CONTENT BASED ADAPTIVE SOCCER VIDEO TRANSMISSION USING QUALITY OF PERCEPTION RANKING

Shenghong Hu[1]
School of Information Management, Hubei University of Economics, Wuhan 430205, China

Abstract. The content based adaptive video transmission (CBAVT) plays a key role in the user oriented multimedia communication, which aims at mapping the importance criterion of user perceived contents to adaptive transmission at different network Quality of Service (QoS). Based on the compositions of information assimilation and entertainment in user perception, this paper proposes a Quality of Perception (QoP) model to predict the importance of soccer video frames, the information metrics and entertainment degree are calculated in each frame by content features and domain knowledge. The adaptive transmission algorithm is designed as two components, intrashot scheme and intershot scheme, intrashot skips low QoP frames in a GoP when slight congestion occurs and intershot performs shot skimming to aggressively reduce bit rate by greedy search when heavy congestion occurs. Experimental results illustrate that the CBAVT performs better objective and subjective quality than those irrespective of video content.

Keywords: Content based adaptive video transmission, Quality of Perception, Intrashot, Intershot

1. INTRODUCTION

In soccer games broadcasting, all kinds of people hold diversity terminals according different bandwidth network to watch video clips. Especially in wireless network, the variable bit rate channel often causes congestion and packet loss in real time, and sacrifices users' perceived content to worse presentation. However, the user pays his/her satisfaction on some important clips, which contain much more information or enjoyment than other insignificant clips. The aim of CBAVT is mapping the importance criterion of user perceived content to adaptive transmission at different network QoS level. When the network degrade into a low bit rate condition, the video server dynamically manipulates encoded stream to yield low bit rate environments, the more important contents annotated after video analysis stage will gained more network resources and perform better satisfaction on user end than others.

In some related works, Chang supposed the video streams can be adapted dynamically by importance criteria in heterogeneous network[1]. Wang proposed a framework for subjective quality optimized adaptation under some low level video features[2]. Xu designed an affective content based personalized adaptation system; the user's preference was divided into three emotion levels to company with three temporal levels composed by I, P or B picture[3]. Cranley designed an optimal adaptation trajectory for spatial and temporal operation on perceived contents in streaming multimedia server[4], Herranz demonstrated a very flexible video summary browsing system based on storyboard and video skimming[5]. All of those researchers provided good ideas to bridge video content analysis and content adaptation in resource constrained environments. A new challenge requires accurately ranking the contents by user characters and mapping to the adaptive transmission on corresponding network condition.

In this paper, we present the QoP model for information content and entertainment content in soccer video, which corresponds to the perception ability of the viewer according to his/her preferences. We also present how to calculate information metrics and entertainment metrics, rank QoP level for every frame, use greedy search to optimize the overall QoP of correctly decoded frames when frames dropping performed in network congestion, finally the objective and subjective evaluations are provided.

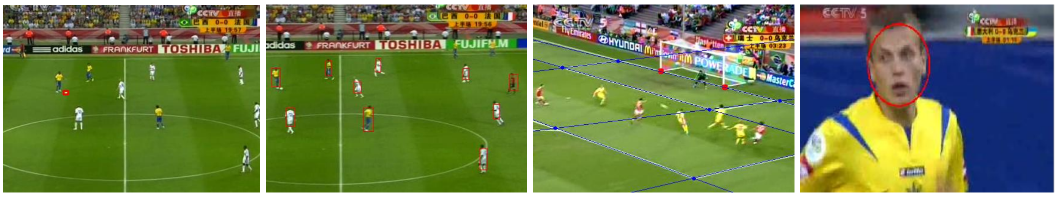
[1]Shenghong Hu (✉)
School of Information Management, Hubei University of Economics, Wuhan 430205, China
e-mail: wuhanhush@126.com

2. CALCULATING QOP FOR SOCCER VIDEO

QoP is a user based definition of quality to measure for the ability to satisfy human needs. Ghinea has used QoP to evaluate the overall subjective quality of multimedia presentations at receiver end, where QoP has defined as two perceptual abilities of information assimilation and the level of enjoyment[6]. However, the perception of video games seems more relations to contents themselves and domain knowledge, namely imposed by semantic features, which can be calculated after content analysis. Considering user perception to perceive information or just enjoy the game when he/she picks up a soccer game, we treat QoP as two metrics of information metrics (QoP_{IM}) and entertainment (QoP_{Ent}) degree, which are linear dependence. In this section, we present an empirical method to calculate the two QoP metrics of every frame in a soccer video.

