

## Spatial-temporal video quality metric based on an estimation of QoE

S. A. Amirshahi<sup>1,2</sup> and M.-C. Larabi<sup>1</sup>, senior member, IEEE

<sup>1</sup> XLIM-SIC, UMR CNRS 6172, University of Poitiers, France

<sup>2</sup> Chair for Computer Vision, Friedrich Schiller University of Jena, Germany

seyed-ali.amirshahi@uni-jena.de

Chaker.larabi@sic.univ-poitiers.fr

### ABSTRACT

*In this work a new Reduced Reference (RR) Video Quality Metric (VQM) is proposed. The method takes advantage of the Human Visual System (HVS) sensitivity to sharp changes in the video. The proposed method has a spatial-temporal approach and because of that it is named as STAQ (Spatial-Temporal Assessment of Quality). In the first step of STAQ we take a temporal approach and find the matching regions in consecutive frames. In the next step, a spatial approach is taken in the way of calculating the quality of the matching regions in the temporal approach. In the last step, the quality of the video is calculated based on the parameters gathered in the spatial and temporal domain and using the motion activity density of the video as a controlling factor. An important improvement lies in taking into account the Quality of Experience (QoE) represented as the motion activity density of the reference video. The results show a great improvement in the case of H.264 and MPEG-2 compressed and IP distorted videos even when compared to state of the art Full Reference (FR) metrics.*

### 1. INTRODUCTION

With the huge amount of video in our daily life, there has been a significant increase in researches on Video Quality Assessment (VQA) [1]. Running subjective tests for video quality evaluation is the main way of evaluating Video Quality (VQ) but it is rather an expensive, time consuming and tedious procedure. All this makes applying objective Video Quality Metrics (VQM) the best option for evaluating VQ for real time applications.

VQM like Image Quality Metrics (IQM) have three main categories. Full Reference (FR) metrics where we have access to the reference video. Reduce Reference (RR) metrics which some information regarding the features of the original video is available. No Reference (NR) metrics that there is no information regarding the reference video. Most VQM that have been proposed so far are FR. These methods range from simple methods such as the MSE and PSNR methods to rather complicated metrics such as metrics proposed in [2-10].

VQM could also be categorized based on the approach they take to evaluate the VQ. For instance, most of them have a spatial approach and often evaluate the VQ by applying IQM on different frames and then by pooling the results achieved from each frame. Few of the metrics have a

temporal approach and others have a Spatial-Temporal approach. In the later case most metrics combine results from a spatial approach and a temporal approach at the last step of the pooling procedure [5]. In the proposed metric we would calculate the VQ based on a Spatial-Temporal. The difference between the proposed metric and other Spatial Temporal metrics is the fact that instead of combining the results from each domain in the polling system our method uses the data from one domain in the other domain. This way the two domains have a close relationship with each other in every step of the work.

There has been an agreement on the importance of the integration of Quality of Experience (QoE) in the quality evaluation community. This has been shown through numerous workshops, conferences and special sessions dealing with this issue. Metrics base on QoE try to bring the observer in the loop and model his/her emotions towards the content of the video. Video content plays an important role when trying to base a VQM on QoE. With respect to what was mentioned, one of the aspects which could be useful when trying to model the observer's opinion is the amount of motion in the video. For example, if we apply the same type and level of noise to two different videos with different motion density the observer will give two different quality ratings to the video. Take the example of a football game and a news anchor. In the first case, the observer has a higher sensitivity and the amount of noise in the video should be low. In the second case, the observer will give a higher quality rating to the video even if we apply a higher amount of noise to the video. This aspect has been discussed and experimented in [23] allowing to find a formulation of the quality function of the motion of a sequence. Keeping this in mind the amount of motion density in videos is what the proposed method is based on and tries to use to take QoE into account.

With respect to the literature, in this paper a new reduced reference metric is proposed. The metric has a spatial-temporal approach and so is named as STAQ (Spatial-Temporal Assessment of Quality).

