Using Visual Attention in Video Quality Assessment

Cristina C. Oprea, Radu O. Preda
Dept. of Telecommunications
Politehnica University of Bucharest
Bucharest, Romania
cristina@comm.pub.ro, radu@comm.pub.ro

***Abstract*—This paper presents the analysis of a modified structural similarity index metric. First, we describe briefly the adaptations we brought. The main idea is to integrate a visual attention model into a metric for perceptual video quality assessment. In order to be able to evaluate our proposed modifications, we brought two other quality metrics: the reduced reference Video Quality Metric (VQM) and the full reference Perceptual Distortion Metric (PDM). Results and conclusions are mentioned in the end, following scatter-plots and correlation coefficients computation.**

Keywords - visual attention; quality evaluation; perceptual assessment.

I. Introduction

Deciding which areas in a given frame have a perceptual significance has several applications. some of the most important applications include: the optimization of the compression scheme in the video coding stage and more effective information hiding in images and video signals in watermarking schemes.. In such applications, the main characteristics and also the thresholds of the visual system can be used to obtain the best performance with respect to visual quality of the output. This paper presents a comparison between three widely used video quality evaluation metrics, all of them trying to integrate a computational model of the human visual system into a tool for quality assessment. Only one of these metrics makes use of a visual attention model by performing a detection of the perceptual important areas.

The paper is structured in five sections. Section II introduces the latest achievements in the area of perceptual region detection and human visual system modelling. Section III contains a detailed presentation of the proposed method, while in the forth section of this paper we briefly introduce the metrics used for comparison. The final section focuses on results and conclusions.

II. Previous Work

The visual assessment task may seem simple, but it actually involves a set of very complex mechanisms that are not completely understood. The visual attention process can be reduced to two physiological mechanisms that combined together result in a usual selection of perceptual significant areas from a natural or artificial scene. Those mechanisms are bottom-up attentional selection and top-down attentional selection. The first mechanism is an automated selection performed very fast, being driven by the visual stimulus itself. The second one is started in the higher cognitive areas of the brain and it is driven by the individual preferences and interests. A complete simulation of both mechanisms can result in a tremendously complex and time-consuming algorithm.

The process of finding the focus of attention in a scene is usually done by building feature maps for that scene, following the feature integration theory developed by Treisman [1]. This theory states that distinct features in a scene are automatically registered by the visual system and coded in parallel channels, before the items in the image are actually identified by the observer. Independent features like orientation, color, spatial frequency, brightness, and motion direction are brought together and analysed in order to find that single object or area being in the focus of attention. Pixel-based, spatial frequency, and region-based models of visual attention are different methods of building feature maps and extracting saliency.

The pixel-based category is represented by Laurent Itti's work concerning the emulation of bottom-up and top-down attentional mechanisms [2]. Another possibility of building feature maps is by applying different filtering operations in the frequency domain. The most common type of such filtering is done using Gabor filters and Difference of Gaussians filters. Meur et al. [3] apply the opponent color theory and use contrast sensitivity functions for high contrast detection. The last category of visual attention models are the region-based algorithms. In this case, a clustering operation is usually performed and then feature maps are computed using these clusters [4].

The main works in human visual system modelling have been studied and evaluated by the Video Quality Experts Group (VQEG), resulting in several video quality metrics that are given as standards at this moment. Such metric is VQM, described by American National Standards Institute (ANSI) in [5].

III. Proposed Algorithm With Attention Modelling

A. Attention model

The first SSIM algorithm developed for video sequences did not take into account in any way the fact that different spatio-temporal regions have specific levels of perceptual

ISBN: 978-1-61208-315-5

importance. Mean-SSIM simply determined an objective score for each frame and at the end of the distorted video sequence computes the mean of all scores. This mean value represented the objective score for the entire video.

Instead of that simple approach, we propose a combination between the visual attention model described in [4] that detects the salient regions in each video frame and the SSIM metric, which is used only for the previously extracted salient areas. This approach saves some computation effort, but not enough to be a major improvement. The positive outcome results come from the determination of the structural distortions only in the regions where the observer's attention is focused. The rest of the video slightly escapes from the natural quality assessment operation performed by the visual system, since it is known that our vision does not process the entire visual information.

