Insertion of Impairments in Test Video Sequences for Quality Assessment Based on Psychovisual Characteristics

J.P. López, J. A. Rodrigo, D. Jiménez and J. M. Menéndez

Abstract — **Assessing video quality is a complex task. While most pixel-based metrics do not present enough correlation between objective and subjective results, algorithms need to correspond to human perception when analyzing quality in a video sequence. For analyzing the perceived quality derived from concrete video artifacts in determined region of interest we present a novel methodology for generating test sequences which allow the analysis of impact of each individual distortion. Through results obtained after subjective assessment it is possible to create psychovisual models based on weighting pixels belonging to different regions of interest distributed by color, position, motion or content. Interesting results are obtained in subjective assessment which demonstrates the necessity of new metrics adapted to human visual system.**

Index Terms—**Artifact, database, distortion, subjective correlation, video quality assessment.**

I. Introduction

Video quality measurement plays a decisive role in most image processing applications, especially related to the Internet and mobile video consumption. Research workgroup "Qualinet" [1] defines Quality of Experience (QoE) as the degree of delight or annoyance of the user of an application or service. This definition does not perfectly comply with classic metrics, such as PSNR (Peak Signal-to-Noise Ratio) or MSE (Mean Square Error), which offer an objective result which differs from the subjective assessment that is the key factor in QoE and perceived quality [2].

Automatically assessing objective image quality that is consistent with human perception in computer vision is a challenge. Image and video sequences need to formulate high-level semantic features to separate the subject and the background and the pixels regions of interest which are more sensitive to human eye [3].

A collection of video artifacts are consequence of compression and video encoding. A classification of video artifacts are collected in Boev et al. articles [4], especially taking into account the distortions of blocking, blurring and ringing, which affect to motion and content itself. These artifacts must be quantified and correlated to human perception.

Understanding experience and finding user satisfaction over objective quality is important [5]. Sometimes this purpose is based only on objective features which are not exactly connected to subjective response. Parameters having an influence on visual quality must be quantified to identify the video sequence as close to the human vision as possible, creating content-weighted [6] to analyze video quality metrics based on psychovisual models.

It is important to have information on what observer really visualize on the screen. Researchers use their own visual systems to detect emerging patterns and trends in graphical representations of datasets, determining where best to employ further statistical analysis. The spatial relationship between different fixations is not a construction of the analysis used; rather, it is their real spatial separation relative to the stimulus used. This makes the analysis of fixation data by fixation maps both powerful and more straight forward. Visualization is only one application of fixation map analysis and is the least quantitative of the applications described here. It does, however, provide a useful and intuitive means of describing and communicating the underlying patterns of otherwise overwhelming datasets [7].

Finding the observer's regions of interest is a key factor to weight the value derived from a quality metric. The graphical representations of the data described are an immediate and powerful demonstration of a key feature of eye movement data: that the fixations are not spread evenly or randomly over the stimulus but cluster into regions of interest according to the features of the stimulus.

This paper describes the methodology for developing databases of video sequences which allows the weighting of each artifact for video quality assessment. These weights may be used for the definition of a psychovisual model, which is required to get a good correlation with the subjective response of the human visual system. This model is based on analyzing five video features: position of the pixels in the frame, motion, level of detail or spatial complexity, face detection and color. These five characteristics define the observer's visual impact allowing the creating a mask that decides the importance of every single pixel. Therefore, quality must be evaluated in the areas which define the regions of interest for human eye. For evaluating the importance of the pixels derived from these factors, sequences have been artificially impaired in order to set the impairments in a determined area of the image, depending on the component being analyzed. This area could be the one where pixels with a high motion can be found, the high-frequency areas or a region of a geometrical division, i.e. pixels located around a corner or a lateral side of the frame, or where a face is detected. This database will be used for subjective quality assessment. Then, results from these subjective tests will be used for weighting the areas of interest when the model to full-reference or no-reference objective metrics is implemented.

II. Methodology for Creating Artificially Impaired Video Sequences

Most studies introduce impairments in the image by encoding the video sequence at different bitrates. This methodology works when assessing the quality of an encoder but the perceived impairment could be hardly distinguished from the impairment hidden by the effect of motion, position of the pixel or high frequencies. Codification introduces impairment in the image in the same way, but in order to analyze individually each of these factors, it is necessary to generate video sequences which have previously been artificially impaired.

