符号化方式に依存しないテレビ会議向け関心領域の映像処理技術

CODEC-free Region of Interest Video Processing Technology for Video Conference Systems

リエン リュウ* ショウモン ワン* ウェイトウ ゴン*
Liyan Liu Xiaomeng Wang Weitao Gong

要　旨

現在，ビデオによる相互通信，例えばビデオ会議などが重要な役割を果たしている．しかし，ネットワークの通信帯域の制限により，画像がはっきり見えなくなるなどの現象を生じることがある．関心領域（ROI）に基づくビデオ処理技術は人間の視覚システム（HSV）を利用して，ユーザーの関心のある範囲を最重視することにより，このような問題を解決する．

本論文では，符号化方式に依存しないROI処理方法を2つ提案する．1つはフィルターに基づくROI処理方法，もう1つはマルチストリームに基づく ROI処理方法であり，どちらもビデオデータの送信量を削減するために，背景領域の品質を犠牲にして，ROI領域の品質を維持する．

可変ビットレート(VBR)と固定ビットレート(CBR)の両方の場合について，それぞれを評価した．その結果，通常のデータ送信量に比べると，本方法はVBRの場合帯域はほぼ40%削減でき，CBRの場合はROI領域の品質は最大2dB以上の向上が見られる．

また，無線環境下でのダイナミックな帯域の変化に対して，本方法を評価した．模擬ネットワーク環境下での評価結果により，当該技術の実用性が示された．

Abstract

Nowadays, conversational video applications, such as video conferencing, have played more and more important role in daily communications. Viewers of these applications, however, may suffer from unclear or jittered video due to restriction of available network bandwidth. Region of interest (ROI) based video processing technology, which utilizes characteristics of human visual system (HSV), by paying more attention to viewers' focusing areas, is of practical use for solving such problems.

In this paper, we propose two ROI-based CODEC-free video processing approaches, which are Filter based ROI Video Processing and Multi-stream based ROI Video Processing, with both preserving quality of ROI area and sacrificing quality of background area, in order to reduce video data transmission volume.

We evaluate each approach by both variable bit rate coding (VBR) and constant bit rate coding (CBR). In our evaluation, compared to "uniform coding" method, our proposed approaches can reduce around 40% of bandwidth consumption in VBR case, or obtain a maximum of more than 2dB increase in quality of ROI area in CBR case.

We also adapt and evaluate the proposed ROI-based approaches for dynamic bandwidth situations in wireless network environments. The evaluation results in a simulated network environment prove the feasibility of this technology in practical use.

* リコーソフトウェア研究所（北京）有限公司
Ricoh Software Research Center(Beijing) Co.,Ltd.

1. Introduction

Nowadays, demands for applications of the digital video communication, such as video conferencing, have increased considerably. However, due to restriction of network bandwidth, sometimes video would be encoded at very low bit rate before transmission, which makes viewers suffer from degradation of video quality, like block effects, jittered video, etc. Although many standards have been proposed and evolved for improving coding efficiency, most implementations adopt "uniform coding" method, which gives equal importance to each block of video frame regardless of its relative importance to the human visual system (HVS).

To address this problem, Region of Interest (ROI) coding was proposed, by which one or more interesting areas in each frame are defined and encoded in priority to preserve quality of ROI area, while quality of other areas are sacrificed to reduce bandwidth consumption. The rationale behind ROI-based video coding relies on the highly non-uniform distribution of photoreceptors on the human retina, by which only a small region of 2–5 visual angles (the fovea) around the center of gaze is captured at high resolution, with logarithmic resolution falloff with eccentricity [1]. Thus, it may not be necessary or useful to encode each video frame with equal quality, since human observers will crisply perceive only a very small fraction of each frame, dependent upon their current point of fixation.

Generally, approaches of ROI coding can be divided into two categories: CODEC free[2][3][4][5][6] and CODEC dependent [7][8][9]. The former precedes encoding stage and can be pipelined with any coding standards, while the latter has closer link with CODEC implementation and usually focuses on quantizer parameter (QP) tuning. Although QP tuning can offer more precise control on video quality, in this paper, our proposed approaches belong to the CODEC-free category because of its flexibility and universality.

