Evaluating Video Quality by Differentiating Between Spatial and Temporal Distortions

Meisam Jamshidi Seikavandi[1] and Seyed Ali Amirshahi[2]

[1] Nikoo Dana Fanavari Delfan, Technology and Science Park of Lorestan, Iran
meisamjam@gmail.com

[2] The Norwegian Colour and Visual Computing Laboratory, Norwegian University of Science and Technology, Gjøvik, Norway
s.ali.amirshahi@ntnu.no

Abstract. To objectively evaluate the quality of videos different state-of-the-art Image Quality Metrics (IQMs) have been used to introduce different Video Quality Metrics (VQM). While such approaches are able to evaluate the spatial quality of the frames in the video they are not able to address the temporal aspects of the video quality. In this study, we introduce a new full-reference VQM which is based on taking advantage of a Convolutional Neural Network (CNN) based IQM to evaluate the quality of the frame. Using other techniques such as visual saliency detection we are then able to differentiate between spatial and temporal distortions and use different pooling techniques to evaluate the quality of the video. Our results show that by detecting the type of distortion (spatial or temporal) affecting the video quality, the proposed VQM can evaluate the quality of the video with a higher accuracy.

Keywords: Video Quality Assessment, Video Saliency, Image Saliency, Spatial Distortion, Temporal Distortion, Temporal pooling.

1 Introduction

With the huge amount of video we have access to in our daily life, evaluating the quality of videos is an essential part of any application that deals with videos. Although subjective assessment is still considered the primary standard for Video Quality Assessment (VQA), it is time-consuming and financially expensive to perform on a regular basis. For this and many other reasons, in the last few decades, objective assessment of video quality has attracted much attention. Objective assessment methods, known as Video Quality Metrics (VQMs) have been widely used to estimate the quality of videos [7]. Depending on the availability of the reference video, VQMs can be classified into full-reference, reduced-reference, and no-reference. Full-reference VQMs need access to the reference video, while reduced-reference metrics required partial information of the reference video and no-reference metrics only have access to the test video.

Full-reference VQMs can be further classified to error sensitivity based methods [10,18], structural similarity based approaches [39], information fidelity based approaches [35], spatial-temporal approaches [34], saliency-based approaches [25], and network-aware approaches [13]. Many of the mentioned methods are an extended version of Image Quality Metrics (IQMs) that generally follow with a pooling structure to bridge over Image Quality Assessment (IQA) and VQA. In this study, we combine different approaches, such as metrics based on the use of Convolutional Neural Networks (CNNs) and saliency techniques, to calculate a series of quality values for the video frames. Using different weighting techniques that depend on the type of distortion (spatial or temporal) affecting the video, the quality of the frames is pooled to represent the video quality score.

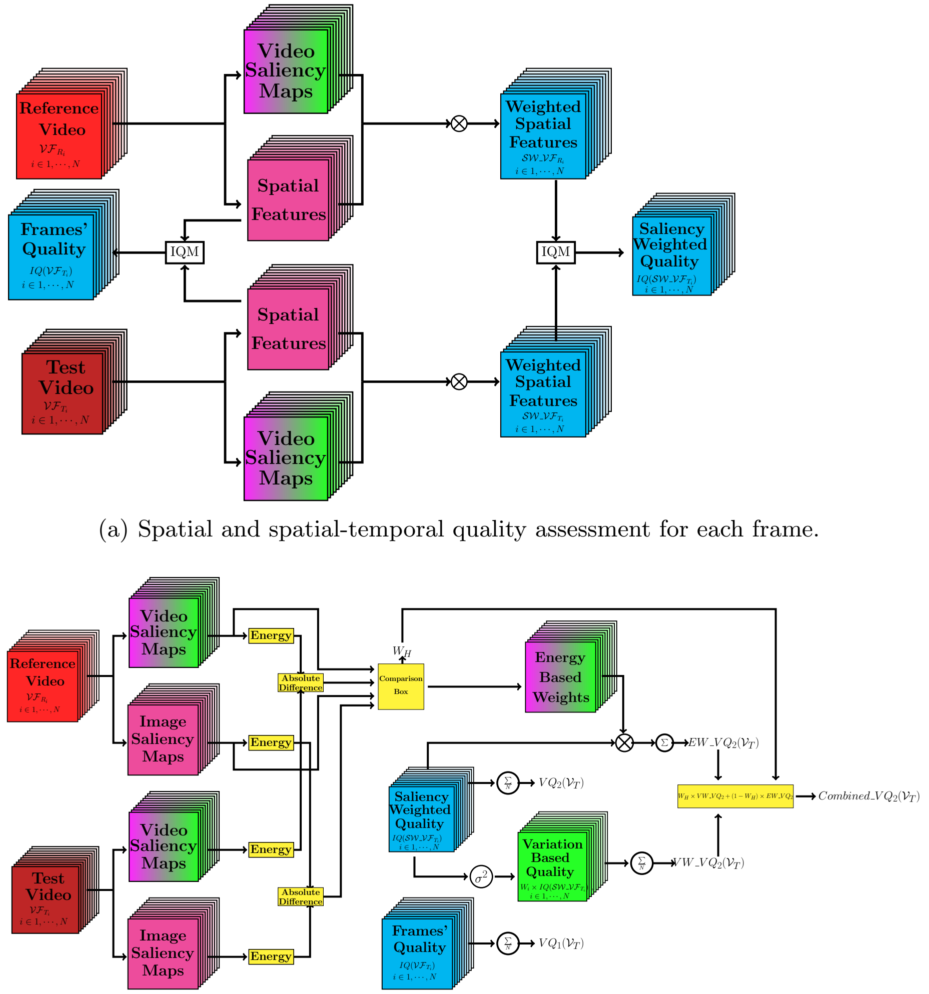
Keeping in mind that VQA has been a field of research for over two decades, it is no surprise that a high number of different VQMs have been introduced. Like any other field of research in image processing and computer vision, early VQMs were based on introducing different single or multiple handcrafted features for VQA. While initially these features were pure mathematical techniques such as Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) [40], overtime a shift towards introducing features that try to model the Human Visual System (HVS) is seen in VQMs [17,29,33]. Unlike the early VQMs which were mostly focused on spatial aspects of the video [10,18,39], with the introduction of temporal features VQMs showed improvement in their performance [15,19,21,24]. Since most VQMs either provide a spatial, temporal, and/or spatial-temporal quality value for the videos, different pooling techniques have been used. As an example, the use of saliency maps for providing different weights to different regions in the frame or different frames have been used to introduce different VQMs [3,6]. Finally, in recent years, with the introduction of state-of-the-art machine learning techniques and especially Convolutional Neural Networks (CNNs), again, a big improvement has been observed in the accuracy of VQMs [2,14,19,41].

