

Perceptual Quality Assessment Based on Visual Attention Analysis

Junyong You, Andrew Perkis
Norwegian University of Science and
Technology
Trondheim, Norway
+47 73592617

junyong.you@q2s.ntnu.no
andrew@iet.ntnu.no

Miska M. Hannuksela
Nokia Research Center
Tampere, Finland
+358 718008000

miska.hannuksela@nokia.com

Moncef Gabbouj
Tampere University of Technology
Tampere, Finland
+358 31153967

moncef.gabbouj@tut.fi

ABSTRACT

Most existing quality metrics do not take the human attention analysis into account. Attention to particular objects or regions is an important attribute of human vision and perception system in measuring perceived image and video qualities. This paper presents an approach for extracting visual attention regions based on a combination of a bottom-up saliency model and semantic image analysis. The use of PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural SIMilarity) in extracted attention regions is analyzed for image/video quality assessment, and a novel quality metric is proposed which can exploit the attributes of visual attention information adequately. The experimental results with respect to the subjective measurement demonstrate that the proposed metric outperforms the current methods.

Categories and Subject Descriptors

I.4.10 [IMAGE PROCESSING AND COMPUTER VISION]:
Image Representation

General Terms: Algorithms, measurement

Keywords: Perceptual quality assessment, visual attention, quality metric

1. INTRODUCTION

The advent of digital visual technology creates a need for automated methods for measuring the perceived quality. Although subjective evaluation is considered to reflect human perception in the most accurate way, it is time-consuming and cannot be done in real time. Moreover, the traditionally widely used metrics, namely MSE and PSNR, have been found not to be credible for measuring the perceived quality because they do not take the characteristics of the human visual system (HVS) into account.

To mimic the process of human vision and perception system for perceived quality, several HVS-based quality metrics have been proposed. Perceptual distortion metric (PDM), proposed by Winkler et al., is a representative example of psychophysical approaches [1]. However, the computation of the psychophysical

approaches is usually very complex due to the complexity of the process of the HVS. Therefore, a number of metrics following engineering approaches have been proposed. These kinds of metrics compare the quality features extracted from the reference signal and the distorted signal by taking certain attributes of the HVS into account. NTIA model proposed by Pinson et al. extracts the features measuring the spatial gradient activity, chrominance, contrast, and temporal information [2]. Based on the assumption that the HVS is highly adapted to extract structural information from the field of view, Wang et al. proposed that a measure of structural similarity (SSIM) can provide a good approximation for perceived image quality [3].

Despite the fact that visual attention is an important attribute of the HVS, it is, however, ignored in most existing quality metrics. Most of the current metrics consider the distortion in all sub-regions or pixels equally. Actually, many physiological and psychological experiments have demonstrated that human attention is not allocated equally to all regions in the visual field, but focused on certain regions known as salient regions [4]. Some tentative work has been done on integrating the human attention analysis into quality assessment. Lu et al. proposed a perceptual quality significance map to reflect the modulatory aftereffects of visual attention and evaluated its application in a Just-Noticeable-Difference (JND) model [5]. Based on the saliency attention model in [4], Feng et al. investigated a few weighting methods on the pixels in the salient regions for MSE, MAD, and SSIM metrics [6]. However, no appropriate metrics that can exploit the characteristics of human attention adequately have been proposed. In this study, we present an appropriate method for extracting visual attention regions from images and investigate adequately the influence of visual attention on the perceived image/video quality assessment. Then, an objective quality metric is proposed and evaluated with respect to the subjective quality measurement.

The rest of this paper is organized as follows. Section 2 presents the extraction of visual attention regions. The detailed explanation and discussions of visual quality metrics based on attention analysis are given in Section 3. Section 4 presents experimental results of the proposed metric with respect to the subjective measurement. Finally, some conclusions are drawn in Section 5.

2. VISUAL ATTENTION ANALYSIS

The studies of visual attention analysis can be divided into two categories: bottom-up and top-down approaches. In the bottom-up approach, a computational model for detecting the visual attention regions is constructed based on the low-level features of visual

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'09, October 19-24, 2009, Beijing, China.