We conduct trials on both filter based ROI processing and multi stream based ROI processing, which can reduce bandwidth consumption in variable bit rates (VBR) situation or improve quality of ROI area in constant bit rates (CBR) situation, compared to traditional uniform coding method.

The rest of this paper is organized as follows. Section II gives detailed description of our ROI processing approaches. Section III presents our experimental results. Conclusions are given in Section IV.

2. ROI Processing Approaches

2-1 Applying ROI processing in video conference scenario

A general flow of CODEC free ROI processing is described in Fig.1.

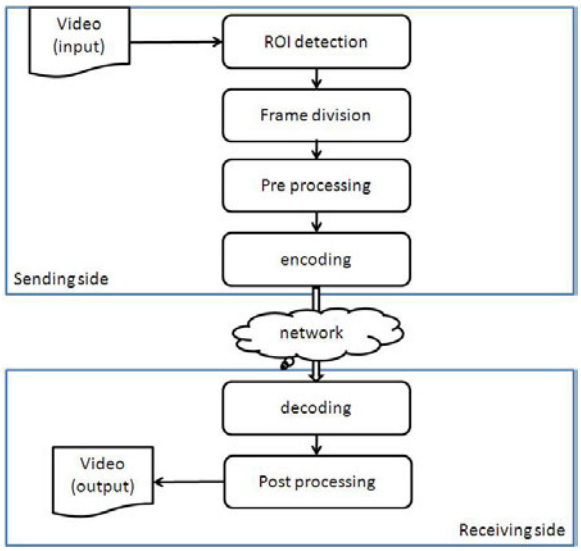


Fig.1 Procedure of applying ROI processing in video conferencing scenario.

In this procedure, ROI detection method is firstly applied to detect interesting area within one video frame following the policy of ROI definition, which is up to the requirements of applications. For example, in a video conferencing scenario, a speaker who is making a

presentation attracts attention from all attendees so the speaker becomes the focus of the scene. Thus, speaker detection or human detection technology is the option to detect the ROI area.

Once the ROI area is detected, the video frame can be divided into two parts: ROI area and non-ROI area, or foreground and background. In our later descriptions, we will not differentiate the two groups. As the core idea of ROI processing is to keep high quality of interesting area and sacrifice quality of the background area, obvious subjective difference can be perceived between these two portions, as shown in Fig.2.

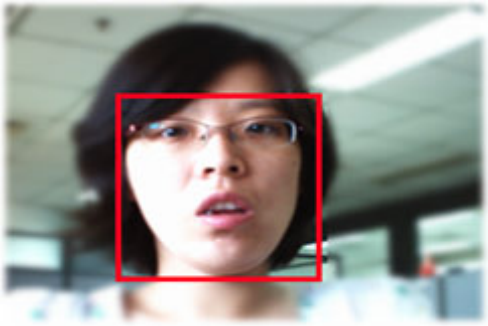


Fig.2 Quality difference between ROI area (red box) and background.

To alleviate such drastic degradation, a transitioning area is introduced between ROI area and non-ROI area – extended region of interest area (X-ROI). It is produced through extending the border of ROI area outward with a predefined distance. Then pre-processing is conducted prior to encoding step. We have two trials in our research: filter based and multi stream based region of interest processing. After that the ROI coded video is encoded and transmitted to the other end. At receiving side, a post processing step is added after the decoding step, though it is optional for filter based approach.

After this brief introduction of ROI processing in video conferencing scenario, next we will focus on the two proposed ROI processing approaches.

2-2 Filter based ROI processing

Filter based Region of Interest processing can be done either spatially[3][5] or temporally[8], or in a hybrid mode[6]. The main purpose of this approach is illustrated as follows:

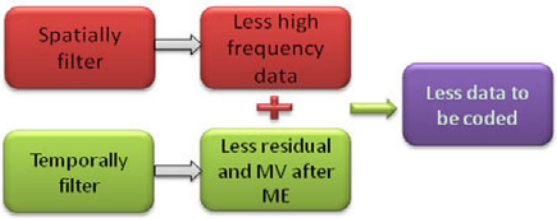


Fig.3 Idea of filtering in ROI processing.
(MV：motion vector; ME: motion estimation)

- Spatial filtering

X-ROI area and non-ROI area are blurred spatially through low-pass filter. By this way, high frequency information is greatly removed from the picture, which results in more zero (high frequency coefficients) in DCT-transformed matrix so less bit rates are needed for later encoding[10](Fig.4).