The reminder of this paper is arranged as followed: we will first introduce STAQ in section 2, present the results in section 3 and finally in section 4 a conclusion of the work is given.

## 2. PROPOSED APPROACH

An overview of STAQ is shown in Figure 1. As it can be seen the proposed metric has three main blocks, temporal block, spatial block and the pooling block. STAQ is based on the fact that the Human Visual System (HVS) is sensitive to sharp changes in the video. This gives the advantage of finding common regions in consecutive frames. In STAQ, we will find these regions in the temporal approach by calculating the appropriate Motion Vectors (MV) for the sub-blocks in the current frame and use their coordinates in the spatial approach to estimate the quality of each frame of the video for different color channels. For the pooling system we use the quality values we have from the spatial block along with other factors that we have gathered in different steps prior to the pooling system. For the developed metric, QoE is introduced as a controlling factor based on the motion activity density of the reference video. A description of each section of the proposed metric is given in the following.

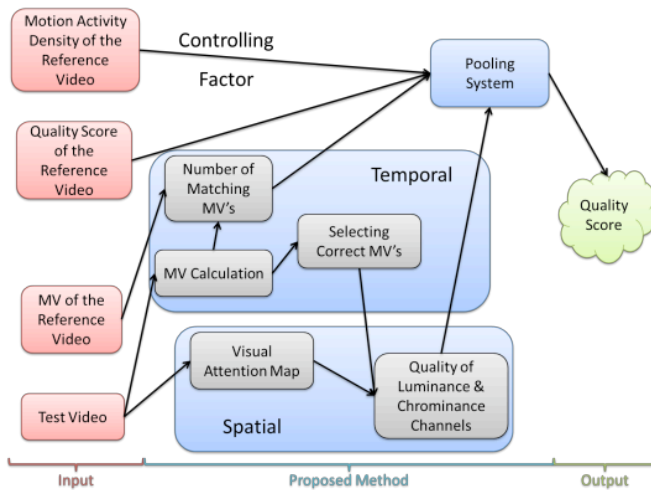


Figure 1 – Overview of STAQ

### 2.1 Temporal Approach

As shown in Figure 1, in the first step of the temporal approach the MV of the test video is calculated resulting in finding matching sub-blocks in consecutive frames. The method used for calculating the MV is the Adaptive Rood Pattern Search (ARPS) [11]. An important change made when calculating the MV compared to previous methods is the use of Complex Wavelet Structural Similarity index (CW-SSIM) [12] instead of using the Mean Absolute Difference (MAD). The reason behind this change is that by using the CW-SSIM method the precision of the results increases. Although by using CW-SSIM instead of MAD we will increase the complexity of the calculation but having results with a higher precision will increase our chances of finding the matching regions with better accuracy. Another important reason for using the CW-SSIM method is the fact that we would use the calculated results in the later stages and so decrease the calculation time for the proposed VQM.

Table 1. shows the MSE values between the original and the motion compensated frame using the CW-SSIM and the MAD method.

After calculating the MV for the test video we match the results with the MV of the reference video. Using the percentage of the matching MV which is named as the Number of Matching Motion Vectors (NMMV), the value of NMMV is sent to the pooling system. Table 2 shows an example of the MV for a frame in the test video and the corresponding frames in the reference video. The last block of the temporal approach is dedicated to check if the sub-block in the test video matches the assigned sub-block in the reference video. Figure 2 gives an example of the calculated MV for a frame; arrows in the figure correspond to the MV for each sub-block in the frame. As it can be seen in Figure 2, MV in a frame tend to have approximately the same size and direction. To separate the correct MV with the incorrect ones we will take the following procedure. Assume that  $MV = \{MV_i | i = 1, \dots, N\}$  is the MV in a specific frame, the total number of MV's been  $N$ . Eq (1) calculates the mean value of the MV. If we assign a scalar value as the threshold to the mean value calculated in Eq (1) named  $T$ , any MV which does not belong to the interval given by Eq (2), is not a correct MV and so the matching sub-block is not used in the spatial approach. Although in the procedure we do not take the angle of the MV into account but experiments have shown that we will still reach promising results. In the case of the MV of Figure 2 we will not accept the dashed MV as the correct ones. In other words the sub-blocks which are shaded grey and correspond to these MV are not used in the spatial domain since they do not match the corresponding sub-blocks in the next frame.

