

No-Reference Video quality assessment of H.264 video streams based on semantic saliency maps

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

The paper contributes to No-Reference video quality assessment of broadcasted HD video over IP networks and DVB. In this work we have enhanced our bottom-up spatio-temporal saliency map model by considering semantics of the visual scene. Thus we propose a new saliency map model based on face detection that we called semantic saliency map. A new fusion method has been proposed to merge the bottom-up saliency maps with the semantic saliency map. We show that our NR metric WMBER weighted by the spatio-temporal-semantic saliency map provides higher results than the WMBER weighted by the bottom-up spatio-temporal saliency map. Tests are performed on two H.264/AVC video databases for video quality assessment over lossy networks.

1. INTRODUCTION

With the introduction of television broadcasting services over IP and DVB networks, the quality assessment of video became an important research topic both for academia and industries. The No-Reference (NR) quality assessment research is of primarily importance for the community because of the wide range of applications and the inherent difficulty of the task [1]. In this paper, as in our previous works, we are interesting in assessing quality of H.264 encoded video transmission over lossy channels. Hence the source of degradation is due to transmission losses, not to encoding. We propose a new model to enhance the visual saliency model of NR video quality assessment metric Weighted Macro-Block Error Rate (WMBER) [1]. The enhancement of visual saliency is obtained by considering the influence of semantics of the visual scene on the visual attention. Recent works [1] [2] [3] [4] have showed that saliency maps are well suited to measure the perceived quality in the context of lossy video broadcasting networks. However visual saliency models [5], [6] are mainly based on a bottom-up approach which does not take into account the semantics of the visual scene. In [7] and [8] the authors showed that semantics increase visual attention especially on faces. It was observed that areas which contain faces grab the attention 16.6 times more than areas without faces and with the same saliency [7]. Furthermore, the authors of [9] also stress that the perceived quality depends on the usefulness of the content. The contribution of our work consists in the integration of a “semantic” saliency focused on human faces and bottom-up saliency model we proposed in [2] on H.264 encoded visual streams. The rest of the paper is organized as follows: in section 2 we introduce the semantic saliency map model based on face detection. In section 3, we propose a fusion method to enhance our bottom-up spatio-temporal saliency maps model by semantic saliency. In section 4 we briefly introduce our NR metric called WMBER based on saliency maps. In section 5 we describe the prediction method

of subjective quality metric MOS from the proposed objective quality metric WMBER. The experiments and the results are described in section 6 while results, conclusion and perspectives are presented in section 7.

2. SEMANTIC SALIENCY MAP MODEL

The authors of [7] and [8] showed that semantics of observed visual scene changes the behavior of human visual attention. Thus, the visual attention is not uniformly attracted by spatio-temporal saliency of the content, obtained on a purely bottom-up manner e.g. local contrast and residual motion. The visual attention also depends on the semantic meaning of salient areas. In [7], it was observed that the visual attention is mainly grabbed by faces. This has the effect of significantly increasing the saliency on face areas. For a large spectrum of quality assessment tasks in broadcasting and IP streaming applications, the study of semantic saliency would be surely genre-dependent. Nevertheless we can still consider the most generic semantic saliency reduced by the presence of human faces across genres and applications. Hence, in this work, we decided to stay focused on face detection to build the semantic saliency model. Face detection in video is a very old research subject [10] as the presence of human faces is of content analysis and mining [11]. In the variety of face detection methods, the detector of Viola & Jones [11] has been chosen due to its availability in the OpenCV library and also the availability of good trained models. The Viola & Jones gives performance of around 0.6 (according to our experiments in TRECVID campaign). Furthermore, the detector of Viola & Jones supplies the results as a bounding box of face features. This crisps detection is not convenient to model the visual attention. Hence, to build the face-based semantic saliency map, we enhance the Viola & Jones results by temporal filtering and model visual saliency by a “psycho-visual” Gaussian.