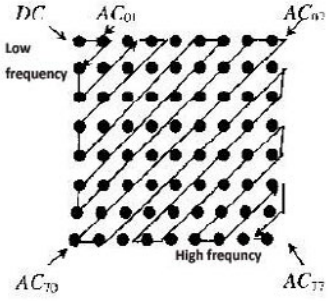


Fig. 4 DCT-transformed matrix.

Different filters can be used here, such as mean filter or Gaussian filter. To smooth transitioning from ROI area to non-ROI area, parameters of the filters are tuned while being applied to X-ROI area and non-ROI area, with the former less blurred than the latter to get gradual quality degradation.

- Temporal filtering

Temporal filter functions similar to spatial filter with the purpose of reducing data to be encoded. Due to continuity of video frames, especially in video conferencing scenario, usually changes between two successive frames at background part are too minor to be perceived, which provides us the chance to do filtering temporally. The simplest way is background skipping. For example, in every two frames only background of the odd frame is preserved, while background of the even frame is skipped. In other words, two successive frames share one background. However, mismatch between ROI area of the current frame and background of the previous frame would occur sometimes because of motions of some objects in the scene. Linear interpolation method is introduced to counter this issue. It is illustrated by following formula:

$$I_i(x,y) = \begin{cases} I_i(x,y), \text{ if } (x,y) \in \text{ROI} \\ (I_i(x,y) + I_{i-1}(x,y)) / 2, \text{ if } (x,y) \in \text{x-ROI} \\ I_{i-1}(x,y), \text{ if } (x,y) \in \text{non-ROI} \end{cases} \quad (1)$$

$I_i(x,y)$: pixel value of (x,y) in i^{th} frame

Either background sharing or background interpolation utilizes a feature of the video coding: motion estimation and motion compensation. For non-key frames (P or B frame), only difference with previous frame is considered for encoding[10](Fig.5). Reduction in difference between two adjacent frames helps significantly in bitrates saving as well.

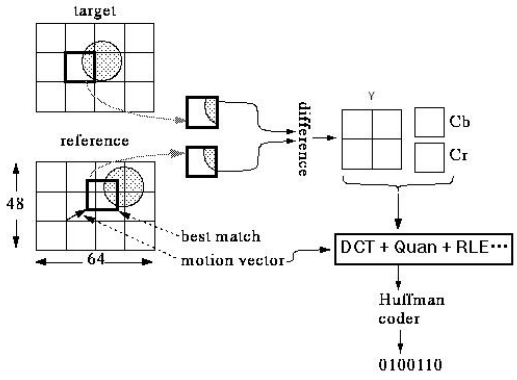


Fig.5 Illustration of motion estimation between two adjacent frames.

2-3 Multi stream based ROI processing

Another approach proposed in our research is to separate one video stream to two or more streams for later ROI processing. The main idea of this approach is shown below(Fig.6). It involves both pre-processing stage and post processing stage.

- Pre-processing

After being detected in video frame, the interesting area and its extension are extracted from original frame to form "ROI stream", and the remaining part becomes "background stream". Separate processing methods are then applied to these two parts.

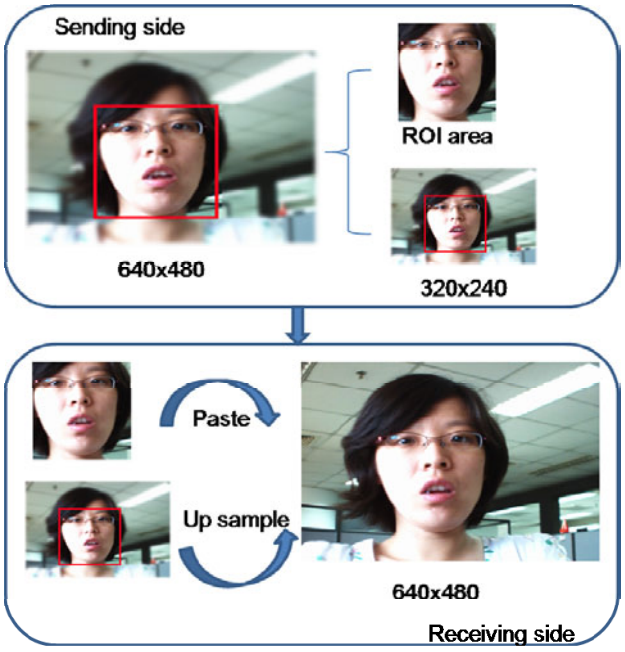


Fig.6 Idea of multi stream based ROI processing approach.

