NR Video Quality Metric with Visual Saliency from H.264 Code Stream in Lossy Network Channels

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

NR Video Quality Metric with Visual Saliency from H.264 Code Stream in Lossy Network Channels

H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, P. Bonnet

Abstract*—The paper contributes to video quality assessment of delivered content without reference. We propose a no-reference video quality assessment metric taking into account the behavior of the human visual system. The proposed metric called Weighted Macro-Block Error Rate (WMBER) is based on macro-block error detection and is weighted by visual saliency maps. Both measures are extracted from a partially decoded H.264 AVC stream. First of all, we propose a new saliency map fusion method to improve the spatiotemporal saliency model. Then a supervised learning method called Similarity Weighted Average is considered to predict subjective MOS from objective video quality metric. The Similarity Weighted Average method is improved in order to be adapted to a training database or a content. The performance of the proposed metric is evaluated on two subjective experimental databases from LaBRI and IRCCyN. The results are compared with two Full-Reference metrics MSE and SSIM. The evaluation shows that the proposed metric provides an accurate prediction of subjective measures.

***Index Terms*—No-reference, video quality assessment, saliency map, supervised learning, visual attention, H.264**

I. Introduction

With the introduction of HDTV Broadcast over DVB-T/S 1&2 [1], and the wide use of IPTV services by Internet Service Providers, the quality assessment of broadcasted video services became an important research topic for both academia and industries. The research on quality assessment was emphasized in the last few years by a need for optimization of bandwidth resources allocation, better system design, and optimal geographical positioning of broadcast equipments. The quality assessment of the delivered HD content would satisfy user requirements and enhance its quality of experience (QoE). Video quality assessment (VQA) was required due to the introduction of lossy video coding standards at the beginning of the 80's. As adopted in ITU-T recommendation [2], the majority of VQA techniques which were proposed so far, had only tackled the degradations induced by encoding process. Today, HD delivery raises new challenges such as how to objectively assess the quality of impaired video stream at the decoder side which may suffer from signal degradation and packet errors. Transmission errors generate strong visual degradations due to simple error resilience mechanisms used in the end-to-end delivery chain. These mechanisms are implemented in typical industrial decoders of the actual HDTV standard H.264/AVC [3]. As delivered HD video is to be perceived by the human vision system (HVS), we believe that quality assessment mechanism has to be designed accordingly to user perception and experience. Towards this objective, we propose an accurate definition of saliency maps in video scenes [4]. Furthermore, our proposed objective quality metric takes into account the loss of blocks in H.264 encoded HD stream during delivery. Many actors are involved in the broadcast delivery chain. For this reason, we believe that a no-reference (NR) VQA scheme will be useful to evaluate transmitted video streams. We also think that a NR VQA scheme will have an important impact on different stakeholders involved in the delivery chain, from content producers to content consumers passing through network operators and service providers.

A very large and extensive research was conducted for full reference (FR) VQA. Its outcomes can be successfully used for designing efficient NR VQA metrics. Specifically, spatial VQA metrics were exhaustively studied [5], [6], [7] on the basis of modeling the sensitivity of HVS. These models of visual attention include local color contrast orientation and global statistics of features in a given image, or only in a Region of Interest (ROI) [6]. A rather complete analysis of these approaches is given in [8]. In [9], [10], and [11], another approach considering motion and its variations along the elapsed time is used for weighting spatial quality indices. This approach is also explicitly incorporated in visual saliency maps for weighting full reference distortion metrics resulting in Weighted Mean Squared Error (WMSE) [12], as well in other objective (FR) weighted metrics such as Weighted Peak Signal-to-Noise Ratio (WPSNR) referenced in [13]. The NR video quality assessment research has made a significant progress and as explained in [13], a general motion-tuned spatiotemporal framework is now proposed for NR VQA. This

Manuscript received January 9, 2012.

H. Boujut is with the LaBRI UMR CNRS 5800 University of Bordeaux, 351 cours de la Liberation 33405 Talence cedex, France (phone: +33 5 40003880; fax: +33 5 40006669; e-mail: boujut@labri.fr).

J. Benois-Pineau is with the LaBRI UMR CNRS 5800 University of Bordeaux, 351 cours de la Liberation 33405 Talence cedex, France (e-mail: benois-p@labri.fr).

T. Ahmed is with the LaBRI UMR CNRS 5800 University of Bordeaux, 351 cours de la Liberation 33405 Talence cedex, France (e-mail: tad@labri.fr).

O. Hadar is with the Communication Systems Engineering Dept. of Ben Gurion University of the Negev, Beer Sheva 84105, Israel (e-mail: hadar@cse.bgu.ac.il)

P. Bonnet is with Audemat Worldcast Systems, 20, avenue Neil Armstrong, Parc d'activite J. F. Kennedy 33700 Bordeaux-Merignac, France (e-mail: bonnet@worldcastsystems.com).

framework incorporates spatial quality and temporal quality indexes which are computed from the decomposition of video signal in Gabor wavelet domain. By a multiplicative pooling, spatial and temporal quality indexes give a global quality index which is called MOVIE index. Another complete framework for spatiotemporal degradations assessment is the TetraVQM algorithm [14]. Here, the authors introduce temporal tracking of degradations on moving objects, imitating in this way user behavior in the perception of video content: Humans focus their attention on degraded objects. The elapsed time during which they gaze at degraded objects, is also modeled. The process results in tree maps: spatial degradations, observation time and reliability, which then are combined via multiplicative pooling. The main principles of TetraVQM are improved in [15] with more sophisticated saliency modeling. Remaining in the framework of FR quality assessment, the authors propose to decrease a VQA metric by a weighted saliency map. In their paper, the authors use a saliency map which was created from gaze patterns to avoid the possible errors in the prediction of saliency.

