# Weighted-MSE based on Saliency map for assessing video quality of H.264 video streams

Hugo Boujut, Jenny Benois-Pineau, Ofer Hadar, Toufik Ahmed, Patrick Bonnet

# Weighted-MSE based on Saliency map for assessing video quality of H.264 video streams

Authors: H. Boujut*, J. Benois-Pineau*, O. Hadar**, T. Ahmed*, and P. Bonnet***

*LABRI UMR CNRS 5800, Universite Bordeaux 1/IPB-Matmeca-Enseirb
351 cours de la Liberation 33405
Talence cedex - France
{boujut, benois-p, tad}@labri.fr

**Communication Systems Engineering Dept.
Ben Gurion University of the Negev
Beer Sheva, Israel, 84105
hadar@cse.bgu.ac.il

***Audemat WorldCast Systems Group
20, av Neil Armstrong, Parc d'activite J.F. Kennedy
33700 Bordeaux-Merignac – France
bonnet@worldcastsystems.com

## ABSTRACT

Human vision system is very complex and has been studied for many years specifically for purposes of efficient encoding of visual, e.g. video content from digital TV. There have been physiological and psychological evidences which indicate that viewers do not pay equal attention to all exposed visual information, but only focus on certain areas known as focus of attention (FOA) or saliency regions. In this work, we propose a novel based objective quality assessment metric, for assessing the perceptual quality of decoded video sequences affected by transmission errors and packed loses. The proposed method weights the Mean Square Error (MSE), Weighted-MSE (WMSE), according to the calculated saliency map at each pixel. Our method was validated trough subjective quality experiments.

## 1 INTRODUCTION

The TV content analysis on the consumer side is of the ever growing importance in the "digital society" nowadays. The consumer not only needs an intelligent and assisted access to the TV content, but the highest quality of the latter, specifically in the HD broadcasting. This is why the questions on quality measurement and assurance through the whole chains of broadcasting to the consumer end are of primarily importance for user satisfaction when navigating through broadcast digital content spaces. The goal of this paper is to contribute to objective video quality assessment of broadcasted video over DVB and IP networks. In this context, we present an objective video quality metric based on saliency map to assess packet loss influence on the perceived quality of broadcasted video streams. This new metric, we called it Weighted-MSE (WMSE), requires the full-reference video like MSE and SSIM [1] metrics. Unlike MSE which does not consider the Human Visual System (HVS), WMSE uses spatio-temporal saliency maps to increase the importance of salient regions. We note that in [2] the authors also propose similar ideas, but they work only on spatial saliency map and use magnitude-error weighting scheme. Despite the fact that SSIM metric and the approach proposed in [2] already take into account the HVS, the WMSE also considers the temporal side of visual perception. The WMSE metrics we propose is designed as a basis of comparison for quality of experience tests [3]. Furthermore, in this paper we contribute as well to a faster saliency map extraction method based on H.264 compressed stream information. The proposed quality metrics (WMSE) predicts perceptual quality of transmitted compressed video over error-channels. The metric has been tested on various video sequences compressed using the H.264 video compression standard with simulating some typical error models of IP networks [4] with different values of Packet Loss Ratios (PLRs). Results show that the proposed metric has better correlation with subjective quality compared to well known metrics such as MSE, PSNR and SSIM.

Several authors proposed methods to automatically predict, in terms of Mean Opinion Score (MOS) [5], visual quality of degraded compressed and transmitted video from objective quality measures. Hence, in [5] the authors predict MOS from calculated MSE using inverse linear relation between MOS and MSE defining the slope of the linear function on the basis of sequence edge strength per macro-block by averaging all macro-blocks in the video. Therefore, they use the intrinsic properties of a video in the pixel domain. Their study extents the known results that visibility of artifacts in highly detailed regions is lower than in lower detailed regions [6].

In [5] an objective full reference quality metric is proposed using content richness fidelity; block fidelity and distortion visibility for compressed video. The authors use natural decrease in entropy of decoded frame due to compression, vertical and horizontal artifacts due the blockiness effect and spatial and temporal masking properties of

human visual system. They tuned the parameters of their metrics by maximizing the Pearson correlation coefficient [7] with respect to the human visual subjective measures. Hence in their paper they try to take into account properties of HVS when defining objective quality metrics.