The size of each frame in "ROI stream" may vary due to changes of ROI area, and this does not conform to rules of coding (every frame in one stream must be of constant size). Consequently, an additional step is necessary before encoding ROI area and X-ROI area. For each ROI frame, a monochroic image with identical size of original frame is prepared and the ROI area and X-ROI area are put on this image at same position in original

frame, as shown in Fig.7. So all frames in ROI stream are "padded" to have an equal size and can be encoded and transmitted.

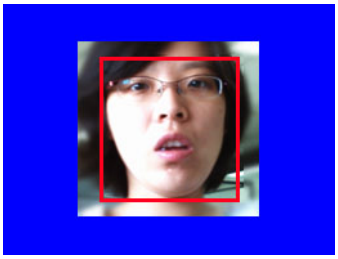


Fig.7 "Padded" ROI frame on blue image.

As the two streams are separated, they are independent to each other. So the processing for background is different from that of ROI part. The background frame sent through "background stream" can be down sampled to a smaller size. For example, if both x direction and y direction are down sampled at a scale of 1/2, then 3/4 data are removed from original background frame. Meanwhile, interpolation based down sampling can remove some high frequency information. All of this helps considerably in reducing bandwidth consumption.

- Post-processing

At receiving side, the two streams are decoded separately, which generates a ROI frame and a background frame. A composition operation is needed to restore the whole video frame before displaying. This is a reverse process compared with that in pre-processing stage. Firstly, ROI area and X-ROI area are detected and extracted from the ROI frame. The background frame is up sampled to restore to its original size. Secondly, the extracted ROI and X-ROI portion should be put back to their original positions on the up-sampled background. However, direct replacement may lead to mismatch on the border between the background and X-ROI area. This is caused by the error introduced in the encoding stage if bit rates are very low, because motion difference between frames is not precisely encoded due to bandwidth shortage. This is illustrated in Fig.8. To remove possible mismatch, a "matching" process is inserted to find best position to put back ROI and X-ROI area on the background.

Fig.8 "mismatch" between foreground (red box) and background.

Furthermore, to smooth the quality degradation from ROI area to background, the overlap area between background and X-ROI area is updated by interpolation which is similar to that in "temporal filtering".

2-4 Adaptive coding

In some cases, video communication is done over wireless network with dynamic bandwidth. To adapt our proposed approaches to such network environments, a real time encoding control is added based on available bandwidth. Each frame is coded under restriction of present usable bit rates to avoid potential packet loss, which may lead to distorted frames. By doing so, we could preserve quality of interesting area and reduce its quality fluctuation.

3. Experimental results

3-1 Test data

The test video used here is of resolution 1024x768. Face detection method is used here for ROI detection and ROI area varies in coverage from 0 (no face detected) to more than 30%. Three video clips are extracted from it with each clip containing relative constant ROI coverage. Details are shown in Table 1:

Table 1 Three Test video clips.

Video clip No.	Frame count	ROI coverage
1	20	2.43%
2	32	20.51%
3	33	28.55%

Each video clip is encoded by three coding methods in both VBR and CBR cases:

- Uniform coding
- Filter based approach: Gaussian filter with kernel value being 7
- Multi stream based approach: with background scaling to 1/4 size of original frame

3-2 Experimental results

In the case of VBR coding, bandwidth consumption is measured for evaluation. ROI processing approaches are expected to reduce consumption of bandwidth after encoding. In the case of CBR coding, available bandwidth is set to be constant and the quality of ROI area becomes a measurement for different methods. Under this condition, ROI processing approaches should produce video clips with higher quality for interesting area compared to the uniform coding approach.

Table 2 shows experimental results in both VBR and CBR cases on the video clip with ROI coverage of 2.43%.

