A Metric for Video Blending Quality Assessment

Zhe Zhu, Hantao Liu, *Member, IEEE*, Jiaming Lu, and Shi-Min Hu, *Member, IEEE*

***Abstract*— We propose an objective approach to assess the quality of video blending. Blending is a fundamental operation in video editing, which can smooth the intensity changes of relevant regions. However blending also generates artefacts such as bleeding and ghosting. To assess the quality of the blended videos, our approach considers the illuminance consistency as a positive aspect while regard the artefacts as a negative aspect. Temporal coherence between frames is also considered. We evaluate our metric on a video blending dataset where the results of subjective evaluation are available. Experimental results validate the effectiveness of our proposed metric, and shows that this metric gives superior performance over existing video quality metrics.**

***Index Terms*— Video quality assessment, video blending.**

I. Introduction

BLENDING is a fundamental operation in video editing. Due to the wide range of applications, various blending techniques [1]–[8] have been proposed. Although these methods were initially developed for images, they can be easily adapted to videos in a frame-by-frame fashion. At the same time, blending techniques specially designed for videos [9], [10] have also been proposed. They were designed to tackle temporal coherence issues to avoid potential flickering in the blended videos. These methods can handle challenging scenes, such as videos having moving objects with uncertain boundaries [9] and stereoscopic videos [10].

The aim of blending is to smooth the illumination inconsistencies between different blending regions. However artefacts can be generated, lowering the blending quality significantly. Two most common types of artefacts in blending are ghosting artefacts and bleeding artefacts, which are illustrated in Figure 1. While directly stitching without blending is not visually pleasing(Figure 1 (a)), artefacts could make the blending results worse(Figure 1(b),(c)). Since blending is a required operation in video composition/fusion, assessing the blending quality is important. However, to the best of our knowledge, no prior work has addressed this problem so far. One possible reason is that there is no video blending dataset with the ground truth of blending quality.

Manuscript received March 25, 2019; revised August 23, 2019 and October 18, 2019; accepted November 14, 2019. Date of publication November 28, 2019; date of current version January 28, 2020. This work was supported by the Natural Science Foundation of China (Project Number 61521002), Research Grant of Beijing Higher Institution Engineering Research Center and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marta Mrak. *(Corresponding author: Zhe Zhu.)*

Z. Zhu, J. Lu, and S.-M. Hu are with TNList, Tsinghua University, Beijing 100084, China (e-mail: ajex1988@gmail.com; loyavaforever@gmail.com, shimin@tsinghua.edu.cn).

H. Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K. (e-mail: LiuH35@cardiff.ac.uk).

Digital Object Identifier 10.1109/TIP.2019.2955294

(a) (b) (c)

Fig. 1. Typical artefacts in image blending(images are from [11]). (a) directly stitching without blending. (b) ghosting artefacts. (c) bleeding artefacts(marked in red rectangle).

Recently Zhu *et al.* [11] performed a comparative study of blending algorithms for videos. A video benchmark which contains videos captured under different conditions was built. Six popular blending algorithms were implemented and applied to the captured videos. Thirty participants were involved in a subjective quality evaluation. Settings based on the ITU-R Recommendation BT.500-13 [12] were followed and a 5-point scale (i.e., 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent) was used for quality scoring. This inspires and enables us to develop a metric for video quality assessment based on this dataset, given that the mean opinion score for each blended video can be used as the ground truth.

In this paper, we propose a metric that can automatically and accurately quantify the quality of blended videos as perceived by humans. We consider three relevant terms: illumination consistency, visual artefacts and temporal coherence. Since smoothing illumination is the aim of blending, we quantify and compare the illumination conditions in different sides of the blending boundary. For visual artefacts we mainly consider ghosting artefacts and bleeding artefacts, and regard them as negative quality effects. We also consider temporal coherence since it represents the stability of the blended video and affects the overall video quality significantly. Our metric is derived by combing all the three aforementioned terms and its performance is validated against the subjective evaluation results as provided in [11].

The rest of the paper is organized as follows. We introduce related work in Section II. In the algorithm part, we first give details of the image blending quality assessment metric in Section III, then describe its extension to videos in Section IV. The performance of the proposed algorithm is evaluated in Section V and we conclude our work in Section VI.

II. Related work

A. Video Quality Assessment

According to [13] video quality assessment approaches have gone through four key stages: Quality of Service(QoS) monitoring, subjective testing, objective quality modeling and data-driven analysis. Early works [14], [15] mainly consider the QoS for video delivery over networks, and choosing optimal QoS parameters [15]–[19] is the most widely adopted strategy. However the application-end QoS is not always consistent with the user-end Quality of Experience (QoE), and their relation is nontrivial. Thus, a direct way to obtain the ground truth of user QoE is through subjective testing. To standardize the subjective evaluation process, the Video Quality Experts Group (VQEG) made detailed plans [12] for conducting subjective tests. Although rather accurate, subjective testing is tedious and costly, and more attention has been paid to objective video quality metrics. Given the subjective test results as ground truth, parameters of objective models [20]–[22] based on how the Human Visual System(HVS) processes the video information can be fine-tuned towards reliable quality prediction. Due to the limited amount of visual stimuli used in the subjective testing under the laboratory environment, objective metrics that relied on these subjective data can not be well applied in general applications. Recently the large amount of videos online as well as the user behavior data(viewing time, return probability etc.) make the data-driven video QoE assessment a new trend. Unlike traditional quality scores, viewing time [23], [24], number of views [25] and return probability [26] become the video QoE measurements. Under some circumstances the user ratings are also available. The models trained on these worldwide datasets are much more accurate and reliable than the traditional models. There are several surveys [27]–[29] of video QoS and QoE assessments and readers can refer to them for more details.

