponds to the lowest quality (bad) and 5 corresponds to the highest quality (excellent)". In total, we had 111 unique observers, of which 31 completed 95% of the images in the dataset and were included in the further analysis. Since some observers evaluated some images more than once (during the first and second trials), their ratings were averaged.

Analysis

For the analysis, we used the differential opinion score instead of the original scores from one to five. It is calculated with Equation 1. Differential opinion score helps normalize the results and neutralize the influence of observers that use the scale differently.

$$d_{i,j} = r_{i,ref}(j) - r_{i,j} \tag{1}$$

In Eq. 1, $r_{i,j}$ is the observer rating i for image j, while $r_{i,ref}(j)$ is the rating for the corresponding original image [11].

To find common patterns in the preferences of the observers, we first modeled their preferences with a quadratic regression model (Eq. 2). The model tries to predict the differential score for each observer using distortion type and distortion change coded in level values, which range from -2 to +2. The coefficients fitted by the model describe observers' preferences and are used to cluster observers. To cluster observers, we used k-means and hierarchical methods. We used the quadratic model over the linear model because the dependence of given ratings is not always linear on the level of distortion. The model coefficients, in turn, help to summarize all quality scores into single numbers, which is more convenient to use for the clustering model than the scores.

$$y = \beta_0 + \beta_1 \times distortion \times level + \beta_2 \times distortion \times level^2 \tag{2}$$

where β_0 is an intercept, β_1 and β_2 are coefficients of covariance of first and second order, while distortion is used as a categorical factor and level as a numerical parameter. The level in this case changes from -2 to 2, where a negative level value corresponds to saturation or contrast decrease and a positive level to increase, respectively.

Observers have also been clustered based on single-distribution modeled preferences. The model in this case only accounts for the level as a numeric parameter (Eq. 3). The normalized β_2 and β_3 coefficients that characterize the slope of the model were then used as parameters for clustering. Since β_0 is just an intercept, we did not include it as a clustering parameter. We normalize the coefficients by dividing them by the corresponding standard deviation of each coefficient among all observers.

$$y = \beta_0 + \beta_1 \times level + \beta_2 \times level^2 \tag{3}$$

Intra- and inter-observer variability

To produce reliable results, we checked intra- and inter-observer variability. The goal of this paper is to find the variability between observers, therefore, we are concentrating on extreme outlier cases. To find it, we check multiple indicators and base the decision on their combined result. For intra-observer variability, we checked repeatability of observers, who completed the experiment twice. For this we computed averaged differences between the scores, given by each observer in the first and second sessions. In addition, we exploit Cohen's kappa, agreement rate, kurtosis, and correlation parameters for each observer.

Cohen's kappa [9] shows the possibility of the agreement between the first and second trials for the same observer happening by chance. The agreement rate measures the repetitiveness of choice for the same category between the first and second trials. Excess kurtosis values indicate the diversity of the chosen categories. While the normal bell-shaped mesokurtic distribution is

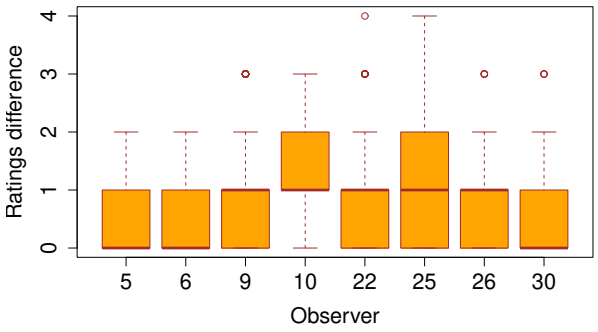


Figure 2. *Consistency for the observers that completed more than 95% of images for two trials. The boxplot shows the difference in ratings between the first and second trial.*

Table 1. Intra-observers variability with different techniques.

