Face Tracking for Optimized Bitrate Control in Low Delay Video Encoding

Chethan Ningaraju

September 9, 2016

Contents

1 Introduction

1.1 Low Delay Bitrate Control

In recent years there is increasing demand for high quality video conferencing solutions. To address this growing need there has been constant improvement in low delay video coding techniques. The need for extremely low end to end delay in video telephony puts additional constraints on video coding which results in compromise of video quality.

The bitrate control module is responsible for controlling the bit-consumption of the encoder to guarantee smooth playback. Bitrate control module is not codec specific and operates independent of any chosen codec. The main purpose of the bitrate control module is to ensure efficient playback of the encoded video. Figure 1 shows the functionality the bitrate control module. It achieves this by controlling the quantization parameter using during the encoding. The decision of quantization parameter is done considering the input bitrate, framerate, input complexity (spatial and temporal activity), acceptable input delay (a measure of VBV buffer size). The module also takes the feedback from the encoder regularly to make better decision. During the low delay video encoding, tools like bi-directional prediction are disabled.

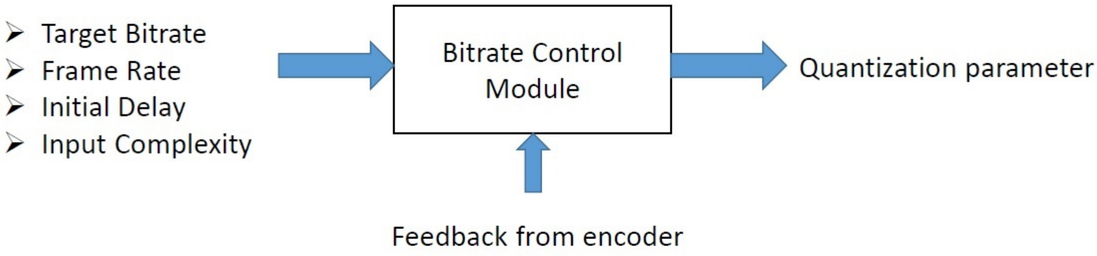

Figure 1: Bitrate Control Module Functionality

1.2 ROI based coding

In conventional video coding, all regions of the frame is considered equally important. It is assumed all regions contribute equally for perceptual quality. Some video codecs use the fact that high frequency components are less important to human visual system and do you preferential coding based on spatial frequency. However, such prefential coding does not take into account the contents the frame to be encoded.

Region of interest based coding is not a common practice in video coding because it is very hard to automatically detect important regions that contribute the most to perceptual quality. However, region of interest in video conferencing content is going to be face region predominantly. Due to recent improvements in face detection algorithms it is possible to detect face with good accuracy. The study in [1] shows how boosting quality of face regions can improve the overall preceived quality of the video. This work aims to study possible ways of improving preceptual quality of the video by detecting face regions and coding it with higher quality than rest of the frame.

In order to ensure real time video communication using standard video codecs (e.g. H.264/AVC) a highly flexible bitrate-control mechanism has to be utilized. This means that the encoder's quantization parameter QP is adapted on a macroblock level in order to ensure that the maximum allowed bitcount for an encoded frame is not exceeded. In a simple approach one would try to distribute the bitrate evenly on every macroblock. Since load efficiency is of high importance it is not advisable to do multipass encoding for optimal bitrate allocation. Therefore over-allocation in one macroblock has to be compensated by under-allocation (using a higher QP) for neighboring macroblocks, regardless of the image content. However, a more intelligent allocation strategy should take the image content into account. Thus parts of the image with higher importance (e.g.faces) should be given a higher percentage of the overall bitcount which results in higher visual quality. Background regions would get a lower proportion of the bitcount.

The goal of this work is to identify the salient region of a frame, which is the face of the participant in a video conference. In a first iteration we assume ideal capture conditions so that the results of the face tracking will be directly used as side information for the H264/AVC encoder's bitrate-control. Face regions should allocate an above-average bitcount and yield a better visual quality than background regions. It is also the aim of this work to develop and extensively evaluate the strategy of uneven bitrate allocation and also to identify its limitations.

