Face Tracking for Optimized Bitrate Control in Low Delay Video Encoding

Chethan Ningaraju

November 15, 2016

Contents

1 Introduction

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 along with better techniques to ensure low-delay transmission reliability at the network level. The tremendous increase in smart phone usage has led to increase in video telephony over cellular networks whose bandwidth is highly constrained. Therefore, it is very important to develop methods of delivering high quality video with less bandwidth requirement. This work explores the methods of region of interest(ROI) based encoding to exploit the available bandwidth to encode regions that are of high importance to perceptual quality with higher quality.

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.

1.1 ROI based coding

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

The study of Human Visual System (HVS) also shows that human eyes can only focus at one area in a frame at any given point in time which is called region of interest. 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 the perceptual quality. This is because the region of interest constantly keeps changing depending on the content. For instance, in a movie, the region of interest can depend on context of the scene. Developing generic techniques for detection of ROI in such videos is very difficult. However, region of interest in the video conferencing content is going to be the face region predominantly. Due to recent improvements in face detection algorithms it is possible to detect the face with good accuracy. The study in [1] shows how boosting quality of face regions can improve the overall perceived quality of the video. This work aims to study possible ways of improving perceptual quality of the video by detecting face regions and coding it with higher quality than rest of the frame.

1.2 Bitrate Control

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. There are various flavors of bitrate control like Constant Bitrate (CBR), Variable Bitrate (VBR) and Average Bitrate (ABR). In this work CBR type of bitrate contol is considered since it is most commonly used in video conferencing and other streaming related solutions.

Figure 1 illustrates the functionality of the bitrate control module. The main purpose of the bitrate control module is to ensure smooth playback of the encoded video under given bandwidth and delay constraints. It achieves this by controlling the quantization parameter (QP) used during the encoding. The decision of quantization parameter is done considering the input bitrate, framerate, input complexity (spatial and temporal activity) and acceptable input delay. The module also takes the feedback from the encoder regularly to make better QP decision. The feedback from the encoder gives information about the content of the video and helps the module to decide the right QP for a given bit budget.

The functionality of the Bitrate control module can be visualized with the help of leaky-bucket model. The output data rate of a video encoder varies depending on the input complexity of the video (motion in the frame). It also depends on the picture type, key/intra frames consume a lot of bits compared to inter pictures. In video streaming scenario considering a constant bitrate channel, the throughput is maximum when data rate is constant and equal to available bandwidth. Therefore, the output data of the encoder is smoothed using a theoretical buffer called Video Buffer Verifier (VBV). The VBV buffer is a virtual buffer modeled by the bitrate control module to ensure that video stream can be correctly buffered, and played back at the decoder end. This is equivalent to a leaky-bucket model, where the output of the encoder with variable rate (bursty output) is stored in a leaky bucket which is draining at a constant rate. Any under-flow or overflow of this buffer causes glitch in the video streaming. To ensure that there is no VBV buffer overflow or underflow, encoder's quantization parameter (QP) is adapted on a macroblock level so that the maximum allowed bitcount for an encoded frame is not exceeded.

As mentioned earlier, the bitrate control allocates bits at every macroblock and adjusts the QP. 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.

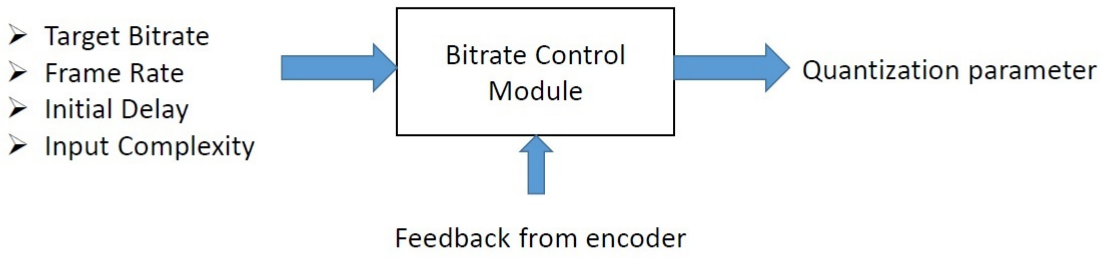

Figure 1: Bitrate Control Module Functionality

2 Literature Survey

The concept of region of interest based encoding has been around for a while.

3 Overview of Bitrate Control Module

This section gives an overview of low-delay bitrate control module used in this work. 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. During the low-delay video encoding, tools like bi-directional prediction are disabled. The delay in video encoding is a direct measure of Video Buffer Verifier (VBV) buffer size. When the size of the VBV buffer is very low (due to low delay), there is less room to accommodate the variation in bitrate of the encoder. This implies that there can be minimum variation in the size of frame irrespective of the content.