To this effect, the methodology must be based on generating masks for each video sequence to cover the pixels where motion is detected, and setting a comparison to the sequence with opposite situation. Just the same occurs in the other two cases. The pixels included in the mask showed the original image and the rest of the image was covered by the impaired image with a sequence containing the artifact to be evaluated, for example, blurring, ringing or blocking, or in general by a sequence encoded at low bitrates, following the procedure in Fig. 1.

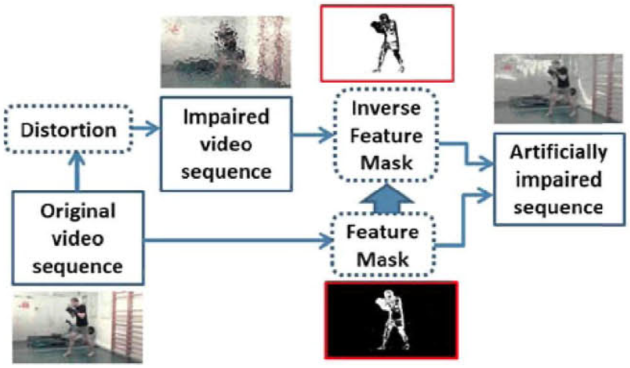


Fig. 1. Scheme of artificially impaired video sequences generation

An example of how each sequence is created is collected in Fig. 2. The sequence combines two different sequences, one with high quality and the other with poor quality, which contains the impairment. Each of the sequences is located in areas where the image feature to analyze is defined.

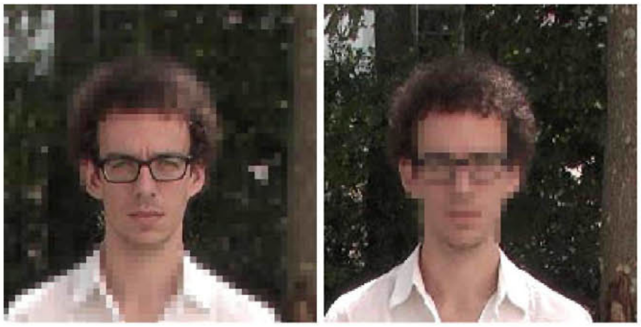


Fig. 2. Example of Sequence "Umbrella" (frame 12) with impairment located in faces ROI and its inverse

III. Impairment Generation and Distortions

The first step is defining a collection of distortions to analyze in video quality assessment for simulating each kind of artifact. From the list of artifacts introduced previously, these ones are selected for distortion simulated. The result of the usage of these techniques is the "Impaired video sequence" which appeared in scheme in Fig. 1.

- Gaussian low-pass filters for blurring simulation.
- JPEG2000 encoder for ringing simulation.
- An 8x8 mosaic filter for MPEG-2 blocking artifact simulation.
- Other encoders at low bitrates for general distortion simulations.

These techniques connect the block diagram of defining an impaired video sequence included in Fig. 3 with the complete scheme for generating artificially impaired sequence with distortion located in a ROI or its corresponding inverse ROI.

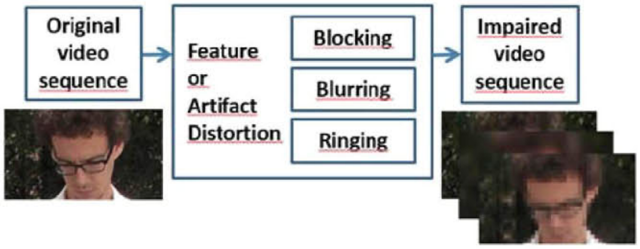


Fig. 3. Block diagram for impairment insertion process

Examples of distortions are collected and explained in Fig. 4. Distortion is applied to the whole sequence equally. This sequence will be discriminated by the mask to locate distortion in only a part of pixels of the whole frame.

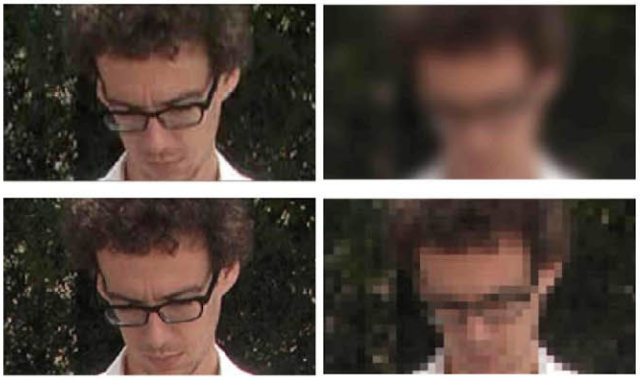


Fig. 4. Detail of frame 125 of sequence "Umbrella" after applying different distortions: original, blurring (Gaussian filter), ringing (JPEG200 encoding), blocking (8x8 mosaic filter).

