SCENE-AWARE SOCCER VIDEO QOE ASSESSMENT - A COMPRESSED-DOMAIN APPROACH

Fan Li[1], *Yixin Mei*[1], *Ziyi Liu*[1], *Pamela Cosman*[2]

1. Xi'an Jiaotong University, Xi'an 710049, China
2. University of California, San Diego, California 92093-0407, USA
lifan@mail.xjtu.edu.cn; {xamichelle, lzy1016}@stu.xjtu.edu.cn; pcosman@eng.ucsd.edu

ABSTRACT

The small screen of mobile devices and bandwidth limitations of communication networks greatly affect users' quality of experience (QoE), especially for soccer video, which is characterized by rapid movement and small objects. In this paper, a Compressed-domain Soccer Video Quality assessment Model (CSVQM) is proposed based on the fact that soccer video includes three distinct scene types, which cause different concerns for viewers. To reduce complexity and operate in real-time, all model parameters are derived from the compressed video stream without resorting to complete video decoding. The validation shows that CSVQM significantly outperforms conventional models in terms of accuracy, consistency, and complexity.

Index Terms— Mobile video, QoE, modeling, soccer video, scene-aware, compressed-domain

1. INTRODUCTION

In mobile video, small viewing size and transmission bandwidth limitations significantly affect a user's quality of experience (QoE), which is particularly important for fast-moving sports. Soccer video has rapid movement and small objects, which often cause unpleasant viewing experiences. However, due to limited network resources, it is necessary to optimize video quality based on user requirements and network conditions. Therefore, quality assessment for soccer video is indispensable to improving users' QoE.

Full reference metrics, such as PSNR, SSIM [1], and MS-SSIM [2] are widely used as QoE metrics, and rely on full access to the original video. No reference metrics operate without any reference video. Pixel-domain no reference metrics use information from reconstructed video to estimate users' QoE at the receiving terminal [3]. However, pixel-domain metrics are unable to meet the requirements for low complexity and real-time operation. For networked video applications, compressed-domain no reference metrics are often preferred as a low complexity, real-time solution [4] , [5] to predict the video quality and optimize end-to-end transmission.

Many video quality assessment metrics consider video content characteristics and content types, such as news, sports, animation, music, comedies and movies [6], [7]. Soccer, like many other sports videos, has distinctive typical scene types. However, it is different from many non-sports video [8], and common quality assessment methods cannot predict soccer video quality accurately because they ignore the influence of different scene types.

In this paper, a Compressed-domain Soccer Video Quality assessment Model (CSVQM) is proposed. The main contributions of this paper can be summarized as follows: 1) The different reactions of audiences to close-up, medium shots, and long shots, are considered in our model to precisely estimate QoE. 2) Our QoE model is a compressed-domain method. The model factors are extracted and estimated directly from the compressed video streams. Compared with pixel-based methods which must fully decode the video, the complexity of our proposed model is substantially reduced. As a result, our QoE model is applicable to transmission applications with stringent real-time requirements.

The rest of the paper is structured as follows. Section 2 introduces the testing methodology. Section 3 presents a compressed-domain soccer video quality assessment model. Section 4 provides validation of the proposed model, and conclusions are drawn in Section 5.

2. TESTING METHODOLOGY

2.1. Test tool and test videos

The Samsung A5 is used as test equipment to represent a general Android mobile phone. The size of the screen is 5 inches, and the display resolution is 1280×720.

We selected 24 high resolution and high bitrate soccer videos of length 8-10 seconds as sources from close-up, medium shot and long shot soccer videos; each scene type includes eight videos. Fifteen of the 24 (five of each scene type) were used to establish the QoE model, as shown in Fig-

This research work was supported in part by National Science Foundation of China (61671365), Natural Science Basic Research Plan in Shaanxi Province of China (2017JM6018), and the Fundamental Research Funds for the Central Universities.