The bitrate control module used in this work in a modified version of [3]. The bitrate control does a frame level bit-allocation based on fullness of the VBV buffer, followed by adapting QP at macroblcok level. The functionality of bitrate control module can be divided into two parts:

- Bit Allocation
- Quantization Parameter (QP) Prediction

3.1 Bit Allocation

The low-delay encoding mode does not favor usage of B-frames and regular key-frames. Therefore, in steady state only P-frames are used in encoding video conferencing content. This makes frame level bit-allocation simpler since there is no need to consider relative complexity between different types of frames during frame level bit-allocation. The key-frame at the beginning is handled using special cases.

As depicted in fig:1, one of the inputs for bitrate control module is a delay/latency parameter (L). This is defined as the maximum permissible delay allowed between encoder and decoder assuming zero transmission delay. In other words, the delay parameter is the maximum allowed time for any encoded frame to be transmitted through a constant bandwidth channel of per-defined bitrate. In this work, delay parameter (L) is configured as,

$$L_0 = 165ms \text{ and } L = \frac{1.5 * 1000}{framerate}$$

Initially a delay of $L_0 = 165ms$ is allowed for the key-frame. This allows, allocation of higher than average bits for the key frame. However, the delay parameter (L) for P frames in steady state is only 1.5 times the average frame delay. For instance, if video is encoded at 30fps, then the time interval between two consecutive frame is 33ms, the permissible delay(L) for frames in steady state is approximately 49ms. The usage of different delay values for first key frame and steady state P-frames is handled by changing the delay value gradually. The large key-frame at the beginning results in huge delay (165ms), this delay is gradually reduced by using less than average bit count for subsequent few frames (half of average bit count per frame). Once the over-consumption of first-key frame has been compensated, the steady state delay of $49ms$ is maintained.

In practical implementation, usage of above average bit-count for I-frame results in few dropped frames subsequently. This is considered an acceptable trade-off to achieve good initial spatial quality.

The bit-allocation module uses VBV buffer fullness and delay parameter (L) to decide the bits allocated for the current frame. The VBV buffer fullness (d_0^{n+1}) after encoding nth frame is calculated as follows based on the size of the encoded frame in bits ($FrameSize_n$)

$$d_0^{n+1} = d_0^n + (FrameSize_n - AvgBitsPerFrame)$$
$$d_0^n = max(d_0^n, 0)$$

where,

$$AvgBitsPerFrame = \frac{bitrate}{framerate}$$

The allocated bits for $(n + 1)th$ frame is maximum amount of bits that can be transferred along with residual bits in the virtual buffer in allowed delay of $49ms$. The maximum acceptable delay in ms is translated to bits using below equation.

$$L_{bits} = \frac{L * birate}{1000}$$

Therefore, allocated bits for current frame (B_{alloc}) is given by

$$B_{alloc} = L_{bits} - d_0^n$$

In practice, rate control QP predictions are not very accurate to exactly consume the bits that was allocated to the frame (B_{alloc}). If a frame consumes more bits than B_{alloc}, it violates the delay conditions. The encoded frame will be unable to reach the decoder in time with specified bitrate. Hence, the frame is not added to the bitstream. These frames which are encoded but not part of the output of the encoder are called *dropped frames*. A slight room for inaccuracy of the QP prediction is considered at the end of the bit-allocation stage to avoid dropped frames. Such dropped frames must be avoided since it causes jerky playback. In practice, the target bits for QP prediction is slightly lesser than B_{alloc} to avoid dropping the frame in case of minor over-consumption.

3.2 QP Prediction

Due to low VBV buffer size, the bitrate control needs to have very quick reaction to any deviation in bitrate to avoid dropped frames. The bitrate control algorithm computes the QP for every macroblock. The two factors considered in computing QP for a macroblock are:

- Complexity of the macroblock
- Instantaneous Virtual Buffer occupancy.

3.2.1 Activity Based delta QP

Firstly, a delta QP (dq) is calculated considering the complexity of the macroblock. The activity of the macroblock is a measure of complexity of the macroblock and hence indicates the bits required to encode the macroblock. After motion compensation with the selected coding mode and motion vectors, the activity of macroblock (i, j) with original pixel value $s(i, j)$ and predicted pixel value of $c(i, j)$ is calculated using equation 1.

$$act_m = \sum_{i,j} \mid s(i,j) - c(i,j) \mid, \quad i,j = 1,2,....16 \tag{1}$$

The relative complexity of the macroblock with respect to entire frame complexity is used in QP adaptation. The ratio of activity of the current macroblock and average activity of the entire frame is used to calculate the delta QP (Eq:2).

$$dq = \begin{cases} -floor(\frac{avj\_act}{act_j} - 1), & 0 < \frac{act_j}{avg\_act} <= 1/2. \\ 0, & 1/2 < \frac{act_j}{avg\_act} <= 2. \\ floor(\frac{act_j}{avj\_act}) - 1, & \frac{act_j}{avg\_act} >= 2 \end{cases} \tag{2}$$

As depicted the above equation, a positive dq is used when current macroblock is complex relative to average frame complexity. This indicates that for relatively complex macroblocks within a frame, higher QP is used. This is to make sure that bits within a frame is equally distributed across all macroblocks. In practice, the average frame activity for entire frame is unavailable until the last macroblock of the frame has been encoded. Therefore, previous frame average activity is used since the average property of two adjacent frames in the video is likely to remain same.

