No-Reference Video Quality Assessment Using Spatial, Temporal, Transform, and Spatiotemporal Features

Yi-Lun Chiang and Jin-Jang Leou

Department of Computer Science and Information Engineering
National Chung Cheng University, Chiayi 621, Taiwan, ROC

Abstract — In this study, a no-reference (NR) video quality assessment approach using spatial, temporal, transform, and spatiotemporal features is proposed. First, three different resolutions of each video sequence by down-sampling are generated and then spatial, temporal, transform, and spatiotemporal features are extracted. Spatial, temporal, and transform feature maps are converted into feature vectors using histogram statistics, whereas principal component analysis (PCA) is employed to perform dimension reduction on spatiotemporal feature vectors. After normalization and concatenation, support vector regression (SVR) is employed to perform machine learning and video quality assessment score estimation. Based on the experimental results obtained in this study, in terms of PLCC and SROCC, the performance of the proposed approach is better than those of eight comparison approaches.

Keywords: no-reference video quality assessment, spatial, temporal, transform, and spatiotemporal features, feature extraction, dimension reduction, support vector regression.

1 Introduction

Nowadays, a lot of video contents are transmitted over the Internet or wireless networks. Due to video compression, noise corruption, and transmission medium properties, various types of distortions may be introduced, which will degrade the quality of experience (QoE) of human users (human perception). Video quality assessment (VQA) (subjective or objective) is very important for video providers and users. Although subjective VQA can achieve higher accuracy of video quality assessment, it is time-consuming and expensive. Additionally, in terms of the use of original video information, objective VQA approaches can be partitioned to three categories, namely, full-reference (FR) [1-2], reduced-reference (RR) [3], and no-reference (NR) [4-7]. For FR VQA approaches, such as peak-signal-to-noise ratio (PSNR), structural similarity index (SSIM), and multi-scale SSIM, the original video sequence and the distorted video sequence are employed (compared) to estimate the video quality assessment score. For RR VQA approaches [3], partial information of the original video sequence and that of the distorted video sequence are used to estimate the video quality assessment score. For NF VQA approaches [4-7], only the distorted video sequence is directly used to estimate the video quality assessment score. In this study, a no-reference video quality assessment approach using spatial, temporal, transform, and spatiotemporal features is proposed.

This paper is organized as follows. The proposed no-reference video quality assessment approach is described in Section 2. Experimental results are addressed in Section 3, followed by concluding remarks.

2 Proposed approach

2.1 System Architecture

As shown in Fig. 1, the proposed no-reference video quality assessment approach contains four main stages. First, three different resolutions of each input video sequence by down-sampling are generated and then spatial, temporal, transform, and spatiotemporal feature maps are extracted. Spatial, temporal, and transform feature maps are converted into feature vectors using histogram statistics, whereas principal component analysis (PCA) is employed to perform dimension reduction on spatiotemporal feature vectors. After feature vector normalization and concatenation, support vector regression (SVR) with polynomial kernel is employed to perform machine learning and video quality assessment score estimation. In this study, including the original video sequence, two corresponding video sequences of reduced resolutions by down-sampling are generated for feature extraction and video quality assessment score estimation by SVR.

2.2 Feature Extraction

To extract beneficial features from a video sequence, based on the characteristics of the human visual system (HVS), four (spatial, temporal, transform, and spatiotemporal) different types of features will be extracted. Assume that a video sequence contains \( n \) frames of size \( M \times N \), \( f_i(x, y) \), \( i = 1, 2, ..., n \), \( 1 \leq x \leq M, 1 \leq y \leq N \). Let \( f_i^1(x, y) \), \( i = 1, 2, ..., n \), \( 1 \leq x \leq M/2, 1 \leq y \leq N/2 \), denote the first down-sampled video sequence and \( f_i^2(x, y) \), \( i = 1, 2, ..., n \), \( 1 \leq x \leq M/4, 1 \leq y \leq N/4 \), denote the second down-sampled video sequence. Four different types of features, namely, spatial, temporal, transform, and spatiotemporal, are extracted, which will be described as follows.

