Comparison of Metrics for Colorized Image Quality Evaluation

Ivana Žeger, Nikolina Bilanović, Gordan Šišul, Sonja Grgić
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
ivana.zeger@fer.hr

2022 International Symposium ELMAR | 978-1-6654-7003-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ELMAR55880.2022.9899824

***Abstract*—Colorization is a process of converting grayscale images into visually acceptable colorized images. It is a complex problem because no unique solution exists, i.e., the objects within the image cannot be associated with one color only. Colorized image quality assessment is also a complex problem because in real life the reference image for comparison with colorized results does not exist. The main purpose of colorization is convincing the observer in the credibility of the colorized image, disregarding color accuracy. For that reason, both subjective and objective evaluation of colorization results show imperfections. In this paper, the objective quality evaluation of colorized images has been conducted with five existing metrics and one new metric. The database of colorized images has been obtained by deep learning colorization methods. Based on the conducted subjective testing, a correlation between objective metrics and subjective grades has been determined. The results indicate that the improvement of objective metrics for the evaluation of colorized images is needed.**

***Keywords*—Colorization; Deep Learning Methods; Image Quality Evaluation; Subjective Testing; Objective Metrics**

I. Introduction

Color is defined as a subjective impression generated by human visual system when retina detects various wavelengths of light [1]. Colorization is a process of adding color to digital grayscale images. It is used mostly for reviving historical black-and-white photography [2]. Intensity of every pixel in grayscale images is represented with a scalar, while the pixels of color images contain complex three-dimensional information about color – RGB (red, green, blue) channels. For colorization, image needs to be transformed to a convenient color space which consists of the values of two chrominance components, which need to be assessed, and the luminance, which stays unmodified (e.g., YUV, CIELAB). In YUV system, chrominance components are U (defined by the difference B-Y, where B denotes the blue image) and V (defined by the difference R-Y, where R denotes the red image), while Y denotes the luminance component. In CIELAB, L denotes the luminance. Component *a* represents the green-red axis, while component *b* represents the blue-yellow axis [3].

Colorization methods are divided into methods which require user intervention (user-guided) and automatic (deep learning) methods [3]. User-guided colorization methods are more precise but unsuitable for real-time application. Deep learning colorization methods are a challenging research area in which image analysis and image processing are unified with machine and deep learning with the goal of obtaining visually persuasive results.

978-1-6654-7003-2/22/$31.00 ©2022 IEEE

The purpose of colorization is to provide color information to the original grayscale image, and to generate believable colorized image.

Image quality assessment has a key role during the system design for image analysis and processing. Methods for image quality assessment are divided into subjective and objective, depending on the end-user (human) involvement. Subjective methods are more precise at quality assessment because they match human impression of quality. However, they are expensive and long-lasting and thus not suitable for real-time application. Consequently, the evaluation of colorization is performed by the objective measures for color image quality assessment.

In this paper, the evaluation of objective methods for colorized image quality assessment is provided. The compared measures are peak signal-to-noise ratio (PSNR) [4], structural similarity index (SSIM) [5], patch-based contrast quality index (PCQI) [6], underwater image quality measure (UIQM) [7], underwater color image quality evaluation metric (UCIQE) [8] and the newly introduced chroma error ratio (CER). The image database consists of the original color images and test images made by colorization of grayscale version of the original color images using four methods: Zhang et al. [9], Vitoria et al. [10], Iizuka et al. [11] and Su et al. [12].

The remainder of the paper is organized as follows. In Section II, the used metrics are reviewed, and the proposed metric explained. In Section III, the image database is described. In Section IV, the results of image quality metrics evaluation are presented and discussed. The paper ends with Conclusion.

II. Metrics for Image Quality Evaluation

A. Overview of Selected Metrics

The chosen metrics used in this paper are PSNR, SSIM, PCQI, UIQM, and UCIQE.