B. Image and Video Blending

The most straightforward way to blend two images is to linearly combine the regions that need to be blended. Usually it is called feather blending. Under some conditions the regions to be blended are not well aligned, and ghosting artefacts might appear. To alleviate this problem multi-band blending [1] combines blending results from versions of the images containing different frequencies. In feather blending and multi-band blending, all the pixels in the overlapped regions will be changed after blending. In some applications such as object insertion, only one region needs to be changed to fit the other. Poisson blending [2] elegantly formulates image blending via a Poisson equation. The solution can be obtained by solving a large linear equation, which makes original Poisson blending time consuming. While the original Poisson blending finds the final pixel values directly, an alternative approach is to calculate the offset map first and then add the offset map to the original image. Since the offset map is smooth regardless of the image content, acceleration can be made based on this property. One way is to use the quadtree [3] to approximate the whole offset map, and this can significantly reduce the number of variables in the final equation. The other way to approximate the offset map is to construct a harmonic interpolant from the boundary intensity differences using mean value coordinate (MVC) [4], [7]. There are also other modifications [5], [6] of the original Poisson blending, and these approaches also change all the regions to be blended. Unlike image blending which has been well studied, less attention has been paid to video blending. One way to adapt blending techniques from images to videos is to apply the image blending methods in a frame-by-frame fashion [30]. This strategy is straightforward and effective, but lack of temporal consistency, which may lead to jittering between frames. A practical solution to handle spatial-temporal coherence is to add a smoothness term [31] in the overall energy, but this requires additional computation efforts.

III. Image Blending Quality Metric

We first introduce our image blending quality metric in this section and then extend it to videos in the next section. For simplicity we only consider the situation that two image regions(a source region and a target region) are to be blended, and extension to multiple image blending is straightforward.

Since our goal is to evaluate the blending quality, we do not consider the quality of the original image content(color accuracy, image sharpness etc). The aim of blending is to smooth the illumination discontinuity, so we quantify the illumination conditions of relevant image regions to calculate the illumination consistency. Consistent illumination means the images should look like as they were captured in the same lighting condition with the same camera parameters(such as ISO, exposure time, etc.). Consistent illumination is preferred in image blending since it makes the output look natural. We also quantify bleeding and ghosting artefacts and calculate bleeding degree and ghosting degree as negative effects of blending.

Note that the input are images/videos associated with binary masks. The binary mask indicates the regions to be blended. Then blending boundaries can be calculated on the masks using the approach in [11].

A. Objective

In [29] three conditions were defined for an image quality metric: symmetry, boundedness and unique maximum. For image/video blending quality assessment, as mentioned in [11], some blending algorithms are inherently not symmetric, which means changing blending orders can lead to different blending results. Thus we do not require our blending quality metric to be symmetric. Since blending quality is rather subjective so unique maximum for a blending quality metric does not make much sense. We only require our metric to be bounded, ranging in $(0, 1]$. We would also like larger values indicating better blending results in our metric.

B. Illumination Consistency

To calculate the illumination consistency the illumination condition of the source(S) and target(T) region should be

(a) (b) (c)

Fig. 2. Two typical scenes and their intrinsic decomposition results. Row 1 is a well blended scene while row 2 is a scene without blending(directly trimming and compositing). (a) original image. (b) reflectance layer. (c) shading layer.

recovered first. Since using HVS model [29] to calculate the illumination change is rather difficult, we approximate the illumination condition of an image using the shading layer recovered by intrinsic image decomposition [32]. In intrinsic image decomposition, an image is decomposed into a reflectance layer and a shading layer which multiply to form the original image. Two typical scenes with their intrinsic decomposition results are illustrated in Figure 2. Intrinsic image decomposition can be generalized to intrinsic video decomposition [33] which we will use to calculate the shading layer of the video. Since the illumination condition in different regions of an image can vary, we calculate the illumination consistency in a local region around the blending boundary.