Copyright 2009 ACM 978-1-60558-608-3/09/10...\$10.00.

signals. The top-down approach is usually driven by a certain task, such as searching for a color target, and the model is then built based on the visual features which are correlated with such task. The saliency-based visual attention proposed by Itti et al. [4] is a bottom-up approach, which is based on the map combination of color, intensity, and orientation information. In this paper, we used the SaliencyToolbox 2.1 developed by Itti et al. [7] to detect the salient regions of images and video frames and computed the corresponding saliency map value of each pixel in the salient regions. On the other hand, although the saliency attention model is able to detect the regions which can draw the attention of viewers, the semantic visual information is not taken into account. For example, viewers usually pay more attention to the face and text regions due to the fact that they can provide much information for understanding. Thus, we used the face and text detection algorithms to complement the attention region extraction. As the SaliencyToolbox can compute the saliency map values which express the saliency of each sub-region, and we assume that the face and text regions in an image or a video frame have the highest priority in quality assessment, the attention map values of face and text regions were set as the maximum of the saliency map values computed by the SaliencyToolbox. Figure 1 indicates an original image, detected attention regions (red line: face; blue line: text; yellow line: salient regions detected by SaliencyToolbox), and the corresponding attention map values.

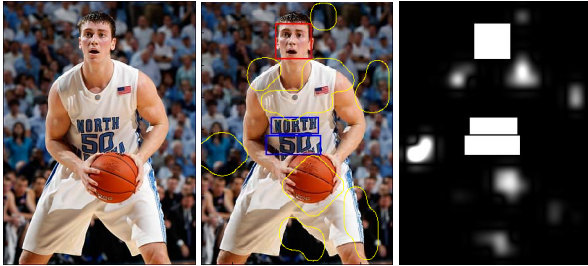


Figure 1. Original image, attention regions and map values.

3. VISUAL ATTENTION BASED QUALITY ASSESSMENT

3.1 Attention based PSNR and SSIM Metrics

To analyze the capability of visual attention analysis in measuring the perceived image/video quality, we first tested combination of the attention analysis into two widely used metrics, namely PSNR and SSIM. After detecting the visual attention regions in an image or a video frame, the corresponding attention map values, which can indicate the attentive degree information at each pixel, were also calculated. The metrics were computed in the attention regions using the following approaches: weighting the quality distortion values at each pixel by the above attention map values; not using weights; extracting the attention regions from the reference and test images; and using a weighted combination between luminance and chrominance information. The image areas outside of the attention regions were not included in the computation. We first tested image quality assessment with respect to the LIVE image database, in which 29 reference images with five distortion types were evaluated by a single-stimulus methodology [8]. A nonlinear regression operation between the metric results (VQ) and the subjective scores (MOS) was performed by the following logistic function:

$$MOS_p = a1/(1 + \exp(-a2 \cdot (VQ - a3))) \quad (1)$$

The nonlinear regression function was used to transform the set of metric values to a set of predicted MOS values, MOS_p , which were compared with the actual subjective scores and then result in two criteria: RMSE (Root MSE) and Pearson correlation coefficient. The analysis results are given in Table 1, where the optimal results among the different weighted linear combinations between luminance and chrominance were chosen, ori denotes the original PSNR and SSIM metrics, refe_w and dist_w denote the weighted results by using the attention map values in the attention regions extracted from the reference and distorted images, and refe_n and dist_n denote that the weights were not used. According to the analysis results, there is no evident performance improvement by integrating the attention analysis into image quality measurement. In our opinion, the reason might be that the viewers had enough time to look at all parts of the images when evaluating the image qualities, such that the influence of attention regions on the overall quality of whole image was not great.

Table 1. Analysis results of attention based PSNR and SSIM metrics for image quality assessment

Criteria		ori	refe_w	dist_w	refe_n	dist_n
PSNR	RMSE	8.67	8.13	7.92	7.65	7.64
	Pearson	0.91	0.92	0.92	0.93	0.93
SSIM	RMSE	5.66	5.78	5.87	4.92	4.95
	Pearson	0.96	0.96	0.96	0.97	0.97

Table 2. Analysis results of attention based PSNR and SSIM metrics for video quality assessment

Criteria		ori	refe_w	dist_w	refe_n	dist_n
PSNR	RMSE	1.24	0.87	0.89	0.96	1.01
	Pearson	0.48	0.72	0.73	0.62	0.63
SSIM	RMSE	0.79	0.80	0.81	0.83	0.81
	Pearson	0.76	0.76	0.75	0.74	0.75

