Video Quality Assessment Based on the Effect of the Estimation of the Spatial Perceptual Information

Carlos D. M. Regis, José V. de M. Cardoso and Marcelo S. Alencar

***Abstract*— The objective video quality assessment is a quick and low cost alternative compared with the subjective evaluation. However, the objective evaluation is not as reliable, because their results are not always according to the perceived quality of the human visual system. This paper presents a new metric for objective video quality assessment, called PW-SSIM, based on the investigation of how the spatial perceptual information can be used as an estimate to predict the visual attention to a particular region of the video and insert a quality weighting according to the spatial perceptual information values. The PW-SSIM presents higher correlation coefficients when compared to popular models (PSNR and SSIM), for a subjective evaluation with 40 participants, considering degraded videos with salt and pepper, blurring and blocking, and 24 participants considering videos subject to Gaussian noise, suggesting that the PW-SSIM has a better ability to predict the perceived video quality for a group of spectators.**

***Keywords*— Video Quality Assessment, Structural Similarity, Visual Attention, Spatial Perceptual Information, Human Visual System.**

I. INTRODUCTION

Video quality assessment methods are subdivided into two categories: objective and subjective. Objective methods are computational models which, using statistical characteristics of the video, estimates the quality score, classified according to the availability of the original signal: *full reference*, in which the original video is compared with the test video; *reduced reference*, whenever only the characteristics of the original video are available for comparison with the test video and *no reference*, in which only the test video is used for assessment of the video quality. However, subjective methodologies assess the video quality via psychophysical experiments with human observers. The observer watches the video sequences and evaluates the video acording to a personal concept of quality.

The subjective approach is the natural way to assess of the video quality [1]. Nevertheless, subjective experiments are complex and time-consuming. Objective metrics are faster and has lower cost than the subjective metrics, because their results may be applied automatically to video systems, to detect imperceptible degradations to the human eye.

Objective video quality assessment constitutes an important sector for video services and processing systems, such as: vigilance systems [2], video on demand [3], spatial transcoding systems [4] and video conferecing [5]. However, the classical objective metrics, such as MSE (Mean Squared Error) and PSNR (Peak Signal to Noise Ratio), present an unsatisfactory correlation with the results provided by subjective evaluation, compromising the reliability of their measures [6].

Currently, the objective metrics that show better correlation with subjective tests are based on the structural similarity approach, proposed by Wang *et al* [7]. In an attempt to improve this approach, many researchers investigate how to introduce characteristcs of the human visual system (HVS) in the objective metrics, in order to raise the correlation with the subjective results. One of the key areas of research that are being investigated to obtain this improvement is the visual attention of the HVS.

Experiments indicate that the human visual attention is not equally distributed throughout the image space, but concentrates in a few regions [8]. It is estimated that the inclusion of methods that can identify the visual attention of a scene, i.e., assign a weight to the visual importance of regions on the image, tends to enhance the measures provided by the objective metrics.

Akamine and Farias [9] investigated the computational modeling of the visual attention peformed by saliency maps that were incorporated in objective metrics (PSNR and SSIM). This technique presents good results, mainly for saliency maps generated from eye-tracking, called subjective salience maps. You *et al* [10] also investigated the visual attention modelated by the saliency map, saliency attention map and GAFFE map [11], as an important factor to assess the objective image quality. Oprea *et al* [12] included elements that attract the attention: color contrast, size, orientation and eccentricity on the image quality assessment.

The authors propose a new objective metric, for *full reference* video quality assessment, derived from the structural similarity index (SSIM), which includes a visual attention model based on the weighting of the spatial perceptual information (SI) of each region. It is called Structural Similarity Index with Perceptual Weighting (PW-SSIM). The proposed metric was compared with MSE, PSNR and SSIM by means of the Pearson Correlation Coefficient (CC) and Spearman Rank-order Correlation Coefficient (SROCC).

