2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)

Visual Quality Assessment for Motion Compensated Frame Interpolation

Hui Men[1], Hanhe Lin[1], Vlad Hosu[1], Daniel Maurer[2], Andrés Bruhn[2], Dietmar Saupe[1]

[1]Department of Computer and Information Science, University of Konstanz, Germany
[2]Institute for Visualization and Interactive Systems, University of Stuttgart, Germany
Email: {hui.3.men, hanhe.lin, dietmar.saupe}@uni-konstanz.de, {bruhn, maurer}@vis.uni-stuttgart.de

***Abstract*—Current benchmarks for optical flow algorithms evaluate the estimation quality by comparing their predicted flow field with the ground truth, and additionally may compare interpolated frames, based on these predictions, with the correct frames from the actual image sequences. For the latter comparisons, objective measures such as mean square errors are applied. However, for applications like image interpolation, the expected user's quality of experience cannot be fully deduced from such simple quality measures. Therefore, we conducted a subjective quality assessment study by crowdsourcing for the interpolated images provided in one of the optical flow benchmarks, the Middlebury benchmark. We used paired comparisons with forced choice and reconstructed absolute quality scale values according to Thurstone's model using the classical least squares method. The results give rise to a re-ranking of 141 participating algorithms w.r.t. visual quality of interpolated frames mostly based on optical flow estimation. Our re-ranking result shows the necessity of visual quality assessment as another evaluation metric for optical flow and frame interpolation benchmarks.**

***Index Terms*—visual quality assessment, optical flow, frame interpolation**

I. Introduction

As one of the basic video processing techniques, frame interpolation, namely computing interpolated in-between images, is a necessary step in numerous applications such as temporal up-sampling for generating slow motion videos and frame rate conversion between broadcast standards [1]. One of the main concepts in frame interpolation is motion compensation. In this context, required frames are obtained by interpolating the image content along the path of motion. Thereby, the apparent motion in terms of the so-called optical flow can be derived in various ways. Typical approaches for this task include block matching techniques [2], frequency-based approaches [1], variational methods [3] or convolutional neural networks [4].

Since the quality of the results heavily depends on the underlying optical flow algorithm, the evaluation of the motion-compensated interpolation results is a critical issue. However, currently, there is only one optical flow benchmark that offers the assessment of interpolated frames: the Middlebury benchmark [5]. Regarding the quality of the motion estimation, it considers angular and endpoint errors between the estimated flow and the ground truth flow. More importantly, it also offers a direct evaluation of the corresponding motion-compensated interpolation results between frame pairs, which is based on the root mean square error (RMSE) and gradient normalized RMSE between the interpolated image and the ground truth image.

In general, the direct evaluation of the frame interpolation results is useful, since the accuracy of the motion estimation is not always directly related to the quality of the motion-compensated frame interpolation. For example, motion errors in low-textured regions are less noticeable in the interpolation result than motion errors in highly-textured regions. However, specifically designed error measures such as the gradient normalized RMSE even revert this relation and penalize interpolation errors in high-textured regions less severely which adapts the error measure to the shortcomings of motion-based frame interpolation techniques instead of trying to assess the frame interpolation quality adequately. Moreover, it is well known that even the standard mean square error can be misleading and may not reliably reflect image quality as perceived by the human visual system (HVS) [6]. This fact also becomes obvious from the Middlebury web-page, where some of the interpolated images have the same RMSE but exhibit obvious differences in image quality (see Fig.1). Evidently, there is a clear need to improve the assessment of motion-compensated interpolation results. Therefore, we propose to change the quality assessment in such a way that the evaluation of the results takes perceived visual quality assessment into consideration.

Regarding visual quality assessment methods, we take full-reference image quality assessment (FR-IQA) into consideration, since ground truth in-between images are available in the Middlebury benchmark. There are several FR-IQA methods that consider the HVS, such as SSIM [6], MS-SSIM [7], FSIM [8], VSI [9]. These methods were designed to estimate image quality degradation due to common artifacts, namely the ones caused by processing such as data compression or by losses in data transmission. However, the artifacts induced by optical flow algorithms lead to interpolated images with different specific distortions.