The semantic saliency map is built as follows for each frame k :

1. Faces are detected with Viola & Jones detector.
2. Detections are filtered along the time axis by a median temporal filter we proposed in [11].
3. For each detected face i , a bounding box $B_{k,i}$ is associated.
4. For each $B_{k,i}$ a two-dimensional Gaussian is drawn at its center (x_0, y_0) (Eq. 1). With σ_x and σ_y respectively equal to the width and the height of $B_{k,i}$ if the width or the height is greater than 2 *vis. deg*. Otherwise the value of 2 *vis. deg* is set to σ_x or σ_y . The value of 2 *vis. deg* depicts the size of the fovea on the screen [12]. The result is stored in a matrix $S_{B_{k,i}}^F(k)$ of the same size as the frame k .

$$f(x, y) = Ae^{-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)}$$

$$\sigma_x = \begin{cases} \text{width}(B_{k,i}) & \text{if } \text{width}(B_{k,i}) > 2 \text{ vis. deg} \\ 2 \text{ vis. deg} & \text{otherwise} \end{cases} \quad (1)$$

$$\sigma_y = \begin{cases} \text{height}(B_{k,i}) & \text{if } \text{height}(B_{k,i}) > 2 \text{ vis. deg} \\ 2 \text{ vis. deg} & \text{otherwise} \end{cases}$$

5. The semantic saliency map $S^F(k)$ is computed by summing up all the $S_{B_{k,i}}^F(k)$ (Eq. 2).

$$S^F(k) = \sum_t S_{B,k,t}^F(k) \quad (2)$$

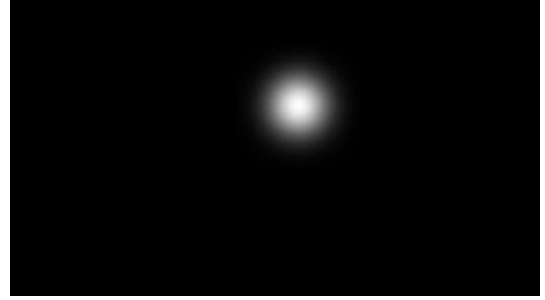
6. $S^F(k)$ is normalized by the maximum value of $S^F(k)$ in the frame (Eq. 3).

$$S^F(k) = \frac{1}{\max(S^F(k))} \times S^F(k) \quad (3)$$

An example of a semantic saliency map is given in Figure 1. Thus build the semantic saliency map does not privilege the faces in a foreground with regards to those in the background of the same frame. In our model, the large faces will have stronger impact on the quality metric as we will see thus further.



Original frame



Semantic saliency map

Figure 1 Example of semantic saliency map

3. FUSION OF BOTTOM-UP AND SEMANTIC SALIENCY MAPS

Bottom-up saliency models [5], [6] follow the computation scheme we proposed in [2]. First of all, the temporal and the spatial saliency maps are processed. The temporal saliency map S^T is mainly based on relative motion and the spatial saliency maps S^{SP} is built from local color contrasts. Then to obtain bottom-up the spatio-temporal saliency map S^{SP-T} , the spatial and the temporal saliency maps are combined by applying a Log-fusion method in our previous work. To merge the bottom-up saliency maps with the semantic saliency map, we have extended the Log fusion method S_{LOG}^{SP-T} introduced in [2] to consider the semantic saliency component. The new Log fusion method S_{LOG}^{SP-T-F} is expressed by Eq. 4 where each map spatial, temporal or semantic is weighted by α, β, γ respectively. The Log fusion method has the advantage to provide stronger weight to areas which have high spatio-temporal and semantic saliency. Unlike the Multiplication fusion method [6], the Log fusion method does not provide null saliency maps when one of the input saliency map is null. This feature of the Log fusion is very important for merging with the semantic saliency as faces are not always present in the frames of the video scene.

$$S_{LOG}^{SP-T-F}(k) = \frac{\alpha}{\alpha + \beta + \gamma} \log(S^{SP}(k) + 1) + \frac{\beta}{\alpha + \beta + \gamma} \log(S^T(k) + 1) + \frac{\gamma}{\alpha + \beta + \gamma} \log(S^F(k) + 1) \quad (4)$$

The weights α, β, γ can be set on the basis of content a priori or trained. In this paper we used $\alpha = \beta = 1, \gamma = 2$ to slightly privilege semantic saliency.