The idea is to extract all spatio-temporal regions that get the most attention from the viewer. This operation is performed frame by frame and similar regions that belong to consecutive frames in a video scene will form a spatio-temporal region. Some areas come foreward in terms of attention because they contain some highly saturated colour, powerfull contrast or a particular object shape. In order to achieve less computational effort, we used only the chromatic contrast saliency map and the dimension saliency map presented in [4]. Both maps are combined using equal weights and the result is a general saliency map.

Extracting chromatic saliency begins with a segmentation procedure applied for all frames in the video sequence. The aim at this point is to obtain regions having one colour or similar colors. These regions do not necessarily need to occupy the same place in space as the objects present in the current frame. One object may correspond to several regions. At this point, one region R_i in a frame has the position parameters and the colour parameter: $R_i(p_{x,y},\ c_j)$. $p_{x,y}$ represent the vector coordinates for the main diagonal of the bounding rectangle and c_j is the j index in the colour palette. Regions corresponding in size, location, and colour from consecutive frames will form a spatio-temporal region, R_i^f, the maximum length for the temporal dimension being $f_{max} = 5$ frames. The number of colors in the color map depends on the chromatic dynamic in the video. Each frame region R_i will also have a set of pointers to the immediate N neighbors and a region perimeter factor, P_i, used to eliminate small unnoticeable patches or large background areas.

The chromatic contrast saliency map is computed considering several scenarios that determine the observer's attention to focus on a specific area. An important aspect of a scene that generates saliency is the color contrast which is differently defined from intensity contrast. Two colors that are situated on opposite sides on the hue color wheel are generating contrast when situated next to one another. There are five color contrast situations evaluated. For a current region R_i^f, each region having an opposite hue brings a score for R_i^f, as well as regions having a sufficiently distant hue. Third situation is the contrast due to the warm and cool colors. Fourth, the saturation contrast, comming from regions having colors with completly different saturations and fifth, the usual intensity contrast. The result will be a chromatic map corresponding to a set of maximum five frames, buid from separate values each corresponding to one region:

$$CM_i^f = \sum_{j=1}^{j=N} k_{op}^{i,j} k_d^{i,j} k_{wc}^{i,j} k_s^{i,j} k_I^{i,j} \left(1 + \Delta S_i^f + \Delta P_i^j\right)$$

A region size map is included in the color contrast map due to the presence of the perimeter factor ΔP_i^j, evaluated for de region R_i^f and its neighbor R_j^f.

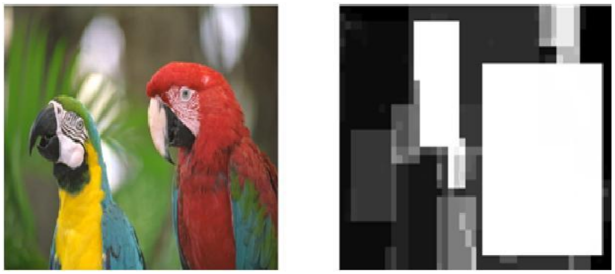


Figure 1. Saliency map for "parrots". Original image taken from [12].

B. *Modified SSIM metric*

The distorted sequence analysis is realised in three steps: at block level, frame level and at the entire video sequence level. Our modifications regard the first level, where blocks of 8x8 pixels are extracted from the salient regions in the previous map. This is completely different from the algorithm SSIM for static images described in [6], where a window was sliding pixel by pixel over the entire image. Our solution leads to less computational effort without accuracy loss, since we analyse only the frame parts that actually are viewed by the human observers. Classic SSIM algorithm for video sequences, without the attention model, is presented in [7], including the frame level and the entire video sequence processing level.

IV. Metrics For Comparison

A. *VQM*

VQM has been presented in detail in [9]. It has been developped for many types of video coding and transmission systems, beeing a reduced-reference metric. This method defines and computes several parameters for the original video sequence, using only a subset of spatio-temporal regions from this sequence. VQM needs a supplementary data of 14% from the transmission band filled by the uncompressed video sequence.