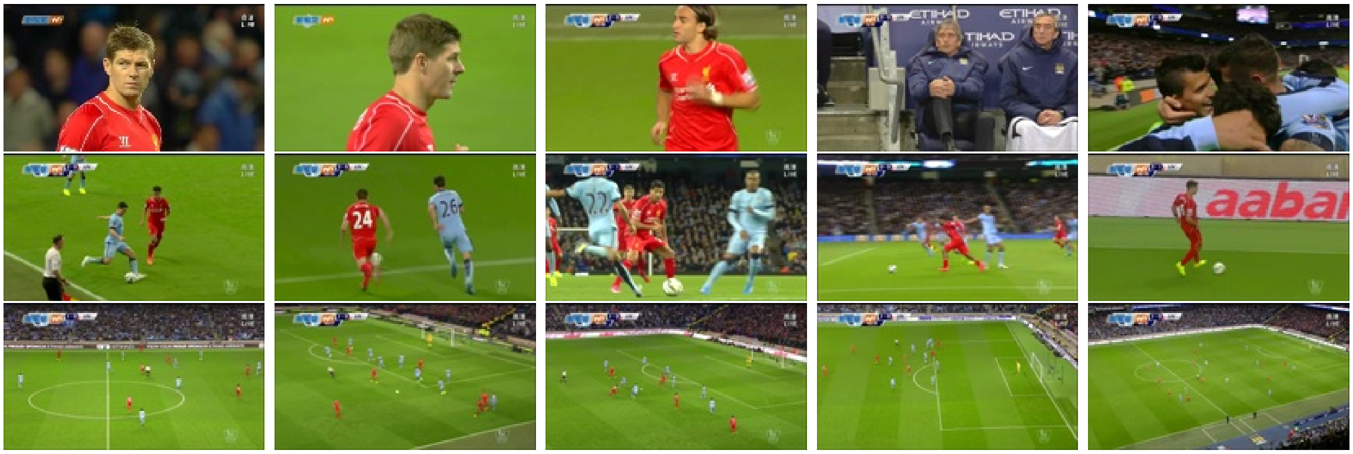

Fig. 1. Soccer video contents. The first group: Close-up soccer videos (Close1 to Close5). The second group: Medium shot videos (Medium1 to Medium5). The last group: Long shot videos (Long1 to Long5).

ure 1. The other nine (three of each scene type) were used for validation.

To produce the test videos, high-resolution (1280×720) videos were converted into low-resolution (848×480, 576×320) videos with the same aspect ratio. Afterwards, the videos were encoded at three spatial resolutions (SR) and seven different bitrates (BR) (listed in Table 1), at 25 fps, using an H.264 coder in High Profile with an IPPP Group of Picture (GOP) structure of size 10. An appropriate starting QP was set for each version. We obtained 504 video sequences in total, with 315 used to establish the QoE model and 189 used to validate it.

Table 1. Encoding parameters

SR	BR(kb/s)						
320p	256	320	384	512	640	768	1024
480p	384	512	640	768	1024	1280	1536
720p	768	1024	1280	1536	2048	2560	3072

2.2. Procedure

Our test method used a stimulus procedure based on Absolute Category Rating (ACR). Before the video quality test, we explained the purpose of the experiment and the rating principles to the participants. Forty non-expert participants, between 20 and 30 years old, including 20 women and 20 men, were involved in the experiment. Eleven levels of MOS were used as the rating scale.

Because of the huge number of test video sequences, participants and video sequences were divided into two groups. Each participant assessed twelve contents, 252 video sequences in total. To reduce fatigue effects, each test was divided into three time periods with a 5-minute break in between. Half of the participants were asked to rate a fixed video set in the first time period, and the other half rated the same video set in the last time period. The distributions of these two rating sets are not significantly different ($p>0.05$) based on a t-test, which revealed there is no significant impact of display order or fatigue on MOS for our testing setup. Test videos were shown in full screen. Participants rated each video sequence immediately after viewing. Tests were performed in a laboratory room with typical office lighting conditions. The subject sits on a chair and the viewing distance is 40 cm from the mobile phone which stands on a table. It took approximately an hour to complete 252 video sequences.

The ratings were used to calculate the MOS of each video sequence. There is inherent variability amongst participants in the quality judgment of a given video. The standard deviation and 95% confidence interval of each MOS are calculated, to determine the degree of uncertainty of the participants' ratings.

