International Journal of Computational and Experimental Science and ENgineering (IJCESEN)
Vol. 5-No.2 (2019) pp. 94-99
http://dergipark.org.tr/ijcesen

ISSN: 2149-9144

Research Article

Full Reference Video Quality Evaluation using Foveated Vision and Multiple Fixation Points

Mario VRANJEŠ*, Snježana RIMAC-DRLJE, Denis VRANJEŠ

Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Department of Communications, Osijek-CROATIA

* **Corresponding Author** : mario.vranjes@ferit.hr
ORCID: 0000-0003-3563-4735

Article Info:

DOI: 10.22399/ijcesen.477034
Received : 31 October 2018
Accepted : 30 July 2019

Keywords

Foveated vision
Video quality
Full reference

Abstract:

In video applications it is necessary to continuously measure the video quality perceived by the end-user. Thus it is desirable to know which parts of video frame, i.e. which contents, attract viewers' attention. If this information is known, then it is possible to estimate perceived video quality in a meaningful way. However, automatic detection of viewers' fixation points is time-consuming process and often is omitted in objective video quality assessment (VQA) metrics. Based on our previous work, in which we proposed Foveation-based content Adaptive Root Mean Squared Error (FARMSE) VQA metric, in this work we propose two new full-reference (FR) VQA metrics called Multi-Point FARMSE (MP-FARMSE), and Simple-FARMSE (S-FARMSE). Both new-proposed metrics are based on foveated-vision features of human visual system and spatio-temporal features of video signal. In MP-FARMSE, by using an engineering approach, we implemented the fact that viewer's attention can be directed out of the center of the frame, thus covering use-cases when objects of interest are not located in the center of the frame. The main idea when creating the S-FARMSE metric was to reduce the computational complexity of the final algorithm and to make S-FARMSE metric capable of processing high-resolution video signals in real-time. Performances of the new-proposed metrics are compared to performances of seven existing VQA metrics on two different video quality databases. The results show that performances achieved by MP-FARMSE and S-FARMSE are quite close to those of state-of-the-art VQA metrics, whereas at the same time their computational complexity level is significantly lower.

1. Introduction

The number of video-based applications, as well as the number of their users, has been rising significantly. Due to the high capacity requirements of the uncompressed video, prior its transmission video has to be compressed, and it introduces different distortions (artefacts) in video. Furthermore, during transmission process, additional degradations are introduced in video, e.g. due to packet loss, packet delay, etc. Thus it is very important to continuously monitor and evaluate the quality of the video perceived at the end-user side, in order to ensure the satisfactory level of end-user Quality of Experience (QoE).

Video quality assessment (VQA) can be performed by using objective and subjective methods. When using subjective methods, human observers evaluate the video quality (quite expensive and time consuming process), whereas in case of objective method the computer algorithm tries to predict the video quality perceived by human observer. That significantly decreases the duration and costs of VQA process, and thus objective VQA methods are very popular research area. Generally, from requirements for reference video point of view, objective VQA metrics can be classified as: full-

reference (FR) (require the entire reference, uncompressed, video to be available), reduced-reference (RR) (require some features of the reference video) and no-reference (do not need information about reference video). Detailed review of different FR, RR and NR VQA metrics, as well as more different classifications of VQA metrics, can be found in [1-3]. In this paper we propose two new FR objective VQA metrics, called Multi-Point Foveation-based content Adaptive Root Mean Squared Error (MP-FARMSE), and Simple-FARMSE (S-FARMSE), which are based on our previous research in which we presented FARMSE metric [1]. Both metrics are primarily based on foveated-vison features of human visual system (HVS) and spatio-temporal features of video.

In Section 2 a brief information about foveated-vision and its usage in VQA metrics is given, whereas the MP-FARMSE and S-FARMSE metrics are presented in Section 3. In Section 4 the comparison of new metrics' performance and seven existing FR VQA metrics performances on two different video quality databases is given, while Section 5 presents the concluding remarks.

2. Foveated-Vision and Objective Video Quality Assessment

There are many HVS characteristics that are often taken into account when designing an accurate and reliable objective VQA metric (multi-channel organization of HVS, contrast and color sensitivity of the HVS, masking effect, …) and some of those metrics can be found in [1-6]. However, there is a limited number of metrics that incorporate foveated-vision based HVS characteristics.