In VQM, the original and distorted video sequences need to be spatially realigned. Then, the algorithm tries to give an estimate for the regions in the distorted video corresponding to the regions from the reference video. The calibration step includes a process of computing the gain and amplitude offsets between the two video sequences and another process for temporal alignment. The objective VQM score is computed from the model quality parameters, which in turn depend on several perceptual features.

B. PDM

PDM stands for Perceptual Distortion Metric and was initially proposed by Winkler in [11]. It is a metric based on a model of the human visual system and needs the reference as well as the distorted videos. The main aspects integrated in PDM are the perception of colours and the opponent colour theory, multi-channel decomposition corresponding to the neural spatio-temporal mechanisms, contrast sensitivity, texture masking and the excitation/inhibition behaviour of the neurons in the primary visual cortex. The algorithm that we used is modified from the original one and has several adaptations. A complete presentation of the modified PDM metric that we used can be found in [12].

V. Results and Conclusions

In order to assess the metrics performances, we used a video sequences database developped by the LIVE laboratory and described in [10] and [11]. Each original video in this database has been subjected to four types of distortions and then subjectively assessed by a group of human observers. The subjectives scores called Difference Mean Opinion Score (DMOS) are then plotted against the objective scores estimated with each algorithm from the ones presented before. The results have been analyzed in Figures 1, 2, and 3 in the form of scatter-plots corresponding to a distortion type and a specific quality evaluation metric. The more condensed are the points in the scatter-plot forming a band or ideally a curve, the more accurate are the metric objectives estimates.

From Figures 1, 2, and 3, it is obvious that no metric performs as well for all distortion types. Although, it can be observed that scatter-plots for PDM and the proposed modified SSIM with attention modelling resemble more to a thin and long cloud. At this point, it is not easy to distinguish which metric performs better. Anyway, PDM is a very complex algorithm that needs a significant amount of computation resources. On the other hand, SSIM with attention modelling is more simple.

TABLE I. Correlation Coefficients

	VQM	SSIM-attention model	PDM
Pearson coefficient	0.73	0.70	0.85
Spearman coefficient	0.72	0.79	0.87

The prediction accuracy can be evaluated through the Pearson correlation coefficient, while the prediction monotony is analysed with the Spearman correlation coefficient. Their values are presented in Table 1 for the wireless distorted videos. It can be observed that our proposed modification for the SSIM has a good outcome, resulting in a greater monotony coefficient than VQM and a similar accuracy coefficient value. In terms of correlation values, PDM remains the best metric of all three serving as a model for comparison. The inconvenience with PDM is that it is not adaptable nor designed for actual use because of its complexity, while our proposed metric has the advantage of beeing practical.

Acknowledgment

This research activity was supported by UEFISCDI Romania under the grant PN-II-RU-TE-2011-3-0261 / No.10 / 2011.

References

[1] A. M. Treisman and G. Gelade, “A feature-integration theory of attention”, Cogn. Psychol., vol. 12, 1980, pp. 97-136.

[2] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention”, IEEE Trans. Image Process., 2004, vol.13, (10), pp. 1304 – 1318.

[3] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, “A coherent computational approach to model bottom-up visual attention”, IEEE Trans.Pattern Anal. Mach. Intell., 2006, vol. 28, pp. 802-817.

[4] C. Oprea, I. Pirnog, C. Paleologu, and M. Udrea, “Perceptual Video Quality Assessment Based on Salient Region Detection”, Proceedings of AICT 2009, May 2009, pp. 232-236.

[5] T1.801.03, American National Standard for Telecommunications – “Parameters for Objective Performance Assessment”. s.l. : American National Standards Institute, 2003.

[6] Z. Wang, L. Lu, and Al. C. Bovik, “Video quality assessment based on structural distortion measurement”, Signal Processing: Image Communication, 2004, Vol. 19 No.2, pp. 121-132.

[7] Al. Bovik, “Handbook of image and video processing”, s.l. : Elsevier Academic Press, 2005. ISBN:0121197921.