More in-deep integration of human perception of video content in designing objective video quality metrics has been recently made on the basis of definition of visual saliency maps [2]. This subject has been extensively studied for quality assessment of still coded images and coded videos. The variety of proposed methods can be split into two large categories, one is the spectrum methods and other is spatial domain methods. In spectral methods only the contrast and direction are taking into account, to model the sensitivity of HVS. This is what we can call a low level interpretation of visual quality. Since the last decade research in visual content understanding has focused more on semantic interpretation of visual scene by humans. Hence, in definition of saliency maps for visual quality assessment, pixel domain methods are more promising. These methods define visual saliency maps on the basis of human perception of motion and local color contrast in pixels domain [8], hence intrinsically simulating semantic interpretation of contents by humans. In [9] the authors derive temporal saliency maps from relative magnitude of local and global camera motion. Therefore, the perception of moving objects in a visual scene by HVS is intrinsically simulated. The saliency maps combing spatial and temporal saliency were proven to be a good approximation of visual saliency maps determined on the basis of psycho-visual experiments. Therefore the natural idea consists in exploring saliency maps for visual quality assessment of video with degradations due to the compression or packet loss during transmission. In [2] they use a spatial saliency map to weight standard objective quality metrics such MSE and SSIM. In this paper we introduced a new metric for video quality assessment which is based on weighting spatial-temporal saliency map. Being in frame work of H.264 HD compressed video transmission via "lossy" channels in this paper, we rather focus on the effect of channel errors than on the compression effect, therefore, we compressed the video streams to a high quality version, i.e., high bit rate, 6 Mbps.

The standard way to measure the goodness of an objective video quality metric consists in computing the Pearson correlation coefficient (PCC) between it and DMOS values [7]. Another contribution of this paper consists in using a supervised learning method for prediction the subjective DMOS from WMSE metric.

For the prediction of the subjective video quality we use a unique algorithm, which is based on a linear classifier with exponential similarity function [10]. As any supervised learning method, this algorithm uses two data sets: training and testing. The first collection of videos served as the training set, consisting of videos for collecting data of subjective rating and objective data (WMSE). The second test served for the evaluation of predicted values of DMOS on the basis of calculated WMSE. Our initial results show a promising way to predict the subjective video quality.

## 2 SALIENCY MAPS AND FOCUS OF ATTENTION (FOA)

The FOA is mostly attracted by salient areas which stand out from visual scene. These salient areas send stimulus to the HVS and sequentially grab the attention. In video scenes the salient stimuli are mainly due to high color, contrasts, motion and edge orientation (Figure 1). In the literature, the saliency of the visual scene is characterized by two saliency maps called "spatial" and "temporal" saliency maps. The spatial saliency map is based on color, contrast. The temporal saliency map is computed with the residual motion in the visual scene with regard to global, camera motion.

Figure 1.a Original frame

Figure 1.b Saliency map

Typically, the saliency map extraction process is performed in two steps. The first one is the extraction of the spatial and the temporal saliency maps. Then, the second step is the fusion of the both saliency maps: spatial and temporal ones.. The result of the fusion step is a spatio-temporal saliency map. Several methods already exist to predict saliency maps ( [8], [9]) from video and images. This is why we have used the algorithms presented in [9] for this paper. However, fusion

algorithms in the literature are usually very simple like the sum or the multiplication of both saliency maps. In the section 3, we propose a new fusion method based on a weighted sum of the saliency maps logarithms.

### 2.1 Spatial saliency map

The spatial saliency map (Figure 2) extraction described in [9] is based on the sum of 7 color contrast descriptors in the HSI domain:
saturation contrast, intensity contrast, hue contrast, opposite color contrast, warm and cold color contrast, dominance of warm colors and dominance of brightness and hue. The seven descriptors $V_\delta$ are computed for each pixels $s_i$ of a frame $i$ using the 8-connected neighborhood. Then, to get the final spatial saliency map $S^{SP}$, the 7 descriptors are combined for each pixel $s_i$ with (1).

$$S^{SP}(s_i) = \frac{1}{7}\sum_{\delta=1}^{7} V_\delta(s_i) \tag{1}$$

Finally, $S^{SP}$is normalized (2) between 0 and 1 according to its maximum value $S_{max}$.

