Reference Free Quality Metric using a Region-Based Attention Model for MPEG Compressed Videos

Rémi Barland, Abdelhakim Saadane, IRCCyN-IVC, UMR CNRS n°659, École polytechnique de l'université de Nantes, Rue Christian Pauc, La Chantrerie, BP50609, 44306 Nantes, France, Mail: {remi.barland, abdelhakim.saadane}@univ-nantes.fr

Abstract

Currently, at low bit rates, the MPEG compression coding can generate some impairments, which can affect the visual video quality. These artifacts such as the blocking, blurring and ringing effects can be exploited in order to design No Reference video quality metric. In this paper, we propose to use an importance map extracted from a region based attention model, to weight distortion measures derived from previous works [1]. This perceptual map is generated from the processed image, combining on the one hand, the simulated results of visual human cell responses and on the other hand, the information of a spatial segmentation. The contribution of these weights is firstly evaluated, in the case of the quality assessment of JPEG and JPEG-2000 compressed images: this perceptual validation allows to assure the relevance of proposed distortion measures. Then, the performance of these combined measures is performed using a database composed of MPEG compressed videos. High correlation between the objective scores of the proposed metric and the subjective assessment ratings has been achieved.

1. Introduction

The new compression techniques have lead to the emergence of new services, such as digital video broadcasting or streaming. However, because of limitations of network bandwidth or storage capacity, these new technologies require a tradeoff between the perceptual quality of the video sequence and the quantity of information transmitted or saved. At low bit rates, the coding techniques such as MPEG or H26-X can create some impairments, which can cause an embarrassment for a human user. To evaluate the contribution brought by an efficient compression technique, the perceptual video quality must be assessed.

The subjective assessment is the reference method to define the perceptual quality of an image or a video sequence. It consists of experiments, where a panel of human observers judges the visual quality of the input video. The Mean Opinion Scores (MOS) corresponding to each test input are the results issued from these subjective tests. The conditions of observations, the choice of observers, the test material, are specified in some recommendations [2-4], proposed by the International Telecommunication (ITU) or the Video Quality Expert Group (VQEG). However, these subjective tests are very long, expensive and difficult to practice. That is why, metrics are developed in parallel.

Most of proposed video quality assessment approaches require the original video sequence as a reference. The most widely used objective image quality metrics is Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE). However, the predicted scores do not well correlate with the subjective ratings: MSE and PSNR do not follow accurately the visual perception of human observers. Moreover, these metrics require the information contained in the original video, which is not possible for applications such as video broadcasting or streaming.

For such technologies, the No Reference (NR) assessment seems to be more suitable. Generally, the NR metrics combine individual distortion measures into a single one, in order to predict quality [5]. Considering the MPEG coded videos, the three most annoying distortions are the blocking, blurring and ringing effects. In the literature, several metrics detecting and measuring blocking effect are proposed [6-10]. On the other hand, blurring and ringing NR metrics are less treated [11-13]. To assess blindly MPEG compressed videos, Cheng et al [14] propose a new distortion measure for each previously cited impairment. The pooling model is based on a linear combination of the three distortion measures and an additional feature, the bit rate. In [15], Caviedes et al. compute blocking, ringing and corner outlining measures. These are first normalized individually, and then combined using Euclidean norm to obtain the predicted quality score. Farias et al [16] develop blocking, blurring and noise NR metrics and design models for overall annoyance of MPEG coded videos, using the Minkowski summation. However, none of the previously quoted models deals with the human visual system.

The goal and the novelty of this paper are to design a new NR metric applied to MPEG compressed videos and including some properties of the Human Visual System (HVS). Distortion measures described in [17], are individually weighted by an importance map issued from a simple algorithm of attention model. The relevance of these artifact metrics is firstly validated on databases containing JPEG and JPEG-2000 compressed images. Then, a NR video quality metric based on the overall annoyance, is described. The paper is designed as follows: section 2 describes the structure of NR distortion measure, while section 3 presents the proposed NR video quality metric. The performance evaluations are discussed in section 4 and conclusions are given in section 5.