Our contribution in this paper can be summarized as: 1) by using video saliency maps we introduce a spatial dimension to a state-of-the-art IQM and use the approach for video quality assessment. 2) By applying temporal and spatial-temporal pooling techniques two different quality scores are calculated for each frame in the video. 3) A new content-based evaluation is introduced that is able to detect the type of distortion (temporal or spatial) and propose a VQM based on the distortion detected.

The rest of the paper is organized as follows: in Section 2, we provide a detailed description of the proposed approach while the experimental results are presented in Section 3. Finally, Section 4 provides a conclusion of the work and what future directions we plan to take to extend the work.

2 Proposed Approach

Our proposed VQM (Figure 1) is based on the extraction of spatial and temporal features from the video. Apart from the main VQM introduced in this study,



(a) Spatial and spatial-temporal quality assessment for each frame.

(b) Different pooling methods used between the quality of each frame for evaluating the quality of the video.

Fig. 1. Pipeline used for calculating different VQMs proposed in this study. In the figure blocks with a magenta shade correspond to spatial features, blocks with a green shade correspond to temporal features, and blocks with a gradient shade of magenta to green correspond to spatial-temporal features.

and to better study the influence of the different features used in our VQM, we also propose other VQMs based solely on one or multiple features introduced.

2.1 Spatial Approach

As pointed out in Section 1, IQA and VQA are closely linked. In fact, when it comes to extracting spatial features from videos, a high number of features used for evaluating the video quality were initially introduced in IQMs. In other words, in the case of spatial features, different VQMs extract spatial features introduced in different IQMs on each frame of the video [2,11,27].

In this study, we aim to introduce a new VQM which takes advantage of the IQM proposed in [8]. In [8] Amirshahi et al. propose a new IQM, which is based on calculating the similarity between extracted feature maps in different convolutional layers of a pre-trained CNN Network. Their hypothesis which was inspired by the use of Pyramid Histogram of Orientation Gradients (PHOG) [12] features for calculating self-similarity in images introduced in [4,9,30] is that the similar feature maps at different convolutional layers are the similar the quality of the test and reference images are. To calculate the similarity between two features maps, they take the following steps:

1. From the reference ($\mathcal{I}_R$) and test ($\mathcal{I}_T$) images feature maps are extracted at different convolutional layers.
2. For the test image $\mathcal{I}_T$ in Convolutional layer n histogram

$$\begin{aligned}&\boldsymbol{h}(\mathcal{I}_T, n, L) = \\ &\quad (\sum_{i=1}^{X}\sum_{j=1}^{Y}\mathcal{F}(\mathcal{I}_T, n, L, 1)(i,j), \sum_{i=1}^{X}\sum_{j=1}^{Y}\mathcal{F}(\mathcal{I}_T, n, L, 2)(i,j), \cdots, \\ &\quad \sum_{i=1}^{X}\sum_{j=1}^{Y}\mathcal{F}(\mathcal{I}_T, n, L, z)(i,j), \cdots, \sum_{i=1}^{X}\sum_{j=1}^{Y}\mathcal{F}(\mathcal{I}_T, n, L, M)(i,j)),\end{aligned} \tag{1}$$

