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 according to a personal concept of quality.

The subjective approach is the natural way to assess 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 conferencing [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 characteristics 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 performed 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 modeled by the saliency map, saliency attention map and GAFSE 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 follows. 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 SIMILARITY 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 \( l(f, h) \), contrast comparison \( c(f, h) \) and structural comparison \( s(f, h) \),

\[
l(f, h) = \frac{2\mu_f\mu_h + C_1}{\mu_f^2 + \mu_h^2 + C_1}, \quad c(f, h) = \frac{2\sigma_f\sigma_h + C_2}{\sigma_f^2 + \sigma_h^2 + C_2}, \quad s(f, h) = \frac{\sigma_f h + C_3}{\sigma_f \sigma_h + C_3},
\]

in which, the partial derivatives are computed by convolution in a window or for an \( 8 \times 8 \) sliding squared window or for an \( 11 \times 11 \) 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

\[
SSIM(f, h) = [l(f, h)]^\alpha \cdot [c(f, h)]^\beta \cdot [s(f, h)]^\gamma,
\]

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

In practice the SSIM is computed for an \( 8 \times 8 \) sliding squared window or for an \( 11 \times 11 \) 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

\[
SSIM(f, h) = \frac{1}{D} \sum_{j=1}^{D} SSIM(f_j, h_j).
\]

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 \( \text{std}([\text{Sobel}(F_n)]) \) 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

\[
SI = \text{max} \{ \text{std}([\text{Sobel}(F_n)]) \}.
\]

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

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

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

\[
\begin{pmatrix}
-1 & 0 & +1 \\
-2 & 0 & +2 \\
-1 & 0 & +1
\end{pmatrix}
\]

and

\[
\begin{pmatrix}
-1 & -2 & -1 \\
0 & 0 & 0 \\
+1 & +2 & +1
\end{pmatrix}
\]

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 magnitude of the gradients, then this frame is partitioned into blocks \( 8 \times 8 \) pixels and compute the SI in each block, as

\[
SI_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left( \mu_{j,i} - |\nabla f_i| \right)^2},
\]

in which, \( \mu_{j,i} \) 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 \( 8 \times 8 \), \( n = 64 \).

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-
SSIM),

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

(11)

The \( \text{SI}_j \) measure is computed using the original video. Furthermore, as shown in 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 \( 3 \times 3 \) 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>
<thead>
<tr>
<th></th>
<th>Excellent</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE I: Discrete scale voting used in ACR method.

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 x 144 pixels) (Fig. 2), available for download from [16]. Their SI and TI values are shown in Fig. 4.

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
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.**

<table>
<thead>
<tr>
<th>Distortion</th>
<th>Intensity</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salt &amp; Pepper Noise</td>
<td>1</td>
<td>Probability 1%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Probability 3%</td>
</tr>
<tr>
<td>Blurring</td>
<td>1</td>
<td>Average Filter with $3 \times 3$ mask (two applications)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Average Filter with $3 \times 3$ mask (four applications)</td>
</tr>
<tr>
<td>Blocking</td>
<td>1</td>
<td>Probability 1%</td>
</tr>
<tr>
<td>Gaussian Noise</td>
<td>1</td>
<td>$\sigma = 0.0002$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$\sigma = 0.003$</td>
</tr>
</tbody>
</table>

**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 subjective 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.

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 performed 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 objective 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Salt &amp; Pepper</th>
<th>Blurring</th>
<th>Gaussian Noise</th>
<th>Blocking</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.828</td>
<td>0.007</td>
<td>0.858</td>
<td>0.697</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.902</td>
<td>0.776</td>
<td><strong>0.951</strong></td>
<td>0.792</td>
</tr>
<tr>
<td>PW-SSIM</td>
<td><strong>0.920</strong></td>
<td><strong>0.866</strong></td>
<td>0.918</td>
<td><strong>0.834</strong></td>
</tr>
</tbody>
</table>

TABLE VI: Spearman correlation coefficients.

<table>
<thead>
<tr>
<th>Model</th>
<th>Salt &amp; Pepper</th>
<th>Blurring</th>
<th>Gaussian Noise</th>
<th>Blocking</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.595</td>
<td>0.738</td>
<td>0.762</td>
<td>0.667</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.929</td>
<td>0.738</td>
<td><strong>0.976</strong></td>
<td><strong>0.881</strong></td>
</tr>
<tr>
<td>PW-SSIM</td>
<td><strong>0.976</strong></td>
<td><strong>0.762</strong></td>
<td>0.976</td>
<td><strong>0.881</strong></td>
</tr>
</tbody>
</table>

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, UFFG/COPELE, IFPB and lecom for providing research support.

REFERENCES


