Using Multiple Spatio-Temporal Features to Estimate Video Quality

Pedro Garcia Freitas[1,a], Welington Y.L. Akamine[b], Mylène C.Q. Farias[b]

University of Brasília (UnB), Brasília, Brazil

[a] *Department of Computer Science*
[b] *Department of Electrical Engineering*

Abstract

In this paper, we propose a new video quality metric based on a set of multiple features that incorporate texture, saliency, spatial activity, and temporal attributes. A random forest regression algorithm is used to combine these features and obtain a video quality score. Experimental results show that the proposed metric has a good performance when tested on several benchmark video quality databases, outperforming current state-of-the-art full-reference video quality metrics.

Keywords: Video Quality Assessment, Texture Operator, Spatio-temporal texture Analysis, Machine Learning, Random Forest Regression.

1. Introduction

Due the wide popularity of services that acquire, store, transmit, and display visual content, digital videos are progressively becoming part of the day-to-day lives of people. According to CISCO™, in 2021, 60 percent of all consumer Internet traffic will be from video applications [1]. Because of this, service providers concentrate a lot of effort trying to improve the quality of their service. This effort involves controlling the quality of the delivered videos, what requires the use of methods for evaluating the perceptual quality of digital videos.

There are two approaches for assessing the quality of videos. The first approach uses subjective video quality assessment (SVQA) methods. SVQA methods collect quality judgments from human viewers by performing psychophysical experiments in controlled laboratory environments. Although reliable, these experiments are typically ponderous, time-consuming, and expensive. Unfortunately, most multimedia applications require feasible quality assessment methodologies that do not require the presence of human viewers.

The second approach for assessing video quality consists of using objective video quality assessment (OVQA) methods. OVQA methods use computer algorithms to automatically estimate video quality. Nevertheless, the design of OVQA methods generally includes tests on a video quality database. These databases include a set of unimpaired video content or source sequences (SRC) and a set of impaired videos or processed video sequences (PVS), which are generated by individually processing each SRC and may contain different types of distortions. In quality databases, all sequences have an associated subjective quality score - the mean opinion score (MOS), which is produced by using a SVQA method [2].

According to Shahid *et al.* [3], OVQA methods can be classified into three categories, depending on the degree of information that is used from SRC. Full reference (FR) methods are based on comparisons of PVS and SRC, when the entire SRC is available as reference. Reduced reference (RR) methods are also based on comparisons of PVS and SRC videos, but in this case only representative features of the SRC (instead of the complete reference information) are used. No-reference (NR) methods do not require access any information about the SRC to assess quality of the PVS. In this paper, we propose a FR OVQA method.

FR OVQA methods use some chosen criteria to measure the quality difference between the PVSs and their corresponding SRCs. Ideally, the qual-

*Corresponding author
Email address: sawp@sawp.com.br (Pedro Garcia Freitas)

ity estimated by an OVQA must be in well correlated with the subjective scores available in reliable quality databases. To achieve this, most successful OVQA methods take into consideration models of the human visual system (HVS), which describe how the spatial and temporal properties of the content affect quality [4].

Spatial-based OVQA methods usually compare the PVS and its corresponding SRC to estimate the presence and strengths of typical spatial distortions [5]. Examples of spatial distortions that are frequently introduced along the communication chain and are, therefore, visible to users in most multimedia applications include ringing, additive noise, blurring, blocking, mosaic patterns, among others. For example, blocking, mosaic patterns, and blurring distortions are introduced by several compression algorithms, such as MPEG-1, MPEG-2, MPEG-4, and H.264 [6]. Ringing, on the other hand, is mostly introduced by compression algorithms that do not use a block-based decomposition, such as Motion JPEG-2000 [7]. Additive noise, which is a grainy perturbation on frame textures, may be introduced during image acquisition and transmission stages.

To assess spatial degradations, a common approach consists of using an image quality assessment (IQA) method to compute a quality score for each PVS frame. Then, a temporal strategy is used to pool these (frame) IQA scores into an overall quality score. For example, Seshadrinathan and Bovik [8] proposed a temporal hysteresis model to pool scores obtained with PSNR and SSIM [9] metrics into an overall video quality score. Using this temporal model provides a better prediction accuracy than taking a simple average of the quality scores of the video frames, an approach which has been shown to be inadequate to video quality [10].

Although it is possible to evaluate video quality using an IQA method to estimate the quality of each individual frame, this approach fails to identify the presence of temporal distortions [10]. Temporal distortions are impairments that change over time and, often, alter the intensity and movement trajectories of the pixels in a video sequence. These distortions may introduce a false perception of movement because of the introduction of additional temporal frequencies [11]. Examples of temporal distortions that are common in multimedia applications include motion compensation mismatch, mosquito noise, ghosting, smearing, and jerkiness, among others [5].

Some OVQA methods [12, 13] measure the amount of temporal distortions using optical flow algorithms [14, 15]. For instance, Manasa and Channappayya [16] proposed an OVQA method that compares optical flow statistics of SRC and PVS videos, assuming that these statistical differences are proportional to the severity of the temporal distortions. Similarly, Seshadrinathan and Bovik [11] proposed a VQA algorithm that uses an optical flow motion estimation algorithm to capture the severity of temporal artifacts in videos.

Digital videos are composed by a sequence of temporally redundant images (frames). Therefore, a natural assumption is that video quality should be modeled using a combination of spatial and temporal information, i.e. the design of OVQA methods should incorporate spatio-temporal HVS models. Seshadrinathan and Bovik [17] proposed the MOVIE index, which is an OVQA method that uses 3-D Gabor filters to assess quality in both spatial and temporal domains. Vu and Chandler [18] proposed an OVQA algorithm that splits the video into spatio-temporal slices to capture both spatial and temporal distortions. Finally, Peng *et al.* [19] proposed a motion-tuning scheme, which captures temporal distortions along motion trajectories by exploiting the space-time texture.

Although the methods cited above are state-of-the-art methods, they are not sensitive to a wide range of video distortions. To overcome this issue, in this paper we propose an OVQA method based on a combination of spatio-temporal features, such as multiscale local binary patterns [20], structural similarity [9], gradient magnitude similarity deviation [21], Riesz pyramids phase-based features [22], and spatial and temporal distortion measures [23]. To predict the overall video quality, these features are used as input to a regression algorithm. The proposed model achieves a good prediction performance when compared with other state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 describes the the proposed OVQA method. Sections 3 and 4 present the experimental setup and results, respectively. Finally, Section 5 presents the conclusions.