This paper is organized as follow. Section II describes the Structural Similarity Index approach. Section III describes the proposed approach to objective video quality assessment. Section IV presents the experiments of subjective evaluation. Section V shows the simulation results and section VI presents the conclusion.

II. SSIM: STRUCTURAL SIMMILARITY INDEX

The Structural SIMilarity Index (SSIM) is a model proposed by Wang et al [13], based on the structural information of the image. Let $f = \{f_i \mid i = 1, 2, 3, \ldots, P\}$ be the original video

signal and $h = \{h_i \mid i = 1, 2, 3, \ldots, P\}$ be the degraded video signal, computed as the set of three measures over the pixel luminance plane: luminance comparison $l(f,h)$, constrast comparison $c(f,h)$ and structural comparison $s(f,h)$,

$$l(f,h) = \frac{2\mu_f\mu_h + C_1}{\mu_f^2 + \mu_h^2 + C_1}, \quad c(f,h) = \frac{2\sigma_f\sigma_h + C_2}{\sigma_f^2 + \sigma_h^2 + C_2}, \quad (1)$$

$$s(f,h) = \frac{\sigma_{fh} + C_3}{\sigma_f\sigma_h + C_3}, \quad (2)$$

$$\mu_f = \frac{1}{P}\sum_{i=1}^{P} f_i, \quad \mu_h = \frac{1}{P}\sum_{i=1}^{P} h_i, \quad (3)$$

$$\sigma_f^2 = \frac{1}{P-1}\sum_{i=1}^{P}(f_i - \mu_f)^2, \quad \sigma_h^2 = \frac{1}{P-1}\sum_{i=1}^{P}(h_i - \mu_h)^2, \quad (4)$$

$$\sigma_{fh} = \frac{1}{P-1}\sum_{i=1}^{P}(f_i - \mu_f)(h_i - \mu_h), \quad (5)$$

in which $C_1 = (0.01 \cdot 255)^2$, $C_2 = 2C_3 = (0.03 \cdot 255)^2$.

The structural similarity index is described as

$$\text{SSIM}(f,h) = [l(f,h)]^\alpha \cdot [c(f,h)]^\beta \cdot [s(f,h)]^\gamma, \quad (6)$$

in which usually $\alpha = \beta = \gamma = 1$ [13].

In pratice the SSIM is computed for an 8×8 sliding squared window or for an 11×11 Gaussian-circular window. The first approach is used in this paper. Then, for two videos, which are subdivided into D blocks, the SSIM is computed as

$$\text{SSIM}(f,h) = \frac{1}{D}\sum_{j=1}^{D}\text{SSIM}(f_j, h_j). \quad (7)$$

III. Perceptual Weighted Video Quality Approach

A. Spatial Perceptual Information

The Spatial Perceptual Information (SI) quantifies the complexity of the spatial details present in a video sequence, and it increases with the spatial complexity of the samples [14]. The SI is computed by means of gradient vectors, which in turn, are computed using the Sobel filters in the n-th video frame (Sobel(F_n)). The standard deviation of the magnitude of the gradient vectors (std[Sobel(F_n)]) is calculated for each video frame. The highest value among the standard deviations represents the SI of the video sample. This process is mathematically represented as

$$\text{SI} = \max\{\text{std}[\text{Sobel}(\text{F}_n)]\}. \quad (8)$$

The gradient vectors (∇f) estimate the rate of change of luminance values of the pixels along the horizontal and vertical directions, their maginitude is computed as

$$\mid \nabla f \mid = \left[\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right]^{1/2}, \quad (9)$$

in which, the partial derivatives are computed by convolution of the video sequence, frame-by-frame, with the Sobel masks:

(a) (b)

Fig. 1: Effect of the gradient vectors computed by Sobel operators

$$\begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, \qquad \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}.$$

Fig. 1b presents the magnitude of the gradient vectors computed by Sobel masks using the orginal video "Glasgow" (Fig. 1a), in which the lighter regions indicate a higher rate of change of luminance.