In this contribution, we show (in Table I) that five of the most popular objective FR-IQA methods have rather low correlations with the evaluations made by human observers, regardless of whether the methods are based on the HVS or

978-1-5386-8212-8/19/$31.00 ©2019 IEEE

TABLE I
SROCC BETWEEN FR-IQA AND GROUND TRUTH (BY SUBJECTIVE STUDY)

FR-IQA	Average	Mequon	Schefflera	Urban	Teddy	Backyard	Basketball	Dumptruck	Evergreen
RMSE	0.598	0.766	0.557	0.854	0.667	0.152	0.534	0.756	0.494
SSIM	0.592	0.747	0.552	0.718	0.566	0.255	0.693	0.788	0.416
MS-SSIM	0.602	0.733	0.491	0.741	0.653	0.260	0.698	0.795	0.444
FSIM	0.599	0.739	0.573	0.783	0.631	0.244	0.553	0.778	0.488
VSI	0.610	0.705	0.558	0.803	0.657	0.204	0.615	0.783	0.555

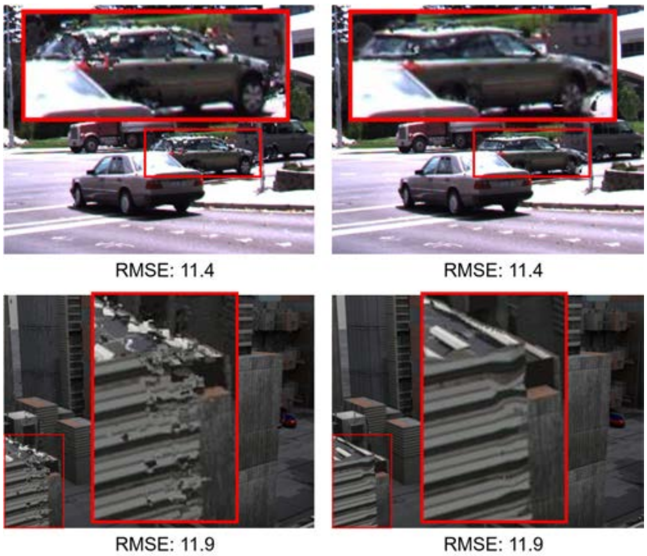


Fig. 1. Two interpolated frames, each with two different methods. RMSE values in each pair are equal, but the visual quality differs in each pair, in particular in the zoomed regions.

just on pixel-wise errors such as RMSE. More specifically, the method VSI, one of the best FR-IQA methods, when trained and tested on the LIVE database [10], yielded a Spearman rank-order correlation coefficient (SROCC) of 0.952 w.r.t. ground truth consisting of MOS from a controlled lab study for the LIVE database. But when applied for the interpolated images by optical flow algorithms, VSI gave only an SROCC of 0.610. This demonstrates that current FR-IQA methods are not able to cope with the specific distortions that arise in interpolated frames produced by optical flow algorithms. Therefore, a new FR-IQA method specifically designed for such images is needed. However, before the research in such FR-IQA methods can proceed, ground truth data, namely subjective quality scores of such images need to be collected, a first set of which is provided by our study. Hence, our contribution is twofold. While revealing the poor performance of existing FR-IQA methods for rating the quality of motion compensated frame interpolation, we also provide subjective quality scores that may serve as a basis for the development of new FR-IQA methods.

Regarding the subjective quality evaluation, lab-studies are well established due to their reliability. In particular, the experimental environment and the evaluation process can be controlled. However, it is time consuming and costly, which severely limits the number of images to be assessed. In contrast, crowdsourcing studies can be less expensive. Moreover, the reliability of crowdsourcing has been proven to be acceptable, if the setup is appropriate and the crowd workers are trained [11]. Therefore, in this paper, we performed subjective quality assessment of the images interpolated by optical flow algorithms with the help of crowdsourcing.