The activity metric used in the equation2 is a complexity metric, hence it can be replaced by similar metrics depicting the complexity of the block. Other metrics like SATD (Sum of Absolute Difference in Transform Domain similar to activity calculation in Eq:1) and cost of macroblock (J) can be used instead of the activity. In this work, cost of the macroblock used in rate-distortion optimization is used as the complexity metric. The cost of the macroblock factors in both amount of residual information to be encoded and bits used for the mode making it accurate in terms of reflecting the complexity of the block.

$$J = D + \lambda R \tag{3}$$

Here the distortion D represents the residual error after prediction measured of sum of absolute difference (SAD), is weighted against the number of bits R associated with motion information using the Lagrange multiplier λ. The least cost of all the evaluated modes is considered as the complexity of the block.

3.2.2 Compute Quantization Parameter

The delta QP (dq) calculated above is added to QP calculated based on current virtual buffer fullness. The virtual buffer fullness corresponds to fullness of the buffer discussed in context of leaky bucket model in the previous section. The virtual buffer occupancy d_0^n is calculated only after entire frame is encoded. Any deviation in bitrate will be reflected in the occupancy of the buffer. In order to account for deviation from in bitrate at macro-block level a global deviation factor is computed at macroblock level based on virtual buffer fullness and size of macroblocks encoded in the current frame. The current virtual buffer occupancy (d_m^n) when encoding the m^{th} macroblock of n^{th} frame is calculated according to equation 4.

$$d_m^{n'} = d_0^{n'} + CurFrameBitCount - \frac{B_{alloc} * m}{M} \tag{4}$$

Where, M is the number of total macroblocks in a frame. The term $d_0^{n'}$ in above equation is directly derived from virtual buffer fulness (d_0^n) after clipping to definite maximum and minimum value keep QP limit in suitable range. This accounts for deviation in bitrate of the encoded video until the last fully encoded frame. It is same as d_m^n at the start of encoding of the current frame. The metric AvgBitsPerMB is not a fixed value based on bitrate and framerate. This value is calculated based on bit-allocation explained above and hence remains constant for a given frame. A positive d_0^n implies over-consumption of bits compared to target bits. CurFrameBitCount is the bit-consumption until the last encoded macroblock. Therefore, CurBufferOccupancy is the measure of deviation in target and achieved bitrate until the current macroblock in the frame. The current virtual buffer occupancy is used to calculate QP parameter for the m^{th} macroblock (Q_m) using equation 5.

$$Q_m = \frac{d_m^{n'} * 31}{r} + dq \tag{5}$$

$$where, r = i * bitrate/framerate$$

The factor dq in eq5 is the delta QP calcuated using equation 2. The factor r, is called the reaction factor. This factor indicates the number of frames over which the deviation in bitrate is to be compensated. The bitrate control module in this work uses $i = 1$.

4 Study setup

4.1 codec configuration

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

4.2 Measurements

One of the crucial aspects in this study is the metric used to evaluate various algorithms in order to choose the best approach. The goal of this study is to improve the quality of the ROI region at the cost of degrading the non-ROI regions. Since the whole approach is to measure the gain in perceptual quality, using frame level PSNR alone as a metric could be misleading.

In this study, the difference in average PSNR of frames and average ROI PSNR is used as one of the metrics to evaluate different algorithms. The expectation is to see an improvement in ROI PSNR, with degradation in PSNR of the non-ROI regions. The shift in the quality of ROI and non-ROI regions should be achieved keeping the bitrate unchanged. Therefore, the second aspect is to measure the deviation in bitrate behavior from the original output. Third aspect involves the measurement of PSNR and QP distribution within a frame. This is to ensure structure of the face is recognizable in the QP map which implies face region information is used to alter the behavior of QP allocation. The PSNR quality map is required to ensure that there is visible difference in quality of ROI without badly degrading the quality of non-ROI.

To measure all these behaviors following metrics are considered

- Quality metrics - PSNR of ROI and non-ROI regions.
- PSNR and QP variation within the frame.
- Delay plot comparison for the entire sequence.

4.2.1 Quality metrics

The initial approach is to find the gain in PSNR of ROI, however finding the desirable extent of improvement in PSNR of 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. The values in Table 1 shows the PSNR values of original output before any modification with respect to ROI based encoding was done. It is clear that the overall PSNR of ROI is much lower compared to that of the overall frame PSNR. This is not a desirable state to be in since the regions that matter most to perceptual quality have lesser PSNR on average.

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

Table 1: Initial PSNR values

The PSNR is calculated using weighted sum of PSNR of individual components per picture ($\mathrm{PSNR_Y}$, $\mathrm{PSNR_U}$ and $\mathrm{PSNR_V}$) [2].