B. Visual Attention Weighted

Some models that use visual attention methods in image quality metrics have been proposed, that use salience maps, regions of interest and visual focus, giving more importance to regions more visually important to the overall index of the image quality [9][10].

The proposed method use the local spatial perceptual information to weigh the most visually important regions. This weighting is obtained as follows: compute the magnitude of the gradient vectors in the original video by means of the Sobel masks, then generate a frame in which the pixel values are the maginitude of the gradients, then this frame is partitioned into blocks 8×8 pixels and compute the SI in each block, as

$$\text{SI}_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\mu_{j,f'} - \mid \nabla f_i \mid)^2}, \quad (10)$$

in which, $\mu_{j,f'}$ represents the average magnitude of the gradients in a j-block and n is a total of gradient vectors in j-th block. For the case that the frames are partitioned uniformly in squares 8×8, $n = 64$.

It should be noted that the SI_j measure indicates the local spatial complexity, computed by means of the rate of change of luminance values of the pixels, and that propositon is used as an estimate of the visual attention of the HVS.

Based on this, the SI_j values were incorporated in the SSIM with a similar approach to that used in the works of Akamine and Farias [9] and Liu and Heynderickx [15]. It computed a weighted average for the SSIM algorithm, in which the weighting coefficients are the SI_j, to give the model called Structural Similarity Index with Perceptual Weighting (PW-

(a) Foreman (b) Glasgow (c) Mobile (d) Mother-Daughter

Fig. 2: Videos sequences used in the experiments.

SSIM),

$$\text{PW-SSIM}(f,h) = \frac{\sum_{j=1}^{D} \text{SSIM}(f_j, h_j) \cdot \text{SI}_j}{\sum_{j=1}^{D} \text{SI}_j}. \qquad (11)$$

The SI_j measure is computed using the original video. Furthermore, as shown the Fig. 3, the spatial perceptual information presents a considerable variation for degraded videos, mainly for videos degraded by blurring, which compromise the identification of important regions of the video. In the Fig. 3b, the terms "2-blurred" and "4-blurred" correspond at two and four applications of the average filter with 3×3 mask, respectively.

IV. Subjective Experiments

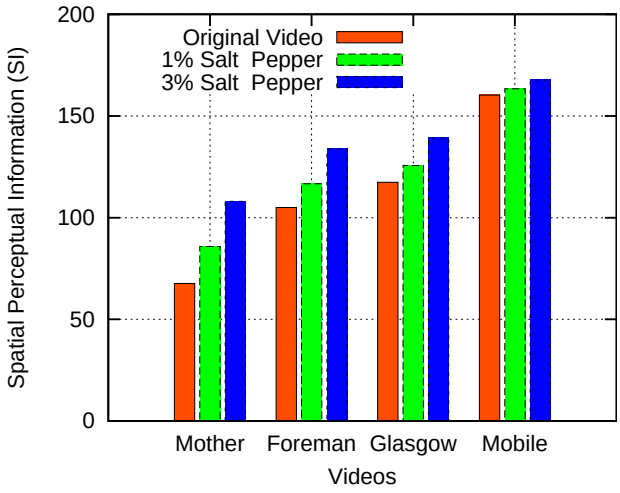
In the subjective evaluation, the observers watch video sequences and choose a score that corresponds to the level of video quality, the average of all subjective scores is called Mean Opinion Score (MOS).

There are two important points that should be considered in the subjective evaluation: the method adopted and the choice of the test material. The literature provides several methods to do a subjective evaluation of video quality [14]. The method used was the Absolute Category Rating (ACR), that is classified as a Single Stimulus Method, i.e., a category of judgement in which the video sequences are presented, one at a time and assessed independently according with a scale of the discrete values, as shown in the Table I. This method was used because it is easy and fast to implement [14].

