BRIDGING THE GAP BETWEEN OBJECTIVE SCORE AND SUBJECTIVE PREFERENCE IN VIDEO QUALITY ASSESSMENT

Qianqian Xu 1, Zhipeng Wu 1, Li Su 1, Lei Qin 2,3, Shuqiang Jiang 2,3, and Qingming Huang 1,2,3

1Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
2Key Lab of Intell. Info. Process., Chinese Academy of Sciences, Beijing 100190, China
3Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
E-mail: {qqxu, zpwu, lsu, lqin, sqjiang, qmhuang }@jdl.ac.cn

ABSTRACT

Nowadays, the issue of objective video quality assessment has been extensively studied. However, the human visual system (HVS) is the ultimate receiver for videos thus leading to a gap between objective scores calculated by computers and subjective preferences given by observers. In this paper, we focus on bridging this gap by introducing a psychological criterion called contrast effect. That is, because of the impression about the quality of previous frame still remaining in observers’ minds, they tend to underestimate or overestimate the quality of the current one. Noticing this fact, we propose a video quality assessment system with an additional revision module to bridge the gap mentioned above. Firstly, the video is described by several representative clips with large entropy values. Then, we present Quality Words (including luminance, contrast, structure and spatio-temporal texture) to evaluate the quality of distorted video. To characterize the spatio-temporal texture, a new descriptor called Rotation Sensitive 3D Texture Pattern (RS-3D) is proposed. Finally, we revise the result in the revision module motivated by contrast effect. Experiments on VQEG Phase I FR-TV test dataset [1] verify the effectiveness of our method.

Keywords—Contrast effect, quality words, rotation sensitive 3D texture pattern

1. INTRODUCTION

With the rapid development and wide application of digital media devices, the number of video resources is growing at an explosive rate. The video quality assessment issue, which may bring about benefits for users in many areas (e.g. enhancement, reconstruction, compression, communication, displaying, registration, printing, watermarking, etc.) has become more and more important and drawn great attention from scientists and researchers.

The existing methods of image/video quality assessment are divided into two categories: subjective assessment and objective assessment. Subjective viewing tests are usually performed according to the standard procedures [2]. However, since the mean opinion scores (MOS) need to be obtained by a large number of observers, the subjective assessment is time-consuming and expensive. Hence, there has been an increasing demand to build intelligent objective quality measurement models.

The most commonly used Full-Reference (FR) objective image/video quality assessment measurements are mean squared error (MSE), signal to noise ratio (SNR), peak signal-to-noise ratio (PSNR), and their relatives. However, it has been well acknowledged that MSE/SNR/PSNR does not always correlate well with the HVS’ perception [3, 4, 5].
Thus, researchers resort to an alternative way which aims to construct perceptual models based on the concept of visual attention [6, 7].

Besides, a unique perspective for image similarity indexing called SSIM (Structure Similarity) [8] has been proved to be a good approximation of perceived image distortion. SSIM extracts structural information from the same spatial patches of the reference and distorted images respectively. This method has also been used in video quality assessment [9]. Although SSIM has been proved to be effective in video quality assessment, it still leaves room for improvement. Firstly, it randomly extracts local areas to represent a frame, which ignores the fact that rather than local patches the whole frame always leaves people a general impression. And whether those randomly selected areas can effectively represent the video content is still unconvincing. Besides, the features taken into consideration only include luminance, contrast and structure. Texture, which may provide important information for quality assessment, is neglected.

Generally, the previous methods (including SSIM) are all objective assessment measures. Since the ultimate receiver for videos is subjective observer, there is therefore a gap between objective scores calculated by computers and subjective preferences given by observers. We believe that it is better to add psychological knowledge into the process of quality assessment, which could make the result more user-oriented.

Motivated by the above analysis, we present an improved video quality assessment method which aims to bridge the gap mentioned above. Fig. 1 illustrates the diagram of the proposed system. The main contributions of this paper are:

1. Rather than individual frames, video clip, which is a collection of sequential frames, always provides additional information (e.g. motion). In our approach, the representative units of the video are several representative clips with large entropy values.

2. Video is treated as a spatio-temporal volume. Thus, a new kind of descriptor called Rotation Sensitive 3D Texture Pattern (RS-3D) is proposed to characterize the spatio-temporal texture of the video.

3. A psychological phenomenon called contrast effect is introduced to add a revision step to bridge the gap.

The rest of the paper is organized as follows. In section 2, we introduce our video quality assessment framework and the revision scheme based on contrast effect. Section 3 presents the experimental results. Finally, the paper is concluded in section 4.

2. THE PROPOSED FRAMEWORK

2.1. Video quality assessment framework

As illustrated in Fig. 1, video quality is measured in four levels: block level, frame level, clip level and video level.

