Region of Interest Encoding in Video Conference Systems

Christopher Bulla and Christian Feldmann
Institut für Nachrichtentechnik
RWTH Aachen University
Aachen, GERMANY
{bulla,feldmann}@ient.rwth-aachen.de

Martin Schink
MainConcept GmbH
Aachen, GERMANY
Martin.Schink@rovicorp.com

***Abstract*—In this paper, we present a region of interest encoding system for video conference applications. We will utilize the fact that the main focus in a typical video conference lies upon the participating persons in order to save bit-rate in less interesting parts of the video. A Viola-Jones face detector will be used to detect the regions of interest. Once a region of interest has been detected it will get tracked across consecutive frames. In order to represent the detected region of interests we use a quality map on the level of macro-blocks. This map allows the encoder to choose its quantization parameter individual for each macro-block. Furthermore, we propose a scene composition concept that is merely based upon the detected regions of interest. The visual quantization artifacts introduced by the encoder thus get irrelevant. Experiments on recorded conference sequences demonstrate the bitrate savings that can be achieved with the proposed system.**

Keywords**-region of interest coding; object detection; object tracking; scene composition; video-conferencing***

I. Introduction

Video-conferencing greatly enhances traditional telephone-conferencing, with applications ranging from every day calls from friends and family to cutting management expenses by replacing business trips with video-conferences. This multi billion dollar market splits mostly into two segments: free applications with decent quality and expensive telepresence systems. Among the free applications Skype is probably the best known application offering decent video quality. When video-conferencing substitutes business trips, the costs for video-conferencing can be several million dollars. Telepresence systems, for example, outfit rooms at individual locations with exact replicas of furniture and life-size displays create an immersive environment which creates the impression of sitting at the same table with other conference participants. All solutions share operating costs for bandwidth as by far the most expensive part of the yearly budget. Naturally, the introduction of the H.264/AVC codec for current generation video-conference systems was a major advantage over legacy systems as it cut bit-rates in half, a pattern that is expected to repeat itself with the introduction of the upcoming HEVC codec.

This paper will present an approach that is also able to achieve a bit-rate reduction by around the same factor by taking the context of the video application into account. Since the main focus of a video conference lies upon the participats our idea is to reduce bitrate in less interesting background areas. We will show how a combination of face detection, tracking, region of interest encoding and scene composition can be used to reduce bitrate while preserving a constant visual quality in the detected regions of interest.

The rest of the paper is organized as follows. In Chapter II we will in detail explain our region of interest encoding concept. Our achieved bitrate savings will be presented in Chapter III. Final conclusions as well as an outlook for future work in this area will be given in Chapter IV.

II. ROI Video encoding

Our region of interest (ROI) video encoding system consists of four key components which interact with each other (see Fig. 1). In our system, the regions of interest correspond to the faces of all participating persons. The detection is done with the Viola Jones object detection framework. Once a face is detected a new tracker will be initialized. The tracker is necessary for two reasons: Our face detection algorithm may not provide a result in every frame, however, the encoder expect a result for each frame. Tracking of the detected persons across consecutive frames will provide the encoder with the necessary information even if the face detection is still active. A second motivation for the use of a tracker is given by the fact that persons may not look into the camera all the time. In this case, the face detector would also not be able to detect these persons which finally result in a classification of theses areas as not of interest and thus in a bad visual quality.

The output of the tracker, which basically correspond to a quality value for each macro block will be forwarded to the encoder. The encoder is then able to encode the detected ROIs in a good and the background in bad quality.

Finally, the encoded video stream will be transmitted to all receiving clients which can then decode it, crop out the ROIs and render them in an arbitrary manner.

A detailed description of each component will be given in the following subsections.

