PERCEPTUAL METRIC FOR COLOR TRANSFER METHODS

Hristina Hristova, Olivier Le Meur, Remi Cozot, Kadi Bouatouch

University of Rennes 1

ABSTRACT

In this paper, we propose a perceptual model for evaluating results from color transfer methods. We conduct a user study, which provides a set of subjective scores for triplets of input, target and result images. Then, for each triplet, we compute a number of image features, which objectively characterize a color transfer. To describe the relationship between these features and the subjective scores, we build a regression model with random forests. An analysis and a cross-validation show that the predictions of our model are highly accurate.

Index Terms— color transfer, evaluation, random forests

1. INTRODUCTION ICIP

Color transfer between input and target images has raised a lot of interest in the past decade. The color transformation of an input image with regards to a target color palette provides solutions for color grading problems, such as example-based image color enhancement [1, 2, 3], time-lapse image hallucination [4], example-based video editing [5], etc.

Different color transfer methods often result in different output images. The process of determining the most plausible output image may be subjective, as it depends on a person's preference. Due to the lack of an objective metric for evaluating results from a color transfer, comparisons between existing methods are often carried out through a user study. Conducting a user study for each newly proposed method may be a tedious and time-consuming task. Moreover, the conditions, under which the study is conducted, and its protocol may vary with the group of people handling the evaluation process. That makes the comparison between different methods and the assessment of their performance challenging.

To ease the evaluation process, in this paper, we propose a model for objective evaluation of the color transfer quality. Our model explains the relationship between users' perception and a number of image features. To account for users' perception, we first conduct a user study on various color transfer results from six state-of-the-art methods. The study combines the aesthetic quality of the result with the quality of the color transfer, as perceived by the users. Then, for each result, we compute nine image features, which objectively describe the quality of a color transfer. We use these features to predict the scores from our user study.

We fit the set of the subjective scores and image features by a regression model with random forests. Our analysis shows that a regression with random forests is more accurate than linear and non-linear regressions. Our model is a general tool for assessing the perceptual performance of a color transfer. To this end, our model introduces an objective metric between three images - the input, the target and the result.

The paper is organized as follows. Section 2 presents commonly used metrics for objective image evaluation. Our user study and our regression model are introduced in section 3.2. An analysis of the proposed model is carried out in section 4. Finally, the last section concludes the paper.

2. RELATED WORKS

The performance evaluation of a color transfer is often addressed from two main perspectives [6]. First, a good color transfer method should not compromise the quality of the input image. Second, it should ensure a good transfer of color from the target image to the result. Hereafter, we present objective metrics, commonly used for color transfer evaluation.

Structural similarity metric (SSIM). The metric [7] measures the perceived quality of an image with regards to the original distortion-free image. In the color transfer context, it is used as a similarity metric between the input image and the result to measure the degree of artifacts in the result [8, 6]. To this end, SSIM is applied on the luminance channel of both the input and the resulting images.

Peak signal-to-noise ratio (PSNR). Like SSIM, PNSR is used as a quality measurement between an original image and a compressed one. And while SSIM outperforms PSNR as a similarity measure, PSNR has been adopted in the context of color dissimilarity by Hwang et al. [9]. The authors evaluate their color transfer method by measuring the difference in color (in terms of PSNR) between the input image and the result. The ultimate goal of Hwang et al,'s method is to maximize the PSNR metric.

Bhattacharya coefficient. The metric measures the amount of overlap between two distributions [10]. It has been adopted to compute the color similarity between a result from a color transfer and a target image [6].

Out-of-gamut metric. The metric has been proposed by Nguyen et al. [11] and used to evaluate their gamut-based color transfer method. The metric computes the distance be-

tween the gamut of the resulting image and the gamut of the target image.

The use of a combination of objective metrics, *e.g.* SSIM and Bhattacharya coefficient [6], SSIM and PSNR [9], may strengthen the objective evaluation by accounting for the quality of both the result and the color transfer. However, figure 1 shows that independent objective metrics may be weak predictors of the color transfer quality. To build a stronger prediction model, we consider an ensemble of features which account for subjective aspects of the color transfer evaluation.