Suppose S and T are two image regions to be blended with a blending boundary B. For each point b_i on the boundary, we sample a $n \times n$ patch in S in the direction perpendicular to the boundary line. There is a trade-off in choosing the size of the patch. Choosing large patch size is inappropriate since we assume that the shading layer is locally smooth. Small patch size is not robust since there may exist noise in a small patch. Through experiment we found $n = 7$ works well for the input resolution of 1024×1024. For other input resolutions the patch size n can be altered accordingly. The sampling strategy is illustrated in Figure 3. Then pixel values in this patch of the shading layer are averaged as b_i^r, representing the illumination value for b_i. Then an illumination feature vector v_s for the source region can be obtained by concatenating the illumination values for all the points along the boundary:

$$v_s = [b_1^r, b_2^r, \ldots, b_i^r, \ldots], \quad b_i \in B \tag{1}$$

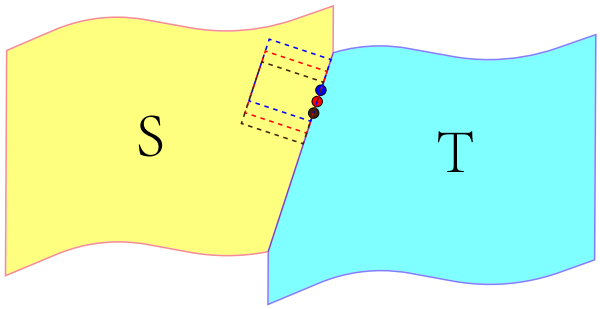


Fig. 3. Sampling strategy for calculating the illumination feature vector. For each side, a rectangle region is sampled for each boundary point. For three adjacent pixels on the blending boundary(blue, red and gray dots), their values in the illumination feature vector are calculated in the corresponding local region(blue, red and gray rectangles). In this figure, we illustrate the sampling strategy in region S, and the same operation should be applied to region T.

Same operation can be applied for the target region side T, and an illumination vector v_t of the same length can be also obtained. Then we calculate the mean value of all the elements of v_s and v_t as μ_s and μ_t respectively. Then the illumination consistency term q_i is defined as:

$$q_i = \frac{2\mu_s\mu_t + \delta}{{\mu_s}^2 + {\mu_t}^2 + \delta} \tag{2}$$

In the above equation δ is used to avoid dividing by zero and in our implementation it is set to be 1e-8. Obviously the illumination consistency term satisfies the boundedness condition.

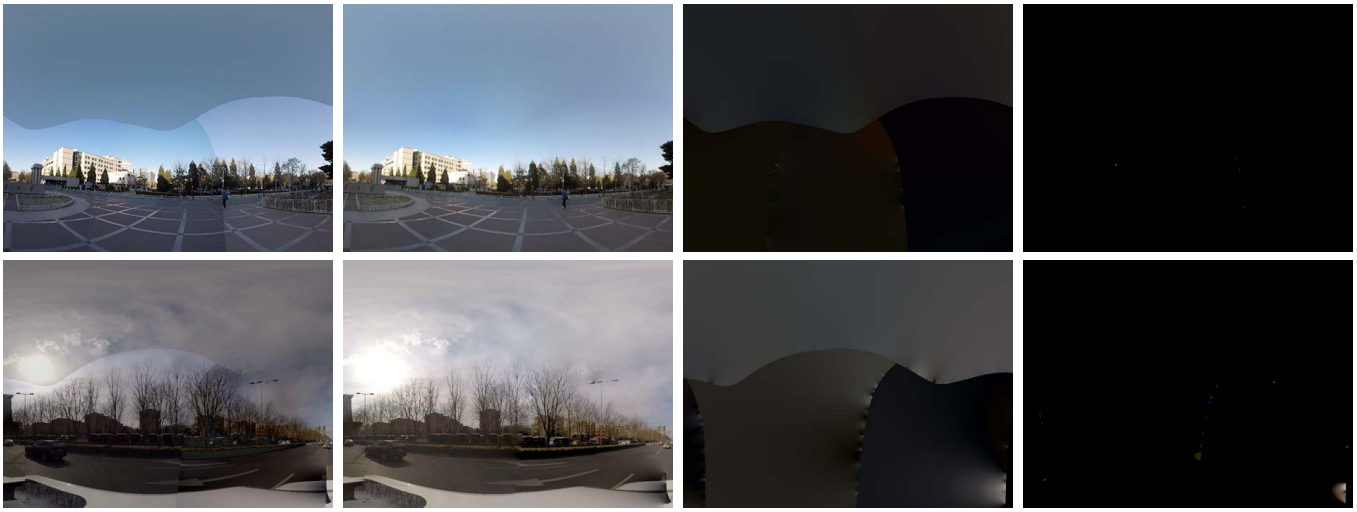


Fig. 4. A scene with visually unnoticeable bleeding artefacts(row 1) and a scene with obvious bleeding artefacts(row 2).(a) unblended image. (b) blended image. (c) offset map. (d) bleeding map. q_i, q_b and q_g for row 1 (a) are 0.82, 0.99, 1.0, for row 1 (b) are 0.98, 0.99, 1.0, for row 2 (a) are 0.87, 0.99, 1.0, for row 2 (b) are 0.97, 0.85, 1.0.

