Region-of-Interest Based Coding Scheme for Synthesized Video

Wenbo Zhao[1], Jingjing Fu[2], Yan Lu[2], Shipeng Li[2], Debin Zhao[1]
[1]Dept. of Computer Science and technology, Harbin Institute of Technology, Harbin, China
[2]Microsoft Research Asia, Beijing, China
{wbzhao,dbzhao}@hit.edu.cn, {jifu,yanlu,spli}@microsoft.com

Abstract—**In many multimedia applications, such as online speech, video chat and online conference, multiple source videos are synthesized in a single scene for explicit presentation and the synthesized video is compressed for transmission. The source video with important contents deserves more compression resources for quality preservation under the bandwidth constraint. To address this problem, a region-of-interest (ROI) based coding scheme for synthesized video is proposed in this paper aiming at achieve better and consistent quality for ROI source videos with the bitrate meeting the constraint bandwidth. In the proposed coding scheme, ROI based rate-distortion (R-D) models are established, in which different R-D models are built for different source video. Then an objective function is defined with respect to the video quality and the consistency of video quality. By minimizing the objective function, the optimal quantization parameters for the ROI and non-ROI source videos are obtained. The experimental results show that the proposed coding scheme achieves better and consistent quality for ROI source videos.**

Index Terms—**Video coding, synthesized video, region of interest, R-D model, consistency**

I. Introduction

With the rapid development of multimedia technologies, synthesized video, which is generated by several source videos, has been widely used in many internet applications, such as online speech, video chat and online conference. In such applications, the synthesized video may contain several source videos and some of them form the region of interest. This kind of source videos is named as ROI videos and the other source videos as non-ROI videos. Generally, if the transmission bandwidth is limited, the encoder should try to keep ROI videos with better and consistent quality.

For ROI based video coding, some approaches have been proposed. Doulamis et al. [1] employed a neural network to detect ROI and allocate more bits to ROI. Yang et al. [2] proposed a rate control scheme that the quantization parameter (QP) of ROI is determined by the interest level. Bulla et al. [3] used a Viola-Jones face detector to detect ROI, then the ROI and non-ROI are set with different QP and the QP distance is a constant. In [4], the encoder allocated more computational resources to ROI by using smaller QP, more referencing frames, and larger search range of motion estimation.

All the above coding schemes are designed for the sake of better resource allocation between ROI and non-ROI on coding computation and bits. As the ROI video sources in the synthesized video often display in the fixed region and contain important information, users are more sensitive to the frequent ROI quality change. Therefore, the consistency of ROI's quality should be taken into account. For example, in the online presentation, the region of slides should be kept high quality even if the bandwidth is limited. In the video chat, the speaker's face is more important than others' and should be clear.

In this paper, a ROI based coding scheme for synthesized video is presented. In the proposed coding scheme, ROI based R-D models are first proposed, different R-D models are built for different source videos. Then, an objective function is defined, considering the video quality and the consistency of video quality. After that, the R-D models are embedded into the objective function. By minimizing the objective function, the optimal quantization parameters for the ROI and non-ROI source videos are obtained and used for encoding the synthesized video.

The rest of paper is organized is as follows. ROI based R-D Models are proposed in Section II. The ROI based coding scheme is presented in Section III. The experimental results are showed in Section IV. Finally, the conclusions are given in Section V.

II. ROI Based R-D Models for Synthesized Video

H.264/AVC is a high-performance video-coding standard and the Lagrangian coder control method [5] is commonly used in the codec to improve the coding performance. In the method, a quadratic R-D model and a linear mean absolute difference (MAD) prediction model are built to solve the rate distortion optimization problem. The R-D model is described as:

$$R = \frac{\beta MAD}{Q_{STEP}} + \frac{\gamma MAD}{{Q_{STEP}}^2} \tag{1}$$

where R is the texture bits cost of the current frame, β and γ are model parameters, and Q_{STEP} denotes quantization step. In H.264/AVC, the relationship between QP and Q_{STEP} is:

$$Q_{STEP} = 2^{(QP-4)/6} \tag{2}$$

The linear MAD prediction model is:

$$MAD = a_1 MAD' + a_2 \tag{3}$$

where MAD is the MAD of current predicted frame, the parameters of the previous frame is represented in the form of $(*)'$, a_1 and a_2 are the parameters of this model.