IV. Definition of Masks For ROI's Description

Based on the five different factors that affect to visual impact and define regions of interest in the video frame, five types of binary masks are developed to quantify the perceived quality in observer's eye. Subjective assessment was used using the databases created with this masks for demonstrating the importance of this methodology, because allows the isolation of each individual artifact or distortion. The different kinds of masks are:

- **Motion mask.** This kind of mask and its inverse is based on temporary information algorithms and object detection which allow analyzing if motion affects to vision or if, on the other hand, it provokes hidden impairments.
- **Level of detail or spatial complexity mask**. Based on the detection of pixels in edges and high frequencies region, this mask is generated by the usage of Canny algorithm [8].
- **Position masks.** The frame is divided into different sections to analyze the importance of pixels derived from their position into the video frame. Three kinds of pixels are defined as belonging to center, lateral or corner areas. Human vision is usually focused on central areas, especially when video is played into wide screens.
- **Face detection.** After studies of eye movement and fixation maps included in [5], it is noticeable that visual impact is important when a human face is detected in the picture. Pixels included on an area where a human face is detected belong to the main regions of interest and should be analyzed as an important area for quality assessment. Haar algorithm is used for face detection.
- **Color masks.** The content itself is primary for understanding human perception [3]. Consequently, human eye is not equally sensitive to different colors, because of the effect of photoreceptors [9]. Color, and additionally bright and contrast, must be analyzed to determine the colors which make artifacts more visual for human eye.

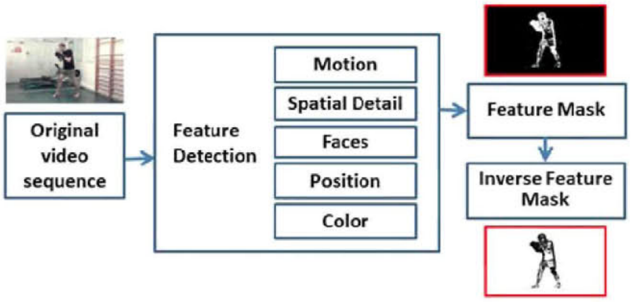


Fig. 5. Diagram for feature mask sequence and its inverse mask

A software specifically used for this purpose was developed containing the necessary tools for analyzing video sequences in each of these five image features, using C# programming language environment. Libraries included in Emgu CV [10] and Open CV platform [11] were used for the process, importing Haar cascade techniques for facial detection included in "haarcascade_frontalface_alt2.xml" and "haarcascade_profileface.xml" files.

A. *Motion masks*

For motion detection, temporal information in consecutive frames is scrutinized. The pixels with variation are analyzed in order to conclude subjectively how pixels with motion weight to different algorithms for quality assessment.

$$Pix_i(x,y) \in .Mask, if \left[F_i(x,y) - F_{framei+1}(x,y)\right] > 0 \quad (1)$$

$Pix_i(x,y)$ represents the value of the pixel located in coordinates (x, y), belonging to the "i" frame in the video sequence,. F indicates the function of that same frame, in this case, the luminance.

The example in Fig. 6 shows how pixels in the motion region of interest are detected over the barrier which is the only element within motion, and it is independent from other effects, such as edges and detailed information.

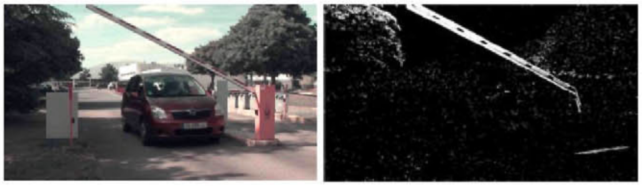


Fig. 6. Barrier sequence after applying motion mask (frame 270)

B. *Level of detail and spatial complexity and masks*

Textures, edges and objects in motion are the source of hiding impairments in cases of blocking algorithms but the opposite happens with blurring algorithms. Canny algorithm is used to create a binary mask which separates the homogenous areas from the high-frequencies area.