3. COMPRESSED-DOMAIN SOCCER VIDEO QUALITY ASSESSMENT MODEL

3.1. Model predictors

3.1.1. Video coding parameters

We select the spatial resolution and bitrate as the predictors of video coding parameters.

LBR (Logarithm Bitrate): As predictor, we use a value $LBR = \log_{10}(BR)$ instead of bit rate. Figure 2 shows the relationship between the *LBR* and QoE of the video sequence Close1. QoE increases with the increase of *LBR* in the beginning and remains the same when the *LBR* is large enough. Even at high bitrate, the highest QoE cannot be achieved because of the limitation of the resolution.

SSR (Scaled Spatial Resolution): QoE increases with the increase of resolution when the coding bitrate is large enough.

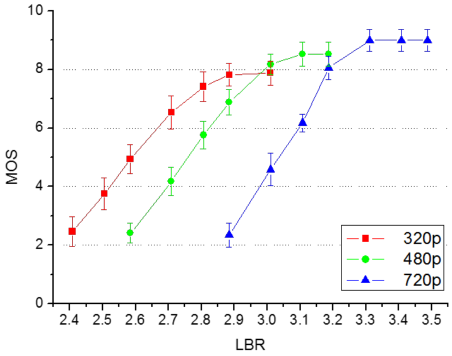

Fig. 2. Impact of LBR and SSR on QoE.

We selected resolutions that are commonly used on the network: 320p, 480p, 720p. Each spatial resolution was scaled by 720p, so the *SSR* values are 0.4, 0.6, and 1.

From Figure 2, only video with 720p resolution and high coding bitrates can achieve Grade nine.

3.1.2. Content characteristic parameters

Video content is strongly related to the compression complexity that the encoder encounters when encoding video sequences with a large amount of detail, complex textures and complex movements [9]. Video with higher compression complexity leads to a lower QoE than other videos under the same coding conditions. Hence, we employed Residual pixel values, Intra 4×4 MBs Proportion in I frames and Motion Vector Magnitude in P frames to represent the richness of spatial texture and the motion characteristic of the video sequence.

RPVI (Residual Pixel Values in I frame): For intra mode, the residual pixel values are used for prediction; therefore, the residual pixel values can predict the spatial statistical dependencies. The prediction *RPVI* is subdivided into 4×4 blocks. In our study, *RPVI* is calculated using the quantized coefficients and QP as follows [6]

$$E_{MB} = \begin{cases} 2^{(Q_{P,MB-4})/3} \times sum\left(Z_{MB} \otimes Z_{MB}\right) & \text{HT OFF} \\ 2^{(Q_{P,MB-4})/3} \times \left(sum\left(Z'_{MB} \otimes Z'_{MB}\right) + sum\left(Z''_{D} \otimes Z''_{D}\right)\right) & \text{HT ON} \end{cases} \tag{1}$$

$$RPVI = \frac{1}{N} \sum_{n \in S_I} E_{MB,n} \ , \tag{2}$$

where Z represents the matrix of quantized coefficients, and $\otimes$ indicates that each element of the left matrix is multiplied by the element in the same position in the right matrix. Here, HT means the Hadamard transform; in the Intra 16×16 mode, the DC coefficients of 4×4 blocks form a new matrix, which will be Hadamard-transformed, and E_{MB} is the estimated energy of an MB. Meanwhile, S_I is the set of MBs in the I frame, and N is the number of pixels we used to calculate *RPVI*.

IMPI (Intra 4×4 MBs Proportion in I frame): The Intra 4×4 mode is based on predicting each 4×4 luma block separately; it works well in coding complex texture. Hence, the average proportion of MBs encoded in Intra 4×4 mode is employed to represent the richness of spatial texture together with *RPVI*.

MVM (Motion Vector Magnitude): The MV of inter-coded MBs in P frames can be employed to evaluate the motion characteristic [10]. We use the magnitude of the MV; a large *MVM* indicates rapid movement, which is related to a low QoE. *MVM* is calculated by

$$MVM = \frac{\sum_{j=1}^{N_t} \sqrt{\left|MV_x^j\right|^2 + \left|MV_y^j\right|^2}}{N_t}, \tag{3}$$

where MV_x^j and MV_y^j denote the horizontal and vertical values of an MV, and N_t is the total number of MVs.