Since different QPs are set to ROI and non-ROI videos, the R-D model in Eqn. (1) is extended to multiple video sources. To choose an optimal pair of QPs, we need to know the accurate bits cost and distortion of each video to prevent the total cost of the frame goes beyond the pre-allocated bits. Building different

R-D models and MAD predictions for different videos is an effective method for solving the problem.

Assume there are n source videos. The ROI based R-D models are described as:

$$\frac{\beta_i MAD_i}{q_i} + \frac{\gamma_i MAD_i}{q_i^2} = R_i,\ i = 1,2,\dots,n \tag{4}$$

where i denotes the number of source, q_i is the Q_{STEP}, R_i is the texture bits, β_i and γ_i are the parameters for different source respectively. And MAD prediction models are built to calculate the MAD_i of *i-th* video source：

$$MAD_i = a_{i1} MAD_i' + a_{i2},\ \ i = 1,2,\dots,n \tag{5}$$

where a_{i1} and a_{i2} are the parameters of this model.

III. ROI Based Coding Scheme

The framework of the proposed scheme is shown in Fig 1. With the bit budget and ROI based R-D models, the optimize pair of QP value corresponding to ROI and non-ROI can be calculated. After exception handling on the QP setting, the QP will be used for encoding the frame.

As each source video has an R-D model, the ROI based R-D models is more complicated than the single source video. Moreover, the quality consistency of ROI video is important to visual experience and should be taken in account. To address this problem, an objective function is defined considering both the video quality and the consistency of it. The synthesized Video with two sources are solved first and then the solution can be applied to the synthesized video with multiple sources.

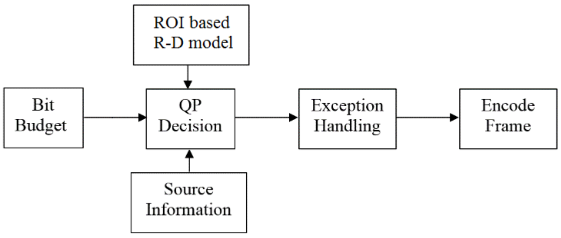


Fig 1: The flowchart of the proposed scheme

A. QP Decision for Synthesized Video with Two Sources

Considering the typical case that there are only two source videos, assuming the first one is marked as ROI and the second is non-ROI. The video quality is measured by the distortion: d_i, and the consistency of video quality is measured by the difference of distortions between successive frames: Δd_i.

For better visual ROI quality, both d_i and Δd_i should be considered. The objective function is defined as a linear combination of d_i and Δd_i. In addition, the channel bandwidth should be taken into account. So our goal is to find an optimal pair of QP: Q_1 and Q_2, by solving the following problems:

$$\begin{aligned} & min(\eta_1 d_1 + \eta_2 \Delta d_1 + \eta_3 d_2 + \eta_4 \Delta d_2) \\ & s.t.\ \ R_1 + R_2 \le R_T \end{aligned} \tag{6}$$

where $\eta_1 \sim \eta_4$ are the weight parameters, R_T is the bit budget.

A distortion-quantization (D-Q) model in [6] is used to solve the problem. In the model, the mean square error (MSE) is used to represent the distortion, and the relation between MSE and Q_{STEP} is a linear model:

$$d_i = \mu_i MAD_i q_i \tag{7}$$

where μ_i is the parameter of the model, $q_i = 2^{(Q_i-4)/6}$.

We substituting Eqn. (4) and (7) in (6), to simplify the problem, the follow marks are introduced:

$$\begin{aligned} \varepsilon_1 &= \eta_1 \mu_1 MAD_1\ , & \varepsilon_2 &= \eta_2 \mu_1 MAD_1\ , \\ \varepsilon_3 &= \eta_3 \mu_2 MAD_2\ , & \varepsilon_4 &= \eta_4 \mu_2 MAD_2\ , \\ \theta_1 &= \beta_1 MAD_1\ , & \theta_2 &= \gamma_1 MAD_1\ , \\ \theta_3 &= \beta_2 MAD_2\ , & \theta_4 &= \gamma_2 MAD_2\ . \end{aligned}$$