Namely, when watching video, viewer fixates a certain part of the frame and the resolution and spatial acuity of HVS are highest at and near the fixation point, while they decrease as the distance from the fixation point increases (this is known as foveated-vision). These observations could be effectively implemented in an image/video quality evaluation algorithm, but it is necessary to know which part of the frame attracts viewers' attention. It is possible to automatically detect viewers' fixation points, but since it is quite time-consuming process, it is usually omitted in objective VQA metrics. In [7-9] authors performed different experiments and concluded that viewers generally more often fixate regions with higher level of contrast and slow movement, and specially regions near the center of the frame. This facts are taken into account when designing FMSE metric [10], which uses the foveation-based contrast sensitivity matrices for moving scenes in quality evaluation process, taking into account the effect of additional spatial acuity reduction due to motion in a video sequence. Furthermore, FARMSE metric presented in [1] is derived from FMSE, but it additionally implements spatial masking in a computationally simple, but very effective way and uses more efficient way for implementation of the retinal image velocity calculation. Both FMSE and FARMSE suppose that center of the frame is fixation point. Thus, degradations located near the center of the frame would more influence perceived video quality. On the other hand, in [11] authors proposed PQI metric, in which M fixation points are predicted by content saliency and the central bias property. Similarly to FMSE and FARMSE, different error sensitivity matrices are then used for the foveal and extra-foveal vision. Few additional metrics that implement the foveated-vision in VQA process can be found in [12-14]. In next section, two new metrics, MP-FARMSE and S-FARMSE, are presented.

3. MP-FARMSE and S-FARMSE Metrics

Since both new metrics presented in this section are based on FARMSE metric, firstly the brief description of FARMSE will be given. Due to the limited space in this paper, only general description of FARMSE steps will be described, without mathematical equations and details. Note that all the details regarding FARMSE metric and its calculation process can be found in [1].

A block-diagram of FARMSE algorithm is given in Fig. 1. The FARMSE metric assumes that absolute DiFference (*DF*) between the reference frame $F_o(x,y,t)$ and the corresponding impaired frame $F_d(x,y,t)$ is a good starting measure of degradation in an impaired video frame.

Furthermore, FARMSE filters obtained *DF* in two separate channels, low-frequency and high-frequency channel (FDF_k - Filtered absolute DiFference, k = 1, 2), thus simulating HVS spatial frequency dependent processing of visual information (multi-channel organization of HVS). Then follows spatial masking implementation, since the distortion can be significantly masked by the video content itself ($MFDF_k$ - Masked Filtered absolute DiFference, k = 1, 2). Key step, which incorporates foveated-vison HVS characteristics, consists of giving different weights to errors occurring at different locations in the frame, assuming that the observer fixates the center of the frame (applying S_{fk}^* matrices with the center of the frame as fixation point, k = 1, 2). By taking it into account the greater weights are given to these distortions that appear in and near the center of the frame, while the weights assigned to distortions appearing in frame parts located far from the center

decrease by increasing the distance from the center (FD_k = Foveated masked filtered absolute Difference, k = 1, 2). To obtain the unique quality

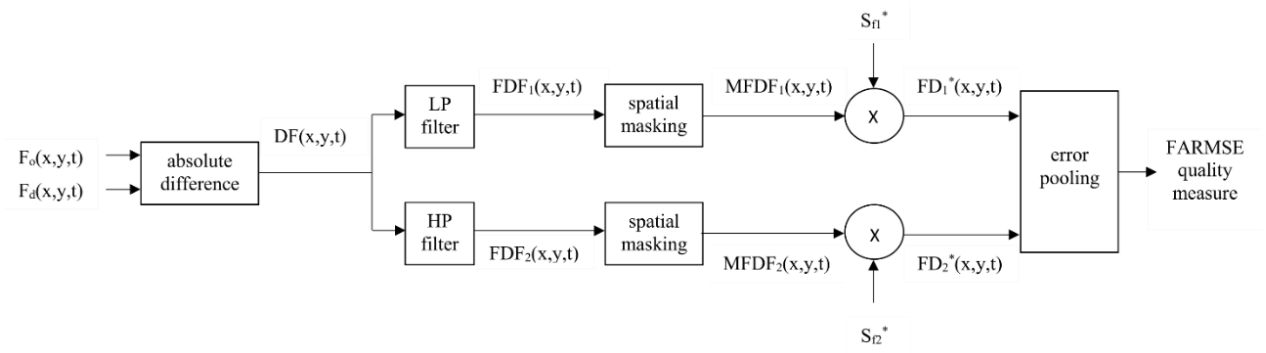

Figure 1. *A block diagram of FARMSE algorithm.*

measure for a given video, firstly spatial pooling of the foveated masked filtered absolute differences within each frame is performed, followed by temporal pooling of the quality measures of all frames from a given video. All details regarding implementation of particular FARMSE step as well as all mathematical equations related to FARMSE final score calculation can be found in [1].