2 Study setup

2.1 codec configuration

This work uses Citrix h264 video codec for the study. The encoder is configured in low delay mode sutiable for video conferencing. The encoder is configured to use IPPP mode with intra/key frames encoded only at the beginning of the sequence. Due to low delay there is no provision to re-encode the frame in case of buffer overflow. The frames are skipped entirely in case of buffer overflow to guarantee smoother playback by maintaining the delay constant.

The quantization parameter is adapted at every macro-block level to meet the overall bitrate precisely and avoid frame drops.

2.2 Measurements

The most crucial aspect in this proposal is the metric used to evaluate various algorithms and choosing the best one. The goal of this proposal is to improve the quality of ROI region at the cost of non-ROI regions. The goal is to keep the bitrate constant. Since the whole approach is to measure the gain in perceptual quality, using frame level PSNR as a metric could be misleading. In this proposal, the metric used is average PSNR of frames and it is compared with average ROI PSNR. The expectation is to see an improvement in ROI level PSNR, with degradation in PSNR of the non-ROI regions.

In order to evaluate different algorithms, we consider different metrics. The gain of PSNR in visual ROI region and non- ROI region is measured as average PSNR of full frame and average PSNR of ROI region. This will also help in measuring the aggressiveness of an algorithm. The goal is to not make background really bad compared to the foreground.

2.2.1 Quality metrics

The initial approach is to find the gain in PSNR in ROI, however finding the desirable extent of improvement in PSNR in ROI along with acceptable drop in PSNR of non-ROI is tricky. The idea here is to find the right balance between quality improvements in ROI with degradation of non-ROI region as to achieve maximum perceptual quality. For the sample video chosen, following was the PSNR measurements. The values in 1 shows the PSNR values The PSNR is calculated using weighted sum of PSNR of individual components per picture (PSNR$_{\mathrm{Y}}$, PSNR$_{\mathrm{U}}$ and PSNR$_{\mathrm{V}}$) [2].

Content	PSNR Avg (dB)	PSNR ROI (dB)
Paul640x480, 250kbps	39.22	37.54
Johny1280x720 750kbps	40.90	39.20

Table 1: Initial PSNR values

$$PSNR_{\mathrm{YUV}} = (6.PSNR_{\mathrm{Y}} + PSNR_{\mathrm{U}} + PSNR_{\mathrm{V}})/8 \tag{1}$$

where individual components are computes as

$$PSNR = 10.log_{10}((2^B - 1)^2/MSE) \tag{2}$$

wher B = 8 is the number of bits per sample of the video and MSE isthe mean squared error.

2.2.2 PSNP and QP variation

The second aspect of measurement in this kind of video is to consider the PSNR and QP distribution within the frame. These are represented as a gray scale image. For a QP scale map, the darker regions in a frame indicate higher quantization. Even though quantization parameter used for a block to be encoded is closely related to the PSNR of the block, it is not the only deciding factor. The PSNR can also vary depending on the content. Therefore, the PSNR distribution within a frame is also represented as gray scale image.

The image in figure 2 shows a frame in the sample video conferencing content. The image in figure 3 shows the same frame when encoded with the codec configurations discussed in ¡section¿ with 250 kbps. The reason for considering a low bitrate of 250