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

where individual components are computes as

$$PSNR = 10.log_{10}((2^B - 1)^2/MSE) \tag{7}$$

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

The change in PSNR of ROI region and non- ROI region is measured as average of PSNR of entire frame and average of PSNR of ROI region of all the frames in the sequence. This measure will also indicate the aggressiveness of an algorithm. The study also included average MSE based PSNR which is calculated by accumulating the MSE over entire sequence and then calculating the PSNR, this metric was found to be heavily influenced by the outliers and hence not presented in the results in this study.

4.2.2 PSNR and QP Variation

The study of PSNR and QP distribution within the frame is important to understand the effect of bit movement from ROI to non-ROI regions. The PSNR and QP is extracted at the macroblock level. It is then stored in the raster scan order which can be used to display as an image to compare the structure with that of the video frame. These values are represented as a gray scale image with values between 0 to 255.

In 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. Generally, the lower frequency regions have better PSNR even when encoded with a higher QP. Also when a region of the frame is static, it tends to have better PSNR even when higher QP is used because of very less new information

Figure 2: A Frame in the sample video

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

to be encoded. Therefore, the PSNR distribution within a frame represented as a gray scale image is also included to analyze the effect of bit movement within the frame.

The image in figure 2 shows a frame in the sample video conferencing content with resolution of 640x480 and 30 frames per second. The image in figure 3 shows the same frame when encoded with the codec configurations discussed in previous section with 250 kbps bitrate. The reason for considering a low bitrate of 250 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

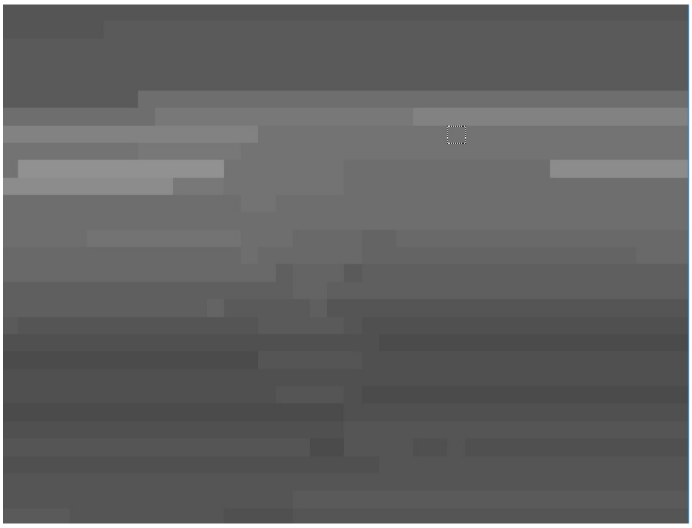

Figure 4: Quantization map

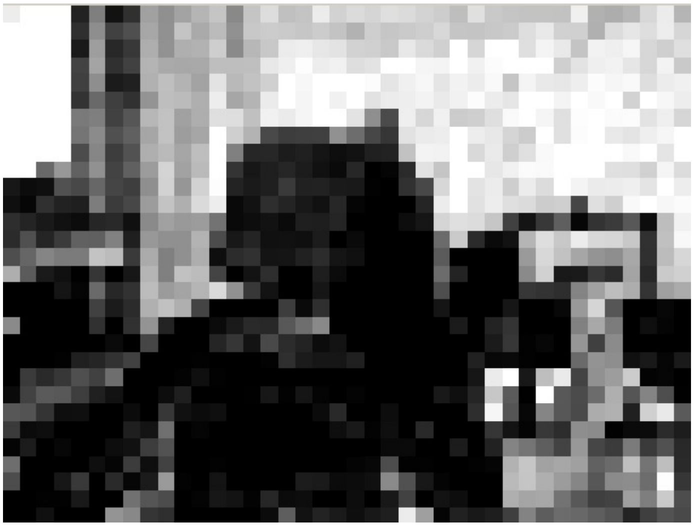

Figure 5: Relative PSNR map

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 random. 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 figure 3. This map is identical to the quantization map, the lighter regions here represent the regions with higher PSNR, the darker regions indicate lower PSNR and 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. Such PSNR maps with absolute scale is also studies to compare results across different configurations.

It is evident that the structure 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 foreground 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 of such preferential encoding is studied in this work. The idea is to reduce the PSNR difference between foreground and background and to boost the quality of face to same level as background if not better.

4.2.3 Bitrate fluctuation - Delay plots

The core idea of this study 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 previous sections, 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 curve appears mostly 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 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.