3.1. MP-FARMSE Metric

Measurements performed for a large number of viewers and a large set of video signals, as mentioned in [7-9], showed that the area of interest within the video frame is generally a wider area around the center of the frame itself (not just one point, as simplified in FARMSE algorithm). The reason for that can be found in a fact that viewers' interests can be different when watching the same scene, but it should be emphasized that these interests are mostly related to the area around the center of the frame. Therefore, when designing MP-FARMSE metric, we consider the case when the foveation-based error sensitivity matrices for scenes with moving objects (S_{fk}^* , k = 1, 2) are calculated using multiple fixation points within one frame. In that way in MP-FARMSE metric the greater weights are given to the distortions located in a wider area around the center of the frame (compared to these in FARMSE). Specifically, in MP-FARMSE algorithm it is supposed that 5 different fixation points exist within one frame, that are distributed in a way presented in Fig. 2. In addition to the center of the frame, four additional fixation points equally distanced from the center of the frame are supposed to be possible fixation points. For each of these points m (m={1,2,3,4,5}) matrices $S_{fk,m}^*$ are calculated (see [1, 15]. After that the corresponding elements of all 5 error-sensitivity matrices ($S_{fk,m}^*$) are summed, and then these obtained final error-sensitivity matrices, $S_{fk,MP}^*$

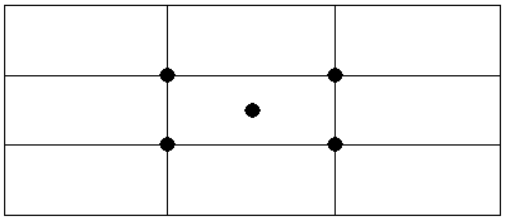

Figure 2. *Fixation points in FARMSE algorithm*

are normed according to Eq. (1), so that maximal value that an element in these matrix can take is equal to 1:

$$S_{fk_{MP}}^* = \frac{\sum_{m=1}^{5} S_{fk_m}^*}{\max\left(\sum_{m=1}^{5} S_{fk_m}^*\right)}, \quad \text{k=1,2} \qquad (1)$$

In that way two unique foveation-based error sensitivity matrices for moving scenes (for two frequency channels), using 5 fixation points are obtained. All other steps of MP-FARMSE algorithm are identical to those of FARMSE algorithm. It is important to note that, due to usage of 5 fixation points, calculation time of the MP-FARMSE metric is two times longer than this of FARMSE.

3.2. S-FARMSE Metric

In order to simplify FARMSE metric from [1] and make it faster, S-FARMSE metric is proposed. As the first simplification of FARMSE, S-FARMSE calculates FARMSE score only for low-frequency channel (k=1), since the HVS contrast sensitivity is the highest for medium and low spatial frequencies. Generally, when performing FARMSE algorithm for both spatial-frequency channels, it is shown that in error-pooling step, the final values of the matrix

FD_2 have lower influence on the final quality score than those from matrix FD_1. In that way, the S-FARMSE calculation complexity is approximately double reduced with respect to FRAMSE.

Second simplification of FARMSE is motivated with the approach used in MOVIE [16] metric, and is related to the number of video frames for which the quality score is calculated. Namely, in order to obtain the final MOVIE score for the entire video, MOVIE algorithm calculates the quality score for each eighth frame. Despite this fact, MOVIE algorithm is shown to be one of the best objective VQA algorithms and thus we applied the same approach for S-FARMSE calculation – we calculated S-FARMSE for each eighth frame. Finally, S-FARMSE is approximately 12 times faster than FARMSE algorithm.

4. Results and Discussion

MP-FARMSE and S-FARMSE performance evaluation is performed using video signals from two different databases, LIVE [17] and ECVQ [2]. These databases contain a total of 240 distorted video signals of 18 distinct contents with a wide range of visual quality. In our experiments each database was randomly divided into two sets, set A (training set) and set B (test set), with an equal number of signals in each set. All details about databases division can be found in [15]. It is important to note that in each database the set A was used for training of the FARMSE, MP-FARMSE and S-FARMSE algorithms in order to obtain the best fitting parameters for particular algorithm. Note that these parameters were then used for quality evaluation of the signals from the set B. New metrics performances are compared to those of seven FR objective video quality metrics: PSNR, VSNR [18], SSIM [19], MS-SSIM [20], FMSE [10], MOVIE [16] and FARMSE [1]. More details about calculation of the final quality score by the particular quality metric can be found in [1]. Performance of the analyzed metrics are examined by using Pearson linear correlation coefficient (PLCC), and prior PLCC calculation a nonlinear mapping between objective (for particular metric) and subjective scores is performed (more details in [1]). The results per distortion type (Wireless, IP, H.264, MPEG-2, MPEG-4 Part 2) are presented in Table 1.