2. Proposed Method

In the proposed OQVA method, separate feature sets are computed independently from each other. These features sets are the following:

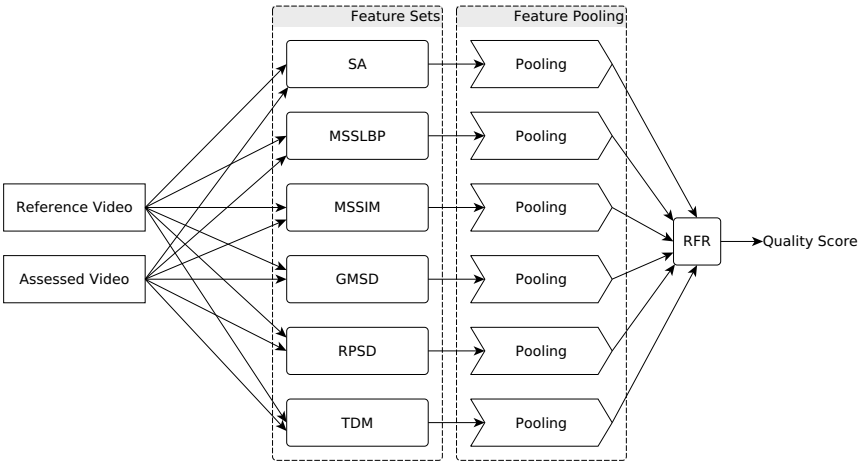


Figure 1: Block diagram of proposed method.

- Multiscale salient local binary patterns (MSLBP),
- Multiscale structural similarity (MSSIM) [9],
- Gradient magnitude similarity deviation (GMSD) [21],
- Riesz pyramids similarity deviation (RPSD),
- Spatial activity (SA) and Temporal distortion measures (TDM) [23].

Each of these feature sets are computed for the reference (SRC) and the test (PVS) videos. For each feature component, a pooling strategy is adopted and the pooled values are concatenated to generate a feature vector. Finally, the feature vector is used as input in a random forest regressor (RFR) to predict the quality score. Fig. 1 depicts a block diagram of the proposed method. In the next sections, we describe each feature set used in the proposed method.

2.1. Spatial Activity

The Spatial Activity (SA) of a pair of frames (SRC and PVS) is computed by taking the root mean square (RMS) difference between the Sobel maps of each of the frames. More specifically, let $\mathcal{S}$ be the Sobel operator [24] defined as:

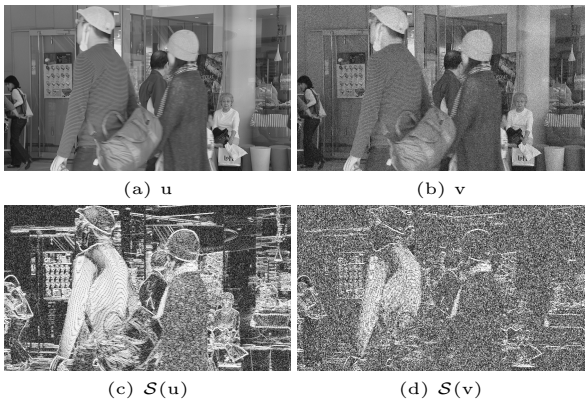
$$\mathcal{S}(z) = \sqrt{(G_1 * z)^2 + (G_1^\top * z)^2}, \qquad (1)$$

where z is the frame picture, $*$ denotes the 2-dimensional convolution operation, G_1 is the vertical Sobel filter, given by:

$$G_1 = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \qquad (2)$$

and $G_1^\top$ is the transpose of G_1 (horizontal Sobel filter).

Fig. 2 illustrates how the Sobel operator captures spatial distortions. Fig. 2-(a) and (b) show the frames with and without distortions, respectively. Their corresponding Sobel maps are shown in Fig. 2-(c) and (d). Notice that the small dif-



(a) u (b) v

(c) $\mathcal{S}$(u) (d) $\mathcal{S}$(v)

Figure 2: Effect of the spatial activity captured by the Sobel filter: (a) reference frame, (b) distorted frame, (c) Sobel map of the reference frame, and (d) Sobel map of the distorted frame.

ferences between distorted and original frames are emphasized by the Sobel operator.

Considering that v is a frame from SRC and u is the same frame from PVS, we first compute the Sobel map of both frames and take the difference between these two maps:

$$s = \mathcal{S}(u) - \mathcal{S}(v). \tag{3}$$

Then, we compute SA using the following equation:

$$SA(v,u) = \sqrt{\frac{1}{MN}\sum_{i,j}|s_{ij}|^2}, \tag{4}$$

where i and j correspond to the horizontal and vertical indices of s, respectively, and M and N are the height and width of the frames, respectively.

2.2. Multiscale Salient Local Binary Patterns

To estimate degradations, the Multiscale Salient Local Binary Patterns (MSLBP) operator makes two main assumptions. The first assumption is that visual distortions alter frame textures and their statistics. So, frames with similar distortions and similar distortion strengths have textures with similar statistical properties [20, 25, 26, 27, 28]. The second assumption is that salient visual areas are more perceptually relevant to users than non-salient areas. Therefore, visual attention models have been shown to improve the performance of OVQA methods [29, 30]. These two assumptions can be combined to design a feature descriptor that can be used to predict quality. We propose a texture descriptor that combines a visual attention model and a multiscale local binary pattern (MLBP) operator [20].

The MLBP operator was proposed by Freitas *et al.* [20] and consists of a combination of several LBP operators with different characteristics. A single LBP is computed as follows:

$$LBP_{R,P}[x,y] = \begin{cases} \sum_{p=0}^{P-1} \varsigma(t_p - t_c), & \Upsilon \leq 2, \\ P+1, & \text{otherwise}, \end{cases} \tag{5}$$

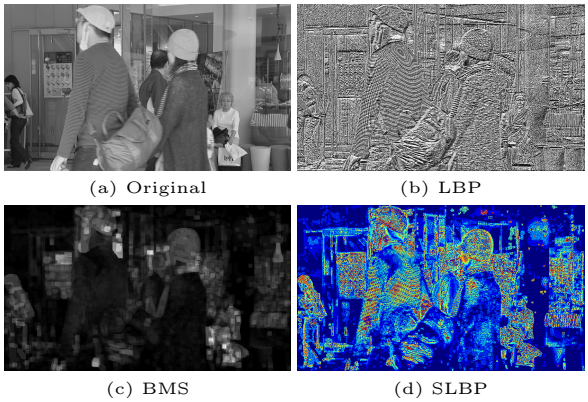
where t_c is the value of the pixel at position (x, y), t_p is the value of its neighbor, P is the total number of considered neighbors, R is the radius of this neighborhood, Υ is given by:

$$\begin{aligned}\Upsilon =& |\varsigma(t_{P-1} - t_c) - \varsigma(t_0 - t_c)| \\ &+ \sum_{p=1}^{P-1} |\varsigma(t_p - t_c) - \varsigma(t_{p-1} - t_c)|,\end{aligned}$$

and the step function $\varsigma(t)$ is given by:

$$\varsigma(t) = \begin{cases} 1 & t \geq 0, \\ 0 & \text{otherwise}. \end{cases} \tag{6}$$

The basic LBP described in the above equations is used in several computer vision applications. However, for each application, the parameters R and P that provide the best performance must be found. Freitas *et al.* [20] proposed a new operator, which is computed by varying the parameters R and P of a single LBP and combining the results. This new operator, called Multiscale Local Binary Patterns (MLBP), has been used with success to estimate image quality.