Figure 2 Spatial saliency map (tractor sequence)

$$S^{SP'}(s_i) = S^{SP}(s_i)/S_{max} \tag{2}$$

### 2.2 Temporal saliency map

We compute our temporal saliency map in the manner similar to [9]. In [9], the temporal saliency map is extracted in four steps. First of all, the optical flow is computed for each pixel $s_i$ of the frame $i$, then the motion is accumulated in $\overrightarrow{V_\theta}(s_i)$. In our case we do not compute optical flow but extract motion vectors from H.264 encoded stream as we will explain this in section 3. Secondly, they estimate global motion $\overrightarrow{V_G}(s_i)$ with collection of evidence approach. We estimate it with the robust estimator we developed in context of rough indexing paradigm [11] to find the 6 affine parameters model. Then, the residual motion $\overrightarrow{V_R}(s_i)$ is computed by (3).

$$\overrightarrow{V_R}(s_i) = \overrightarrow{V_\theta}(s_i) - \overrightarrow{V_G}(s_i) \tag{3}$$

Finally, the temporal saliency map $S^T(s_i)$ is computed by filtering the amount of residual motion in the frame. S. Daly [12] has established that the human eye cannot follow objects with a velocity higher than 80 deg./s. In this case, the saliency is null. S. Daly has also demonstrated that the saliency reaches its maximum with motion values between 6 deg./s and 30 deg./s. According to this psycho-visual constraints, the filter proposed in [9] is given by (4).

$$S^T(s_i) = \begin{cases} \frac{1}{7}\overrightarrow{V_R}(s_i) & if \quad 0 \le \overrightarrow{V_R}(s_i) < \vec{v}_1 \\ 1 & if \quad \vec{v}_1 \le \overrightarrow{V_R}(s_i) < \vec{v}_2 \\ \frac{1}{60}\overrightarrow{V_R}(s_i) + \frac{8}{5} & if \quad \vec{v}_2 \le \overrightarrow{V_R}(s_i) < \vec{v}_{max} \\ 0 & if \quad \overrightarrow{V_R}(s_i) \ge \vec{v}_{max} \end{cases} \tag{4}$$

with $\vec{v}_1 = 6\ deg./s$, $\vec{v}_2 = 30\ deg./s$ and $\vec{v}_{max} = 80\ deg./s$.
Hence we followed this filtering scheme in temporal saliency map computation.

### 2.3 Saliency map fusion

In the literature, the common fusion method is the sum of the temporal and the spatial saliency map weighted by a 2 dimensional Gaussian centered at the centre of the frame. In [9] they have established that for HD TV content the best Gaussian spread is 5 visual degrees. A comparison of fusion methods in [8] has shown that the multiplication of saliency

maps gives better results than the additive fusion. In this paper, we use the multiplication fusion method $S_{mul}^{SP-T}(s_i)$ weighted with a 5 visual deg. 2D Gaussian $2DGauss(s)$ (5) to compare our fusion method proposed in sec. 3.2.

$$S_{mul}^{SP-T}(s_i) = S^{T}(s_i) \times S^{SP}(s_i) \times 2DGauss(s) \quad (5)$$

In the next section we propose an enhanced method to extract the temporal saliency map which takes advantage of the H.264 compressed stream and a new fusion method.

## 3 SPATIO-TEMPORAL SALIENCY MAP DERIVATION FROM PARTIALLY DECODING H.264 VIDEO STREAMS

When disposing an encoded stream at the decoder end, the very seducing idea consists in the re-use of motion and spatial information already embedded in the stream and to compute the spatio-temporal saliency map without full decoding of it thus saving computational workload. Unfortunately the use of spatial information in Integer bloc transform domain and specific spatial prediction schemes in H.264 is not so straightforward as in MPEG2 stream [13]. Despite new methods has recently appeared for re-covering approximate spatial information without full decoding of H.264 streams, this remains a challenge for precise contrast estimation. Hence to compute spatial saliency map we have to decode the video frames and then to apply the method [9] described in section 2.1.