$$m = \frac{\sum_{i=1}^N MV_i}{N} \quad (1)$$

$$|MV_i| < Tm \quad (2)$$

Table 1 – MSE values between the original and the motion compensated result using the CW-SSIM and MAD method.

Method used	MSE
CWSSIM	29.3877
MAD	37.22332

Table 2 – Example of calculating the value for NMMV.

Frame number	MV of a frame in the test video	MV of a frame in the reference video	Matching MV
1	2,3	2,3	Yes
2	1,2	0,2	No
3	-2,4	-2,4	Yes
4	2,3	2,2	No
5	2,3	2,3	Yes
6	0,0	-1,2	No
Result	3 MV match $\Rightarrow$ NMMV = 50%		

#### 2.1.1 Adaptive Rood Pattern Search (ARPS) Method [11]

ARPS method is based on the fact that we expect the MV of a sub-block to have the same direction and size as the MV of

the neighboring sub-blocks. In [11] we will have a predicted MV for each sub-block with the same size and direction of the sub-block on the left of the current sub-block. In the first step, the predicted sub-block is checked along with the rood pattern distributed points. Then, the best matching sub-block is selected and a Small Diamond Search Pattern (SDSP) is used until the MV of the sub-block is found. As mentioned before STAQ uses the strategy introduced by the ARPS method for finding the MV with a small change. That is, instead of using the MAD method in STAQ uses the CW-SSIM method.

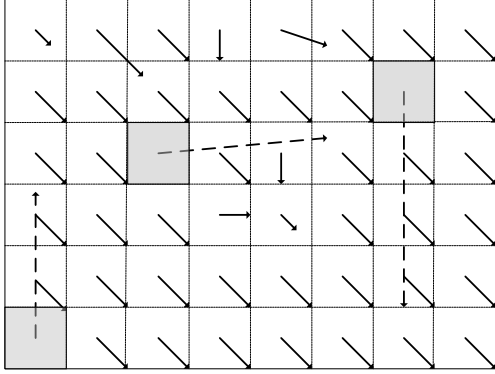


Figure 2 – example of calculated MV for a frame. Arrows show the MV.

### 2.1.2 Complex Wavelet Structural Similarity index (CW-SSIM) [12]

CW-SSIM IQM is based on the SSIM [13] IQM. As it is mentioned by the authors in [12] “the CW-SSIM index is an extension of the SSIM method to the complex wavelet domain”. This will make the CW-SSIM method not sensitive to nonstructured geometrical image distortions which we will face in the video environment such as rotation, scaling and etc.

CW-SSIM is calculated using Eq (3). In the equation  $c_x = \{c_{x,i} | i = 1, \dots, N\}$  and  $c_y = \{c_{y,i} | i = 1, \dots, N\}$  are two sets of coefficients extracted at the same spatial location in the same wavelet subbands of the two images being compressed,  $N$  representing the number of coefficients.  $c^*$  is the complex conjugate of  $c$  and  $K$  is a small positive constant mainly used to improve the results when Signal to Noise Ratio (SNR) is low. Like SSIM, the maximum value of CW-SSIM occurs when the two images compared are identical. In this case the value of CW-SSIM is 1.

$$S(c_x, c_y) = \frac{2|\sum_{i=1}^N c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^N |c_{x,i}|^2 + \sum_{i=1}^N |c_{y,i}|^2 + K} \quad (3)$$

## 2.2 Spatial Approach

In the spatial approach the first block calculates the Visual Attention Map (VAM) for each frame using the Achanta et al [14] method. The VAM will later be used to weigh each sub-block in the luminance channel according to the importance it has in the frame.