---

+ This work was supported in part by Ministry of Science and Technology, Taiwan, Republic of China under Grant MOST 106-2221-E-194-057-MY2.
Both edges possessing strong intensity contrasts in a frame and blockiness being a type of distortion are usually attractive to human visual perception. Here, Laplacian of Gaussian (LoG) filter with zero crossing is employed to construct three LoG edge maps $L_1(x, y), L_2(x, y)$, and $L_3(x, y)$, for $f_1(x, y), f_2(x, y)$, and $f_3(x, y)$, respectively, $i = 1, 2, ..., n$. To generate six edge maps from $f_1(x, y), f_2^1(x, y)$, and $f_2^2(x, y), i = 1, 2, ..., n$, Sobel operator is employed, i.e.,

$$S_{ix} = f_1 * g_x, S_{iy} = f_1 * g_y;$$

$$S_{1x}^1 = f_1^1 * g_x, S_{1y}^1 = f_1^1 * g_y,$$

$$S_{2x}^2 = f_2^2 * g_x, S_{2y}^2 = f_2^2 * g_y,$$

where * denotes the convolution operator and

$$g_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix},$$

$$g_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}.$$

Based on entropy defined as

$$E_{ik} = - \sum p_{i,k}(x) \log_2 p_{i,k}(x),$$

where $p_{i,k}(x)$ is the probability mass function of luminance in the $k$-th block of frame $i$, six edge maps are converted into six edge entropy feature maps, namely, $E_{L,1}, E_{L,1}^1, E_{L,1}^2, E_{S,1}, E_{S,1}^1, E_{S,1}^2$, and $E_{S,1}^2$ by entropy computation.

In this study, blockiness measure proposed in [8] is employed. As shown in Fig. 2, blockiness measure of block $A$ is based on the horizontal block difference between blocks $A$ and $B$ and the vertical block difference between blocks $A$ and $C$. The horizontal block difference between blocks $A$ and $B$ is defined as

$$B_{i,k,h} = \begin{cases} \frac{N_h}{S_h} \text{if } S_h \neq 0, \\ 0 \text{if } S_h = 0 \text{ or } |b - \bar{a}| < tm(b - \bar{a}/2), \end{cases}$$

where

$$N_h = r_D \times \sum_{i=1}^{8} |a_{i1} - b_{i1}|,$$

$$S_h = \sum_{i=1}^{8} (\sum_{j=1}^{4} |b_{ij} - b_{ij}| + \sum_{j=1}^{4} |a_{ij} - a_{ij}|),$$

$$\bar{a} = (\sum_{i=1}^{8} \sum_{j=1}^{4} a_{ij})/32,$$

$$\bar{b} = (\sum_{i=1}^{8} \sum_{j=1}^{4} b_{ij})/32,$$

$$tm(X) = \begin{cases} 17 \times (1 - \sqrt{X/127}) + 3, \text{if } X \leq 127, \\ 3 \times (X - 127)/128 + 3, \text{if } X > 127, \end{cases}$$

where $r_D$, a control coefficient, is set to 10. The vertical block difference $B_{i,k,v}$ between blocks $A$ and $C$ can be calculated similarly. The blockiness measure of block $A$ (the $k$-th block of frame $i$) is defined as

$$B_{i,k} = \frac{B_{i,k,h} + B_{i,k,v}}{2}.$$

All block difference values of frame $i$ are combined as its blockiness feature map. Finally, blockiness feature maps for $f_1(x, y), f_1^1(x, y)$, and $f_1^2(x, y)$ are obtained and denoted as $B_{block,i}, B_{block,i}^1$, and $B_{block,i}^2$, respectively.