2.1.1 Representative clips

We first segment the reference video into \( N \) non-overlapping clips by a shot boundary detection scheme [10]. Then we select representative video clips based on the content information calculated by entropy value. Assuming that all pixels in a clip are independent to each other, the entropy is calculated by the normalized grayscale histogram \( p \) of the whole clip:

\[
E = -\sum_{i=0}^{bin} p(i) \log p(i)
\]

where \( bin \) is the number of gray levels (histogram bins). We choose the top \( k \) entropy value clips as the processing units (as shown in Fig. 1). To implement FR video quality assessment, representative clips are also extracted from the corresponding spatio-temporal locations of the distorted video sequences (reference and distorted video have already been aligned in the dataset). Intuitively, the clip selection step is mainly focusing on reducing the computational cost while still maintaining the richness of the video content.

Besides, the final video quality score is a sum of weighted clip quality scores. It has been proved that after rapid scene change or large temporal difference, the sensitivity of the HVS to spatial details is lowered [11]. Therefore, we set small weights to the clips with large motion values (estimated by optical flow [12]) which seem to be weakly perceived by the subjective observers.

2.1.2 Quality words

We evaluate the video quality by comparing the perception similarities from the clips in the distorted video to their origins in the reference. Because clip is a collection of sequential frames, frame-level quality assessment is the prerequisite. Firstly, image blocks are extracted at the same locations from frames being compared. Then we propose Quality Words (QW), which contain luminance, contrast, structure, and spatio-temporal texture to describe the block quality. The luminance, contrast, and structure are defined as follows [9]:

\[
l(\text{Block}_i, \text{Block}_j) = \frac{2 \mu_i \mu_j}{\mu_i^2 + \mu_j^2}
\]

\[
c(\text{Block}_i, \text{Block}_j) = \frac{2 \sigma_i \sigma_j}{\sigma_i^2 + \sigma_j^2}
\]

\[
s(\text{Block}_i, \text{Block}_j) = \frac{\sigma_{i,j}^2}{\sigma_i \sigma_j}
\]

where \( \mu_i, \mu_j, \sigma_i^2, \sigma_j^2 \) and \( \sigma_{i,j} \) are the mean of \( \text{Block}_i \), the mean of \( \text{Block}_j \), the variance of \( \text{Block}_i \), the variance of \( \text{Block}_j \), and the covariance of \( \text{Block}_i \) and \( \text{Block}_j \).

Taking a further step beyond the features in [9], video is always treated as a spatio-temporal volume. Therefore, there is a demand of spatio-temporal texture descriptor
which combines motion and visual appearance for video quality assessment.

In [13], Ojala et al. found that certain Local Binary Patterns (LBP, obtained by binarization of the neighbor pixels with the value of the center pixel) called “uniform patterns” are fundamental properties of local image texture. It is reported that these patterns can cast a vast majority, sometimes over 90% of the local texture situations. As shown in Fig. 2 (a), a pattern is considered uniform if it contains no more than two spatial transitions (bitwise 0/1 changes) when counting the bit patterns circularly. For example, pattern 00000000\( _2 \) and 11111111\( _2 \) have transition value of 0, while 00111111\( _2 \) and 11000000\( _2 \) have transition value of 2. This operator has the advantage of rotation invariant, thus is suitable for rotation invariant texture analysis. However, since the HVS is sensitive to rotation, uniform pattern may lose its discriminative power in video quality assessment. Take uniform pattern 1 for example. As shown in Fig. 2 (b), although all the 8 patterns belong to the same uniform pattern 1, the HVS’ reactions for them are totally different. In other words, we need to distinguish them and bring a rotation sensitive texture descriptor in video quality assessment.

![Fig. 2. (a) The 9 uniform patterns in the circularly symmetric neighbor set \((N = 8)\). (b) The 8 patterns belonging to the uniform pattern 1.](image)

In this paper, a novel descriptor called Rotation Sensitive 3D Texture Pattern (RS-3D) is designed to characterize the spatio-temporal texture. Instead of an image, our descriptor operates on a spatio-temporal volume, which is the combination of three orthogonal planes (XY, XT and YT).

According to Tab. 1, we calculate the RS-3D histogram of an input block. Then, we use \( \chi^2 \) distance to measure the distance between histogram \( RSH_i \) and \( RSH_j \). Fig. 3 illustrates the process of spatio-temporal texture measuring between two blocks:

\[
\chi^2(\text{Block}_i, \text{Block}_j) = \exp\{-\chi^2(RSH_i, RSH_j)\} \quad (5)
\]

Based on the above analysis, by incorporating luminance, contrast, structure and spatio-temporal texture, both the appearance and motion information are taken into account for video quality assessment. The QW between block \( B_i \) and block \( B_j \) is defined as:

\[
QW(B_i, B_j) = l(B_i, B_j) \times c(B_i, B_j) \times s(B_i, B_j) \times t(B_i, B_j) \quad (6)
\]

where \( l, c, s, t \) stands for luminance, contrast, structural and RS-3D texture similarity, respectively.