Fig. 1. Results from two color transfer methods: (a) [2] and (b) [6]. SSIM and Bhattacharya coefficient, computed for both results, are inconsistent with our subjective evaluation.

3. OUR METHOD

We propose a model for predicting the human judgment on results from color transfer methods. The flowchart of our method is illustrated in figure 2. Given triplets of images (input and target images and a result), we conduct a user study and obtain subjective scores for each triplet. Then, for each triplet, we extract several image features. We apply a regression method between the image features and the subjective scores. That way, we build a model, predicting the color transfer quality of any image triplet, given its image features.

3.1. User study.

Protocol. We conducted a subjective evaluation of results from 6 color transfer methods, *i.e.* two non-parametric methods [12, 13], two global [1, 2] and two local [5, 6] parametric methods. We computed 20 results from each method, for a total of 120 results.

The participants were first shown tuples of input and target images. They had 5 seconds (prior to displaying the result) to get familiar with the images and to imagine what result they would expect to see. The evaluation of the results was guided by two criteria, encompassing the main aspects of the human judgment on a color transfer [6]. We were interested, first, in the way users perceived the match in color between the result and the target image and second, in users' judgment on the aesthetic quality of the result. Based on these two criteria, users were asked to give a single score reflecting their expectation about the result. Five-point scale (5-excellent, 4-good, 3-acceptable, 2-poor, 1-bad) was used in the evaluation. To avoid any possible bias, we used two repetitions per result.

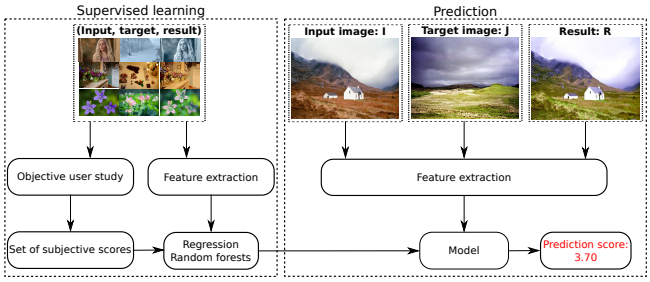

Fig. 2. Flowchart of our method. A regression model, allowing to predict the quality of a color transfer, is built from subjective scores and a set of image features.

Furthermore, an extra triplet, called a baseline, was inserted among the 120 results and shown randomly to each participant. The baseline was the only triplet, for which the colors of the result and the target image differed significantly one from another. Therefore, users were expected to give the baseline the lowest score. That way, the evaluation of the baseline identified how trustworthy a user's judgment was.

The number of participants in the user study was 20 and the majority of them had average image editing expertise. Each user evaluated the results individually using an online platform. The order of displaying the results was random and different for each participant. A short training session took place before the real test in order for the users to get familiar with the presented task and the platform.

Data. For each image triplet, we computed a final subjective score as the weighted mean over the scores of all participants. We used weights, which were inversely proportional to the baseline score, given by a participant. Figure 3 shows the score distribution for each of the six color transfer methods. According to our user study, Hristova et al.'s method [6] outperforms the others by obtaining the highest mean score. Furthermore, we observe differences in the score distribution respectively between the two non-parametric methods and the two local parametric methods. In contrast, the performance of the two global parametric methods is similar.

3.2. The regression model

Objective features. We extract nine image features for each image triplet (**I**, **J**, **R**) in our image dataset, where **I** denotes the input image, **J** the target image and **R** the result. We denote the channels of an image **H** in the CIE Lab color space by $\mathbf{H}_L$, $\mathbf{H}_a$ and $\mathbf{H}_b$ (**H** stands for either of the images **I**, **J** and **R**). The Bhattacharya coefficient between two image channels $\mathbf{C}_1$ and $\mathbf{C}_2$ is denoted by $BC(\mathbf{C}_1, \mathbf{C}_2)$. The features, considered in our model, are defined below:

- SSIM(**R**, **I**) measures the degree of artifacts in the result, caused by the color transfer and it is computed as in [6];
- Luminance histogram similarity, computed between the images **R** and **J** as $BC(\mathbf{R}_L, \mathbf{J}_L)$;