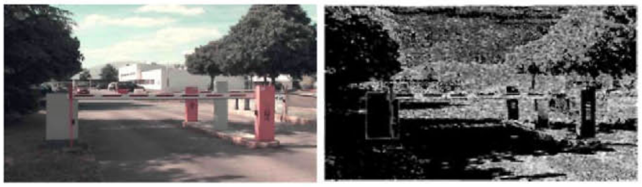


Fig. 7. Canny algorithm for spatial complexity mask

C. *Position of pixels in spatial sections*

The image is divided into nine sections (Fig. 9), as indicated in research by Nojiri et al. [12]. The idea of these sequences is to analyze the focus on different sequences, independently of their content. Then, it is possible to analyze if there is a high influence on the fact that a pixel is located on a corner, or on a lateral or central area (Fig. 8).

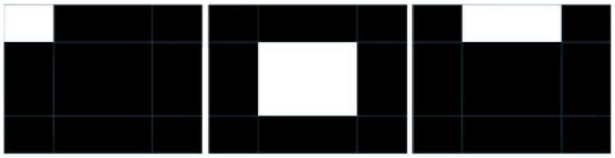
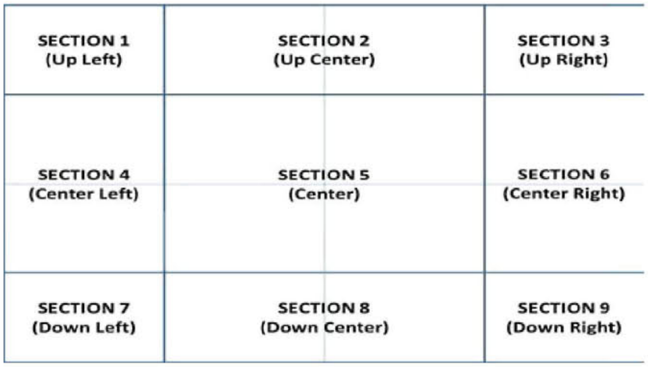


Fig. 8. Corner, center and lateral masks



SECTION 1 **(Up Left)**	**SECTION 2** **(Up Center)**	**SECTION 3** **(Up Right)**
SECTION 4 **(Center Left)**	**SECTION 5** **(Center)**	**SECTION 6** **(Center Right)**
SECTION 7 **(Down Left)**	**SECTION 8** **(Down Center)**	**SECTION 9** **(Down Right)**

Fig. 9. Distribution of sections to divide video frames

D. *Face detection*

Haar algorithm based on a boosted cascade of simple features [13] is used for face detection. Pixels belonging to a facial region of interest are important because are part of the visual impact derived from user experience.

When analyzing video quality on a video sequence, pixels in face regions must be treated separately because the visual quality is increased in areas where visual impact is higher. Observer tends to be more exigent in areas where visual attention is focused, concretely in human faces regions.

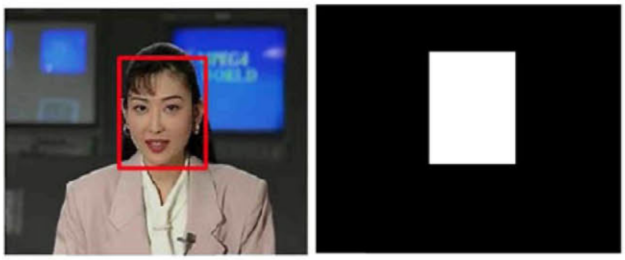


Fig. 10. Face detected in "Akiyo" sequence (left) and its facial mask (right)

For face detection Haar techniques in cascade included in files "haarcascade_frontalface_alt2.xml" and "haarcascade_profileface.xml" were used as the core of this platform.

E. *Range of colors detection*

Through the optics of the eye, two different types of photoreceptors, rods and cones, exist. Cones are responsible of the photopic vision and the distinguish colors, with different normalized spectral sensitivities to spectrum wavelengths for the three cone types: L-cones, M-cones, and S-cones [9].

Some artifacts allow the visualization of artifacts more easily by human eye. For that reason, range of colors should be analyzed to determine the weight of this factor to visual algorithms. Three ranges of colors are developed for masks: red, blue and green. The mask contains pixels corresponding to the determined color and the ones with a similarity related to a threshold, as seen in Fig. 11.