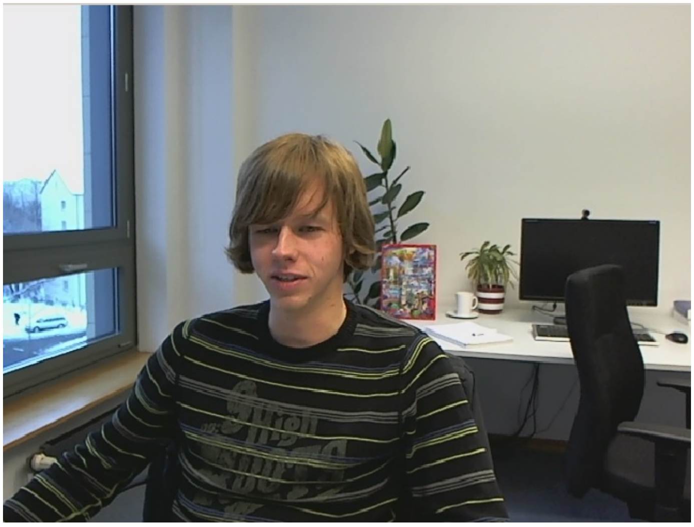

Figure 2: A Frame in the sample video

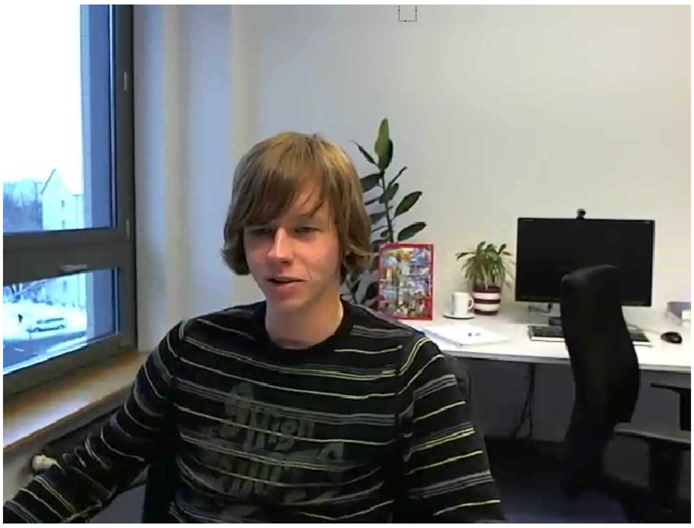

Figure 3: Sample frame in encoded video (250kbps)

kbps is that it will help in making the improvement in face region and decrease in quality of background more evident and hence will be useful in evaluating different algorithms.

The image in figure 4 shows the quant map of the frame in figure 3. The darker regions in this map indicate usage of higher quantization parameter compared to the lighter regions. It can be noticed that since no information about region of interest is used while encoding the frame, the pattern of quantization appears almost normal. The shape of the original content is almost not recognizable from the quantization map.

The image in figure 5 shows the PSNR for the frame in figure3. This map is identical to quantization map, the lighter regions here represent the regions with higher PSNR,