(a) Original (b) LBP (c) BMS (d) SLBP

Figure 3: (a) Original frame, (b) a sample LBP, (c) corresponding BMS map, and (d) resulting SLBP map.

In this work, we propose a modification to the MLBP operator, which consists of incorporating visual saliency. The proposed operator, the Multiscale Salient Local Binary Patterns (MSLBP), is computed as follows. First, we estimate the saliency of the different areas of a frame using a computational visual attention model. To keep the computational complexity low, we chose the Boolean map-based saliency (BMS) model [31]. When compared with other state-of-the-art visual attention models, BMS is noticeably faster, while still providing a good performance.

Next, we compute the $MLBP$, which gives the local texture associated to each frame pixel, for a set of parameters R and P. Fig. 3-(a), (b), and (c) depict the original frame, one sample LBP image, and the corresponding BMS map, respectively. The BMS map is used to give a weight to each pixel of the LBP maps, what is achieved by multiplying each pixel of the LBP maps by the correspond-

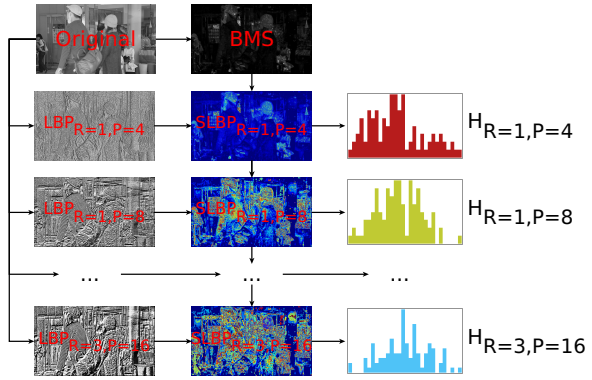


Figure 4: Multiple histogram generation from SLBP.

ing value of the BMS map. This weighting process generates the salient local binary patterns (SLBP) maps. Fig. 3-(d) shows a sample SLBP map for Fig. 3-(a).

Then, we generate the histogram of the LBP maps weighted by the BMS maps. The histogram $\mathbf{H} = \{h[0], h[1], \cdots, h[P+1]\}$ is given by the following expression:

$$h[\phi] = \sum_i \sum_j BMS[i,j] \cdot \Delta(LBP[i,j], \phi), \quad (7)$$

where

$$\Delta(v,u) = \begin{cases} 1 & v = u, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

As mentioned earlier, the MLBP operator generates different LBP maps at different scales. We multiply each of the generated LBP maps by the BMS map in order to generate multiple SLBP at different scales. Therefore, we acquire a set of MSLBP maps and, then, compute the histograms for each of these maps.

The MSLBP histograms describe the textures of the videos, but not the differences in quality between the sequences u (SRC) and v (PVS). Since the size of these histograms can be an issue, especially when we take into account the several frames of a video, we compute the similarity between the histograms of u, $H^U_{R,P}$, and v, $H^V_{R,P}$, using the following metric:

$$JSD(p,q) = \frac{KLD(p,r) + KLD(q,r)}{2}, \quad (9)$$

where

$$KLD(p,q) = \sum p(x) \log\left(\frac{p(x)}{q(x)}\right), \quad (10)$$

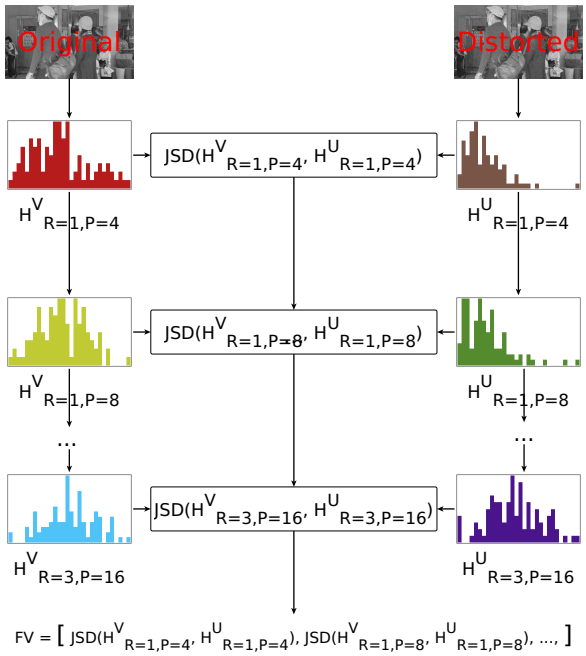


Figure 5: Generation of the MSLBP feature set.

and

$$r = \frac{p(x) + q(x)}{2}. \quad (11)$$

In the above equations, $p(x)$ and $q(x)$ are the probability mass functions (PMF) representing the bins of $H^U_{R,P}$ and $H^V_{R,P}$, respectively. JSD is the Jensen-Shannon divergence [32] and KLD is the Kullback-Leibler divergence [33]. JSD was chosen because it is a symmetric version of the mutual information (KLD) and it is always a finite value [34].

Fig. 5 illustrates the construction of the MSLBP feature set. After performing the steps depicted in Fig. 4, for both u and v frames, JSD is used to compute the divergences between the histograms of u and v. The divergence values compose the MSLBP feature vector (FV).

2.3. Multiscale Structural Similarity

The Structural SIMilarity (SSIM) index is a popular IQA method based on luminance, contrast, and structure measures [9]. It is calculated using the following equation:

$$SSIM(u,v) = \frac{(2\mu_u\mu_v + C_1)(2\sigma_{uv} + C_2)}{(\mu_u^2 + \mu_v^2 + C_1)(\sigma_u^2 + \sigma_v^2 + C_2)}, \quad (12)$$

where μ_f, σ_f are the average and standard deviation of the frame f, σ_{fg} is the covariance of frames

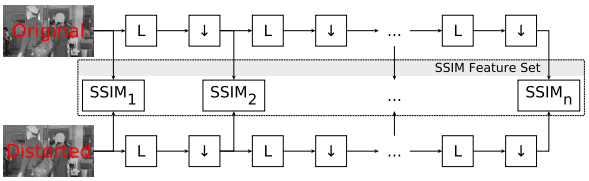


Figure 6: Multiscale structural similarity measurement index. L: low-pass filtering, ↓: downsampling by factor 2.

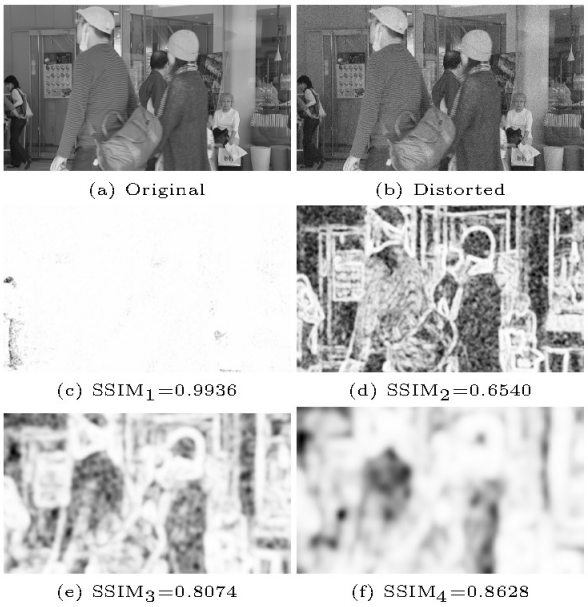


Figure 7: Multi-scale structural similarity index maps.

f and g, and C_1 and C_2 are constants used to stabilize divisions with a weak denominator. In this work, we use the mean of the SSIM index map to quantify the quality difference between of u and v, at different scales [35]. Fig. 6 depicts the feature extraction using the SSIM index for different scales, while Fig. 7 shows the SSIM maps for four different scales.