The results show that both filter based approach and multi stream based approach perform better than traditional uniform coding. In VBR situation, bandwidth consumed by the video clip after encoding is compared, and 22% and 43% of bandwidth consumption are reduced respectively; while in CBR case, PSNR value of encoded video clip is calculated. ROI processing approaches improve the quality of interesting area by 1.11dB and 2.63dB respectively, which is in accordance with our expectations.

The results also indicate that multi stream based ROI processing approach can achieve more bitrates gain compared with filter based approach, though it is more sensitive to ROI coverage in the video frame. With ROI area accounts for more in video frame, multi stream based approach gradually loses its advantage over filter based approach. Table 3 and Fig.9 show this trend.

Another point to be considered in multi stream based approach is the bitrates allocation policy between ROI stream and background stream. Different proportions are tried in our experiments. If more bitrates are allocated to ROI stream, the quality of ROI area gets higher at sacrifice of worse quality of background area.

Table 2 Experimental results on the video clip with ROI coverage of 2.43%.

	VBR	CBR (384KB)
Uniform coding	617KB	40.65dB
Filter based approach	484KB (78%)*	41.76dB (1.11dB ↑)**
Multi stream based approach	356KB (57%)	43.28dB (2.63dB ↑)

*: the percentage is calculated by comparing with result of uniform coding

**: the difference of PSNR value is calculated by comparing with result of uniform coding

Table 3 Bandwidth consumption comparison between video clips with different ROI coverage.

	Clip 1	Clip 2	Clip 3
Uniform coding	617KB	1016KB	1076KB
Filter based approach	484KB (78%)*	847KB(83%)	921KB(85%)
Multi stream based approach	356KB (57%)	817KB(80%)	920KB(85%)

However there is an upper limit for bitrates allocated to ROI stream, beyond which quality of ROI area remains almost constant.

Table 4 gives results of video quality of ROI area and background area under different bitrates allocation proportion on three video clips. For video clip with ROI coverage of 2.43%, the quality of ROI area reaches the peak at proportion of 1:3 (background area: ROI area). Even if more bitrates are allocated, quality of ROI area remains unchanged. However, with ROI coverage increasing in video frame, the "upper limit" increases as well. So with size of ROI area varying in video frame, the proportion between two parts should be adjusted to reach a balance between ROI area and background area so as to make full use of available bandwidth.

To evaluate ROI processing in variable network situations, adaptive coding is conducted under a simulated network environment. The simulation is done by NS2[11], assuming 22 applications in the environment starting and stopping randomly and repeatedly in 100 seconds. Fig.10-a shows the simulated result. To match duration of our test video clip (around 50s), a segment of the simulated result is extracted (red box, 24ths -73rds) and shown in Fig.10-b.

As mentioned earlier, in this dynamic network environment, we hope not only generate higher quality of ROI area than uniform coding method, but also decrease the influence by fluctuation of available bandwidth and keep the quality as stable as possible. This is to be realized by adjusting parameters of ROI processing approaches. In the case of filter based approach, the kernel value of Gaussian filter is the only tunable parameter; while in multi stream based approach this is done through adjusting bitrates allocation between ROI area and background area.

With available bandwidth decreasing, we try to keep

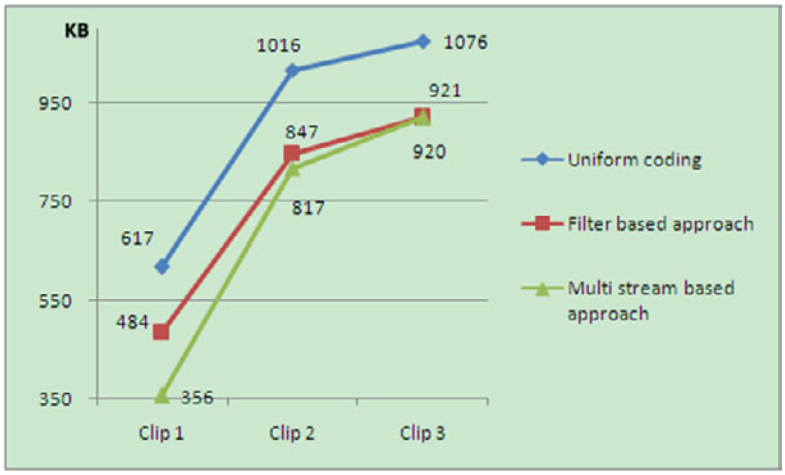


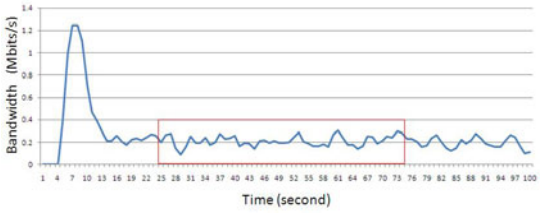
Fig.9 Bandwidth consumption comparison by three methods on three video clips.