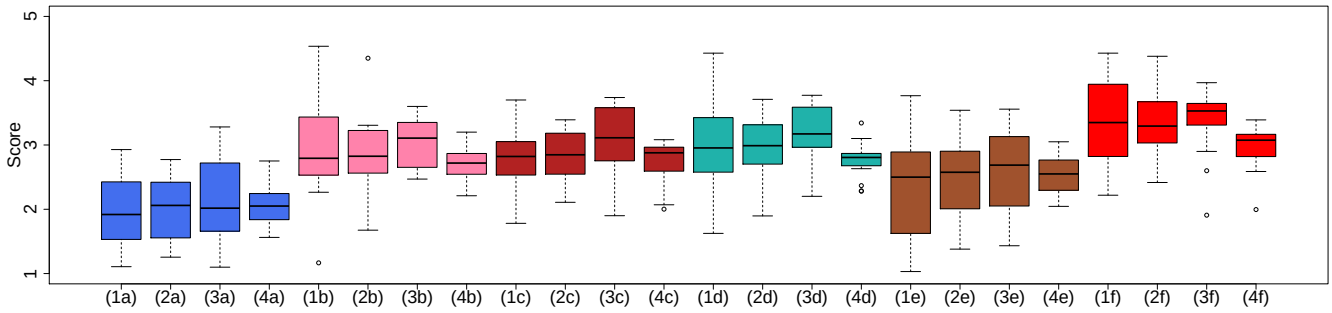

Fig. 3. Box plots for: (1) the scores from our user study, (2) the predictions from regression with random forests, (3) the predictions from non-linear regression and (4) the predictions from linear regression. Each group of four identically-colored box plots represents one of six color transfer methods: (a) [12], (b) [13], (c) [1], (d) [2], (e) [5], (f) [6].

- Color histogram similarity, computed as the mean of $BC(\mathbf{R}_a, \mathbf{J}_a)$ and $BC(\mathbf{R}_b, \mathbf{J}_b)$;
- Saliency map similarity between the images **R** and **J**, computing the similarity between two saliency maps, as presented in [14]. For the sake of robustness, discussed in [15], the saliency map of each image is computed as the mean of two saliency maps, obtained with the methods in [16] and [17];
- Histogram similarity of color appearance attributes, computed as $BC(\mathbf{R}_l, \mathbf{J}_l)$, where l denotes one of five color appearance attributes, *i.e.* brightness, lightness, chroma, colorfulness and saturation [18].

The first three metrics represent the basic objective evaluation, previously used in [6]. They identify how the degree of artifacts in the result and the match in color and light between the result and the target image influence users' perception. Comparing saliency maps allows to test the influence of a color transfer on salient areas. In the best case, the saliency maps of the input image and the result should be the same.

The channels of the CIE Lab color space represent well the light and color distributions of an image. However, they do not account for color appearance phenomena [18], *e.g.* chromatic adaptation, Hunt effect, Stevens effect, etc., which occur with a change in the viewing conditions. In contrast, the color appearance attributes describe the perceptual aspects of color. The nine objective features represent the set of the independent variables in our model.

To apply a regression method, we first carry out a proper preprocessing of the data, collected from our user study, as presented hereafter.

Quantization and over-sampling. Due to the use of baseline weights, the set of subjective scores is highly undersampled as there is an insufficient number of triplets per score. To tackle this issue, we round the scores up to the nearest 0.5 so that at least two triplets have then the same score. Despite the performed quantization, the set of subjective scores remains imbalanced. As discussed in [19], imbalanced data may compromise the performance of regression methods, such as discriminant analysis and decision trees. Therefore, we balance our data by using the synthetic minority oversampling technique [20]. In our model, the balanced set of scores represents the set of dependent variables.

Regression methods. To fit our data by a simple linear model, we first use linear regression. Our experiments with linear regression, however, have pointed out strong non-linearities in the set of subjective scores and image features. That is why, we also apply non-linear regression. We use support vector regression with radial basis function kernel [21].