To this end, we implemented paired comparisons of the interpolated images given by optical flow algorithms in the Middlebury interpolation benchmark with the help of crowdsouring and re-ranked them accordingly. Comparing the old ranking according to RMSE in the Middlebury benchmark and the re-ranking according to our subjective study then allows us to judge the suitability of existing quality metrics. The outcome of our study is clear. It demonstrates that current FR-IQA methods are not suitable for assessing the perceived visual quality of interpolated frames that have been created by using optical flow algorithms.

II. Related Work

As the recent literature on frame interpolation shows, there is mainly one benchmark that is used for evaluating the performance of frame interpolation: the Middlebury benchmark. Originally designed for the evaluation of optical flow algorithms, this benchmark also offers an evaluation of motion-based frame interpolation results based on the calculated optical flow. To this end, it compares the interpolated frames with the ground truth in-between images that have been obtained by rendering or recording the original image sequence with a higher frame rate. Hence, despite of its original focus on evaluating optical flow methods, in the last few years this benchmark has become increasingly popular for the evaluation and comparison of frame interpolation algorithms [3], [1], [12].

Apart from the Middlebury benchmark, some interpolation algorithms like [13], [14] used the UCF 101 dataset [15] for training and testing. Others like [3], [16], [17] used the videos from [18], [19].

Regarding the assessment of the interpolation quality, both the Middlebury benchmark and the other data sets rely on standard metrics, such as MSE, PSNR, or SSIM to measure the differences between the interpolated result and the ground truth in-between image. However, to the best of our knowledge, there have been no attempts so far to analyze how useful these metrics actually are to measure the quality of motion-compensated frame interpolation.

III. Subjective Study using Paired Comparisons

A. Subjective Study

Absolute Category Rating (ACR) is a type of subjective testing where the test items are presented one at a time and are rated independently on one of five possible ordinal scales, i.e., Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5 [20].

ACR is easy and fast to implement, however, it has several drawbacks [21]. Participants may be confused when the categories of the rating scale have not been explained sufficiently well. They may also have different interpretations of the ACR scale, in particular in crowdsourcing experiments because of the wide range of cultural backgrounds and perceptual experiences of the crowd workers from around the world. Moreover, the perceptual distances between two consecutive scale values, e.g., between 1 and 2, should ideally be the same. However, in practice this can hardly be achieved [22]. Also it is not easy to detect when a participant intentionally or carelessly gives false ratings.

Alternatively, *paired comparisons* (PC) can solve some of the problems of ACR. In a PC test, items to be evaluated are presented as pairs. In a forced choice setting, one of the items must be chosen as the preferred one. The main advantage of this strategy is that it is highly discriminatory, which is very relevant when test items have nearly the same quality.

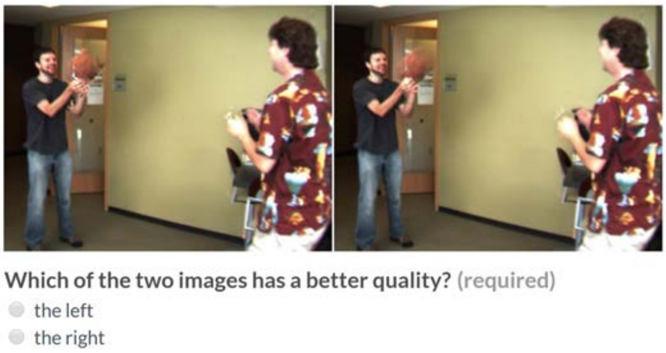
However, when implemented naively, comparing N items would require $\binom{N}{2}$ comparisons, too many to be practical, when N is on the order of 100, for example. In our case, for each of the 8 sequences, we would have to compare $N = 141$ images, giving a total of 78,960 pairs.

A practical solution to this problem is to resort to the concept of randomly paired comparisons that is based on randomly choosing a fraction of all possible paired comparisons. This strategy is not only more efficient, it also has been proven to be as reliable as full comparisons [23]. After obtaining results from the (randomly) paired comparisons, subjective scores have to be reconstructed. This can be done based on Thurstone's model [24], [25] or the Bradley-Terry model [26].