Table 4 Video quality under different bitrates allocation proportion in multi stream based approach. (CBR, 384KB) (BG: background, ROI: region of interest)

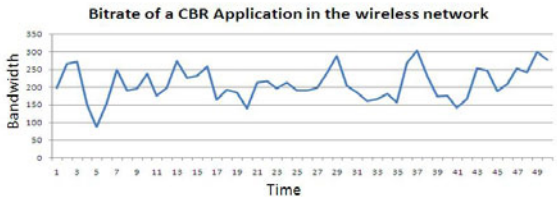
Video clip / Bitrates allocation (BG:ROI)	1 (2.43%)		2 (20.51%)		3 (28.55%)	
	ROI part	BG part	ROI part	BG part	ROI part	BG part
1:1	43.08dB	37.75 dB	37.72 dB	37.83 dB	38.69 dB	37.40 dB
1:3	43.28 dB	36.28 dB	39.20 dB	36.74 dB	41.08 dB	35.75 dB
1:5	43.28 dB	35.61 dB	39.68 dB	35.78 dB	41.31 dB	34.29 dB

stable quality of ROI area by increasing the kernel value of Gaussian filter (hopefully, a larger kernel value removes more information of background area, so less bitrates are needed for encoding background part). However, experimental results prove no feasibility of this method (Table 5). With bandwidth down from 1200kbps to 300 kbps and the kernel value growing up from 5 to 13 correspondingly, the quality of ROI area still gets lower by 1.28dB, which does not behave as our expectation (This relates to the characteristics of video conferencing scenario, where background is relatively simple and contains little high frequency information. So even if we increase kernel value, no more bitrates can be reduced from background part and reallocated to ROI part)

But in case of multi stream based approach, due to separation of ROI stream and background stream, much



(a)



(b)

Fig.10

(a) simulated bandwidth variation by NS2 with duration of 100s.

(b) extracted segment of (a) with duration of 50s.

Table 5 Quality of ROI area under different bitrates by filter based approach. (dB)

bitrates	Gaussian filter kernel value			
	5	7	9	13
300Kbit/s	26.51	26.59	26.59	26.53
600Kbit/s	27.33	27.39	27.42	27.44
900Kbit/s	27.62	27.65	27.67	27.69
1200Kbit/s	27.81	27.86	27.88	27.90

Table 6 Quality of ROI area under different bitrates by multi stream based approach. (dB)

(a) Quality of ROI area set to constant, remaining bandwidth allocated to background.

bitrates	Quality of ROI area	Quality of background area
300Kbit/s	27.52	21.09
600Kbit/s	27.52	22.19
900Kbit/s	27.52	22.34
1200Kbit/s	27.52	22.34

(b) ROI area encoded by “best” effort, remaining bandwidth allocated to background.

bitrates	Quality of ROI area	Quality of background area
300Kbit/s	27.52	21.09
600Kbit/s	27.97	21.70
900Kbit/s	28.14	22.06
1200Kbit/s	28.20	22.19

more flexibility is provided in bitrates controlling. Table 6 shows two kinds of policies in bitrates allocation: the quality of ROI area can remain unchanged by setting to a constant or ROI part is always encoded by "best" effort, and then the background part is encoded with the remaining bandwidth. Comparing between results by filter based method and multi stream based method, with bandwidth down from 1200kbps to 300kbps, the quality of ROI area by former decreases from 27.81dB to 26.53dB (1.28dB↓); while for latter, the value is either of no change (as constant as 27.52dB, Table 6(a)) or from 28.20dB to 27.52dB(0.68dB↓,Table 6(b)). This indicates better adaptability of multi stream based approach. As a result, this approach is selected to be applied to dynamic network situation. Table 7 shows the mean value and the standard deviation value of the quality of ROI area in simulated network environments. And Fig.11 illustrates results of the actual bandwidth consumption. Most bandwidth is allocated to ROI area to guarantee its quality and the remaining bandwidth is allocated to the background stream.

