Hacking VMAF with Video Color and Contrast Distortion

A. Zvezdakova[1], S. Zvezdakov[1], D. Kulikov[1,2], D. Vatolin[1]
azvezdakova@graphics.cs.msu.ru|szvezdakov@graphics.cs.msu.ru|dkulikov@graphics.cs.msu.ru|
dmitriy@graphics.cs.msu.ru
[1]Lomonosov Moscow State University, Moscow, Russia;
[2]Dubna State University, Dubna, Russia

Video quality measurement takes an important role in many applications. Full-reference quality metrics which are usually used in video codecs comparisons are expected to reflect any changes in videos. In this article, we consider different color corrections of compressed videos which increase the values of full-reference metric VMAF and almost don't decrease other widely-used metric SSIM. The proposed video contrast enhancement approach shows the metric in-applicability in some cases for video codecs comparisons, as it may be used for cheating in the comparisons via tuning to improve this metric values.

***Keywords**: video quality, quality measuring, video-codec comparison, quality tuning, reference metrics, color correction.*

1. Introduction

At the moment, video content takes a significant part of worldwide network traffic and its share is expected to grow up to 71% by 2021 [1]. Therefore, the quality of encoded videos is becoming increasingly important, which leads to growing of an interest in the area of new video quality assessment methods development. As new video codec standards appear, the existing standards are being improved. In order to choose one or another video encoding solution, it is necessary to have appropriate tools for video quality assessment. Since the best method of video quality assessment is a subjective evaluation, which is quite expensive in terms of time and cost of its implementation, all other objective methods are improving in an attempt to approach the ground truth-solution (subjective evaluation).

Methods for evaluating encoded videos quality can be divided into 3 categories [9]: full-reference, reduced-reference and no-reference. Full-reference metrics are the most common, as their results are easily interpreted — usually as an assessment of the degree of distortions in the video and their visibility to the observer. The only drawback of this approach compared to the others is the need to have the original video for comparison with the encoded, which is often not available.

One of the widely-used full-reference metrics which is gaining popularity in the area of video quality assessment is Video Multimethod Fusion Approach (VMAF)[5], announced by Netflix. It is an open-source learning-based solution. Its main idea is to combine multiple elementary video quality features, such as Visual Information Fidelity (VIF)[12], Detail Loss Metric (DLM)[11] and temporal information (TI) – the difference between two neighboring frames, and then to train support vector machine (SVM) regres-sion on subjective data. The resulting regressor is used for estimating per-frame quality scores on new videos. The scheme [7] of this metric is shown in Fig. 1.

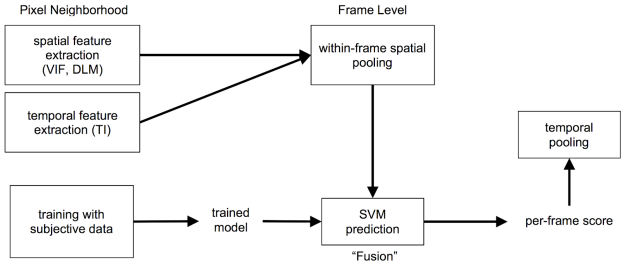


Fig. 1. The scheme of VMAF algorithm.

Despite increasing attention to this metric, many video quality analysis projects, such as Moscow State University's (MSU) Annual Video Codec Comparison [2], still use other common metrics developed many years ago, such as structural similarity (SSIM) and even peak signal-to-noise ratio (PSNR), which are based only on difference characteristics of two images. At the same time, many readers of the reports of these comparisons send requests to use new metrics of VMAF type. The main obstacle for the full tran-sition to the use of VMAF metrics is non-versatility of this metric and not fully adequate results on some types of videos [4].

The main goal of our investigation was to prove the no-universality of the current version of VMAF algorithm. In this paper, we describe video color and contrast transformations which increase VMAF-score with keeping SSIM score the same or better. The possibility to improve full-reference metric score after adding any transformations to distorted im-age means that the metric can be cheated in some cases. Such transformations may allow competitors, for example, to cheat in video codecs comparisons, if they "tune" their codecs for increasing VMAF qual-ity scores. Types of video distortions that we were looking for change the visual quality of the video, which should lead to a decrease in the value of any full-reference metric. The fact that they lead to an increase in the value of VMAF, is a significant obstacle to using VMAF for all types of videos as the main quality indicator and proves the need of modification of the original VMAF algorithm.