Observer	Cohen's kappa	Agreement rate (%)	Kurtosis	Correlation
5	0.37	60	-0.92	0.44
6	0.35	58	-1.03	0.52
9	0.20	48	0.00	0.25
10	-0.01	22	-0.30	0.48
22	0.33	47	-1.35	0.44
25	0.19	37	-0.86	0.34
26	0.09	34	-1.98	0.49
30	0.03	60	1.60	0.19

defined by 0, platykurtic distribution can reach -2 and leptokurtic may vary up to +3 [17]. Negative kurtosis values indicate non-normal distribution of observers scores, for example bimodal distribution. High numbers are more alerting and suggest that the observer's majority choice falls under the same category. Lastly, we check the correlation between the ratings of each observer for each image and the median score of the rest of the observers, excluding the current one.

Results

Intra- and inter-observer variability

In total we collected 7136 individual scores from 31 observer who completed more than 95% of images. eight out of 31 observers completed 95% of images twice. Their results have been used to test consistency. Figure 2 visualizes the average differences between the ratings, given by each observer in the first and second trials. While some outliers exist, we can see that in general observers do not have average difference greater than one. Observers 5, 6 and 22 have the least differences (being the most consistent in the two sessions), while observers 10, 25, 26 and 30 have larger variabilities (less consistency between the two sessions). In order to analyze intra-observer variability more in detail, we computed Cohen's Kappa, agreement rate, correlation, and kurtosis. Correlation and kurtosis in this case are calculated for all ratings given in both trials (Table 1). From the results, we can see that the agreement rate is proportional to the average variability (Figure 2).

Cohen's kappa value interpretation by Landis and Koch [9] indicates that most of the observers belong to fair and moderate reliability groups, while observers 10, 30 and 26 have slight agreement. To further investigate these observers, we compare additional excess kurtosis and correlation parameters. Here we can see that observer 30 has the lowest correlation with others, in addition to a high kurtosis value, suggesting that this observer prefers one category over others. Looking at this data more closely we found that 67% and 88% of chosen answers belonged to the same category in the first and second round, respectively, while on average, other observers chose this category in 28% and 45% respectively. Therefore, we excluded this observer from further analysis.

In addition, we checked the reliability of observer results. Inter-observer variability tests include correlation and kurtosis comparison for finding outliers. The results show a similar behavior of observer 31, who has a preference for one particular category, with 2.68 kurtosis and 32% of correlation with others, compared to the mean of -0.42 kurtosis and 48% correlation among others. While the other observers chose this specific category in 12% cases in general, observer 31 preferred it in 67% of the cases. So, we discard observer 31 for further analysis, considering this observer as an outlier.

Distortion based preference patterns

We analyzed the preferences of the observers for each distortion and found groups with similar preferences.