B. Thurstone's Model

Thurstone's model provides the basis for a psychometric method for assigning scale values to options on a one-dimensional continuum from paired comparisons data. It assumes that an option's quality is a Gaussian random variable, thereby accommodating differing opinions about the quality of an option. Then each option's latent quality score is revealed by the mean of the corresponding Gaussian.

The result of a paired comparison experiment is a square count matrix C denoting the number of times that each option was preferred over any other option. More specifically, for n comparisons of option A_i with option A_j, $C_{i,j}$ gives the number of times A_i was preferred over A_j. Similarly, $C_{j,i}$ in the count matrix denotes the number of times that A_j was preferred over A_i, and we have $C_{i,j} + C_{j,i} = n$.



Which of the two images has a better quality? (required)

○ the left

○ the right

Fig. 2. Crowdsourcing interface.

According to Thurstone's Case V, subjective qualities about two options A and B are modelled as uncorrelated Gaussian random variables A and B with mean opinions μ_A, μ_B and variances ${\sigma_A}^2, {\sigma_B}^2$, respectively. When individuals decide which of the two options is better, they draw realizations from their quality distributions, and then choose the option with higher quality. More specifically, they choose option A over option B if their sample from the random variable $A-B$ (with mean $\mu_{AB} = \mu_A - \mu_B$ and variance ${\sigma_{AB}}^2 = {\sigma_A}^2 + {\sigma_B}^2$) is greater than 0. Therefore, the probability of a subject to prefer option A over B is:

$$P(A > B) = P(A - B > 0) = \Phi\left(\frac{\mu_{AB}}{\sigma_{AB}}\right), \qquad (1)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF).

Thurstone proposed to estimate $P(A > B)$ by the empirical proportion of people preferring A over B, which can be derived from the count matrix C as:

$$P(A > B) \approx \frac{C_{A,B}}{C_{A,B} + C_{B,A}}.$$

The estimated quality difference $\hat{\mu}_{AB}$ can be derived from inverting Eq. 1, giving:

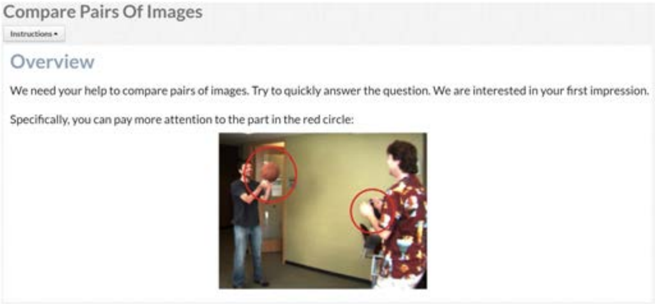
$$\hat{\mu}_{AB} = \sigma_{AB}\Phi^{-1}\left(\frac{C_{A,B}}{C_{A,B} + C_{B,A}}\right)$$

known as Thurstone's Law of Comparative Judgment, where $\Phi(\cdot)^{-1}$ is the inverse standard normal CDF, or z-score. Least-squares fitting or maximum likelihood estimation (MLE) can be then applied to estimate the scale values μ_A for all involved stimuli A. For more details we refer to [27].

C. Study Design

In order to re-rank the methods in the Middlebury benchmark, we implemented paired comparisons based on Thurstone's model with least-squares estimation to obtain subjective judgments of the image qualities. In the benchmark, there are 8 sets of 141 interpolated images each, most of which generated by optical flow methods.[1] Therefore, in our experiment, for each set of 141 interpolated images, we generated a random sparse graph with degree of 6 (i.e., each image was to be randomly compared to 6 other images), which resulted in 423 pairs of images. We ran the experiment using the Figure Eight [28] platform. In our crowdsourcing interface as shown in Fig. 2, crowd workers were asked to identify and select the image with better quality for each image pair (forced binary choice).

[1] When we ran the experiments in June 2018, there were altogether 141 methods in the Middlebury benchmark, which now includes a number of additional, new methods.



Compare Pairs Of Images

Overview

We need your help to compare pairs of images. Try to quickly answer the question. We are interested in your first impression.

Specifically, you can pay more attention to the part in the red circle:

Fig. 3. Instructions of the crowdsourcing experiment.