2. Study Method

During testing of VMAF algorithm for video codecs comparisons, we noticed that it reacts on contrast distortions, so we chose color and contrast adjustments as basic types of the searched video transformations. Two famous and common approaches for color adjustments were tested to find the best strat-egy for VMAF scores increasing. Two cases of transformations application to the video were tested: applying transformation before and after video encoding. In general, there was no significant difference between these options, because the compression step can be omitted for increasing VMAF with color enhance-ment. Therefore, further we will describe only the first case with adjustment before compression, and we leave the compression step because in our work VMAF tuning is considered in case of video-codec compar-isons.

We chose 4 videos which represent different spatial and temporal complexity [8], content and contrast to test transformations which may influence VMAF scores. All videos have FullHD resolution and high bit rate. *Bay time-lapse* and *Red kayak* were filmed in flat colors, which usually require color post-processing. Three of the videos (*Crowd run*, *Red kayak* and *Speed bag*) were taken from open video collection on media.xiph.org and one was taken from MSU video collection used for selecting testing video sets for annual video codecs comparison [2]. The description (and sources) of the first three videos can be found on site [6], and the rest *Bay time-lapse* video sequence contained a scene with water and grass and the grass and waves on the water.

Three versions of VMAF were tested: 0.6.1, 0.6.2, 0.6.3. The implementations of all three metric versions from MSU Video Quality Measurement Tool [3] were used. The results did not differ much, so the following plots are presented for the latest (0.6.3) version of VMAF.

3. Proposed Tuning Algorithm

For color and brightness adjustment, two known and widespread image processing algorithms were chosen: unsharp mask and histogram equalization. We used the implementations of these algorithms which are available in open-source scikit-image [13] library. In this library, unsharp mask has two parameters which influence image levels: radius (the radius of Gaussian blur) and amount (how much contrast was added at the edges). For histogram equalization, a parameter of clipping limit was analyzed. In order to find optimal configurations of equalization parame-ters, a multi-objective optimization algorithm NSGA-II [10] was used. Only the limits for the parameters were set to the genetic algorithm, and it was applied to find the best parameters for each testing video.

SSIM and VMAF scores were calculated for each video processed with the considered color enhancement algorithms with different parameters. As it was mentioned before, after color correction the videos were compressed with medium preset of x264 encoder on 3 Mbps. Then, the difference between metric scores of processed videos and original video were calculated to compare, how color corrections influenced quality scores. Fig. 2 shows this difference for SSIM metric of *Bay time-lapse* video sequence for different parameter values of unsharp mask algorithm. The similarity scores for VMAF quality metric are presented in Fig. 3.

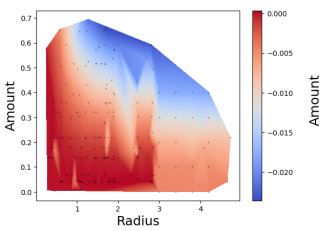


Fig. 2. SSIM scores for different parameters of unsharp mask on *Bay time-lapse* video sequence.

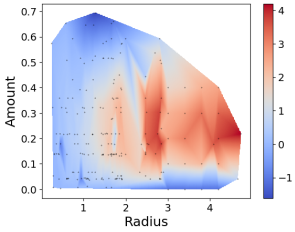


Fig. 3. VMAF scores for different parameters of unsharp mask on *Bay time-lapse* video sequence.

On these plots, higher values mean that the objective quality of the color-adjusted video was better according to the metric. VMAF shows better scores for high radius and a medium amount of unsharp mask, and SSIM becomes worse for high radius and high amount. The optimal values of the algorithm parameters can be estimated on the difference in these plots. For another color adjustment algorithm (histogram equalization), one parameter was optimized and the results are presented on Fig. 4 together with the results of unsharp mask.

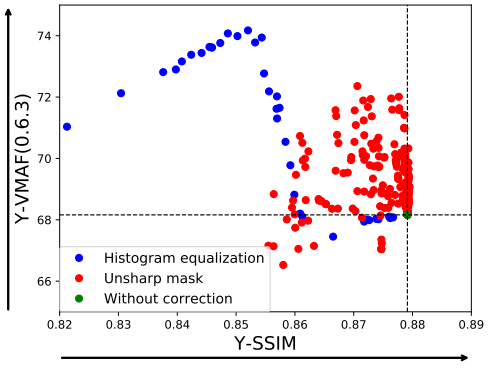
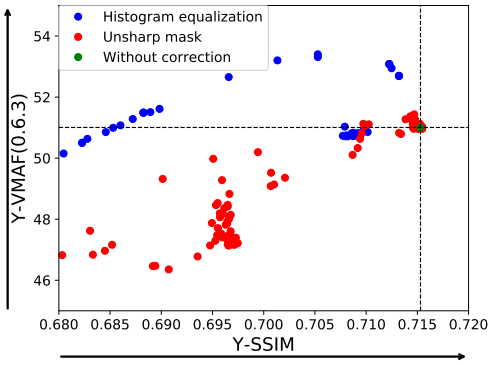
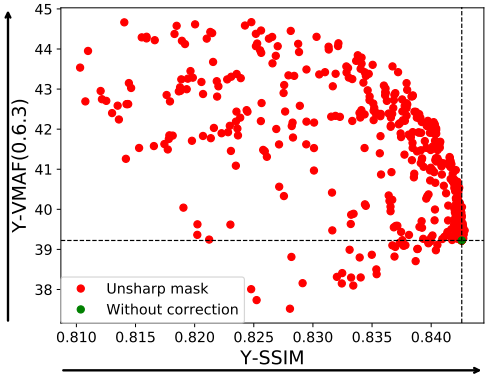