The NR VQA metrics still remain a research challenge and contrary to FR metrics [16], they are not yet normalized in recommendations of ITU-T. Furthermore, as it is stated in the position paper [17], a NR Quality Estimator has to be designed "to achieve the required accuracy for its application over the set of input content and artifacts for which it was designed". The contribution of our paper consists in proposing a new NR VQA metric called Weighted Macro-Block Error Rate (WMBER) to assess the quality of HD video encoded in H.264/AVC for video delivery via broadcast channels and IP networks. We evaluate the quality loss resulting from transmission and decoding. Specifically, the metric is designed to detect degradations induced by macro-block losses. The temporal artifacts caused by transmission delays such as jerkiness, have not been yet considered. However, these temporal artifacts can be modeled as loss. The three steps of measuring, pooling and quality score mapping [17] are developed in this framework. In [17], the proposed "paramount" condition for NR metrics is to incorporate as much information about HVS as possible. Therefore, we weighted a simple and fast measure on a compressed stream by a spatiotemporal saliency map built at the decoder end. In our work, we have decided to only evaluate video quality degradation caused by transmission artifacts. As a result, the reference video in our case – the Source Reference Channel (SRC) in VQEG terminology – is an "ideally" delivered code-stream, that is a decoded H.264/AVC video code-stream transmitted without any loss. The impaired video we evaluate – processed video stream (PVS) in VQEG terminology – are sequences which are decoded after transmission via a noisy channel and concealed by the error concealment mechanism from the decoder. The transmission loss models – the Hypothetic Reference Circuit (HRC) in VQEG terminology – are a combination of network loss profile and video decoder with their own error concealment mechanism.

To build our NR VQA metric WMBER [18], our first step is to define a visual saliency map computed from compressed H.264/AVC code-stream. It includes motion, transmission errors and pixels. Motion is directly extracted from code-stream motion vectors. Transmission errors are identified when parsing the code stream at the decoder end. Finally pixel values are obtained after decoding. The automatic saliency map obtained is evaluated against subjective saliency maps built from gaze tracking reference databases (IRCCyN [15] and LaBRI [12]). Theses subjective saliency maps are computed with a method inspired from D. S. Wooding [19]. To measure the performance of the automatic saliency map model, two evaluations are performed. First, the automatic saliency map "objective" quality is analyzed. This analysis is achieved by comparing objective saliency maps with subjective saliency maps. Then the performances of the WMBER weighted by the *automatic* saliency maps are compared with the WMBER weighted by the *subjective* saliency maps.

The gradient energy and the count of erroneous blocks are also required to be implemented at the decoder end for WMBER calculation. The whole block diagram of the method is presented in Fig. 1.

The pooling strategy for WMBER computation is composed of two levels: saliency map pooling (see the block "Saliency map pooling" in Fig. 1) and error measurement pooling. The error measurement pooling is performed with the erroneous block count, the error map and gradient (see block "error pooling" in Fig. 1).

Spatial and temporal saliencies are also pooled when building the spatiotemporal saliency map. After error pooling process, are only kept damaged macro-blocks with artifacts caused by transmission loss on transformed coefficients or motion vectors.

The *mapping-to-quality* strategy still remains an open issue. Despite the cubic polynomial regression proposed in VQEG report [20], the prediction of quality score was implemented in recent works with other forms of regression, such as exponential regression [7], [21], or with classical machine learning tools such as neural networks [22], [23]. Here we choose a machine learning approach based on a classifier weighted by an exponential similarity measure. This machine learning approach is called "similarity weighted average" and is introduced in [24]. The purpose of the similarity weighted average is to predict Mean Opinion Score (MOS) from objective quality measurements (i.e. WMBER scores). Then, predicted MOS are compared with MOS obtained in psycho-visual experiments accordingly to ITU-R BT500.11 protocol on 69 subjects.

For these experiments, several packet loss profiles are simulated in the scenario of video transmission over IP networks. For broadcast applications, other loss models are explored, such as Bit-Error Rate (BER) due to signal fading and interference.

The remainder of our paper is organized as follows. Section II describes the method for building an automatic saliency maps and the evaluation methodology with an NSS measure as proposed in [25]. In section III, we introduce a new low-cost, high accurate NR VQA metric WMBER which uses the

saliency map we compute at the decoder end. Section IV presents the learning approach for prediction of subjective metrics (MOS). Evaluation results for both proposed saliency maps and NR VQA metric are shown in section V. Section VI concludes this work and outlines its perspectives.

II. SPATIOTEMPORAL VISUAL SALIENCY MAP FROM H.264 CODE-STREAM

The HVS has the property of focusing the attention on narrow areas in the visual scene called salient areas. These salient areas send stimulus to the HVS. Inside video scenes, salient stimuli are characterized by high color contrasts, motion, and edge orientation.

In the literature, the saliency of a visual scene is generally depicted by two saliency maps, the "spatial" and the "temporal" saliency maps [25]. We will focus on both aspects to propose a saliency map model built on the decoder side.

A. Spatial saliency map built from the impaired decoded stream.

Spatial and/or temporal saliency maps are always computed from the original video for FR quality metrics and in other visual attention studies. The saliency of video content changes in the presence of compression and transmission artifacts. In [26], the authors study the influence of MPEG2 video compression on the human observation of video content in the specific cases of target detection and tracking. They state that gaze patterns become more random and unfocused on target as the compression ratio increases. The reason is the attraction of human eyes by strong compression artifacts outside the target region. The same statement is done by the authors of [27] in JPEG-compressed still images. Finally, the authors of [28] also study the influence of compression ratio to the gaze attraction. The authors conclude that the focus of attention is more concerned by the dynamic contents of the scene. Nevertheless, the results shown for high distortion video confirm the same phenomenon.

In the case of transmission errors, conventional decoders apply error concealment mechanisms, which correct errors on video frames with more or less success. Therefore, our proposal for computation of spatial saliency map on the decoded video stream consists in applying exactly the same method as if the spatial saliency map was computed on the original non-degraded video.

The spatial saliency map S^{SP} is mainly based on color, contrast and luminance [11]. If the decoder error concealment mechanism manages to recover a transmission error on a block without inducing any discontinuity with surrounding blocks, then the eye will not be attracted by the artifact and no saliency will be induced by the transmission and decoding. If, on the contrary, a visible artifact is induced, then the area becomes salient.

To build spatial saliency map on the decoded frames we used the method from [4].

The spatial saliency map extraction is based on the sum of 7 color contrast descriptors in the HSI domain. The 7 color contrast descriptors are: hue contrast, saturation contrast, intensity contrast, opposite color contrast, warm and cold color contrast, dominance of warm colors, and dominance of brightness and hue. The seven descriptors V_δ are computed for each pixel 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 accordingly to (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}.

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

As we stated above, we compute the spatial saliency maps on the decoded frames. In Fig. 2 we present two extreme cases for H.264/AVC compressed stream in typical broadcast applications: one with no transmission artifacts (upper row) and another one with strong transmission artifacts (lower row). As a result, a complementary saliency appears in the areas of loss.

