Full-Reference SSIM Metric for Video Quality Assessment with Saliency-Based Features

Eduardo Romani[1(✉)], Wyllian Bezerra da Silva[2], Keiko Verônica Ono Fonseca[1], Dubravko Culibrk[3], and Alexandre de Almeida Prado Pohl[1]

[1] Graduate Program on Electrical and Computer Engineering (CPGEI), Federal Technological University–Paraná (UTFPR), Curitiba, Brazil
romaniseduardo@gmail.com, {keiko,pohl}@utfpr.edu.br
[2] Federal University of Santa Catarina – Santa Catarina (UFSC), Florianópolis, Brazil
willianbs@gmail.com
[3] Department of Industrial Engineering - Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
dubravko.culibrk@gmail.com

Abstract. This paper uses models of visual attention in order to estimate the human visual perception and thus improve metrics of Video Quality Assessment. This work reports on the use of the saliency based model in a full-reference structural similarity metric for creating new metrics that take into account regions that greatly attract the human attention. Correlation results with the differential mean opinion score values from the LIVE Video Quality Database are presented and discussed.

Keywords: Video quality assessment · Salient model · Human visual system

1 Introduction

Today there is a growing quantity of available multimedia content, particularly videos, creating a need for their quality evaluation and measurement. Video Quality Assessment (VQA) techniques concern the development of metrics that attempt to estimate the video quality as perceived by the Human Visual System (HVS). They are meant to assess the degradations related to lossy compression, losses in transmission and/or reception processes or even due to problems related to the original video content, for example. In order to determine the efficiency of a metric at predicting the human perception of video quality, a correlation between objective and subjective of video sequences is needed, which is achieved by calculating the correlation between objective and subjective scores (Mean Opinion Score - MOS [1] or Differential Mean Opinion Score - DMOS [1]).

VQA metrics can be classified as Full-Reference (FR), Reduce-Reference (RR) or No-Reference (NR), depending if the whole, reduced or no information about the original video is available to estimate the level of degradation. Many studies have already been developed based on encoding artifacts for all types of metrics. For instance, video

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 547–554, 2015.
DOI: 10.1007/978-3-319-23222-5_66

degradation can be assessed by relating the information about encoding characteristics, such as blocking and blurring features [2]. Several FR metrics have been developed, going from the most simple ones, such as the Peak Signal-to-Noise Ratio (PSNR), obtained by the ratio between the maximum power of a signal and its noise, to techniques that involve the structural similarity with motion associated weighting, such as the Speed-Weighted Structural SIMilarity Index (SW-SSIM) [3] and to more complex ones, such as the Motion-based Video Integrity Evaluation (MOVIE) [4], based in the spatio-temporal domains using the Gabor filter. Among RR metrics, there is the Spatial-Temporal Reduce Reference Entropic Differences (STRRED) [5], calculated by the entropic differences. And concerning NR metrics, the No-Reference Video Quality Assessment based on the Extreme Learning Machine algorithm (NRVQA-ELM) [6, p.8] uses the spatio-temporal features with a Neural Network in the prediction method.

A good prediction method should emulate the perceived quality of the HVS. Therefore, in addition to the analysis of encoding effects, VQA techniques have been recently developed, which includes the temporal analysis and the human attention models. For instance, it is known that the visual system focus the attention on objects of interest [7]. This is an evolutionary ability that permits the human being to rapidly localize salient

objects in a scene, such as a possible prey or predators.

In this way, salient models are being included in VQA techniques, as they are able to highlight video regions with discrepancies, which greatly attracts the human attention since movements, differences in texture and artifacts are better perceived by the HVS in visual focus regions than in regions of peripheral vision. Models of visual attention are widely used in computer vision [7], in eye tracking [8], and recently they have also been used in VQAs [9], [10].

This work presents results on the use of the saliency based model to improve the quality assessment. As a starting point, the FR metric proposed by Wang [11] is used, which is based on the characteristics between the frame structure of the original video and the degraded video. The metric is then modified by introducing the saliency-based model in the processing, resulting in metrics that contain the salient and the not-salient characteristics. Results show that there is an improvement in prediction compared to the original metric when salient features are used. In the experiments the LIVE Video Quality Database [12] is employed, which follows the recommendations of the Video Quality Experts Group (VQEG) [1] for subjective assessments and is widely used in the VQA testing. For comparison purposes the Pearson Linear Correlation Coefficient (PLCC) [13] and Spearman's Rank-Order Correlation Coefficient (SROCC) [13] are considered.

This work is divided, as follows: section 2 contains a basic description of the saliency model and section 3 describes the modified FR metric SSIM using the same model. Next, section 4 presents the experimental results and the analysis, followed by the conclusion in section 5.