PSNR [4] evaluates the difference between the original and the distorted image. The result is expressed in decibels. The metric cannot be used if the reference (original color) image is not available. PSNR value depends on the image content. For that reason, images with different content cannot be compared using PSNR. If used for colorized image quality assessment, PSNR is calculated separately for U component (PSNR_U) and V component (PSNR_V). Ultimately, the mean of these values is specified (PSNR_UV).

SSIM [5] is a more advanced metric because it takes into consideration the difference in luminance, contrast, and structure between the original image and the colorized image. SSIM values range between 0 and 1. Larger value indicates higher image quality. If used for colorized image quality assessment, SSIM is calculated separately for U component (SSIM_U) and V component (SSIM_V). Finally, the mean of these values is determined (SSIM_UV).

PCQI [6] also requires a reference image. Values of PCQI range from 0 to 1. The higher the PCQI value, the better the image regarding quality, similar as SSIM.

UIQM [7] is developed for quality assessment of underwater images. No reference image is required. UIQM is a linear combination of three components inspired by the properties of human visual system: colorfulness, sharpness, and contrast. Values tend to be between 0 and 1. Larger value indicates better image quality. The values may also be slightly larger than 1 (if the constants that define the measure are not precisely defined).

UCIQE [8] is another metric for the evaluation of underwater images. It is based on CIELAB chromatic system. It is primarily used for the evaluation of the deformities characteristic for underwater images. The reference image is unnecessary [13]. Values belong to the interval [0, 1]. Larger value implies higher image quality.

B. Proposed metric - CER

After image transformation from RGB to YUV color space, the image consists of three channels: Y, the luminance channel, and U and V, the chrominance channels. Both the original color image and the colorized image are transformed from RGB to YUV. Results of colorization are evaluated using the difference in U and V channels between the original color image and the colorized image. Fig. 1 shows the UV coordinate system. Each point in this system represents *u* and *v* values of pixels in U and V channels of an image. Values (*u, v*) form the chrominance vector. Values (u_i, v_i) represent the chrominance vectors of the original image, and (u_k, v_k) represent the chrominance vectors of the colorized image. The difference between the chrominance vector of the original image and the chrominance vector of the colorized image makes the difference vector, which indicates how much the original and the colorized image differ regarding the chrominance components. The new method is named chroma error ratio (CER). It is calculated as:

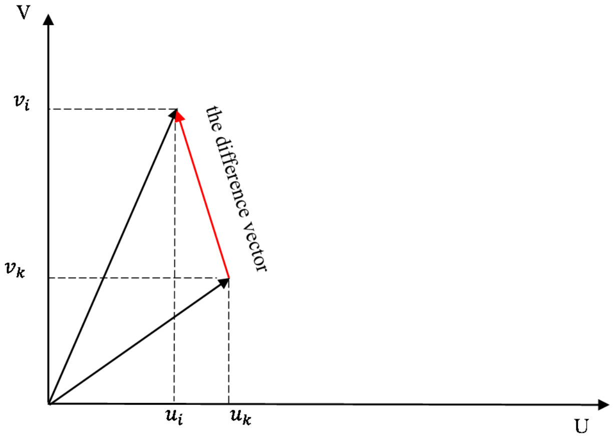


Figure 1. Graphical representation of the chrominance vectors and the difference vector

$$CER(dB) = 10\log\frac{\frac{1}{N}\sum_{k=0}^{N-1}[u_k^2 + v_k^2]}{\frac{1}{N}\sum_{k=0}^{N-1}[(u_i - u_k)^2 + (v_i - v_k)^2]}, \quad (1)$$

where *N* is the number of pixels in the original image and the colorized image. The results are given in decibels.

III. Image Database

For the evaluation of image quality assessment procedure, a database consisting of 34 color images of different content has been created. The resolution of images is 320x240. The images from the database and the corresponding names are shown in Table I.