B. Temporal saliency maps

The temporal saliency map S^T models the attraction of attention to motion singularities in a scene. Two sources for saliency can be used. The first source is the transformed domain, such as Gabor decomposition of both, video frames and optical flow field [13]. The second is the baseband pixel domain, as shown in [12]. Temporal saliency maps were recently proposed on the basis of residual motion with respect to global model [25]. The latter is estimated using image signal on pixel basis. In our proposal, we take profit of motion information already present in a video code-stream. The primary motion features such as macro-block and sub macro-block motion vectors of H.264 are used to estimate the global model and compute the residual motion. The global scheme for temporal saliency map computation is presented in Fig. 3.

The main step here is to estimate a global motion model. In this work, we follow the preliminary study from [12] and use a complete first order affine model (3):

$$\begin{aligned} dx_i &= a_1 + a_2(x_i - x_O) + a_3(y_i - y_O) \\ dy_i &= a_4 + a_5(y_i - y_O) + a_6(y_i - y_O) \end{aligned} \tag{3}$$

Here $\theta = (a_1, a_2, \dots, a_6)^T$ is the parameter vector of the global model (3) and $(dx_i, dy_i)^T$ is the motion vector of a macro-block. To estimate this model, we used robust least square estimator presented in [29]. We denote this motion vector $\overrightarrow{V_\theta}(s_i)$. Our goal is now to extract the local motion in video frames i.e. residual motion with regard to model (3). Therefore, we need to extract reliable motion vectors from H.264 code stream to compare with $\overrightarrow{V_\theta}(s_i)$. A 3 steps processing of H.264 motion vectors based on the standard architecture is thus fulfilled. First, in H.264/AVC a macro-

block or a sub macro-block may refer to multiple frames. This is the reason why all the vectors extracted from code stream have to be normalized by the distance from reference frame to current frame. The purpose of this normalization is to express the correct instantaneous motion in the current frame. We implement this normalization by simply weighting the motion vector by inverse temporal distance to the reference frame of each block.

The second particularity of H.264 is that some blocks can be intra-coded. We need local motion; hence we will recover a mean value of its motion vector from $2\Delta T$ frames: ΔT from the past and ΔT from the future. With a short ΔT, we use a simple 0-order prediction of block positions in previous and future frame for computational cost saving. This is depicted in Fig. 3 by motion interpolation block.

Taking into account that H.264 allows variable size of blocks, we interpolate MB motion vectors up to the smallest size of block (4x4 pixels). This is done by simple zero order interpolation of motion field.

The motion vectors from H.264 code stream are obtained at a 4x4 pixel resolution. We denote this motion vector $\overrightarrow{V_C}(s_i)$.

The residual motion is computed as a difference between sub macro-block motion vectors and estimated global motion vectors.

$$\overrightarrow{V_R}(s_i) = \overrightarrow{V_\theta}(s_i) - \overrightarrow{V_C}(s_i) \tag{4}$$

Another problem we face is the filtering of flat areas (see the block "Flat area filtering" in Fig. 4). Indeed, due to the ill-posed problem of motion estimation, the code stream contains erroneous vectors on flat areas. These motion vectors are usually very noisy and yield a strong residual motion. We remove these improper values by detecting flat area located in the background. Detecting flat areas is performed by computing and thresholding the gradient energy. Here we use a simple region growing algorithm. Starting from a single macro-block for which the energy of the gradient $\overline{\|\nabla\|^2}$ is lower than a threshold. We stress that the process of removing flat areas has to be applied to the background. If a small size flat area is a part of foreground object, its saliency should not be reduced. Coming back to the foundations of digital image processing we use the definition of a background by Azriel Rosenfeld [30]: *"the background"* component of a visual scene touches at least one of the borders. Thus our region-growing algorithm starts from the borders of a frame and agglutinates blocks whose gradient energy is lower than a threshold. The propagation stops in the direction of the first encountered "non-flat" block. The remaining blocks on borders are explored until all flat areas have been removed. Obviously, such an assumption can be criticized when an object enters in the camera field, but in this case we hope to get the saliency by a spatial contrast.

In Fig. 4 we illustrate the contribution of flat areas filtering. The original frame is presented in Fig. 4 (a). Then in Fig. 4 (b) we show the temporal saliency without flat area detection and in Fig. 4 (c) the temporal saliency map with the flat area removal we proposed.

Finally, the temporal saliency map $S^T(s_i)$ is computed by filtering the amount of residual motion in the frame. The authors of [4] reported, as established by S. Daly, that the human eye cannot follow objects with a velocity higher than 80 deg./s [31]. 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 [4] is given by (5).

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

with $\vec{v}_1 = 6\ deg./s$, $\vec{v}_2 = 30\ deg./s$ and $\vec{v}_{max} = 80\ deg./s$.

We follow this filtering scheme in temporal saliency map computation.

Due to the use of motion vectors from H.264 code stream and simple but efficient interpolation schemes, the computation of temporal saliency map is faster than real time (i.e. full decoding time).

C. Spatiotemporal saliency map

A spatiotemporal saliency map may be produced by combining the spatial and temporal saliency. Spatiotemporal saliency map fusion methods present in the literature remain simple. In [25], the authors review spatiotemporal saliency maps such as the "mean", "max" and a multiplicative "and" maps obtained on spatial and temporal maps as arguments. To obtain an integrated spatiotemporal saliency map, three steps are generally required. The two first steps rely on extracting both spatial and temporal saliency maps. The last step is the fusion. Several models which give good results already exist [4], [25] to predict the saliency of a video scene. In [12], we proposed a new method for fusion of saliency maps in a log-space. In this paper, we introduce a faster alternative: a squared sum of both spatial and temporal saliency maps. We denote resulting saliency maps respectively S_{LOG}^{SP-T} and S_{SQUARE}^{SP-T}. The S_{LOG}^{SP-T} [12] is defined by (6) with $\alpha = 0.5$. This fusion method has the same advantage as multiplicative saliency map S_{mul}^{SP-T} [18] that gives more importance to regions which have both high spatial and high temporal saliencies. S_{LOG}^{SP-T} is better than S_{mul}^{SP-T} as it still exhibits saliency when one of the primary saliencies, the temporal saliency, is low (we note that in video, null spatial saliency with high temporal saliency is improbable, because motion is perceived through luminance changes).

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

The squared fusion method S_{SQUARE}^{SP-T} we propose in this paper is defined by (7). This fusion method has similar fusion properties as S_{LOG}^{SP-T} when the temporal saliency is null. Its advantage is an obvious computational time-saving.

$$S_{SQUARE}^{SP-T}(s_i) = (S^{SP}(s_i) + S^{T}(s_i))^2 \quad (7)$$

The proposed spatiotemporal saliency map is an "objective" map which is built on features extracted from the video. It models the human perception dependency on low-level features which are contrasts and motion. In order to assess how spatiotemporal saliency maps are correlated to human visual attention, we compare these automatic saliency maps with the "subjective" visual saliency maps $S_{subj}(s_i)$. Subjective saliency maps can be obtained during experiments on the human subjects with *eye-trackers*. Such saliency maps are used in this work as the benchmark for "objective" saliency maps we build at the decoder end.