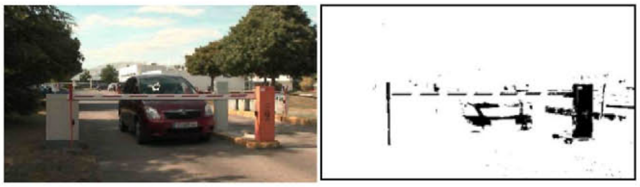


Fig. 11. Range of red colors mask in sequence "Barrier" (frame 125)

V. Results Derived From Subjective Tests

Methodology followed in subjective tests is described in [14], with tests sequences included in three databases: SVT, [15], NAMA3DS1-COSPAD1 [16] and EBU [17]. Assessment follows the indications included in recommendation ITU P.910 [18] such as the necessity of including a collection of sequences with enough variety. For this purpose, spatio-temporal diagram is included in Fig. 12.

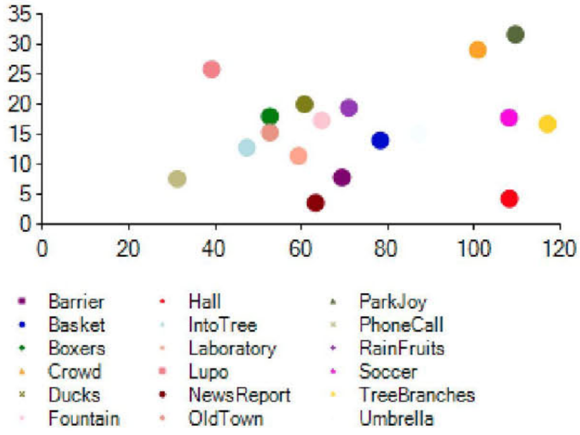


Fig. 12. Spatio-temporal diagram for the sequences used

After introducing the video sequences used for the study, results based on subjective tests are included and analyzed in this section. For these three examples, the combination of sequences includes a video sequence encoded at 75Mbps in H.264, which is a high quality, and poor quality derives from a sequences encoded at 500Kbps. All sequences are in 1920x1080 progressive formats at 25fps. "D" indicates that the distortion is located in a determined ROI and "Inv." indicates the inverse sequence of previous case of study.

First example refers to sequence "News Report" which represents two talking heads from a news channel. When distortion is located in the areas corresponding to human faces, the subjective MOS values are lower (1.33) than when located in the rest of the picture and faces appear sharp (3.78). This effect is completely opposite to PSNR (46.82 vs. 34.52) or MSE's behavior (0.10 vs. 2.30) as seen in Table 1.

Table 1. Results for impairment located in faces ROI for artificially distorted sequence "News Report"

Sequence	FR Metric	H.264		Impairment located in Faces ROI.	
		75Mbps	500Kbps	D.	Inv.
News Report	PSNR	47.93	37.58	46.82	34.52
	Blur	0.44	3.63	0.38	5.17
	MSE	0.67	1.93	0.10	2.30
	MOS	4.81	1.54	1.33	3.78

A similar situation occurs when analyzing motion in "Barrier sequence" as seen in values collected in Table 2.

Table 2. Results for impairment located in motion ROI's for artificially distorted sequence "Barrier"

Sequence	FR Metric	H.264		Impairment located in Motion ROI.	
		75Mbps	500Kbps	D.	Inv.
Barrier	PSNR	49.82	33.19	39.85	34.24
	Blur	0.27	8.36	1.97	6.24
	MSE	0.51	3.34	0.359	2.98
	MOS	4.77	1.33	3.11	3.89

Finally, Table 3 contains the results when comparing distortions located in a corner, a lateral or the center area in sequence "Crowd". For observers, a high distortion located in a corner is insignificant. On the other hand, when impairment is located in central area, opinion scores decrease to 1.44. The PSNR and MSE reveals the distortion related to the size of the impaired area, while the influence in human eye is related to the position of that impaired area.