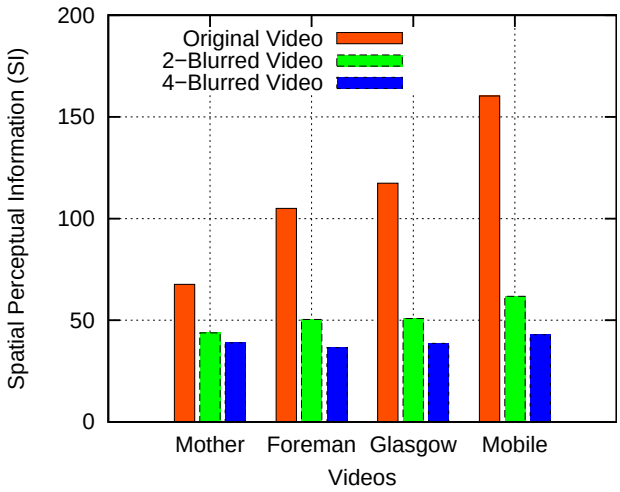
TABLE I: Discrete scale voting used in ACR method.

5	Excelent
4	Good
3	Fair
2	Poor
1	Bad

The proprieties of the Spatial perceptual Information (SI) and Temporal perceptual Information (TI) of the videos were considered to compose a set of video sequences for subjective evaluation. Therefore, it is important that the videos present a variety of values of SI and TI. The selected videos were: "Foreman", "Glasgow", "Mobile & Calendar"and "Mother and Daughter", in QCIF format (176×144 pixels) (Fig. 2), available for download from [16]. Their SI and TI values are shown in Fig. 4.



(a)



(b)

Fig. 3: a) Salt and Pepper Noise and Spatial Perceptual Information; b) Blurring and Spatial Perceptual Information

The selected videos were submitted to four types of degradation: Gaussian noise, salt and pepper noise, blurring and blocking (Fig. 5), producing a total of 32 simulated videos that were evaluated by 40 people for salt and pepper, blurring and blocking and 24 people for Gaussian noise. These types of degradation often occur in video processing systems and

(a) Blocking (b) Blurring (c) Gaussian Noise (d) Salt and Pepper Noise

Fig. 5: Samples of the simulated distortions used in the evaluation.

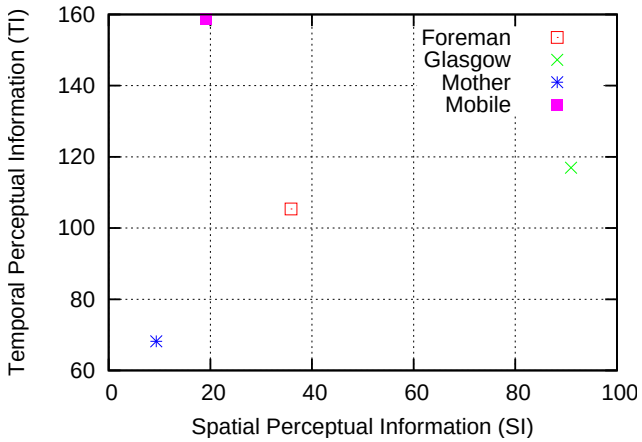


Fig. 4: SI and TI values for the set of the selected videos.

therefore it is important that they be simulated in controlled environments [17]. Table II shows the parameters used in the simulation.

TABLE II: Distortion parameters obtained from experiments.

Distortion	Intensity	Parameters
Salt & Pepper Noise	1	Probability 1%
	2	Probability 3%
Blurring	1	Average Filter with 3×3 mask (two applications)
	2	Average Filter with 3×3 mask (four applications)
Blocking	1	Probability 1%
	2	Probability 3%
Gaussian Noise	1	$\sigma = 0.0002$
	2	$\sigma = 0.003$

V. Experimental Results

To evaluate the metrics, the videos "Foreman", "Glasgow", "Mobile & Calendar" and "Mother and Daughter" were used. All videos have been degraded by noise presented in Table II.

