Real-time Video Quality Assessment Methods for Video Conference System

Xinchen Zhang, Yuan Fang, Hao Jiang, Yin Wang and Lihua Wu

Abstract—After a deep investigation on the performance measure of the video conference and the video quality assessment methods, two type plans of video quality assessment for video conference system are proposed. If we can input the sequence data to the video conference, the FR VQA method can be used to assess the video quality as human doing. To solve the problem of which frame is the reference one, the special block marks were added to the original sequences. Without reference, an new NR VQA method is introduced, which is the composite assessment based on coding layer and packet layer information. The experimental results show that two methods can obtain the dependable score which close to the subjective assessment.

Index Terms—Video quality assessment (VQA), Full reference, No reference, Subjective assessment, Objective quality assessment, Video conference.

I. INTRODUCTION

It is important to measure the video quality accurately in real-time applications such as video conference, video on demand, during the growing digital video applications. Video conference on IP network is different to other video communication applications, so there is a urgent need for researching on special video quality assessment methods in order to help to evaluate the performance of the given video conference system.

In the past years, there are many research results about how to measure the capacity performance of video conference. These introduced us that the schemes and implements of measuring the capacity of MCU in video conference system and the detecting design basements are aims at the logic topology of the control units connection, network transmission bandwide and the conference devices inter-connection in different levels. Those are the key parameters of the video conference system, but they can not replace the feeling of the user who use the conference system. Now we want to find an effective way to test the video quality which the user look on.

These past achievements are helpful to measure the performance of the important component equipment of the video conference and also are useful to improve and guarantee the quality of the video in the end-device if we find the video quality had been dropped down.

In practice application, the received video may be a degraded one of the original sequence because we can not believe there is not error in the delivered video data to the users. Therefore, the service provider might be interested in monitoring the actual video quality at the end-user terminal. With the development of the video technology, the need to discover ways of video quality assessment (VQA) has led to several solutions of this question. Generally, VQA methods can be classified into subjective methods and objective methods [1]. Paper [2] discussed the conventional objective methods for VQA. The objective VQA can be classified into three categories, such as full reference (FR) methods, Reduced reference (RR) and no reference (NR). There are several subjective VQA methods, such as Double Stimulus Impairment Scale (DSIS), Double Stimulus Continuous Quality Scale (DSCQS), Single Stimulus Methods (SSM), and Single Stimulus Continuous Quality Evaluation (SSCQE) [3]-[4].

Obviously, the subjective VQA method is too inconvenient and too expensive to be conducted in real-time applications. But it’s results can be as the standard to modify the evaluation value of the objective methods. When the video conference is running officially, because of the lack of corresponding reference video, the NR VQA method is the only way to suit this solution. However NR VQA is relatively difficult because the limited understanding of the Human Visual System (HVS) and the corresponding cognitive aspects of the brain [3]. The main contribution of this paper is to resolve two key questions in VQA for video conference. The first question is how to use FR objective VQA to evaluate the quality and to adjust it close to the subjective result as user feeling if we can input special video to one video terminal in conference. The second question is how to get an available result by using NF VQA in video conference.

This paper is organized as follows. In section 2, the related RF VQA method and the novel frame-mark way are given. In section 3, the proposed NF VQA method based on video bit-stream analysis is presented. The experimental results of the two methods proposed in this paper are also presented in section 4. Finally, our work of this paper and future work are summarized in the last section.

Manuscript received June 19, 2012. This work was supported by the self-determined research funds of CCNU from the colleges’ basic research and operation of MOE.

Xinchen Zhang is with the School of Physical Science and Technology, Central China Normal University, Wuhan, 430079 China (phone: 86-13971488777; e-mail: zhangxc@phy.ccnu.edu.cn).
Yuan Fang, Hao Jiang, Yin Wang, Lihua Wu are with Wuhan University, Wuhan, 430072 China. (e-mail: zxc9501@hotmail.com).
II. FR VQA Method for Video Conference

When video conference is working, we could use a device to send the original video sequence as the camera to one video conference terminal. Because we can use the original video as the full reference source, the captured videos form different test points were computed the objective distortion to the original by Mean Squared Error (MSE), Peak Signal-to-Noise Ration (PSNR) and video Structural Similarity Metric (SSIM), and so on. The different test points give us the distortion of the video processing in the conference which caused by different devices and reasons.

As the figure 1 shows, the capture point 1 can measure the distortion which is caused by the conference terminal encoder, and the point 2 is the distortion caused by encoder, transmission and decoder. So the first distortion means the conference encoder performance, and the second distortion means the delivered video quality.