3.1.3. Semantic scene parameters

Soccer video can be divided into three categories. Close-up soccer scenes focus on the facial expressions and the movement of the upper bodies of the players. Medium shots always have rapid movement showing players shooting or running with the ball. Long shots show half the field or the full field.

ST (Scene Type): *ST*, including *STcloseup*, *STmedium*, *STlong*, is used to denote whether a video belongs to the scene type of close-up, medium shot or long shot. Figure 3 illustrates users' QoE for these three scene types at three spatial resolutions. Figure 3(a) shows that long shot videos have the lowest perceived quality at 768k and 1024k. For long shots, viewers require high definition because of the small foreground objects. High resolution is not necessary for close-ups because of the large foreground object. Figure 3(b) and 3(c) show that medium shots nearly always have the lowest perceived quality among the three scene types. That is because medium shot soccer videos have rapid movement. Hence, we conclude that both background proportion and motion activity intensity have a great influence on the users' QoE, so we use them to represent the characteristics of the three scene types.

To represent the background proportion and motion activity intensity of the video, we introduce Horizontal and Vertical Motion Vectors Proportion (*HVMVP*) and *MVM*. While shooting long shots, the camera usually moves either horizontally or vertically; hence, the proportion of horizontal and vertical MVs can be utilized to approximate the proportion of background. Horizontal MVs include all MVs within the angle range of 345° to 15° and 165° to 195° and vertical MVs include all MVs within the angle range of 75° to 105° and 255° to 285°, as shown in Figure 4. *HVMVP* is calculated as

$$HVMVP = \frac{\sum_{j=1}^{N_{sq}} (N_{hor}^j + N_{ver}^j)/N_{fra}^j}{N_{sq}}, \tag{4}$$

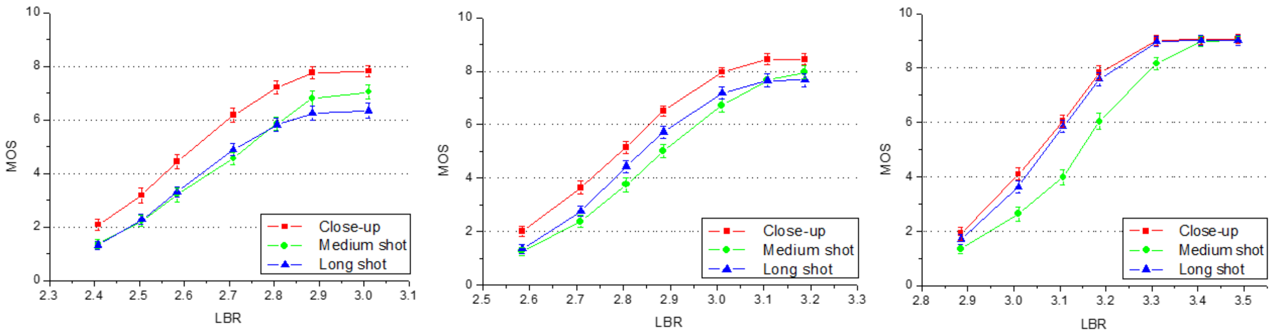

Fig. 3. Impact of video scene types on QoE: (a) 320p; (b) 480p; (c) 720p.

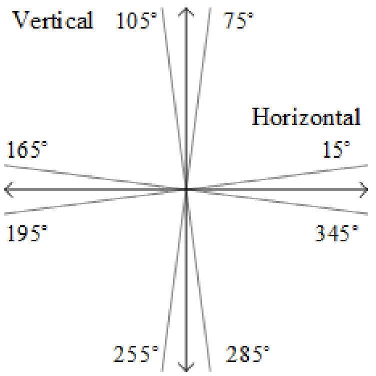

Fig. 4. Horizontal and vertical MVs

where N^j_{hor} and N^j_{ver} are the number of horizontal or vertical MVs in frame *j*, N^j_{fra} is the total number of MVs in frame *j*, and N_{sq} is the number of frames in a sequence.