The results obtained by the objective and subjetive evaluation, shown in Table III, indicate that the blurring degradation interferes more heavily with the visual quality for videos with high spatial complexity, such as "Mobile", suggesting that the comparison of spatial information may be regarded as a quality indicator.

TABLE III: Objective and Subjective Results for Blurred Videos.

Video	Intensity	PSNR	SSIM	PW-SSIM	MOS
Foreman	1	25.561	0.824	0.788	2.0455
Foreman	2	23.658	0.728	0.668	1.5303
Mobile	1	19.971	0.598	0.593	1.7692
Mobile	2	18.606	0.446	0.434	1.3051
Mother	1	29.218	0.831	0.768	2.0508
Mother	2	27.321	0.751	0.667	1.5500
Glasgow	1	23.789	0.705	0.653	1.7692
Glasgow	2	22.439	0.598	0.523	1.4118

TABLE IV: Objective and Subjective Results for Video with Salt and Pepper Distortion

Video	Intensity	PSNR	SSIM	PW-SSIM	MOS
Foreman	1	25.19	0.812	0.865	2.1364
Foreman	2	20.45	0.582	0.682	1.7424
Mobile	1	25.18	0.903	0.916	2.7167
Mobile	2	20.44	0.761	0.788	1.9661
Mother	1	25.67	0.746	0.825	2.3158
Mother	2	20.93	0.469	0.601	1.6481
Glasgow	1	25.17	0.831	0.876	2.4340
Glasgow	2	20.46	0.618	0.702	1.9423

The results obtained for the salt & pepper noise, as shown in Table IV, suggest that this degradation affects, more intensively, videos with low spatial information, such as "Mother and Daughter".

The comparison between the PW-SSIM, SSIM and PSNR was peformed by the computation of the Pearson Correlation Coefficient (CC) and Spearman Correlation Coefficient (SROCC) using the MOS obtained from the subjective evaluation and the objective measures. Table V shows the Pearson Correlation Coefficients for the subjective and objectives measures.

It is observed that the PW-SSIM presents a significant improvement for video sequences subject to blurring, blocking, and salt & pepper noise compared to the SSIM and PSNR. For the Gaussian noise, the SSIM metric provided the best correlation. This fact can be justified by the low ratio between the Gaussian noise and the Spatial Perceptual Information, as shown in Fig. 6, used in the weighting for the PW-SSIM.

TABLE V: Pearson correlation coefficients.

Model	Salt & Pepper	Blurring	Gaussian Noise	Blocking
PSNR	0.828	0.607	0.858	0.697
SSIM	0.902	0.776	**0.931**	0.792
PW-SSIM	**0.920**	**0.866**	0.918	**0.834**

TABLE VI: Spearman correlation coefficients.

Model	Salt & Pepper	Blurring	Gaussian Noise	Blocking
PSNR	0.595	0.738	0.762	0.667
SSIM	0.929	0.738	**0.976**	**0.881**
PW-SSIM	**0.976**	**0.762**	**0.976**	**0.881**

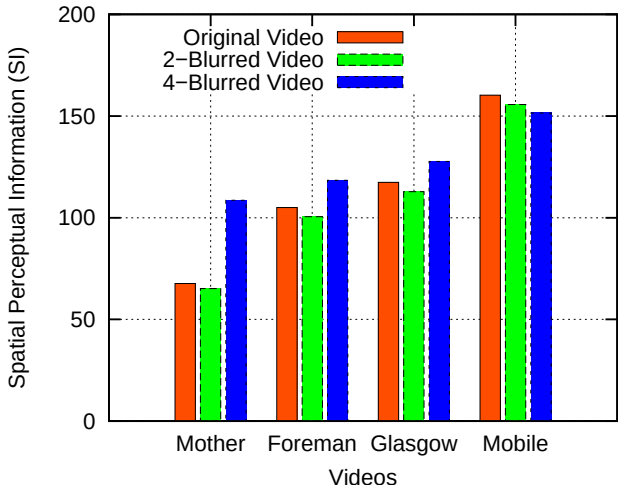


Fig. 6: Ratio between Gaussian noise and SI.