2. NR Perceptual Distortion Measures

2.1 Generation of the Importance Map

The human observer selects Regions Of Interest (ROI), using two different mechanisms called bottom-up (signal-dependent) and top-down (task-dependent) controls. The proposed algorithm only incorporates the first mechanism. After a color transformation in the Krauskopf color space [18] (one

Achromatic A and two Chromatic Cr_1, Cr_2 components), the model (figure 1) computes on the one hand, stimulus salience from achromatic contrast using a center-surround simulation [19] and on the other hand, spatial features [20], which are known to influence visual attention.

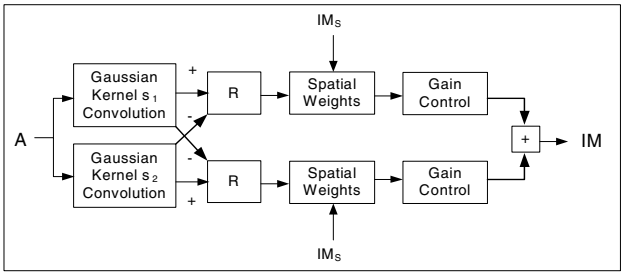

Figure 1: Block diagram of the overall importance map generation

To detect achromatic contrast, the behavior of the Classical Receptive Field (CRF) localized in the human visual cells, is emulated. The interactions between center-surround are computed as follows:

$$CS_{On/Off}=R(LPF(A,\sigma_C)-LPF(A,\sigma_S)) \quad (1)$$

$$CS_{OFF/ON}=R(LPF(A,\sigma_S)-LPF(A,\sigma_C)) \quad (2)$$

Where $R(x)=0$ if $x\leq 0$ otherwise $R(x)=x$. $LPF(A,\sigma)$ is a low-pass filter, defined by the convolution of the achromatic (A) image with a Gaussian kernel ($\sigma_C=0.4$ and $\sigma_S=2.4$).

These outputs are then weighted by a spatial importance map (IM_S) generated by the block diagram illustrated by figure 2. The algorithm begins by an unsupervised segmentation of color-texture regions [21] performed on A, Cr_1 and Cr_2 components. The segmented image is then analyzed by a number of different spatial features (shape, location, size, background), known to influence the visual attention [20]. An importance feature map for each considered factor is generated. The feature maps are combined by a Minkowski summation, to produce the spatial Importance Map (IM_S), which is used to weight results of Center-Surround interactions.

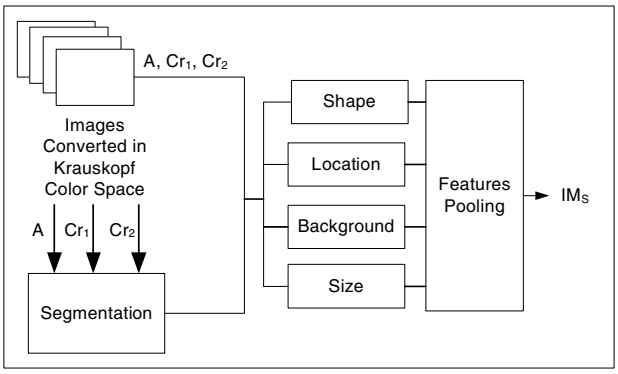

Figure 2: Block diagram of spatial importance map (IM_S) generation

This weighting is performed for each pixel (m,n) and is expressed by:

$$r_{On/Off}(m,n)=CS_{On/Off}(m,n).(1+IM_S(m,n)) \quad (3)$$

$$r_{Off/On}(m,n)=CS_{Off/On}(m,n).(1+IM_S(m,n)) \quad (4)$$

Divisive inhibition is applied to these responses. The normalized response is given by:

$$N(r)=\frac{r^2}{h*r^2+b^2} \quad (5)$$

Where r is the response at a given location, b is a saturation constant to prevent division by zero and h is a Gaussian kernel. The normalized responses of on-center-off-surround and off-center-on-surround receptive fields are then summed and normalized in order to obtain a two-dimensional map representing the conspicuous location.

2.2 NR Distortion Measures

The blocking artifact visually creates an artificial discontinuity between neighboring blocks in an image, caused by a severe and independent quantization of DCT coefficients of each block. This impairment can be amplified by the contrast between the neighboring blocks. Considering these facts, the local blocking measure LBM(k,l) for a block (k,l) can be formulated as follows:

$$LBM(k,l)=\frac{R_H(k,l)+R_V(k,l)}{2}.S(k,l) \quad (6)$$

where $R_H(k,l)$ (respectively $R_V(k,l)$) defines the contrast reinforcement produced by the horizontal (respectively vertical) neighboring blocks and S(k,l), the quantization severity.