To evaluate the capability of visual attention analysis in measuring video quality, 60 Temporal Scalability test clips and the corresponding subjective scores were used. The Temporal Scalability test was a subjective measurement to compare the performance of different temporal scalability parameters in the H.264/AVC codec. Four different video contents were employed: City (VGA, 8s duration), Ice (VGA, 10s duration), Foreman and Mobile (QVGA, 10s duration). For SSIM in video quality measurement, we first computed the SSIM values in each frame, and the Minkowski summation over all frames was calculated as the overall quality of the video sequence. In addition, because video sequence usually has strong temporal correlation, it is unnecessary to extract the attention regions in every frame. We extracted the attention regions once every 5 frames, and the analysis results demonstrated the performance in this way was even a little bit better than extracting attention regions in every frame. Table 2 gives the analysis results of video quality assessment. According to the analysis results on video quality assessment, the visual attention has prominent influence on the quality measurement. In contrast to the image quality assessment, viewers usually focus on the attention regions when evaluating the video quality. However, SSIM seems to be unsuitable for integrating with attention analysis. These analysis results motivated us to develop an effective metric which can make use of the attributes of visual attention for video quality assessment.

3.2 Visual Attention based Quality Metric

The quality features are derived for the attention regions by taking into account not only the luminance, but also the chrominance information; not only the spatial information, but also the temporal activity. First, a video sequence is divided into groups of pictures (GOP). Each GOP contains a certain number of frames with 50% overlap with the adjacent GOP. Such a GOP structure is able to express the fore-and-aft influence in human visual perception approximately. The GOP length is determined according to the temporal perception information (TI) defined as:

$$TI = \max_{time} \{std_{space}(F_n - F_{n-1})\} \quad (2)$$

where F_n denotes the n -th frame luminance image. If the temporal change of a sequence or a scene is big, then the GOP length will be short; otherwise, a longer GOP structure will be defined. The visual attention regions are extracted in the first frame of a GOP and the attention map values are also computed as the weights. The attention regions are then divided into spatial blocks with $N \times N$ (pixels) sizes, and N is determined based on the spatial perceptual information (SI) defined as:

$$SI = \max_{time} \{std_{space}[Sobel(F_n)]\} \quad (3)$$

where $Sobel(F_n)$ denotes the filtered result of F_n by Sobel filter. If the image content is complex, then the attention regions will be divided into smaller blocks; otherwise, bigger block sizes will be used. The quality features are computed for each spatial-temporal (S-T) block which is a set of image blocks located in the same position within a GOP. Our experiments demonstrate that appropriate selection of S-T block sizes can reduce the computing time greatly whilst keeping the performance satisfactory.

The first quality feature of an S-T block is defined as the PSNR of luminance component, marked as Y_PSNR . The second quality feature is to measure the chrominance degradation, which is defined as the Euclidean distance of the chrominance components (Cb and Cr) between the reference ($_r$) and distorted videos ($_d$):

$$C_Eu = 0.5 \cdot [mean(\|Cb_r, Cb_d\|) + mean(\|Cr_r, Cr_d\|)] \quad (4)$$

As many publications and subjective quality measurements have demonstrated that the most two prominent factors of visual quality degradation are blockiness and blurring, the third and fourth quality features are constructed to detect such two distortions. The video frames are first filtered by the vertical and horizontal Sobel filters, which result in two enhanced spatial gradients: SV in the horizontal direction and SH in the vertical direction. The polar coordinates of SH and SV are converted by:

$$\begin{cases} \rho = \sqrt{SH^2 + SV^2} \\ \theta = \text{atan}(SV/SH) \end{cases} \quad (5)$$

Then, two quality indices in an S-T block are defined to detect the change of spatial activity as:

$$\begin{cases} SA = std(\rho) \\ HV = mean(HV_1) / mean(HV_2) \end{cases} \quad (6)$$

$$\text{where } HV_1 = \begin{cases} \rho, \rho \geq 20, m\frac{\pi}{2} - 0.113 < \theta < m\frac{\pi}{2} + 0.113 & (m=0,1,2,3) \\ 0, & \text{else} \end{cases}$$

$$\text{and } HV_2 = \begin{cases} \rho, \rho \geq 20, m\frac{\pi}{2} + 0.113 \leq \theta \leq (m+1)\frac{\pi}{2} - 0.113 & (m=0,1,2,3) \\ 0, & \text{else} \end{cases}$$