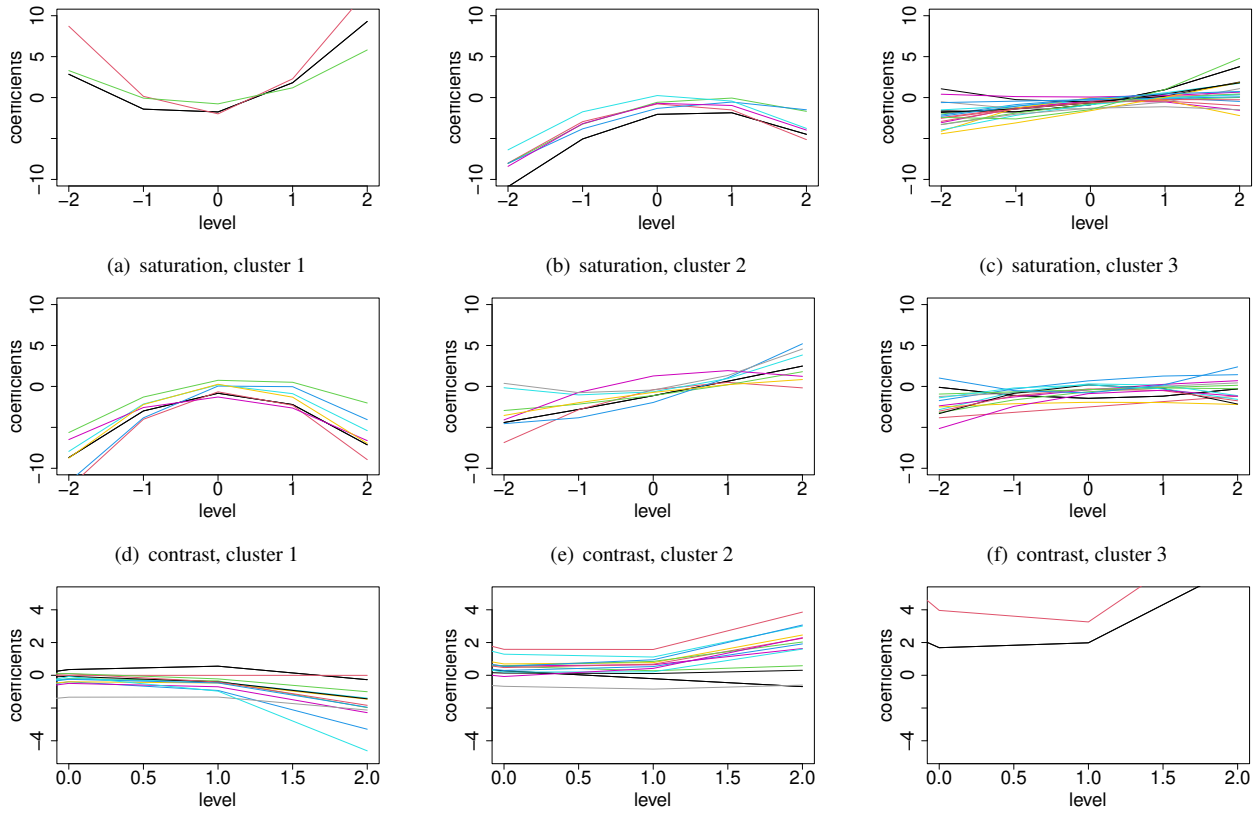
Saturation

In order to understand observers' preferences, we compute quadratic regression models, which predict differential opinion score based on the level of distortion. We utilize the coefficients of the models to find similarities between observers. Furthermore, we employ k-means and hierarchical clustering to divide observers into groups based on their modeled preferences. The results are shown in Figure 4 for k-means and in Figure 5 for hierarchical clustering. The k-means clustering shows a general division, while hierarchical clustering helps to understand the distances between observers. Two methods show similar results and particularly detach observers (3, 7, 8, 12, 18, and 29) and (21, 25, and 15) from the other observers.

We visualized the observers' preference models for each cluster in Figures 3(a), 3(b), and 3(c). The level changes from -2 to +2 along the x-axes, and the y-axes correspond to the coefficients of the model. From the plots, we can see two types of preference trends, which create convex and concave shapes of the models. Observers 15, 21 and 25 particularly like desaturated and oversaturated images (corresponding to higher values for level -2 and +2). We see the opposite behavior for observers 3, 7, 8, 12, 18, and 29 in the second cluster, who prefer the original images (level 0) over too saturated or desaturated images. The rest of the observers did not have such distinct preferences and were grouped together, but we can see some variability in their preference models. We can also notice that some observers in the third group prefer slightly saturated images, while they did not like desaturated images.

Contrast

The clustering results for contrast distortion are presented in Figures 6 and 7 and the corresponding models are visualized in Figures 3(d), 3(e), and 3(f). We can see different patterns depend-



(a) saturation, cluster 1 (b) saturation, cluster 2 (c) saturation, cluster 3

(d) contrast, cluster 1 (e) contrast, cluster 2 (f) contrast, cluster 3

(g) color quantization, cluster 1 (h) color quantization, cluster 2 (i) color quantization, cluster 3

Figure 3. *Visualized regression models for each cluster. Each line represent the model for a single observer.*

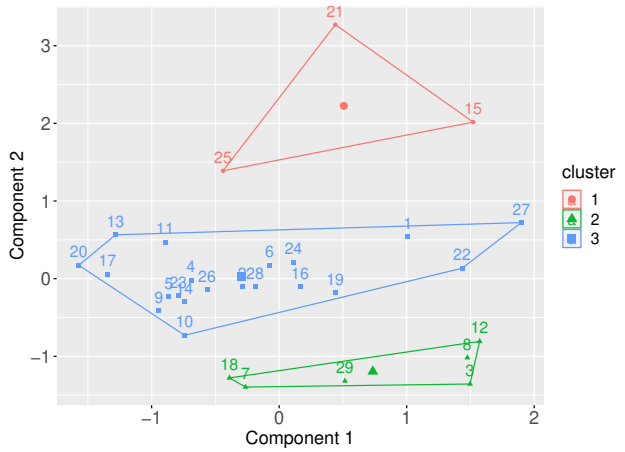


Figure 4. *K-means clustering based on regression model coefficients for saturation distortion.*