   is calculated. In Eq. (1), L corresponds to the level in spatial pyramid the histograms are calculated at and $\mathcal{F}(\mathcal{I}_T, n, L, z)$ corresponds to feature map z in the n^{th} convolutional layer of image $\mathcal{I}_T$ at level L with a size of $X \times Y$. To take a pyramid approach, Amirshahi et al. divided feature maps to four equal sub-regions resulting in different $\boldsymbol{h}$ histograms (Eq. (1)) at different levels (L) of the spatial resolution. The division and calculation of $\boldsymbol{h}$ continues to the point that the smallest side of the smallest sub-region is equal or larger than seven pixels.
3. The quality of the test image at level L for the convolutional layer n is then calculated by

$$\begin{aligned}m_{IQM}(\mathcal{I}_T, n, L) &= d_{HIK}(\boldsymbol{h}(\mathcal{I}_T(n, L)), \boldsymbol{h}(\mathcal{I}_R(n, L))) \\ &= \sum_{i=1}^{n}\min(h_i(\mathcal{I}_T, n, L)), h_i(\mathcal{I}_R, n, L)).\end{aligned} \tag{2}$$

4. The concatenation of all $m_{IQM}(\mathcal{I}_T, n, l)$ values

$$\begin{aligned}\boldsymbol{m}_{IQM}(\mathcal{I}_T, n) = (&m_{IQM}(\mathcal{I}_T, n, 1), m_{IQM}(\mathcal{I}_T, n, 2), \\ &\cdots, m_{IQM}(\mathcal{I}_T, n, z), \cdots, m_{IQM}(\mathcal{I}_T, n, L)),\end{aligned} \tag{3}$$

would then be used by

$$IQ(\mathcal{I}_T, n) = \frac{1 - \sigma(\boldsymbol{m}_{IQM}(\mathcal{I}_T, n))}{\sum_{l=1}^{L} \frac{1}{l}} \sum_{l=1}^{L} \frac{1}{l} \cdot m_{IQM}(\mathcal{I}_T, n, l) \tag{4}$$

to calculate the quality of the test image at convolutional layer n. In Eq. (4), $\sigma(\boldsymbol{m}_{IQM}(\mathcal{I}_T, n))$ corresponds to the standard deviation among the values in $\boldsymbol{m}_{IQM}(\mathcal{I}_T, n)$.

5. Finally, the overall quality of the test image is calculated using a geometric mean of all quality scores at different convolutional layers

$$IQ(\mathcal{I}_T) = \sqrt[N]{\prod_{n=1}^{N} IQ(\mathcal{I}_T, n)} \tag{5}$$

where N corresponds to the total number of convolutional layers.

While the study presented in [8] was mainly focused on the use of the Alexnet model [20], nevertheless, it was shown that it would be possible to use other deeper CNN models such as VGG16 and VGG19 [36]. Different studies have shown the flexibility of the mentioned IQM and how it can be extended to improve the performance of other IQMs [5]. For this reason, in this study, we would take advantage of this IQM to evaluate the spatial quality of the video frames (Figure 1).

To evaluate the spatial quality of the test video ($\mathcal{V}_T$), the average quality of the frames

$$VQ_1(\mathcal{V}_T) = \frac{\sum_{i=1}^{N} IQ(\mathcal{VF}_{T_i})}{N}. \tag{6}$$

could be used. In Eq. (6), $\mathcal{VF}_{T_i}$ represents the i^{th} frame and N corresponds to the number of frames in the test video.

2.2 Spatial-Temporal Approach

It is clear that without taking into account the temporal aspects of a video, any VQM would be lacking accuracy. In our approach the first feature extracted from the videos is visual saliency which is linked to the spatial-temporal aspects of the video quality. Different studies have shown the important role visual saliency plays in IQA and VQA [3]. While there are a considerable number of different methods to calculate the saliency maps of images and video, the Graph-Based Visual Saliency (GBVS) [16] approach is one of the well-known techniques which has shown good accuracy for image and video saliency detection. It is important to point out that while image saliency calculation is purely based on spatial features, when it comes to video saliency, temporal aspects of the video are also considered and so video saliency calculation could be linked to the spatial-temporal properties of the video.

In our approach, we first calculate the saliency map for the test and reference videos. The saliency map of each frame is then resized to the size of the input of the network. Similar to the layers of the pre-trained CNN model used, we apply max-pooling to the calculated saliency maps in each frame resulting in different saliency maps, each corresponding to the size of the feature maps at each convolutional layer of our model. The calculated feature maps are then used as pixel-wise weights for the features in different convolutional layers. This will allow us to give higher weights to regions in the feature maps that are more salient to the observer. The quality of the video is then calculated by

$$VQ_2(\mathcal{V}_T) = \frac{\sum_{i=1}^{N} IQ(\mathcal{SW}\_\mathcal{VF}_{T_i})}{N} \tag{7}$$

in which $IQ(\mathcal{SW}\_\mathcal{VF}_{T_i})$ corresponds to the quality of $\mathcal{VF}_{T_i}$ where the saliency map of the frames has been used as a weighting function on the feature maps at each convolutional layer (Figure 1).