ISBN: 978-1-61208-265-3

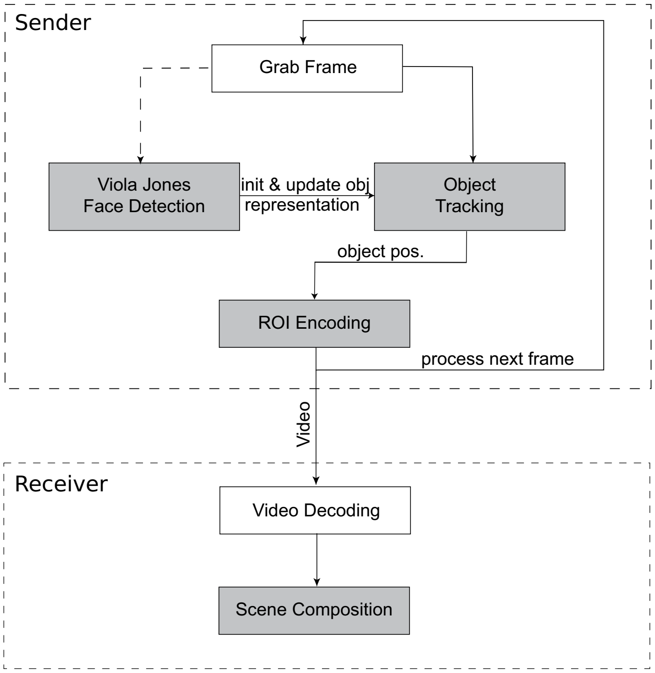


Figure 1. System overview. Interaction of face detection, tracking, video encoding and scene composition in sending and receiving client

A. *Face detection*

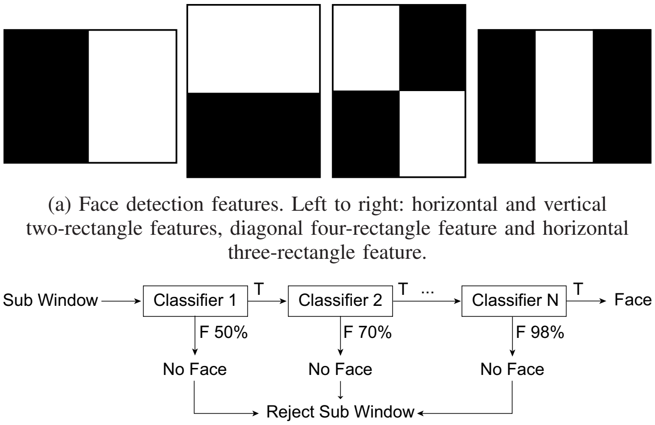
Our face detection algorithm is based on the Viola-Jones object detection framework [1]. It has three key components that will be briefly explained in the following. In a first step, a learning algorithm selects significant features in order to build efficient classifiers. The features used in this classifiers are Haar like and can be computed efficiently using an integral image representation. In order to speed up the classification process the single classifiers will be combined in a cascade.

The features that were used in the object detection system are exemplary depicted in Fig. 2a. The response of each feature is the sum of all pixel inside the black area subtracted form the sum of all pixel inside the white area. Using an alternative image representation, the integral image $II(x, y)$, these features can be computed very efficiently:

$$II(x,y) = \sum_{x' \leq x, y' \leq y} I(x', y') \ , \tag{1}$$

with $I(x', y')$ denoting the original image.

The integral image allows for the computation of the sum of all pixel inside a rectangle with only four memory access operations. The response of each feature can thus be computed very efficiently. The features are so called weak features, that means, that a classifier based on each single feature is only able to distinguish between a face and something else in a limited degree. However, a combination of these weak classifiers can yield to a strong classifier.



(a) Face detection features. Left to right: horizontal and vertical two-rectangle features, diagonal four-rectangle feature and horizontal three-rectangle feature.

(b) Cascaded classifier structure. Simple classifier reject many negative sub-windows while complex classifiers reduce the false positive rate

Figure 2. Rectangle features and cascaded classifier structure used in the face detection process

For a detection window of 24x24 pixel the entire set of possible rectangle features is about 45000. Since not all of them are necessary to detect faces in an image, a set of significant features have to be selected from all possible features what is done by AdaBoost [3].

Given a set of positive and negative training examples, the rectangle features that best separate the positive and negative examples need to be selected. The learning algorithm therefore determines the optimal threshold for a classification function such that the minimum number of examples are misclassified. The weak classifier $h_j(\mathbf{x})$ is then given by the function:

$$h_j(\mathbf{x}) = \begin{cases} 1, & \text{if } p_j f_j(\mathbf{x}) \leq p_j \theta_j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

with f_j denoting the feature, θ_j a threshold, p_j a parity for the direction of the inequality and $\mathbf{x}$ a patch of the image.