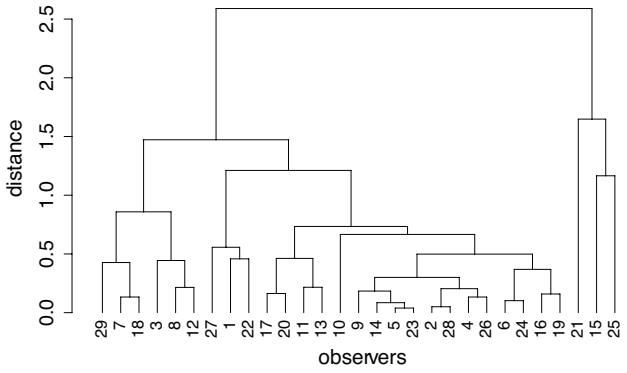


Figure 5. *Hierarchical clustering based on regression model coefficients for saturation distortion.*

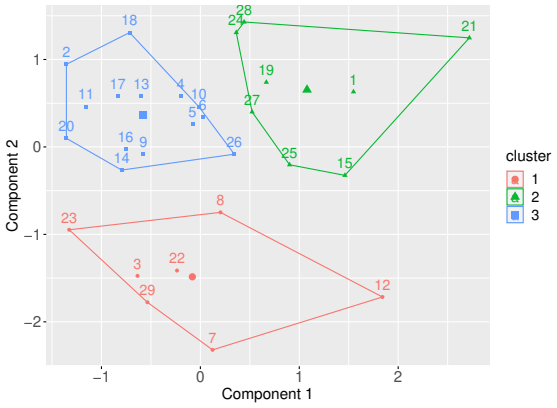


Figure 6. *K-means clustering based on regression model coefficients for contrast distortion.*

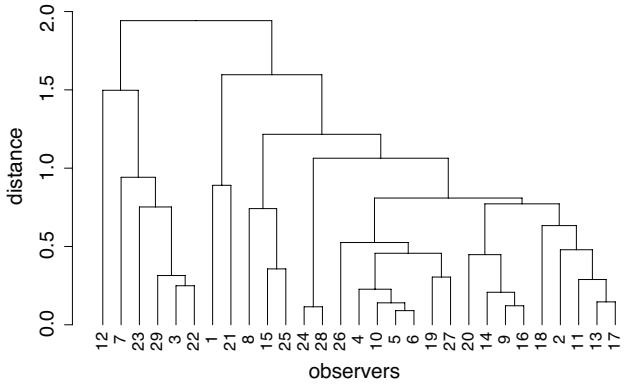


Figure 7. *Hierarchical clustering based on regression model coefficients for contrast distortion.*

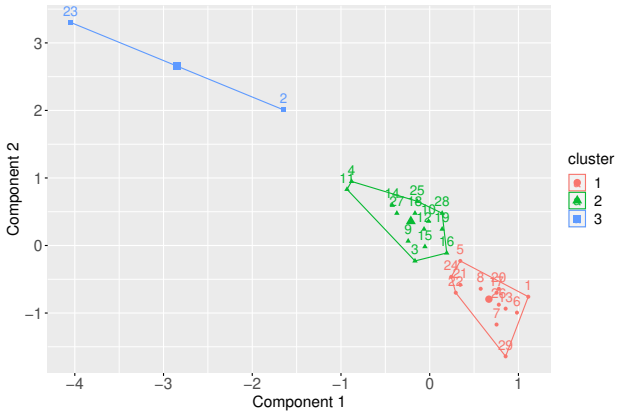


Figure 8. *K-means clustering based on regression model coefficients for color quantization distortion.*