In the second block, we calculate the quality of the channels in the spatial approach. The chrominance channels are evaluated using the MSE and the luminance channel by using

the CW-SSIM method. The reason behind the two approaches is the higher amount of information in the luminance channel compared to the chrominance channels. For this reason we will need a metric with a higher accuracy for the luminance channel. The procedure is done by calculating the quality score,  $QSC(i, j)$  of each sub-block in the frames using Eq (4).

$$QSC(i, j) = IQM(sbC(i, j), sbC(i + 1, j)) \quad (4)$$

In Eq (4)  $sbC(i, j)$  is the sub-block  $j$  in frame  $i$ . depending on the channel we are assessing  $IQM$  is either MSE or CW-SSIM. Each quality score ( $WQSC(i, j)$ ) of each sub-block is weighted using Eq (5).

$$WQSC(i, j) = QSC(i, j) W(i, j) \quad (5)$$

$W(i, j)$  is the weigh associated to sub-block  $j$  in frame  $i$ . As mentioned before we would weigh the luminance channel based on the VAM we have calculated. Due to the low amount of information in the chrominance channel we will give a weigh of 1 to all sub-blocks in the frame. In the next step the mean value ( $FSC(i)$ ) of all sub-blocks ( $N$ ) for each frame is calculated as shown in equation (6)

$$FSC(i) = \frac{\sum_{j=1}^N WQSC(i, j)}{N} \quad (6)$$

Finally, we calculate the quality score of each channel named  $CS$ , over the  $M$  frames shown in Eq (7).

$$CS = \frac{\sum_{i=1}^M FSC(i)}{M} \quad (7)$$

For a video in the YUV colorspace we will have  $YS$ ,  $US$  and  $VS$  representing the quality of the channels.

### 2.2.1 Visual Attention Map

At the speed the frames are being played the observer does not have enough time to pay attention to all the details in the video. This is why we would give a higher weight to regions which according to the VAM will take the focus of the observers.

Achanta et al [14] introduced a VAM for coloured images. The method could be summarized in Eq (8).  $s(x, y)$  is the saliency score for each pixel  $(x, y)$ ,  $I_\mu$  is the mean image feature vector and  $I_{\omega hc}(x, y)$  is the corresponding image pixel vector value in the Gaussian blurred version of the reference image, here we use a  $5 \times 5$  separable binomial kernel. Since the saliency map is calculated in the Lab colorspace  $\| \cdot \|$  is the Euclidean distance.

$$S(x, y) = \| I_\mu - I_{\omega hc}(x, y) \| \quad (8)$$

The last step in the calculation of the VAM is the normalization procedure. In this step we will divide the pixel values in the VAM by the maximum value they have in the frame. We would then calculate the mean value for each sub-block. This way the sub-block with the highest importance will have the maximum value. As the importance decrease the value for each pixel will decrease as well. Figure 3 shows an example of a frame in the video along with a normalized VAM for the frame which we will use in the weighting procedure in Eq (5).

### 2.3 Motion Activity Density Group

As mentioned before, an important improvement in our metric compared to the previous VQM lies in the use of QoE.

In [23], we demonstrated the relation existing between quality and the motion. So QoE is introduced as a function depending on the motion activity density group of the video i.e. Very Low, Low, Medium, High and Very High motion activity.

$$Q = f(m), f: M \rightarrow MOS$$

where M is the set of possible motion activity classes and MOS=[1 : 5]. A value of 1 means that the quality will be rated as poor by all users while a value of 5 indicates an excellent quality.

Most motion activity density algorithms which are based on statistical values calculated from the MV. The motion activity group is determined based on the median, max2 and the variance value of the MV magnitude as introduced by Peker et al. [15]. Max2 is the maximum MV magnitude after removing the top 10% values.