D. *Quality Assurance and Quality Control*

Before the actual paired comparison, there was a training session, in which workers were instructed how to compare the quality of images. Since the visual differences between some images are not that obvious, in the instructions as shown in Fig. 3, we highlighted those parts in the images that are more degraded and hence showed more differences.

In our experiment, we had eight separate jobs, each containing the 423 image pairs of the same series. In each job, there were 22 tasks, each consists of one page. In order to make sure that crowd workers' performance was not effected by exhaustion, we showed 20 pairs of images per page, and payments were initiated for each completed page. For each pair of images to be compared, we collected 30 votes from the crowd workers.

In order to assure the reliability of the crowd workers, the unreliable ones need to be detected and disallowed to continue. This was done by requiring crowd workers to answer test questions. For the test questions we chose image pairs with the ground truth in-between image as one of the images, and the other image of bad quality. Then the expected, correct answer, was obviously given by choosing the ground truth image as the one with a better quality.

Before crowd workers were allowed to start a job, they had to pass a quiz which was composed entirely of test questions. This ensures that only capable crowd workers that proved to be able to work on the subject matter of the job, would be able to enter the job. Crowd workers that failed the quiz were permanently disqualified from working on the job. After passing the quiz, crowd workers were admitted to start the real job. During the job, they had to answer further hidden test questions. Once a crowd worker failed more than 30% of the hidden test questions, he or she was disqualified and removed from the job. Only crowd workers who passed the quiz and showed an accuracy above 70% on the hidden test questions were regarded as reliable.

IV. Results

A. *Statistics for the Crowdsourcing Experiment*

In our experiment, there were eight jobs, each one comparing 141 interpolated images pairwise (423 pairs per job). The average run time was 29 hours per job. In total, 3189 crowd workers participated in our experiment, some of them took multiple jobs. Before the real experiment, 54% of them did not pass the quiz thus were not allowed to contribute to the job. During the real job, 14% of them failed more than 30% of the test questions, and thus were disregarded. In the end, 1033 crowd workers were accepted as trusted workers. Among the trusted workers, 79% had an accuracy of 90%–100% (where 100% means they passed all test questions), 10% of them had an accuracy of 80%–90% and 10% of them had an accuracy of 70%–80%.

B. *Re-ranking Results*

Given the results of the paired comparisons, we reconstructed the corresponding quality values based on Thurstone's model using the code provided by [27]. In order to make the results of the 8 separate jobs comparable, we added two fictitious images as anchors. One of them represents the worst quality among all the images, and the other one is like the ground truth image, with a quality that is better than that of all the other images. After reconstruction of the scale values for the 141 + 2 images in each series, we linearly rescaled the quality scores such that the quality of the imaginary worst quality image became 0, and that of the ground truth image became 1. In this way, we rescaled the reconstructed scores to the interval $[0, 1]$.[2] Each set was scaled and ranked separately. Then the average quality of a method was obtained by taking the mean of the (scaled) quality values of the 8 series, which resulted in an overall rank.

The best three methods ranked by the subjective study (i.e., SuperSlomo [30], CtxSyn [12], and DeepFlow2 [31]), ranked 1st, 5th, and 9th in the Middlebury benchmark, respectively. Overall, 36 methods showed rank differences up to 5. However, 30 methods gave differences of more than 30 between their re-rankings and the rankings in the Middlebury benchmark.[3]

Fig. 4 shows the subjective qualities of the methods ranked the highest 20 and the lowest 20. The quality scores of the best two methods, SuperSlomo and CtxSyn, are better than those of the rest by a large margin.

As an overall analysis, Table II shows the bootstrapped (after 1000 iterations) SROCC correlation values accompanied with confidence intervals (CI, 95%) between the ranking in the Middlebury benchmark (i.e., ranking according to RMSE) and the re-ranking according to our subjective study. Note that the CI of SROCC was computed by transforming the rank correlation score into an approximate z-score using the Fisher transform [32]. In a nutshell, a CI of probability p is given by $\tanh(\arctan r \pm \Phi^{-1}(p)/\sqrt{n-3})$, where r denotes the estimated SROCC, n is the sampling size, and Φ^{-1} is the inverse of the standard normal CDF. In order to visualize the result, we computed the disagreement level as $1 -$ SROCC as shown in Fig. 5.