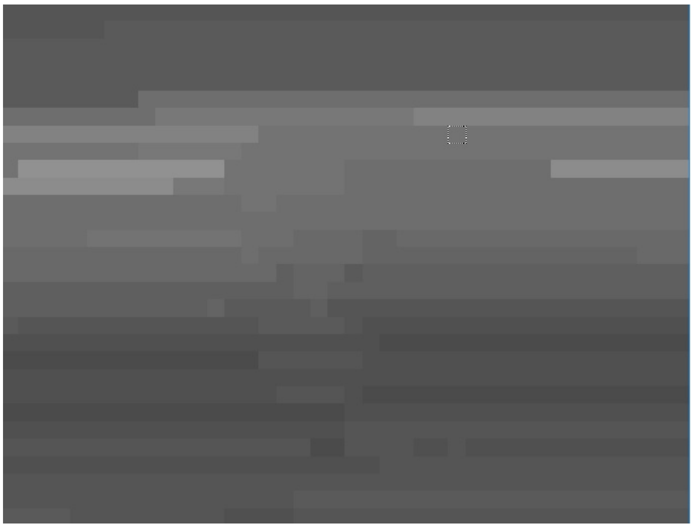

Figure 4: Quantization map

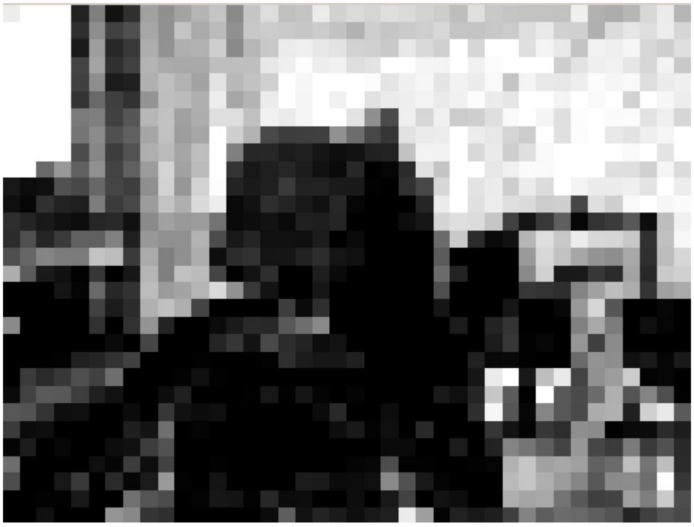

Figure 5: Relative PSNR map

the darker regions indicate lower PSNR hence worser quality. This map is relative within the frame and does not represent absolute quality. This map is generated by considering the full range of PSNR of the image after removal of outliers. The map is generated by mapping the PSNR range between 10th percentile and 90th percentile of the whole frame to value between 0 to 255.

It is evident that shape of the original content is preserved in the PSNR map. The background regions have better PSNR, the foreground has worser quality and the difference in quality is quite huge. The difference in the quality is due to the fact that, background in a video conferencing content is mostly static and hence gets encoded better with every frame. On the other hand, the foregound has motion and new data to be encoded, and hence it cannot achieve the same quality as background. Since the focus of attention during video telephony is foreground or the face region, improving the face region must help in improving overall perceptual quality. The effect on such preferential encoding is studied in this work. The idea is to reduce the PSNR difference between foreground and background.

2.2.3 Bitrate fluctuation - Delay plots

The core idea of this proposal is to efficiently use the bits within the frame to encode region of interest. The algorithms used to achieve that purpose should not alter the overall behavior of the codec in terms of bit consumption. As mentioned in ¡section ¿ the codec skips the frame in order to maintain strict VBV buffer compliance. Skip frames result in jerky playback and hence should be avoided as much as possible. The intelligent bit-allocation scheme should not contribute to more skip frames if not reduce them.

A plot of measuring the delay of each frame is used to verify this. Figure 6 is the delay plot of bitstream encoded with 250 kbps (a sample frame in figure 3). Every point in the plot specifies the time taken by the corresponding frame of x-axis to reach the decoder assuming zero transmission delay. The cure appears moslty smooth except for sudden drops (zero values). These zero value points indicate skip frames. Since these frames are not included in the final bitstream and hence not transmitted, the delay is indicated as zero. Ideally, an algorithm with intelligent bit allocation within a frame should not alter the shape of this graph. It is also desirable to not have any increase

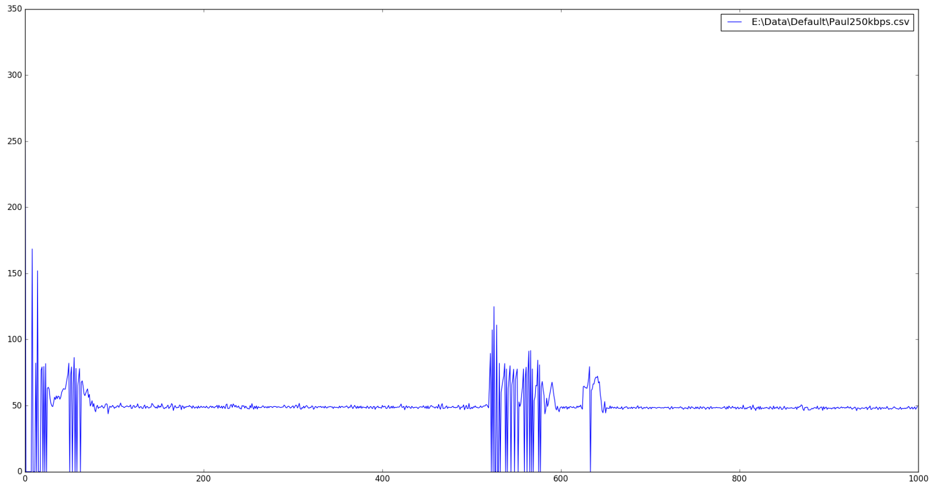

Figure 6: Delay plot

in the number of skip frames. Since, the algorithm is not expected to change overall behavior it is not expected to have reduction in number of skip frames.