#### 2.4 Pooling Procedure

As mentioned in the previous sections, for each video we have YS, US, VS and NMMV along with the motion activity group in our pooling system. The quality score of the reference video is also used in the pooling system. This constitutes a part of the reduced reference.

For pooling the data, a weight is assigned to the values sent to the pooling system and the values are added up shown in Eq (9).



(9)



(b)

Figure 3 – (a) original frame. (b) VAM normalized and divided into blocks.

$$MeanS = w_1 \cdot YS + w_2 \cdot US + w_3 \cdot VS + w_4 NMMV \quad (9)$$

In the last step, the difference between the *MeanS* value calculated in Eq(9) for the test video is compared to the

corresponding value for the reference video (referred to as the quality score of the reference video) as shown in Eq (10).

$$meanS_{st} = MeanS_{reference} - MeanS_{Test} \quad (10)$$

For pooling the data we use a neural-fuzzy system having *MeanS<sub>st</sub>* as its input and objective score as output. The system has five different pooling functions each representing one of the motion activity groups described previously. For this we use Matlab's fuzzy toolbox and its ANFIS (Adaptive Neuro-Fuzzy Inference System) structure. For training the algorithm we will use a hybrid optimization method with three two sided Gaussian membership functions (MF's) as its input and a linear output.

### 3. RESULTS

The proposed method was tested on the H.264 and MPEG-2 compressed and IP distorted videos from the "LIVE Video Quality Database" [15-17]. The sub-blocks used were in the size of  $8 \times 8$  pixels. Also based on different experiments we used a value of  $w_1 = w_2 = w_3 = w_4 = 0.25$  for the current database. Figure 4 shows the scatter plots between the objective and subjective results for the mentioned categories of videos.

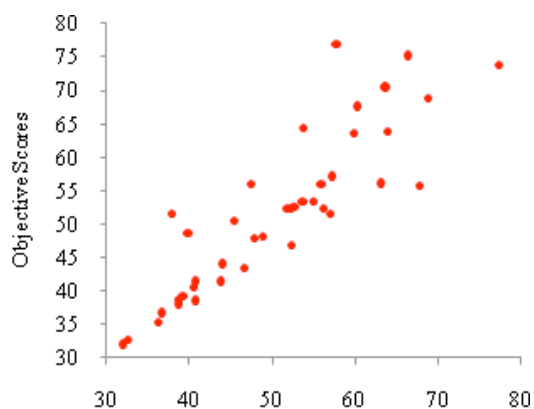
Table 3 shows the Spearman and table 4 presents the Pearson correlation between the objective and subjective results for ten state of the art FR metrics and also the STAQ metric which is a RR metric. As it can be seen in the case of H.264 compressed videos the results of STAQ are ranked first compared to the others. Especially in the case of the Spearman correlation, we see a large increase in the accuracy of the results. For MPEG-2 compressed videos, STAQ is ranked second and third compared to the other metrics. Note that the results are so close and STAQ is a RR metric and not a FR metric which is an advantage for the proposed metric. In the case of IP distorted videos STAQ is ranked first in the case of Spearman correlation but in the case of Pearson correlation it is ranked sixth. This is mainly because of the small difference between the results.