D. Spatiotemporal visual saliency map from gaze tracking

The construction of "subjective" visual saliency maps from observation of human gaze fixations on images or video is still an open issue and there is not any unified methodology. The building of such saliency maps is based on the so-called fixation maps, which were probably the most exhaustively studied by D. S. Wooding [19]. Fixation maps are successfully used for studying human attention in visual analysis tasks [32].

In [19], Wooding proposes a fixation map as a landscape of Gaussians centered on fixation points. In the case of eye-trackers with a high sampling rate such as 1000 Hz (see [32]), a supplementary processing is needed to extract fixation points. In our case as in the experiments in [15], a standard eye-tracker with a sampling rate of 60Hz is used. Since the video frame rate (25Hz) and the eye-tracker sampling rate (60Hz) are very closed, we consider each eye-tracker measurement as a fixation point. On each fixation point, a 2 dimensional Gaussian is centered with a 2 visual degrees standard deviation as in the Wooding's method. The height of the Gaussian is unitary at the fixation point. Then, for a given frame, all Gaussians from fixation points of all observers are cumulated into a "landscape" and the sum is normalized by the maximum in a frame.

Examples of saliency maps from gaze-tracking are presented in Fig. 5.

In order to compare "subjective" and automatic "objective" saliency maps, different strategies can be considered. Some of these strategies are referenced in [19]. Wooding proposes a differential saliency map obtained by a simple subtraction of normalized saliency maps S^{SP-T} and S_{subj}. S^{SP-T} and S_{subj} can also be compared with the Normalized Scan Path (NSS), which was introduced in [33] and used in [25]. We use the NSS as well and present the results in section V-B.

III. A NO-REFERENCE VIDEO QUALITY ASSESSMENT METRIC: WMBER

In this section, we describe the proposed NR quality assessment metric Weighted Macro Block Error Rate (WMBER). The block-diagram of WMBER computation is presented in Fig. 1.

The method is based on MB error detection. During the decoding process, the first step is to detect error location. This could be done by extracting the errors in the compressed stream. After recognizing the error in the compressed stream, we find the address of the MB forming a so-called MB Error Map (see "Error Maps" on Fig. 1). It means that if only one coefficient or motion vector is damaged in the MB, the whole MB is labeled as damaged. When the MB type is P or B, it is also labeled as damaged if the motion vector points to a damaged MB on the reference frame. The characteristic function of a MB is thus defined as Err_i. Err_i equals to one for damaged blocks and 0 otherwise. Then, the standard H.264/AVC spatiotemporal error concealment is applied. The purpose of our algorithm is to measure video quality in networks with transmission loss and not to measure the quality of compression. According to section II, we compute the spatiotemporal saliency map for all frames after error concealment. To improve the results, we need to take another parameter into account, which is the norm of the gradient in a block. It is well known that the human visual system is sensitive to low spatial frequencies and surrounding edges. If we consider a strong visible artifact on the block border, then it will be expressed in the higher gradient energy. In case of strongly textured blocks, the visible artifacts are possible due to the encoding inside a block. In this case we cannot make any distinction between the loss or the coding process. We found, that considering gradient energy for saliency computation inside a block enhances the saliency due to network transmission errors. Blockiness generated by transmission loss is for instance very noticeable by HVS on regions with low spatial activity. Hence, the norm of gradient $\|\nabla I\|$ is computed in the whole error-concealed frame I and normalized between 0 and 1. The gradient is computed on Y component of YUV decoded frames by Sobel operator [34]. This step is shown by the block « Gradient Energy » in Fig. 1. For each labeled macro-block, the mean of the normalized norm of the gradient in this block $\left\|\overline{\nabla_{mb_l}}\right\|$ is computed (8).

$$\left\|\overline{\nabla_{mb_l}}\right\| = \frac{1}{|mb_l|}\sum_{s\in mb_l}\frac{\|\nabla I(s)\|}{max_s(\|\nabla I(s)\|)} \quad (8)$$

With $mb_l -$ a macro-block and $s -$ a pixel of mb_l.

The saliency measure $\overline{S_l}$ for a block B_l is derived from the spatiotemporal saliency map as an average saliency of all pixels in a block. For WMBER computation we weight the saliency by the average gradient norm from (8). Therefore, areas with high gradient on block borders will get more weight in the final decision on saliency. Finally the WMBER is computed by (9) for each decoded frame I_j except IDR frames because the temporal saliency map is not available.

$$WMBER_l = 1 - \frac{\sum_{l=1}^{N_{mb_l}} \left[Err_l \cdot \overline{\|\nabla_{mb_l}\|} \cdot \overline{S_l} \right]}{\sum \overline{S_l}} \quad (9)$$

With N_{mb_l} the number of macro-blocks in a frame l.

Here $\overline{S_l}$ is a mean saliency of a block computed from pixel-based saliency in (7). According to equation (9), when $WMBER_l$ is close to 1, the quality of the frame is high. On the contrary, when $WMBER_l$ is close to 0, the quality of the frame is poor. To compute the WMBER for the whole video sequence, the average WMBER of all frames is calculated, as this is usually done for other VQ metrics (see e.g. [21]).

IV. MOS PREDICTION FOR EVALUATION OF SALIENCY BASED METRICS

In this section, we propose to use supervised learning method called similarity-weighted average classifier [24] to predict subjective quality (MOS) values from objective quality metrics, such as WMBER and FR VQA SSIM or MSE. This prediction method requires a training data set of n known pairs (x_i, y_i) to predict unknown y from measured x. Here, (x_i, y_i) pairs are objective metric values (WMBER, SSIM or MSE) x_i, associated with subjective metric values (MOS) y_i. The prediction of y given x is performed using equation (10) known as a weighted mean classifier with similarity function (11).