3 Face Detection

Face detection algorithms are used to mark the region of interest in the current frame. All the algorithms considered for intelligent bit allocation involve improving the region of interest at the cost of rest of the frame. Therefore it is very important to have high accuracy with face detection. Any false detection will lead to degradation of the region of interest compared to normal encoding, this should be avoided. The damage cause by false detection is higher than the loss due due to not detecting any face. Therefore, a high threshold must be used to declare any region of the frame as face.

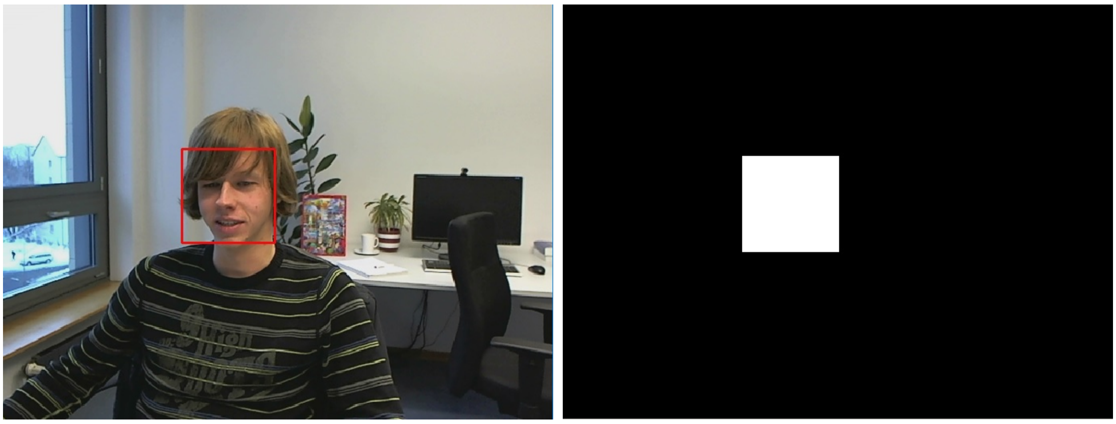

Figure 7: Face Detection binary map

The face detection module itself shall not be part of encoder, but the input from face detection is a binary file with face regions marked as input. In this work ,the face detection module input YUV to the encoder and marks the region of interest at mack block level. Each byte value represents a macro-block scanned in raster scan order. A value of 0xff signifies macro-block being part of the face or region of interest and 0x00 represents a normal macro-block. Figure 7 represents the face map generated for the frame shown in 2. The region in white is considered region of interest, this information is used inside the bitrate control module of the encoder to perform a intelligent bit allocation.

Different approaches are used to detect the face. There is always a tradeoff between accuracy in face detection and complexity. The work presented here is mostly relevant to real time systems. Any added complexity due to additional module of face detection will cause significant delay which is totally unacceptable. Therefore the algorithm chosen for face detection must be light weight and reasonably accurate in all lighting conditions.

3.1 Spatial Domain Face Detection

The face detection algorithm works directly on pixels. This approach is simple in terms of implementation. Many open-source solutions like OpenCV offers a ready to use solutions that can be integrated with the codec library. It has large set of trained classifiers considering many types of faces and angles. However, this is very computation intensive and almost impractical to use in the final solution.

3.2 Compressed domain Face Detection

TBD LATER

4 ROI based intelligent bitrate control - Approaches

There are many ways of using the additional information of knowing the region of interest. These methods should increase the quality of the frame to gain maximum perceptual quality.

4.1 QP offset

The simplest and straight-forward way of creating a bias in quality for ROI and non-ROI is by using a QP offset between ROI based regions in the existing bitrate control module. The feedback from the encoder to rate control will ensure that final bitrate is still met. A negative QP offset is used for ROI regions, which triggers increase in QP of non-ROI regions.