2.1 Calculating information metrics of QoP

In ongoing soccer games, Human eyes are attracted by significant changing on objects arising, moving, distribution in the pictures, then they can be aware of what have happened in the playfield, why the hurt player seems suffering so much, who the ball has been passed to. The cameramen also use different type shots to help the viewer know the detail of important scenes and events. So the shot classification and objects in the pictures contribute to stimulating the ability to assimilate game information, we calculate the contribution of shot classification and semantic objects in soccer video as information metrics (or IM, for short). Learned from some related works[7, 8], we have succeeded to detect shot classification and most of objects in soccer videos, such as gate, play position, ball and player's face. Then, we compute shot type IM and object IM separately.



(c) Gate (d) player position (a) Ball (b) Player's face

Figure 1. Object detection in playfield

2.1.1 Shot Type information metrics

Different shot type contains the cameraman's intention to reveal or emphasis the certain semantics. The long view is used to convey the ongoing game in a global view, but it would be broken by a medium view or close-up when the cameraman wants to show the detail of a certain event or the appearance of the appearance of a famous star. Meanwhile, most part of long view is boring but the start few frames and the end few frames, the information include in the medium view and close-up view arouse the viewer's attention at begin and scatter to the left frames as the nature of exponential distribution. So, the shot type information metrics of frame *i* can be computed as:

$$STIM(i) = \lambda \cdot F(i)$$

$$F(i) = \begin{cases} 1 - e^{-\left|k - ShotLen/2\right|/(ShotLen/2)}, i \in longview \\ e^{-k/ShotLen}, i \in medium, close-up, else \end{cases} \tag{1}$$

Here, λ is the initial value for different type shot, *2* is for long view, *3* is for medium view, *4* is for close-up, and *1* is for else. *F(i)* is the temporal distribution function of the information metrics, *k* is the index in the shot frame *i* belongs to, and $k \in [0, ShotLen-1]$.

2.1.2 Ball information metrics

The viewer mostly wants to know who the ball is passed to. The different locations of ball between two adjacent frames often drive the viewer's eye and information capacity, which can be measured by the distance of the two coordinates:

$$BIM(i) = \frac{\sqrt{(y_i - y_{i-1})^2 + (x_i - x_{i-1})^2}}{\sqrt{W^2 + H^2}} \tag{2}$$

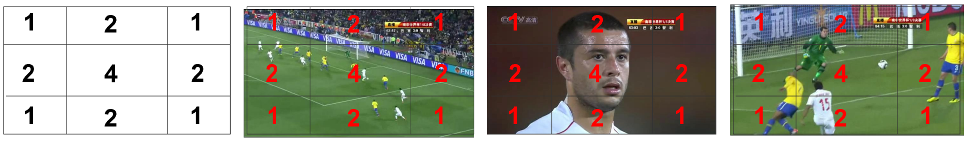
Here, (x_i, y_i) is the ball coordinate of frame *i*, *W* and *H* are the width and height of video resolution.

2.1.3 Players information metrics

The players are under the focus of audience all along, while they scatter into or gather off the picture, a new event occurs. The information metrics is related to the number of players between two adjacent frames and the positions in the field at that time：

$$PIM(i)=\frac{Abs(\sum_{k=1}^{N(i)}P(i,k)\bullet\frac{\varpi_{pos}(k)}{16}-\sum_{k=1}^{N(i-1)}P(i-1,k)\bullet\frac{\varpi_{pos}(k)}{16})}{Abs(\sum_{k=1}^{N(i)}P(i,k)-\sum_{k=1}^{N(i-1)}P(i-1,k))} \tag{3}$$

Here, *P(i, k)*is the position of player *k* in frame *i*, ω_{pos} is the object location weights mapping into a template of the picture. The template is given as:



1	2	1
2	4	2
1	2	1

Figure 2. Template for PIM, FIM

2.1.4 Gate information metrics

The gate mouth arising in the camera's view is large probability event to catch an important semantic event for viewer, where attack, goal, free-kick or corner-kick is coming soon. Any action around the gate is important information in frame *i*, we assume *GIM(i)=1* if the gate mouth is detected in frame *i*, or not *GIM(i)=0*.