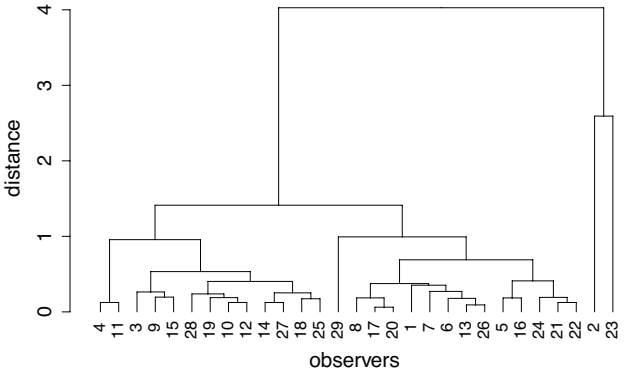


Figure 9. *Hierarchical clustering based on regression model coefficients for color quantization distortion.*

ing on the preferred contrast level (optimal point), after which the trend changes from ascending to descending with further increase in contrast. Observers 3, 7, 8, 12, 22, 23, 29 of the first cluster prefer images with average (original) contrast level, while observers 1, 15, 19, 21, 24, 27, 28 of the second cluster prefer more contrast images more clearly, and for some of them, an optimal point might be higher than the maximal level we used in our experiment. The preferences of the observers of the third cluster gradually increase with increasing image contrast. In addition, we can see that some observers like images with a lower level of contrast. From this we can make a conclusion of having at least two groups of people: those who prefer more contrast images and those who prefer more natural-looking images.

The finding of an optimal point of contrast, after which subjective judgement has a negative correlation with the perceived contrast level, is similar to [6]. In addition to the research by Bringier et al. [6], where they explored the dependence of MOS and perceived contrast, we found different groups of observers with different optimal points in contrast. The same applies to saturation.

Color quantization

There are only three levels for color quantization distortion: original (0), average (1), and high (2). The K-means and the hierarchical clustering show two observers, which appear more like outliers in this case (Figures 8 and 9). They have not been grouped together previously, so we do not see any repetitive pattern with saturation and contrast. Looking at the model plots in Figures 3(g), 3(h), and 3(i) we can see two general trends of observers who are more sensitive to color quantization distortion (observers 1, 5, 6, 7, 8, 13, 17, 20, 22, 24, 29) and those who are less sensitive to it (observers 3, 4, 9, 10, 11, 12, 14, 15, 16, 18, 19, 25, 27, 28). In this case, observers with high (Figure 3(h)) and low (Figure 3(g)) tolerance were grouped together.

Joint analysis

Previous clustering results demonstrate several groups of observers with different preferences for each distortion. Here we analyze their inter-relations. Figure 10 shows the hierarchical clustering based on the three combined distortions. We chose hierarchical clustering for comparison because it allows grouping observers based on distances between them, while not restricted to a number of clusters, which helps offset the impact of outliers. In Figure 10 three main clusters of observers (4, 5, 9, 10, 11, 14, 16, 18, and 20), (1, 6, 13, 15, 17, 19, 24, and 27), and (3, 7, 8, 12, 22, 26, and 29). If we compare these results to distortion-based clusters, we can notice that observers 3, 7, 8, 12, 22, and 29 of the third cluster prefer images with original level of contrast, while at the same time prefer original levels of saturation, and most of them do not like color quantization distortion effect.

When comparing contrast and saturation preferences (Figures 5 and 7), we can see that observers 21, 25, and 15 who prefer high saturation in images (Figure 3(a)) also prefer high contrast (Figure 3(e)). Other observers, who prefer high contrast (Figure 3(e)), also have a growing preference trend towards higher saturation (Figure 3(c)). Most of them are not sensitive to color quantization, but preferences vary in that case.

In addition, we checked the influence of content on individual preferences, but did not find a strong connection in this case.