2.3 Temporal Approach

Although different VQMs try to take into account the spatial and temporal aspects of the video, nevertheless, most VQMs provide a single quality score for each video. To reach this single quality score different pooling techniques are used to combine quality scores of all video frames. While careful attention has been paid on how the quality scores for each frame is calculated, most, if not all, pooling approaches are based on some version of averaging the quality scores for all frames. The average value, geometric mean, harmonic mean, and Minkowski mean are some of the different types of averaging used in different VQMs. It is clear that using any type of averaging on the quality values of the frame could result in disregarding different aspects of the video that could be linked to the HVS. In this study, to better link the video quality score to how observers react to the change of quality in a video clip we try a new approach for pooling the quality scores of the frames.

Recent studies such as [28] have suggested that the overall perceptual quality of a video is highly dependent on the temporal variation of the video quality. That is, with an increase in the temporal variation of the video quality along the video sequence, the video quality declines. To address this aspect, we use the variation of the quality scores of the frames in our pooling approach. While the variance of the quality scores of the frames could be a good description of the quality fluctuation in the video, it only provides a general description of the video quality. To better consider the temporal variation of the video quality, we calculate the variance of quality scores in a specific time frame. That is, for the i^{th} frame of the test video ($\mathcal{V}_{T_i}$) we calculate

$$Local_{Var}(\mathcal{V}_{T_i}) = \sigma^2(\mathcal{VQ}_{T_{i-L}}, \cdots, \mathcal{VQ}_{T_i}, \cdots, \mathcal{VQ}_{T_{i+L}}). \tag{8}$$

In Eq. (8), the length of the local window in which we calculate the variance (σ^2) of the frame quality score is $2L+1$. Based on our experimental results, the best value for L is 2 resulting in a window of five frames. To introduce a better regional representation of the quality score for the video we calculate the video quality using

$$\begin{aligned} VW\_VQ(\mathcal{V}_T) &= \frac{\sum_{i=1}^{N} W_i \times IQ(\mathcal{VF}_{T_i})}{\sum_{i=1}^{N} W_i}, \\ W_i &= \begin{cases} 1 & if \quad Local_{Var}(\mathcal{V}_{T_i}) < Global_{Var}(\mathcal{V}_T) \\ 0 & if \quad Local_{Var}(\mathcal{V}_{T_i}) > Global_{Var}(\mathcal{V}_T) \end{cases}. \end{aligned} \tag{9}$$

In Eq. (9) W_i corresponds to the weight given to the quality score of the i^{th} frame in the video ($\mathcal{VF}_{T_i}$) and $Global_{Var}(\mathcal{V}_T)$ represents the variance of all the frame quality scores in the test video. From Eq. (9) it is clear that if the variance in the local quality score is larger than the variance of the global quality score a weight of one is given to the frame quality but if the variance of the local quality is lower than the variance of the global quality score a weight of zero is given to the quality frame. Simply said, the quality of a frame is only considered if the change of video quality in a given local interval ($[\mathcal{V}_{T_{i-L}}, \mathcal{V}_{T_{i+L}}]$) is bigger than the change of frame quality over the total duration of the video.

2.4 Spatial vs. Spatial-Temporal Distortion Detection

Although saliency maps have mostly been used to detect salient regions in the image and/or videos, studies such as [26,31] have used saliency maps to differentiate between salient and non-salient frames. While this labeling process is simply done by calculating the total energy of the saliency map in each frame, we take one step further. That is, by comparing saliency maps calculated for frames using the GBVS video and image saliency techniques we will be able to differentiate between frames that are mostly influenced by spatial or spatial-temporal distortions (See Section 2.2 for a detailed description of the difference between saliency maps calculated for frames using the image and video saliency techniques). The following steps are taken for this process:

1. Assuming the total energy of the saliency in the i^{th} frame in the test video ($\mathcal{VF}_{T_i}$) is equal to

$$EV_{T_i} = \sum_{x=1}^{X} \sum_{y=1}^{Y} \left(\text{Video\_Sal}(\mathcal{VF}_{T_i})(x,y)\right)^2, \tag{10}$$

   we calculate similar values for EV_{R_i}, EI_{T_i}, and EI_{R_i} which represent the total energy of the i^{th} frame in the reference video using a video saliency approach, the total energy of the i^{th} frame in the test video using an image saliency approach, and the total energy of the i^{th} frame in the reference video using an image saliency approach respectively. In Eq. (10), the i^{th} frame has a size of $X \times Y$ and Video_Sal represents the video saliency function used.
2. The difference between the total salient energy of the reference and test frame using video and image saliency is calculated by

$$dEV_{T_i} = \frac{\mid EV_{T_i} - EV_{R_i} \mid}{EV_{R_i}}, \tag{11}$$

$$dEI_{T_i} = \frac{\mid EI_{T_i} - EI_{R_i} \mid}{EI_{R_i}}. \tag{12}$$

While dEV_{T_i} could be linked to spatial-temporal aspects of the video, dEI_{T_i} is linked to the spatial aspects.

