https://doi.org/10.2352/EI.2022.34.9.IQSP-358

Temporal MTF evaluation of slow motion mode in mobile phones

Lin Luo, Celalettin Yurdakul, Kaijun Feng, Dong Eun Seo, Fangwen Tu, and Bo Mu
OmniVision Technologies, Santa Clara, California 95054, USA

Abstract

While slow motion has become a standard feature in mainstream cell phones, a fast approach without relying on specific training datasets to assess slow motion video quality is not available. This manuscript proposes a fast and generalized no-reference objective metric based on temporal loss in modulation transfer function (MTF) to evaluate interpolated slow motion mode in mobile phones. First, a standard chart embodying slanted edges is used to capture a slow-motion video. Second, the edge spread function is extracted from a region of interest in individual slow motion frames. Third, the line spread functions and MTFs are calculated. Finally, reference and interpolated frames are identified. Sharpness loss in slow motion mode is quantified by the MTF area difference between minimum interpolated frame and reference frame MTF scores. The proposed approach is evaluated in simulated and experimentally captured slow motion videos. In experiments, slow-motion videos are captured by moving mobile phones mounted on a motorized linear stage apart from the test chart at a constant speed while keeping the test chart still. The proposed MTF scores of several mainstream cell phones are analyzed and compared.

Introduction

Slow motion videos are those captured at a rate much higher frame rates, that is 120 frames per second (fps), and playbacked at normal speed, that is 30 fps, by time-stretching so that time appears to be slower than a human observer can perceive. Nowadays slow motion has become a standard feature in mainstream mobile phones and is considered a performance metric for mobile phones. Mobile image sensors typically support up to 240 fps recording frame rate. In order to achieve a higher frame rate, which is larger than 1000 fps, the captured video speed is often computationally upscaled by 2x or 4x using frame interpolation algorithms. Interpolated frames are estimated from reference frames along with the temporal distance to consecutive frames. The interpolation process introduces sharpness loss in the interpolated frames. Figure 1 illustrates the 4x interpolation that produces 120 fps interpolated slow motion videos from 30 fps captured videos. In recent years, numerous methods have been demonstrated for video interpolation, including optical flow [1], deep neural network [2], and motion estimation and compensation [3, 4]. To assess the performance of each method as well as the whole slow motion mode in terms, a quantitative metric for efficiently evaluating the interpolated video quality is highly desirable. Unfortunately, conventional video quality metrics are designed for normal speed videos and thus suffer from providing reliable results for slow motion videos [5].

In this study, we aim to solve the problem of slow motion video quality metric by proposing a general approach based on temporal MTF evaluation for mobile phones. The goal of this method is to provide an objective measurement of slow motion video quality without relying on specific training datasets and full-reference frames. Our approach could be used as an easy and efficient baseline score for further evaluation of slow motion video quality.

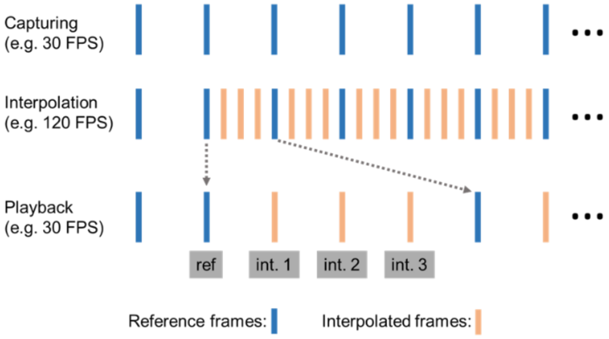

Figure 1: *Schematic of interpolated slow motion mode in mobile phones.*

Related work

The current video quality assessment (VQA) of normal speed videos is primarily focusing on data-driven approaches, which include the training and evaluation phase. As shown in Figure 2, firstly, a dataset relevant to real use cases with different quality losses is collected in the training phase. Secondly, expert or non-expert observers are invited to score the quality of the videos in the dataset. Thirdly, the subjective scores from all observers are pre-processed and then normalized to the range of 0 to 100 or 1.0, with the score of 100/1.0 for perfect video. Meanwhile, objective metrics are utilized to extract quality features of the videos in the dataset. To evaluate quality artifacts, widespread human perception based objective metrics including Visual Information Fidelity (VIF) and Detail Loss Metric (DLM) [6] are adopted. Finally, a machine-learning such as Support Vector Machine (SVM) or deep-learning model is then used to map the objective quality metrics to subjective scores. In the evaluation phase, the trained model is used to evaluate a video after the same objective metrics are extracted. The accuracy of data-driven approaches heavily depends on the training dataset. When a video is not included in the training dataset, the assessment score may be far from human evaluation results. Another limitation of data-driven approaches is the cost to conduct experiments to obtain reliable subjective scores. Furthermore, subjective experiments are time-consuming due to the need for human observers. Thus, classical data-driven approaches, such as Video Multimethod Assessment Fusion (VMAF) are complex and hard to generalize to widespan of videos due to the use of a limited number of distorted videos (e.g., 300) in the training phase [7]. Alternatively, pixel-based metrics, such as mean square error (MSE), peak signal-to-noise ratio (PSNR), or structural similarity index (SSIM), can be utilized to evaluate slow motion mode in mobile phones. There are two major limitations of these pixel-based metrics: (i) a full-reference video with the same frames as slow motion video is required and (ii) the motion blur level scores, which is one of the important quality factors when assessing a slow motion video, cannot be correlated well with the aforementioned objective metrics.