To further improve the accuracy of the fit, we use random forests [22]. Random forests is a learning method, which constructs an ensemble of decision trees. Although decision trees provide a classical model for fitting various data, they tend to over-fit the training sets and introduce a high variance. Random forests aim to correct the high variance of decision trees and provide a more accurate prediction by using bootstrapping. Random forests take into account strong non-linearities in the data and are not sensitive to correlated predictors.

4. RESULTS

Cross-validation. To analyze the behavior of each of the three regression methods, we performed k-fold cross-validation with 10 splits. For each regression method, table 1 shows the mean correlation between predicted scores and actual subjective scores, and the mean square error (MSE), computed over all test sets. Linear regression is the least correlated with our data method. The regression with random forests provides the most correlated with the subjective scores prediction.

The same conclusion is drawn from the box plots in figure 3. The predicted score distributions from the linear regression is characterized by a low variance and significantly differs from the distribution of subjective scores per color trans-

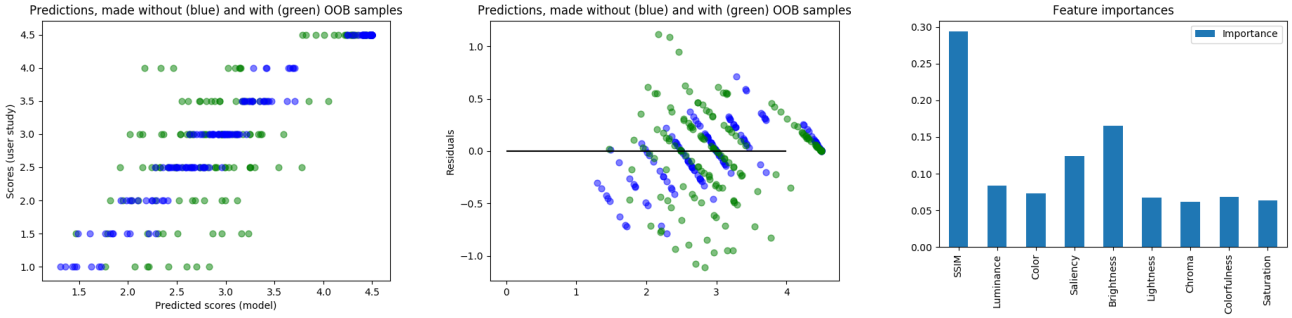

Fig. 4. Plots from left to right: relationship between the predicted scores from our model and the actual scores from our user study; residual distribution in our model; bar plots of the importance of each objective feature in our model.

	Random forests	**Linear**	**Non-linear**
Corr	0.765 ± 0.133	0.567 ± 0.135	0.644 ± 0.157
MSE	0.472	0.808	0.889

Table 1. Correlation $\pm$ standard deviation and MSE over all 10 test sets in our cross-validation.

fer method. Overall, the prediction, made with non-linear regression, is much more accurate than the prediction, made with linear regression. However, non-linear regression fails for Hristova et al.'s method [6]. The most consistent regression method is the random forests. It provides the best fit for the score distributions of all six color transfer methods.

Furthermore, figure 5 illustrates the actual and predicted scores for a representative triplet of images from our user study. Random forests provide a very accurate score, close to the actual subjective score. The linear regression gives the poorest prediction, followed by the non-linear regression.

Our analysis has shown that regression with random forests describes very precisely the relationship between the users' evaluation and the nine objective image features and therefore, we adopt it in our model. Hereafter, we discuss in more details the accuracy of this regression model.

Fig. 5. Score predictions from random forests, linear and non-linear regressions. Color transfer method used: [6].

Random forests accuracy. The accuracy is characterized by the coefficient of determination equal to 0.94, the out-of-bag (OOB) error equal to 0.58, and the mean square error (MSE) of the prediction equal to 0.063. The high coefficient of determination indicates that a major part of the variability of the subjective scores is explained by our model. The left-most plot in figure 4 illustrates a strong correlation between the predictions (made with and without using OOB samples) and the actual user scores. Figure 4 also shows the residuals of the predictions (made with and without using OOB samples). The residuals are symmetrically distributed around the x-axis and are clustered around -1.5 and 1.5 on the y-axis. The random patterns in the residual plot indicate that our model provides a decent fit of the data.