Table 1. *Pearson linear correlation coefficient for (a) LIVE (b) ECVQ database (a)*

Algorithm	*Wireless*		*IP*		*H.264*		*MPEG-2*		*All sequences*	
	set A	**set B**	**set A**	**set B**	**set A**	**set B**	**set A**	**set B**	**set A**	**set B**
PSNR	0.7206	0.6588	0.4859	0.5901	0.6108	0.6018	0.3127	0.4108	0.5290	0.5809
VSNR	0.7517	0.6385	**0.6711**	**0.8698**	0.6319	0.6486	0.6395	0.7468	0.6277	0.7218
SSIM	0.6739	0.5531	0.2026	0.8109	0.6564	0.6648	0.4886	0.6075	0.4878	0.5906
MS-SSIM	0.7983	0.6561	0.4856	0.8526	0.6966	0.7421	0.7606	0.7979	0.7029	0.7810
FMSE	0.9052	0.7820	0.6459	0.7830	0.8039	0.8200	0.7493	0.7466	0.7710	0.7784
MOVIE	0.8574	**0.8392**	0.4987	0.8561	0.7985	0.7459	0.8616	**0.8800**	0.7775	**0.8351**
FARMSE	**0.9094**	0.7513	0.6604	0.8172	0.8463	**0.8623**	0.8368	0.7755	0.7917	0.7916
MP-FARMSE	0.8968	0.7609	0.6604	0.8311	0.8224	0.8254	**0.9040**	0.8640	**0.8005**	0.7974
S-FARMSE	0.8942	0.7395	0.4624	0.7117	**0.8498**	0.8573	0.8395	0.7674	0.7500	0.7835

(b)

Algorithm	*H.264*		*MPEG-4 Part 2*		*All sequences*	
	set A	**set B**	**set A**	**set B**	**set A**	**set B**
PSNR	0.7156	0.7136	0.7272	0.7674	0.7356	0.7911
VSNR	0.8572	0.8771	0.7971	0.7971	0.8338	0.8402
SSIM	0.9113	0.9057	0.8876	0.8905	0.8986	**0.9015**
MS-SSIM	0.9344	0.9132	0.8784	0.8295	0.9082	0.8782
FMSE	0.9489	0.9303	0.8962	0.8568	0.9108	0.8678
MOVIE	0.9389	0.9214	**0.9269**	**0.8927**	**0.9332**	0.8517
FARMSE	**0.9784**	0.9598	0.8650	0.8141	0.9120	0.8697
MP-FARMSE	0.9735	0.9526	0.8381	0.8200	0.8919	0.8537
S-FARMSE	0.9775	**0.9615**	0.8654	0.8152	0.9075	0.8612

It can be seen that generally four metrics, i.e. MOVIE, FARMSE, MP-FARMSE and S-FARMSE, outperform other analyzed metrics (except VSNR for IP distortion type and SSIM for All sequences in ECVQ database). The reason for such results can be found in a fact that those four metrics implement a larger number of HVS characteristics than other analyzed metrics. However it should be noted that the calculation time of MP-FARMSE and S-FARMSE metric is not significantly higher than this of other used metrics, while MOVIE calculation time is approximately 115 times longer than this of MP-FARMSE. More detailed discussion about the results presented in Table 1, as well as the statistical analysis of all metrics results, can be found in [2].

When only FARMSE, MP-FARMSE and S-FARMSE performances are compared, it can be seen that MP-FARMSE achieves the highest performance for LIVE database (FARMSE is very close), and FARMSE for ECVQ database (S-FARMSE is very close). In LIVE database the videos are of higher resolution (768x432) than those in ECVQ (352x288) Thus extending the region of interest (with 5 fixation points) around the center of the frame, which is used in MP-FARMSE algorithm, has led to the better estimate of perceived video quality. Because of the relatively larger frame resolution, the viewer more often switches his attention further than the center of the frame, what MP-FARMSE covers with additional assumed fixation points. On the other hand, for smaller resolution (ECVQ database), extending the region of interest around the center of the frame does not influence the final quality score of the video, since the frame is of smaller dimension. Therefore, the assumption with a single fixation point (FARMSE and S-FARMSE) leads to quite accurate results. Regarding S-FARMSE, it is shown that the high performance, relatively close to this of FARMSE, can be achieved, while the calculation time in that case is significantly shorter.