The first quality index is sensitive to changes in the overall amount of spatial activity within an S-T block. For example, localized blurring usually produces a reduction in the amount of spatial activity, while noise causes an increase. The second index is sensitive to changes in the angular distribution or orientation of spatial activity. If the horizontal and vertical edges suffer more blurring than diagonal edges, HV of the degraded video will be smaller than HV of the reference. On the other hand, if erroneous horizontal or vertical edges are introduced, say in the form of blockiness or tiling distortion, then HV of the degraded video will be higher than that of the reference video. The third and fourth quality features are derived based on the ratio of these two indices between the reference and distorted video as follows, where R and D denote the reference and distorted videos, respectively.

$$\begin{cases} SI_Q = \min\left\{\frac{SA(D) - SA(R)}{SA(R)}, 0\right\} + \max\left\{\log_{10} \frac{SA(D)}{SA(R)}, 0\right\} \\ HV_Q = \min\left\{\frac{HV(D) - HV(R)}{HV(R)}, 0\right\} + \max\left\{\log_{10} \frac{HV(D)}{HV(R)}, 0\right\} \end{cases} \quad (7)$$

The last quality feature (TI_Q) is to detect the changes of temporal activity, such as the blocks or slices freeze and clearly incorrectly reconstructed blocks, both appearing because of packet losses and unsatisfactory error concealment. TI_Q is defined as follows for not only S-T blocks in the attention regions but also other regions, since the freeze or bad blocks have a great influence on the perceived quality even in non-attention regions.

$$TI_Q = \left\| \log_{10} \frac{f[F_n(D) - F_{n-1}(D)]}{f[F_n(R) - F_{n-1}(R)]} \right\| \quad (8)$$

where f denotes temporal masking function plotted in Figure 2.

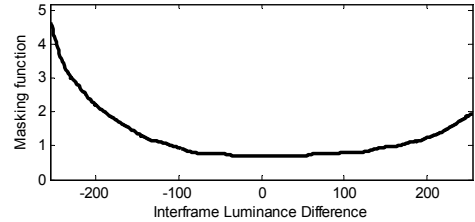


Figure 2. Temporal masking function.

After all above five quality features are computed, the first four features are averaged over all S-T blocks in the attention regions within a GOP, by using the attention map values as the weights, respectively. TI_Q of all S-T blocks in a GOP are also combined by the weights, where the attention map values are taken as the weights for attention regions, and the minimal value of the map values is taken as the weight for non-attention regions. In this way, the weighted averages of these five quality features are calculated for a GOP, and then the Minkowski summation over all GOPs is performed on these five quality features, respectively. In this study, the index of Minkowski summation is set as 2. Finally, a linear combination of these five quality features is computed as the overall perceived quality of the video sequence as following:

$$VQ = 0.0234 \cdot Y_PSNR - 0.3895 \cdot C_Eu + 1.6987 \cdot SI_Q - 0.4823 \cdot HV_Q - 0.5594 \cdot TI_Q \quad (9)$$

where the linear coefficients are derived by using the Levenberg-Marquardt algorithm on the training video sequences.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed video quality metric, a total of 392 video clips and the corresponding subjective measurements were evaluated, which included 320 VQEG FR-TV Phase I test clips [9], 60 Temporal Scalability test clips as explained in Section 3.1, and 12 Mobile test clips. The Mobile test was a double stimulus continuous quality scale measurement which employed three different contents with QCIF resolution to test the performance of H.264/AVC codec at four bit rates: 24, 32, 40, 48 kbps. Because the different subjective experiments usually have different rating scales, different testing conditions, and many other test variables that change from one laboratory to another, it is difficult to compare or combine results of two or more subjective experiments directly. Pinson et al. proposed an objective method for combining multiple subjective data sets which can map the multiple subjective data sets onto a single common scale using the iterated nested least squares algorithm (INLSA) [10]. We used the INLSA to map all the subjective results to the range of [0, 1].