The original images are transformed into their grayscale versions, as shown in Fig 2, and then colorized using the chosen automatic methods: Zhang et al. [9], Vitoria et al. [10], Iizuka et al. [11] and Su et al. [12]. The final version of the image database used for the evaluation of image quality metrics colorization consists of 34 original color images and 4*34 colorized images, which gives 170 images in total.

To show that the selected original color images have different characteristics, vector representation of colors (vectorscope) is shown for all the images, as can be seen in Fig. 3. Similar to Fig. 1, vectorscopes in Fig. 3 show the chrominance vectors for every pixel in an image. The x-axis of a vectorscope shows U component, while the y-axis shows V component. Vector length defines color saturation, and vector phase defines hue. By observing vectorscopes, range of colors and the dominant color can be assessed. The image with wide color range is Graffiti1, while Text2 is an example of scarce color range. The examples of colorized images obtained with Vitoria et al. [10] are given in Fig. 4. The corresponding vectorscopes are also shown. By comparing vectorscopes of original color images (Fig. 3) and vectorscopes of colorized images (Fig. 4), a significant reduction of color range is noticed.

IV. Results of Image Quality Evaluation

Quality evaluation of colorized images has been conducted. Six image quality metrics have been calculated.

Results of the objective metrics for all colorized images obtained by the mentioned colorization methods are shown in Fig. 5.

It can be seen that the highest results of PSNR_UV and CER are associated with the same images, i.e., Animals2 for methods of Zhang et al. [9] and Vitoria et al. [10] and Buildings for methods of Iizuka et al. [11] and Su et al. [12].

UCIQE assigns the highest result to Balloons. UIQM gave the highest score to Balloons in case of Zhang et al. [9] and Vitoria et al. [10], and to Sea1 in case of Iizuka et al. [11] and Su et al. [12]. SSIM_UV favors Sea4 for every method, except for Iizuka et al. [11], for which the highest value is obtained for Furniture2. The highest PCQI values are obtained for Food3 (for

TABLE I. ORIGINAL IMAGES DATABASE WITH THE NAMES OF IMAGES

Image name	Image	Image name	Image	Image name	Image	Image name	Image
1) Cars1		10) Food1		19) People1		28) Shop3	
2) Cars2		11) Food2		20) People2		29) Shop4	
3) Balloons		12) Food3		21) River		30) Buildings	
4) Flowers1		13) Sea1		22) Watches		31) Animals1	
5) Flowers2		14) Sea2		23) Text1		32) Animals2	
6) Tree		15) Sea3		24) Text2		33) Animals3	
7) Graffiti1		16) Sea4		25) Text3		34) Animals4	
8) Graffiti2		17) Furniture1		26) Shop1			
9) Graffiti3		18) Furniture2		27) Shop2			

Vitoria et al. [10] and Iizuka et al. [11]) and for Graffiti2 and Furniture2 (Zhang et al. [9] and Su et al. [12]).

To estimate which of the objective metrics coincides best with the subjective human impression, the correlation between the results of subjective quality evaluation and objective metrics needs to be calculated. For this purpose, subjective testing has been conducted. The double stimulus impairment scale method [14] has been used. Both the original color image and the colorized image were shown to the observers. It was pointed out to the observers that in addition to the accuracy of colors, the credibility of colorization needs to be estimated. The scale used for grading spans from 1 to 5 with the possibility of giving half grades (possible grades are: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5 and 5). Subjective testing was performed at University of Zagreb, Faculty of Electrical Engineering and Computing. 10 observers, non-experts aged between 20 and 60, participated in the research. Each observer had 136 images to grade. The duration of the test was 20 minutes per observer. For each image, grades from all the observers were averaged. The results of subjective testing are given in Fig. 6. Subjective testing gave the highest score to Sea2 for all methods except Su et al. [12], in which case Sea1 is rated with the highest score.

Table II. shows the Pearson correlation coefficient [15] between the results of subjective testing and the results of given objective metrics for every image.

The results show that the metrics for underwater image quality evaluation, UIQM and UCIQE have very weak correlation with subjective testing and thereby with subjective human impression of image quality. In other words, metrics that do not require a reference image show low level of correlation with the subjective judgement of quality for colorization purposes.