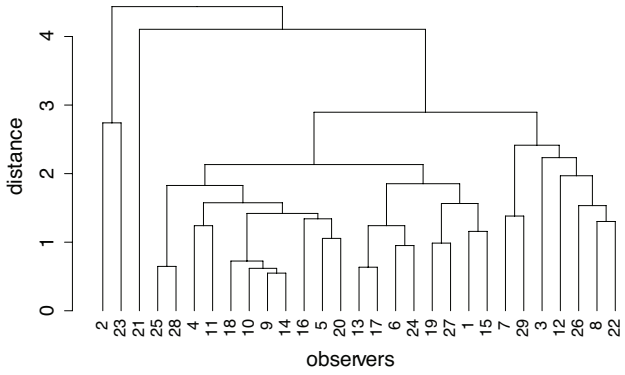


Figure 10. *Hierarchical clustering based on regression model coefficients for all distortions.*

Conclusion

In this work, we analyzed personal preferences in judging saturation, contrast, and color quantization distortions. We found groups of observers, clustered based on preference patterns, characterized by preferences of certain distortion levels. For saturation distortion we found two opposite trends in the observers' preferences. The first group prefers desaturated and over-saturated images, while the second group prefers more natural-looking images. We also found the third group, who preferred slightly saturated images.

In the case of contrast distortion, we found that preferences of different groups varied based on an optimal point, after which contrast increase did not further increase observers' opinion about image quality. While one group prefers natural-looking images, another group liked images with higher contrast. The difference was only in the level of contrast, which observers prefer. Color quantization distortion has shown observers who were more sensitive to visible artifacts and those who were less sensitive to them. We found that observers who liked over-saturated images also preferred images with higher contrast. Most of them were less sensitive toward color quantization, but some of the opinions differed. Another group of observers preferred original images without a change in the level of contrast or saturation.

Furthermore, we created a dataset, containing individual observers scores, in contrast to MOS based datasets available now. The dataset and the collected data will be publicly available and could be used to test various image quality metrics in Software and Data section in www.colourlab.no.

References

[1] Amirshahi, S.A., Del Pin, S.H.: Subjective quality evaluation: what can be learnt from cognitive science? In: CEUR Workshop Proceedings (2022)

[2] Amirshahi, S.A., Pedersen, M., Beghdadi, A.: Reviving traditional image quality metrics using cnns. In: Color and Imaging Conference. vol. 2018, pp. 241–246. Society for Imaging Science and Technology (2018)

[3] Amirshahi, S.A., Pedersen, M., Yu, S.X.: Image quality assessment by comparing cnn features between images. Journal of Imaging Science and Technology **60**(6), 60410–1 (2016)

[4] Bianco, S., Celona, L., Napoletano, P., Schettini, R.: On the use of deep learning for blind image quality assessment. Signal, Image and Video Processing **12**(2), 355–362 (2018)

[5] Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on image processing **27**(1), 206–219 (2017)

[6] Bringier, B., Richard, N., Larabi, M.C., Fernandez-Maloigne, C.: No-reference perceptual quality assessment of colour image. In: 2006 14th European Signal Processing Conference. pp. 1–5. IEEE (2006)

[7] Calabria, A.J., Fairchild, M.D.: Perceived image contrast and observer preference i. the effects of lightness, chroma, and sharpness manipulations on contrast perception. Journal of imaging Science and Technology **47**(6), 479–493 (2003)

[8] Cherepkova, O., Amirshahi, S.A., Pedersen, M.: Analyzing the variability of subjective image quality ratings for different distortions. In: 2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6 (2022). https://doi.org/10.1109/IPTA54936.2022.9784120

[9] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)

[10] Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: 2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–3. IEEE (2019)

[11] Mohammadi, P., Ebrahimi-Moghadam, A., Shirani, S.: Subjective and objective quality assessment of image: A survey. arXiv preprint arXiv:1406.7799 (2014)

[12] Ninassi, A., Le Callet, P., Autrusseau, F.: Pseudo no reference image quality metric using perceptual data hiding. In: Human vision and electronic imaging XI. vol. 6057, p. 60570G. International Society for Optics and Photonics (2006)

[13] Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Peculiarities, results and perspectives. Signal processing: Image communication **30**, 57–77 (2015)

[14] Sun, W., Zhou, F., Liao, Q.: Mdid: A multiply distorted image database for image quality assessment. Pattern Recognition **61**, 153–168 (2017)

[15] Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., Häkkinen, J.: Cid2013: A database for evaluating no-reference image quality assessment algorithms. IEEE Transactions on Image Processing **24**(1), 390–402 (2014)

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

[17] West, S.G., Finch, J.F., Curran, P.J.: Structural equation models with nonnormal variables: Problems and remedies. (1995)