When constructing the proposed metric, half of the reference video scenarios and their corresponding distorted clips were used in training the linear coefficients, and then the metric performance was evaluated with respect to the subjective measurements on the remaining sequences. To compare the performance of the proposed metrics, the NTIA general model in [2] was also computed on the same video sequences. As a traditionally and widely used metric, PSNR was taken as a benchmark. Furthermore, we also tested two other approaches, marked as A1 and A2, respectively. The first was to compute the 5 quality features in Section 3.2 in the attention regions while not using the attention map values as the weights; the second was to still use these 5 features while not considering the attention regions, i.e. these features were computed in the whole images. The linear coefficients in these two approaches were retrained to achieve the optimal performance. As we found there is no big difference in computing attention regions between reference and distorted videos, the results coming from attention regions in the reference videos are reported here. In addition, SSIM and PSNR weighted with the saliency map were also tested on these video sequences, and the results were similar to those in Table 2.

After getting the predictive qualities using these metrics on the remaining clips, the logistic function in Eq. (1) was used to fit the MOS mapped by INLSA and metric results. Pearson correlation coefficient and RMSE were selected to evaluate the prediction accuracy of the quality metrics. In addition, the quality model should predict a change in predictive quality that has the same sign as the change in subjective MOS, so Spearman rank-order correlation coefficient was also taken as an evaluation metric. Table 3 gives the evaluation results on these quality metrics.

Table 3. Evaluation results of visual attention based video quality metrics and other methods

Criteria	PSNR	NTIA	Proposed	A1	A2
RMSE	0.149	0.086	0.081	0.093	0.102
Pearson	0.57	0.88	0.90	0.85	0.81
Spearman	0.48	0.83	0.84	0.79	0.80

According to the evaluation results, the performance of the proposed metric is improved greatly compared with PSNR according to both RMSE and correlation, and the influence of visual attention on video quality assessment is as evident as we

expected, while the proposed quality features can be still improved since the results of A2 are not satisfactory enough. In addition, the proposed metric is better than the NTIA model, while our method has simpler computation. In addition, we also tested other two methods which are similar to [5] and [6] on the same video sequences. The first was to integrate the attention model in this work into a JND metric, but the performance was worse than the proposed metric. The second was to use Saliency-Toolbox only to extract the salient regions, then, MSE, MAD and SSIM were computed in the salient regions with different weights, but the performance was worse than that in Table 2.

5. CONCLUSIONS

The visual attention and its capability in perceptual quality assessment were analyzed in this paper, and an objective video quality metric which took the visual attention into account was proposed. Our analysis indicated that the visual attention is more suitable to video quality assessment rather than image quality measurement. Five effective quality features were developed based on a detailed analysis on visual quality degradation and an accurate metric was constructed based on these quality features. The experimental results with respect to the subjective measurements demonstrated the promising performance of the proposed metric compared with the existing methods. The future work is to find and design more suitable quality features and consider the temporal attention as well as the audiovisual attention analysis in the joint audiovisual quality measurements.

6. REFERENCES

- [1] Winkler, S. 2005. Digital video quality: vision models and metrics, John Wiley & Sons Press.
- [2] Pinson, M. and Wolf S. 2004. A new standardized method for objectively measuring video quality. IEEE Trans. Broadcasting, 50 (Sep. 2004), 312-322.
- [3] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing, 13 (Apr. 2004), 600-612.
- [4] Itti L. and Koch C. 2001. Computational modeling of visual attention, Nat. Rev. Neurosci., 2 (Mar. 2001), 194-203.
- [5] Lu Z., Lin W., Yang X., et al. 2005. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. IEEE Trans. Image Processing, 14 (Nov. 2005), 1928-1942.
- [6] Feng X., Liu T., Yang D., and Wang Y. 2008. Saliency based objective quality assessment of decoded video affected by packet losses. In Proceedings of IEEE Int. Conf. Image Processing (California, USA, Oct. 12-15, 2008), 2560-2563.
- [7] SaliencyToolbox 2.1, <http://www.saliencytoolbox.net>.
- [8] Sheikh H. R., Wang, Z., Cormack L., and Bovik A. C. LIVE Image Quality Assessment Database. <http://live.ece.utexas.edu/research/quality>.
- [9] VQEG Sequence, <ftp://ftp.crc.ca/crc/vqeg/TestSequences/>.
- [10] Pinson, M. and Wolf, S. 2003. An Objective Method for Combining Multiple Subjective Data Sets. In Proc. SPIE Video Communication and Image Processing Conf. (Lugano, Switzerland, Jul. 2003), 583-592.