C. Bleeding Artefacts

We follow the idea in [11] and quantify the bleeding artefacts using the offset map. The offset map is calculated by subtracting the blended image from the unblended image and calculating the absolute value of each pixel. In the blended image a bleeding artefact manifests as a particular color leaking to its surroundings(Figure 4 (b)) while in the offset map it appears as highlighted regions(Figure 4 (c)). We first detect these "bleeding regions" and then calculate the energy of these regions. We observed that the bleeding regions usually have much lower or higher intensities and only occupy a small portion of the whole offset map, and the rest of the offset map is of very smooth regions. Thus the bleeding regions can be detected by truncating the smooth regions by setting a proper intensity threshold τ:

$$\tau = \alpha \frac{sum(M_e)}{C_n(M_e) + \delta} \tag{3}$$

In the above equation M_e is the binarized offset map calculated using Otsu method [34]. $sum(M_e)$ is the sum of the energies at the non-zero positions in the binarized map and $C_n(M_e)$ is the number of non-zero values in the offset map. α is used for safe thresholding and in our implementation it is set to be 2. Same as above, δ is used to avoid dividing by zero and is set to be 1e-8.

Then the bleeding map(see examples in Figure 4 (d)) M_b is calculated by:

$$M_b(x) = \max(0, x - \tau) \tag{4}$$

Here x denotes the pixels of the blended region. The final bleeding degree is calculated as follows:

$$q_b = e^{-\frac{sumsqr(M_b)}{C_n(M_b)+\delta}} \tag{5}$$

Here $sumsqr()$ is the sum of squared elements operation. This term also satisfies the boundedness condition.

D. Ghosting Artefacts

Ghosting artefacts appear in overlapping regions. Imagine that the images are perfectly blended, in the overlapping regions although the pixel intensities may change, image gradients should remain the same. Based on this observation we use the gradient differences of the overlapping regions before and after blending to quantify ghosting artefacts. Formally suppose S and T are two regions to be blended, and $O = S \cap T$ is the overlapping region. R is the blended result. Then the ghosting degree is defined as:

$$q_g = e^{-\frac{\sum_{i \in O} (R_g(i) - T_g(i))^2 + (R_g(i) - S_g(i))^2}{C_n(O)+\delta}} \tag{6}$$

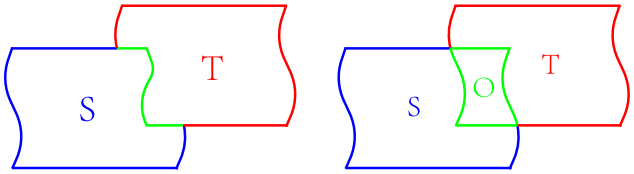
Here $C_n()$ calculates the number of all the elements of a region and i indicates pixel position. S_g, T_g and R_g indicate the gradient of the source, target and blended image respectively. The ghosting term satisfies the boundedness condition.

E. Single Frame Blending Quality Metric

The illumination consistency, bleeding degree and ghosting degree are relatively independent. For example, the change of illumination consistency will not affect the bleeding degree. Thus we multiply these terms [29] to calculate the overall quality. Given the illumination consistency, bleeding degree and ghosting degree, the blending quality of a single frame(one image) q is calculated by:

$$q = {q_i}^{\lambda_i} {q_b}^{\lambda_b} {q_g}^{\lambda_g} \tag{7}$$

In the above equation, λ_i, λ_b and λ_g are the weights balancing their corresponding terms. In our implementation we empirically set λ_i to be 0.2, λ_b to be 0.5 and λ_g to be 0.5, with the aim of selecting the combination of the weights that achieves the best predictive performance. Note that lower weight indicates higher influence to the overall value. We give



(a) (b)

Fig. 5. Two types of formulation in blending. (a) One region changes to fit the other region. (b) There is an overlapping region and the blending was done in the overlapping region.

lower weight to illumination consistency term since it has larger influence on the overall blending quality.

F. Implementation Details

We have introduced the calculation of blending quality metric of two neighbouring regions. In real world scenarios there may be multiple regions to be blended. For example, in [11] the final panorama is obtained by blending 6 images. Besides, as mentioned in [11] there are two different types of formulations in blending:

1) *Type A*:Either source region or target region changes to fit the other region. We illustrate this in Figure 5 (a).
2) *Type B*:Only the overlapping region changes. We illustrate this in Figure 5 (b).

Thus our algorithm should consider multiple blending regions and different types of blending algorithms. We would like the calculation of the blending quality metric of different types of blending algorithms with any number of blending regions to be done in a unified framework. We next discuss the implementation in detail.

To calculate illumination consistency score when multiple regions are blended, our strategy is to calculate the illumination consistency score for each blending boundary and then take the average. Under the condition of *Type A* blending boundaries appear between different regions. Under the condition of *Type B* blending boundaries appear as the boundary of overlapping regions.

For bleeding degree calculation with multiple regions, we first calculate the bleeding score for each candidate region. A candidate region can be a region to be blended or an overlapping region. The overall bleeding degree is calculated by multiplying the bleeding degrees of all candidate regions. For regions that remain unchanged, the bleeding degree is 1 so they do not affect the overall bleeding degree.

Ghosting degree calculation is done in the same way as bleeding degree calculation. The ghosting degree of each candidate region is calculated first and the overall ghosting degree is calculated by multiplying the ghosting degrees of all candidate regions. Regions that remain unchanged also have the ghosting degree of 1.