$$y = \frac{\sum_{i=1}^{n} sim(x_i, x) y_i}{\sum_{i=1}^{n} s(x_i, x)} \quad (10)$$

$$sim(z, x) = exp[-|x - z|] \quad (11)$$

In their original paper [24], the authors show good generalization properties of the classifier due to the monotonicity of the exponential similarity measure (11). This is the reason why we choose this prediction scheme. The other reason is that similarity weighted average does not require a heavy training as it is the case for many classifiers such as Neuronal Networks and SVMs. However, the similarity function (11) depends on the value range of the objective metric. Indeed, the minimum similarity value that we denote ε is obtained when the distance d between z and x reaches theoretical maximum value (12). We denote this distance d_{max}. For instance, $d_{max} = 1$ for WMBER and SSIM metrics.

$$sim(d_{max}) = exp[-d_{max}] = \varepsilon \quad (12)$$

This means that the similarity weight in the classifier for samples x_i which are very far ($d_{max} = 1$) from current measure, still remains high. Note that for $d_{max} = 1$, $sim(d_{max}) \approx 0.36$. Hence, taking into account the normalization of VQ metrics WMBER, SSIM and normalized MSE, we propose to modify the similarity measure as defined in (13).

$$sim'(z, x) = exp[-\lambda|x - z|] \quad (13)$$

From (12) and (13), the normalization parameter λ can be obtained for maximal theoretical values d_{max} as:

$$\lambda = \frac{-\ln(\varepsilon)}{d_{max}} \quad (14)$$

Furthermore, the normalization parameter λ can be adapted to a kind of content or a training database. The problem, here, is to optimize the parameter λ maximizing a payoff. We define the following logical payoff function $F(\lambda)$ measuring prediction qualities such as Pearson Correlation Coefficient (PCC), Spearman Rank-Order Correlation Coefficient (SROCC), Root Mean-Squared Error (RMSE) and Outlier Ratio (OR) evaluated on the training database:

$$F(\lambda) = \begin{cases} \begin{cases} 0, if\ PCC'(\lambda) < 0 \\ 1, otherwize \end{cases} \\ + \\ \begin{cases} 0, if\ SROCC'(\lambda) < 0 \\ 1, otherwize \end{cases} \\ + \\ \begin{cases} 0, if\ RMSE'(\lambda) > 0 \\ 1, otherwize \end{cases} \\ + \\ \begin{cases} 0, if\ OR'(\lambda) > 0 \\ 1, otherwize \end{cases} \end{cases} \quad (15)$$

with $\lambda > 0$ and $R'(\lambda)$ is the derivative of a function $R(\lambda)$.

The optimization with regards to λ was implemented by bisection method.

In the VQEG Report on the Validation of the Video Quality Models for High Definition Video Content [20], the authors propose to use a cubic polynomial function (16) to map the objective metric values x to the MOS y. In [21] and [7] exponential regression is used as it is recommended in former reports of VQEG for standard definition streams [35]. In section V we compare the prediction results of MOS with supervised learning approach and the polynomial regression method.

$$y = ax^3 + bx^2 + cx + d \quad (16)$$

V. EVALUATION AND RESULTS

A. Subjective experiments

To evaluate our work, we have generated a database at LaBRI from H.264 decoded full HD video; we further call it "LaBRI DB". The LaBRI DB is partially generated from a non-compressed open source video available in [36]. Since the availability of reference databases with subjective quality assessment data (MOS) for the VQA of H.264 encoded

streams is coming true [37], we are able to use the video DB produced by IRCCyN [15]. We call it "IRCCyN DB".

LaBRI DB

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 at the University of Bordeaux and Communication Systems Engineering Dept. at the Ben Gurion University of the Negev (BGU). Ten 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 [38]. Video sequences come from four different corpora: The Open Video Project [36], NTIA/ITS [39] and TUM/Taurus Media Technik [40].

Video sequences were encoded into the H.264/AVC format [3] using the x264 [41] software with a bit-rate of 6000Kb/s. Two models of transmission impairments were applied to each video sequence (Table 1). The first one, called IP model, simulates IP packet networks according to ITU-T Rec. G.1050 [42]. Hence, three kinds of networks: managed, semi-managed and unmanaged were simulated using five packet loss profiles. The second model, called 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 10 video sources (SRC) with the 8 impairment profiles, 80 processed video sequences (PVS) were generated. So, the total number of video sequences assessed by the experiment participants was 90.

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 [20]. The experiment room and the lightning conditions were compliant with the ITU-R Rec. BT.500-11 [2]. 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 30 minutes. To avoid the "learning effect" each participant has seen the video sequences in a random order and a "warm-up" session of 5 minutes was done before starting the experiment. Hence, 39 participants were gathered: 11 at LaBRI and 28 at BGU. MOS and DMOS subjective metrics were computed by using methods described in [20] and [2]. Eye-tracking measurements are only available for 13 subjects.

IRCCyN DB

In this paper, we also evaluate our method on a video database provided by IRCCyN Lab [15]. The IRCCyN DB is composed of 20 SD resolution video sequences. As in LaBRI DB all the video sequences were encoded with H.264/AVC and we consider the encoded video sequence without transmission impairments as the source (SRC). Four packet loss profiles were applied on each SRC. The subjective experiment was carried out by following ITU-R BT500.11 [2] protocol and 30 subjects have participated. MOS and DMOS metrics were computed as described in [2].

B. Evaluation of saliency maps

In this section, we compare the "objective quality" of the objective spatiotemporal saliency maps S_{mul}^{SP-T}, S_{LOG}^{SP-T} and S_{SQUARE}^{SP-T} with regard to the subjective spatiotemporal saliency map S_{subj}. Here, we use the NSS metric that was proposed in [25] instead of the PCC. In fact the correlation coefficient is very dependent on the Gaussian that was applied to build the subjective saliency map S_{subj} from gaze positions. NSS is a Z-Score that expresses the divergence of the subjective saliency map from the objective saliency maps. The NSS computation for a frame l is depicted by equation (17). Here, S_{obj}^{N} denotes the objective saliency map S_{obj} normalized to have a zero mean and a unit standard deviation, $\bar{X}$ means an average. When $\overline{S_{subj_l} \times S_{obj_l}^{N}}$ is higher than the average objective saliency, the NSS is positive; it means that the gaze locations are inside the saliency depicted by the objective saliency map. In other words, higher the NSS is, more objective and subjective saliency map are similar.

$$NSS_l = \frac{\overline{S_{subj_l} \times S_{obj_l}^{N}} - \overline{S_{obj_l}}}{\sigma\left(S_{obj_l}\right)} \tag{17}$$

The NSS score for a video sequence is obtained by computing the average of NSS for all frames as in [25]. Then the overall NSS score on each video database is the average NSS of all video sequences. Results are presented in Table 2. For both databases, the results of the proposed fusion method (i.e. Square fusion) are better than the state of the art fusion method (i.e. Multiplicative fusion). Log fusion despite good visual results gives low NSS due to scale change with regard to the second evaluation on the target task of VQA.