The final classifier $h(\mathbf{x})$ is then a linear combination of the selected weak classifiers:

$$h(\mathbf{x}) = \begin{cases} 1, & \text{if } \sum_{j=1}^{J} w_j h_j(\mathbf{x}) \leq \frac{1}{2} \sum_{j=1}^{J} w_j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

with J denoting the total number of weak classifier and w_j a specific weight for each weak classifier. More information on the determination of the weights can be found in [1].

In order to reduce computation time and increase the detection performance the classifiers are arranged in a cascaded structure. An example of such a structure is depicted in Fig. 2b. Classifiers with relatively large false positive rates at the beginning of the cascade can be used to reject many negative sub-windows. Computationally more complex classifiers are then used at the remaining sub-windows to reduce the false positive rate. The idea is motivated by the fact that many sub-windows within an image won't contain a face.

B. Mean Shift Tracking

Since the face detection doesn't provide a detection result for each frame, a tracking of the face positions across consecutive frames is necessary. In the general case, given the object location and its representation in frame t we want to estimate the object location in frame $t+1$. We will use a Mean Shift based tracking algorithm in order to fulfill this task. Mean Shift is an iterative technique for locating the mode of a density estimation based on sample observations $\{\mathbf{x}_n\}$ [2]. In the context of video object tracking, the samples $\{\mathbf{x}_n\}$ represent the pixel positions within the object region. In the following we will refer to the object that will be tracked as target, while possible locations of that object will be denoted as target candidates.

Let a kernel function G be given, the Mean Shift procedure estimates the new position of the target candidate $\mathbf{y}_j$ based on a previous estimate of the target candidate position $\mathbf{y}_{j-1}$ as follows:

$$\mathbf{y}_j = \frac{\sum_{n=1}^{N} w_n \mathbf{x}_n G\left(\frac{\mathbf{y}_{j-1}-\mathbf{x}_n}{h}\right)}{\sum_{n=1}^{N} w_n G\left(\frac{\mathbf{y}_{j-1}-\mathbf{x}_n}{h}\right)} \tag{4}$$

Here, N denotes the number of pixel within the object region, h the width of the kernel and w_n the weight at pixel position $\mathbf{x}_n$. The actual weight is given by:

$$w_n = \sum_{u=1}^{M} \sqrt{\frac{q_u}{p_u(\mathbf{y}_0)}\delta(b(\mathbf{x}_n) - u)} \ , \tag{5}$$

with the normalized kernel-weighted M-bin target and candidate histograms $\mathbf{q} = \{q_u\}_{u=1,\ldots,M}$ and $\mathbf{p}(\mathbf{y}) = \{p_u(\mathbf{y})\}_{u=1,\ldots,M}$:

$$q_u = C \cdot \sum_{n=1}^{N} K(\mathbf{y}_0 - \mathbf{x}_n)\delta(b(\mathbf{x_n} - u)) \tag{6}$$

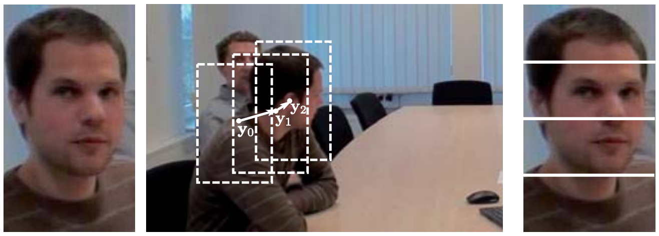
$$p_u(\mathbf{y}) = C_h \cdot \sum_{n=1}^{N} K\left(\frac{\mathbf{y} - \mathbf{x}_n}{h}\right)\delta(b(\mathbf{x_n} - u)) \ . \tag{7}$$

Here, u denotes an index of a histogram bin, $b(\cdot)$ yields the bin index of the color at pixel location $\mathbf{x}_n$, $\delta(\cdot)$ is the Kronecker delta function and C and C_h are normalization constants.
The kernel functions $K(\mathbf{x})$ and $G(\mathbf{x})$ are connected through their individual profiles $k(x)$ and $g(x)$ for which $g(x) = -k'(x)$ holds [2].