The gain in performance can be obtained in temporal map extraction. Indeed, the quality of motion vectors in H.264 has been drastically improved due to the possibility of splitting blocks up to very small sizes (4x4 pixels). Furthermore, the motion estimators used in conventional encoders today (diamond search and full search) ensure better quality of motion field. Last but not least, the robust global motion estimation scheme with outlier rejection we developed in [11] is a powerful tool for a proper estimation of affine 6 parameter global model.

Hence for extraction of temporal saliency map, we propose a new method (Figure 3) operating on the compressed stream. It starts with decoding motion vectors from H.264 stream. Due to the fact that H.264 encoder allows from multiple reference frame, we cumulate motion vectors for a given pixel in the current frame up to the same reference frame – the most recent IDR. Then a motion vector characterizes a pixel normalized by the time distance to the IDR. Another problem we address is filtering of flat areas. Indeed due to the ill-posed problem of motion estimation, the bit stream contains erroneous vectors on flat areas. We remove these improper values by detecting flat area located in the background.

Coming into the very early stage of digital image processing to fundamental works by Azriel Rosenfeld [14], “the background” component of a visual scene touches at least one of the borders. Thus we apply a region growing algorithm which starts from the borders of the image and stops when the norm of the gradient is not null (Figure 4).

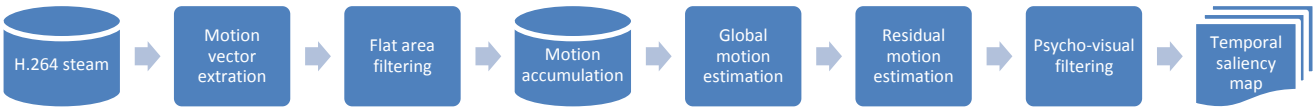
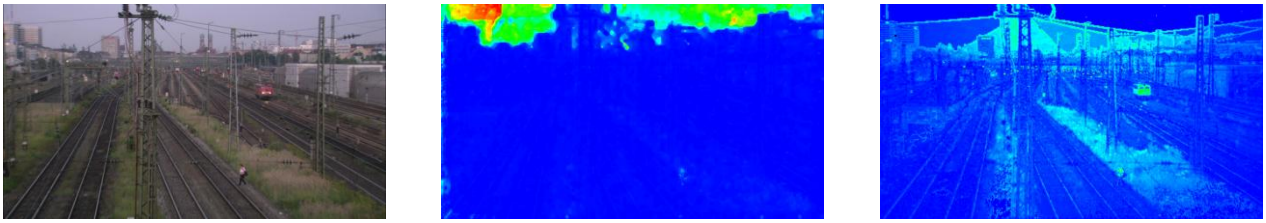



Figure 3 Temporal saliency map extraction process



a. Original frame b. Temporal saliency map before filtering c. Temporal saliency map after filtering

Figure 4 Flat area filtering (station2 sequence)

Finally, to produce the spatio-temporal saliency map we also propose a new fusion method $S_{LOG}^{SP-T}$ defined by (6) with $\alpha = 0.5$. This new method has the same advantage as $S_{mul}^{SP-T}$ that gives more importance to regions which have both high spatial saliency and high temporal saliency. Unlike $S_{LOG}^{SP-T}$, $S_{mul}^{SP-T}$ provide null spatio-temporal saliency maps when the temporal saliency is very low.

$$S_{LOG}^{SP-T}(s_i) = \alpha \log(S^{SP}(s_i) + 1) + (1 - \alpha) \log(S^{T}(s_i) + 1) \tag{6}$$