2.4. Gradient Magnitude Similarity Deviation

The Gradient Magnitude Similarity Deviation (GMSD) is an IQA method based on the standard deviation of the gradient magnitude similarity (GMS) map [21]. The GMS map is computed as follows:

$$GMS(u,v) = \frac{2 \cdot m(u) \cdot m(v) + c}{m(u)^2 + m(v)^2 + c}, \qquad (13)$$

where u is the SRC frame, v is the PVS frame, c is a positive constant that garantees numerical stability,

Figure 8: Gradient magnitude similarity map.

and $m(z)$ is defined as:

$$m(z) = \sqrt{(z * G_2)^2 + (z * G_2^\top)^2}. \qquad (14)$$

In the above equation, $*$ denotes the convolution operation, G_2 represents the Prewitt filter along the vertical direction, which is defined as:

$$G_2 = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix}, \qquad (15)$$

and $G_2^\top$ is the transpose of G_2, which corresponds to the Prewitt filter along the horizontal direction.

Fig. 8-(c) depicts the GMS map, which serves as a local quality map of the distorted frame. The GMSD index is computed as follows:

$$GMSD(u,v) = \sqrt{\frac{1}{NM} \sum_{i,j} \left(GMS(u,v) - \overline{GMS(u,v)} \right)^2}, \qquad (16)$$

where $\overline{GMS(u,v)}$ is the gradient magnitude similarity mean, computed as follows:

$$\overline{GMS(u,v)} = \frac{1}{NM} \sum_{i,j} GMS(u,v). \qquad (17)$$

2.5. Riesz Pyramids Similarity Deviation

Wadhwa *et al.* proposed a technique technique to represent images, which is called Riesz pyramids [22]. Their work was inspired by the work of Simoncelli and Freeman [36]. The Riesz pyramids make use of a highpass filter $h_H[n]$ and a lowpass filter $h_L[n]$. First, the frame is highpassed to generate the top level of the pyramid. Next, the frame is lowpassed and downsampled. This process is recursively applied to the downsampled image to generate the pyramid representation, as illustrated in Fig. 9.

To generate the features using the Riesz pyramids, we compare the highpassed v frames with the highpassed u frames at each pyramid level. More

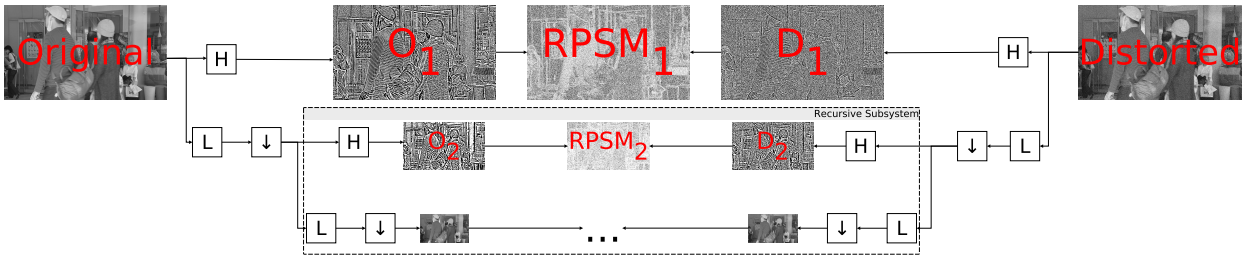


Figure 9: Diagram of RPSD pyramid construction of both original and distorted frames. The lowpass and highpass filters can be recursively used with subsampling to produce a sequence of critically bandpassed frames. The blocks ↓ denote downsampling by a factor of 2. L and H denote linear shift invariant lowpass and highpass filters, respectively.

specifically, for each level i, we compute the high-passed version of the u (u_i^h) and v (v_i^h) frames. Using u_i^h and v_i^h, we generate the i-th Riesz pyramid similarity map (RPSM_i), as follows:

$$\text{RPSM}_i(u,v) = \frac{2 \cdot n(u_i^h) \cdot n(v_i^h) + c}{n(u_i^h)^2 + n(v_i^h)^2 + c}, \tag{18}$$

where

$$n(z) = \sqrt{(z * G_3)^2 + (z * G_3^\top)^2} \tag{19}$$

and

$$G_3 = \begin{bmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix}. \tag{20}$$

From the the i-th Riesz pyramid similarity map, we compute the Riesz Pyramids Similarity Deviation (RPSD), as follows:

$$\text{RPSD}_i(u,v) = \sqrt{\frac{1}{NM} \sum_{i,j} \left(\text{RPSM}_i(u,v) - \overline{\text{RPSM}_i(u,v)}\right)^2}, \tag{21}$$

where $\overline{\text{RPSM}_i(u,v)}$ is the mean of i-th RPSM map, computed as follows:

$$\overline{\text{RPSM}_i(u,v)} = \frac{1}{NM} \sum_{i,j} \text{RPSM}_i(u,v). \tag{22}$$

Finally, the RPSD feature set is composed by all i-th RPSD values, given by:

$$RPSD(u,v) = \{\text{RPSD}_1(u,v), \text{RPSD}_2(u,v), \cdots\}.$$

2.6. Temporal Distortion Measures

Temporal distortion measures (TDM) are based on the spatio-temporal texture representation proposed by Derpanis & Wilders [23]. The texture representation is computed using a bank of spatio-temporal 3-rd derivative Gaussian filters, defined as follows:

$$\mathcal{G}_{3_\theta} = \frac{\partial^3}{\partial \theta^3} k e^{-(x^2+y^2+t^2)}, \tag{23}$$

where θ is the unit vector that corresponds to the spatio-temporal direction of the filter and k is a normalization factor [37].

First, we compute the sum of the pointwise squared response of the filter output over a space-time region, Ω, producing the following energy measurement:

$$\mathcal{E}_{3_\theta}(x,y,t) = \sum_{x,y,t \in \Omega} (\mathcal{G}_{3_\theta} * V(x,y,t))^2, \tag{24}$$

where V is the input spatio-temporal signal (video).