In the video conference application environment as the other real time video communication system, it is not suit to use FR VQA methods. If we can overcome the problem of no original reference video, there is another problem that we do not know that the present received video frame is which frame in original video sequence, as the figure 2 shows. After we design a way to decide which number is the frame of the original video to the encoder, the distortion between the received and decodered frame and encoder frame could be computed by using different computation formulas. In FR VQA method for the video conference, the last question is how to modify the objective result close to the subjective one, in order to evaluate the watching video quality according to the human visual and to replace the user feeling.

At the present time, the mark way of the frame number is based on the semantic method in the video compress standard, which can guide the frame in the encoder and decoder. But because of variable bitrate and bandwide transmission channel, the encoder always used variable frame rate and bitrate algorithm to fit the channel wave. Those make the original frame’s order is different to the encoder frame’s order, so we can not using the order in decoded video as the original sequence. Then we design a method to add additional content in the original video. Considering not to change the original video content and easy to judge the additional mark, one novel special mark-added block method in the luminance to replace the serial number of the video is designed and proposed.

Figure 3 presents the method of adding special block mark in the luminance. We change the 16x16 block values of the luminance component of the original video in the space order to the 4 possible values. For example, we want to indicate 256 number frames by changing 4 blocks' values. The 4 possible values we can select are as 0, 85, 170, 255. Then 4 blocks and 4 possible values combinate 256 number, as $4^4 = 256$. Through encoding, transmitting and decoding, the value of those 4 block must be changed by adding encoding error and transmission error. But they would not be changed to exceed the boundary for judgement, because of the magnanimous range to distinguish the different values. In order to compute the value in binary system conveniently, 00 is equivalent to 0 in the luminance, 01 is to 85, 10 is to 170, and 11 is to 255. In instance, the order number 60 in decimal system can be converted to 00111100 in binary system. The 4 added-mark blocks’ values are (0, 255, 255, 0). Through coding, transfering and decoding, we can compute the frame order number by judging the average value of the added-mark blocks belong to which region. The computing rule is that if $A = \sum_{i=0}^{15} \sum_{j=0}^{15} x_{ij}/256$, $i=0,1,2,\cdots,15$, $j=0,1,2,\cdots,15$, $A \in [0,42]$, then the mark of the block $16 \times 16$ is 00. In the same breath, if $A \in [43,126]$, the mark is 01; if $A \in [127,212]$, the mark is 10; if $A \in [213,255]$, the mark is 11. In
order to prove the availability of the novel method, we input the video adding the marks to the encoder with extreme QP and transmission parameters setting, it also can promote correct the order of the video frame. We had got the patent of this method in China.

VQEG report [6] provides us a good experimental series of video source, and also gives us the subjective score of DMOS regardless of which human measure. The differential mean opinion score (DMOS) scores are plotted for each of the distorted video sequences. Because now in video conference we uses FR VQA method, the DMOS score represents the perceived quality of the video between the original and the assessed video. In the later part of the paper, we will use Mean Opinion Score (MOS) to represent the perceived quality of the video without reference video. Paper [5] recommended 8 objective VQA algorithms. Considering the accuracy and the complexity, we select the MS-SSIM [6] as the VQA method, because the MS-SSIM index performs better (relative to human opinion) than the SS-SSIM index on images [7]. The scatter plots for MS-SSIM algorithms need to focus to one regressed function. From the past experimental results, we use non-linear regression as prescribed in [6] and [7]. The 4-parameter logistic function is represented as the formula (1).

$$Q_j = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-\beta_3[Q_j]}}$$ (1)

When we confirm the 4 parameters of the formula, we can convert objective score to subjective score which is video quality assessment as the user feeling. In section 4, this paper will continue to discuss the residual details.

III. NR VQA Method for Video Conference

Actually, it is not suitable to assess the terminal video quality by using FR or RR, because of the lack of corresponding reference video when the video conference is running only with camera or other picture input devices. In order to solve the VQA requirement in this environment, the No-Reference VQA method must be discussed. As the user feeling, they consume the video conference to concern the real video quality they watch, without watching the same content video in different quality two times. So for user, they only consider the playing video in terminal do or not make them dislike it. We propose one method based on NR VQA using bitstream analysis, video frame’s temporal relationship and still image feature to obtain the score as the human subjective feeling.