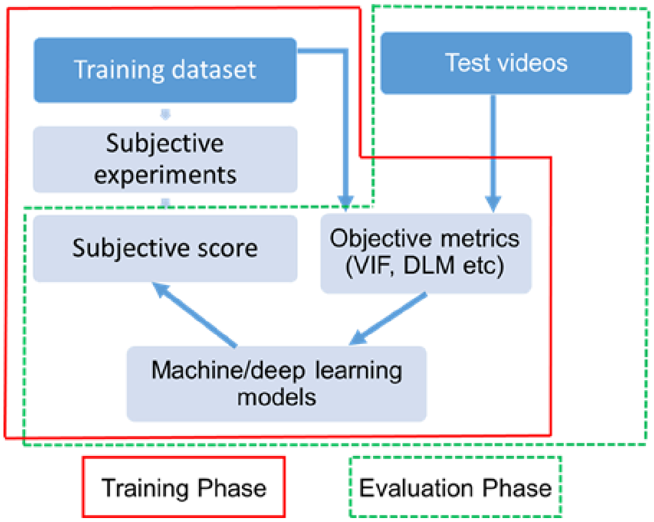

Figure 2: *Example flowchart of data-driven VQA approach (adapted from VMAF [7]).*

Proposed approach

We seek fast and efficient slow motion video quality assessment metric based on temporal MTF loss. We propose an objective metric that does not require full-reference (ground truth) which is often hard to obtain. Our approach relies on the sharpness loss which can be quantified by means of degradation in the MTF over the video sequences.

Different metrics in different frequency ranges, such as MTF50/MTF30/(normalized MTF area) can be used to quantify the sharpness loss in interpolated slow motion frames. The peak normalized MTF area is relatively insensitive to oversharpening compared to MTF50/MTF20 owing to the integral over all frequencies up to Nyquist rate (0.5 c/p). Therefore, we used the normalized MTF area metric to quantify the image sharpness. The MTF area score can be expressed as follows:

$$MTF_{Area} = \int_0^{0.5} MTF(f)df$$

where f denotes the frequency in units of cycles/pixel (c/p). The sharpness loss is then defined by the difference of MTF scores between the reference and interpolated frames. This can be referred to as MTF loss:

$$MTF_{loss}(\%) = 100\frac{MTFArea_{ref} - \min[MTFArea_{int}]}{MTFArea_{ref}}$$

where $MTFArea_{ref}$ and $MTFArea_{int}$ denote the MTF area calculated from MTFs of the reference frame and interpolated frames, respectively. min[] operator denotes the minimum MTF area among the interpolated frames. It should be noted that MTF loss is calculated using consecutive reference frames and interpolated frames from those reference frames. Therefore, MTF loss metric do not require full-reference compared with other objective video quality metrics such as PSNR, SSIM, and VIF.

Methods

Synthetic slow-motion video generation

To obtain slow motion reference frames, a synthetic slanted edge image is generated. Then, a blur kernel that mimics the effect of the camera optics is applied. A diffraction-limited point spread function (PSF) corresponds to camera lens f/2.0 at 550 nm illumination is calculated to be as the blur kernel. After the convolution, Gaussian noise with variance of 0.0001 is added. Figure 3 shows the snapshots from generated images. The reference video sequence is generated by circularly shifting blurred reference frames from left to right at a 5px/frame speed. Synthetic slow-motion video is then generated using deep learning-based frame interpolation algorithms developed by Jiang *et al.* [2]. The network input is fed by the generated reference video sequence. The input video speed is upscaled by 4x, interpolating 3 frames in between reference frame sequences.