[2] All reconstructed quality values, accompanied by their corresponding rankings, are shown in tables in [29]. The differences between the re-ranking (ranked according to subjective study) and their corresponding ranking in the Middlebury benchmark (ranked according to RMSE) are also available in [29].

[3] Some methods are specifically tailored for frame interpolation (e.g., SuperSlomo) and some are only tailored for optical flow estimation (e.g., DeepFlow2).

TABLE II
BOOTSTRAPPED CORRELATIONS BETWEEN RMSE AND GROUND TRUTH (BY SUBJECTIVE STUDY) AFTER 1000 ITERATIONS.

RMSE	Average	Mequon	Schefflera	Urban	Teddy	Backyard	Basketball	Dumptruck	Evergreen
SROCC	0.598	0.766	0.557	0.854	0.667	0.152	0.534	0.756	0.494
CI (95%)	[0.507,0.674]	[0.699,0.816]	[0.454,0.647]	[0.813,0.888]	[0.581,0.737]	[0.015,0.283]	[0.419,0.618]	[0.695,0.813]	[0.382,0.593]

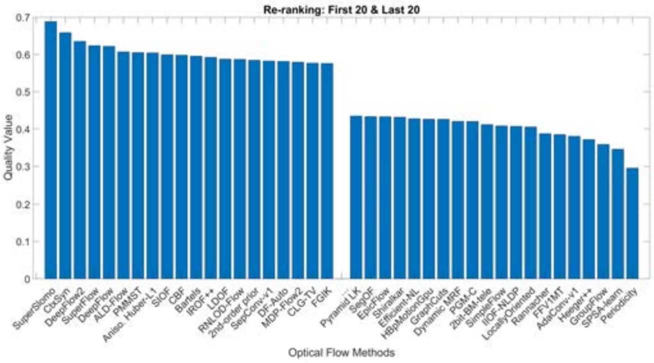


Fig. 4. The methods ranking in the top 20 and the bottom 20 by the subjective study. The x-axis shows the names of the methods. The y-axis denotes the value of the average subjective scores.

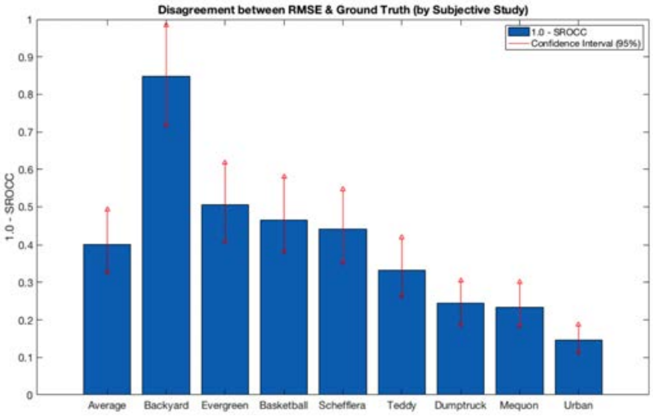


Fig. 5. Disagreement level with 95% confidence interval.

Fig. 6 shows the scatter plots of the RMSE values compared to subjective scores of each optical flow method. Those eight plots are displayed in descending order of SROCC between RMSE and subjective scores. It can be seen that starting from *Urban* with the highest SROCC (0.854) down to the lowest SROCC (0.152) given by *Backyard*, the scattered plots become more sparse, as to be expected from their decreasing correlation.

V. LIMITATIONS

The interpolated images available from the Middlebury benchmark are compressed and slightly downsized from the original images. The original interpolated images could not be made available by the maintainers of the Middlebury benchmark. The differences between their resolutions are shown in Table III. Since we used the down-scaled, compressed public version of the images for the crowdsourcing study, our results may be biased to a small extent.

Another limitation of the experiment is the difficulty of the subjective study. The quality differences between some images are quite hard to distinguish. Therefore, in the instructions of the crowdsourcing experiment, we highlighted the main degraded parts according to a simple visual inspection to help the crowd workers to focus on the critical parts of the images. We believe, that this can be further improved in future studies, e.g., by providing zoomed image portions that contain the most noticeable artifacts.