Table 2 shows the background proportion (BP), *HVMVP* and *MVM* for example video contents. We calculated the proportion of background MBs in the whole picture. From Table 2, we observe that *HVMVP* can approximately represent the proportion of background. A video sequence with a large or a small value of *HVMVP* can be determined to be a long shot or a close-up. When the value of *HVMVP* is neither large nor small enough, *MVM*, which reflects the intensity of temporal change, is selected to resolve this situation. When the *MVM* of a video sequence is not large enough, we can discriminate it as a non-medium shot.

Table 2. BP, HVMVP and MVM for example video contents

Video contents	*BP*(%)	*HVMVP*(%)	*MVM*
Close1	68.56~79.50	63.21	12.44
Close3	67.92~74.97	79.58	35.55
Close4	74.91~75.60	75.27	2.38
Medium1	82.77~86.98	84.87	51.50
Medium2	76.86~89.06	71.89	51.15
Medium3	69.68~89.12	80.87	67.77
Long1	94.97~95.72	93.82	12.40
Long2	91.63~98.23	92.38	45.12
Long4	93.96~95.22	84.59	6.07

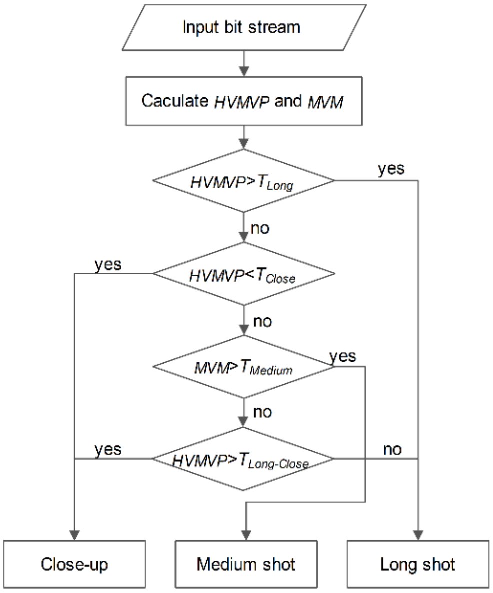

Fig. 5. Soccer scene classification method

Our soccer scene classification method is shown in Figure 5. We set the thresholds based on our experiments. T_{Long}, T_{Close} and $T_{Long-Close}$ are set to be 88%, 71% and 81%, respectively. T_{Medium} is set to 36, 40, and 44 for 320p, 480p, and 720p, respectively. Using these thresholds, this soccer scene classification method achieves 98.02% accuracy for all 504 video sequences. Even if the method fails to classify the soccer video scene in some cases, the scene type we determined for the soccer video still has the characteristics that have impact on the users' QoE.

3.2. Model establishment

The next step of QoE modeling is to map the relationship between QoE and the various parameters. We use *LBR* to be our main QoE prediction parameter and the MOS to be the outcome factor. According to the relationship between MOS and *LBR* shown in Figure 2, the QoE can be formulated as

$$QoE = \alpha \times \left\{ - \exp\,(\frac{\min(LBR - \beta, 0)}{\gamma})^2 \right\}, \qquad (5)$$

where α, β, γ are empirical parameters. The α parameter estimates the maximum value of QoE. From Figure 2, we know that α varies under different spatial resolutions, spatial textures and scene types. So α is defined as

$$\alpha = 9 + \log(\mathit{SSR}) \times (a{+}b \times \mathit{IMPI})\ , \tag{6}$$

and it is affected by *SSR*, *IMPI* and *ST*, and achieves Grade 9 (highest MOS for an encoded video) only when *SSR*=1.

Parameter β estimates the *LBR* value above which viewers can no longer have better QoE. β is defined as

$$\beta = c + d \times \mathit{SSR} + e \times \mathit{MVM}, \tag{7}$$

where β is affected by motion and coding factors, including *SSR* and *MVM*. We use a linear parameter function to represent the relationship between β and QoE predictors because of its simplicity and accuracy.