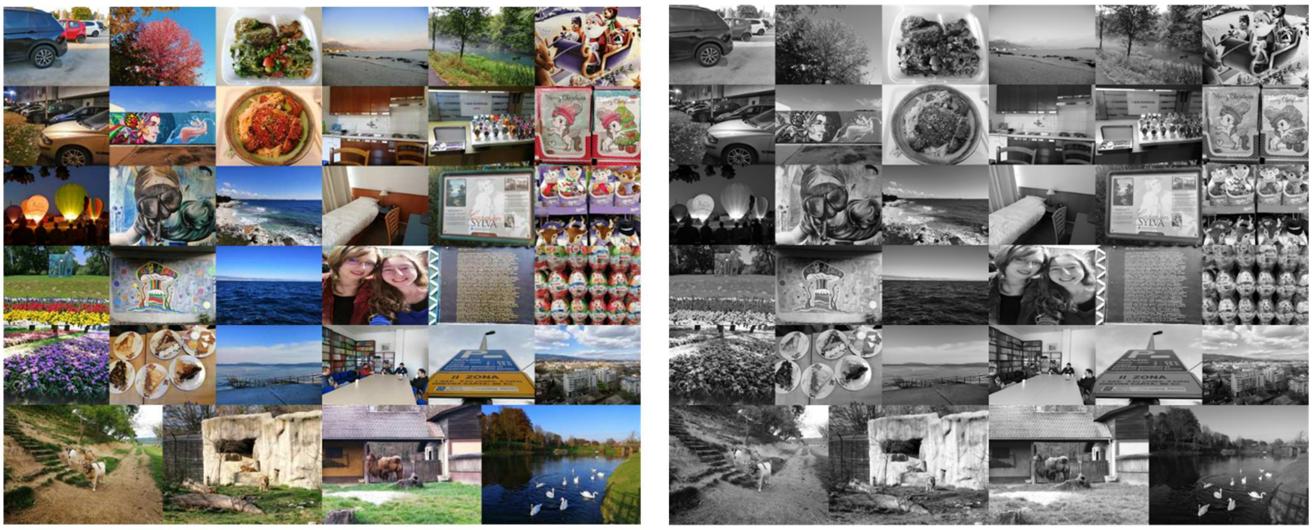


Figure 2. Original color images and their grayscale versions