C. Evaluation of NR saliency based metric WMBER

This section presents the evaluation results of the proposed NR metric WMBER. The performance of WMBER is compared with the FR metrics PSNR and SSIM. Several saliency maps are tested to compute the WMBER:

- The temporal saliency map (Temporal)
- The spatial saliency map (Spatial)
- The Multiplication spatiotemporal saliency map (Multiplication)
- The Log spatiotemporal saliency map (Log)
- The Square spatiotemporal saliency map (Square)
- The saliency map from eye-tracking measurements (EyeTracker).

To compare the metrics, we use four evaluation criteria: PCC, SROCC, RMSE and OR [7]. The Similarity Weighted

Average MOS prediction method is also compared with the polynomial regression method (16).

All the results are presented in Table 3 and Table 4. For both databases and both prediction methods, the WMBER with Squared saliency map provides the best results for the fully automatic methods.

The "target" comparison of our proposed saliency map with regards to the subjective saliency map in ultimate quality assessment task shows very good performances of the proposed method. In average, on IRCCyN DB with SD content, the target evaluation criteria values PCC, SROCC, RMSE are only 10% lower than those obtained with eye-tracker saliency map. On the LaBRI database, the results obtained with our proposed saliency map are even 5% better on all four PCC, SROCC, RMSE and OR. The difference between IRCCyN and LaBRI database results can be explained that LaBRI database is in full HD. Our method provides finer saliency with full HD.

Overall the proposed NR VQA WMBER metric gives for all cases better results than MSE and SSIM FR metrics.

We also test the contribution of the new prediction scheme by weighted average classifier into global quality assessment scheme. In Fig. 6 we show the improvement in terms of four quality criteria with adequate choice of λ, see section IV. Fig. 6 illustrates how the minimal accepted similarity value (see equation (12) in section IV) impacts the quality of prediction. With decrease of this parameter a clear improvement is observed. In Fig. 7, we show a scatter plot of prediction of MOS with standard $\lambda = 1$ (see equation (11)) and optimal λ computed according to method we propose in section IV. One can see that the optimal λ ensures almost one to one correspondence between MOS and predicted MOS. The experiments were conducted using 64 bits double precision floats which allow fine tuning of λ parameter.

Meanwhile the proposed prediction scheme gives approximately the same results as baseline polynomial regression. Since the similarity weighted average prediction method is based on an incremental learning scheme, it can easily be improved by enriching the training database.

VI. Conclusion and perspectives

In this paper, we were interested in NR quality assessment of video content delivery over IP and radio-frequency broadcasting networks which are subject to loss. For both HD and SD video quality content encoded in actual H.264/AVC, we proposed a new NR quality metric truly using the information contained in transmitted stream such as error detection and motion vectors extraction, combined with image signal from decoded frames. The metric is based on visual saliency maps built at the decoder end. To predict subjective video quality metric in terms of MOS, we used a supervised learning approach such as weighted average classifier with exponential similarity function and we proposed a new normalization approach for this similarity measure.

The new metrics and the prediction methods were exhaustively tested with 11 loss profiles on IRCCyN and LaBRI databases. The experiments conducted on both databases according to the VQEG evaluation protocol show that the proposed No-Reference metric WMBER provides more accurate results than classical Full-Reference metrics such as MSE and SSIM.

We show that the proposed method for visual saliency maps construction outperforms multiplicative saliency pooling and gives very close results to subjective saliency maps build from eye-tracker data.

The proposed prediction method on the basis of similarity weighted average give similar results as VQEG polynomial regression with our modification proposal.

In conclusion, we believe that the proposed NR quality metric has very good applications perspective as it doesn't require any modification of actual broadcast chain. All the intelligence is at the decoder side. Furthermore, the reuse of information during the decoding process gives seducing perspective for real-time implementation. Moreover we see an interesting perspective in the study of supervised learning approaches for prediction of video quality as a function of content genre and "usefulness". This is, to our knowledge, one of the objective of the VQA community.

Acknowledgment

This work was partially supported by French national CIFRE grant to Audemat and LaBRI.

References

[1] European Broadcasting Union, "Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television," ETSI European Standard ETSI EN 300 744 V1.6.1, 2009.

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

[3] ISO/IEC, "H.264 Advanced Video Coding," in *Information technology - Coding of audio-visual objects*, 2004, ch. Part 10.

[4] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483-2498, Sep. 2007.

[5] 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.

[6] A. Panayides, et al., "Atherosclerotic plaque ultrasound video encoding, wireless transmission, and quality assessment using H.264," *IEEE Trans Inf Technol Biomed*, vol. 3, no. 15, pp. 387-397, May 2011.

[7] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, "Balancing Attended and Global Stimuli in Perceived Video Quality Assessment," *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1269-1285, Dec. 2011.

[8] R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transaction on Image Processing*, vol. 6, no. 19, pp. 1427-1441, Jun. 2010.

[9] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[10] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal Optical Society America / Optics Image Science Vision*, vol. 24, no. 12, pp. B61-B69, Dec. 2007.

[11] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering

temporal variation of spatial visual distorsions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266-279, 2009.

[12] H. Boujut, O. Hadar, J. Benois-Pineau, T. Ahmed, and P. Bonnet, "Weighted-MSE based on Saliency map for assessing video quality of H.264 video streams," *IS&T/SPIE Electronic Imaging*, Jan. 2011.

[13] K. Seshadrinathan and A. C. Bovik, "Motion Tuned Spatio-temporal Quality Assessment of Natural Videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[14] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 2560-2563, Oct. 2008.

[15] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling Saliency Awarness for Objective Video Quality Assessment," *QoMEX*, 2010.

[16] International Telecommunication Union, "ITU-T Rec. J.247 Objective perceptual multimedia video quality measurement in the presence of a full reference," IEC Recommendation, 2008.

[17] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Processing: Image Communication*, vol. 25, pp. 469-481, Aug. 2010.

[18] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet, "A Metric For No Reference video quality assessment for HD TV Delivery based on Saliency Maps," *ICME 2011, Workshop on Hot Topics in Multimedia Delivery*, Jul. 2011.

[19] D. S. Wooding, "Eye Movements of Large Populations: II. Deriving Regions of Interest, Coverage, and Similarity using Fixation maps," *Behavior Research Methods*, vol. 34, no. 4, pp. 518-528, 2002.

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

[21] 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.