### 4. CONCLUSION

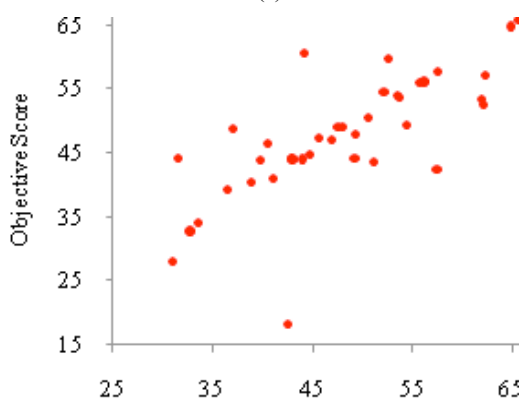
In this paper a new Reduced Reference VQM named as STAQ (Spatial-Temporal Assessment of Quality) which is based on QoE is proposed. One of the most important aspects of STAQ is the fact that we take Motion Activity Density into account and so try to present a metric with a QoE approach. STAQ has been compared to some of the state of the art FR metrics and it shows great improvement in the correlation between the subjective and objective scores. In the case of H.264 videos the results have a higher correlation compared to 10 different state of the art FR metrics. In the case of MPEG-2 and IP distorted videos the results are highly competitive with FR metrics. Keeping in mind that STAQ is a RR metrics this shows the advantage STAQ has compared to other metrics.

## 5. REFERENCE

- [1] Z. Wang, H. R. Sheikh and A. C. Bovik, *The handbook of video databases: design and application, Chapter 41: Objective video quality assessment*, B. F. a. O. Marques, Ed. CRC Press, 2003.
- [2] Z. Wang, I. Lu and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Special Issue on Objective Video Quality Metric*, vol. 19, no. 2, pp. 121-132, Feb. 2004.
- [3] f. Zhang, J. Li, G. Chen and J. Man, "Assessment of color video quality with singular value decomposition of complex matrix," in *International Conference on Information Assurance and Security*, 2009.
- [4] D. C. lin, P. M. Chau, "Objective human visual system based video quality assessment metric for low bit-rate video communication systems," in *IEEE Workshop on Multimedia Signal Processing*, 2006.
- [5] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transaction on Image Processing*, vol. 19, no. 2, pp. 335-350, Feb. 2010.
- [6] Y. Fu-zheng, W. Xin-dai, C. Yi-lin and W. Shuai, "A no-reference video quality assessment method based on digital watermark," in *International Symposium on Personal, Indoor and Mobile Radio Communication Proceedings.*, 2003.
- [7] F. Yang, S. Wan, Y. Chang and H. R. Wu, "A novel objective No-Reference metric for digital video quality assessment," *IEEE Signal Processing Letters*, vol. 4, no. 11, pp. 685-688, Oct. 2005.
- [8] C. Opreal, I. Pirnig, C. Paleologu and M. Udrea, "Perceptual video quality assessment based on salient region detection," in *International Conference on Telecommunication*, 2009.
- [9] A. Maalouf and M. C. Larabi, "A No-Reference color video quality metric based on a 3d multispectral wavelet transform".
- [10] Wang, Z. and Li, Q., "Video quality assessment using a statistical model of human visual speed perception", *Journal of the Optical Society of America A - Optics, Image Science and Vision* 24, B61-B69 (Dec 2007).
- [11] Y. Nie, K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Transaction on Image Processing*, vol. 11, no. 12, pp. 1442-1449, Dec. 2002.
- [12] M. P. Sapat, Z. Wang, A. c. Bovik and M. K. Markey, "complex wavelet structural similarity: a new image similarity index," *IEEE Transaction on Image Processing*, vol. 18, no. 11, pp. 2385-2401, Nov. 2009.
- [13] Wang, Z. and Bovik, A. C., "A universal image quality index," *IEEE Signal Processing Letters* 9(3), 81-84 (2002).
- [14] R. Achanta, S. Hemami, F. Estrada and S. Ssstrunk, "Frequency-tuned Salient Region Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.
- [15] K. A. Peker and A. Divakaran, "Framework for measurement of intensity of motion activity of video segments," *Journal of Visual Communication & Image Representation*, vol. 15, pp. 265-284, 2004.
- [16] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transaction on Image Processing*, 2009.
- [17] K. Seshadrinathan, R. Soundarajan, A. C. Bovik and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," in *SPIE proceeding Human Vision and Electronic Imaging*, 2010.
- [18] Live Video Quality Database. [Online] [http://live.ece.utexas.edu/research/quality/live\\_video.html](http://live.ece.utexas.edu/research/quality/live_video.html)
- [19] Wang, Z., Simoncelli, E. P. ., Bovik, A. C., and Matthews, M. B., "Multiscale structural similarity for image quality assessment," in [*IEEE Asilomar Conf. Signals, Sys. and Comp.* ], (2003).