For the block (k,l), the quantization severity is defined as:

$$S(k,l)=\frac{1}{1+STD(k,l).a} \quad (7)$$

Where STD(k,l) is the standard deviation of the block (k,l), and a, a constant. The horizontal contrast reinforcement is computed as follows:

$$R_H(k,l)=1+C_H(k,l) \quad (8)$$

With

$$C_H(k,l)=\frac{|A(k,l)-A(k,l-1)|+|A(k,l)-A(k,l+1)|}{2.\max\{|A(k,l)-A(k,l-1)|;|A(k,l)-A(k,l+1)|\}} \quad (9)$$

Where A(k,l) (respectively A(k,l-1) and A(k,l+1)) is the mean value of the achromatic component for the block (k,l) (respectively block (k,l-1) and block (k,l+1)). The vertical contrast reinforcement is defined with the same formula, but considering the vertical neighboring blocks.

The final blocking measure BlM is obtained using a Minkowski summation:

$$BM=\left(\frac{1}{NB_V.NB_H}\sum_{k=1}^{NB_V}\sum_{l=1}^{NB_H}(IM(k,l).LBM(k,l))^p\right)^{1/p} \quad (10)$$

Where NB_V (respectively NB_H) represents the number of vertical (respectively horizontal) blocks in the processed image, IM is the importance map described in the section 2.1.

The blurring effects correspond to a total distortion on the whole image, characterized by an increase of the spread of edges and spatial details, while the ringing effect locally produces haloes and/or rings near sharp object edges in the image. The blurring measure (BlM) is formulated with a ratio using spatial information, pixel activity and weighting given by the Importance Map (IM) values:

$$BlM = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} IM(i,j).A'_{Edge}(i,j).I_A^2(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} IM(i,j)A_{Edge}(i,j).I_A^2(i,j)} \cdot \frac{\frac{N(A'_{Edge})}{MxN}}{\frac{N(A_{Edge})}{MxN}} \quad (11)$$

Where $I_A(i,j)$ is the pixel (i,j) intensity of the Achromatic (A) component of size MxN pixels. A_{Edge} is the binary image resulting from A edge detection. A'_{Edge} is the A_{Edge} complementary image. $N(A_{Edge})$ (respectively $N(A'_{Edge})$) is the number of non-null pixel values of A_{Edge} (respectively A'_{Edge}).

Before measuring the ringing distortion, the areas around edges, called "ringing regions", must be identified ($A_{Ringing\ Mask}$ image). These are computed by using a binary "ringing mask" on the current image, resulting from the detection and the dilatation of strong edges. Then, a measure of ringing artifact is computed, defined by the ratio of regions activities of middle low and middle high frequencies, localized in these "ringing regions". Each part of the defined ratio is locally weighted by the IM values:

$$RM = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} IM(i,j).A'_{RM\ Edge}(i,j).I_{ARM}^2(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} IM(i,j).A_{RM\ Edge}(i,j).I_{ARM}^2(i,j)} \cdot \frac{\frac{N(A'_{RM\ Edge})}{N(RingingMask)}}{\frac{N(A_{RM\ Edge})}{N(RingingMask)}} \quad (12)$$

Where $I_{ARM}(i,j)$ is the pixel (i,j) intensity of $A_{Ringing\ Mask}$ image of size MxN pixels. $A_{RM\ Edge}$ is the binary image resulting from $A_{Ringing\ Mask}$ image edge detection. $A'_{RM\ Edge}$ is the combination (XOR operator) of $A_{RM\ Edge}$ complementary binary image and "Ringing Mask" binary image. $N(A_{RM\ Edge})$ (respectively $N(A'_{RM\ Edge})$ or N(Ringing Mask)) is the number of non-null pixel values of $A_{RM\ Edge}$ (respectively $A'_{RM\ Edge}$ or Ringing Mask) binary image.