2.1.5 Face information metrics

The medium view and close-up are often used by cameraman to send the information on action detail or appearance of the stars, which is just the star fan's want to know. We sum the face area tradeoff by its position, which is denoted as:

$$FIM(i)=\sum_{k=1}^{N}\frac{A_{face}(i,k)}{W\bullet H}\bullet\frac{\omega_{pos}(k)}{16} \tag{4}$$

Here, $A_{face}(i, k)$is the area of face *k* in frame *i*, ω_{pos} is the face location weights mapping into the template of the picture which has been used in *PIM(i)*.

After all information entropies have been calculated, the whole information metrics of frame *i* is computed in terms of the shot type it belongs to, which is denoted as:

$$QoP_{IM}(i)=\begin{cases}STIM(i)+BIM(i)+PIM(i)+GIM(i), i\in longview\\ STIM(i)+FIM(i), i\in mediumview, close-up\\ STIM(i),\ else\end{cases} \tag{5}$$

2.2 Calculating entertainment degree of QoP

The entertainment composition of user perception is satisfied by exciting contents in soccer videos, which arouse the user perception to emotion change. The entertainment degree is used in QoP by us to measure the enjoyment ability of every frame; we calculate entertainment degree according to *Highlight Time Curve* proposed by Hanjlic[9]. The motion activity, shot changing rate, and audio energy are selected.

2.2.1 Motion Activity

The motion activity, defined as a total motion in the scene including both the object and camera motion. The motion activity, *MA(i)* at video frame *i* , can be computed as the average magnitudes of all motion vectors:

$$MA(i)=(\sum_{k=1}^{B}\overrightarrow{v_k}(i))/(B\bullet|\overrightarrow{v_{\max}}(i)|) \tag{6}$$

Here, *B* is the number of blocks within a frame and $\overrightarrow{v_k}(i)$ is the motion vector of the block k in frame *i*.

2.2.2 Shot change density

In order to measure the density of shots, we defined the shot density function *SC(k)*:

$$SC(i) = e^{((1-n(i))/\gamma)} \tag{7}$$

Here, *n(i)* is the amount of frames if the shot frame *i* belongs to. The parameter *γ* is a constant empirically set as *300*, in which the *SC(i)* values are distributed on the scale between *0* and *1*.

2.2.3 Short time energy of sound

The short time energy of frame *i*, *STE(i)*, sums up all spectral values of its *N* audio samples:

$$STE(i) = \sum_{k=1}^{N} x^2(k) \tag{8}$$

Here, The power spectrum *x(k)* is computed for each consecutive segment of the audio signal containing N samples.

After these three features have been calculated, the entertainment degree in QoP of frame *i* is denoted as：

$$QoP_{Ent}(i) = (MA(i) + SC(i) + STE(i))/3 \tag{9}$$

2.3 Ranking information metrics and entertainment degree

After these two types of perception value have been calculated, we quantize them into five levels equivalent to network QoS. The user perception level of frame *i* is denoted as:

$$QoP_{IM|Ent}(i) = \left[QoP_{IM|Ent}(i) - Min(QoP_{IM|Ent}(i))\right] \cdot 5 / \left[Max(QoP_{IM|Ent}(i)) - Min(QoP_{IM|Ent}(i))\right] \tag{10}$$

Here both QoP_{IM} and QoP_{Ent} has been quantized to the interval [1, 5].

3. OVERALL QOP WEIGHTED BY USER PREFERENCE

Although we have calculated information metrics and entertainment degree for each frame, we don't know which perception type is vital to a user, and the user will change perceptual needs within different contexts. So, we send two questions typically indicating what the user perceives to collect user preferences when the video is playing.

Table 1. Two questions to user preferences

Questions	Selections
1. Would you prefer more details on picture information?	Allways(1.0), often(0.75), Normal(0.5), Few(0.25)
2. Would you prefer more duration on highlight clips?	Allways(1.0), often(0.75), Normal(0.5), Few(0.25)

As depicted in *table 1*, the first question is desired to measure the user perception on information assimilation content, and the other is desired to measure the user perception on entertainment. User can submit them according anytime he/she wants. Each question is assigned four answers to user with different perception scores, which range from *0.25* to *1.0*. With a relevant feedback mechanism, the user perception score can be collected as *α* and *β*, they also weight the information metrics and entertainment degree of QoP. If the two questions have been submitted *N* times by user, α_m or β_m is related to scores at *m-1* time they have been answered, the initial value is set by 0.5. The equation is noted as:

$$\alpha_m \mid \beta_m = \begin{cases} 0.5, & m = 0 \\ (\alpha_{m-1} \mid \beta_{m-1}) \cdot (\frac{1}{M} \sum_{m=1}^{M} Score_{\alpha|\beta}(m)), & m > 0 \end{cases} \tag{11}$$