Feature importances. Despite the lack of an analytical model in the random forests, a summary of the importance of each predictor can be computed [22]. The right-most plot in figure 4 shows the importance of each image feature in our dataset. SSIM has a significantly greater importance than the second most important feature, the brightness histogram similarity. This shows that users are unlikely to give high scores to a color transfer, which compromises the integrity of the result. The first three most important features, *i.e.* SSIM, brightness histogram similarity and saliency map similarity, are all perceptual features. Surprisingly, the color histogram similarity plays a much less significant role in our model. This is due to the fact that the Bhattacharya coefficient represents the overlap between color distributions in the CIE Lab color space without accounting for various color appearance effects.

5. CONCLUSION

In this paper, we have presented a regression model, based on random forests, for objective evaluation of results from color transfer methods. Our method describes the relationship between the scores from our user study and nine image features, used for objective image evaluation. Our analysis has shown that perceptual image features, such as SSIM and saliency, play a main role in predicting the color transfer quality. Furthermore, a cross-validation indicates a high correlation between predicted and actual scores. To this end, our regression model can be used as a general prediction of a user's judgment on any color transfer result.

6. REFERENCES

[1] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[2] François Pitié and Anil Kokaram, "The linear monge-kantorovitch linear colour mapping for example-based colour transfer," 2007.

[3] Hasan Sheikh Faridul, Tania Pouli, Christel Chamaret, Jürgen Stauder, Alain Trémeau, Erik Reinhard, et al., "A survey of color mapping and its applications.," in *Eurographics (State of the Art Reports)*, 2014, pp. 43–67.

[4] Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman, "Data-driven hallucination of different times of day from a single outdoor photo," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, pp. 200, 2013.

[5] Nicolas Bonneel, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister, "Example-based video color grading.," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 39–1, 2013.

[6] Hristina Hristova, Olivier Le Meur, Rémi Cozot, and Kadi Bouatouch, "Style-aware robust color transfer," in *Proceedings of the workshop on Computational Aesthetics*. Eurographics Association, 2015, pp. 67–77.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] Oriel Frigo, Neus Sabater, Vincent Demoulin, and Pierre Hellier, "Optimal transportation for example-guided color transfer," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 655–670.

[9] Youngbae Hwang, Joon-Young Lee, In So Kweon, and Seon Joo Kim, "Color transfer using probabilistic moving least squares," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3342–3349.

[10] Frank J Aherne, Neil A Thacker, and Peter I Rockett, "The bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998.

[11] RMH Nguyen, SJ Kim, and MS Brown, "Illuminant aware gamut-based color transfer," in *Computer Graphics Forum*. Wiley Online Library, 2014, vol. 33, pp. 319–328.

[12] Tania Pouli and Erik Reinhard, "Progressive histogram reshaping for creative color transfer and tone reproduction," in *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*. ACM, 2010, pp. 81–90.

[13] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot, "N-dimensional probability density function transfer and its application to color transfer," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. IEEE, 2005, vol. 2, pp. 1434–1439.

[14] Olivier Le Meur and Thierry Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior research methods*, vol. 45, no. 1, pp. 251–266, 2013.

[15] Olivier Le Meur and Zhi Liu, "Saliency aggregation: Does unity make strength?," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 18–32.

[16] Jonathan Harel, Christof Koch, Pietro Perona, et al., "Graph-based visual saliency," in *NIPS*, 2006, vol. 1, p. 5.

[17] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.

[18] Mark D Fairchild and Garrett M Johnson, "icam framework for image appearance, differences, and quality," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 126–138, 2004.

[19] Sven F Crone and Steven Finlay, "Instance sampling in credit scoring: An empirical study of sample size and balancing," *International Journal of Forecasting*, vol. 28, no. 1, pp. 224–238, 2012.

[20] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[21] Jean-Philippe Vert, Koji Tsuda, and Bernhard Schölkopf, "A primer on kernel methods," *Kernel Methods in Computational Biology*, pp. 35–70, 2004.

[22] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.