## 4 WEIGHTED MSE BASED ON SALIENCY MAPS

Considering that the visual attention is focused on the saliency map would suppose that video transmission artifacts would be more annoying in these areas. However, transmission artifacts may change the saliency map, hence we propose to extract the saliency map from the already broadcasted disturbed video stream. WMSE metric is given by (7) where $\aleph\gamma$ is the reference frame, $\hat{\gamma}$ the disturbed frame and $W_{\hat{\gamma}}$ the spatio-temporal saliency map of $\hat{\gamma}$, E is the mathematical expectation operator.

$$WMSE(\hat{\gamma}) = E\left[W_{\hat{\gamma}} * (\hat{\gamma} - \gamma)^2\right] \tag{7}$$

Hence, to get a WMSE value for a video sequence, we propose to compute for each video sequence $K$ the average WMSE (8). Where $T_K$ is the number of frames in the video sequence, $W(t)$ the spatio-temporal saliency map, $I_d(t)$ the disturbed frame and $I_0(t)$ the reference frame.

$$\overline{WMSE}_K = \frac{1}{T_K} \sum_{t=1}^{T_K} \left[W(t) \times \left(I_d(t) - I_0(t)\right)\right]^2 \tag{8}$$

## 5 PREDICTION OF DMOS FROM WMSE

### 5.1 Subjective experiment

We carried out subjective experiments to measure the quality of HDTV transmitted over lossy networks. To get more participants and more reliable results, the experiment was done in two research laboratories: LaBRI (University of Bordeaux) and Communication Systems Engineering Dept. (Ben Gurion University of the Negev (BGU)). Twenty different video sequences of 10 seconds were selected to compose a representative sample of broadcasted HDTV programs. The selection of video sequences was done according to two features called spatial and temporal information, described in ITU-T Rec. P.910 [15]. Video sequences come from four different corpuses: The Open Video Project [16], NTIA/ITS [17], TUM/Taurus Media Technik [18] and French HDTV. According to copyrights, video sequences from the French HDTV corpus are not available outside France. Video sequences were encoded into the H.264/AVC format [19] using the x264 [20] software with a bit-rate of 6000kb/s. Two models of transmission impairments were applied to each video sequences (Table 1). The first one, we called it IP model, simulates IP packet networks according to ITU-T Rec. G.1050 [4]. Hence, three kinds of networks: managed, semi-managed and unmanaged were simulated using five packet loss profiles. The second model, we called it RF model, simulates radio frequency transmission impairments by introducing bit corruption in Transport Stream (TS) packets. To simulate the RF model, three levels of bit corruption were chosen. After processing the 20 video sources (SRC) with the 8 impairment profiles, 160 processed video sequences (PVS) were generated. So, the total number of video sequences assessed by the experiment participants was 180.

| | Profile | Loss | Burst |
|---|---|---|---|
| **IP Model** | 0 | 0.05% | No |
| | 1 | 1% | No |
| | 2 | 1% | Yes |
| | 3 | 5% | No |
| | 4 | 5% | Yes |
| **RF Model** | 5 | 0.01% | No |
| | 6 | 0.1% | No |
| | 7 | 1% | No |

Table 1 Loss profiles

Figure 5 Experiment room

The experiment was carried out by following the ACR-HR experimental protocol described in the VQEG Report on the Validation of the Video Quality Models for High Definition Video Content [7]. The experiment room (Figure 5) and the lightning conditions were compliant with the ITU-R Rec. BT.500-11 [21]. The distance between the subject head and the screen was three times the height of the screen. The video sequences were displayed with a resolution of 1920x1080 pixels using a HDMI cable. In order to be compliant with ITU-R Rec. BT500.11, the experimentation time was reduced to 30 minutes by splitting the video dataset in two parts. Therefore, each participant has seen only 90 videos, i.e. 10 SRC with the 8 related PVS. Moreover, twice more participants were required to carry out the whole experiment. The experiment was done with the two video sub-datasets at LaBRI and one video sub-dataset at BGU. Due to copyrights, the sub-dataset used at BGU was not composed from French HDTV videos. To avoid the "leaning effect" each participant has seen the video sequences in a unique order and a "warm-up" session of 5 minutes was done before starting the experiment. Hence, 22 participants were gathered at LaBRI, i.e. 11 for each sub-set and 13 at BGU. MOS and DMOS subjective metrics were computed by using methods described in [7] and [21].