The measuring processing of the method is as figure 4 shows. The distortion of the delivered video is mainly caused by video coding, network transmission and the device which captures the picture. The first distortion can make blocking artifacts and blur to the HVS which is determined by the bits per pixel and the video compression standard. The second distortion can make the quality degradation and stillness because of the transmission error and packet loss. The third distortion is related to the quality of the input video, as the any digital capture device leads error in the AD processing. As everyone know, good capture device makes the picture pleasing to both the eye and the mind. In our method, we can not appraise the artness value of the video or picture, but we can induce the special features of the still picture and define the appropriate bounds to represent the good picture for human normally. If we can evaluate the main three distortions, it is easy to establish the compositive assessment model to get the final score.

The proposed method is based on analysis on the received packet, video bit stream data and the primary picture features. Paper [7] provide us an available way to compute the primary features of the still image which is our prevenient research. We can choose the lightness, chroma, definition, contrast, and histogram as the primary features to define the content of one still image which is or not adapt to human’s eye. Because like or dislike is related to the watcher’s mind, we can not consider that this image is better than that image only by simply computation. We only define the ranges of the feature parameter to represent the favorable image. If the image’s feature parameters exceed the range, which means that something make user uncomfortable, we can modify the final score based on the effect on the user feeling from the previous experiments.

Our VQA method is for the IP video conference. Then the input data is IP packet data. Firstly, it is necessary to parse the arrival time of the video payload packet and the bit rate of the delivery video, which can be used to determine the video playing fluency and playback detail quality. Also the channel bit error and packet loss is another improtant factor to effect the playing quality. Secondly, except the transmission factors, the video encoding information can be analysed from the video header. The current frame is which type in the encoder and the current QP is used at the marcoblock or slice in the picture. To study the realitionship between the QP and the coding distortion, the experimental results show the score of the MOS and the QP have an approximate linear relationship with variations in slopes for different video requance4. Therefore, we can obtain the assessment score by using the QP parameter which is sought from the bit stream. I would like to emphasize that QP affect the quality assessment is different in the different coding frame type, we need to establish the model respectively. Paper [8] used a database for calculating and adjusting the frame quality to the SSCQE scores. In our method, the computation sketch map of the video coding information and the packet analysis is as the figure 5 shows.

![Fig. 4. The processing procedure of the proposed VQA.](image-url)
The estimation models include two parts: the no-error model and the packet loss model. To the packet loss channel, there are many researches on the H.264 in loss channel. Because the packet loss is infrequently occurred in video conference network, we only mention this question but not resolve it in practice.

IV. EXPERIMENTAL RESULTS

In this section, there are three experiments about the frame order number, converting the FR objective result to subjective score and No-Reference VQA score.

In the first experiment, to evaluate the availability of the proposed block-marked method, we use the video test sequences [9] (CIF 352x288, YUV 4:2:0, 30fps) as the original video. We used ZXV10 T600 as the video conference MCU and one terminal, other terminals were personal computers. The PC terminals run our video conference software which can support the H.323 and SIP protocol standards. We used H.264 codec as the video compression standard at the terminal, also T600 can provide H.264 codec. At first, we add the special block mark to the video sequence in the first 150 frames. Then we encode the mark-added video aiming to the lower bitrate as in the bad transferring environment. Figure 6 shows us the mark-added video and Figure 7 shows us the received video. We can find that the special feature of the added blocks were saved in the received terminal, which can be extracted the order number.

After we got the decoded video picture in the terminal and find the right order number in the reference sequence, it is convenient to compute the objective quality parameter in the FR method. In our experiments, the MS-SSIM is chose as the objective parameter, the reasons was given in the section 2. We used the test video sequences down from VQEG and LIVE database [10]. We selected 10 video sequences named “Blue Sky”, “River Bed”, “Pedestrian area”, “Tractor”, “Sunflower”, “Rush hour”, “Station”, “Park run”, “Shields”, “Mobile and Calendar” in the Table I. The format is 768x432, YUV 4:2:0. Above the results of previous studies, we selected 4 parameters of the formula in the section 2 as: \( \beta_1 = -1563.646 \), \( \beta_2 = 68.847 \), \( \beta_3 = 112.587 \), \( \beta_4 = 3.427 \). Every original video was processed by H.264 encoder and the different packet loss rates. There were 15 new videos form one sequence to delegate the
different quality video by the different reasons. Figure 8 shows the results of the the “best-fit” regressed curve.

For performance evaluation suggested by the VQEG, we used three metrics to evaluate the performance of the proposed method. The root-mean-squared error (RMSE), the Pearson correlation coefficient (PCC), the spearman rank ordered correlation coefficient (SROCC) were computed between the fitted objective data and the corresponding subjective data.

Table II shows the performance of this FR video quality assessment method.