(a)



(b)

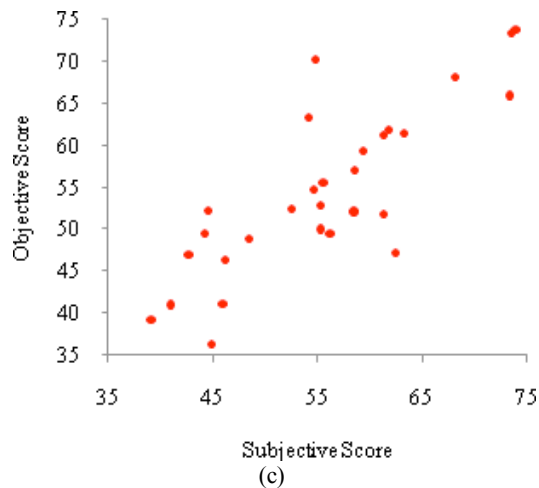


Figure 4 – Objective vs. subjective results for (a) H.264 and (b) MPEG-2 compressed videos (c) IP distorted videos.

Prediction Model	H-264	MPEG2-2	IP
PSNR	0.4385	0.3856	0.4108
SSIM [2]	0.6656	0.5491	0.5119
MS-SSIM [19]	0.6919	0.6604	0.7219
Speed-SSIM [10]	0.7206	0.6270	0.5587
VSNR [20]	0.6216	0.5980	0.7341
VQM [21]	0.6459	0.7860	0.6480
V-VIF [22]	0.6911	0.6145	0.5102
Spatial MOVIE [5]	0.7252	0.6587	0.7378
Temporal MOVIE [5]	0.7920	<b>0.8252</b>	0.7383
MOVIE [5]	0.7902	0.7595	<b>0.7622</b>
STAQ	<b>0.8778</b> (1)	0.7988 (2)	0.7080 (6)

- [20] Chandler, D.M. and Hemami, S. S., “VSNR: A wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Transactions on Image Processing* 16(9), 2284–2298 (2007).
- [21] Pinson, M. H. and Wolf, S., “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting* 50, 312–322 (Sept. 2004).
- [22] Seshadrinathan, K. and Bovik, A. C., “Motion-based perceptual quality assessment of video,” in *[Proc. SPIE - Human Vision and Electronic Imaging]*, (2009).
- [23] Larabi M.-C., Quoirin L., “Relation between bitrate, motion, and framerate for scoring of image sequences”, *IS&T/SPIE Image Quality and System Performance V* (2008).

Table 3 –Spearman correlation between the subjective results and 10 state of the art FR metrics results along with the results for the proposed metric for H.264 and MPEG-2 compressed and IP distorted videos.

Prediction Model	H-264	MPEG-2	IP
PSNR	0.4296	0.3588	0.3206
SSIM [2]	0.6514	0.5545	0.4550
MS-SSIM [19]	0.7051	0.6617	0.6534
Speed-SSIM [10]	0.7086	0.6185	0.4727
VSNR [20]	0.6460	0.5915	0.6894
VQM [21]	0.6520	0.7810	0.6383
V-VIF [22]	0.6807	0.6116	0.4736
Spatial MOVIE [5]	0.7066	0.6911	0.7046
Temporal MOVIE [5]	0.7797	<b>0.8170</b>	0.7192
MOVIE [5]	0.7664	0.7733	0.7157
STAQ	<b>0.9132</b> (1)	0.76515 (4)	<b>0.7202</b> (1)

Table 4 – Pearson correlation between the subjective results and 10 state of the art FR metrics results along with the results for the proposed metric for IP distorted videos.