We assume that the D-Q models between adjacent frames are similar. That is to say: $\mu_i \approx \mu_i', \varepsilon_i \approx \varepsilon_i'$. Then the optimization problem is converted to:

$$\begin{aligned} & Min(\varepsilon_1 q_1 + \varepsilon_2 |q_1 - q_1'| + \varepsilon_3 q_2 + \varepsilon_4 |q_2 - q_2'|) \\ & s.t.\ \ \frac{\theta_1}{q_1} + \frac{\theta_2}{q_1^2} + \frac{\theta_3}{q_2} + \frac{\theta_4}{q_2^2} \le R_T \end{aligned} \tag{8}$$

To solve the problems in Eqn. (8) analytically, we need to consider the relationship between q_1 and q_1', $q_2$and q_2'. Then we have four combination cases:

Case 1: $q_1 \ge q_1',\ q_2 \ge q_2'$,

Case 2: $q_1 < q_1',\ q_2 < q_2'$,

Case 3: $q_1 \ge q_1',\ q_2 < q_2'$,

Case 4: $q_1 < q_1',\ q_2 \ge q_2'$.

Here we discuss Case 4 as an example, the optimization problem is converted to:

$$\begin{aligned} & min(\varepsilon_1 q_1 - \varepsilon_2 q_1 + \varepsilon_2 q_1' + \varepsilon_3 q_2 + \varepsilon_4\, q_2 - \varepsilon_4 q_2') \\ & s.t.\ \ \frac{\theta_1}{q_1} + \frac{\theta_2}{q_1^2} + \frac{\theta_3}{q_2} + \frac{\theta_4}{q_2^2} \le R_T \\ & \qquad q_1 \le q_1' \\ & \qquad q_2' \le q_2 \end{aligned} \tag{9}$$

This problem can be solved by Karush-Kuhn-Tucker optimality conditions, the gradient equation is given by:

$$\begin{pmatrix} \varepsilon_1 - \varepsilon_2 \\ \varepsilon_3 + \varepsilon_4 \end{pmatrix} + \lambda_1 \begin{pmatrix} -\frac{\theta_1}{q_1^2} - 2\frac{\theta_2}{q_1^3} \\ -\frac{\theta_3}{q_2^2} - 2\frac{\theta_4}{q_2^3} \end{pmatrix} + \lambda_2 \binom{1}{0} + \lambda_3 \binom{0}{-1} = 0 \tag{10}$$

where λ_m is non-negative Lagrange multiplier.
The Complementary slackness conditions are:

$$\lambda_1 \left(\frac{\theta_1}{q_1} + \frac{\theta_2}{q_1^2} + \frac{\theta_3}{q_2} + \frac{\theta_4}{q_2^2} - R_T \right) = 0\ , \tag{11}$$

$$\lambda_2 (q_1 - q_1') = 0\ , \tag{12}$$

$$\lambda_3 (q_2' - q_2) = 0\ , \tag{13}$$

$$\lambda_1,\ \lambda_2,\ \lambda_3 \ge 0\ . \tag{14}$$

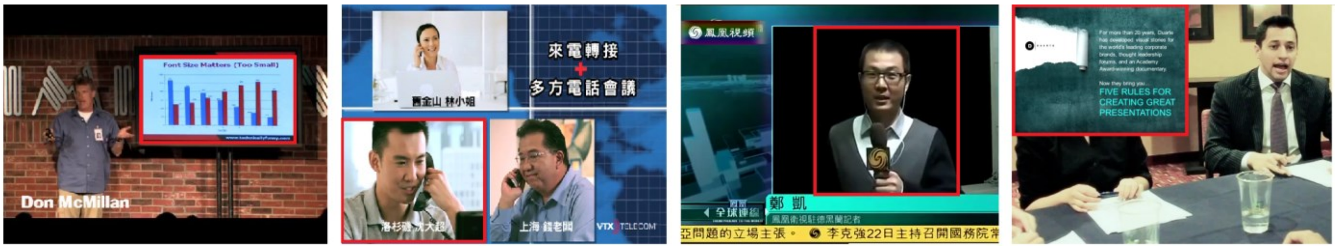
After solving the problems in Eqn. (10) ~ (14), we can get the best QP for ROI and non-ROI in this frame for Case 4. The same method can also be used for the remaining three cases to get the final solution which can minimize the objective function.