Such QP offset will ensure the quality difference and shall not be overruled by bitrate control mechanisms. However, this approach might result in bitrate control over-reacting for the blocks around the ROI and encode them with very low quality. The magnitude of QP offset shall dictate the magnitude of shift in quality for the ROI. This approach is used to find right QP offset for best perceptual quality.

It is perhaps a better idea to link the confidence quotient of face detection algorithms with the QP offset used. It should also be dependent on the area of the face, if the area of the ROI is considerably large in a video then the quality difference should be minimized.

The images in 8 shows the comparison between original images, PSNR, Quant map and corresponding attributes after encoding the ROI with a better lower QP of 4. The overall bitrate of the streams were almost constant, the result is only due to movement in the bits. The number of dropped or skip frames were found to be the same. Figure9 shows the comparison of delay of original bitstream with the bitstream with QP offset for ROI. It can be seen that the delay behavior does not change significantly with skip/dropped frames at same locations.

4.2 Reaction Factor - Buffer control

The second approach of using ROI information to enhance quality of ROI is by using different buffer controls in the inside the birate control module. The bitrate control

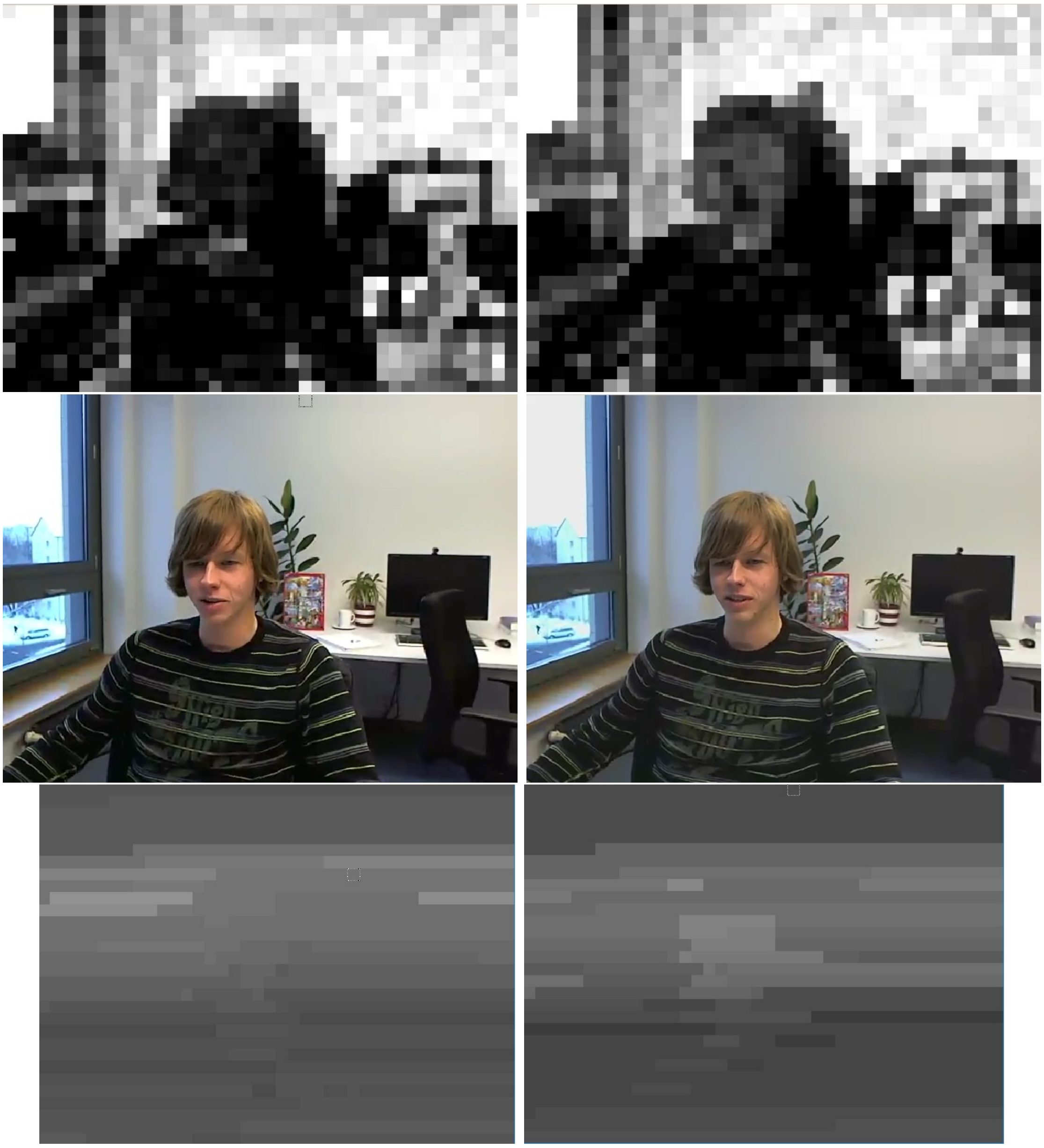

Figure 8: Comparing the images with QP offset of 4 for ROI

module used in this study is compensates for the overconsumption or underconsumption of bits in the past by adjusting the delta bits for future frames.This corrected allocation happens at every macroblock level. For instance, if there is excessive consumption of bits in the past macro-blocks, the excess is subtracted from certain number of future frames known as reaction factor. If the reaction factor is low, the excess or shortage of bits is shared by a large number of macro-blocks.

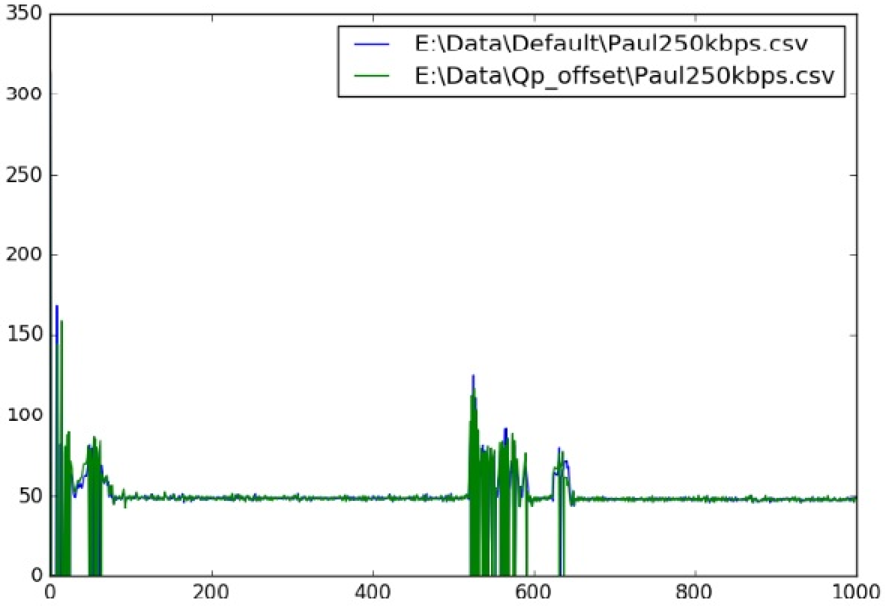

Figure 9: Delay Comparison of images with QP offset of 4 for ROI

The reaction factor is used to allocate additional bits to ROI by changing the reaction for macro-blocks belonging to ROI.

References

[1] Manzur Murshed and James 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.

[2] J. Ohm. G. J. Sullivan, H. Schwarz, Thiow Keng Tan and T. Wiegand. *Comparison of the Coding Efficiency of Video Coding Standards—Including High Efficiency Video Coding (HEVC).* IEEE Transactions on Circuits and Systems for Video Technology (Volume: 22, Issue: 12, Dec. 2012)

[3] Knuth: Computers and Typesetting,
`http://www-cs-faculty.stanford.edu/~uno/abcde.html`