2 The Salient Model

The salient model used in this paper has been developed by Culibrk [9], and is based on the bottom-up processing [14] for modeling visual attention, in which the presented stimulus process establishes a region of saliencies, i.e., a region with outliers in a given context. In this type of processing a disparate region is detected through its outliers in relation to its adjacent region, as, for instance, a red sign 'Stop' in a landscape with trees, and in a scene in which an object moves in one direction and at a different speed from the rest of the content, such as a moving car in a highway.

The algorithm employs the principles of multi-scale processing in the background, where two background frames are obtained by Infinite Impulse Response (IIR) filters for each frame. The principles of cross-scale motion consistency, temporal coherence and outlier detection are also used. The multiscale processing is similar to that proposed by Itti [7], in which a Gaussian pyramid is formed with scales and backgrounds frames. The method uses the maps of crominance, intensity and orientation to extract the conspicuity maps of the frame, then, a linear combination of these maps create a saliency map of the spatial information for one frame. A novelty filter in the form of Mexican hat function is applied to the two background and the current frame to extract the motion information, hereupon one single image obtained by the temporal filtering is formed. Finally, an outlier detector is used to detect the salient regions, defining regions that differ significantly from the context of the video sequence. This approach makes it possible to consider cross-scale consistency and the spatial coherence.

Through this method a binary map of saliencies is extracted for each frame. The processed frame of the video sequence is showed in Figure 1(a). Figure 1(b) shows the map of saliencies extracted of the same frame. A filter is proposed in this paper to create blocks with size of 8x8 pixels. These blocks can then be characterized as salients or not. They are used to eliminate small regions of saliency (sizes less then 8 pixels salients per block) and to create macro salient regions. The filter focus on regions mostly affected by the HVS attention, aiming at reducing small noise and optimizing the salient algorithm used to improve FR and NR metrics predictions. Figure 1(c) presents how the filter works: white regions are the superposition of the salient map from Figure 1(b) and the filtered salient map of Figure 1(d), the gray regions are areas formed by just one of that two maps, and the black areas have no salient regions in both maps. In the zoom area showed by the red square it is possible to see the amount of pixels in one block and, by the definition of the filter, blocks with less than 8 pixels salients per block do not integrate a salient block. The result of this process can be seen in Figure 1(d).

Looking to the final result in Figure 1(d) it is easy to see some examples of salient (white) regions determined by the crominance discrepancy in the two wheels of the tractor, the orientation differences in the boundaries of the object (tractor) and intensity enhancement in the ceiling of the tractor. There are also some salient regions caused by the motion information, which can not be perceived, like the spatial features, because it is generated by the temporal filter that requires more than just one frame. The salient map described in this section is used to characterize the new metrics approach in section 3.

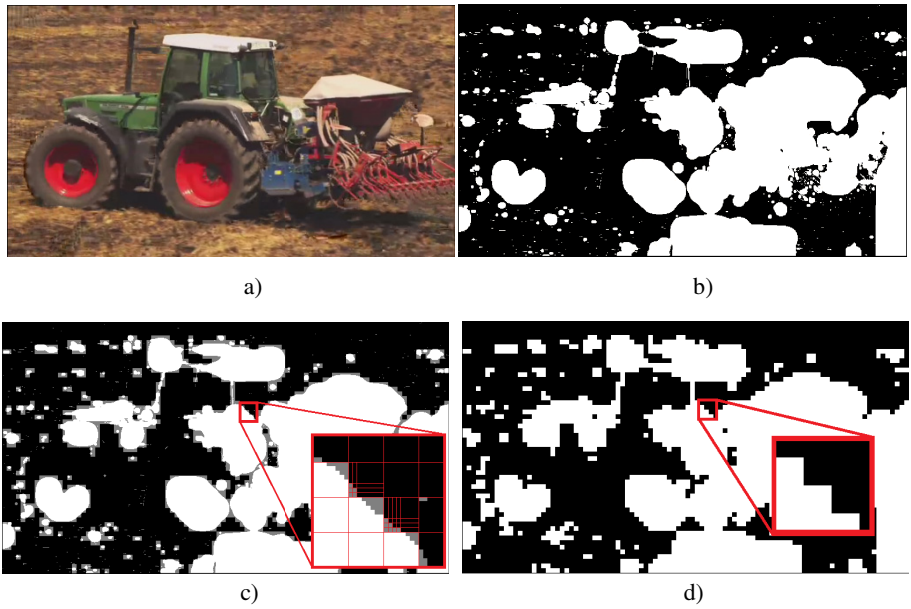

Fig. 1. Salient model Extraction. a) Processed frame. b) Salient map after using [9] model. c) Filter 8x8 for the extraction of macro regions of saliencies. d) Filtered salient map.