3. Assuming that the reference video has a higher quality compared to the test video in the cases in which dEV_{T_i} is larger than dEI_{T_i}, it can be interpreted that the distortion has likely a higher spatial-temporal effect on the video than just a spatial effect.

While in Section 2.3 the variance weighted VQM was introduced (VW_VQ), in this section by detecting the type of distortion (spatial or spatial-temporal) we introduce the energy weighted VQM

$$\begin{aligned} EW\_VQ(\mathcal{V}_T) &= \frac{\sum_{i=1}^{N} EW_i \times IQ(\mathcal{VF}_{T_i})}{\sum_{i=1}^{N} EW_i}, \\ EW_i &= \begin{cases} dEV_{T_i} & if \quad dEI_{T_i} < dEV_{T_i} \\ dEI_{T_i} & if \quad dEI_{T_i} > dEV_{T_i} \end{cases}. \end{aligned} \tag{13}$$

While until now we have introduce two video quality values for each video (the variance weighted video quality, and the energy weighted video quality) we believe that since the two methods use different approaches, it is highly possible to find situations that one of the methods perform better than the other. Although finding a perfect metric that ideally detects this issue is challenging, nevertheless, as a first step we introduce the following two parameters

$$\mathbf{dEV}_{ALL} = \sum dEV_{T_i}, \tag{14}$$

$$\mathbf{dEI}_{ALL} = \sum dEI_{T_i}. \tag{15}$$

The final video quality score which we refer to as the combined video quality is then calculated by

$$\begin{aligned} combined\_VQ(\mathcal{V}_T) &= (W_H \times VW\_VQ + (1 - W_H) \times EW\_VQ), \\ W_H &= \begin{cases} 1 & if \quad dEI_{ALL} < dEV_{ALL} \\ 0 & if \quad dEI_{ALL} > dEV_{ALL} \end{cases} \end{aligned} \tag{16}$$

using dEV_{ALL} and dEI_{ALL} values introduced earlier. Obviously finding a better weighting approach than a simple zero and one for the W_H values is a better option which we will address in the next sections.

3 Experimental Results

To evaluate the performance of the proposed VQMs we calculate the correlation between the subjective scores in different subjective datasets and the objective quality scores from the VQMs.

3.1 Datasets Used

To test the accuracy of our proposed VQMs, two different datasets which are widely used in the scientific community are used. While one dataset (CSIQ) is focused on covering different types of distortion, the other (NETFLIX) is mainly focused on including videos and distortions in video streaming for entertainment use.

Computational and Subjective Image Quality (CSIQ) video dataset [1] contains 12 reference videos and 216 distorted videos from six different types of distortion. All videos in the dataset are in the raw YUV420 format with a resolution of 832 × 480 pixels, and with a duration of 10 seconds at different frame rates (24, 25, 30, 50, or 60 fps). Among the six distortions, four are linked to different compression-based distortions: H.264 compression (H.264), HEVC/H.265 compression (HEVC), Motion JPEG compression (MJPEG), and Wavelet-based compression using the Snow codec (SNOW). The PLoss and WNoise distortions are the other two types of distortions covered in this dataset.

Netflix public dataset used the Double Stimulus Impairment Scale (DSIS) method to collect their subjective scores. In the DSIS method the reference and distorted videos are displayed sequentially. Since the focus of this study was to evaluate the quality of video streams focused on entertainment in the subjective experiments, a consumer-grade TV under controlled ambient lighting was used. The distorted videos with lower resolution than the reference was upscaled to the source resolution before displaying on the TV. Observers evaluated the quality of the videos while sitting on a couch in a living room-like environment and were asked to assess the impairment on a scale of one (very annoying) to five (not noticeable). The scores from all observers were combined to generate a Differential Mean Opinion Score (DMOS) for each distorted video and results were normalized in the range of zero to 100 were it was assumed that the reference video has a subjective quality score of 100 [23].

3.2 Results and Discussion

To calculate the accuracy of the proposed VQM in our experiments the linear Spearman and leaner and non-linear Pearson correlations were calculated between our objective scores and the subjective scores provided in different datasets. In this paper and due to space limitations, we would only provide the non-linear Pearson correlation results. From the results we can observe that:

- In the case of each separate distortion, the proposed spatial based VQM (VQ_1) is able to evaluate the video quality with a relatively high correlation (Table 1). This correlation value (average of .89) drops dramatically (.77) when videos independent of their distortions are evaluated. This finding can be linked to the fact that depending on the type of distortion, different spatial features affect the video quality.