4. NO REFERENCE VIDEO QUALITY ASSESSMENT METRIC

In this section we will describe the metric Weighted Macro Block Error Rate (WMBER) we proposed in [1] for NR quality assessment and the related method. The block-diagram of WMBER computation is presented in Figure 2 below.

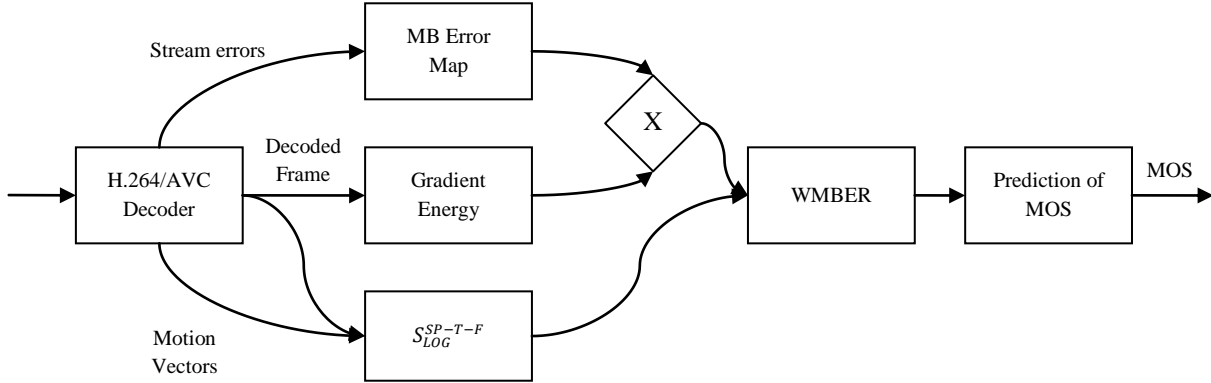


Figure 2 WMBER computation block-diagram

The method is based on MB error detection, during the decoding process. The first step here 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. It means that if only one coefficient or a motion vector is damaged in the MB, the whole MB is labeled as damaged. The characteristic function of a MB is thus defined as Err_i . It equals to 1 for damaged blocks and 0 otherwise. Then the standard H.264/AVC spatio-temporal error concealment is applied. Our algorithm is designed to measure video quality on networks with transmission loss and not to measure the quality of compression. According to section 3, we compute the spatio-temporal-semantic saliency map S_{LOG}^{SP-T-F} for all frames after error concealment. To improve the results, we need to take into account another parameter 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 distinction between the loss and the coding process. We found that considering gradient energy for saliency computation inside a block enhances the saliency due to network transmission errors. This is hold especially in regions with low spatial activity where blockiness due to transmission is very noticeable by HVS. Hence, the norm of the gradient $\|\nabla I\|$ is computed in the whole error-concealed frame I and normalized between 0 and 1. This step corresponds to the block « Gradient Energy » in Figure 2. For each labeled macro-block, the mean of the normalized norm of the gradient in this block $\|\overline{\nabla_{mb_i}}\|$ is computed. The saliency measure for a block is derived from the spatio-temporal-semantic saliency map (Eq. 4) as an average saliency of all pixels in a block. For WMBER computation we weight the saliency by the average gradient norm (Eq. 5). In this way areas with high gradient on block borders will get more weight in the final decision on saliency. We are especially interested in regions with high energy on MB borders and low energy inside and in the surrounding of the MB. Finally the WMBER is computed by Eq. 5:

$$WMBER = 1 - \frac{\sum Err_i \cdot \|\overline{\nabla_{mb_i}}\| \cdot \overline{S_{LOG}^{SP-T-F}}_i}{\sum \overline{S_{LOG}^{SP-T-F}}_i} \quad (5)$$

Here $\overline{S_{LOG}^{SP-T-F}}_i$ is a mean saliency of a block computed from pixel-based saliency in Eq. 4.