VI. Conclusions

A new metric for objective assessment of the video was proposed, combining the full reference metric (SSIM) added to the aspect of visual attention and Spatial Perceptual Information, called PW-SSIM.

The addition of these techniques obtained the best results, since there is a strong correlation between the video degradation and the quality, to the blurring, blocking, Gaussian noise and salt & pepper noise.

The proposed metric was compared with the PSNR and SSIM metrics and the correlation coefficients presented attained a better ability to predict the visual quality.

As future work, the aspect of visual attention and Spatial Perceptual Information to develop metrics, mainly *no reference*, can be used to evaluate the quality of the videos for specific types of degradation.

Acknowledgments

The authors would like to thank CNPq/PIBITI, UFCG/COPELE, IFPB and Iecom for providing research support.

References

[1] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98 –117, jan. 2009.

[2] H. U. Keval, "Effective, design, configuration, and use of digital cctv." Ph.D. dissertation, Department of Computer Science, University College London, 2009.

[3] R. F. Lopes, C. D. M. Regis, W. T. A. Lopes, and M. S. Alencar, "Adaptvod - an adaptive video-on-demand platform for mobile devices," in *5th FTRA International Conference on Multimedia and Ubiquitous Engineering (MUE)*, june 2011, pp. 257 –262.

[4] C. D. M. Regis, "Avaliação de técnicas de redução da resolução espacial de vídeos para dispositivos móveis," Dissertação de Mestrado, Universidade Federal de Campina Grande, Campina Grande, Brasil, 2009.

[5] C. R. D. Estrada, "Avaliação automática de qualidade de videoconferências de alta definição," Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2009.

[6] J. V. de Miranda Cardoso, A. C. S. Mariano, C. D. M. Regis, and M. S. Alencar, "Comparação das métricas objetivas baseadas na similaridade estrutural e na sensibilidade ao erro," *Revista de Tecnologia da Informação e Comunicação (RTIC)*, no. 2, pp. 33–40, April 2012.

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

[8] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001. [Online]. Available: http://papers.klab.caltech.edu/83/1/391.pdf

[9] M. C. Q. F. Wellington Y. L. Akamine, "Incorporating visual attention models into image quality metrics," in *Proceedings of the Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2012.

[10] J. You, A. Perkis, and M. Gabbouj, "Improving image quality assessment with modeling visual attention," in *2nd European Workshop on Visual Information Processing (EUVIP)*, july 2010, pp. 177 –182.

[11] U. Rajashekar, I. van der Linde, A. Bovik, and L. Cormack, "Gaffe: A gaze-attentive fixation finding engine," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 564 –573, april 2008.

[12] C. Oprea, I. Pirnog, C. Paleologu, and M. Udrea, "Perceptual video quality assessment based on salient region detection," in *Fifth Advanced International Conference on Telecommunications (AICT '09)*, may 2009, pp. 232 –236.

[13] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment using structural distortion measurement," in *International Conference on Image Processing.*, vol. 3, 2002, pp. III–65 – III–68 vol.3.

[14] ITU-T, "ITU-T recommendation P.910, subjective video quality assessment methods for multimedia applications," September 1999.

[15] H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *16th IEEE International Conference on Image Processing*, nov. 2009, pp. 3097 –3100.

[16] W. Trace, "YUV Video Sequences," *http://trace.eas.asu.edu/yuv/index.html*, Abril 2008.

[17] F. L. P. Albini, "Geração e avaliação de artefatos em vídeo digital," Dissertação de Mestrado, Universidade Tecnológica Federal do Paraná, Curitiba, Brasil, Março 2009.