Let $\mathcal{E}^v_{k_\theta}$ and $\mathcal{E}^u_{k_\theta}$ be the energy measurements, along the direction k_θ of u and v, respectively. The temporal distortion (TD) measure at (x,y,t) is obtained by computing the distance between the two corresponding energy distributions in u and v:

$$\text{TD}(x,y,t) = \sqrt{\sum_k (\mathcal{E}^v_{k_\theta}(x,y,t) - \mathcal{E}^u_{k_\theta}(x,y,t))^2}. \tag{25}$$

Finally, the TDM is computed along time:

$$\text{TDM}(t) = \sum_{x,y} \text{TD}(x,y,t). \tag{26}$$

2.7. Feature Pooling and Mapping

The formulation presented in the previous sections describes the generation of a set of values corresponding to each feature. To convert these sets of values into a single score per feature, we use a feature pooling strategy based on the Minkowski norm. The Minkowski norm is computed as follows:

$$Q_k = \sqrt[4]{\sum_t f_k(t)^4}, \tag{27}$$

where $f_k(t)$ corresponds to the k-th feature at its t-th value.

Next, the pooled features $Q_1, Q_2, \cdots Q_n$ are treated as inputs to a random forest regression (RFR) algorithm, which gives an estimated video quality score. We choose the RFR method, rather than the popular support vector regression (SVR), because RFR does not require the hyper-parameter tuning. Furthermore, RFR has been successfully used in several pattern recognition applications [38].

3. Experimental Setup and Protocol

There are a number of existing databases created for the evaluation of video quality. In this work, we use the following databases:

- Computational and Subjective Image Quality (CSIQ) Video Database [39]: The database contains 12 high-quality reference videos and 216 distorted videos from 6 different types of distortion. All videos are in raw YUV420 format with a resolution of 832x480 pixels, a duration of 10 seconds at 24, 25, 30, 50, or 60 fps. The distortion types consist of 4 compression-based distortion types including H.264 compression (H.264), HEVC/H.265 compression (HEVC), Motion JPEG compression (MJPEG), and Wavelet-based compression using the Snow codec (SNOW). The database also include 2 transmission-based distortion types, namely wireless transmission loss (WIRELESS), and additive white noise (WN).

- Image and Video Processing Laboratory (IVPL) Database [40]: The database contains 10 reference videos and 128 distorted videos from 4 different types of distortion. All videos are in raw YUV420 format with a resolution of 1920x1088 (progressive) at 25 fps. The distortion types consist of 3 compression-based distortion types including H.264 compression (H.264), Dirac coding (DIRAC), and MPEG2. The database also include 1 transmission-based distortion (IP).

- MediaCommLab Video (MCL-V) Database [41]: The database contains 12 uncompressed source video clips with HD resolution (1080p). The database captures two typical video distortion types in video streaming services, including compression (H.264) and image size scaling (SD H.264). Four distortion levels are adopted for each distortion type. There are 96 distorted video clips in total.

- Laboratory for Image & Video Engineering (LIVE) Video Database [42]: The database contains 10 high-quality reference videos with 15 distorted videos per reference in a total of 150 test videos. The videos files have planar YUV420 format with spatial resolution of 768x432 pixels. The distortion types consist of 2 compression-based distortion types, including H.264 compression (H.264) and MPEG2. The database also include 2 transmission-based distortion, including simulated transmission of H.264 compressed bitstreams through (1) error-prone IP networks and (2) error-prone wireless networks. These two distortions are grouped into a single category, named "transmission errors" (TE).

- LIVE Public-Domain Subjective Mobile Video Quality Database (LIVE-M) [43, 44]: The database consists of 10 raw HD reference videos and 200 distorted videos (4 compression + 4 wireless packet-loss + 4 frame-freezes + 3 rate-adapted + 5 temporal dynamics per reference), each of resolution 1280x720 at a frame rate of 30 fps, and of duration 15 seconds each. For testing purposes, we excluded the frame-freezes distortions.

We compared the proposed method with a set of publicly available standard-of-the-art VQA methods. The chosen VQA methods are SSTS-GMSD [45], STRRED [46], and ViS3 [18]. Additionally, we also compared the proposed algorithm with three well-established IQA metrics, namely PSNR, SSIM [9], GMSD [21].

In our test methodology, we adopted a grouped shuffle split cross-validation approach. This approach consists of dividing each single database into two content-independent subsets (training and testing), where videos generated from one reference (same content) in the testing subset are not present in the training subset, and vice-versa. This division is illustrated in Fig. 10. Each reference video and its corresponding distorted versions belong to the same group of scenes. After grouping videos by content (versions of the same reference), 80% of groups are randomly selected for training and the remaining 20% are used for testing. As depicted

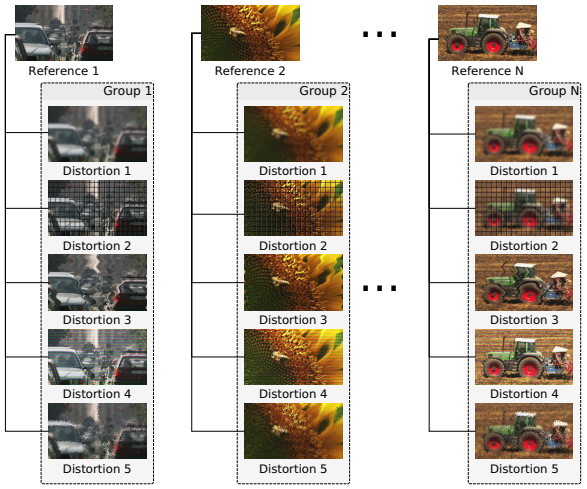


Figure 10: Grouping of images used on testing and training procedure.

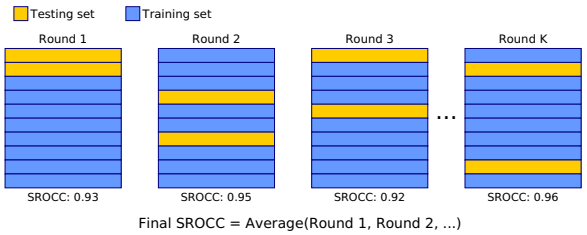


Figure 11: Visual representation of the k training and tent simulations (k-fold). The reported result is the average of the k results.

in Fig. 11, this 80-20 split is repeated for 1,000 rounds of simulation and the average correlation is reported. For the methods that are not based on training, we adopted the same strategy of splitting, but considered only the 20% testing group and discarded the 80% group reserved for training.

To assess the performance of the tested methods, each simulation consisted of selecting a set of videos, predicting the quality score using each tested VQA method, and then comparing the scores predicted with the VQA method and the subjective scores provided by the databases. To compare the predicted and subjective quality scores, we used Spearman's rank correlation coefficient (SROCC) and Pearson's linear correlation coefficient (LCC).