With respect to our semantic saliency map, the following holds: faces of large size will impact the metric (Eq. 5) more, as they will contribute into WMBER computation in several blocks.

5. MOS PREDICTION BY SUPERVISED LEARNING

In our recent work [2] we proposed a supervised learning method for prediction of subjective score from objective quality metric. 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 objective metrics output values associated with MOS values from the subjective experiment. y is the predicted MOS from a given objective metric output value x . The prediction is performed using equation (Eq. 6) known as Similarity Weighted Average classifier (Eq. 7).

$$y = \frac{\sum_{i=1}^n s(x_i, x) y_i}{\sum_{i=1}^n s(x_i, x)} \quad (\text{Eq. 6})$$

$$s(z, x) = \exp[-|x - z|] \quad (\text{Eq. 7})$$

In the original paper [13] the authors show good generalization properties due to the monotonicity of the exponential similarity measure (Eq. 7), 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 classifiers such as Neuronal Networks and SVMs and proved to be more accurate than the polynomial fitting usually employed [14].

6. TESTS AND EVALUATIONS

6.1. Subjective experiments

We remind that we are interested in quality assessment of video transmitted over lossy channels. Thus for us the reference source SRC [15] is ideal decoded, without any errors and any error concealment, from the H.264 compressed stream.

6.1.1. LaBRI database

We used the LaBRI database described in [2] for subjective HD video quality assessment for lossy networks. This database contains 20 different video sources (SRC) with a resolution of 1920x1080 pixels encoded in H.264 at 6000 kb/s. Eight network loss profiles described in [16] were applied on each SRC. Mean Opinion Score (MOS) values were computed from the votes of 35 participants for all the videos present in the database. However, in this paper, we are only interested in video content with faces. So in this evaluation context, we have only kept the 9 SRC containing faces with the 8 impaired versions.

6.1.2. IRCCyN database

The IRCCyN/IVC Eyetracker SD 2009_12 Database [4] is also used in this experiment. This database contains 20 SRC and 4 network loss profiles were applied on each SRC. The videos of the database have a resolution of 720x576 pixels and were encoded in H.264. The MOS values were computed from the votes of 30 participants. However, for this study we have only kept the 10 SRC containing faces with their 4 impaired versions.

6.2. Evaluation

In this section, we compare three objective video quality metrics with the results of the two subjective experiments described in sections 6.1.1 and 6.1.2. The first one is the Mean Squared Error (MSE) computed between the original non degraded video and its degraded version. It is a Full Reference (FR) metric. The second one is the SSIM [17], it is also a FR metric. The third one is WMBER, which is a NR metric. For the WMBER metric we have tested three different weight methods which are:

- $S_{LOG}^{SP-T-F(V\&J)}$ the spatio-temporal-semantic saliency maps using the Viola & Jones face detector.
- $S_{LOG}^{SP-T-F(GT)}$ the spatio-temporal-semantic saliency maps using the manual annotation of faces.
- S_{LOG}^{SP-T} the spatio-temporal saliency maps without semantic map.

The Similarity-Weighted method described in section 5 is used to predict the MOS. Therefore, to train and evaluate the prediction methods, a dataset of objective_metric/MOS pairs is built for each metric. To validate the results of the metrics, the 10-Fold cross-validation method is applied. This method randomly splits the dataset into 10 equal parts, 9 parts are used for training the prediction method and the last one is used for the evaluation. Then, the evaluation subset is used for training and one of the 9 training subsets is used for evaluation. The process is run to validate each metric until that each subset has been used for evaluation i.e. 10 times. The evaluation is performed by computing the Pearson Correlation Coefficient (PCC) (Eq. 8) denoted by R and the Root Mean Squared Error (RMSE) (Eq. 9)

$$R = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (8)$$

where x_i is the *MOS*, y_i the predicted MOS_p and N the number of data pairs in the evaluation dataset. The final performance score is the mean of the 10 R values.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2} \quad (9)$$

where x_i is the *MOS*, y_i the predicted MOS_p and N the number of data pairs in the evaluation dataset. The final performance score is the mean of the 10 RMSE values.