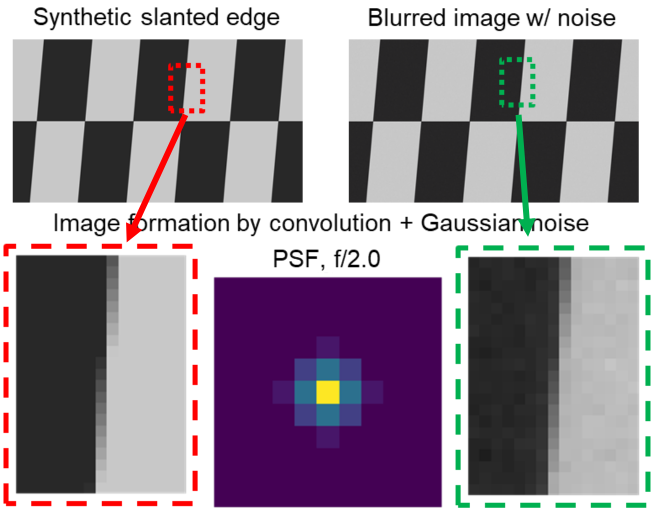

Figure 3: *Synthetic slanted edge frame generation. Blurred frames are generated using a diffraction-limited airy disk kernel corresponding to f/2.0 at 550 nm illumination wavelength. Additive Gaussian noise with var = 0.0001 is also added. The slanted edge angle is 5°.*

Experimental setup

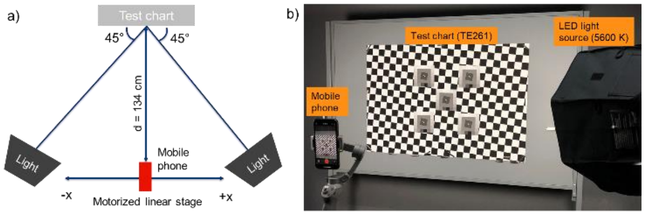

Figure 4: *(a) Schematic of the Slow-motion video capturing setup. (b) Side-view of the experimental setup. Two LED light sources at 5600 K illuminates the test chart at 45° angle. The mobile phone is mounted atop of the motorized linear stage which is translated at a constant speed during the video capturing.*

Figure 4 depicts the experimental setup schematic used for capturing slow motion videos for temporal MTF analysis. A test chart (TE261, Image Engineering) consisting of a tilted checkerboard background and low contrast slanted edges surrounded with gray patches is used for MTF evaluation. Two daylight light-emitting diode (LED) floodlight with 5600 K color temperature is placed at 45° angle with respect to test chart normal. To improve the illumination uniformity, LED floodlights are bundled with softboxes. The illuminance at the chart surface is measured by a light meter as 416 lux. The tested mobile phone is held by a phone holder mounted on top of a motorized linear stage which can control the speed and displacement. The stage is placed 134 cm away from the test chart. This distance is more than 100x of camera module lens focal length to ensure measurements are not limited by printing MTF. Slow-motion videos are captured by translating mobile phones at a constant speed of 1 m/s while keeping the test chart still.

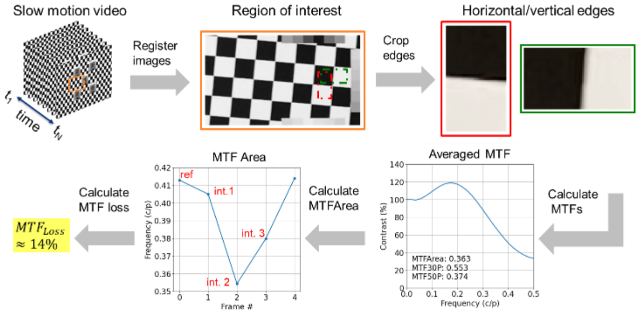

***Figure 5**: Temporal MTF e-SFR analysis flowchart. Slow motion videos recorded in mp4 format are first pre-processed. Identified ROI image is registered at each video frame followed by cropping horizontal and vertical edge patches. Horizontal and vertical e-SFRs are calculated using ISO12333 standard algorithm and averaged. Normalized MTF area is calculated at each frame and then the MTF loss score is quantified after identifying reference and interpolated frames.*

Temporal MTF analysis framework

Mobile phones save the recorded slow motion videos in mp4 format. We first preprocessed video frames to extract horizontal and vertical slanted edge patches from an identified region of interest (ROI) image located at the chart center. To do so, images are read from video files on Python. At each video frame, the corresponding ROI image is registered using 2-dimensional cross-correlation. Then, pre-defined horizontal and vertical edge patches are cropped. MTFs from both horizontal and vertical slanted edge patches are calculated using the e-SFR slanted-edge algorithm complying with ISO12333 standards. The MATLAB implementation of this algorithm named "sfrmat4" by Burns *et al.* is available open-source [8, 9]. Edge patches are fed into the algorithm in RGB color space which is later to be converted to luminance channel using default color channel weights. The calculated horizontal and vertical edge MTFs are then averaged. We refer to this averaged MTF as MTF, unless otherwise noted. Peak normalized MTF area from each MTF plot is then calculated and temporal MTF area plot is extracted. Finally, the MTF loss score is calculated using reference and interpolated frames.