### 5.2 Prediction method

In this section, we propose to use supervised learning method called similarity-weighted average to predict DMOS values from WMSE or MSE. This prediction method requires a training data set of $n$ known pairs $(x_i, y_i)$ to be able to predict $y$ from $x$. Here $(x_i, y_i)$ pairs are WMSE or MSE values associated with DMOS values from the subjective experiment. $y$ is the predicted DMOS from a given WMSE/MSE $x$. The prediction is performed using equation (9) known as a weighted mean classifier and (10).

$$y = \frac{\sum_{i=1}^{n} s(x_i, x) y_i}{\sum_{i=1}^{n} s(x_i, x)} \tag{9}$$

$$s(z, x) = exp[-|x - z|] \tag{10}$$

In the original paper [10] the authors show good generalization properties due to the monotonicity of the exponential similarity measure (10), this was a reason for us to choose this prediction scheme. The other reason is that it does not require a heavy training as it is the case of many classifies such as Neuronal Networks and SVMs.

In the VQEG Report on the Validation of the Video Quality Models for High Definition Video Content, they propose to use a cubic polynomial function (11) to map the WMSE/MSE values $x$ to the DMOS $y$. In the next section we compare the performance of those two prediction methods.

$$y = ax^3 + bx^2 + cx + d \tag{11}$$

### 5.3 Results and evaluation

In this section, we compare three objective video quality metrics: MSE, the proposed method WMSE using the multiplication fusion ($WMSE_{mul}$) and the log sum fusion ($WMSE_{log}$) with the results of the subjective experiment described in section 5.1. For all the 160 PVS of the subjective experiment, a DMOS value is computed. The similarity-weighted method and the cubic polynomial function are used to predict the DMOS. Hence, to train and evaluate the prediction methods, a dataset of 160 data pairs MSE/DMOS, $WMSE_{mul}$/DMOS or $WMSE_{log}$/DMOS is built for each objective metric. Then, each dataset split into two equal parts, one part is used for training the prediction method and the other is used for the evaluation. The evaluation is performed by computing the PCC (12) denoted by R

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})}} \tag{12}$$

where $x_i$ is the DMOS, $y_i$ the predicted DMOS and $N$ the number of data pairs in the evaluation dataset. The PCC results for the three objective metrics and the two prediction methods are given in table 2.

| Methods | similarity-weighted average | cubic polynomial similarity function |
|---|---|---|
| **MSE** | 0,9917 | 0.6394 |
| **$WMSE_{log}$** | 0,9235 | 0.4161 |
| **$WMSE_{mul}$** | 0,7151 | 0.3311 |

Table 2 PCC between objective metrics and prediction methods

A cross-validation algorithm was also applied to appreciate the similarity-weighted average prediction stability. It showed the PCC are stable with a standard deviation of 0.0033 for MSE, 0.0181 for $WMSE_{log}$ and 0.0325 for $WMSE_{mul}$. The results show that the proposed fusion method $WMSE_{log}$ gives better performance than $WMSE_{mul}$. Results also shows that $WMSE_{log}$ has close results to the reference metric MSE. Moreover, the similarity-weighted average prediction method provides very good results for the prediction of DMOS compared to the traditional method based on a cubic polynomial similarity.

## 6 CONCLUSION

Hence in this paper we were interested in the problem of objective assessment of subjective quality of video scenes transmitted over lossy channels. We followed the recent trends in the definition of spatio-temporal saliency maps for FOA and derived a new full reference quality metric: the WMSE based on saliency map. We brought some new solutions for saliency map computation, the main contribution being the re-use of H.264 motion vectors for temporal saliency definition. Furthermore, we proposed a new logarithmic fusion of spatial and temporal saliency maps into the global spatio-temporal saliency map. Finally, instead of polynomial fitting for prediction of subjective quality metric DMOS from objective quality metrics, we proposed to use a supervised learning approach with a weighted average classifier and exponential similarity measure. Due to the monotonicity of this measure, the prediction quality is better than with a conventional approach for quality assessment.