Parameter γ estimates the slope factor of the model. It is mainly affected by the spatial textures and coding factors, including *IMPI*, *RPVI* and *SSR*. We also use a linear parameter function to represent the relationship between γ and QoE predictors. So parameter γ is defined as

$$\gamma = f + g \times \mathit{SSR} + h \times \mathit{RPVI} + i \times \mathit{IMPI}. \tag{8}$$

Consequently, the QoE is formulated as

$$\begin{aligned} QoE &= (9 + \log(\mathit{SSR}) \times (a{+}b \times \mathit{IMPI})) \\ &\times \left\{ -\exp\left(\frac{\min(\mathit{LBR}-(c{+}d{\times}\mathit{SSR}{+}e{\times}\mathit{MVM}),0)}{f{+}g{\times}\mathit{SSR}{+}h{\times}\mathit{RPVI}{+}i{\times}\mathit{IMPI}} \right)^2 \right\}. \end{aligned} \tag{9}$$

We obtain the model coefficients shown in Table 3. Coefficient *a* has three values depending on the three scene types. The R^2 value of 0.9765 indicates a good fit.

Table 3. Coefficient values

a	b	c	d	e
0.8881/2.1419/2.8420	3.5012	2.5821	0.6749	2.7722×10^{-3}
	f	g	h	i
	0.6588	-0.2486	-2.5145×10^{-4}	-0.3465

4. VALIDATION OF THE PROPOSED MODEL

4.1. Model validation over other videos

We validate CSVQM over 189 test videos mentioned in Section 2.1. Figure 6 shows the performance of CSVQM. The x and y axes represent the model-predicted MOS and the observed MOS, respectively. These scatter plots are distributed near the diagonal line. This indicates that, using CSVQM, the perceived soccer video quality can be accurately measured.

To further evaluate the performance of the model, four statistical evaluation metrics suggested by the VQEG were employed. The Pearson linear correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC), the Root Mean Square Error (RMSE) and the Outlier Ratio (OR) are computed between the observed values and the model-predicted values. The RMSE is used to test the accuracy of the models; smaller RMSE means greater accuracy. The PCC is considered as a metric to measure the linearity between two variables. The SROCC indicates monotonicity between two variables. The values of PCC or SROCC are between -1 and 1; a higher absolute value means a relatively stronger linear relationship or monotonicity. The OR tests consistency; a smaller value of OR means the predictions of the model are more consistent. These metrics do not take into account the subjective uncertainty.

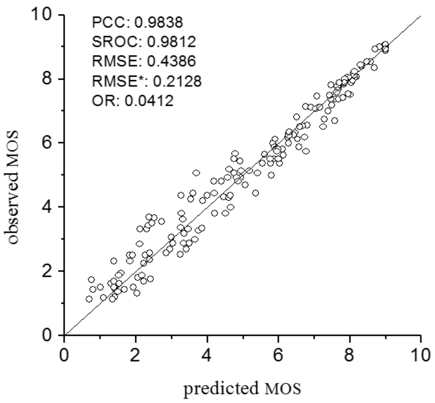

Fig. 6. Scatter plot of observed MOS and predicted MOS for validation

To provide more insight into the subjective test, we evaluate the model performance using the epsilon-insensitive RMSE (RMSE*) statistical metric, which considers the uncertainty of the subjective scores [11]. This is a RMSE that considers the confidence interval of the individual MOS scores. It is calculated like the traditional RMSE, but small differences to the target value are not counted. This RMSE* considers only differences related to an epsilon-wide band around the target value. This epsilon is defined as the 95% confidence interval of the subjective MOS value. As shown in Figure 6, the PCC and the SROCC each has a value close to 1, RMSE is 0.439, RMSE* is 0.213 and OR is 0.041.

4.2. Prediction performance comparison

We compare the prediction performance of CSVQM with several other metrics, including the pixel-domain full reference video quality metric SSIM [1], the pixel-domain no reference VMOS model [3], the compressed-domain no reference metric P.1202.1 [4] and the hybrid model in [12]. The performance evaluation in terms of PCC, SROCC, RMSE, RMSE* and OR is provided in Table 4. The improvement of our model in terms of accuracy, consistency, linearity and monotonicity compared to other models is statistically significant.