VI. CONCLUSION AND FUTURE WORK

We have adopted visual quality assessment to the Middlebury benchmark for frame interpolation based mostly on optical flow methods. Our study confirms that only using RMSE as an evaluation metric for image interpolation performance is not representative of visual quality. Also current FR-IQA methods do not provide satisfying results on those interpolated images. This is due to the fact that such images, especially the ones generated by optical flow algorithms have specific distortions that are quite different from artifacts commonly addressed by conventional IQA methods.

Therefore, we plan to develop a domain specific FR-IQA for frame interpolation based on optical flow estimation. Since both the ground truth frame and the corresponding flow are available, we can extract features from both sources to train the corresponding quality assessment model. As a result, we obtain a FR-IQA method with side information given by the optical flow. Evidently, such a method could serve as a visual quality metric in future optical flow benchmarks. Moreover, we plan to apply VQA methods on the videos generated by frame interpolation as a further study. This, in turn, will allow us to consider temporal aspects in the quality assessment.

ACKNOWLEDGMENT

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 251654672 – TRR 161 (Project A05 and B04).

REFERENCES

[1] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, "Phase-based frame interpolation for video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1410–1418.

[2] T. Ha, S. Lee, and J. Kim, "Motion compensated frame interpolation by new block-based motion estimation algorithm," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 2, pp. 752–759, 2004.

[3] L. L. Rakêt, L. Roholm, A. Bruhn, and J. Weickert, "Motion compensated frame interpolation with a symmetric optical flow constraint," in *International Symposium on Visual Computing*. Springer, 2012, pp. 447–457.

[4] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *arXiv e-prints*, p. arXiv:1810.08768, Oct. 2018.

[5] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

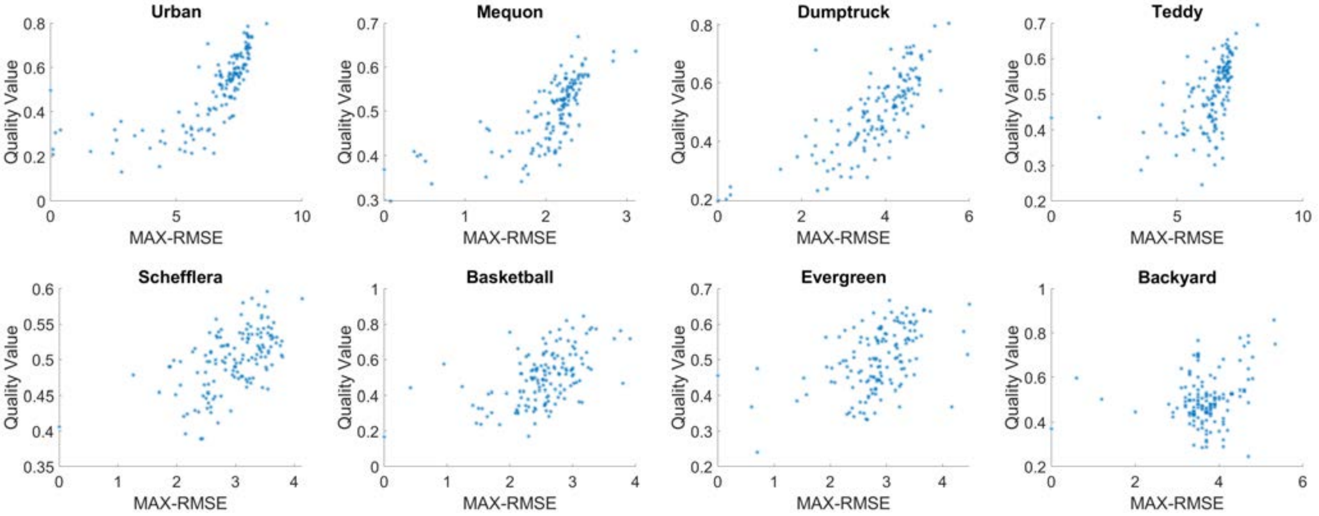


Fig. 6. Scatter plots of RMSE and subjective scores. To show positive correlations, we used the difference between maximum and individual RMSE as the x-axis values.