Fig. 4. Comparison of VMAF and SSIM scores for different configurations of unsharp mask and histogram equalization on *Bay time-lapse* video sequence. The results in the second quadrant, where SSIM values weren't changed and VMAF values increased, are interesting for us.

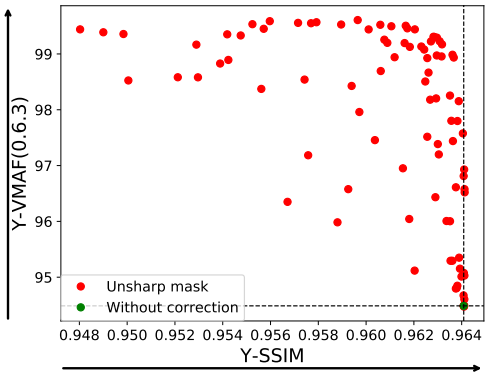
According to these results, for some configurations of histogram equalization VMAF become significantly better (from 68 to 74) and SSIM doesn't change a lot (decrease from 0.88 to 0.86). The results slightly dif-fer for other videos. On *Crowd run* video sequence, VMAF was not increased by unsharp mask (Fig. 5a) and was increased a little by histogram equalization. For *Red kayak* and *Speed bag* videos, unsharp mask could significantly increase VMAF and just slightly decrease SSIM (Fig. 5b and Fig. 5c)



(a) Color tuning results for *Crowd run* video sequence.



(b) Color tuning results for *Red kayak* video sequence.

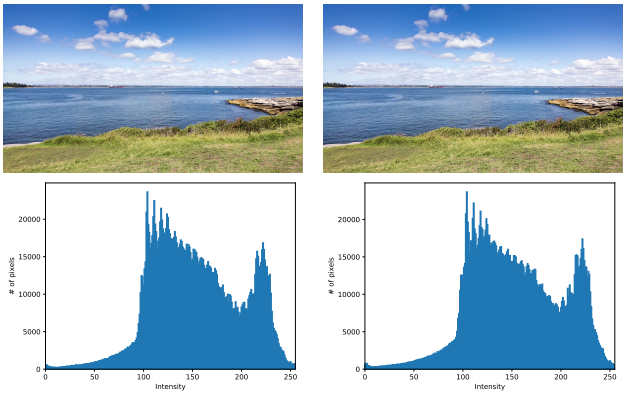


(c) Color tuning results for *Speed bag* video sequence.

Fig. 5. Comparison of VMAF and SSIM scores for different configurations of unsharp mask and histogram equalization on tested video sequences. The results in the second quadrant, where SSIM values weren't changed and VMAF values increased, are interesting for us.

4. Results

The following examples of frames from the testing videos demonstrate color corrections which increased VMAF and almost did not influence the values of SSIM. Unsharp mask with *radius* = 2.843 and *amount* = 0.179 increased VMAF without significant decrease of SSIM for *Bay time-lapse* (Fig. 6a and Fig. 6b). The images before and after masking look equivalent (a comparison in a checkerboard view is in Fig. 7) and have similar SSIM score, while VMAF score is better after the transformation.