As in this work we were interested in quality assessment for video due to the transmission artifacts and not compression artifacts, we built a significant video corpus with degradations according to state-of-the-art models of packet loss and RF loss and tested our approach on this data set.

The first results are promising, with the new quality metric for full-reference scheme we explored. We intend to improve the saliency model in order to better consider transmission artifacts.

We plan to use these first results for the no-reference QA we are working on now and obviously to collect more experimental data on subjective quality assessment.

# 7 REFERENCES


[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[2] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency Based Objective Quality Assessment of Decoded Video Affected by Packet Loss," *ICIP*, pp. 2560-2563, 2008.

[3] ITU-T, "New Appendix I - Definition of Quality of Experience (QoE)," *G.100/P.10 Amendment 1*, Jan. 2007.

[4] International Telecommunication Union, "ITU-T Rec. G.1050 Network model for evaluating multimedia transmission performance over Internet Protocol," Recommendation, 2007.

[5] A. Bhat, I. Richardson, and K. Sampath, "A new perceptual quality metric for compressed video," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 933-936, Apr. 2009.

[6] E. P. Ong, W. Lin, L. Zhongkang, S. Yao, and M. H. Loke, "Perceptual Quality Metric for H.264 Low Bit Rate Videos," *IEEE International Conference on Multimedia and Expo*, pp. 677-680, Jul. 2006.

[7] VQEG (Video Quality Experts Group), "Report on the Validation of Video Quality Models for High Definition Video Content," Report, 2010.

[8] S. Marat, et al., "MODELLING SPATIO-TEMPORAL SALIENCY TO PREDICT GAZE DIRECTION FOR SHORT VIDEOS," *IJCV*, no. 82, pp. 231-243, Mar. 2009.

[9] O. Brouard, V. Ricordel, and D. Barba, "Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif," *CORESA*, Feb. 2009.

[10] A. Billot, I. Gilboa, and D. Schmeidler, "Axiomatization of an exponential similarity function," *Mathematical Social Sciences*, no. 55, pp. 107-115, 2008.

[11] M. Durik and J. Benois-Pineau, "Robust Motion characterisation for video indexing based on MPEG2 optical flow," *CBMI*, pp. 57-64, Sep. 2001.

[12] S. Daly, "Engineering Observations from Spatio-velocity and Spatiotemporal Visual Models," *IS&T/SPIE Conference on Human Vision and Electronic Imaging III*, vol. 3299, pp. 180-191, Jan. 1998.

[13] F. Manerba, J. Benois-Pineau, R. Leonardi, and B. Mansencal, "Multiple Moving Object Detection for Fast Video Content Description in Compressed Domain," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, p. 15, Aug. 2008.

[14] A. Rosenfeld, "Digital Topology," *The American Mathematical Monthly*, vol. 86, no. 8, pp. 621-630, Oct. 1979.

[15] International Telecommunication Union, "ITU-T Rec. P.910 Subjective video quality assessment methods for multimedia applications," Recommendation, 1999.

[16] The Open Video Project. (2010, Nov.) LABRI-ANR ICOS-HD. [Online]. http://www.open-video.org/collection_detail.php?cid=23

[17] NTIA/ITS. (2010, Nov.) VQEG FTP - NTIA source. [Online]. ftp://vqeg.its.bldrdoc.gov/HDTV/NTIA_source/HDTV_Readme.doc

[18] TUM / Taurus Media Technik. (2010, Nov.) HD test sequences Taurus Media Technik. [Online]. ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/1080p/ReadMe_1080p.txt

[19] ISO/IEC, "Advanced Video Coding," in *Information technology - Coding of audio-visual objects*, 2004, ch. Part 10.

[20] Videolan. (2010, Nov.) x264 - a free h264/avc encoder. [Online]. http://www.videolan.org/developers/x264.html

[21] International Telecommunication Union, "ITU-R BT.500-11 Methodology for the subjective assessment of the quality of television pictures," Recommendation, 2002.