[22] P. Le Callet, C. Viard-Gaudin, and D. Barba, "A Convolutional Neural Network Approach for Objective Video Quality Assessment," *IEEE trans. on neural networls*, vol. 17, no. 5, pp. 1316-1327, 2006.

[23] C. Li, A. C. Bovik, and X. Wu, "Blind Image Quality Assessment Using a General Regression Neural Network," *IEEE Transaction on Neural Networks*, vol. 22, no. 5, pp. 793-799, May 2011.

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

[25] S. Marat, et al., "Modelling Spatio-Temporal Saliency To Predict Gaze Direction For Short Videos," *IJCV*, no. 82, pp. 231-243, Mar. 2009.

[26] O. Hadar, E. Goldberg, and E. Topchik, "The influence of image compression on target acquisition," *Electronic Imaging 2008*, no. 68060U, p. 6806, Jan. 2008.

[27] T. Shoham, D. Gill, and C. Sharon, "A novel perceptual image quality measure for block based image compression," *IS&T/SPIE Electronic Imaging*, vol. 7867, Jan. 2011.

[28] A. Mittal, A. K. Moorthy, W. S. Geisler, and A. C. Bovik, "Task Dependence of Visual Attention on Compressed Videos: Point of Gaze Statistics and Analysis," *IS&T SPIE Human Vision and Electronic Imaging*, no. 84444, Jan. 2011.

[29] P. Krämer, J. Benois-Pineau, and J.-P. Domenger, "Scene similarity measure for video content segmentation in the framework of rough indexing paradigm," *International Journal of Intelligent Systems*, vol. 21, no. 7, pp. 765-783, 2006.

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

[31] 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.

[32] V. Yanulevskaya, J.-B. Marsman, F. Cornelissen, and J.-M. Greusebroek, "An image statistics based model for fixation prediction," *Cognitive Computation*, vol. 3, no. 1, pp. 94-104, Mar. 2011.

[33] R. J. Peters, A. Lyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, pp. 2397-2416, 2005.

[34] I. Sobel, "Neighbourhood coding of binary images fast contour following and general array binary processing," *Computer Graphics and image Processing*, vol. 8, pp. 127-135, Aug. 1978.

[35] VQEG (Video Quality Experts Group), "Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase 1, FR-TV 1," Report, 2000.

[36] The Open Video Project. (2011, Dec.) LABRI-ANR ICOS-HD. [Online]. http://www.open-video.org/collection_detail.php?cid=23

[37] F. De Simone, et al., "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," *International Workshop on Quality of Multimedia Experience, QoMEX 2009*, no. 5246952, pp. 204-209, 2009.

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

[39] NTIA/ITS. (2011, Dec.) VQEG FTP - NTIA source. [Online]. ftp://vqeg.its.bldrdoc.gov/HDTV/NTIA_source/HDTV_Readme.doc

[40] C. Keimel, J. Habigt, T. Habigt, M. Rothbucher, and K. Diepold, "Visual quality of current coding technologies at high definition IPTV bitrates," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, p. 390–393, 2010.

[41] Videolan. (2011, Dec.) x264 - a free h264/avc encoder. [Online]. http://www.videolan.org/developers/x264.html

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

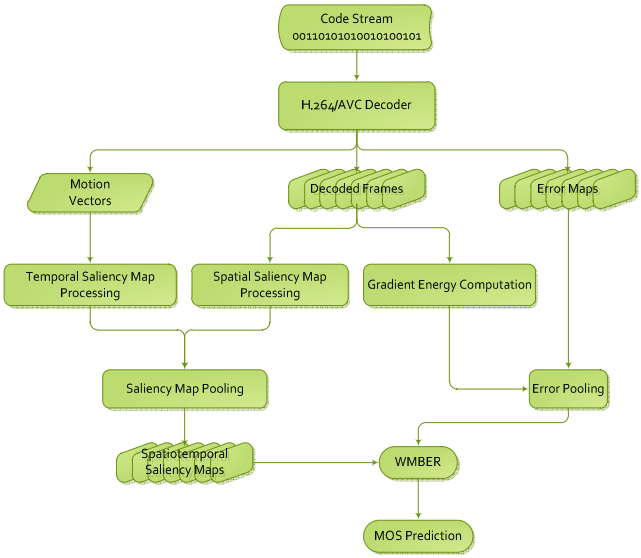
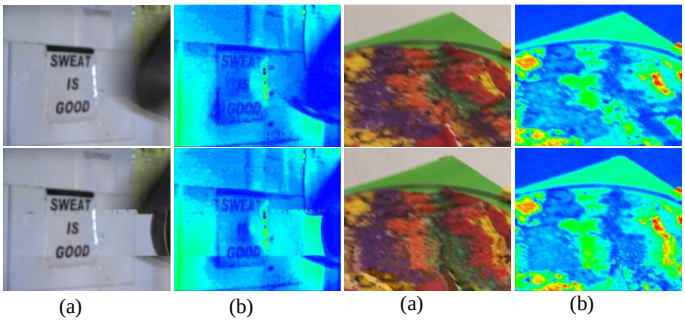


Fig. 1. Block-diagrams for WMBER computation and MOS prediction



(a) (b) (a) (b)

Fig. 2. Illustration of spatial saliency map construction on decoded video: (a) original decoded frame, (b) spatial saliency map. Sequences 7 and 12, LaBRI database at 6 MBps. First row: SRC, second: PVS