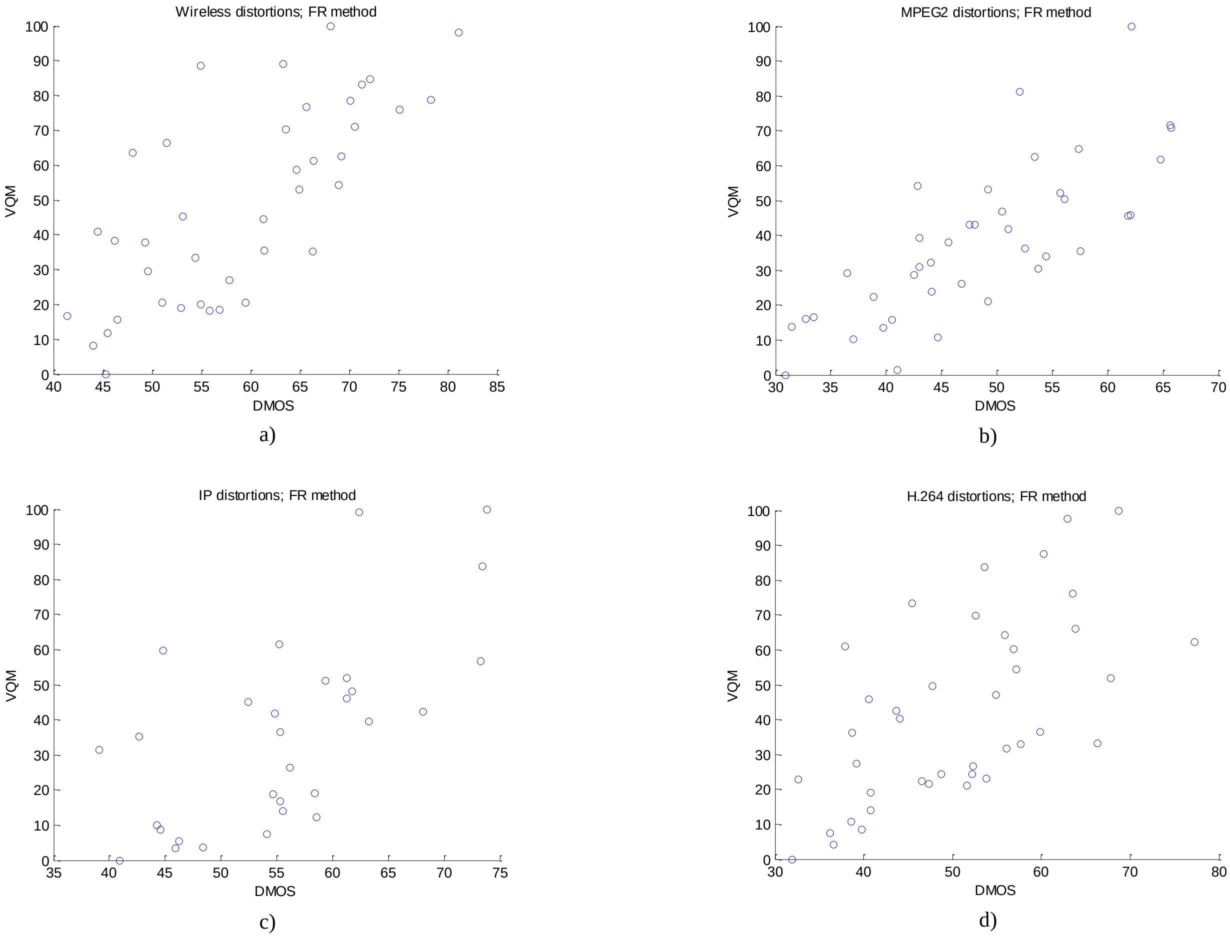
[8] S. Winkler, “Digital video quality: vision models and metrics”, s.l. : John Wiley & Sons, 2005.

[9] C. Oprea, C. Paleologu, I. Pirnog, and M. Udrea, ”Saliency detection making use of human visual perception modelling”, International Journal on Advances in Life Sciences, vol. 2, nr. 3&4, 2010, ISSN: 1942-2660, pp. 200-208.

[10] K. Seshadrinathan, et al. “A subjective study to evaluate video quality assessment algorithms”, s.l. : SPIE Proceedings Human Vision and Electronic Imaging, 2010.

[11] http://live.ece.utexas.edu/research/quality/live_video.html, accesses September 2012.