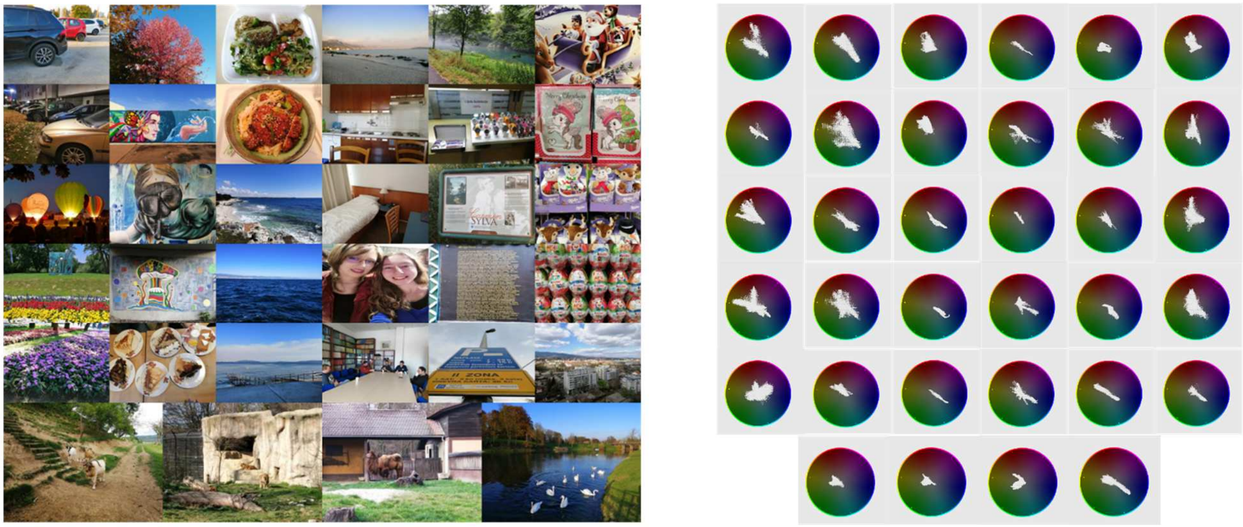


Figure 3. Original color images and their vectorscopes