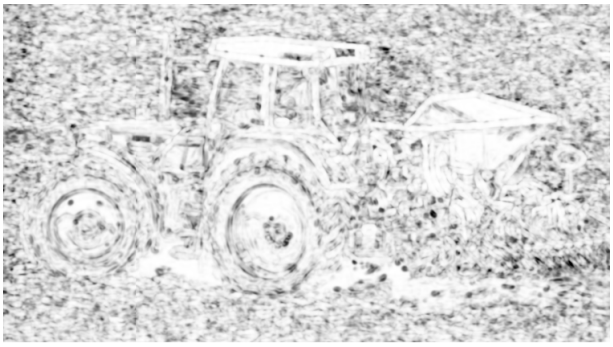

Fig. 2. SSIM map extracted from the processed frame of Figure 1(a).

3 Salient-Based SSIM

Considering that the HVS is adapted for extracting structural information from a scene, the FR SSIM metric [11] analyses the structural similarity between the processed and the reference video. The SSIM uses the fact that the luminance of a frame is composed by illumination and reflection of objects, but the structure is not affected by the illumination, so it can be isolated from the structure of the frame.

Assuming the signal x as the reference frame and signal y as the degraded frame, Wang [11, p. 605] proposed a method comparing three components: luminance, contrast and structure, given by

$$SSIM(x,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma}, \quad (1)$$

where α, β and γ are parameters used to adjust the importance of each component. In this paper, α, β and γ received value '1' like suggested by Bovik [2] to simplify the equation. With (1) it is possible to plot the SSIM map presented in Figure 2 for the degraded frame of Figure 1(a).

A visual inspection in the structural similarity map represented in Figure 2 reveals the structural differences between two consecutives frames of the video sequence, outlining the object boundaries, in this case, the tractor.

In (2) the overall frame quality is presented.

$$MSSIM(X,Y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x_j, y_j), \quad (2)$$

To create versions of salient and not-salient SSIM it is necessary to calculate separately the corresponding regions of interest. In this paper the salient map extracted from the degraded frame is used. Considering MS_j as the binary Map of Saliencies exemplified in Figure 1(d) and counting the total of numbers "1" in the binary frame the total number of salient regions in the frame can be obtained, as follows

$$M_s = \sum_{j=1}^{M} MS_j, \quad (3)$$

Then, (4) and (5) calculate the overall SSIM value of the salient regions and of the not-salient regions of the frame, respectively.

$$MSSIM\_S = \frac{1}{M_s}\sum_{j=1}^{M} SSIM(x_j, y_j), \qquad if\ MS_j = 1, \quad (4)$$

$$MSSIM\_NS = \frac{1}{M-M_s}\sum_{j=1}^{M} SSIM(x_j, y_j), \qquad if\ MS_j = 0. \quad (5)$$

Finally, to obtain the metric score of the video sequence, the average of the MSSIM of all frames is calculated.

$$SSIM\_std = \frac{1}{N_f}\sum_{i=1}^{N_f} MSSIM_i\,, \quad (6)$$

$$SSIM\_S = \frac{1}{N_f}\sum_{i=1}^{N_f} MSSIM\_S_i\,, \quad (7)$$

$$SSIM\_NS = \frac{1}{N_f}\sum_{i=1}^{N_f} MSSIM\_NS_i\,, \quad (8)$$

where N_f is the total number of frames of the video sequence, SSIM_std (6) is the standard metric as proposed by Wang [11], and SSIM_S (7) and SSIM_NS (8) are the proposed salient and not-salient SSIM modified metrics, respectively.

The analysis of the results of these three metrics is correlated to the DMOS values of video sequences using the PLCC and SROCC coefficients. The data is presented in Table I.

4 Experimental Results

The correlation of DMOS for videos in the LIVE Database with the saliency-based metrics is analyzed using PLCC (accuracy) and SROCC (monotonicity). Table I shows the results of the standard SSIM_std [11] metric and the proposed SSIM_S (7) and SSIM_NS (8) for all frames of the videos of LIVE database, subdividing in categories according to different types of videos. Numbers in bold emphasizes the best outcomes.

The SSIM_S metric presents an improvement of SROCC in all categories listed in the table, when compared to the results of the SSIM_std. The most significant SROCC results were obtained with the SSIM_S metric for Wireless and IP videos with an increase of 21.5% and 12.8%, respectively. Looking at the SROCC values, when all videos are analyzed together (last row of Table I), an improvement of 11.05% using the proposed salient method is observed, i.e., the relationship between the SSIM_S and the DMOS by a monotonic function presents a better correlation than the one using the original SSIM_std method.

For the PLCC, a significant improvement in the Wireless (24,5%), IP (9,3%) and "all" (11,2%) video categories is observed. However, in the case of H.264 and MPEG videos the best values were presented with the SSIM_NS method, but in these cases the difference lies within 1% and does not affect the general result in a significant way.