Figure 3 and Figure 4 below depict the results of evaluation of our metric WMBER with mixed saliency and semantic saliency against FR metrics and WMBER with bottom-up saliency maps only. On both databases, the results of WMBER with mixed saliency and semantic are the bests. The results of all the evaluated metrics are lower on the IRCCyN database. This is due to a lower number of loss profiles and a narrow range of loss profiles.

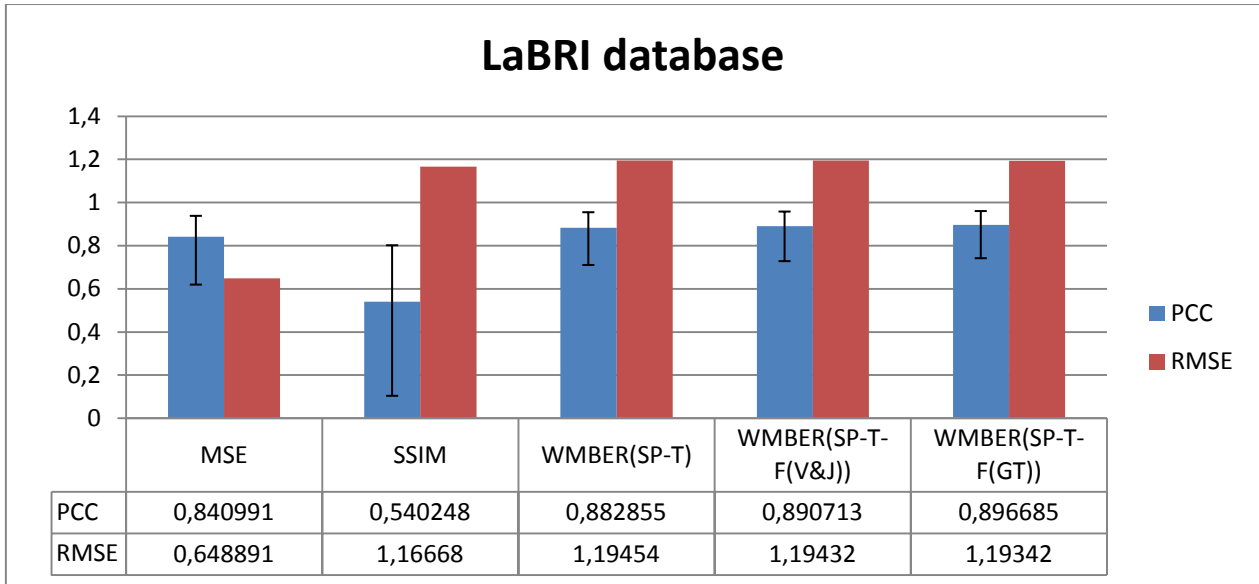


Figure 3 Objective metrics performance on LaBRI database

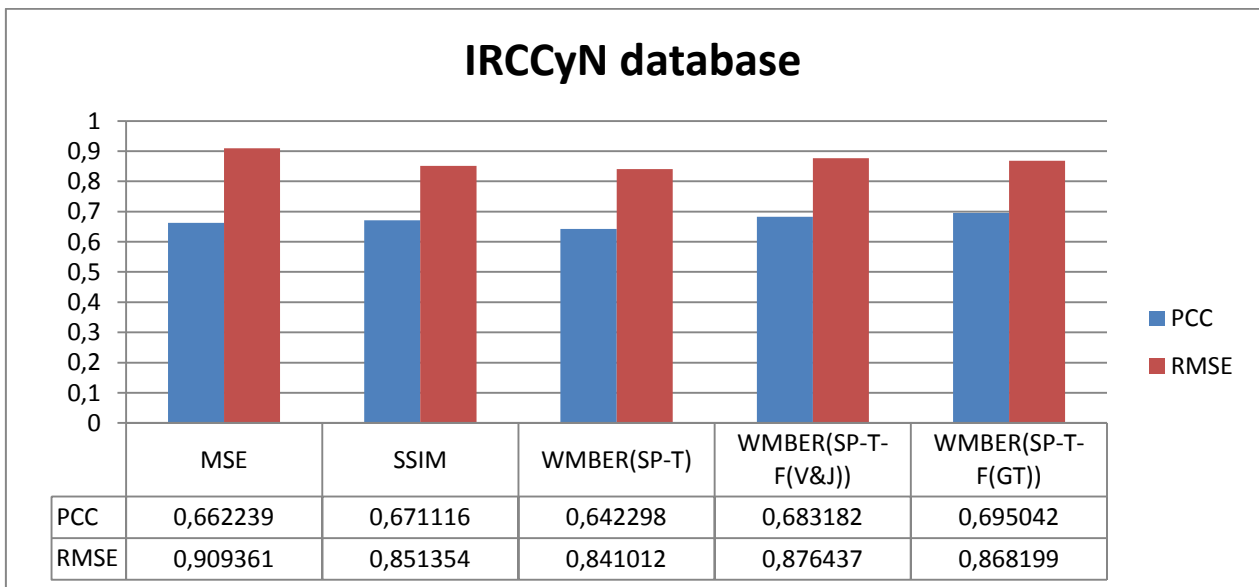


Figure 4 Objective metrics performance on IRCCyN database

7. CONCLUSION

In this paper, we were interested in the problem of NR video quality assessment over lossy channels. We proposed a new semantic saliency map based on human faces to consider the semantics of the visual scene. A new fusion method was also proposed in this work to combine bottom-up saliency maps with semantic saliency map. We showed with this first experiment that the spatio-temporal and semantic saliency map with WMBER provides better results than the bottom-up

spatio-temporal saliency map. However the Log-fusion weights can be tuned as a function of the content genre e.g. operas should have stronger weight on the semantic saliency map instead of content showing landscapes that should have a low weight on the semantic saliency map. These weights can be also obtained by a training method. This will be considered in our future studies of this work in order to improve the NR quality metric performance.

8. REFERENCES

- [1] 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.
- [2] 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.
- [3] 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.
- [4] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling Saliency Awareness for Objective Video Quality Assessment," *QoMEX*, 2010.