Table 3. Results for impairment located in position ROI's for artificially distorted sequence "Crowd"

Seq.	FR Metric	H.264		Impairment located in Position ROI's					
		75 Mbps	500 Kbps	Center		Lateral		Corner	
				D.	Inv.	D.	Inv.	D.	Inv.
Crowd	PSNR	34.33	25.34	30.74	26.82	33.87	26.00	35.95	25.88
	Blur	3.44	22.55	6.27	15.33	2.60	19.44	0.95	22.47
	MSE	3.55	8.76	2.30	6.21	1.21	7.30	0.64	7.87
	MOS	4.68	1.22	1.44	2.44	3.78	1.33	4.11	1.22

VI. CONCLUSIONS

Characterization of video sequences to concrete artifacts and definition of visual attention models is a key factor when developing a useful database for assessing quality.

Video database developed and methodologies used allow the research in the field of video quality assessment. This database is useful in subjective tests in order to decide the human eye's response to impairments. Due to the difficulty of isolating impairment to analyze individually each artifact, this methodology is necessary for characterizing the video sequence and consequently weighting the impact of pixels affecting to a quality algorithm. Pixel-based metrics based on difference with a reference does not correspond perfectly to human eye as demonstrated in results derived from subjective tests. Motion, spatial complexity, position of the pixel in the frame, face detention or a range of color is primary for human eye's final response. This response is the bases of QoE and consequently of viewer's satisfaction, which is the real final purpose of this work. When understanding the human perception, the correlation may be assured for objective algorithm to evaluate quality and subjective scores.

For future work, the development of psychovisual models with weighted pixels based on these image features is primary, with the conclusions obtained through this study. This models of weighting will be used for impact over metrics and mathematical algorithms in order to assess the quality of a video sequence when analyzing a concrete artifact or distortion. The model should be used especially in no-reference metrics, additional to common full-reference algorithms.

ACKNOWLEDGMENT

The work developed in Universidad Politécnica de Madrid performed in the framework of project TEC2012-38402-C04-01 HORFI, which is partially funded by the Spanish Ministry of Science and Innovation. The authors would like to acknowledge the Spanish broadcaster RTVE and University of Nantes for the cession of the stereoscopic database.

REFERENCES

[1] Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems andServices (COST Action IC 1003), Patrick Le Callet, Sebastian Möller andAndrew Perkis, eds., Lausanne, Switzerland, Version 1.2, March 2013.

[2] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). The SSIM Index for Image Quality Assessment. https://ece.uwaterloo.ca/~z70wang/research/ssim/.

[3] Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. In *Computer Vision–ECCV 2008* (pp. 386-399). Springer Berlin Heidelberg.

[4] Boev, Atanas, Danilo Hollosi, and Atanas Gotchev. "Classification of stereoscopic artefacts." MOBILE3DTV Project 216503 (2008).

[5] Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. Behavior Research Methods, Instruments, & Computers, 34(4), 518-528.

[6] Li, Chaofeng, Alan Conrad Bovik. "Content-weighted video quality assessment using a three-component image model." Journal of Electronic Imaging 19.1 (2010): 011003-011003.

[7] Wooding, D. S. (2002, March). Fixation maps: quantifying eye-movement traces. In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 31-36). ACM.

[8] Canny, J. (1986). A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (6), 679-698.

[9] Winkler, S. (1999). Issues in vision modeling for perceptual video quality assessment. Signal Processing, 78(2), 231-252.

[10] Emgu, C. V. (2013). Emgu CV: OpenCV in .NET (C#, VB, C++ and more). Online: http://www.emgu.com

[11] Bradski, G., & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.".

[12] Nojiri, Y., Yamanoue, H., Ide, S., Yano, S., Okana, F. (2006). Parallax distribution and visual comfort on stereoscopic HDTV. In Proceedings of IBC (No. 3, pp. 373-380).

[13] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-511). IEEE.

[14] López, J.P., Rodrigo, J.A, Jiménez, D. and Menéndez, J.M. (2015). Subjective Quality Assessment in Stereoscopic Video Based on Analyzing Parallax and Disparity. IEEE Proceedings on ICCE 2015.

[15] L. Haglund, ”The SVT High Definition Multi Format Test Set.“ [Online]. Available at: ftp://vqeg.its.bldrdoc.gov.

[16] Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Le Callet, P., ... & García, N. (2012, July). NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on (pp. 109-114). IEEE.

[17] European Broadcasting Union (EBU). Test sequences. https://tech.ebu.ch/testsequences.

[18] Recommendation, I. T. U. T. P. 910. “Subjective video quality assessment methods for multimedia applications”, Recommendations of the ITU (Telecommunication Standardization Sector).