IV. Video Blending Quality Metric

We have introduced the blending quality metric for a single frame(image). We next extend our approach to videos. Since video is composed of consecutive frames, temporal coherence has a large impact on the overall video quality and must be taken into consideration.

A. Temporal Coherence

We follow the idea in [35] to evaluate the temporal coherence of the blended videos. Suppose there are n frames of a video and each frame is represented as I^j where j indicates the index of the frame. While each frame has the same resolution we denote the number of pixels of each frame as N. The temporal coherence score Q_c of a video is defined as:

$$Q_c = e^{-\omega \frac{1}{n-1}\frac{1}{N}\sum_j \|warp(I^{j-1})-I^j\|} \tag{8}$$

Here $\|\|$ is the operation calculating the SSD(Sum of Squared Differences), and $warp()$ uses backward flow to advect the previous frame towards the current frame. The correspondence could be obtained by calculating the optical flow [36] between the two frames. In our implementation since the pixel intensity ranges in $[0, 1]$ and each frame is a 3-channel RGB image, the weight ω is set to be $1/3$.

B. Overall Metric

Given the blending quality score of each single frame and the temporal coherence score, the overall blending quality score Q for the video is defined as follows:

$$Q = \left(\frac{1}{n}\sum_j q^j\right)^{\beta} {Q_c}^{\chi} \tag{9}$$

In the above equation q^j denotes the image blending quality score of the jth frame. β and χ are the weights to balance the relative importance of different terms, with the intuition that the average quality score of individual frames should have a larger influence than the temporal coherence score. We empirically set the weight β and χ to be 0.8 and 0.5 in our implementation.

V. Experiment

We evaluate our proposed video blending quality metric using the benchmark in [11]. Our experiments were performed on a PC with an Intel i7-6700 3.4GHz CPU with 32GB memory. We implemented our method in MATLAB.

A. Benchmark

A subjective quality assessment database of video blending was created in our previous study [11]. This database contained 6 different scenes; and each scene yielded the results of 7 blending algorithms, including Feather Blending (FB), Multi-Band Blending (MBB), MVC Blending (MVCB), Convolution Pyramid Blending (CPB), Multi-Spline Blending (MSB), Modified Poisson Blending (MPB), and simple stitching without blending (NoB). Each stimulus was evaluated by 30 subjects. Now, we further detail the processing of subjective data. First, the raw scores were transformed to z-scores in order to account for the differences between subjects in the

use of the scoring scale. By doing so, the raw scores were calibrated towards the same mean and standard deviation:

$$z_{ij} = (s_{ij} - u_i)/\sigma_i \tag{10}$$

where s_{ij} denotes the raw score given by the ith subject to the jth stimulus, u_i is the mean of all scores given by the subject i, and σ_i is the standard deviation. A standard outlier detection and subject exclusion procedure was applied to the z-scores. Scores more than two standard deviations from the mean score for a stimulus were considered to be outliers; an individual subject was an outlier if more than 1/4 of scores submitted were outliers. This caused one subject to be rejected. After removing outliers, the remaining scores were linearly mapped to [0, 100]. Finally, the mean opinion score (MOS) of each blended video was computed as the average of the rescaled z-scores over all subjects:

$$MOS_j = \frac{1}{s}\sum_{i=1}^{s} z'_{ij} \tag{11}$$

where z'_{ij} is the filtered and rescaled z-score, and s is the number of subjects.

B. Performance Evaluation

The proposed metric for video blending quality assessment is validated against the benchmark. We also want to evaluate whether the existing general-purpose video quality metrics can be used to assess the quality of blended videos. Before being able to do so, we need to clarify a significant difference between the settings of blending quality assessment and of video quality assessment. In the context of video quality assessment, the original reference is considered to be of perfect/maximum quality; a video quality metric predicts the quality of a distorted video using the reference (i.e., referred to as full-reference) or without using the reference (i.e., referred to as no-reference). It should be noted that such reference of perfect quality does not exist in the scenario of video blending. The original images before blending either as individuals or as a whole (in the case of simple stitching) are not of perfect quality at all. Therefore, if the existing video quality metrics were to be used for assessing blended videos, they would have to be no-reference metrics and full-reference metrics wouldn't be applicable. We used BIQI [37], BRISQUE [38], FRIQUEE [39], NIQE [40], SSEQ [41] and VIIDEO [42] in our comparative study. For the metrics that were originally designed for image quality assessment, a conventional process had been applied: a frame-level quality is computed and averaged over all frames to give an overall quality of the entire video sequence. [Note, sophisticated weighting assignment of frames is avoided in order to ensure a fair comparison and arguably optimal weighting assignment is difficult because many psychological aspects are involved, which may depend on the content and context of the video sequence being observed.] In selecting metrics for comparison, we also avoided machine-learning based metrics for the following reasons: first, there is no adequate (video blending) data for training a model so any form of comparisons is meaningless; second, our model is not based on machine learning so in fairness we intend to select metrics which use similar approaches for video quality assessment. It should be noted that we already individually fine-tuned the parameters of these metrics towards the highest performance possible for the benchmark. This is done to ensure a fair comparison between the results of different metrics. Each metric was applied to assess the quality of the 42 blended videos in the benchmark, resulting in an objective video quality rating (VQR) per video.