5. Conclusion

Since VQA is required in many video-based applications, it is necessary to design accurate and reliable objective VQA metric. In this paper we extend our research presented in [2] and propose two new objective FR VQA metrics, named MP-FARMSE and S-FARMSE. Besides lot of very often used HVS characteristics, MP-FARMSE and S-FARMSE implement foveated-vision based HVS characteristics, and it is shown that each of the proposed metrics can be used in a specific case (video resolution, video content type,…) and achieve high performance. In future work we would like to examine the ways of efficient implementation of automatic fixation point detection in our proposed VQA metrics, what should lead to higher correlation with subjective quality scores.

Acknowledgement

This work was supported by J.J. Strossmayer University of Osijek business fund through the internal competition for the research and artistic projects „IZIP-2016-55“.

References

[1] M. Vranješ, S. Rimac-Drlje, D. Vranješ, “Foveation-based content adaptive root mean square error for video quality assessment”, Multimedia Tools and Applications, Vol. 77, 21053-21082, **(2018)**

[2] M. Vranješ M, S. Rimac-Drlje, K. Grgić, “Review of objective video quality metrics and performance comparison using different databases”, Signal Processing: Image Communication, Vol. 28, 1–19, **(2013)**

[3] S. Chikkerur, V. Sundaram, M. Reisslein, L.J. Karam, “Objective video quality assessment: a classification, review, and performance comparison”, IEEE Transactions on Broadcasting, Vol. 57, 165–182, **(2011)**

[4] S.H. Bae, M. Kim, “DCT-QM: A DCT-based quality degradation metric for image quality optimization problems”, IEEE Transactions on Image Processing, Vol. 25, 4916–4930, **(2016)**

[5] F. Zhang, D.R. Bull, “A perception-based hybrid model for video quality assessment”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26, 1017–1028, **(2016)**

[6] S. Li, L. Ma, K.N. Ngan, “Full-reference video quality assessment by decoupling detail losses and additive impairments”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22, 1100–1112, **(2012)**

[7] J. You, J. Korhonen, A. Perkis, “Attention modelling for video quality assessment: balancing global quality and local quality”, Proceedings of International Conference on Multimedia and Expo (ICME 2010), **(2010)** July 19-23; Suntec City-Singapore

[8] I. Van der Linde, U. Rajashekar, A.C. Bovik, L.K. Cormack, “DOVES: a database of visual eye movements”, Spatial Vision, Vol. 22, 161–177, **(2009)**

[9] A. Mittal, A.K. Moorthy, W.S. Geisler, A.C. Bovik, “Task dependence of visual attention on compressed videos: points of gaze statistics and analysis”, Proceedings of the SPIE 7685 **(2011)**: 786505–786510

[10] S. Rimac-Drlje, M. Vranješ, D. Žagar, "Foveated mean squared error–a novel video quality metric", Multimedia Tools and Applications, Vol. 49, 425-445, **(2010)**

[11] Y. Zhao, L. Yu, Z. Chen, C. Zhu, "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 21, 1890–1902, **(2011)**

[12] J. You, T. Ebrahimi, S. Perkis, "Attention driven foveated video quality assessment", IEEE Transactions on Image Processing, Vol. 23, 200–213, **(2014)**

[13] Z. Wang, A.C. Bovik, L. Lu, J. Kouloheris, "Foveated wavelet image quality index", Proceedings of the SPIE 4472 (2001), 1–11

[14] S. Lee, M.S. Pattchis, A.C. Bovik, "Foveated video quality assessment", IEEE Transactions on Multimedia, Vol. 4 129–132, **(2002)**

[15] M. Vranješ, "Objective image quality metric based on spatio-temporal features of video signal and foveated vision", PhD Thesis, Josip Juraj Strossmayer University of Osijek, Croatia, **(2012)**

[16] K. Seshadrinathan, A.C. Bovik, "Motion-tuned spatio-temproal quality assessment of natural videos", IEEE Transactions on Image Processing, Vol. 19, 335–350, **(2010)**

[17] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, "Study of subjective and objective quality assessment of video", IEEE Transactions on Image Processing, Vol. 19, 1427-1441, **(2010)**

[18] D.M. Chandler, S. Hemami, "VSNR; A wavelet based visual signal-to-noise-ratio for nature images", IEEE Transactions on Image Processing, Vol. 16, 2284-2297, **(2007)**

[19] Z. Wang, A.C. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, Vol. 13, 600-612, **(2004)**

[20] Z. Wang, E. Simoncelli, A.C. Bovik, "Multi-scale structural similarity for image quality assessment", Proceedings of 37th Asilomar Conference on Signals, Systems and Computers (ACSSC 2003), **(2003)** November 9-12, Pacific Grove, USA