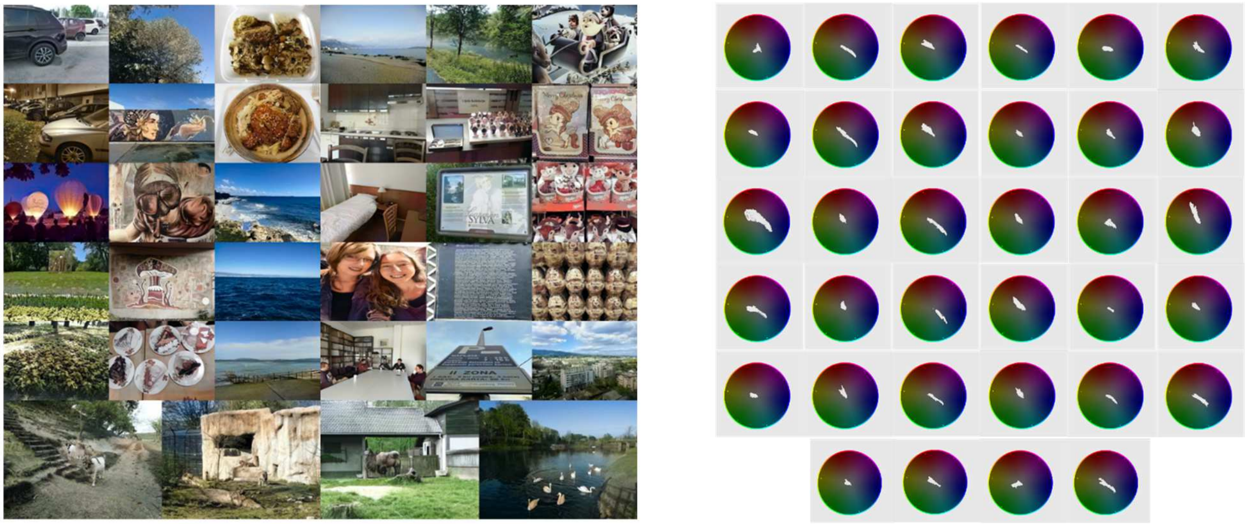
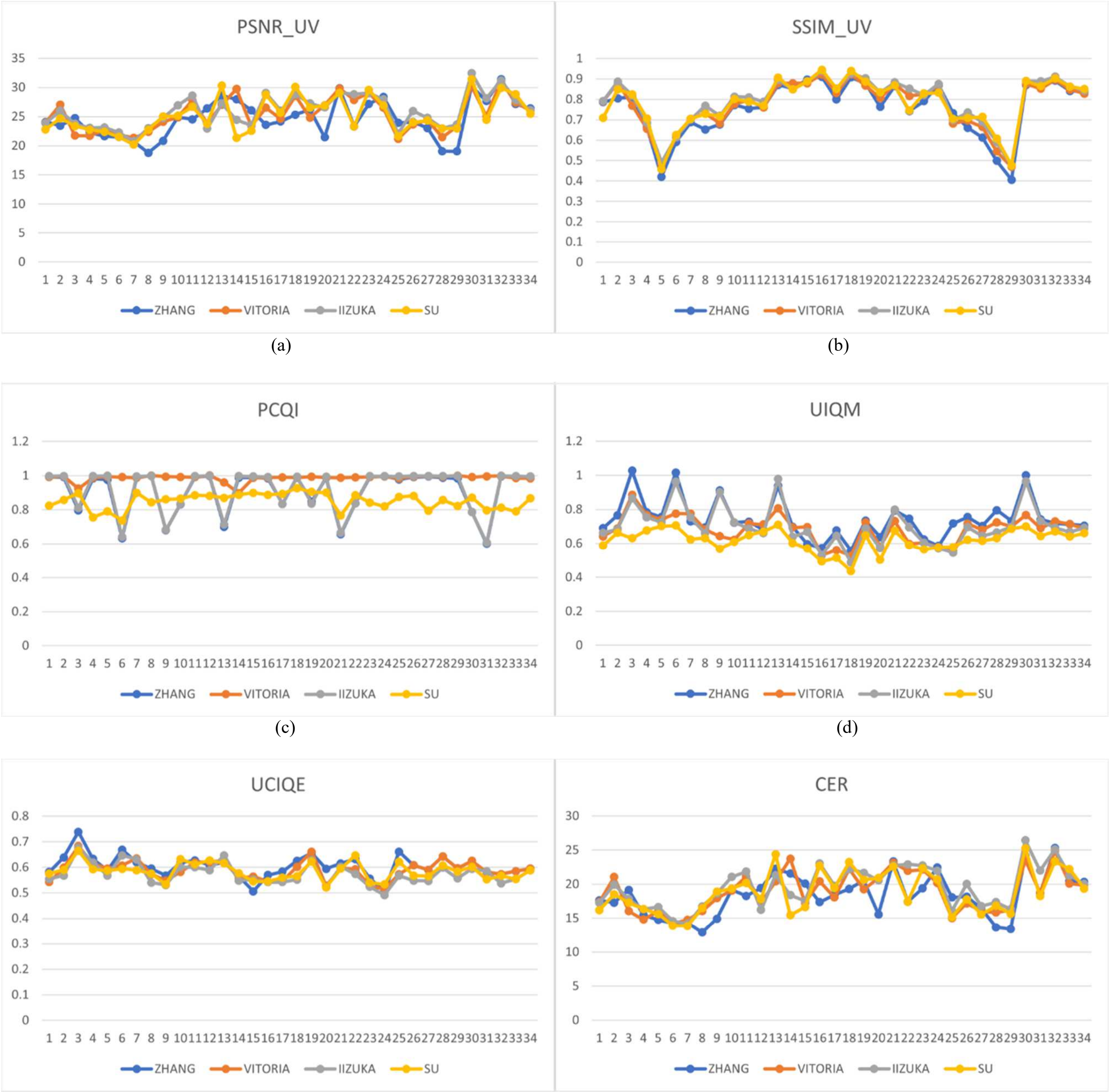