Table 1. Non-linear Pearson correlation values for different distortions using the VQ_1 values at different convolutional layers in the Alexnet model.

dataset	distortion	CONV 1	CONV 2	CONV 3	CONV 4	CONV 5	All
CSIQ dataset	H.264	.95	.97	.96	.96	.96	.96
	PLoss	.74	.79	.77	.77	.76	.79
	MJPEG	.45	.89	.93	.93	.89	.90
	Wavelet	.89	.87	.86	.85	.85	.86
	WNoise	.87	.92	.91	.92	.92	.92
	HEVC	.89	.92	.90	.91	.90	.91
	ALL	.70	.77	.75	.76	.77	.77
Netflix	-	.77	.83	.84	.84	.86	.84

Table 2. Non-linear Pearson correlation values for different distortions using the VQ_2 values at different convolutional layers in the Alexnet model.

dataset	distortion	CONV 1	CONV 2	CONV 3	CONV 4	CONV 5	All
CSIQ dataset	H.264	.90	.95	.94	.93	.95	.94
	PLoss	.74	.84	.83	.84	.86	.84
	MJPEG	.50	.92	.91	.90	.90	.90
	Wavelet	.88	.91	.90	.91	.92	.92
	WNoise	.91	.94	.93	.93	.92	.95
	HEVC	.93	.93	.92	.93	.91	.93
	ALL	.77	.82	.82	.82	.82	.82
Netflix	-	.78	.90	.88	.89	.92	.90

- Similar to the IQM proposed in [8], quality scores using VQ_1 in the mid-convolutional layers (CONV3 and CONV4 in the case of the Alexnet model) show a higher correlation value (Table 1). Amirshahi et al. have linked this issue in the case of images to the nature of deeper convolutional layers which are more focused on patterns and textures seen in the image.
- Results from calculating VQ_2 for different distortions in the case of the CSIQ dataset (Table 2) show an average of 0.02 increase in correlation values compared to VQ_1 VQM. From the results, it is interesting to observe the most significant increase in the correlation value from the VQ_1 to VQ_2 VQMs is in the case of PLoss (0.05) and WNoise (0.03).
- When it comes to the case of all videos in the dataset independent of the type of distortion, VQ_2 shows a better performance than VQ_1. This increase of approximately 0.06 shows that by simply giving a higher weight to more salient regions of the feature map we could increase the accuracy of the VQM.
- We can see that compared to VQ_2, results from VW_VQ_2 (Table 3) decreases for all individual distortions in the CSIQ dataset while the overall result do not show any changes.

Table 3. Non-linear Pearson correlation values for different distortions using the VW_VQ_2 values at different convolutional layers in the Alexnet model.

dataset	distortion	CONV 1	CONV 2	CONV 3	CONV 4	CONV 5	All
CSIQ dataset	H.264	.89	.94	.93	.94	.95	.94
	PLoss	.70	.81	.77	.79	.85	.80
	MJPEG	.51	.91	.90	.89	.89	.90
	Wavelet	.88	.90	.89	.90	.92	.91
	WNoise	.89	.90	.88	.89	.87	.90
	HEVC	.93	.93	.93	.93	.91	.93
	ALL	.77	.82	.80	.80	.82	.82
Netflix	-	.71	.87	.86	.87	.90	.87

Table 4. Non-linear Pearson correlation values for different distortions using the EW_VQ_2 values at different convolutional layers in the Alexnet model

dataset	distortion	CONV 1	CONV 2	CONV 3	CONV 4	CONV 5	All
CSIQ dataset	H.264	.90	.94	.93	.93	.96	.95
	PLoss	.71	.79	.80	.80	.83	.81
	MJPEG	.51	.88	.87	.86	.89	.88
	Wavelet	.87	.91	.88	.89	.92	.90
	WNoise	.89	.92	.92	.92	.91	.92
	HEVC	.93	.94	.93	.93	.92	.94
	ALL	.75	.83	.81	.81	.83	.83
Netflix	-	.85	.90	.90	.90	.91	.91

- Using saliency-based weighting (EW_VQ_2), show a small improvement in the performance of H.264, HEVC, and all distortions by 0.01 (Table 4). This can be linked to the fact that by using saliency in the case of EW_VQ_2 the VQM covers both spatial and temporal aspects of the video quality.
- In the case of $combined\_VQ_2$ (Table 5), results show an increase in CONV1 and CONV2 layers. Comparing EW_VQ_2 to VW_VQ2, an improvement can also be seen in CONV3 and CONV4 layers. We can observe that $combined\_VQ_2$ has a better or the same performance in the case of H.264, MJPEG, and HEVC compressions, which could imply that the mentioned compressions are more discriminative for W_H to detect.
- Compared to other state-of-the-art VQMS (table 6), the proposed approach has better or as good as a performance in the case of H.264 and MJPEG compressions. This can be linked to the compatibility of our method with the structure of such compression methods and how W_H is able to discriminate between these compression methods. In the case of the WNoise and PLoss distortions, our proposed approach does not show a competitive performance. This could be linked to the fact that the saliency methods used

Table 5. Non-linear Pearson correlation values for different distortions using the $Combined\_VQ_2$ values at different convolutional layers in the Alexnet model

dataset	distortion	CONV 1	CONV 2	CONV 3	CONV 4	CONV 5	All
CSIQ dataset	H.264	.90	.95	.93	.94	.95	.95
	PLoss	.63	.73	.70	.71	.78	.72
	MJPEG	.50	.91	.90	.90	.91	.91
	Wavelet	.88	.91	.90	.90	.92	.90
	WNoise	.89	.91	.90	.91	.90	.91
	HEVC	.94	.94	.93	.94	.92	.94
	ALL	.78	.83	.81	.81	.82	.83
Netflix	-	.84	.89	.90	.90	.91	.90

Table 6. Non-linear Pearson correlation values for different distortions in the CSIQ dataset in comparison with state-of-the-art VQMs.