Since we have obtained two weights *(α, β)* for preferences acquisition, the overall perception level of frame *i* can be re-ranked as:

$$QoP(i) = \alpha * QoP_{IM}(i) + \beta * QoP_{Ent}(i) \tag{12}$$

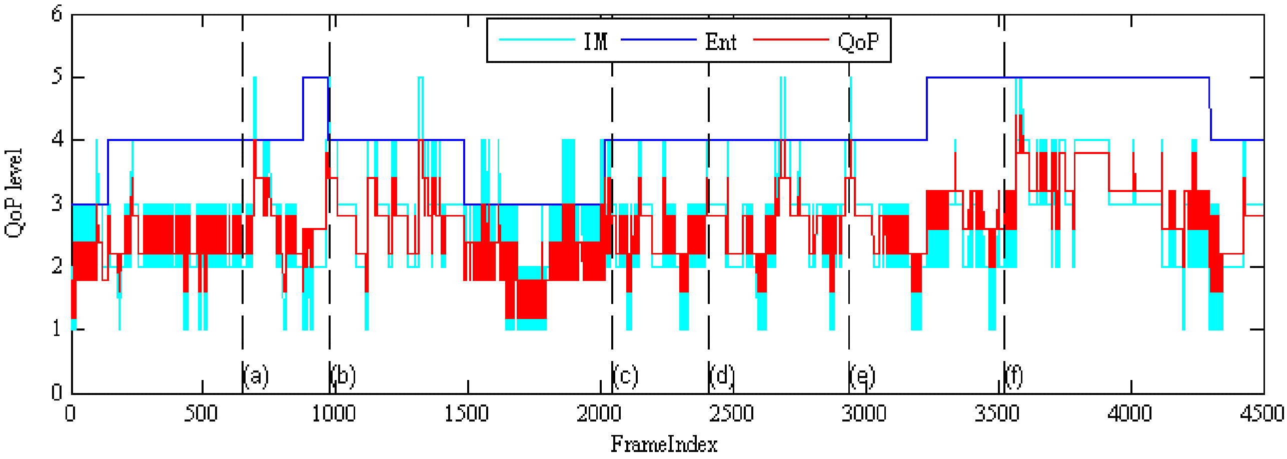


Figure 3. An example for overall QoP ranked by user preferences, several events occur at: (a) shoot but not goal; (b) playback for shoot; (c) medium shot for a neat move; (d) medium shot for a long pass; (e) close-up for coach; (f) shoot and goal.

4. THE QOP OPTIMIZED ALGORITHM OF CBAVT

Video transmissions over Best-effort IP networks are unreliable and unpredictable particularly in wireless networks. QoS is a general measure to quantify the performance of an end to end channel via parameters such as delay, loss and jitter. Network congestion results in lost video packets, which, as a consequence, produces poor quality video and causes unsatisfied perceptual needs simultaneously. Any information that could exceed the available bandwidth would be delayed or discarded by the network. At the application layer, A user often prefers higher QoP contents in better quality while other contents corrupted due to packet loss, adaptive operations are used to manipulate content with different temporal level is in accordance with the available bandwidth, the proposed adaptation operation tradeoff QoP and QoS parameters to preserve more perceived contents by user. When the network congestion occurs, the less important temporal units annotated by low QoP value are skipped immediately to search the next higher QoP units for transmission. This can be modeled as a QoP optimized problem constrained by QoS as:

$$\begin{aligned} &Maxmize \sum_{i \in SD(i)} QoP(i) \\ &s.t.(\sum_{i \in SD(i)} i)/N \le adaptRatio(QoS, QoP) \end{aligned} \tag{13}$$

Where *SD(i)* is the successfully decoded frames set, *adaptRatio(QoS, QoP)* is the temporal operation parameter mapped by QoS and QoP in a temporal unit. If we consider a video encoded with *N* frames can separate into successive temporal units. In a semantic structure, the temporal units are divided into shot level, group of pictures (GoP) level, frame level. Shot skimming will aggressively reduce the send rate than skipping a GoP or a frame. The shot level is also an individual semantic unit, skipping a shot may hardly hurt the experience of viewer, so the shot skimming should be used with more critical conditions. We treat the temporal adaptation as two schemes, *InterShot* scheme and *IntraShot* scheme. In *InterShot* scheme, the frame skimming condition is referred by shot QoP value, which is the average QoP value of all frames in the shot. Greedy search algorithm is employed to find the highest QoP shot to send in next *M* shots. When the correct frame location has been found, it should be normalized to the latest I frame to prevent decoding error.