As prescribed by the Video Quality Experts Group (VQEG) [43], we evaluate the performance of metrics by quantifying their ability to predict subjective ratings (i.e., MOS) contained in our benchmark, using Pearson linear correlation (CC), Spearman rank order correlation (SROCC) and Root Mean Square Error (RMSE). Note subjective testing can produce nonlinear quality rating compression at the extremes of the scoring range, e.g., a possible saturation effect at high quality. Therefore, the relationship between the metric outputs and subjective ratings does not need to be linear. It is not the linearity of the relationship that is critical, but the stability of the relationship and a data set's error-variance from the relationship that determine predictive usefulness. As suggested by VQEG [43], to account for any nonlinearity due to the subjective rating process and to facilitate comparison of metrics in a common analysis space, a nonlinear regression is fitted to the [MOS, VQR], using the following logistic function:

$$MOS_p = b_1/(1 + \exp(-b_2 * (VQR - b_3))) \tag{12}$$

where MOS_p indicates the predicted MOS values, and b_1, b_2 and b_3 indicate the parameters for fitting of logistic regression. This nonlinear regression function essentially transforms the set of raw VQR values from a video quality metric to a set of predicted MOS values, which will then be compared with the actual MOS values from the subjective tests. Once the nonlinear transformation was applied, the CC, SROCC and RMSE are computed between the subjectively measured DMOS and the predicted MOS_p.

Fig 6 shows the scatter plots of the MOS versus BIQI, BRISQUE, FRIQUEE, NIQE, SSEQ, VIIDEO and our proposed metric, respectively. The logistic curves are also illustrated. Table I lists the results of the CC, SROCC and RMSE. Fig 6 and Table I demonstrate that our proposed metric outperforms the existing metrics in the prediction of the quality of video blending. In comparison to the best metric (i.e., VIIDEO) in the literature, our metric shows a higher correlation with the subjective ratings, i.e., the increase in the CC and SROCC is 10%, and lower prediction error as measured by RMSE. To verify whether the performance comparison, as shown in Table I, is statistically significant, hypothesis testing is conducted. As suggested in [43], the test is based on the residuals between the MOS and the quality predicted by a metric (i.e., referred to as M-MOS residuals). First, we evaluate the assumption of normality of the M-MOS residuals. The results of the test for normality are summarised as follows: the M-MOS residuals are normal for BIQI, FRIQUEE, NIQE, SSEQ, and proposed; and are not normal for BRISQUE and VIIDEO. When paired M-DMOS residuals are both normally distributed, an independent