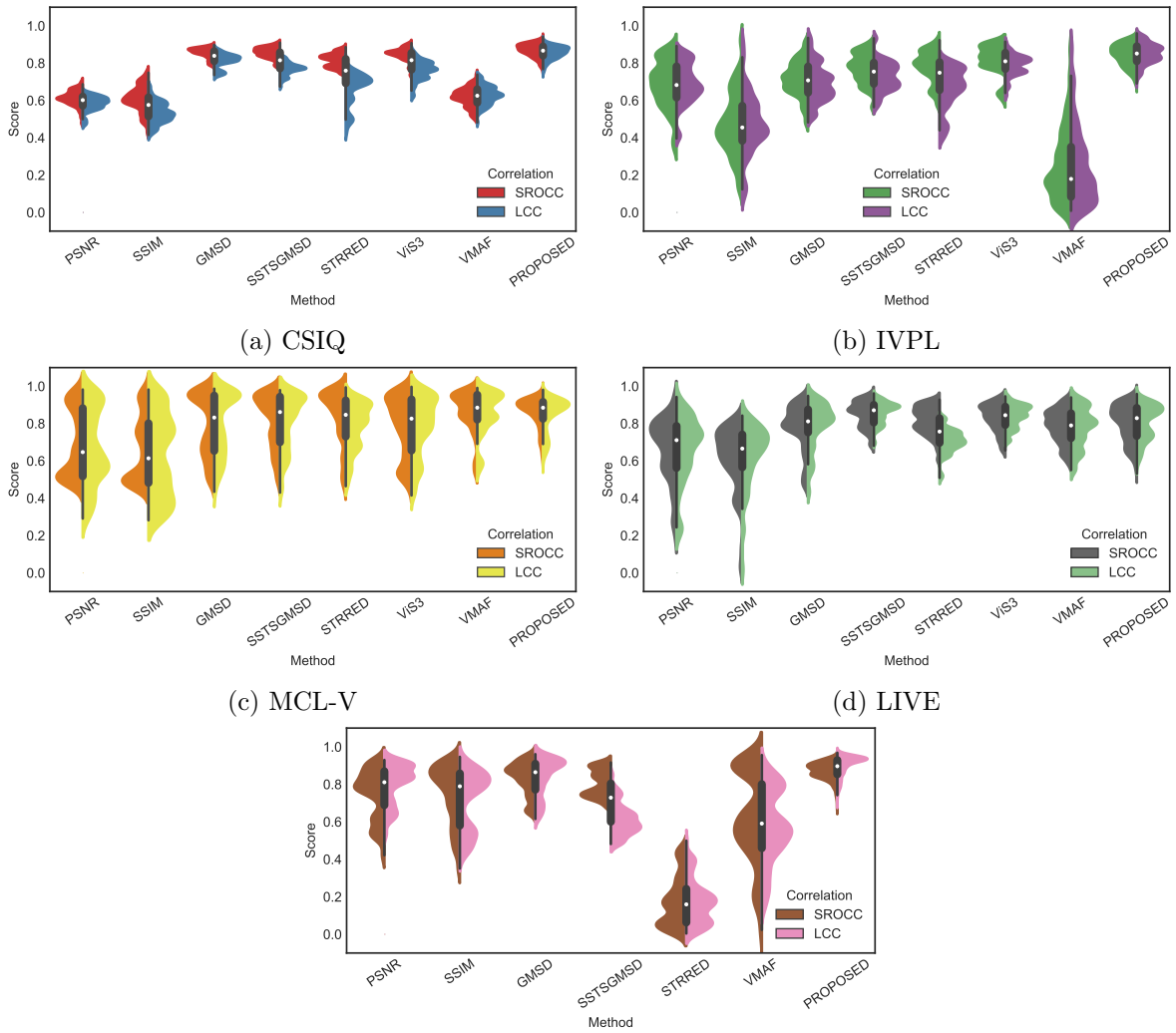
4. Experimental Results

Fig. 12 depicts the distribution of the correlation scores over 1,000 rounds of simulations. In order to evaluate if values are clustered around the median, we choose the violin plot [47] to illustrate the data. The violin plot was chosen because it combines the convenience of a box plot, which summarizes important statistics (e.g. median, range and quartiles), and of a kernel density plot, which shows details of the data distribution. In Fig. 12, the white dots represent the median, the wider bars in the center represent the interquartile range, and the fine lines represent the 95% confidence interval. On the left side of the gray lines is the kernel estimation of the distribution of SROCC scores. Similarly, on the right side of the gray lines is the kernel estimation of the distribution of LCC scores. In both sides, larger sections of the violin plots depict a higher probability of achieving these correlation scores, while narrower sections depict a lower probability.

Fig. 12-(a) presents the SROCC ans LCC violin plot for the CSIQ dataset. Notice that the proposed method shows the highest SROCC value, when compared to the state-of-the-art metrics, followed by SSTSGMSD, GMSD, ViS3, and STRRED. Since CSIQ contains two transmission-based distortions, it is expected that IQA methods present a worse performance, which explains the differences between the PSNR and SSIM results when compared with other methods. Surprisingly, GMSD presents a competitive performance, having a performance similar to its video-based version, SSTSGMSD.

Among the different tested quality metrics, the proposed approach has the highest SROCC in the IVPL and LIVE-M datasets, as displayed in Figs. 12-(b) and (e). Notice that both mean and median correlation values are higher for the proposed method. Taking a closer look at the interquartile range of the methods, we notice that the proposed method also presents a smaller dispersion, what indicates that it is more stable along multiple simulations. Therefore, the proposed method is significantly better than all tested methods on IVPL and LIVE-M datasets.

Fig. 12-(c) shows the SROCC violin plot for the MCL-V dataset. Notice that the performance of the proposed method is among the best performances, although its SROCC median value overlaps with the SROCC median value for STRRED and SSTSGMSD. However, observing again the interquartile range, we notice that the proposed method has a narrow range and, therefore, represents a smaller dispersion. Taking this into consideration, the performance of the proposed method is slightly better than the performance of STRRED and SSTSGMSD



(a) CSIQ

(b) IVPL

(c) MCL-V

(d) LIVE

(e) LIVE-M

Figure 12: Violin plots of SROCC results of the tested OVQA methods for CSIQ, IVPL, MCL-V, LIVE, and LIVE-M databases.

and significantly better than the performance of ViS3, GMSD, SSIM, and PSNR.

Fig. 12-(d) shows the SROCC violin plot for the LIVE dataset. In this case, the proposed method has one of the highest median SROCC values. While this performance is not significantly better than the performance of ViS3, STRRED, and SSTSGMSD, it outperforms the other tested state-of-the-art VQA methods.

Finally, Fig. 12-(e) depicts the SROCC violin plot for the LIVE-M dataset. Notice that, in this database, the proposed method represents a substantial prediction improvement in relation to the other metrics. In addition to a higher median value, the proposed method has a smaller confidence interval. More specifically, the interquartile range indicates that the proposed method has a narrower spread of the correlation values.

Table 1 depicts the average SROCC results, separated for the different distortions in each database. In this table, each row of the first column specifies the database, while the second column lists the distortion of the given database. The distortion named as 'ALL' corresponds to the general case that includes all types of distortions, i.e. it corresponds to the results presented in Fig. 12. In each line, the highlighted values (in bold) represent the best average SROCC values for each distortion.