B. QP Decision for Synthesized Video with Multiple Sources

For the videos which are generated by n source videos, the objective function can be described as:

$$\begin{aligned} & min(\textstyle\sum_{i=1}^{n} (\eta_{2i-1} d_i + \eta_{2i} \Delta d_i)) \\ & s.t.\ \ \textstyle\sum_{i=1}^{n} (R_i) \le R_T. \end{aligned} \tag{15}$$

For simplicity, we can assume that all the ROI sources are encoded with the same Q_{STEP}: q_1, the subscript of ROI sources range from 1 to m, and all the non-ROI sources are encoded with the same Q_{STEP}: q_2, the subscript of non-ROI sources



(a) Online speech (b) Video chat (c) News (d) Online conference

Fig 2: the ROI region of each video sequence.

range from m+1 to n. The following marks are introduced:

$$\varepsilon_1 = \sum_{i=1}^{m}(\eta_{2i-1}\mu_i MAD_i)\,, \qquad \varepsilon_2 = \sum_{i=1}^{m}(\eta_{2i}\mu_i MAD_i)\,,$$

$$\varepsilon_3 = \sum_{i=m+1}^{n}(\eta_{2i-1}\mu_i MAD_i)\,,\ \varepsilon_4 = \sum_{i=m+1}^{n}(\eta_{2i}\mu_i MAD_i)\,,$$

$$\theta_1 = \sum_{i=1}^{m}(\beta_i MAD_i)\,, \qquad \theta_2 = \sum_{i=1}^{m}(\gamma_i MAD_i)\,,$$

$$\theta_3 = \sum_{i=m+1}^{n}(\beta_i MAD_i)\,,\ \theta_4 = \sum_{i=m+1}^{n}(\gamma_i MAD_i)\,.$$

Then the optimal problems in Eqn. (15) are converted to the same problems in Eqn. (6) and can be solved with the above method.

C. Exception Handling

Some thresholds are pre-defined to prevent Q_1 and Q_2 from changing too much. The max difference between Q_1 and Q_1' is 2, while the max difference between Q_2 and Q_2' is 4. In addition, the distance range between Q_1 and Q_2 is $[max\left(\frac{Q_1}{5} - 1,4\right), 12]$.

If all the four combinations fails to find one solution under above conditions, we enumerate all the pair of legal value of Q_1,Q_2, and substitute them into Eqn. (8), then choose a pair that minimize the function and satisfy the bit budget.

If R_T is too small to find a pair of QP that satisfy the bit budget. Then $Q_1 = Q_1' + 2$, $Q_2 = Q_2' + 4$.

IV. Experimental Results

The performance of the proposed rate control scheme for synthesized video is evaluated in this section. The proposed scheme is implemented on JM15.1 and the bit budget is calculated by the GOP and frame level rate control in JM. The information about the location of sources and which sources are ROI is input by the user.

Several test sequences with 480P spatial resolution are used in the experiments: "Online speech", "Video chat", "News", and "Online conference". All the sequences are made up of two source and the ROI regions are shown in Fig 2.

In the experiments, four bitrate levels are considered at 30fps for each test sequence: 800Kbps, 600Kbps, 400Kbps and 200Kbps. The GOP type is IPPP, the GOP size and sequence length are both 150. To achieve high consistency of quality in ROI, the weight parameter of $\Delta d_1, \eta_2$ should be big enough. The parameter can be adjust to meet the requirement of the user. In our experiments: η_1 and η_3 is set to 0.75 and 0.2, η_2 is set to 12.5 and η_4 is 0.5, respectively.

For comparison, we also conduct the experiment with the JM 15.1 codec and the bitrate savings in [3] (denote by BS). We define C_0 as the initial QP value of the video that encoded by the original JM codec. Four QP level are considered corresponding to the bitrate: 20, 25, 30 and 35. For the video encoded by the proposed scheme, the QP of ROI should be $C_0 - C_1$ in order to increase the quality of ROI, and the QP of non-ROI will be $C_0 + C_2$. As mentioned in [7]: $C_2 = nC_1/(N-n)$, here N denotes the number of MB in a frame and n denotes the number of MB in ROI, and $C_1 = 3$. The QP distance between ROI and non-ROI in BS is 6.

TABLE 1: BITRATE ERROR COMPARISON

Scheme	Bitrate error in different bitrate				Average Error
	800Kbps	600Kbps	400Kbps	200Kbps	
Proposed	0.13%	0.22%	**0.30%**	**0.39%**	0.26%
BS	**0.08%**	**0.05%**	0.33%	0.46%	**0.23%**
JM	0.10%	0.24%	0.31%	1.01%	0.41%

As shown in Table 1, the average bitrate error of videos encoded by the proposed scheme is 0.26% while the error is 0.41% encoded by JM and 0.23% encoded by BS, the proposed scheme performs better in rate control than JM.