### 2.2. Contrast effect

The inspiration of our work comes from a psychological discovery which is called contrast effect. It mainly includes simultaneous contrast and successive contrast [14]. First, let us take a look at Fig. 4. Which ball in the center is bigger? Most observers may feel that the red ball on the left is bigger than the one on the right. However, they are of the same size. This is a classic example of simultaneous contrast, which results from the different sizes of the surrounding blue balls. Besides, [15] found that when individuals first lift a heavy weight, they underestimate the weight of lighter weights they are subsequently asked to lift.
This is a good proof of successive contrast. Another simple way of illustrating this concept is to put one hand into hot water and the other into cold water, then move both of them to lukewarm water. The cold hand will feel hot and the hot hand will feel cold.

Similarly, when human observers assess the quality of video/image sequence, because of the impression of the previous frame/image shown to them, do they tend to underestimate or overestimate the quality of the current one? To demonstrate this discovery, we design 3 tests.

First of all, although we believe that all the frames in a specific video stand on the same quality level, there still exist obvious quality differences among individual frames. In other words, when observers are giving subjective scores to the current frame, they may be influenced by the contrast effect phenomenon. Test #1 simply proves the existence of this problem.

**Test #1**

**Data:** VQEG Phase I FR-TV test set [1] 525_src13-src22.  
**Purpose:** Verify that the qualities of individual frames in the same video are different.  
**Method:** Compute SSIM [8] between reference and distorted video for all the frames. Then find the maximum and minimum frame quality scores in each video.

**Fig. 5. Results of test #1.**

From Fig. 5 shown above, we notice that quality differences do exist among individual frames in the same video. In other words, video can be treated as a sequence of frames with different qualities. Subsequently, test #2 will demonstrate that when observers assess the video quality, they can be influenced by the adjacent quality differences. That is to say, contrast effect can make an impact on their judgment.

**Test #2**

**Data:** 400 high quality images (HQ) \((H_1, H_2, H_3, ..., H_{400})\) from www.dpchallenge.com. 400 low quality images (LQ) \((L_1, L_2, L_3, ..., L_{400})\) generated from HQ.  
- Form \(L_1, H_1, L_2, H_2, L_3, ..., L_{399}, H_{400}\) into image sequence 1.  
- Form \(L_2, L_3, L_4, H_3, L_5, L_6, ..., L_{400}\) into image sequence 2.  
**Purpose:** Prove that observers can be influenced by the contrast effect.  
**Method:** Compare the subjective scores of sequence 1 and sequence 2 for all the images at odd positions.

In test #2, the images at odd positions in sequence 1 and sequence 2 are the same. However, the quality scores given by observers are different (as shown in Fig. 6). It is obvious that affected by the high quality impression of the previous image, observers tend to underestimate the quality of the current one. After verifying the existence of contrast effect in the process of subjective video quality assessment, we expect to add a revision step for the scores given by computer to make them more user-oriented.

Supposing that we have a sequence of objective frame scores given by the computer \((...O_i, O_{i+1}, ...)\), our target is...
We have to bridge the gap between $O$ and $S$:

$$\text{Gap} = G_i = S_i - O_i$$

According to the above analysis, the gap between $O$ and $S$ is related to the quality of the previous image $i-1$. Without loss of generality, we use the objective score $O_{i-1}$ to estimate the quality of the previous image. Then, the target equation can be written as:

$$G_i = F(O_{i-1})$$

(7)

$$S_i = F(O_{i-1}) + O_i$$

(8)

Test #3 is designed to find the modified function $F$ in the above equation.

**Test #3**

**Data:** To implement FR video quality assessment, sequence 1 in test #2 is treated as reference sequence $R$ ($R_{1}, R_{2}, R_{3}, ..., R_{440}$). Several distortions (e.g. noise interference, low pass filtering, lossy compression) are operated on sequence $R$, resulting in the distorted sequence $D$ ($D_{1}, D_{2}, D_{3}, ..., D_{440}$).

**Purpose:** Obtain the modified function $F$.

**Method:** We calculate the objective score sequence ($O_{1}, O_{2}, O_{3}, ..., O_{440}$) by QW for all the images in sequence $D$. Then, 5 individuals are invited to grade the 400 images in $D$ (80 images per individual), providing the subjective score sequence ($S_{1}, S_{2}, S_{3}, ..., S_{440}$). We use the objective-subjective gap $S_i - O_i$, which indicates the adjustment value as y-axis; use $O_{i-1}$ which indicates the objective score of the previous image as x-axis, and obtain the modified function $F$ by curve fitting.