3. NR Video Quality Metric

The proposed video quality metric is designed as follows: after a conversion in a perceptual color space of each image of the video sequence, the three distinct distortion measures of blocking (BM_i), blurring (BlM_i) and ringing (RM_i) effects, described in the section 2.2, are computed. Then, for each impairment, a temporal pooling using a Minkowski summation is performed.

Then, the final predicted quality score (pMOS) for an entire video sequence can be obtained by a linear combination of temporal distortion measures:

$$pMOS = a_0 + a_1.BM + a_2.BlM + a_3.RM + a_4.BlM.BM + a_5.BlM.RM \quad (13)$$

Where a_i, i=0..5 are the weights to be optimized. The three first terms of the final pooling model correspond to the distortion caused by each artifact, while the others define the combined actions of blocking/blurring and blurring/ringing effects.

4. Experiments and Results

In order to confirm the relevance of the three proposed distortion measures, a conjoint measure of blocking and blurring effects is validated using a database composed of JPEG compressed images. Indeed, the blurring and blocking artifacts are the two most annoying impairments engendered by JPEG coding, at low bit rates. On the other hand, the blurring and ringing effects are the visual distortions observed by human observers for JPEG-2000 compressed images. That is why, a conjoint measure of blurring and ringing is validated using a database containing JPEG-2000 compressed images.

Currently, VQEG proposes some statistical tools [22], in order to quantify the performance of a quality metric: the Pearson linear correlation and the Root Mean Square Error (RMSE) for the accuracy, the Spearman rank order correlation for the monotonicity, the outlier percentage for the consistency and the Kappa coefficient for the agreement. These statistical tests are performed comparing the predicted quality scores to the MOS of the input images.

The image database we use in the experiments is from [23]. It consists of 29 original high-resolution 24-bits/pixel RGB color images (typically 768x512) and their JPEG and JPEG-2000 compressed versions with different compressed ratios. The bit rates used for compression are in the range of 0.03 to 3.2 bits per pixel, chosen such that the resulting distribution of quality scores is roughly uniform over the entire range. About 25 human observers assess the quality of each image as "Bad", "Poor", "Fair", and "Good" and "Excellent". The raw scores for each subject are normalized by the mean and the variance of that subject and then scaled and shifted by the mean and the variance of the entire subject pool to the full range (1 to 100). Mean scores are then computed for each image after removing outliers.

4.1 Validation of Blocking, Blurring Measures

The conjoint measure (CM_1) of blocking (BM) and blurring (BlM) effects consists of a linear combination of the first order terms of the blocking and blurring measures, plus the associated crossed term. Hence, the predicted quality score of an image may be written as:

$$CM_1 = a_0 + a_1.BM + a_2.BlM + a_3.BM.BlM \quad (14)$$

Where a_i, i=0..3 are the weights to be optimized. The JPEG image database is divided in two parts: one for training and the second one for the test (75 images). The table 1 presents the different results of proposed artifacts measures, while the table 2 presents the results of distortions measures computed without the weights introduced by the importance map.

	Pearson	RMSE	Spearman	Kappa	Outlier
BM	0.935	0.610	0.937	0.696	0.133
BlM	0.925	0.650	0.926	0.659	0.133
CM_1	0.964	0.482	0.948	0.749	0.053

Table 1: Correlation results of the proposed blocking (BM), blurring (BlM) and conjoint (CM_1) measures.

	Pearson	RMSE	Spearman	Kappa	Outlier
BM	0.91	0.865	0.92	0.448	0.12
BlM	0.91	0.712	0.90	0.518	0.133
CM_1	0.927	0.795	0.919	0.519	0.12

Table 2: Correlation results of the proposed blocking (BM), blurring (BlM) and conjoint (CM_1) measures computed without the weights of the importance map.

The correlation results issued from the separate use of each distortion measure allows to demonstrate the contribution of the importance map: the Pearson correlation obtained by the proposed weighted artifact measures indicates a better ability to predict subjective scores with a minimum average error than the same measures, neglecting the weighting. The monotonic relationship (Spearman rank-order correlation) is respected. The small obtained outlier ratio means that the proposed distortion metrics have a good ability to provide consistently accurate predictions for all types of compressed images and not fail excessively for a subset of images. The Kappa coefficient is a measure of agreement. Usually, a Kappa coefficient superior to 0.4 is a good value; so the proposed artifact metrics obtain a good agreement between subjective and predicted scores. Comparing separately these tables, a second important fact appears: the proposed pooling model increases largely the prediction performances.