In the NR VQA method for video conference, in order to confirm the relationship between the bitrate and video quality, we used two models to establish the relationship. One model named “T-V” [11] was described as the formula (2), which was used to study the Mpeg4 video quality.

\[
I_e = (PQ_H - PQ_L)(1 - e^{-a(b-Br_e)}) + PQ_L - 1
\] (2)

Another using the next was recommended in G.1070 [12] by ITU-T, as the formula (3) shows.

\[
I_e = v_i(1 \cdot \frac{1}{1 + (\frac{b}{v_e})^{n}})
\] (3)

In the two formulas, \(b\) represents bitrate, and the other parameters can be fixed by the experiments results.

In T-V model, we use \(a_1\), \(a_2\), \(a_3\) to build the relationship of the 4 parameters as the formula (4) shows.

\[
a_1 = (PQ_H - PQ_L)e^{aBR_e} \\
a_2 = \alpha \\
a_3 = PQ_H - 1
\] (4)

Using “Blue Sky” videos in the different bitrates, we recommend the value of the parameters as Table 3 shows. The results by Matlab tools simulation were in the Figure 9 and Figure 10.

### Table I

<table>
<thead>
<tr>
<th>Name</th>
<th>Time</th>
<th>Frame Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Sky</td>
<td>8.68s</td>
<td>217</td>
</tr>
<tr>
<td>River Bed</td>
<td>10s</td>
<td>250</td>
</tr>
<tr>
<td>Pedestrian area</td>
<td>10s</td>
<td>250</td>
</tr>
<tr>
<td>Tractor</td>
<td>10s</td>
<td>250</td>
</tr>
<tr>
<td>Sunflower</td>
<td>10s</td>
<td>250</td>
</tr>
<tr>
<td>Rush hour</td>
<td>10s</td>
<td>250</td>
</tr>
<tr>
<td>Station</td>
<td>10s</td>
<td>250</td>
</tr>
<tr>
<td>Park run</td>
<td>10s</td>
<td>500</td>
</tr>
<tr>
<td>Shields</td>
<td>10s</td>
<td>500</td>
</tr>
<tr>
<td>Mobile&amp;Calendar</td>
<td>10s</td>
<td>500</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Model</th>
<th>PCC</th>
<th>SROCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-SSIM</td>
<td>0.8116</td>
<td>0.7890</td>
<td>6.519</td>
</tr>
</tbody>
</table>

Table II shows the performance of this FR video quality assessment method.

V. CONCLUSIONS

This research has been motivated at certain positions in the video conference system, such as the conference terminal. We created an new method to mark the order of the video sequence in each frame, and make the FR VQA to fit the real-time measure. By using FR VQA, we can calculate the accurate objective quality of the delivered video. Above the previous studied, we also give a viable way to convert the objective result to subjective DMOS score. For some condition we can not obtain the reference video, the NR VQA methos was also studied in the paper. We proposed the compositive assessment method based on the received bit stream analysis with using still image features, transmission performance and video encoding information. To establish the relationship between the video quality and the bit rate, we use the two models to regress the curve for the different video format and the frame coding type. Finally, according to the still image features, we can modify the subjective score by satisfying or not the appropriate range which was defined by human eye. The experimental results proved that the proposed methods are satisfiable and it produces the score...
close to those of the subjective assessment. For video conference, we give the resolution to the practical applications with reference video or without it. In the future, we want to study the NR VQA model in the packet loss condition.

ACKNOWLEDGMENT

The authors would like to thank Prof. Kun-Mean HOU in the LIOMS (Laboratory of Informatics, Modeling and Optimization of Systems, University Blasise Pascal, France) for his support and encouragement.

REFERENCES


Xinchen Zhang (M’11) became a Member of IEEE in 2011. He was born in Wuhan, China in 1977. He graduated from Wuhan University with a B.Sc. Degree in Eelectronic and Information System and M.Eng. Degree in Circuit and System in 1999 and 2001. He earned a Ph.D in Photographic Measuremene and Remote Sense in 2006. Now he is Associate Professor, Pyscical Science Technology School, Central China Normal University. He is awarded China Scholarship Council Sponsorship to visit Lioms, ISIMA, UBP, Clermont-ferrand, France in 2011-2012.


Lihua Wu was born in Hubei, China in 1987. He graduated from Wuhan University with B.Eng. Degree in Comm. Engineering in 2009. He is a Ph.D student in School of Electronic Information, Wuhan University.

Yuan Fang was born in Hubei, China in 1988. She graduated from CCNU with B.Eng. Degree in Information Technology in 2010. Now She is a Master student in School of Electronic Information, Wuhan University.