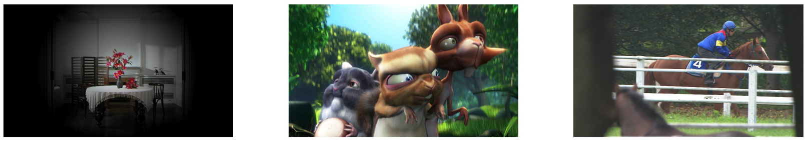
	H.264	PLoss	MJPEG	Wavelet	WNoise	HEVC	ALL
SSIM	.95	.84	.80	.89	.97	.96	.76
VIF [35]	.95	.92	.91	.92	.96	.96	.72
STMAD [38]	.96	.87	.89	.87	.89	.92	.82
ViS3 [37]	.93	.82	.81	.93	.93	.96	.81
MOVIE [34]	.90	.88	.87	.89	.85	.93	.78
V-BLIINDS [32]	.94	.76	.85	.90	.93	.92	.84
SACONVA [22]	.91	.81	.85	.85	.90	.90	.86
VQ_1	.96	.79	.90	.86	.92	.91	.77
VQ_2	.94	.84	.90	.92	.95	.93	.82
$VWVQ_2$	.94	.80	.90	.91	.90	.93	.82
$EWVQ_2$	.95	.81	.88	.90	.92	.94	.83
$Combined\_VQ_2$	.95	.72	.91	.90	.91	.94	.83

could not follow imposed transformation loss as well as selected compression distortions.

- Finally, our experiments showed that like the case of the IQM introduced in [8] the depth of the network (in our case, the use of VGG-16 and VGG-32 [36]) did not have any significant impact on the performance of the proposed VQM.

3.3 Content and Compression Analysis

Our experiments show a link between the content and compression method with the video and image saliency and so the performance of our VQM. To be more specific, the difference between video saliency and image saliency can provide a better understanding of the content. Likewise, the difference between image or video saliency of test and reference video provides information about the



(a) Flowervase (b) Chipmunks (c) Keiba

Fig. 2. Sample frames from three video clips in the CSIQ dataset.

compression method used in the video. Thus, W_H would include information about video quality, content, and distortion. Experimental results show that instead of having a value of zero and one for W_H, a fuzzy approach for selecting the value of W_H could result in improving the accuracy of our proposed VQM. That is, depending on the amount of temporal and structural variations of the video W_H could have different values. For example, our initial study has shown that in the case of the CSIQ database W_H would have a low value in Flowervase video (Figure 2(a)) while the Chipmunks and Keiba videos (Figures 2(b) and (c) respectively) would be assigned a high W_H values.

4 Conclusion and Future works

In conclusion, we proposed a set of different VQMs that are inspired by a CNN-based IQM which assesses the spatial features effectively. Saliency maps of videos added a spatial-temporal approach to our method, yielding to a series of quality scores for each frame in the video. Different schemes are then applied to these quality scores to introduced two different video quality scores for the video. Finally, using a saliency based approach to compare spatial and temporal distortions, one of the two mentioned scores are presented as the final video quality. The proposed measure was tested on the CSIQ and the Netflix public dataset. Our experimental results show that by simply differentiating between spatial and temporal distortions, our VQM could have a better accuracy. The proposed approach performs well in the case of compression based distortions while its accuracy drops in the case of distortions infected by transformation loss.

As we discussed in Section 3.3, finding the content and distortion type of a video based on spatial-temporal and spatial saliency could also improve the performance of the VQM. Further study of this issue and selecting the perfect weighting function for the two spatial and spatial-temporal VQM would be part of the future work we plan to perform.

References

1. CSIQ video quality database, http://vision.eng.shizuoka.ac.jp
2. Ahn, S., Lee, S.: Deep blind video quality assessment based on temporal human perception. In: ICIP. pp. 619–623 (2018)