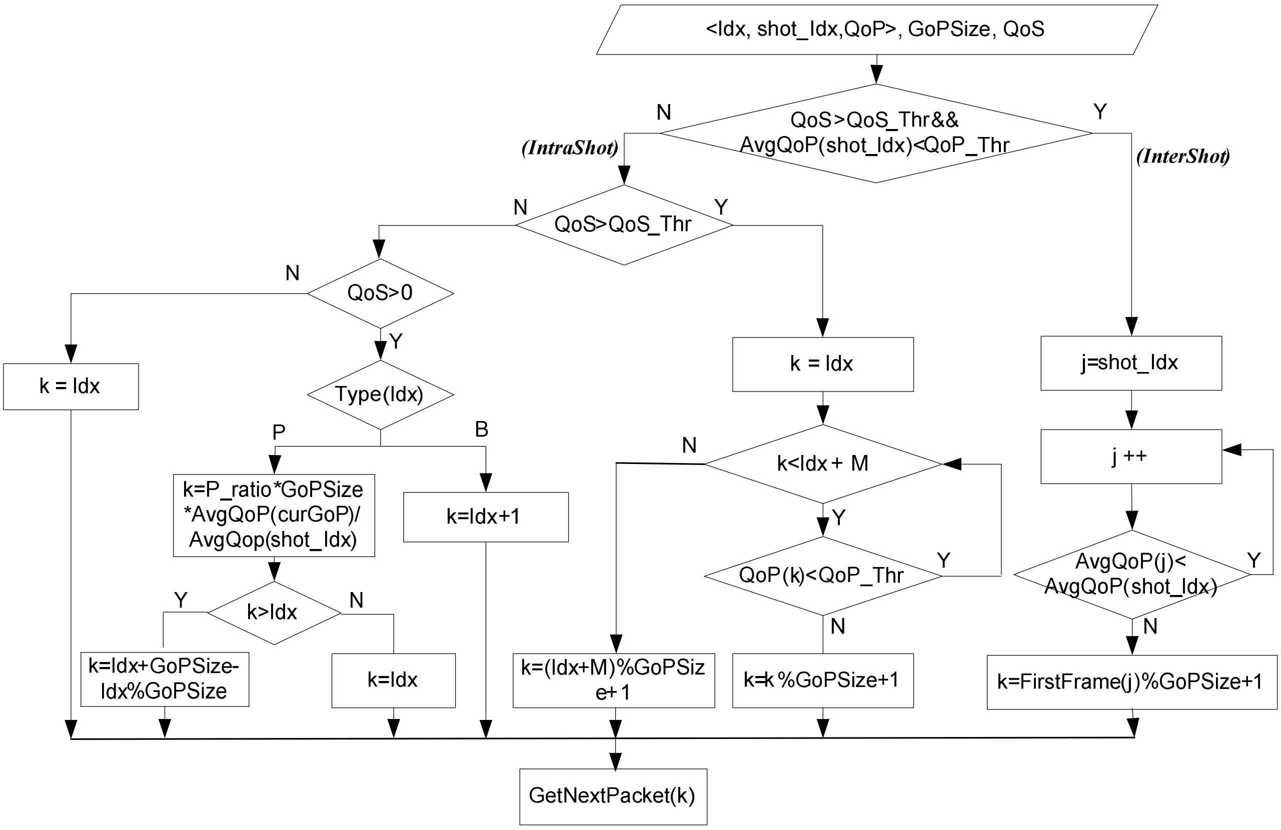


Figure 4. The process chart of QoP maximization algorithm

In *IntraShot* scheme, the frames in a GoP can be dropped until only I frame left in conventional method, the remaining rate of P frame (*P_ratio*) or B frame only because of QoS value. Now, the frame drop condition is referred by frame QoP value and QoS value together. The frame remaining percent of P or B frames mapped by QoS value needs to be recalculated by the ratio of average QoP of associated GoP to average QoP of associated shot. Two thresholds of *QoS_Thr* and *QoP_Thr* are designed to switch the proposed schemes, only if current QoS value is greater than *QoS_Thr* and average QoP of current shot is less than *QoP_Thr*， the *IntraShot* scheme will switch to *InterShot* sheme.