5 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 reliability with face detection. Any false detection will lead to degradation of the actual region of interest compared to normal encoding, this should be avoided in all scenarios. The damage caused by false detection is higher than the loss due to not detecting any

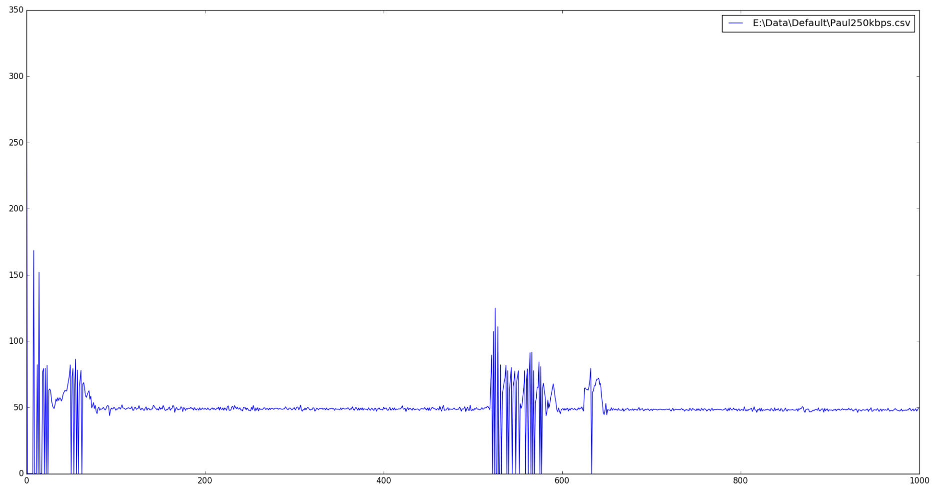

Figure 6: Delay plot

face. Therefore, a high threshold must be used to declare any region of the frame as face.

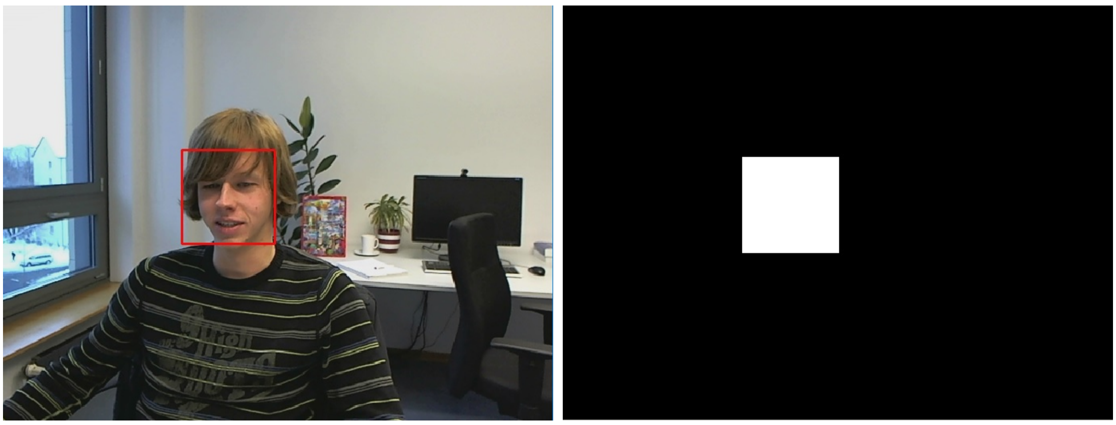

Figure 7: Face Detection binary map

The face detection module itself shall not be a part of the encoder, but the output from the face detection is a binary file with face regions marked is used as input by the encoder. In this work, the face detection module uses input YUV to the encoder and marks the region of interest at the macro block level. Each byte value in the output file of face detection module 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 Figure 2. The region in white is considered as region of interest, this information is used inside the bitrate control module of the encoder to perform intelligent bit allocation.

Different approaches are used to detect the face region in the video. There is always a trade-off between accuracy of face detection algorithm and its 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 in such systems. Therefore, the algorithm chosen for face detection must be light weight and reasonably accurate in all lighting conditions.

5.1 Spatial Domain Face Detection

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

5.2 Compressed domain Face Detection

Most available face detection algorithms work on the pixel domain. These algorithms provide good level of detection accuracy. The main drawback of this approach is that they are computationally intensive. As discussed earlier, the use case considered in this work has very less room for additional computations. Therefore, in this work ways of compressed domain face detection is explored to detect faces with less computational requirements. Many works have been published TBD LATER

6 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 ROI to gain maximum perceptual quality.

6.1 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 macroblocks belonging to these regions in the existing bitrate control module. The feedback from the encoder to rate control module 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 due to the feedback.

Such QP offsets will ensure the quality difference and shall not be overruled by bitrate control mechanisms. However, this approach might result in bitrate control overreacting for the blocks around the ROI and encode them with extremely 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. This is because there will be lesser non-ROI regions to compensate for additional bits used in ROI.