Table 7 Quality of ROI area by multi stream based approach in simulated network environment.

	Mean value	Standard deviation value
Adaptive multi stream based approach	28.14dB(↑)	0.08dB (↓)
Uniform coding	26.98dB	0.19dB

4. Conclusions and Future Work

In this paper, two CODEC-free ROI processing approaches are presented: filter based and multi stream based approach. The former is a pre-processing step prior to encoding stage with background being blurred by filters; and the latter covers both pre-processing and post processing stages, with ROI area and background area being separated for different processing. The two approaches can be combined with any standard encoder and decoder because it is independent of any concrete implementation of them. We evaluate the two proposed methods in VBR and CBR situations and the results show advantages over traditional uniform coding method.

In dynamic network situations, the multi stream based approach proves its feasibility. Due to independency between background and ROI area, it offers more flexibility than filter based approach with free bitrates allocation.

Consequently, in future, the intelligent bitrates allocation between ROI and background area is to be studied to make full use of available bandwidth. Furthermore, if more than one ROI areas exist in vide frame, prioritized encoding and transmitting can also be studied on basis of independency feature of this approach. And because the quality of background area is sacrificed in pre-processing stage, video enhancement techniques can be applied to restore the quality of the background part.

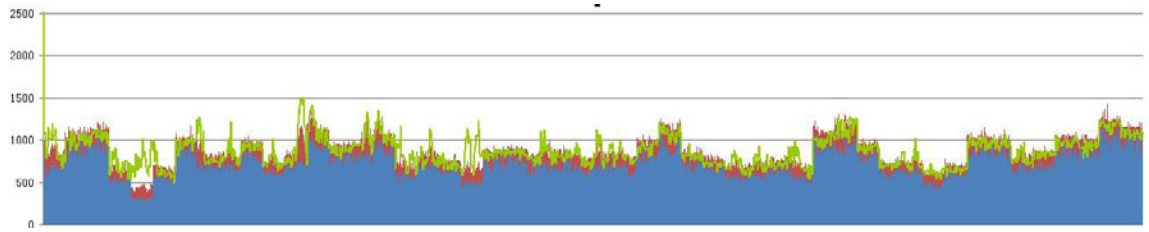


Fig.11 Actual bandwidth consumption in simulated network environment
– ROI area – background area – uniform coding.

Reference

1) B.Wandell : Foundations of Vision. 1st edition, Sinauer Associates, (1995).
2) Chen et al.: Using a region based blurring method and bits reallocation to enhance quality on face region in very low bitrate video, Proc. of the 1998 IEEE Int. Symp. on Circuits and Systems, vol. 4, (1998), pp. 134-137.
3) Chen et al.: ROI video coding based on H.263+ with robust skin-Color detection technique, IEEE Transactions on Consumer Electronics, (2003), pp. 724-730.
4) Cavallaro, A. et al: Perceptual prefiltering for video coding, ISIMP'04, (2004), pp. 510-513.
5) Nicolas Tsapatsoulis et al.: Visual attention based region of interest coding for video-telephony applications, 5th International Symposium on Communication Systems, Networks and Digital Signal Processing, (2006).
6) Linda S. Karlsson: Spatio-temporal filter for ROI video coding, (2006).
7) Chung-Ming Huang et al.: Multiple priority region of interest h.264 video compression using constraint variable bitrate control for video surveillance, Optical Engineering, vol. 48, issue 4, (2009), pp. 47004-47005.
8) Haohong Wang et al: Real time region of interest video coding using content-adaptive background skipping with dynamic bit reallocation, ICASSP'06, (2006), pp. 45-48
9) Yang Liu et al: Region of interest based resource allocation for conversational video communication of h.264/avc, IEEE transactions on circuits and systems for video technology, Vol. 18, No. 1, (2008), pp. 134-139
10) Iain E. G. Richardson: Video CODEC design, Wiley, (2002).
11) Network simulator – NS2, http://www.isi.edu/nsnam/ns/