- [5] O. Le Meur and P. Le Callet, "What we see is most likely to be what matters: Visual attention and applications," *ICIP*, no. 16, pp. 3085-3088, Nov. 2009.
- [6] S. Marat, et al., "Modelling Spatio-Temporal Saliency To Predict Gaze Direction For Short Videos," *IJCV*, no. 82, pp. 231-243, Mar. 2009.
- [7] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, pp. 1-15, 2009.
- [8] C. Li and A. C. Bovik, "Content-weighted video quality assessment using a three-component image model," *J. Electronic Imaging*, no. 19, Jan. 2010.
- [9] 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.
- [10] M. Kapfer and J. Benois-Pineau, "Detection of human faces in color image sequences with arbitrary motions for very low bit-rate videophone coding," *Pattern Recognition Letters*, vol. 14, no. 18, pp. 1503-1518, 1997.
- [11] G. Quénot, et al., "Rushes summarization by IRIM consortium: Redundancy removal and multi-feature fusion," *Proceedings of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops*, pp. 80-84, 2008.
- [12] 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.

- [13] A. Billot, I. Gilboa, and D. Schmeidler, "Axiomatization of an exponential similarity function," *Mathematical Social Sciences*, no. 55, pp. 107-115, 2008.
- [14] VQEG (Video Quality Experts Group), "Report on the Validation of Video Quality Models for High Definition Video Content," Report, 2010.
- [15] International Telecommunication Union, "ITU-R BT.500-11 Methodology for the subjective assessment of the quality of television pictures," Recommendation, 2002.
- [16] International Telecommunication Union, "ITU-T Rec. G.1050 Network model for evaluating multimedia transmission performance over Internet Protocol," Recommendation, 2007.
- [17] 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.