Figure 4. Images colorized using Vitoria et al. [10] method and the vectorscopes of colorized images



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. Results of objective quality metrics for different colorization methods: (a) PSNR, (b) SSIM, (c) PCQI, (d) UIQM, (e) UCIQE, (f) CER

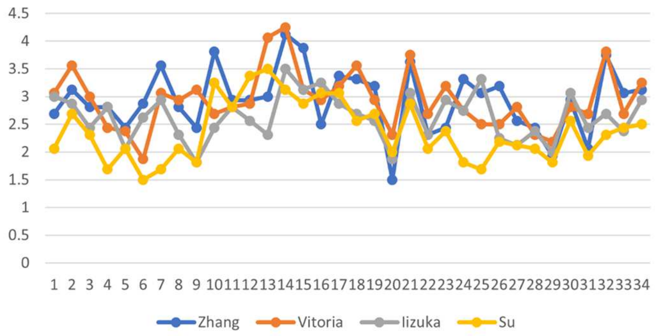


Figure 6. Results of subjective testing for different colorization methods

TABLE II. Pearson Correlation Coefficient Between Subjective Testing and Objective Metrics

	PSNR_UV	SSIM_UV	PCQI	UIQM	UCIQE	CER
Correlation coefficient	0.370	0.456	0.227	0.084	0.076	0.407

Low level of correlation between UIQM and UCIQE with the subjective judgement of quality can be a consequence of the method used for the subjective testing that is adjusted to the metrics which require a reference image. PCQI also has weak correlation with the subjective estimation of quality. SSIM_UV and CER show good correlation with subjective testing, as well as PSNR_UV. Correlation results indicate that SSIM_UV, CER and PSNR_UV better estimate quality of colorized images.

V. Conclusion

Colorization is a process of adding colors to grayscale images. The quality estimation of colorized images is a challenging field. There is no unique solution. Subjective testing is most successful in estimating colorization achievements because people are end users of the images. Subjective methods are precise and reliable, but also slow, impractical, and expensive. For that reason, objective methods are being used. Objective methods are fast and affordable. However, they are not designed for colorization quality evaluation. Most objective methods require both original and distorted image. In real-word application, the reference color image generally does not exist. Because of that, objective metrics which demand the original image can be used only in the process of colorization method development and testing.

In this paper, the quality of images colorized using methods [9-12] was evaluated. Colorized images were evaluated with PSNR, SSIM, PCQI, UIQM, UCIQE and CER. PSNR and SSIM were modified to assess quality of U and V channels. CER is a proposed metric based on the difference of the chrominance vectors. Subjective testing was also conducted.

PCQI, UIQM and UCIQE showed low correlation with the results of subjective testing. PSNR, SSIM and CER showed better results. SSIM and PSNR require RGB to YUV transformation, separate computation for U and V channels and calculation of mean (SSIM_UV and PSNR_UV). CER is simpler than PSNR and SSIM. CER works with U and V channels simultaneously; there is no need for the calculation of mean. PSNR, SSIM and CER demand a reference image. However, if quality of colorized images produced from grayscale images needs to be assessed, when the reference color image is not available, no-reference image quality metrics should be used. In future work, we plan to perform subjective evaluation with more viewers, include more metrics (especially no-reference) and carry out more detailed statistical analysis of the results.

Acknowledgment

This work was supported in part by Croatian Science Foundation (HRZZ) under the project number IP-2019-04-1064.

References

[1] R. Mukamal. (2017). *How humans see in color*. Accessed: May 3rd, 2022. [Online]. Available: https://www.aao.org/eye-health/tips-prevention/how-humans-see-in-color.

[2] I. Žeger and S. Grgić, "An overview of grayscale image colorization methods," *2020 International Symposium ELMAR*, 2020, pp. 109-112.

[3] I. Žeger, S. Grgic, J. Vuković and G. Šišul, "Grayscale image colorization methods: Overview and evaluation," in *IEEE Access*, vol. 9, pp. 113326-113346, 2021.

[4] National Instruments Corp. (Updated in 2020). *Peak signal-to-noise ratio as an image quality metric* [White paper]. https://www.ni.com/en-rs/innovations/white-papers/11/peak-signal-to-noise-ratio-as-an-image-quality-metric.html.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity, " *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, 2004.

[6] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2387-2390, 2015.

[7] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541-551, 2016.

[8] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062-6071, 2015.

[9] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649-666.

[10] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 2445-2454.

[11] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1-11, 2016.

[12] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7965-7974.

[13] G. Hou, T. Lu, Y. Li, Z. Pan, and B. Huang, "No-reference quality assessment for underwater images," http://dx.doi.org/10.2139/ssrn.4089412, 2022.

[14] Recommendation ITU-R BT.500-13 (01/2012). Methodology for the subjective assessment of the quality of television pictures.

[15] A. Zarić et al., "VCL@FER Image Quality Assessment Database", *Automatika*, vol.53, no. 4, pp. 344-354, 2012. [Online]. https://doi.org/10.7305/automatika.53-4.241