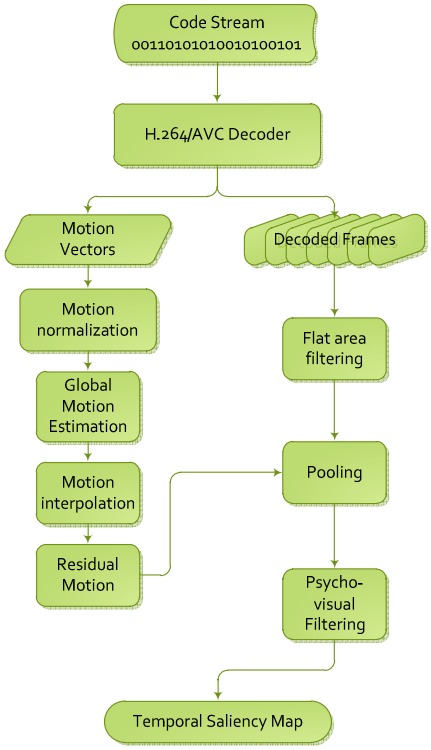
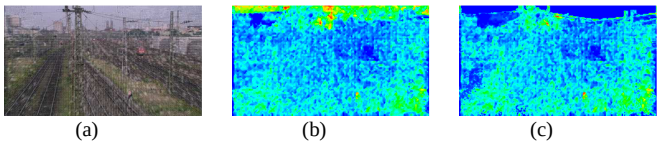
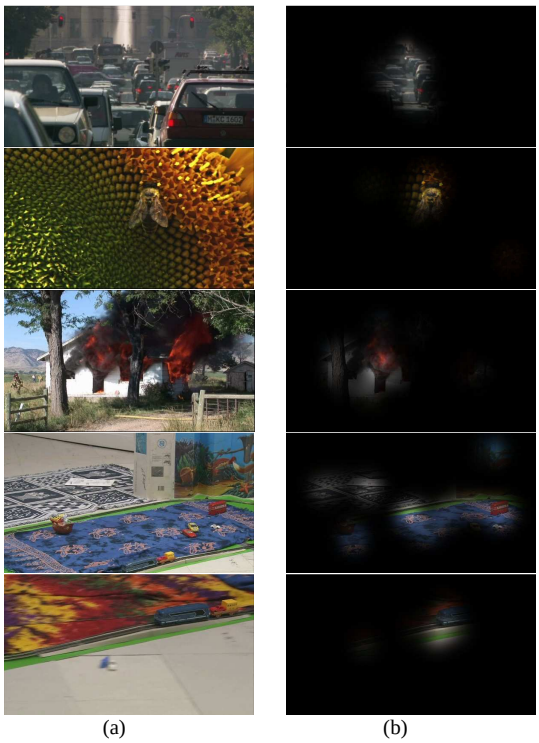


Fig. 3. Block-diagram for temporal saliency map estimation.



(a) (b) (c)

Fig. 4. Residual motion computation, “station 2 – TUM”, LaBRI DB at 6MBps: (a) Original frame, (b) “Heat-map” of temporal saliency map before flat area filtering, (c) “Heat map“ of temporal saliency map after flat area filtering.



(a) (b)

Fig. 5. Example of saliency maps from gaze tracking. (a) Original frames, (b) Saliency maps.

TABLE 1
LOSS PROFILES OF LABRI DB

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

TABLE 2
NSS RESULTS OF FUSION METHODS

	Multiplication	Log	Square
LaBRI DB	0.773	0.042	0.994
IRCCyN DB	0.024	-0,545	1,059

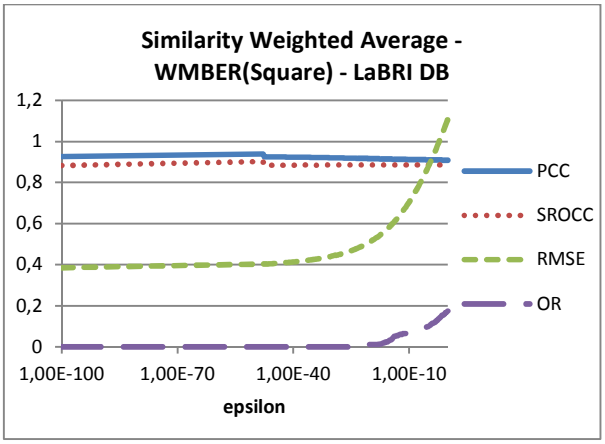


Fig. 6 Similarity Weighted Average performance on LaBRI DB with WMBER(Square)

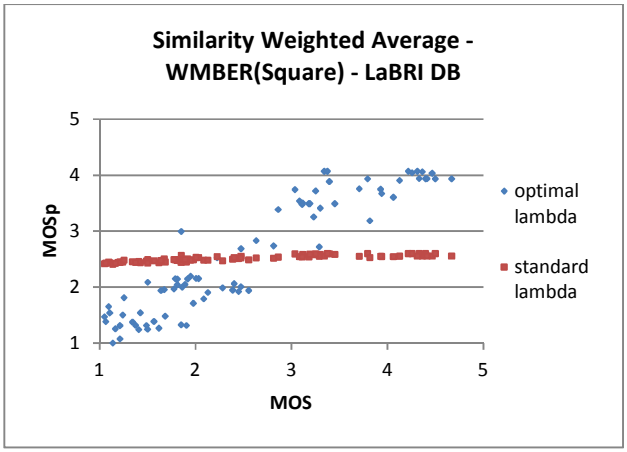


Fig. 7 MOS vs MOSp of Similarity Weighted Average on LaBRI DB with WMBER(Square)

TABLE 3
METRICS RESULTS ON LABRI DB

MOS prediction		MSE	SSIM	WMBER (Temporal)	WMBER (Spatial)	WMBER (Multiplication)	WMBER (Log)	WMBER (Square)	WMBER (EyeTracker)
Similarity Weighted Average	PCC	0.831	0.768	0.928	0.925	0.930	0.935	0.938	0.887
	SROCC	0.782	0.765	0.890	0.871	0.879	0.882	0.903	0.841
	RMSE	0.581	0.750	0.410	0.424	0.432	0.402	0.404	0.467
	OR	0.044	0.073	0.013	0.000	0.000	0.000	0.000	0.024
Polynomial regression	PCC	0.799	0.746	0.930	0.921	0.930	0.931	0.940	0.897
	SROCC	0.819	0.684	0.881	0.868	0.869	0.879	0.881	0.833
	RMSE	0.648	0.698	0.387	0.432	0.398	0.389	0.369	0.468
	OR	0.019	0.074	0.013	0.013	0.002	0.002	0.000	0.015

TABLE 4
METRICS RESULTS ON IRCCYN DB

MOS prediction		MSE	SSIM	WMBER (Temporal)	WMBER (Spatial)	WMBER (Multiplication)	WMBER (Log)	WMBER (Square)	WMBER (EyeTracker)
Similarity Weighted Average	PCC	0.367	0.369	0.351	0.407	0.393	0.402	0.470	0.567
	SROCC	0.367	0.409	0.364	0.355	0.364	0.393	0.443	0.558
	RMSE	0.893	0.825	0.821	0.811	0.820	0.810	0.793	0.734
	OR	0.184	0.173	0.173	0.149	0.167	0.144	0.153	0.087
Polynomial regression	PCC	0.426	0.391	0.358	0.347	0.348	0.387	0.468	0.527
	SROCC	0.458	0.407	0.326	0.345	0.341	0.362	0.444	0.495
	RMSE	1.604	0.811	0.892	1.193	1.406	1.139	0.790	0.707
	OR	0.123	0.119	0.206	0.171	0.191	0.171	0.133	0.079