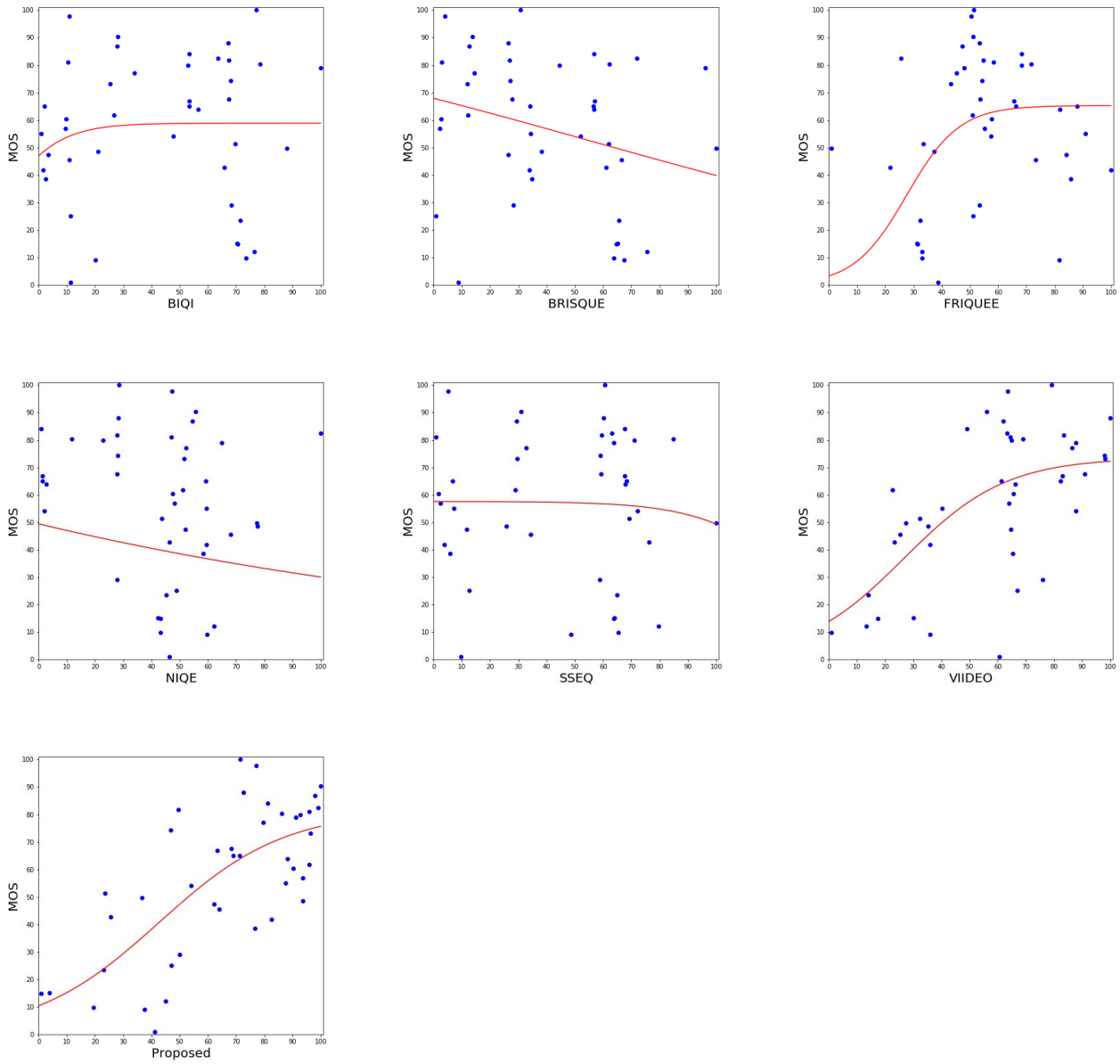


Fig. 6. Scatter plots of MOS versus the BIQI, BRISQUE, FRIQUEE, NIQE, SSEQ, VIIDEO and our proposed metric, respectively. Curves show the regression lines of nonlinear logistic fitting. X-axis shows the predicted score and y-axis shows the observers' MOS. It can be seen from the graphs that except our metric these metrics fail to provide scores that consistently predict the MOS ratings from observers. For example, focusing on the BIQI graph on the upper left corner, for BIQI values around 70, the "ground-truth" MOS values range anywhere from 10 to 90. On the contrary, our method provides much more consistent predictions with ground truth MOS.

TABLE I

PERFORMANCE COMPARISON OF SEVEN QUALITY METRICS FOR VIDEO BLENDING

	BIQI	BRISQUE	FRIQUEE	NIQE	SSEQ	VIIDEO	Proposed
CC	0.13	0.27	0.33	0.21	0.05	0.62	**0.71**
SROCC	0.05	0.31	0.14	0.2	0.02	0.53	**0.64**
RMSE	26.19	25.41	25.43	30.69	26.37	20.64	**18.65**

TABLE II

RESULTS OF STATISTICAL SIGNIFICANCE TESTING BASED ON M-MOS RESIDUALS. 1 MEANS THAT THE DIFFERENCE(AS SHOWN IN TABLE I) IN PERFORMANCE IS STATISTICALLY SIGNIFICANT. 0 MEANS THAT THE DIFFERENCE(AS SHOWN IN TABLE I) IS NOT STATISTICALLY SIGNIFICANT

versus	BIQI	BRISQU	FRIQUEE	NIQE	SSEQ	VIIDEO
Proposed	1(Sig.)	1(Sig.)	1(Sig.)	1(Sig.)	1(Sig.)	1(Sig.)

samples t-test is performed; otherwise, in the case of non-normality, a nonparametric version (i.e., Mann-Whitney U test) analogy to a t-test is conducted. The test results are given in Table II. This means the proposed metric is statistically significantly better than all other six state-of-the-art metrics.

VI. Conclusion

We present a video blending quality assessment metric and the effectiveness of our proposed metric is validated on a subjective quality assessment dataset. This is the first video blending quality assessment metric and it exhibits few limitations. Since there is a lot of computation in intrinsic video decomposition and optical flow estimation, it takes time to calculate the final blending quality score. The future work will focus on the reduction of the metric's computational complexity. In addition, we will investigate improving the metric's performance by considering more perceptually relevant features.

References

[1] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Trans. Graph.*, vol. 2, no. 4, pp. 217–236, 1983.

[2] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 2003.

[3] A. Agarwala, "Efficient gradient-domain compositing using quadtrees," *ACM Trans. Graph.*, vol. 26, no. 3, p. 94, 2007.

[4] Z. Farbman, G. Hoffer, Y. Lipman, D. Cohen-Or, and D. Lischinski, "Coordinates for instant image cloning," *ACM Trans. Graph.*, vol. 28, no. 3, p. 67, 2009.

[5] M. Tanaka, R. Kamio, and M. Okutomi, "Seamless image cloning by a closed form solution of a modified Poisson problem," in *Proc. SIGGRAPH Asia Posters (SA)*. New York, NY, USA: ACM, 2012, pp. 15:1–15:1, doi: 10.1145/2407156.2407173.

[6] R. Szeliski, M. Uyttendaele, and D. Steedly, "Fast Poisson blending using multi-splines," in *Proc. Int. Conf. Comput. Photogr. (ICCP)*, Apr. 2011, pp. 1–8.