Because the appearance of the target may change over time (eg. due to a change in the lighting or a change of the 3D object pose), we will update the target representation in each frame:

$$\mathbf{q}_t = \alpha \mathbf{q}_{t-1} + (1-\alpha)\mathbf{p}(\mathbf{y}_{final})_t \ , \ 0 \leq \alpha \leq 1 \ . \tag{8}$$

Fig. 3 shows an example of the iterative Mean Shift procedure in a possible conference scenario. The target is depicted in Fig. 3a, the target candidates and the estimated locations as well as the final object location in Fig. 3b.



(a) holistic target representation

(b) target candidates and their position estimations

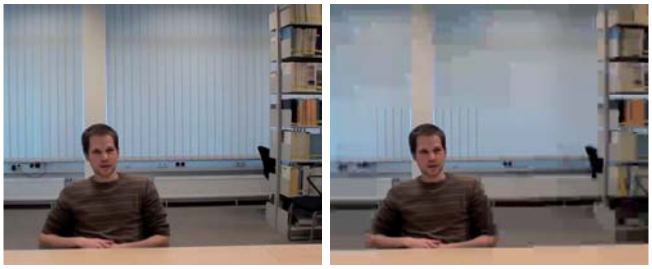
(c) multi-part target representation

Figure 3. Target representation and new location estimation by iterative mean shift updates

In order to get a more distinct object representation and thus an improved and robust tracking result, we divide our object representations according to [6] into parts which will be tracked separately. Fig. 3c shows an example of such a multi-part object representation. In contrast to the holistic representation illustrated in Fig. 3a, a multi-part representation provides information about the distribution of features for each subregion of the object.

C. ROI encoding

Implementing a region of interest algorithm alters the behavior of encoders and creates greatly different visual results. A traditional H.264/AVC encoder compresses a video stream, composed by a sequence of frames, by representing the content of these frames in a more efficient way; Although this compression is lossy, resulting in non-recoverable loss of image content, the effects are usually barely noticeable to the viewer. Rate distortion optimization makes sure that content with high importance to viewers perception of the videos quality, e.g., high frequency parts like the contours of a face or the pattern on a plant, is compressed less aggressively than content that contributes little to the viewers perception of the videos quality. Fig. 4a shows a scene with a person at a desk, and a bookshelf in the background; the scene is compressed with a standard H.264/AVC encoder and shows both the person and the bookshelves in about the same visual quality - the contours of both the person and the bookshelf are clearly identifiable, because both contribute equally to the overall visual quality. While this approach is very natural and pleasing to the human eye, it does not take the viewers attention into account: in a video-conference setting we are more interested in the person talking than in the books on the shelves. Taking the viewers attention into account means that the encoder should increase the quality of objects that are currently capturing the viewers attention, while paying for this increase in quality with lower quality on anything that is not important to the viewer; consequently, the goal of region of interest encoding is to redistribute bits for image



(a) QP 26 in ROI and background

(b) QP 26 in ROI and 51 background

Figure 4. Comparison of image qualities within and outside of region of interest

compression from areas with little interest to areas with high interest. Fig. 4b shows a very extreme case of ROI encoding, where the bookshelf is now encoded in a much lower quality than the face of the person.

A region of interest is in its simplest form a rectangle containing the object of highest interest. In the case of video conferencing this is the face of the person currently speaking and the immediate area around it. However, the shape of the ROI is not limited to a rectangle but is flexible in shape as well as in the distribution of weights within the region.

A final thought should be given to H.264/AVC standard compliance. While it is possible to implement proprietary solutions that require an encoder and decoder pair capable of understanding the implemented region of interest algorithm, it is much preferred to make do without such requirements. Video-conferencing, just like telephone-conferencing, first and foremost requires interoperability. Consequently, a region of interest implementation may only modify the encoder, but must leave the decoder untouched, resulting in decodable content by every standard compliant vendor.