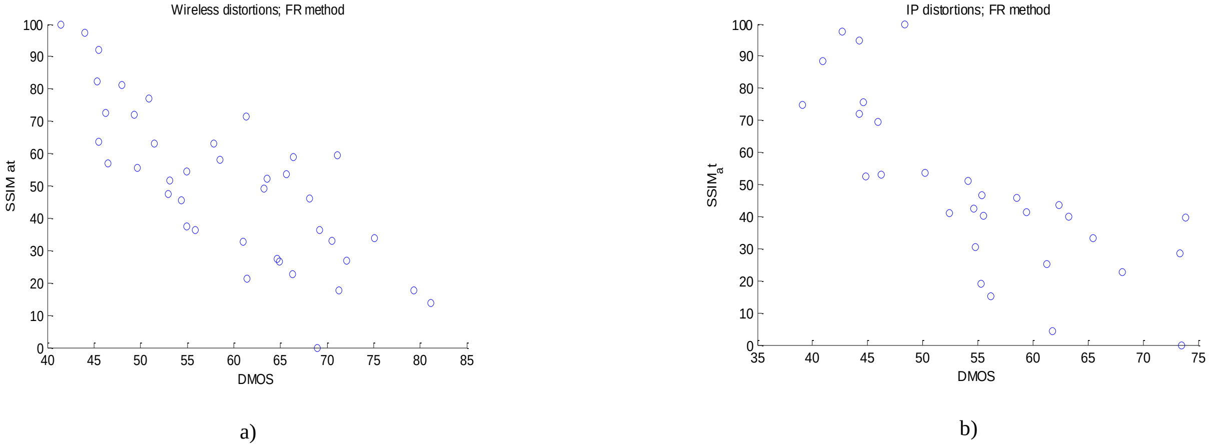
[12] G. Kootstra, A. Nederveen, and B. Leeds, “Paying Attention to Symmetry” UK. : Proceedings of the British Machine Vision Conference (BMVC2008), 2008. pp. 1115-1125.



a) b)

c) d)

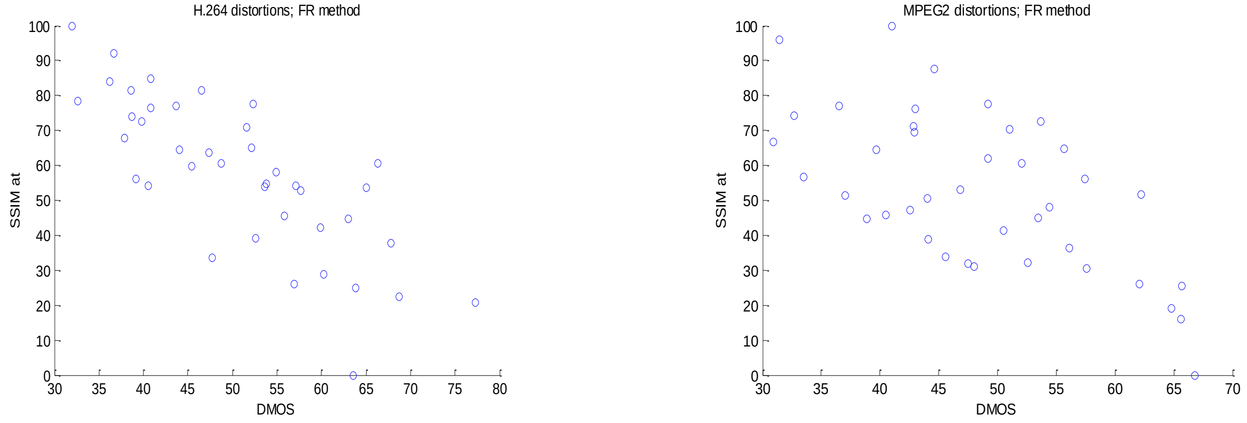
Figure 1. Scatter-plots obtained for the LIVE video database and VQM metric: a) videos with wireless distortions; b) MPEG2 distortions; c) IP distortions; d) H.264 distortions.



a) b)

Figure 2. Scatter-plots obtained for the LIVE video database and the proposed metric, SSIM with attention modelling: a) videos with wireless distortions; b) IP distortions.