(a) Without color correction $VMAF = 68.160$, $SSIM = 0.879$

(b) After unsharp mask $VMAF = 72.009$, $SSIM = 0.878$

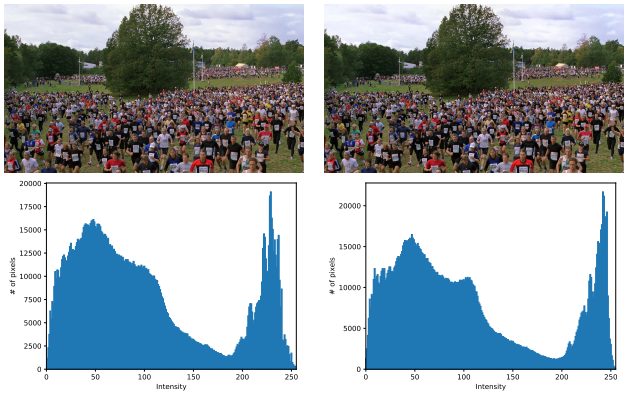
Fig. 6. Frame 5 from *Bay time-lapse* video sequence and its histogram with and without contrast correction. Two images and their histograms look equivalent.

Fig. 7. Checkerboard comparison of frame 5 from *Bay time-lapse* video sequence before and after distortions. Two images look almost equivalent.

For *Crowd run* sequence, histogram equalization with *kernelsize* = 8 and *cliplimit* = 0.00419 also increased VMAF (Fig. 8a and Fig. 8b). The video is more contrasted, so the decrease in SSIM was more significant. However, tho images also look similar (Fig. 9) and have similar SSIM score, while VMAF showed better score after contrast transformation.

Red kayak looked better according to VMAF after unsharp mask with *radius* = 9.436, *amount* = 0.045.

For *Speed bag*, the following parameters of unsharp mask allowed to increase VMAF greatly without influencing SSIM: *radius* = 9.429, *amount* = 0.114.



(a) Without color correction $VMAF = 51.005$, $SSIM = 0.715$

(b) After histogram equalization $VMAF = 53.083$, $SSIM = 0.712$

Fig. 8. Frame 1 from *Crowd run* video sequence and its histogram with and without color correction. Two images and their histograms look almost similar.

Fig. 9. Checkerboard comparison of frame 1 from *Crowd run* video sequence before and after distortions.

5. Conclusion

Video quality reference metrics are used to show the difference between original and distorted streams and are expected to take worse values when any transformations were applied to the original video. However, sometimes it is possible to deceive objective metrics. In our article, we described the way to increase the values of popular full-reference metric VMAF. If the video is *not* contrasted, VMAF can be increased by color adjustments without influencing SSIM. In another case, contrasted video can also be tuned for VMAF but with little SSIM worsening.

Although VMAF has become popular and important, particularly for video codec developers and customers, there are still a number of issues in its application. This is why SSIM is used in many competitions, as well as in MSU Video-Codec Comparisons, as a main objective quality metric.

We wanted to pay attention to this problem and hope to see the progress in this are, which is likely to happen since the metric is being actively developed. Our further research will involve a subjective comparison of the proposed color adjustments to the original videos and the development of novel approaches for metric tuning.

6. Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research under Grant 19-01-00785a.

7. References

[1] Cisco Visual Networking Index: Forecast and Methodology. 2016-2021.

[2] HEVC Video Codec Comparison 2018 (Thirteen MSU Video Codec Comparison) http://compression.ru/video/codec_ comparison/ hevc_2018/

[3] MSU Quality Measurement Tool: Download Page http://compression.ru/video/quality_ measure/ vqmt_download.html

[4] Perceptual Video Quality Metrics: Are they Ready for the Real World? Available online: https://www.ittiam.com/perceptual-video-quality-metrics-ready-real-world

[5] VMAF: Perceptual video quality assessment based on multi-method fusion, Netflix, Inc., 2017 https://github.com/Netflix/vmaf.

[6] Xiph.org Video Test Media [derf's collection] https://media.xiph.org/video/derf/

[7] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," in *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[8] C. Chen, S. Inguva, A. Rankin, and A. Kokaram, "A subjective study for the design of multi-resolution ABR video streams with the VP9 codec," in *Electronic Imaging*, 2016(2), pp. 1-5.

[9] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment meth-ods: A classification, review, and performance comparison," in *IEEE Transactions on Broadcast-ing*, 57(2), pp. 165–182, 2011.

[10] K. Deb, A. Pratap, S. Agarwal, and T. A. M. T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in *IEEE transactions on evolutionary computation*, 6(2), pp.182-197, 2002.

[11] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments", in *IEEE Transactions on Multimedia*, 2011, 13(5), pp. 935-949.

[12] H. R. Sheikh and A. C. Bovik, “Image informa-tion and visual quality,” in *IEEE International Conference on Acoustics, Speech, and Signal Pro-cessing*, 2004, . 3. – . iii-709.

[13] S. van der Walt, J. L. Schonberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. scikit-image: Image processing in Python. PeerJ 2:e453, 2014.