1) ROI encoding using quantization parameters: Taking all these conditions into account, we chose the modification of the quantization parameters for each individual macroblock (MB), similar to the approach by Ferreira et al. [4]. In H.264/AVC each frame is divided into MBs, each with a dimension of 16x16 pixels. These MBs are then transformed into the frequency domain using the discrete cosine transform (DCT), and are then quantized before entropy encoding [5]; the decoder performs the inverse steps to recover the final frame. Quantization is used to increase compression efficiency by mapping a large set of data to a smaller set of data. This operation is lossy and introduces a quantization error into the reconstructed values. By applying this technique to the transform coefficients the amount of coded data as well as the quality of the reconstructed picture can be controlled. In H.264/AVC, the quantization can be controlled by a quantization parameter ranging from 0 to 51, 0 being the finest quantization and 51 the coarsest.

We implemented ROI encoding in the MainConcept H.264/AVC encoder by quantizing the MBs within areas of low interest very coarsely, e.g., with QPs in the range from 40 to 51, while quantizing MBs of interesting parts more finely to preserve as much of the original values as possible. Our approach generalizes the approach by Ferreira et al. [4] by allowing arbitrary values for the region of interest. As an example region of interest may include fading, e.g., values of 22 on the MBs covering the face of the active speaker, values of 28 in the MBs adjacent to the face and then QPs of 51 for the remaining background regions. Another reason for allowing a more flexible quantization of the MBs describing a region of interest are our two main use cases for video-conferencing: Without scene composition one will always view the entire frame in contrast to scene composition where parts of the frame are cropped, typically only showing the person and immediately adjacent content; since large parts of the frame aren't even seen during scene composition the quantization can easily be set to 51 for the background region that will be discarded during scene composition; likewise, without scene composition the less interesting MBs would probably not be quantized so harshly because they are clearly seen and are, while arguably less interesting, still negatively impacting the perception of quality due to the blocky nature of coarsely quantized MBs.

The quantization parameters for each MB are stored in an array which is the output of the face tracking algorithm. For convenience and to give extra to room rate distortion optimization and rate-control, we changed the values from 0 to 51 to 100 to 0, indicating the percentage of interest the viewer has in a MB - with a value of 0 resulting in the coarsest quantization and a value of 100 resulting in the finest quantization available. We choose to receive a QP array every frame, to allow for maximum flexibility for a region of interest, even though the region typically does not change rapidly due to the fact that people are rarely moving dramatically to warrant constant changes in the ROI.

The benefit of this approach is a very flexible region of interest, implemented in a H.264/AVC standard compliant manner. The downside of this approach is the MB based structure which can create blocky artifacts particularly with a very coarse quantization. Furthermore, a region of interest that resembles the exact contours of a face is also not possible due to the block based approach.

D. Scene Composition

The proposed region of interest concept offers at the receiving client the possibility to compose a video based on the detected persons. Inspired by the idea of a telepresence video conference systems, which creates the impression that all conference participants are sitting on the same table, and the fact that the focus of interest in a typical conference scenario is up to the participating persons, an alternative video composition could be achieved by showing only the detected persons. Each person is then scaled and placed side by side at the receiving client. This concept can be

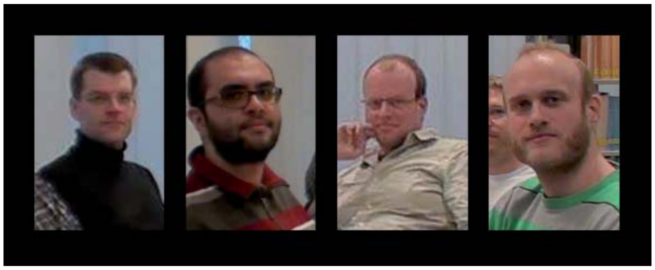


Figure 5. Exemplary scene composition of four participants

extended in that way, that only the n most active speakers will be displayed at the receiving client. Determining the active speaker can be achieved through a combined audio and video analysis. The decision which person gets rendered at which client will be made by a central mixing component that analyzes an activity index of all participants.

Fig. 5 shows an example of the scene composition with four active participants. In addition to the advantage that our proposed scene composition depicts only relevant and active conference participants, the rough quantized background gets discarded and the visual quantization artifacts depicted in Fig. 4 can be neglected. This kind of scene composition thus allows a very coarse quantization of the background.