Simulation Results

We first evaluated the simulated slow-motion videos described in the methods section. Figure 6a shows cropped ROIs from reference and interpolated frames. The sharpness loss in the second interpolated frame is more obviously perceived among the others. The interpolated frame ESFs have broadened profiles resulting in lower MTF profiles. The analyzed MTF area plot shows that the most blurred image frame is the interpolated frame 2 which is consistent with the visual observation. The MTF loss is calculated as 14%. These results indicate that MTF loss could be used as a metric to evaluate the sharpness loss in slow motion videos.

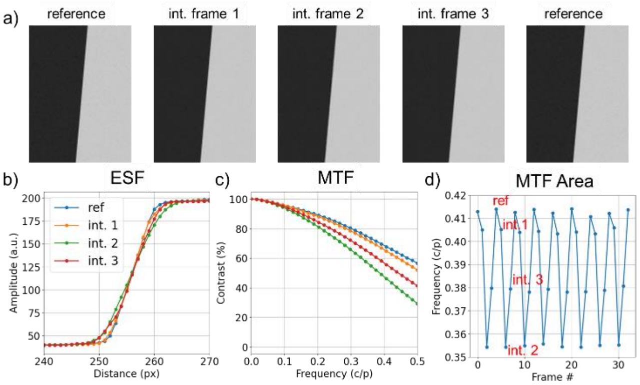

***Figure 6:** MTF loss evaluation of simulated slow motion videos. (a) Cropped slanted edge ROIs from simulated slow motion video frames. Reference frames are generated using parameters in Figure 3. Three frames are interpolated between the adjacent reference frames. Calculated (b) Edge spread functions and (c) MTF plots. (d) Temporal MTF area plot for the total of 33 frames (9 reference and 24 interpolated frames).*

Experimental Results

Next, we experimentally evaluated temporal MTF loss using two flagship mobile phones referred to as phone A and phone B. Phone A can achieve 7680 fps at 720p resolution by 4x interpolation from video frames captured at 1920 fps. Phone B can achieve 960 fps at 720p resolution by 2x interpolation from video frames captured at 480 fps. Figure 7 shows the experimental results obtained by these phones. The ESF plot from phone A has ringing artifacts due to oversharpening which be easily seen in corresponding MTF plots. Interestingly, the MTF difference between reference and interpolated frames is very small leading to a lower MTF loss score of 1.2%. In contrast, phone B has smaller ringing artifacts but still has software sharpening. MTF differences are much more obvious and overall MTF loss is 10%. This demonstrates that phone B has more than 8x worse slow motion video quality performance than that of phone A. We further speculate that such a large performance difference could be caused by video capturing frame rate differences. Phone A has a 4x higher video capturing frame rate, thus is less likely to be affected by motion blur.

Conclusion

We reported an MTF-based objective video quality assessment metric for interpolated slow motion videos captured by mobile phones. Our approach relies on temporal sharpness loss between the reference and interpolated frames. MTF area difference score is defined to quantify this sharpness loss. We simulated interpolated slow-motion videos and analyzed the sharpness loss. In experiments, we evaluated two flag-ship mobile phones, respectively phone A (7680 fps, 720p, 4x frame rate interpolation) and phone B (960 fps, 720p, 2x frame rate interpolation). We demonstrated interpolated frames have inferior sharpness compared with reference frames. The experimental results show consistency with the simulations. Phone B has 8 times more sharpness loss compared with Phone A. We speculate that high-speed video capturing could reduce the sharpness loss in interpolations.

Our approach could be further extended to use other objective quality metrics including noise. Since MTF only quality metrics do not correlate well with human perception, human visual system-dependent metrics including contrast sensitivity function in both spatial and temporal domain could be integrated. Speed-dependent motion blur effect can be also investigated. Furthermore, our study is limited to one-dimensional translation. Various motion effects including transverse and rotation could be also investigated in future studies.