Because intensity variations in a video sequence is sensitive to human visual perception, the absolute DC coefficient difference between two adjacent frames is computed as

$$\text{Diff}_{f_1} = |DC_{i+1,k} - DC_{i,k}|,$$

where $DC_{i+1,k}$ and $DC_{i,k}$ denote DC coefficients of the $k$-th blocks of frames $f_{i+1}$ and $f_i$, respectively. Then, we can obtain three absolute DC coefficient difference feature maps, namely, $\text{Diff}_{f_1}, \text{Diff}_{f_1}^1$, and $\text{Diff}_{f_1}^2$, for $f_1(x, y), f_1^1(x, y)$, and $f_1^2(x, y)$, respectively. In addition, to track temporal variations between two adjacent frames, Kullback Leibler distance (KLD) calculating the distance between two probability distributions is employed, which is defined as

$$\text{KLD}_i = \sum p_i(u) \log(p_i(u)/p_{i+1}(u)),$$

where $\mu$ denotes the luminance and $p_i(u)$ denotes the probability distribution of luminance for frame $f_i$. Tree KLD feature maps, namely, $\text{KLD}_1, \text{KLD}_1^1$, and $\text{KLD}_1^2$ for $f_1(x, y), f_1^1(x, y)$, and $f_1^2(x, y)$, respectively, can be obtained.
Additionally, motion is an important cue in a video sequence. Here, nearest neighbor field and random sample consensus (RANSAC) algorithm are employed to generate a dominant similarity matrix $d$ for the motion correspondences between two adjacent frames. By subtracting the corresponding backgrounds of two adjacent frames, motions between two adjacent frames are obtained. By entropy computation, three motion track maps, namely, $M_1$, $M_1^1$, and $M_1^2$ for $f_1(x,y)$, $f_1^1(x,y)$, and $f_1^2(x,y)$, respectively, can be obtained.

By using DCT (discrete cosine transform), we can obtain the DCT feature map $DCT_i$ for $f_1(x,y)$. To obtain temporal variations of local phase structures [9], continuous wavelet transform with 8 orientations $\theta \in\{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}$ is employed and the average of all oriented subbands of continuous wavelet transform is calculated as continuous wavelet transform map $CWT_i$, which may be converted into continuous wavelet transform energy feature map $E_{CWT,i}$ for $f_1(x,y)$. In this study, pairwise products [10] are used to obtain relationships between neighboring feature map values along four orientations, namely, horizontal H, vertical V, diagonal ND, and off-diagonal FD. Thus, DCT pairwise feature maps along four orientations, namely, $H_{DCT,i}$, $V_{DCT,i}$, $ND_{DCT,i}$, and $FD_{DCT,i}$, and continuous wavelet transform pairwise feature maps along four orientations, namely, $H_{CWT,i}$, $V_{CWT,i}$, $ND_{CWT,i}$, and $FD_{CWT,i}$, are obtained.

To extract spatiotemporal features, based on improved dense trajectories, histogram of oriented gradients (HOG), histogram of optical flow (HOF), motion boundary histogram (MBH) [11], and 3-D (three-dimensional) spatiotemporal DCT [12] are employed. First, HOG based on gradient orientations can extract static information of local appearance and shape of a frame. Second, HOF based on orientations of optical flow can extract motion information for a video sequence. MBH can suppress camera motions in a video sequence, which can be converted into optical flow gradients. Motion boundary histograms [12] are employed. First, HOG based on gradient orientations can extract static information of local appearance and shape of a frame. Second, HOF based on orientations of optical flow can extract motion information for a video sequence. MBH can suppress camera motions in a video sequence, which can be converted into optical flow separately.

Due to inherent multidimensional statistical characteristics and excellent spatiotemporal decorrelation ability [12], 3-D DCT is utilized to extract 3-D features in spatiotemporal volumes. In this study, the averages of all 3D-DCT coefficients for spatiotemporal volumes are extracted as 3D-DCT feature vector $ThrDCT$.

### 2.3 Histogram statistics

After spatial, temporal, and transform feature maps are obtained, histogram statistics are used to convert these feature maps into feature vectors. First, each mean feature map is obtained as the average feature maps over all frames in a video sequence. Second, each mean feature map is quantified into histogram with 40 bins. Feature vectors, $E_L$, $E_L^1$, $E_L^2$, $E_S$, $E_S^1$, $E_S^2$, $B_{block}$, $B_{block}^1$, $B_{block}^2$, $diff$, $diff^1$, $diff^2$, $KLD$, $KLD^1$, $KLD^2$, $DCT$, $EG_{CWT}$, $H_{CWT}$, $V_{CWT}$, $ND_{CWT}$, $FD_{CWT}$, $H_{DCT}$, $V_{DCT}$, $ND_{DCT}$, and $FD_{DCT}$ are obtained by histogram statistics. Additionally, HOG, HOF, and MBH feature maps are quantized into histograms with 96, 96, and 192 dimensions, which are denoted as $HOG$, $HOF$, and $MBH$, respectively.