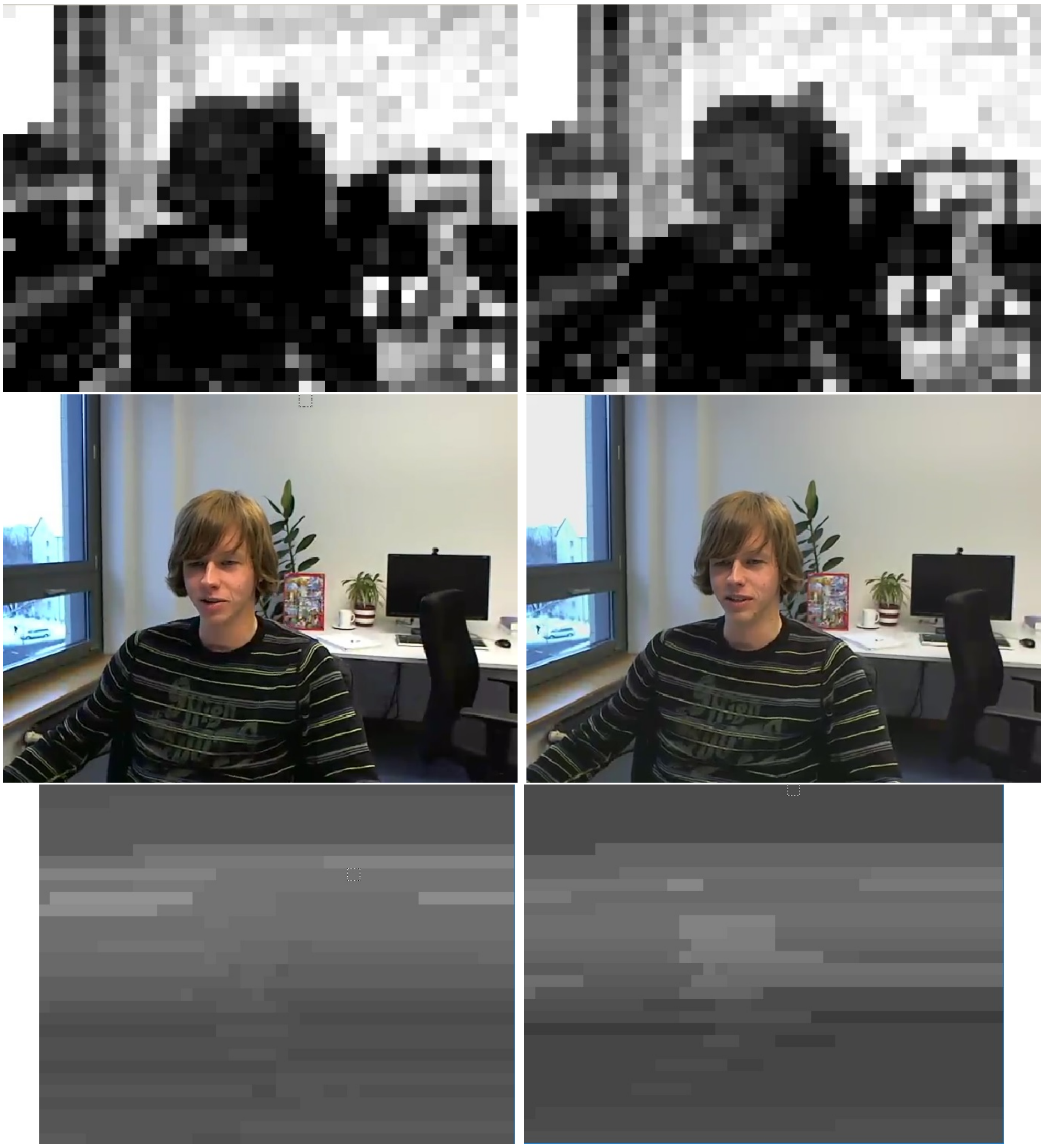

Figure 8: Comparing the images with QP offset of 4 for ROI

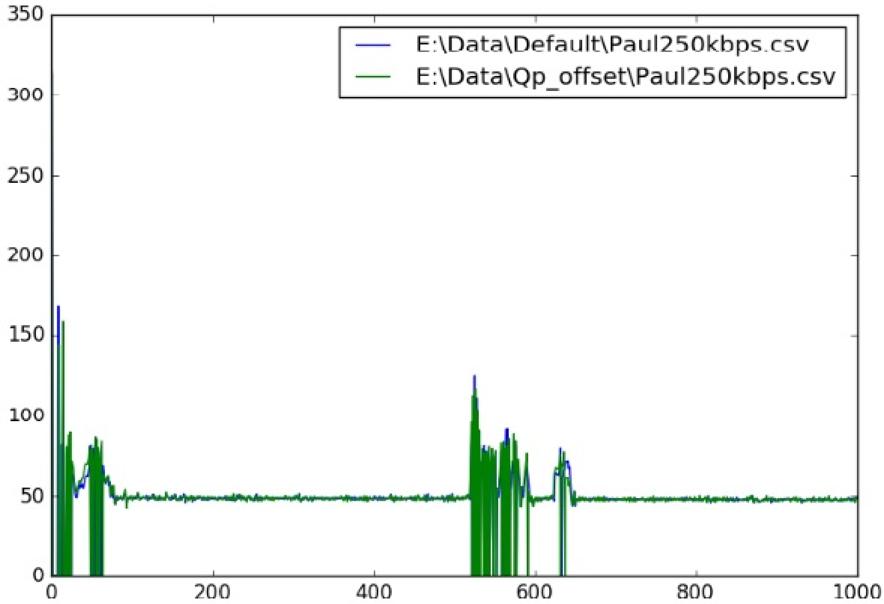

Figure 9: Delay Comparison of images with QP offset of 4 for ROI

The images in 8 shows the comparison between sample frame from original video with video encoded using QP offset of -4 for the ROI region and their corresponding attributes like PSNR map, Quant map. It can be seen that the QP in ROI region is considerably low compared to non ROI regions (marked with lighter shade of gray). The PSNR of the face is now more closer to the background region. The face appears much sharper due to additional boost in quality. The overall bitrate of the streams remained almost the same, the difference in the result is only due to the movement of bits. The number of dropped or skip frames were found to be same as original video. Figure 9 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.

The results of ROI based PSNR is tabulated in Table 2. It can be noticed that the ROI PSNR now approaches the overall frame PSNR. This difference can be further altered by tuning the QP offset used.

Methodology	Content	PSNR Avg (dB)	PSNR ROI (dB)
QP Offset	Paul640x480, 250kbps	38.90	38.88
	Johny1280x720 750kbps	40.49	40.22
Reaction Factor	Paul640x480, 250kbps	38.22	39.50
	Johny1280x720 750kbps	39.71	40.74

Table 2: PSNR Comparison for different approaches

6.1.1 Tuning QP Offset