TABLE III
RESOLUTIONS OF THE ORIGINAL IMAGES AND THE AVAILABLE ONES USED FOR SUBJECTIVE STUDY

	Mequon	Schefflera	Urban	Teddy	Backyard	Basketball	Dumptruck	Evergreen
Original	584×388	584×388	640×480	420×360	640×480	640×480	640×480	640×480
Available	467×310	467×310	512×384	336×288	512×384	512×384	512×384	512×384

[7] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proceeding of the IEEE Asilomar Conference on Signals (ACSSC), Systems & Computers*, 2003, pp. 1398–1402.

[8] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: a feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[9] L. Zhang, Y. Shen, and H. Li, “VSI: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[10] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[11] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, “Crowd workers proven useful: A comparative study of subjective video quality assessment,” in *the 8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[12] S. Niklaus and F. Liu, “Context-aware synthesis for video frame interpolation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1701–1710.

[13] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, “Video frame synthesis using deep voxel flow,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4463–4471.

[14] Z. Gong and Z. Yang, “Video frame interpolation and extrapolation,” Stanford University, Tech. Rep., 2017. [Online]. Available: http://cs231n.stanford.edu/reports/2017/pdfs/714.pdf

[15] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” *arXiv e-prints*, p. arXiv:1212.0402, Dec. 2012.

[16] J. Zhai, K. Yu, J. Li, and S. Li, “A low complexity motion compensated frame interpolation method,” in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, 2005, pp. 4927–4930.

[17] R. C. Ghutke, C. Naveen, and V. R. Satpute, “A novel approach for video frame interpolation using cubic motion compensation technique,” *International Journal of Applied Engineering Research*, vol. 11, no. 10, pp. 7139–7146, 2016.

[18] http://see.xidian.edu.cn/vipsl/dataset.html.

[19] http://see.xidian.edu.cn/vipsl/database_Video.html.

[20] P. ITU-T RECOMMENDATION, “Subjective video quality assessment methods for multimedia applications,” Tech. Rep., 1999. [Online]. Available: https://www.itu.int/rec/T-REC-P.910-200804-I

[21] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, “A crowdsourceable QoE evaluation framework for multimedia content,” in *Proceedings of the ACM International Conference on Multimedia (MM)*, 2009, pp. 491–500.

[22] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, “QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS,” *Quality and User Experience*, vol. 1, no. 2, pp. 1–23, 2016.

[23] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, “Hodgerank on random graphs for subjective video quality assessment,” *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, 2012.

[24] L. L. Thurstone, “A law of comparative judgment.” *Psychological Review*, vol. 34, no. 4, p. 273, 1927.

[25] R. D. Luce, “Thurstone and sensory scaling: Then and now.” *US: American Psychological Association*, vol. 101(2), pp. 271–277, 1994.

[26] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. the method of paired comparisons,” *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[27] K. Tsukida and M. R. Gupta, “How to analyze paired comparison data,” University of Washington, Tech. Rep., 2011. [Online]. Available: https://www.itu.int/rec/T-REC-P.910-200804-I

[28] https://www.figure-eight.com/.

[29] H. Men, H. Lin, V. Hosu, D. Maurer, A. Bruhn, and D. Saupe, “Technical Report on Visual Quality Assessment for Frame Interpolation,” *arXiv e-prints*, p. arXiv:1901.05362, Jan 2019.

[30] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, “Super slomo: High quality estimation of multiple intermediate frames for video interpolation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9000–9008.

[31] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, “Deepmatching: Hierarchical deformable dense matching,” *International Journal of Computer Vision*, vol. 120, no. 3, pp. 300–323, 2016.

[32] J. Ruscio, “Constructing confidence intervals for spearmans rank correlation with ordinal data: a simulation study comparing analytic and bootstrap methods,” *Journal of Modern Applied Statistical Methods*, vol. 7, no. 2, p. 7, 2008.