### 2.4 Dimension reduction, normalization, and concatenation

In this study, principal component analysis (PCA) is employed to perform feature vector dimension reduction and maintain feature variability as much as possible. Since feature vector values may have different ranges and distributions, a simple normalization is employed, which is realized as

$$ V' = \frac{(V - \min(V))}{(\max(V) - \min(V))}, $$

where $V'$ and $V$ denote the normalized and original feature vectors, respectively, and $\min(\cdot)$ and $\max(\cdot)$ return the minimum and maximum feature vector values, respectively. After normalization, the final feature vectors for machine training and video quality assessment score estimation are

$$ V_{con} = [E_L, E_L^1, E_L^2, E_S, E_S^1, E_S^2, B_{block}, B_{block}^1, B_{block}^2, \text{diff}, \text{diff}^1, \text{diff}^2, \text{KLD}, \text{KLD}^1, \text{KLD}^2, \text{DCT}, \text{EG}_{CWT}, \text{H}_{CWT}, \text{V}_{CWT}, \text{ND}_{CWT}, \text{FD}_{CWT}, \text{HOG}, \text{HOF}, \text{MBH}, \text{ThrDCT}]. $$

Finally, based on the final feature vectors, linear SVR support vector regression with polynomial kernel LIBSVM [13] is used to perform machine learning and video quality assessment score estimation.

### 3 Experimental results

The proposed approach is implemented on an Intel Core i7-4790K 4GHz CPU with 32GB main memory for Windows 10 64-bit platform using MATLAB 9.0 (R2016a). LIVE video quality assessment database [14] containing 150 distorted video sequences is employed, which includes four distortion categories, namely, MPEG-2 compression, H.264 compression, transmission over IP network, and transmission over wireless network. To evaluate the performance of the proposed approach, eight comparison approaches including three FR VQA approaches, one RR VQA approach, and four NR VQA approaches are employed. In terms of Pearson linear correlation coefficients (PLCC) and Spearman rank correlation coefficients (SROCC), performance comparisons of the eight comparison approaches and the proposed approach are illustrated in Table 1.

### 4 Concluding remarks

In this study, a no-reference video quality assessment approach using spatial, temporal, transform, and spatiotemporal features is proposed. Based on the experimental results obtained in this study, in terms of PLCC and SROCC, the performance of the proposed approach is better than those of eight comparison approaches.
Table 1. In terms of PLCC and SROCC, performance comparisons of the eight comparison approaches and the proposed approach.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Approaches</th>
<th>PLCC</th>
<th>SROCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>PSNR</td>
<td>0.715</td>
<td>0.694</td>
</tr>
<tr>
<td>FR</td>
<td>SSIM [1]</td>
<td>0.646</td>
<td>0.652</td>
</tr>
<tr>
<td>FR</td>
<td>MOVIE [2]</td>
<td>0.852</td>
<td>0.807</td>
</tr>
<tr>
<td>RR</td>
<td>STRRED [3]</td>
<td>0.767</td>
<td>0.813</td>
</tr>
<tr>
<td>NR</td>
<td>NVIE [4]</td>
<td>0.620</td>
<td>0.604</td>
</tr>
<tr>
<td>NR</td>
<td>Video-BLINDS</td>
<td>0.853</td>
<td>0.741</td>
</tr>
<tr>
<td>NR</td>
<td>Galea and Farrugia [6]</td>
<td>0.732</td>
<td>0.703</td>
</tr>
<tr>
<td>NR</td>
<td>Video-CORNIA</td>
<td>0.693</td>
<td>0.740</td>
</tr>
<tr>
<td>NR</td>
<td>Proposed</td>
<td>0.861</td>
<td>0.833</td>
</tr>
</tbody>
</table>

5 References