4.2 Validation of Blurring, Ringing Measures

The conjoint measure (CM_2) of blurring (BlM) and ringing (RM) effects consists of a linear combination of the first order terms of the blurring and ringing measures, plus the associated crossed term. Hence, the predicted quality score of an image may be written as:

$$CM_2 = a_0 + a_1.BlM + a_2.RM + a_3.BlM.RM \qquad (15)$$

Where a_i, i=0..3 are the weights to be optimized. The JPEG-2000 image database is divided in two parts: one for training and the second one for the test (84 images). The table 3 presents the different results of proposed artifacts measures, while the table 4 presents the results of distortions measures computed without the weights introduced by the importance map.

	Pearson	RMSE	Spearman	Kappa	Outlier
BlM	0.895	0.855	0.958	0.381	0.095
RM	0.887	0.833	0.937	0.351	0.107
CM_2	0.918	0.692	0.944	0.668	0.059

Table 3: Correlation results of the proposed blurring (BlM), ringing (RM) and conjoint (CM_2) measures.

	Pearson	RMSE	Spearman	Kappa	Outlier
BlM	0.848	0.912	0.886	0.302	0.190
RM	0.802	0.966	0.848	0.263	0.226
CM_2	0.866	0.819	0.903	0.60	0.142

Table 4: Correlation results of the proposed blurring (BlM), ringing (RM) and conjoint (CM_2) measures computed without the weights of the importance map.

The comparison between these two tables allows to confirm the conclusions observed in section 4.1. Hence, both of these experimental validations demonstrate the relevance of the distortion measures proposed in this paper.

4.3 Validation of the NR video quality metric

The proposed video quality metric is tested using a video database. This set consists of 35 video sequences derived from 7 original scenes. These clips contain a wide range of entertainment content from TV news to sport event. Each original video sequence is compressed using XVID coder (a free MPEG-4 coder) at five different bit rates ranging from 1.0 Mbps to 5 Mbps. Subjective ratings of the compressed videos are obtained using psychophysical experiment and following the recommendation ITU-T BT.500.10 [2]. In our experiment, the database is divided randomly into two sets: 3 training videos and 4 testing videos, together with their compressed versions.

The weights a_i of the pooling model (section 3, equation 13) are estimated from training videos using minimal mean squared error estimate between quality predictions and subjective scores. Then, the proposed trained quality metric is validated on the test database. The quality predictions resulting from this assessment are compared with scores given by human observers. The table 5 presents the correlation results of the proposed NR video quality metric (M_1) described in section 3. The other results correspond to a metric (M_2) based on the same distortion measures, but not taking into account the weights of the importance map.

	Pearson	RMSE	Spearman	Kappa	Outlier
M_1	0.939	0.561	0.959	0.666	0.15
M_2	0.892	0.792	0.920	0.533	0.2

Table 5: Correlation results of the NR video quality metric (M_1) and a second one (M_2), only based on the distortion measures (the weights of the importance map are neglected)

The different performance metrics of VQEG recommendations are satisfied, which demonstrates the efficiency of the proposed metric, in the case of the video quality assessment. The integration of a perceptual importance map significantly increases the correlation between the predicted and subjective quality scores.

5. Conclusions

In this paper, we have presented a new reference free quality metric to assess the quality of MPEG compressed video sequences. The proposed method is based on the exploitation of separate distortion measures, specifically tuned to certain type of distortion (blocking, blurring and ringing). Each artifact measure is weighted by a perceptual importance map. This is generated by a combination of two distinct mechanisms: on the one hand, a stimulus salience from achromatic contrast using a center-surround simulation and on the other hand, spatial features (shape, size, background, location), which are known to influence visual attention. Each algorithm of distortion measure is previously validated with subjective ratings, which assures the relevance and the efficiency of the proposed approach. The experimental results, computed from a MPEG video database, confirm the efficiency of the NR video quality metric based on the combination of these impairment measures.