3. EXPERIMENTAL RESULTS

In this section, we demonstrate the effectiveness of our method using the most publicly-accessible database for video quality assessment, VQEG Phase I FR-TV test dataset [1], which includes 20 different reference videos and 320 distorted videos. Performance of our approach is evaluated following the procedures employed in the VQEG Phase I FR-TV test [16]. To facilitate comparison of the models in a common analysis space, a nonlinear regression between the model’s objective scores and the corresponding DMOS (Difference Mean Opinion Score) values is necessary to estimate. In [16], four metrics are adopted to measure the model’s performance:

Metric 1 ($M_1$): The correlation coefficient between objective/subjective scores after variance-weighted regression analysis.

Metric 2 ($M_2$): The correlation coefficient between objective/subjective scores after nonlinear regression analysis.

Metric 3 ($M_3$): The Spearman rank-order correlation coefficient between the objective/subjective scores.

Metric 4 ($M_4$): The outlier ratio (percentage of the number of predictions outside the range of ±2 times of the standard deviations) of the predictions after the nonlinear mapping.

Here, $M_1$ and $M_2$ provide an evaluation of prediction accuracy, while $M_3$ and $M_4$ are employed to measure prediction monotonicity and prediction consistency, respectively.
Fig. 8 shows the scatter plots of DMOS (y-axis) versus objective scores (x-axis) on all test video sequences provided by VQEG Phase I FR-TV test dataset [1].

In this paper, a new objective video quality assessment scheme is presented. The highlight of the proposed method is the use of psychological knowledge: contrast effect, which effectively bridges the gap between objective scores and subjective preferences. Besides, to provide fast solution for quality assessment system, we select representative clips with large entropy values in the video. Moreover, a new descriptor called RS-3D is designed to characterize the spatio-temporal texture of the video. Experiments on VQEG Phase I FR-TV dataset [1] verify the effectiveness of our method. Future work will be deployed on large dataset. Also, new perspective for video quality assessment will be studied.

5. REFERENCES


4. CONCLUSION

Table 2. Performance comparison of video quality assessment models [16] on VQEG Phase I FR-TV test dataset [1].

<table>
<thead>
<tr>
<th>Model</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0 (PSNR)</td>
<td>0.804</td>
<td>0.779</td>
<td>0.786</td>
<td>0.678</td>
</tr>
<tr>
<td>P1 (CPqD)</td>
<td>0.777</td>
<td>0.794</td>
<td>0.781</td>
<td>0.650</td>
</tr>
<tr>
<td>P2 (Tektronix/Sarnoff)</td>
<td>0.792</td>
<td>0.805</td>
<td>0.792</td>
<td>0.656</td>
</tr>
<tr>
<td>P3 (NHK/Mitsubishi)</td>
<td>0.726</td>
<td>0.751</td>
<td>0.718</td>
<td>0.725</td>
</tr>
<tr>
<td>P4 (KDD)</td>
<td>0.622</td>
<td>0.624</td>
<td>0.645</td>
<td>0.703</td>
</tr>
<tr>
<td>P5 (EPFL)</td>
<td>0.778</td>
<td>0.777</td>
<td>0.784</td>
<td>0.611</td>
</tr>
<tr>
<td>P6 (TAPESTRIES)</td>
<td>0.277</td>
<td>0.310</td>
<td>0.248</td>
<td>0.844</td>
</tr>
<tr>
<td>P7 (NASA)</td>
<td>0.792</td>
<td>0.770</td>
<td>0.786</td>
<td>0.636</td>
</tr>
<tr>
<td>P8 (KPN/Swisscom CT)</td>
<td>0.845</td>
<td>0.827</td>
<td>0.803</td>
<td>0.578</td>
</tr>
<tr>
<td>P9 (NTIA)</td>
<td>0.781</td>
<td>0.782</td>
<td>0.775</td>
<td>0.711</td>
</tr>
<tr>
<td>SSIM (Without adjustment)</td>
<td>0.830</td>
<td>0.820</td>
<td>0.788</td>
<td>0.597</td>
</tr>
<tr>
<td>Proposed (RS-3D, Unrevised)</td>
<td>0.852</td>
<td>0.837</td>
<td>0.787</td>
<td>0.587</td>
</tr>
<tr>
<td>SSIM (With adjustment)</td>
<td>0.864</td>
<td>0.849</td>
<td>0.812</td>
<td>0.578</td>
</tr>
<tr>
<td>Proposed (Revised)</td>
<td>0.865</td>
<td>0.866</td>
<td>0.851</td>
<td>0.588</td>
</tr>
</tbody>
</table>