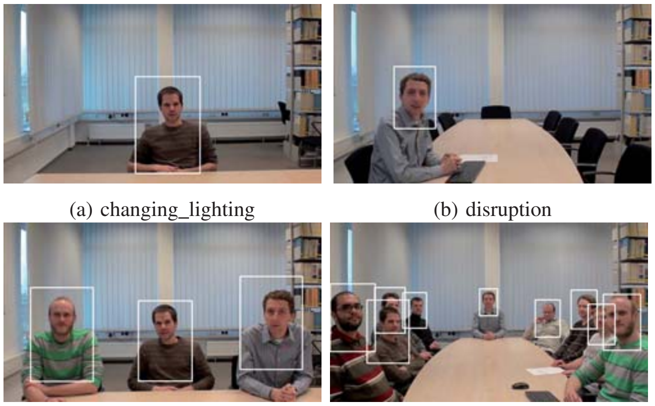
III. Evaluation

Our investigations focus on the bitrate savings achievable through region of interest (ROI) encoding in a videoconference. We thereby assume that the result of the detection and tracking algorithm is reliable. A separate evaluation of the performance of the face detection and tracking algorithm will not be the subject of this paper. Detailed information about the tracker performance for different object representation is given in [7].

Our goals for visual quality differ when scene composition is turned on or off: in case of scene composition, most of the video is cropped, so the ROI should achieve high bitrate reduction without regard to visual quality outside the ROI; without scene composition the effects of ROI encoding are directly visible to the viewer so our goal here was to find a sweet spot where bitrate savings and visual quality outside the ROI are in balance.

A. Test environment

In order to show the efficiency of our region of interest encoding approach we captured several videos with typical video-conferencing conditions. All of these videos have been recorded with a high-end consumer grade camera at a resolution of 720p and 50fps. All of the videos are 900 frames long. The videos *changing_lighting*, *disruption*, *next_to_each_other*, and *individual_spakers* show scenes with one, one, three and nine tracked people in them. Fig. 6 shows a typical frame from each of these videos. In addition to changing the number of tracked faces, we also included a change of light in the video *changing_lighting*: mid way through the video the light is turned off suddenly and gradually faded back in. Additionally, we included movement of a person in the video *next_to_each_other*. The area covered by the ROI box is 6% for *disruption*, 13% for *changing_lighting*, 23% for *next_to_each_other*, and 26% for the nine people video *individual_spakers*. For the quantization parameters of the ROI only two values have been chosen: all MBs inside the ROI have the same quantization value, just like anything outside has the same values.



(a) changing_lighting

(b) disruption

(c) next_to_each_other

(d) individual_spakers

Figure 6. Sample Images of our test sequences with detected ROIs

The face tracker generates a box shaped region of interest sized with respect to the individual faces, showing head and shoulders. The region of interest encoding has been implemented in MainConcept's H.264/AVC encoder, based on MainConcept Codec SDK 9.5. The encoder itself has been configured to a low-delay setting suitable for videoconferences: no B-frames have been used, base profile, GOP length of 300, constant quantization instead of rate-control, and deblocking turned on. The long GOP of 300 allows some IDR frames to improve the robustness against network errors, but does not allow frequent joining of a conference; whenever a new user joins the video-conference a new IDR is requested. Deblocking helps improve the visual quality for highly compressed areas so it has been turned on for all videos. To further evaluate the efficiency of different profiles we also evaluated the *next_to_each_other* video with main profile (replacing CAVLC with CABAC) and high profile (enabling 8x8 transform and PCM prediction).

The quantization parameters inside the ROI ranged from 18 to 34; values below 18 no longer provide improved visual quality for the viewer, values above 34 produce artefacts that make reading facial expressions difficult. The outside of the ROI is quantized with a step size which is a multiple of six; The quantization parameters outside the ROI range from +0, to create a non-ROI reference, until they reach +18 for a very coarse quantization.

B. *Results*

In Fig. 7 the encoder performance for different quantization values for the ROI and the non ROI region are shown. Each graph represents a constant QP difference between the ROI and the non ROI area. For QP Difference 0 the ROI and the non ROI regions use the same quantization so this is the reference for encoding not using ROI information. With higher QP difference values the quality of the non ROI region decreases. The PSNR measure only takes the PSNR inside of the ROI into account.