5. EXPERIMENTAL RESULTS & EVALUATIONS

5.1 Implementation in Darwin Streaming Server (DSS)

In our experiments, the apple's DSS5.5.5 is used as a stream server, and the QuickTime 7.62 is used as the receiver. The Nistnet2.0.12b is installed in a Linux Router to generate a variable bandwidth channel between the DSS server and QuickTime client. A 3-minute long soccer clip is selected from world cup 2006 to evaluate the proposed adaptation algorithm. The original input video rate before adapting is h.264 encoded at 1400kbps, the frame rate of 25 fps, and GoP Size is 15 with the sequence of "IBPBPBPBPBPBPBB" hinted by QuickTime.

We compare the network throughput performance of three adaptive transmission methods. The first is CBAVT with the network method based on Reliable UDP (RUDP), which is implemented as a TCP-Friendly Rate Control Protocol in DSS 5.5.5; the second is additive increase/multiplicative-decrease (AIMD); the third method is merely RUDP. An actual bit rate channel is set in the Nistnet router varying between 512kbps and 1024kbps, the throughput from the DSS server is collected in 180 seconds.

Among three adaptive transmission methods, the throughput of CBAVT is near to RUDP; too much higher than AIMD, proves high performance in network layer.

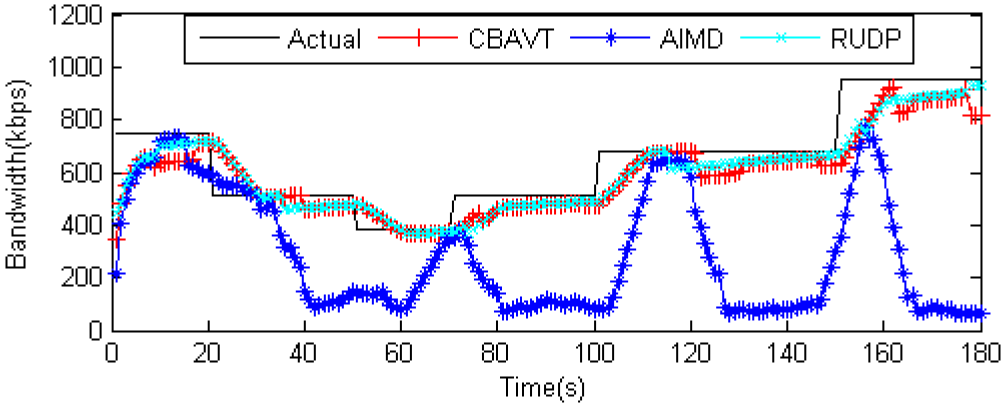


Figure 5. Throughputs of CBAVT, AIMD, RUDP

5.2 Evaluation of CBAVT

The Quality in application layer of CBAVT can be evaluated by objective measurement and subjective measurement.

5.2.1 Objective measurement

Objective measurement often uses Peak Signal-to-Noise Ratio (PSNR) to evaluate the decoded images. We capture five pictures at the same timestamp in the playback videos by three methods of CBAVT, AIMD and RUDP. The five pictures are also annotated by five level QoP values.

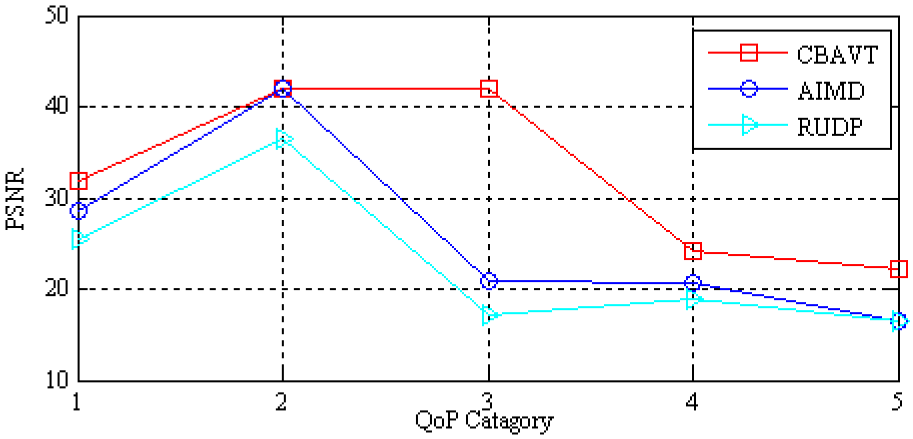


Figure 6. PNSR evaluations of CBAVT, AIMD, RUDP

When the low level QoP contents are transmitted, the three methods performed much the same, but when high QoP contents are transmitted, the CBAVT performs much better than the others.