[7] Z. Farbman, R. Fattal, and D. Lischinski, "Convolution pyramids," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 175:1–175:8, Dec. 2011, doi: 10.1145/2070781.2024209.

[8] M. Wang, Z. Zhu, S. Zhang, R. Martin, and S. -M. Hu, "Avoiding bleeding in image blending," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2139–2143.

[9] T. Chen, J.-Y. Zhu, A. Shamir, and S.-M. Hu, "Motion-aware gradient domain video composition," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2532–2544, Jul. 2013.

[10] Z. Wang, X. Chen, and D. Zou, "Copy and paste: Temporally consistent stereoscopic video blending," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3053–3065, Oct. 2018.

[11] Z. Zhu *et al.*, "A comparative study of algorithms for realtime panoramic video blending," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2952–2965, Jun. 2018.

[12] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Rec. BT.500-13., Jan. 2012.

[13] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1126–1165, 2nd Quart., 2015.

[14] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "End-to-end QoS for video delivery over wireless Internet," *Proc. IEEE*, vol. 93, no. 1, pp. 123–134, Jan. 2005.

[15] B. Vandalore, W.-C. Feng, R. Jain, and S. Fahmy, "A survey of application layer techniques for adaptive streaming of multimedia," *Real-Time Imag.*, vol. 7, no. 3, pp. 221–235, Jun. 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077201401902244

[16] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[17] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 68–78, Jan. 2007.

[18] A. M. Adas, "Using adaptive linear prediction to support real-time VBR video under RCBR network service model," *IEEE/ACM Trans. Netw.*, vol. 6, no. 5, pp. 635–644, Oct. 1998, doi: 10.1109/90.731200.

[19] M. Wu, R. A. Joyce, H.-S. Wong, L. Guan, and S.-Y. Kung, "Dynamic resource allocation via video content and short-term traffic statistics," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 186–199, Jun. 2001.

[20] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[21] M. A. Saad and A. C. Bovik, "Blind quality assessment of videos using a model of natural scene statistics and motion coherency," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 332–336.

[22] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Three dimensional scalable video adaptation via user-end perceptual quality assessment," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 719–727, Sep. 2008.

[23] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for Internet video," in *Proc. ACM SIGCOMM Conf. SIGCOMM (SIGCOMM)*. New York, NY, USA: ACM, 2013, pp. 339–350, doi: 10.1145/2486001.2486025.

[24] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "A quest for an Internet video quality-of-experience metric," in *Proc. 11th ACM Workshop Hot Topics Netw. (HotNets-XI)*. New York, NY, USA: ACM, 2012, pp. 97–102, doi: 10.1145/2390231.2390248.

[25] F. Dobrian *et al.*, "Understanding the impact of video quality on user engagement," in *Proc. ACM SIGCOMM Conf. (SIGCOMM)*. New York, NY, USA: ACM, 2011, pp. 362–373, doi: 10.1145/2018436.2018478.

[26] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 2001–2014, Dec. 2013.

[27] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[28] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[29] 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, Apr. 2004.

[30] J. Kopf, M. F. Cohen, and R. Szeliski, "First-person hyper-lapse videos," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 78-1–78-10, 2014, doi: 10.1145/2601097.2601195.

[31] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross, "Practical temporal consistency for image-based graphics applications," *ACM Trans. Graph*, vol. 31, no. 4, p. 34, 2012, doi: 10.1145/2185520.2185530.

[32] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 159.

[33] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt, "Live intrinsic video," *ACM Trans. Graph.*, vol. 35, no. 4, Jul. 2016, Art. no. 109. [Online]. Available: http://gvv.mpi-inf.mpg.de/projects/LiveIntrinsicVideo/

[34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[35] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister, "Blind video temporal consistency," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 196:1–196:9, Nov. 2015, doi: 10.1145/2816795.2818107.

[36] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011, doi: 10.1007/s11263-010-0390-2.

[37] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[38] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[39] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, 2016.

[40] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'Completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[41] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process., Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.

[42] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.
[43] Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II (FR-TV2)," Video Quality Experts Group (VQEG), Tech. Rep. 32, 2003. Accessed: Nov. 26, 2019. [Online]. Available: ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Approved_Final_Reports/VQEGII_Final_Report.pdf

Zhe Zhu received the bachelor's degree from Wuhan University in 2011 and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, in 2017. He is currently a Senior Research Associate with Duke University. His research interests include computer vision and computer graphics.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is serving for the IEEE MMTC, as the Chair of the Interest Group on Quality of Experience for Multimedia Communications. He is currently an Associate Editor of the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and the IEEE TRANSACTIONS ON MULTIMEDIA.

Jiaming Lu is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include computer vision and fluid simulation.

Shi-Min Hu received the Ph.D. degree from Zhejiang University in 1996. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing. He has published over 100 articles in journals and refereed conference. His research interests include digital geometry processing, video processing, rendering, computer animation, and computer-aided geometric design. He is a Senior Member of the ACM. He is the Editor-in-Chief of *Computational Visual Media*, and on editorial board of several journals, including the IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, *Computer Aided Design*, and *Computer and Graphics*.