As mentioned earlier, the methods discussed in this work only aim to re-distribute the bits within a frame based on the region of interest. The extent of re-distribution should be carefully chosen to avoid degradation of the background to an extent that artifacts become noticeable to the viewer even though they are not expected to concentrate on those regions. Ideal level of redistribution will make sure that there is maximum transfer of bits from non-ROI region to ROI without creating any visible artifacts in the image.

Figure 10: Tuning QP offset - Trial QP offsets used are 2, 4, 6 and 8 in raster scan order

In this work, various offsets were used to study the effect of magnitude of QP offset on perceptual quality. The results shown in Figure 10 shows that, as QP offset increases the face region appears more sharper which improves the overall perceptual quality of the frame. The image with QP offset of 8 appears to not have any noticeable blockiness in the background. However, using a very high QP offset triggers another problems that is not shown in the sample images. Due to increased QP offset, the encoder is forced to use lower QP even when virtual buffer is critically full. This increases the number of skipped/dropped frames in the entires sequence. The sample video consisted of 1000 frames and a total of 44, 46, 46 and 50 frames were dropped in the sequence with QP

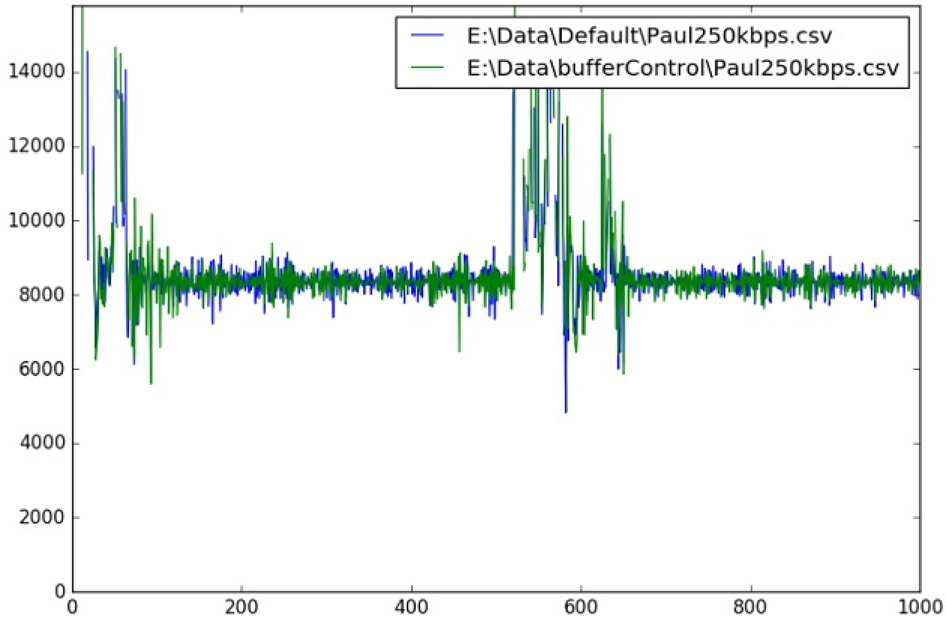

Figure 11: Delay Comparison of original video with ROI based reaction factor modification

offset of 2, 4, 6 and 8 respectively. This increase in skip frames reduces the smoothness of the playback which is annoying to the viewer. Therefore, the increase in QP offset should tuned not only considering the degradation of the background quality but also by assessing any other side effects like increase in dropped frames.