5.2.2 Subjective measurement

We have invited five subjects to evaluate the subjective quality of the transmitted videos. Each subject will watch the transmitting soccer games and give the scores of vision quality, smoothness and contents completeness, which are guided by following descriptions:

(1) Vision quality: Almost all the pictures have few visual distortions;

(2) Smoothness: Few highly motion clips were felt into halting or few jerks have happened;

(3) Contents completeness: all the information presented is grasped without anything else much eager to know.

Five scales are given with 5 corresponding to strongly satisfied, 4 mostly satisfied, 3 accept, 2 reject and 1 strongly reject. To make a comparison, the subjects will watch the origin video and all three transmitted videos by CBAVT, AIMD and RUDP. The average scores from all subjects are given in *table 2*.

Table 2. Average scores in subjective evaluation

Subjects	CBAVT	AIMD	RUDP
vision quality	4.2	2.7	1.2
smoothness	3.9	2.4	1.5
contents completeness	4.2	2.8	1.9

6. CONCLUSIONS

In this work, a QoP model for user perceived contents in each frame of a soccer video has been presented; temporal operation has been divided into two schemes, *intrashot* and *intershot*, which are applied in a CBAVT system to satisfy QoS and user perceptual needs. All results of objective evaluation and subjective evaluation perform better than AIMD and RUDP. However, many aspects can be improved. A key issue to achieve QoP is the accuracy of content analysis; more best semantic features will be adopted in future works, and hopefully achieve QoP value more close to user perception. Another issue that is expected to improve QoP optimized algorithm, extend to a synthetically utility on temporal-spatial-quality optimization for h.264 SVC streaming.

ACKNOWLEDGEMENTS

We are gratefully acknowledge the supports received from the Natural Science Fund Committee of Hubei Province (2012FFC01301) and Major Project of Education Department in Hubei Province (D20132205) in China.

REFERENCES

1. S. F. Chang, D. Zhong, and R. Kumar, “Real-time content based adaptive streaming of sports video”, in proceedings of IEEE CVPR conference, Hawaii, 148-158(2001).
2. Y. Wang, Jae-Gon Kim, and S.-F. Chang, "Utility-Based Video Adaptation for Universal Multimedia Access (UMA) and Content-Based Utility Function Prediction for Real-Time Video Transcoding", IEEE Transactions on Multimedia, 9(2), 213-220(2007).
3. Min Xu, Jesse S. Jin, and Suhuai Luo, "Personalized Video Adaptation Based on Video Content Analysis". MDM/KDD'08, Las Vegas, Nevada, USA, 26-35(2008).
4. N. Cranley, P. Perry and L. Murphy, "User perception of adapting video quality", International Journal of Human-Computer Studies, Elsevier, 64(8), 637-647(2006).
5. L. Herranz, and J. M. Martı´nez, “An integrated approach to summarization and adaptation using H.264/MPEG-4” SVC, Signal Process: Image Communication, Elsevier, 24(6), 499-509(2009).
6. G. Ghinea, J. P. Thomas, Quality of Perception: User Quality of Service in Multimedia Presentations, IEEE Transactions on Multimedia, 7(4), 786-789(2005).
7. Junqing Yu, Yang Tang, and Zhifang Wang, Playfield and Ball Detection in Soccer Video. Advances in Visual Computing, Springer Berlin, Heidelberg, 4842, 387-396(2007).
8. Junqing Yu, Yunfeng He, and Kai Sun, "Semantic Analysis and Retrieval of Sports Video," Frontier of Computer Science and Technology, Japan-China Joint Workshop, 97-108(2006).
9. A. Hanjalic. “Adaptive extraction of highlights from a sport video based on excitement modeling,” IEEE Transactions on Multimedia, 7(6), 1114-1122(2005).