References

1. Barland, R. and A. Saadane. A New Reference Free Approach for the Quality Assessment of MPEG Coded Videos. in 7th International Conference Advanced Concepts for Intelligent Vision Systems (ACIVS). 2005. Antwerp, Belgium: Springer-Verlag GmbH.
2. ITU-R Recommendation BT.500-10, Methodology for the Subjective Assessment of the Quality of Television Pictures. 2000, ITU: Geneva.

3. ITU-T Recommendation P.910, Subjective Video Quality Assessment Methods for Multimedia Application. 1999, ITU: Geneva, Switzerland.
4. ITU-T Recommendation P.920, Interactive Test Methods for Audiovisual Communications. 2000, ITU: Geneva, Switzerland.
5. Farias, M., S.K. Mitra, et al. Perceptual contributions of blocky, blurry and noisy artifacts to overall annoyance. in International Conference on Multimedia and Expo. 2003. Balitmore, Maryland USA.
6. Triantafyllidis, G., D. Tzovaras, et al., Blocking Artifact Detection and Reduction in Compressed Data. IEEE Transactions on Circuits and Systems for Video Technology, 2002. 12(10): p. 877-891.
7. Vlachos, T., Detection of Blocking Artifacts in Compressed Video. Electronics Letters, 2000. 36(13): p. 1106-1108.
8. Wu, H.R., Z. Yu, et al. Impairment metrics for MC/DPCM/DCT encoded digital video. in Picture Coding Symposium. 2001. Seoul, Korea.
9. Wang, Z., A. Bovik, et al. Blind Measurement of Blocking Artifacts in Images. in IEEE International Conference on Image Processing. 2000.
10. Gao, W., C. Mermer, et al., A De-Blocking Algorithm and Blockiness Metric for Highly Compressed Images. IEEE Transactions on Circuits and Systems for Video Technology, 2002. 12(12): p. 1150-1159.
11. Marichal, X., W.-Y. ma, et al. Blur Determination in the Compressed Domain Using DCT Information. in IEEE International Conference on Image Processing. 1999. Kobe, Japan.
12. Marziliano, P., F. Dufaux, et al. Perceptual metrics for JPEG2000 Coded Images. in MDWESPIC. 2003.
13. Caviedes, J. and S. Gurbuz. No-Reference Sharpness Metric Based on Local Edge Kurtosis. in IEEE International Conference on Image Processing. 2002. Rochestern, New York, USA.
14. Cheng, H. and J. Lubin, Reference free objective quality metrics for MPEG coded video. Visual Communications and Image Processing, 2003.
15. Caviedes, J. and J. Jung. No-Reference Metric for a Video Quality Control Loop. in World Multi-Conference on Systems Cybernetics and Informatics Broadcasting Convention. 2001.
16. Farias, M. and S.K. Mitra. No Reference Video Quality Metric Based on Artifact Measurements. in International Conference on Image Processing. 2005. Genoa, Italy.
17. Barland, R. and A. Saadane. A Reference Free Quality Metric for Compressed Images. in Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics. 2006. Scottsdale, Arizona, USA.
18. Williams, D.R., J. Krauskopf, et al., Cardinal Directions of Color Space. Vision Research, 1982. 22: p. 1123-1131.
19. Parkhurst, D. and E. Niebur, Texture Contrast Attracts Overt Visual Attention in Natural Scenes. European Journal of Neuroscience, 2004. 19: p. 783-789.
20. Osberger, W. and A.M. Rohaly. Automatic Detection of Regions of Interest in Complex Video Sequences. in SPIE Human Vision and Electronic Imaging. 2001. San Jose, California, USA.
21. Deng, Y. and B.S. Manjunath, Unsupervised Segmentation of Color-Texture Regions in Images and Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. 23(8): p. 800-810.
22. Rohaly, A.M., P. Corriveau, et al. Video Quality Experts Group: Current Results and Future Directions. in Proceedings of Visual Communications and Images Processing. 2000.
23. Sheikh, H., Z. Wang, et al., LIVE Image Quality Assessment Database. http://live.ece.utexas.edu/research/quality.

Author Biography

Remi Barland is a PhD student in image processing from the University of Nantes (France). He obtained in September 2003, the engineering degree in computer sciences (Multimedia and Vision). Teaching assistant at the University of Nantes, its current centers of research are the quality assessment without reference for image and video processing.