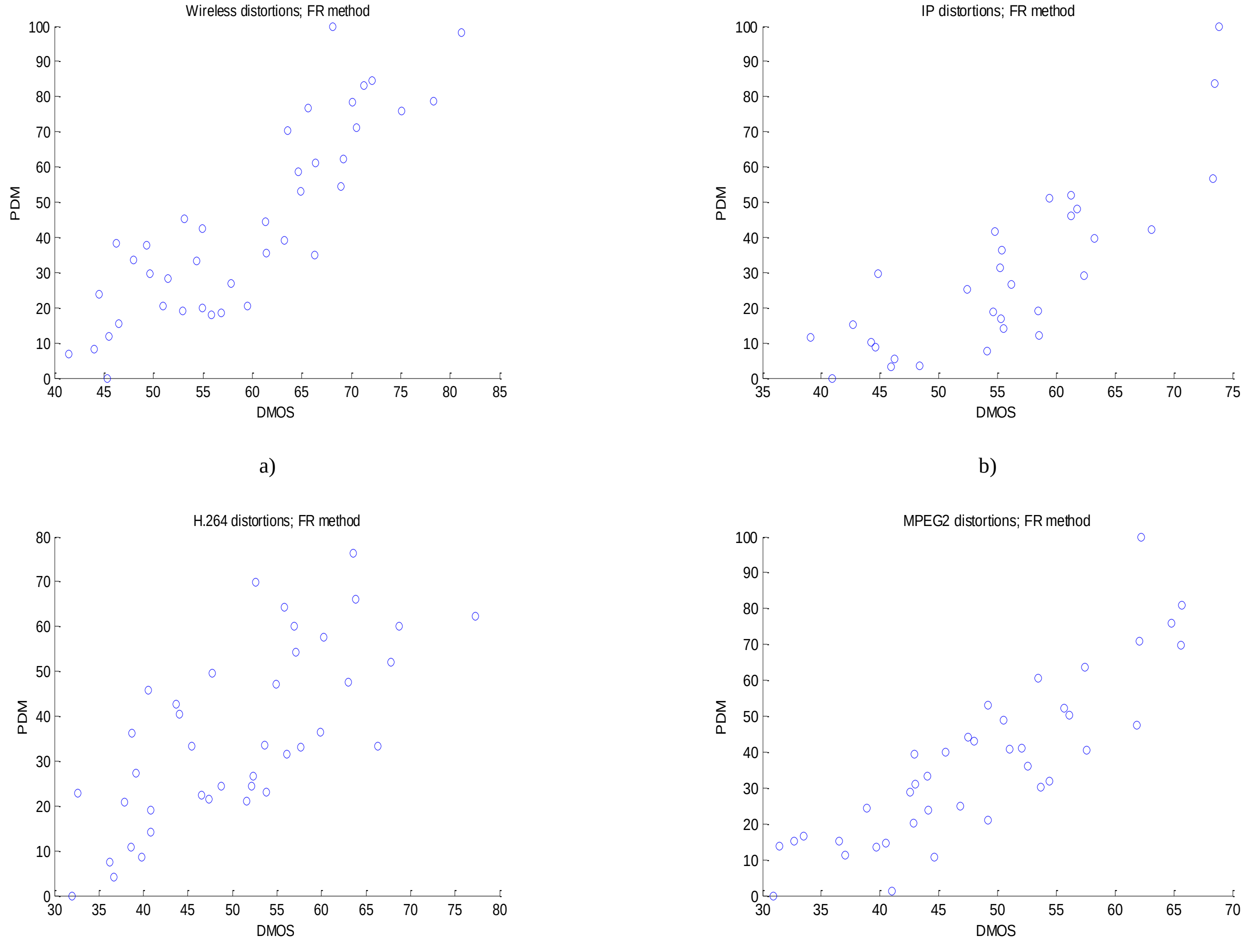
ISBN: 978-1-61208-315-5



a)

b)

Figure 3. Scatter-plots obtained for the LIVE video database and the proposed metric, SSIM with attention modelling: a) videos with H.264 distortions; b) MPEG2 distortions.



a)

b)

c)

d)

Figure 4. Scatter-plots obtained for the LIVE video database and the adapted PDM metric: a) videos with wireless distortions; b) IP distortions; c) H.264 distortions; d) MPEG2 distortions.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-315-5