We can see that especially at high bitrates the bandwidth savings using a coarser quantization for the background are enormous. For example, for the highest data point (ROI QP 22) we save about 77% using a QP of 28 for the background (QP difference 6) or 86% using a QP of 34 (QP difference 12). However, such high bitrates are unrealistic to be used in video conferencing applications. A more realistic QP range is between QP 26 and 30 where the conventional video coding approach uses a bitrate of about 1-2 Mbit/sec. In this area our ROI based encoding approach yields a coding gain of approximately 50%.

TABLE I. AVERAGE BD-RATE SAVINGS FOR THE TEST SET AT DIFFERENT QP DIFFERENCES.

QP Difference	Y	U	V
6	-43.51%	-48.75%	-48.45%
12	-46.41%	-52.70%	-52.21%
18	-44.90%	-51.25%	-52.28%

In Table I the average BD-savings in our test set are shown at different QP differences. In the table as well as in Fig. 7 one can see that the rate savings do not grow with the chosen QP difference. While a QP difference of 6 already gives great rate savings a difference of 12 or more does not further decrease the bitrate by the same magnitude. However, the perceived image quality of the not ROI regions suffers badly when the QP difference is increased to 12 or even 18.

IV. Conclusion and Future Work

In this paper, we proposed a system that combines face detection, tracking and region of interest based encoding to improve the users video conferencing experience. By choosing a coarser quantization for the non ROI regions we can either save a significant amount of bandwidth or increase the quality of the video inside the ROI. When this system is combined with our proposed scene composition, the non ROI regions and their coding artifacts are removed which improves the quality of the video conference. However, also without scene composition the user experience is enhanced by shifting the encoder focus into the regions that are interesting to the user.

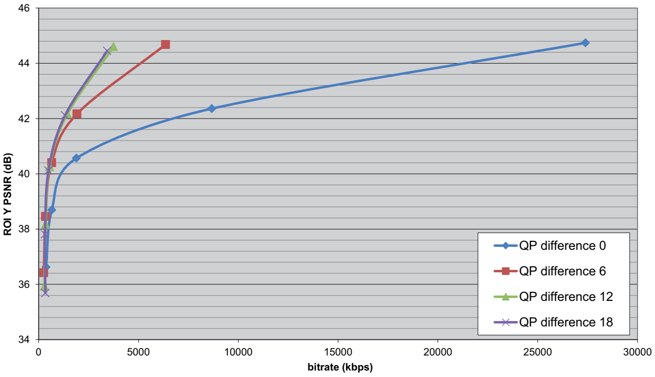


Figure 7. ROI Y-PSNR vs Bitrate for the sequence *changing_lighting* and different differences relations between the QP inside and outside of the ROI.

In future works, the accuracy of the face detection and tracking can be further improved to provide reliable information also in difficult environments. Additionaly, the shape of the ROI region can be better adapted to the speaker (e.g. give a higher priority to the face) then choosing a constant QP in a rectangular region around the face.

Acknowledgment

This work was co-funded by the German federal state North Rhine Westphalia (NRW) and the European Union (European Regional Development Fund: Investing In Your Future).

References

[1] P. Viola and M.J. Jones, “Robust real-time face detection,” International Journal of Computer Vision, vol. 57, 2004, pp. 137–154.

[2] Y. Cheng, “Mean shift, mode seeking, and clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, 1995, pp. 790–799.

[3] Y. Freund and R. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” Journal of Computer and System Sciences, vol. 55, 1997, pp. 119–139.

[4] L. Ferreira, L. Cruz and P.A. Assunção, “H. 264/SVC ROI encoding with spatial scalability,” in Proc. of International Conference on Signal Processing and Multimedia Applications, 2008, pp. 212–215.

[5] T. Wiegand, G.J. Sullivan, G. Bjontegaard and A. Luthra, “Overview of the H. 264/AVC video coding standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, 2003, pp. 560–576.

[6] D. Caulfield and K. Dawson-Howe, “Evaluation of multipart models for mean-shift tracking,” in Proc. of International Machine Vision and Image Processing Conference, 2008, pp. 77–82.

[7] P. Hosten, A. Steiger, C. Feldmann and C. Bulla, “Performance evaluation of object representations in mean shift tracking,” in Proc. of International Conferences on Advances in Multimedia, 2013.