From Table 2, the total average PSNR gain of ROI region is 3.19dB in our scheme, while 2.69dB in BS. To show the optimal quality of ROI region, an example of subjective visual quality comparison is shown in Fig 3, the quality of ROI is effectively improved. Therefore, the proposed scheme demonstrates high video quality in ROI region relative to other schemes in different scenes.

Fig 4 shows the performance comparison of the consistency of quality. It can be observed that the QP of ROI keeps constant in our scheme, and this leads to the quality of ROI is consistent.

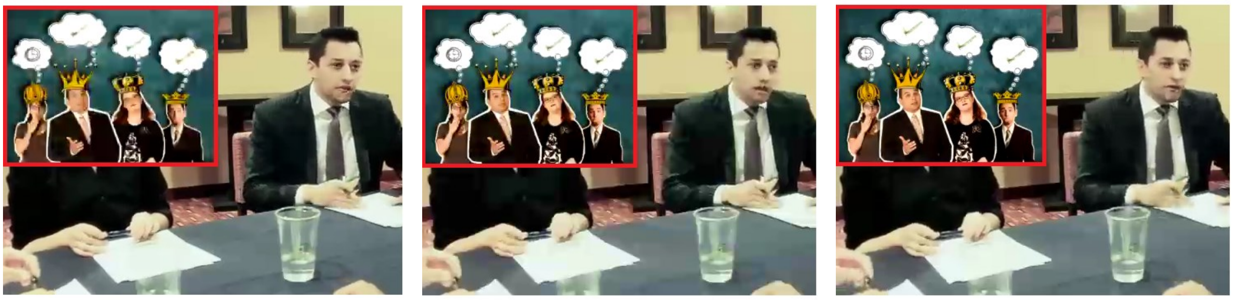
The results show that the proposed scheme is able to achieve better quality and consistency in ROI. With the optimized selected QP, the QP of ROI keeps low in the proposed scheme, thus leads to a better perceived quality, also the stability of QP ensures the consistency of quality in ROI. Our scheme can allocate more bits to ROI than other schemes, even when the bandwidth is limited. What's more, because of the ROI based R-D models, the proposed scheme performs better rate control on most test sequences than JM.

V. Conclusion

In this work, we propose an ROI coding scheme for synthesized video. R-D models with different parameters are assigned to different source video to make an accurately estimation for the bit cost of each source. An objective function in which quality and consistency of different regions are set different weight is built, the encoder calculates an optimize pair of QP of ROI and non-ROI videos to minimize the value of the function. This scheme can work well in different scenes, and achieves not only better quality and consistency in ROI region but also less bitrate error than JM 15.1.

TABLE 2: VISUAL QUALITY (PSNR) OF ROI AND NON-ROI COMPARISON

Test Seq.	Scheme	Average PSNR of ROI in different bitrate (dB)				Average PSNR (dB)	Average PSNR of non-ROI in different bitrate (dB)				Average PSNR (dB)
		800Kbps	600Kbps	400Kbps	200Kbps		800Kbps	600Kbps	400Kbps	200Kbps	
Online speech	Proposed	**41.02**	**38.89**	36.41	**33.48**	**37.45**	39.32	37.49	35.92	32.28	36.25
	BS	40.09	38.52	**36.48**	33.10	37.05	39.42	37.94	35.96	32.68	36.50
	JM	37.93	36.36	34.21	30.98	34.87	**40.84**	**39.48**	**37.48**	**34.07**	**37.97**
Video chat	Proposed	**56.80**	**54.30**	**51.00**	**44.98**	**51.77**	49.80	47.09	43.48	37.43	44.45
	BS	55.90	53.57	50.32	44.95	51.19	50.47	48.11	44.21	37.59	45.10
	JM	53.60	51.43	48.09	41.69	48.70	**52.38**	**49.87**	**46.09**	**39.29**	**46.91**
News	Proposed	**49.33**	**47.51**	**44.34**	**40.71**	**45.47**	42.50	40.83	38.33	34.02	38.92
	BS	48.01	46.46	44.07	40.03	44.64	42.81	41.10	38.30	34.22	39.11
	JM	45.18	43.30	40.52	37.14	41.54	**43.87**	**41.93**	**39.13**	**35.03**	**39.99**
Online conference	Proposed	**49.38**	**46.85**	42.37	36.78	**43.85**	44.72	42.37	39.67	35.02	40.45
	BS	48.55	46.14	**42.73**	**37.10**	43.63	45.41	43.04	39.71	34.98	40.79
	JM	45.75	43.20	39.50	34.12	40.64	**46.89**	**44.49**	**41.13**	**36.30**	**42.20**