Table 1. PLCC (accuracy) and SROCC (monotonicity) between DMOS of LIVE database [12] and the metrics SSIM [11], SSIM Salient and SSIM Not-Salient.

Video	SSIM_std [11]		SSIM_S (7)		SSIM_NS (8)	
Type	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Wireless	0,5283	0,5221	**0,6577**	**0,6345**	0,4371	0,4163
IP	0,6137	0,4812	**0,6709**	**0,5430**	0,5812	0,4888
H.264	0,6900	0,6503	0,6831	**0,6598**	**0,7024**	0,6398
MPEG	0,5767	0,5581	0,5813	**0,5699**	**0,5826**	0,5527
ALL	0,5440	0,5248	**0,6048**	**0,5828**	0,4979	0,4714

The overall analysis points to the fact that the metric SSIM_S presents the best results for both PLCC and SROCC. We can also note that the main variations occur in the lowest correlation values of the original metric SSIM_std, given by the Wireless, IP and ALL video categories. The most significant improvement occurs when one does not know the specific video category and the SSIM_S metric is applied to the the set of "All" videos.

Observing the results of the SSIM_NS, the PLCC values of H.264 and MPEG videos are higher than the other metrics correlations, but don`t have a significant improvement. In general, the metric shows very low results, which reinforces the fact that the salient regions have more influence in the human visual system than the not-salient regions.

The strength of the metric arises from the fact that the salient model highlight the regions of the video with the higher discrepancies in orientation, crominance, intensity and motion, leading to a better correlation to the HVS evaluation.

5 Conclusion

The impact of saliencies in VQA metrics has been explored in this paper. Saliency-based features are employed in the standard FR SSIM [11] metric, resulting in two alternative metrics: SSIM_S based in the characteristics of the region of frames with the salient information and SSIM_NS with the complementary not-salient information. The comparison with the original metric is performed with the LIVE Video Quality Database.

The results obtained with the FR SSIM_S show consistent improvements compared to the original metric, as it can be noticed from values shown in Tables I. The use of the saliency-based model increases the correlation with the human perception, therefore enhancing the benefits of using saliencies in VQAs. This study also shows that the technique can be used in future approaches, such as the ones that employ NR metrics, which have been developed in past researches of the group, such as the NRVQA-ELM [6] and NRVQA-LM [15] techniques.

References

1. Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R BT.500 Std. (2002)
2. Wang, Z., Sheikh, H.R., Bovik, A.C.: No-reference perceptual quality assessment of JPEG compressed images. In: Proc. IEEE Int. Conf. Image Process., pp. 477–480 (2002)
3. Wang, Z., Li, Q.: Video quality assessment using a statistical model of human visual speed perception. J. Opt. Soc. Amer. A **24**(12), B61–B69 (2007)
4. Seshadrinathan, K., Bovik, A.C.: Motion-based perceptual quality assessment of video. IEEE Trans. Image Process. **19**(2), 335–350 (2010)
5. Soundararajan, R., Bovik, A.C.: Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Transactions on Circuits and Systems for Video Technology **3**(4), 684–694 (2013)
6. Silva, W.B.: Métodos sem referência baseados em características espaço-temporais para avaliação objetiva de qualidade de vídeo digital. Ph.D. dissertation, Federal Technological University - Paraná, March 2013
7. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience **2**(3), 194–203 (2001)
8. Liang, Z., Fu, H., Chi, Z., Feng, D.: Refining a region based attention model using eye tracking data. In: Proc. IEEE Int. Conf. Image Process., pp. 1105–1008, September 2010

9. Culibrk, D., Mirkovic, M., Zlokolica, V., Pokric, M., Crnojevic, V., Kukolj, D.: Salient motion features for video quality assessment. IEEE Transactions on Image Processing **20**(4), 948–958 (2011)
10. Wang, Y., Jiang, T., Ma, S., Gao, W.: Novel sapatio-temporal structural information based video quality metric. IEEE Transactions on Circuits and Systems for Video Technology **22**(7), 989–998 (2012)
11. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
12. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing **19**(16), 1427–1441 (2010)
13. O'Rourke, N., Hatcher, L., Stepanski, E.J.: A Step-by-Step Approach to Using SAS for Univariate & Multivariate Statistics, 2nd edn. Wiley-Interscience New York, NY (2008)
14. Connor, C., Egeth, H., Yantis, S.: Visual attention: Bottom-up versus top-down. Current Biology **14**(19), R850–R852 (2004)
15. Silva, W.B., Pohl, A.A.P.: No-reference video quality assessment method based on levenberg-marquardt minimization. In: XXX Simpósio Brasileiro de Telecomunicações (SBrT 2012), September 2012