Table 4. Performance Comparisons of Model

Metric	PCC	SROCC	RMSE	OR	RMSE*
SSIM [1]	0.876	0.880	0.920	0.137	0.752
VMOS [3]	0.912	0.927	0.637	0.084	0.498
P.1202.1 [4]	0.943	0.903	0.646	0.086	0.454
Hybrid [12]	0.961	0.942	0.562	0.055	0.329
CSVQM	0.984	0.981	0.439	0.041	0.213

4.3. Model complexity comparison

Model complexity includes complexity for decoding and for obtaining QoE predictors. To implement CSVQM, we only need partial decoding of the encoded video and obtain compressed-domain information from the bitstream, which reduces the decoding complexity compared to the FR, pixel-domain NR and hybrid metrics. The complexity for obtaining the QoE predictors of our model is macroblock level. To obtain the QoE predictors, we only need to calculate the information of each macroblock. However, to obtain the predictors of SSIM, VMOS, and the hybrid model in [12], we need to calculate pixel-level information, which requires 256 (16×16) times as many computations as the macroblock level. Hence, the computational complexity of our model is much lower.

5. CONCLUSION

We propose a compressed-domain soccer video QoE assessment model that considers three scene types (close-up, long shot, medium shot). All the model prediction factors can be obtained from the compressed domain without resorting to full decoding, enabling real-time use to predict users' QoE in advance. Residual pixel values, Intra 4×4 MBs Proportion in I frames and Motion Vector Magnitude in P frames are employed to represent video spatial texture and motion characteristics. Soccer scene types are carried through into our model because of viewers' different ratings of different scene types. To distinguish the soccer scene types, a classification method is proposed, by calculating MVs. The compressed-domain soccer video quality assessment model maps the relationship between the multiple QoE prediction factors and MOS.

Our CSVQM can achieve excellent prediction performance; the improvement of our model in terms of accuracy, consistency, linearity and monotonicity compared to other models is statistically significant. Additionally, our model has low implementation complexity in the aspects of decoding and obtaining QoE predictors.

6. REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers*, Nov 2003, pp. 1398–1402.

[3] Y. Shen, Y. Liu, N. Qiao, L. Sang, and D. Yang, "QoE-based evaluation model on video streaming service quality," in *IEEE Globecom Workshops*, Dec 2012, pp. 1314–1318.

[4] Rec. ITU-T P.1202, *Parametric non-intrusive bitstream assessment of video media streaming quality*, 2008.

[5] H. Zhang, F. Li, and N. Li, "Compressed-domain-based no-reference video quality assessment model considering fast motion and scene change," *Multimedia Tools and Applications*, vol. 76, no. 7, pp. 9485–9520, May 2017.

[6] H. Su and F. Yang, "Content-adaptive bitstream-layer model for coding distortion assessment of H.264/AVC networked video," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1199–1208, 2014.

[7] F. Li, S. Fu, Z. Liu, and X. Qian, "A cost-constrained video quality satisfaction study on mobile devices," *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM.2017.2764329, 2017.

[8] S. Hu, L. Sun, C. Xiao, and C. Gui, "Semantic-aware adaptation scheme for soccer video over MPEG-DASH," in *IEEE International Conference on Multimedia and Expo (ICME)*, July 2017, pp. 493–498.

[9] M. N. Garcia, A. Raake, and P. List, "Towards content-related features for parametric video quality prediction of IPTV services," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 757–760.

[10] L. Anegekuh, L. Sun, E. Jammeh, I. H. Mkwawa, and E. Ifeachor, "Content-based video quality prediction for HEVC encoded videos streamed over packet networks," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1323–1334, Aug 2015.

[11] Rec. ITU-T P.1401, *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, 2012.

[12] K. Yamagishi, T. Kawano, and T. Hayashi, "Hybrid video-quality-estimation model for IPTV services," in *IEEE Global Telecommunications Conference (GLOBECOM)*, 2009, pp. 1–5.