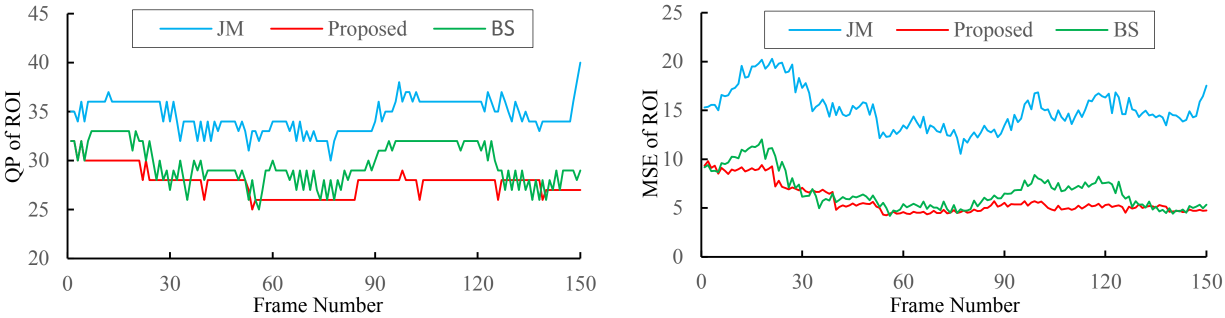


（a） JM
PSNR: ROI: 24.9dB, Non-ROI: 32.5dB

(b) BS
PSNR: ROI: 27.2dB, Non-ROI: 32.3dB

(c) Proposed
PSNR: ROI: 29.5dB, Non-ROI: 31.6dB

Fig 3: Performance in the ROI of "online conference" sequence (Bitrate: 200Kbps)



(a) QP of ROI

(b) MSE of ROI

Fig 4: Consistency of quality in "Video chat" sequence (Bitrate: 200Kbps)

ACKNOWLEDGEMENT

This work has been supported in part by the Major State Basic Research Development Program of China (973 Program 2015CB351804), the National Science Foundation of China under Grant No. 61272386。

REFERENCES

[1] Doulamis, N., Doulamis, A., Kalogeras, D., Kollias, S., "Low bit-rate coding of image sequences using adaptive regions of interest," Circuits and Systems for Video Technology, IEEE Transactions on, vol.8, no.8, pp. 928-934, Dec 1998.

[2] Ling Yang, Li Zhang, Siwei Ma, Debin Zhao, "A ROI quality adjustable rate control scheme for low bitrate video coding," Picture Coding Symposium, 2009. PCS 2009, pp. 1-4, May 2009.

[3] Bulla, Christopher, Christian Feldmann, and Martin Schink, "Region of Interest Encoding in Video Conference Systems," The Fifth International Conferences on Advances in Multimedia, pp. 119-124, Apr 2013.

[4] Yang Liu, Zheng Guo Li, Yeng Chai Soh, "Region-of-Interest Based Resource Allocation for Conversational Video Communication of H.264/AVC," Circuits and Systems for Video Technology, IEEE Transactions on, vol.18, no.1, pp. 134-139, Jan 2008.

[5] Sullivan, G.J., Wiegand, T., "Rate-distortion optimization for video compression," Signal Processing Magazine, IEEE, vol.15, no.6, pp. 74-90, Nov 1998.

[6] Yongjun Chang, Munchurl Kim, "A Joint Rate Control Scheme in a Hybrid Stereoscopic Video Codec System for 3DTV Broadcasting," Broadcasting, IEEE Transactions on, vol.59, no.2, pp. 265-280, Jun 2013.

[7] Murshed, M.M., Siddique, M.A.R., Islam, S., Ali, M., Lu, G., Villanueva, E.V., Brown, "High Quality Region-of-Interest Coding for Video Conferencing based Remote General Practitioner Training," The Fifth International Conference on eHealth, Telemedicine, and Social Medicine, pp. 240-245, Feb 2013.A.