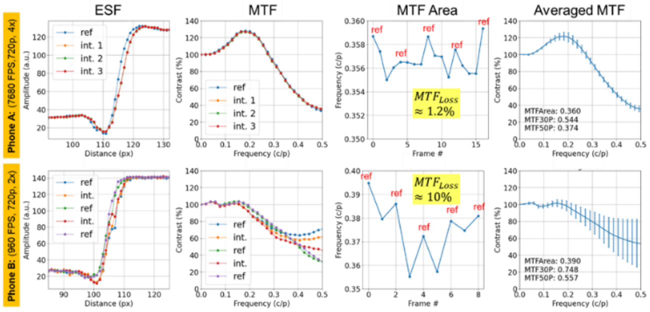

Figure 7: MTF loss evaluation of experimentally captured slow motion videos by two flagship phones. From left to right, calculated ESF, MTF, temporal MTF Area, and averaged MTF plots. Error bars in averaged MTF come from the MTF variation across multiple video frames.

References

[1] E. Herbst, S. Seitz and S. Baker, "Occlusion reasoning for temporal interpolation using optical flow," Technical Report, August 2009.

[2] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 9000-9008*, 2018.

[3] 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," in *TPAMI*, 2019.

[4] H. Liu, R. Xiong, D. Zhao, S. Ma and W. Gao, "Multiple hypotheses Bayesian frame rate up-conversion by adaptive fusion of motion-compensated interpolations," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 22, pp. 1188-1198, Aug. 2012.

[5] H. Men, V. Hosu, H. Lin, A. Bruhn and D. Saupe, "Visual Quality Assessment for Interpolated Slow-Motion Videos Based on a Novel Database," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020.

[6] S. Li, F. Zhang, L. Ma and K. N. Ngan, "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments," *IEEE Trans. Multimedia,* vol. 13, no. 5, pp. 935-949, Oct. 2011.

[7] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy and M. Manohara, "Toward a practical perceptual video quality metric," The Netflix Tech Blog, June 2016. [Online]. Available: netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[8] P. D. Burns and D. Williams, "Camera resolution and distortion: Advanced edge fitting," in *IS&T Internat. Symp. Electronic Imaging, Image Quality and System Performance XV*, 2018.

[9] P. D. Burns. [Online]. Available: burnsdigitalimaging.com/Software/sfrmat4/.

Author Biography

***Lin Luo** received his B.S. degree in Biomedical Engineering and M.S. degree in Computer Science from Southeast University, China, in 2007 and 2010, and his Ph.D. degree in Color Science from The Hong Kong Polytechnic University, in 2015. Currently he is a Staff Color Imaging Scientist at OmniVision Technologies in Santa Clara, California, USA. His interests lie in the area of digital color imaging and image/video quality evaluation.*

***Celalettin Yurdakul** received the B.S. degree in Electrical and Electronics Engineering from Koç University, Istanbul, Turkey, in 2016, and the Ph.D. degree in Electrical Engineering from Boston University, Boston, MA, USA, in 2021. He then joined OmniVision Technologies in Santa Clara, California, USA, and is currently focusing on physical system modeling of next-generation CMOS image sensors.*

***Kaijun Feng** received the B.S. degree in Microelectronics from Peking University, Beijing, China, in 2013, and the Ph.D. degree in Electrical Engineering from University of Notre Dame, South Bend, Indiana, USA in 2019. From 2019 to 2020 he was with Seagate Technology in Fremont, California where he developed machine vision systems for advanced manufacturing. In 2020 he joined the CTO office of OmniVision Technologies in Santa Clara, California, USA, and is currently focusing on system modeling, data processing and application research for novel image sensors.*

***Dong Eun Seo** received the B.S. degree in Electronic Engineering from Sogang University, Seoul, Korea, in 2007 and the M.S. degree in Electronic Engineering from Sogang University, Seoul, Korea, in 2009. He is currently working as a Staff Algorithm Engineer in OmniVision Technologies Singapore Pte. Ltd. His research interests include image processing, imaging, and computer vision.*

***Fangwen Tu** received the B.E. degree from Dalian University of Technology, Dalian, China, 2012 and the PhD. degree in Electrical Engineering from National University of Singapore, Singapore, 2017. He is currently working as a Staff Algorithm Engineer in OmniVision Technologies Singapore Pte. Ltd. His research interests include machine learning, image processing, and sensor fusion.*

***Bo Mu** received his Ph.D. degree in Imaging Science from Rochester Institute of Technology, Rochester, NY, USA in 2007. He is currently a Director of Algorithm Development, OmniVision Technologies in Santa Clara, California, USA. His research interests include color image and video processing, computer vision, computational imaging and image quality metric.*