Notice that the proposed method is among the top best three metrics. It is worth mentioning that, among all subsets, the proposed methods is the only one that has average SROCC values greater than 0.8 for almost all distortions. The only exception is the 'Temporal Dynamics' distortion of the LIVE-M dataset. These results indicate that the proposed method is the adequate for most practical multimedia scenarios, where several types of distortions are present (besides compression and transmission distortions). Furthermore, the proposed method presents the best results for 20 out of the 24 cases (83.3%), what makes it the method with the highest prediction accuracy.

5. Conclusions

In this study, we proposed a new full reference video quality assessment method. The proposed method is a machine learning based method that uses multiple spatio-temporal features. A random forest regression algorithm is used to map the multiple features into subjective scores. Based on the calculated Spearman correlation values, the proposed approach outperforms state-of-the-art video quality metrics for most datasets and distortion types. In cases where the proposed approach does not have the best performance, it is among the 3 best performing metrics, providing a competitive prediction performance. In future works, we plan to investigate how to adapt the feature sets to reduced and no-reference scenarios.

Achnowledgment

This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), by the Coordenacção de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and by the University of Brasília.

References

[1] Cisco, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 (Feb. 2017).

[2] M. H. Pinson, L. Janowski, Z. Papir, Video quality assessment: Subjective testing of entertainment scenes, IEEE Signal Processing Magazine 32 (1) (2015) 101–114.

[3] M. Shahid, A. Rossholm, B. Lövström, H.-J. Zepernick, No-reference image and video quality assessment: a classification and review of recent approaches, EURASIP Journal on image and Video Processing 2014 (1) (2014) 40.

[4] B. A. Wandell, Foundations of vision., Sinauer Associates, 1995.

[5] M. Yuen, H. Wu, A survey of hybrid MC/DPCM/DCT video coding distortions, Signal processing 70 (3) (1998) 247–278.

[6] A. Leontaris, A. R. Reibman, Comparison of blocking and blurring metrics for video compression, in: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, Vol. 2, IEEE, 2005, pp. ii–585.

[7] T. Fukuhara, K. Katoh, S. Kimura, K. Hosaka, A. Leung, Motion-JPEG2000 standardization and target market, in: Image Processing, 2000. Proceedings. 2000 International Conference on, Vol. 2, IEEE, 2000, pp. 57–60.

[8] K. Seshadrinathan, A. C. Bovik, Temporal hysteresis model of time varying subjective video quality, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011, pp. 1153–1156.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.

[10] M. Narwaria, W. Lin, A. Liu, Low-complexity video quality assessment using temporal quality variations, IEEE Transactions on Multimedia 14 (3) (2012) 525–535.

[11] K. Seshadrinathan, A. C. Bovik, Motion-based perceptual quality assessment of video, in: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, 2009, pp. 72400X–72400X.

[12] Z. Wang, Q. Li, Video quality assessment using a statistical model of human visual speed perception, JOSA A 24 (12) (2007) B61–B69.

[13] M. C. Farias, M. Carli, A. Neri, S. K. Mitra, Video quality assessment based on data hiding driven by optical flow information, in: Electronic Imaging 2004, International Society for Optics and Photonics, 2003, pp. 190–200.

[14] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, Image analysis (2003) 363–370.

[15] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid, Deepflow: Large displacement optical flow with deep matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1385–1392.

[16] K. Manasa, S. S. Channappayya, An optical flow-based full reference video quality assessment algorithm, IEEE Transactions on Image Processing 25 (6) (2016) 2480–2492.

[17] K. Seshadrinathan, A. C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE transactions on image processing 19 (2) (2010) 335–350.

[18] P. V. Vu, D. M. Chandler, ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices, Journal of Electronic Imaging 23 (1) (2014) 013016–013016.

[19] P. Peng, K. Cannons, Z.-N. Li, Efficient video quality assessment based on spacetime texture representation, in: Proceedings of the 21st ACM international conference on Multimedia, ACM, 2013, pp. 641–644.

[20] P. G. Freitas, W. Y. Akamine, M. C. Farias, Blind image quality assessment using multiscale local binary patterns, Journal of Imaging Science and Technology 60 (6) (2016) 60405–1.

Table 1: Average LCC and SROCC of 1,000 runs of rounds of simulations on tested databases

Database	Distortion	PSNR		SSIM		GMSD		SSTSGMSD		STRRED		ViS3		VMAF		PROPOSED	
		SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC
CSIQ	H.264	0.8023	0.8221	0.8440	0.8453	0.9436	0.9251	0.9137	0.8944	**0.9768**	0.8741	0.9194	0.8695	0.9284	0.9359	0.9166	***0.9419***
	HEVC	0.7846	0.7969	0.8136	0.7886	0.9418	0.9351	0.9292	0.8972	0.9135	0.7292	0.9173	0.8632	0.9274	0.9113	**0.9501**	***0.9631***
	MJPEG	0.5086	0.4858	0.7969	0.8007	0.8842	0.8806	0.8803	0.8227	0.7289	0.7375	0.7348	0.7414	**0.8871**	0.7273	0.8833	***0.9066***
	TE	0.8512	***0.8267***	0.8317	0.7722	**0.8680**	0.8245	0.8187	0.7449	0.8476	0.7375	0.8442	0.8069	0.7701	0.7689	0.7833	0.8122
	WC	0.7586	0.7742	0.7539	0.7461	0.8659	0.8743	0.8584	0.8291	**0.9459**	0.8438	0.8998	0.8516	0.8957	0.8919	0.8833	***0.9071***
	WN	0.9063	0.9464	0.9300	0.9302	0.9094	0.9089	0.8818	0.8731	**0.9305**	0.8325	0.9202	0.9001	0.8831	0.8641	0.9166	***0.9492***
	ALL	0.6104	0.5832	0.6077	0.5342	0.8540	0.8125	0.8546	0.7112	0.8134	0.6627	0.8483	0.7402	0.6160	0.6282	**0.8688**	***0.8564***
IVPL	DIRAC	0.8532	0.8469	0.7846	0.7289	0.8229	0.8290	0.8140	0.8011	0.8527	0.7017	**0.9132**	0.8578	0.8858	0.8995	0.8857	***0.9289***
	H.264	0.8154	0.7998	0.6636	0.6931	**0.8690**	0.8597	0.8435	0.8323	0.8614	0.5892	0.8425	0.8582	0.8662	***0.8845***	0.8571	0.8781
	MPEG2	0.6974	0.6431	0.5884	0.6269	**0.8318**	0.8440	0.7864	0.7431	0.6774	0.6488	0.7939	0.8134	0.7570	0.8434	0.8285	***0.9165***
	TE	0.6283	0.5861	0.0481	0.0410	0.7093	0.6168	0.7695	0.6335	0.6650	0.2914	0.7246	0.6080	0.3103	0.3448	**0.8333**	***0.8692***
	ALL	0.6864	0.6933	0.4710	0.4725	0.7154	0.6935	0.7751	0.7267	0.7796	0.6645	0.8199	0.7885	0.2557	0.2404	**0.8433**	***0.8525***
MCL-V	H.264	0.4215	0.4654	0.3545	0.3766	0.6419	0.6528	0.6946	0.6855	0.7716	0.5948	0.5868	0.6171	0.7938	0.7786	**0.8671**	***0.8755***
	SD H.264	0.4925	0.4630	0.4400	0.3774	0.6376	0.6300	0.6817	0.6732	0.7040	0.5883	0.6890	0.6479	0.7524	0.7590	**0.8741**	***0.8861***
	ALL	0.6912	0.6676	0.6619	0.6204	0.7925	0.7932	0.8191	0.8072	0.8439	0.7824	0.7849	0.7855	**0.8654**	***0.8578***	0.8638	0.8485
LIVE	H.264	0.4729	0.5390	0.6561	0.6079	0.6471	0.6273	0.7938	0.8087	0.8193	0.7649	0.7685	0.7890	0.7476	0.7286	**0.8809**	***0.8877***
	MPEG2	0.3830	0.4009	0.5609	0.5743	0.6915	0.6627	0.8123	0.8135	0.7193	0.7425	0.7362	0.7510	0.7025	0.6967	**0.8809**	***0.8819***
	TE	0.5798	0.5938	0.5151	0.5051	0.7457	0.7621	0.8157	0.8300	0.7934	0.7554	**0.8372**	0.8473	0.7257	0.6791	0.8285	***0.8721***
	ALL	0.6614	0.6631	0.5251	0.4997	0.7262	0.7291	**0.8387**	***0.8458***	0.8007	0.6922	0.8168	0.8263	0.7521	0.7288	0.8246	0.8367
LIVE-M	COMP	0.8270	0.7945	0.7172	0.7508	0.8662	0.8608	0.8713	0.8170	0.0881	0.0795	0.8607	0.8231	**0.9555**	0.9531	0.9523	***0.9621***
	RA	0.6353	0.5556	0.6014	0.6001	0.7312	0.7438	0.7666	0.7359	0.0629	0.0424	0.7550	0.7491	0.8968	0.8956	**0.9428**	***0.9044***
	TD	0.2917	0.3671	0.2850	0.3366	0.3649	0.4488	0.3752	0.4019	0.1979	0.2021	0.3746	0.4257	**0.4836**	***0.5104***	0.2999	0.3619
	WPL	0.7897	0.7709	0.6929	0.7109	0.8468	0.8379	0.8286	0.7917	0.1014	0.0834	0.8394	0.8018	0.9514	0.9578	**0.9523**	***0.9661***
	ALL	0.7437	0.8005	0.7522	0.7096	0.8321	0.8392	0.8035	0.6153	0.1762	0.1706	0.8248	0.6927	0.6242	0.5766	**0.8713**	***0.9023***