3. Amirshahi, S.A.: Towards a perceptual metric for video quality assessment. Master's thesis, Norwegian University of Science and Technology (NTNU) (2010)
4. Amirshahi, S.A.: Aesthetic quality assessment of paintings. Verlag Dr. Hut (2015)
5. Amirshahi, S.A., Kadyrova, A., Pedersen, M.: How do image quality metrics perform on contrast enhanced images? In: EUVIP. pp. 232–237 (2019)
6. Amirshahi, S.A., Larabi, M.C.: Spatial-temporal video quality metric based on an estimation of qoe. In: QoMEX. pp. 84–89 (2011)
7. Amirshahi, S.A., Pedersen, M.: Future directions in image quality. In: CIC. vol. 2019, pp. 399–403 (2019)
8. Amirshahi, S.A., Pedersen, M., Yu, S.X.: Image quality assessment by comparing cnn features between images. J ELECTRON IMAGING **2017**(12), 42–51 (2017)
9. Amirshahi, S.A., Redies, C., Denzler, J.: How self-similar are artworks at different levels of spatial resolution? In: CAE. pp. 93–100 (2013)
10. Antkowiak, J., Jamal Baina, T., Baroncini, F.V., Chateau, N., FranceTelecom, F., Pessoa, A.C.F., Stephanie Colonnese, F., Contin, I.L., Caviedes, J., Philips, F.: Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000 (2000)
11. Bampis, C.G., Li, Z., Bovik, A.C.: Spatiotemporal feature integration and model fusion for full reference video quality assessment. IEEE T CIRC SYST VID **29**(8), 2256–2270 (2018)
12. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: CIVR. pp. 401–408 (2007)
13. Chan, A., Zeng, K., Mohapatra, P., Lee, S.J., Banerjee, S.: Metrics for evaluating video streaming quality in lossy ieee 802.11 wireless networks. In: Infocom. pp. 1–9 (2010)
14. Dendi, S.V.R., Krishnappa, G., Channappayya, S.S.: Full-reference video quality assessment using deep 3d convolutional neural networks. In: NCC. pp. 1–5 (2019)
15. Freitas, P.G., Akamine, W.Y., Farias, M.C.: Using multiple spatio-temporal features to estimate video quality. Signal Process. Image Commun. **64**, 1–10 (2018)
16. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS. pp. 545–552 (2007)
17. Hekstra, A.P., Beerends, J.G., Ledermann, D., De Caluwe, F., Kohler, S., Koenen, R., Rihs, S., Ehrsam, M., Schlauss, D.: Pvqm–a perceptual video quality measure. Signal Process. Image Commun. **17**(10), 781–798 (2002)
18. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electron. Lett. **44**(13), 800–801 (2008)
19. Kim, W., Kim, J., Ahn, S., Kim, J., Lee, S.: Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In: ECCV. pp. 219–234 (2018)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
21. Li, X., Guo, Q., Lu, X.: Spatiotemporal statistics for video quality assessment. IEEE T IMAGE PROCESS **25**(7), 3329–3342 (2016)
22. Li, Y., Po, L.M., Cheung, C.H., Xu, X., Feng, L., Yuan, F., Cheung, K.W.: No-reference video quality assessment with 3d shearlet transform and convolutional neural networks. IEEE T CIRC SYST VID **26**(6), 1044–1057 (2015)
23. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., Manohara, M.: Toward a practical perceptual video quality metric. The Netflix Tech Blog **6** (2016)
24. Liu, K.H., Liu, T.J., Liu, H.H., Pei, S.C.: Spatio-temporal interactive laws feature correlation method to video quality assessment. In: ICMEW. pp. 1–6 (2018)

25. Ma, Q., Zhang, L., Wang, B.: New strategy for image and video quality assessment. J ELECTRON IMAGING **19**(1), 011019 (2010)
26. Maczyta, L., Bouthemy, P., Le Meur, O.: Cnn-based temporal detection of motion saliency in videos. Pattern Recognit. Lett. **128**, 298–305 (2019)
27. Men, H., Lin, H., Saupe, D.: Spatiotemporal feature combination model for no-reference video quality assessment. In: QoMEX. pp. 1–3 (2018)
28. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D.: Considering temporal variations of spatial visual distortions in video quality assessment. IEEE J. Sel. Topics Signal Process. **3**(2), 253–265 (2009)
29. Ong, E., Lin, W., Lu, Z., Yao, S.: Colour perceptual video quality metric. In: ICIP. vol. 3, pp. III–1172 (2005)
30. Redies, C., Amirshahi, S.A., Koch, M., Denzler, J.: Phog-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects. In: ECCV. pp. 522–531 (2012)
31. Roja, B., Sandhya, B.: Saliency based assessment of videos from frame-wise quality measures. In: IACC. pp. 639–644 (2017)
32. Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE T IMAGE PROCESS **23**(3), 1352–1365 (2014)
33. Sector, I.T.S.: Objective perceptual multimedia video quality measurement in the presence of a full reference. ITU-T Recommendation J **247**, 18 (2008)
34. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE T IMAGE PROCESS **19**(2), 335–350 (2009)
35. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE T IMAGE PROCESS **15**(2), 430–444 (2006)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Vu, P.V., Chandler, D.M.: Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. J ELECTRON IMAGING **23**(1), 013016 (2014)
38. Vu, P.V., Vu, C.T., Chandler, D.M.: A spatiotemporal most-apparent-distortion model for video quality assessment. In: ICIP. pp. 2505–2508 (2011)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE T IMAGE PROCESS **13**(4), 600–612 (2004)
40. Winkler, S.: Digital video quality: vision models and metrics. John Wiley & Sons (2005)
41. You, J., Korhonen, J.: Deep neural networks for no-reference video quality assessment. In: ICIP. pp. 2349–2353 (2019)