6.1.2 Area of Region of Interest

The variation in QP offset discussed in the above section correspond to same sequence where the ratio of number of ROI macroblocks to number of non-ROI macroblocks is almost same. In the sample sequence used, 46 to 50 macroblocks out of 1200 total macroblocks belong to face (ROI). Therefore, the ROI is less than 5 percent of the entire video frame. This is comparatively smaller ROI, it is possible to use very high QP offsets since there are large number of non-ROI frames to compensate for the consumption of bits at ROI. However, based on the focal length of the camera and distance from camera, the area of face in video conference can change considerably. For instance, when the area of ROI is half of the entire frame, usage of higher QP offsets will cause severe degradation in the quality of non-ROI macroblocks. In order severe degradation of the background, the magnitude of the QP offset should be inversely proportional to the ratio of ROI to non-ROI.

The algorithm implemented to use area of ROI, calculates the QP offset using linear relationship described in equation 8

$$\begin{aligned} dq_{roi} &= -round\Big(\frac{M}{M_{roi} * 3}\Big) \\ dq_{roi} &= clip(dq_{roi}, -1, -6) \end{aligned} \tag{8}$$

where, dq_{roi} is the offset of used for ROI blocks, M is the total number of macroblocks in the frame, M_{roi} is the total number of macroblocks marked as region-of-interest. The negative sign in the equation implies the calculated offset is negative, which results is QP lower non-ROI blocks.

It is evident from equation 8, there is no QP offset for ROI region if ROI region covers more than two-third of the whole frame. Subsequently, the offset increases linearly with increase in ROI area. A offset is clipped between -1 to -6 to avoid extreme offsets.

The QP offsets calculated so far is only applied to ROI region. The bitrate control module is help responsible to increase QP of the non-ROI. The over-consumption of bits in ROI triggers the reaction by Bitrate control to increase the QP for non-ROI blocks. However, this approach of one-sided offset results in the artifact event in figure TBD. It can be noticed the the blocks above ROI belonging to non-ROI generally have lower QP compared to non-ROI blocks below the face region. This is because, the QP of the non-ROI blocks is increased by bitrate control module only after it sees the over-consumption after encoding the ROI blocks. This results in usage of average QP for non-ROI blocks before encoding ROI. Once the ROI is encoded with lower QP, the bits over-consumption is compensated for by increasing the QP of non-ROI blocks encoded after ROI blocks.

A two sided QP offset is used to avoid the behavior described above. The non-ROI blocks are assigned with positive QP offset, which can compensate the over-consumption in ROI blocks. Since the non-ROI blocks from start of the frame are encoded with higher QP, there will be surplus of bits already which can be used in encoding ROI blocks. In ideal scenario, the negative and positive QP offsets must negate each others effect resulting in frame level bits being unchanged had the frame been encoded without any offset.

The study in [4] suggests that, for frame level bitcount to be constant, the average QP of the frame must remain unchanged before and after adding the offsets. This is an observation made after multiple experiments. The equation 9 shows computing the offset for non-ROI based on offset calculated for non-ROI in eq 8

$$dq_{non-roi} = \frac{M_{roi} * dq_{roi}}{M - M_{roi}} \tag{9}$$

where, $dq_{non-roi}$ is a positive QP offset used for non-ROI blocks when negative QP offset of dq_{roi} is used for ROI blocks.

6.2 Reaction Factor - Buffer control

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

Figure 12: Comparing the images with modified reaction for ROI macroblcoks

used in this study compensates for the overconsumption or underconsumption of bits in the past by adjusting the delta bits during bit allocation of the 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 bit budget of 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.

The reaction factor is used to allocate additional bits to ROI by using different reaction factors for macro-blocks belonging to the ROI. For instance, in case of underconsumption in the past, the excess bits available for future macroblocks is forced to be used aggressively for macroblocks belonging to ROI. On the other hand whenever there is overconsumption in the past, the bits are reduced more aggressively for macroblocks belonging to non-ROI. The advantage of this approach over QP offset approach is that the bit-allocation is still controlled within the rate control module and its decisions are not overridden by external offsets. This guarantees better buffer compliance. The results shown in Figure 12 compares the output with modified reaction factor based on region of interest and original video. It can be noticed that PSNR of the face regions is very close to that of the background. The data in Table 2 shows that PSNR of ROI region is actually higher than the overall frame PSNR. The delay plots in 11 shows there is no significant changes in the delay or bitrate consumption.

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] Siwei Ma, Wen Gao, Yan Lu and Hanqing Lu. *Proposed draft description of rate control on JVT standard.* Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6) 6h Meeting: Awaji, 5-13, December, 2002.

[4] Manzur Murshed, Md. Atiur Rahman Siddique, Saikat Islam and James Brown. *High Quality Region-of-Interest Coding for Video Conferencing based Remote General Practitioner Training.* eTELEMED 2013 : The Fifth International Conference on eHealth, Telemedicine, and Social Medicine