[21] W. Xue, L. Zhang, X. Mou, A. C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Transactions on Image Processing 23 (2) (2014) 684–695.

[22] N. Wadhwa, M. Rubinstein, F. Durand, W. T. Freeman, Riesz pyramids for fast phase-based video magnification, in: Computational Photography (ICCP), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–10.

[23] K. G. Derpanis, R. Wildes, Spacetime texture representation and recognition based on a spatiotemporal orientation analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (6) (2012) 1193–1205.

[24] I. Sobel, An isotropic 3× 3 image gradient operator, Machine Vision for three-demensional Sciences.

[25] Y. Zhang, J. Wu, X. Xie, G. Shi, Blind image quality assessment based on local quantized pattern, in: Pacific Rim Conference on Multimedia, Springer, 2016, pp. 241–251.

[26] Q. Li, W. Lin, Y. Fang, BSD: Blind image quality assessment based on structural degradation, Neurocomputing 236 (2017) 93–103.

[27] J. Wu, W. Lin, G. Shi, Image quality assessment with degradation on spatial structure, IEEE Signal processing letters 21 (4) (2014) 437–440.

[28] P. G. Freitas, W. Y. Akamine, M. C. Farias, No-reference image quality assessment based on statistics of local ternary pattern, in: Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on, IEEE, 2016, pp. 1–6.

[29] M. C. Farias, W. Y. Akamine, On performance of image quality metrics enhanced with visual attention computational models, Electronics letters 48 (11) (2012) 631–633.

[30] W. Y. Akamine, M. C. Farias, Video quality assessment using visual attention computational models, Journal of Electronic Imaging 23 (6) (2014) 061107–061107.

[31] J. Zhang, S. Sclaroff, Exploiting surroundedness for saliency detection: a Boolean map approach, IEEE transactions on pattern analysis and machine intelligence 38 (5) (2016) 889–902.

[32] B. Fuglede, F. Topsoe, Jensen-Shannon divergence and hilbert space embedding, in: Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on, IEEE, 2004, p. 31.

[33] J. M. Joyce, Kullback-Leibler divergence, in: International Encyclopedia of Statistical Science, Springer, 2011, pp. 720–722.

[34] J. A. Thomas, T. M. Cover, Elements of information theory, John Wiley & Sons, 2006.

[35] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on, Vol. 2, IEEE, 2003, pp. 1398–1402.

[36] E. P. Simoncelli, W. T. Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, in: Image Processing, 1995. Proceedings., International Conference on, Vol. 3, IEEE, 1995, pp. 444–447.

[37] K. G. Derpanis, J. M. Gryn, Three-dimensional nth derivative of gaussian separable steerable filters, in: Image Processing, 2005. ICIP 2005. IEEE International Conference on, Vol. 3, IEEE, 2005, pp. III–553.

[38] M. Liu, M. Wang, J. Wang, D. Li, Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and chinese vinegar, Sensors and Actuators B: Chemical 177 (2013) 970–980.

[39] CSIQ video quality database, http://vision.eng.shizuoka.ac.jp, accessed: 2017-05-09.

[40] IVP Subjective Quality Video Database, http://ivp.ee.cuhk.edu.hk/research/database/subjective,

accessed: 2017-05-09.

[41] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, C.-C. J. Kuo, MCL-V: A streaming video quality assessment database, Journal of Visual Communication and Image Representation 30 (2015) 1–9.

[42] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, L. K. Cormack, Study of subjective and objective quality assessment of video, IEEE transactions on Image Processing 19 (6) (2010) 1427–1441.

[43] A. K. Moorthy, L. K. Choi, A. C. Bovik, G. De Veciana, Video quality assessment on mobile devices: Subjective, behavioral and objective studies, IEEE Journal of Selected Topics in Signal Processing 6 (6) (2012) 652–671.

[44] A. K. Moorthy, L. K. Choi, G. De Veciana, A. C. Bovik, Subjective analysis of video quality on mobile devices, in: Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, Arizona, Citeseer, 2012.

[45] P. Yan, X. Mou, W. Xue, Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices, in: SPIE/IS&T Electronic Imaging, International Society for Optics and Photonics, 2015, pp. 94110M–94110M.

[46] R. Soundararajan, A. C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Transactions on Circuits and Systems for Video Technology 23 (4) (2013) 684–694.

[47] J. L. Hintze, R. D. Nelson, Violin plots: a box plot-density trace synergism, The American Statistician 52 (2) (1998) 181–184.