High Quality Video Conferencing: Region of Interest Encoding and Joint Video/Audio Analysis

Christopher Bulla*, Christian Feldmann*, Magnus Schäfer†, Florian Heese†, Thomas Schlien†, and Martin Schink‡
*Institut für Nachrichtentechnik, RWTH Aachen University, Aachen, Germany,
Email: {bulla,feldmann}@ient.rwth-aachen.de
†Institute of Communication Systems and Data Processing, RWTH Aachen University, Aachen, Germany,
Email: {schaefer,heese,schlien}@ind.rwth-aachen.de
‡MainConcept GmbH, Aachen, Germany, Email: Martin.Schink@rovicorp.com

***Abstract*— In this paper, we present a high quality video conferencing system, that has been developed in the collaborative project "Connected Visual Reality (CoVR)[1] – High Quality Visual Communication in Heterogeneous Networks" and was designed to reduce bitrate while preserving a constant visual quality. We utilize the fact that the main focus in a typical video conference lies upon the participating persons to save bitrate in less interesting parts of the video and introduce a scene composition concept that is merely based on the detected regions of interest. The region of interest encoding and the scene composition will be supported by a joint video and audio analysis. On the video analysis side we use a Viola-Jones face detector to detect, and a MeanShift tracker to track the regions of interest. The audio analysis exploits the information from the video analysis about the detected participants by a beamforming algorithm and creates an activity index for each participant. To represent the detected region of interests for the encoder we use a quality map on the level of macro-blocks, which allows the encoder to choose its quantization parameter individually for each macro-block. Finally, the proposed scene composition omits the background and shows only the most active participants of the conference, thus visual quantization artifacts introduced by the encoder get irrelevant. Experiments on recorded conference sequences demonstrate bitrate savings up to 50% that can be achieved with the proposed system.**

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

I. Introduction

Video conferencing greatly enhances traditional telephone conferencing, with applications ranging from every day calls of friends and family to cutting management expenses by replacing business trips with video conferences. The solutions for such systems range from Telepresence systems, e.g., by Tandberg or Polycom, specially designed rooms for video conferences with high-performance hardware which create the impression to sit at the same table with the far end participants, to mobile clients or classical dial-in telephones. Till now high acquisition costs, limited quality or bad usability often prevent the acceptance of this systems. Another drawback, which has already been addressed in our previous work [1], is the very high operation costs caused by the high bandwidth requirements.

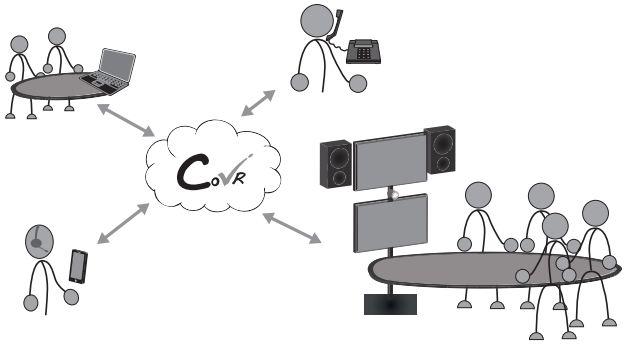


Fig. 1. CoVR - Video Conference.

The collaborative project "Connected Visual Reality (COVR) – High Quality Visual Communication in Heterogeneous Networks" [2] was working on the goal of improving the audio and video quality and interoperability of video conference systems like shown in Fig. 1. The participants of such a conference can differ significantly from each other in terms of their location, their network connectivity and in terms of the hardware. For a high heterogeneity of systems the project aims the best possible quality for every participant.

For this purpose, new methods of audio and video signal processing, coding and transmission have been developed and investigated. In the field of video processing, the project partners worked on the detection of persons, the development and standardization of HEVC (ITU-T/MPEG High Efficiency Video Coding) and Region of Interest (ROI) coding, among others [3] [4]. In the area of audio signal enhancement, algorithms for echo cancellation, noise reduction, dereverberation for multi-channel communication and artificial bandwidth extension have been considered [5].

This paper focuses on combining different parts of the work of the CoVR project. Most of the operation costs of yearly expenses for a video conference system has to be spend on the provision of the necessary bandwidth. So the saving of bandwidth by developing more efficient video codecs, like the HEVC, is a big advantage. Another approach to

[1]CoVR was funded by the NRW Ziel 2-Programm "Regionale Wettbewerbsfähigkeit und Beschäftigung" 2007-2013 and the ERDF 'European Regional Development Fund'. Participating project partners are Ericsson GmbH, MainConcept GmbH, part of Rovi, as well as two institutes of the RWTH Aachen, Institut für Nachrichtentechnik and Institute of Communication Systems and Data Processing.

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

save bandwidth is the ROI encoding, which can save up to 50% of the necessary bandwidth by coding the regions of interest (participants) with high quality and the background with low quality. This is particularly interesting for the scene composition developed in CoVR. The combination of face detection, tracking, region of interest encoding and activity index calculation allows for a scene composition which only shows the most active participants of the conference and omits the background. This way the bitrate can be reduced while preserving a constant visual quality.

The paper is organized as follows. In Section II, we will explain our video analysis, region of interest encoding concept, multi channel audio analysis and interaction between the video and audio analysis for the scene composition in detail. The evaluation of the achieved bitrate savings and audio analysis will be presented in Section III. Final conclusions as well as an outlook for future work will be given in Section IV.

II. High Quality Video Conferencing

In classical multi party video conferencing approaches each connected endpoint has one camera. The captured video is encoded into two video streams: One with a high resolution (e.g., 720p) and a second one with a lower resolution which is used as a thumbnail. Both streams are transmitted to a server, which decides upon the most active party and forwards the high resolution video stream of this party to all the other parties. The thumbnail views are always routed to all endpoints. Of course, the active party does not receive the high resolution video of itself but the high resolution video of the last active party. Hereby, each party can see a high resolution video of the active speaker and thumbnails of the other parties. But, especially in the uplink, this classical approach wastes a lot of bandwidth, and is not capable of handling simultaneous activity of multiple participants at different terminals.

The CoVR high quality video conferencing system overcomes the above mentioned limitations. Taking the requirements w.r.t. bandwidth and audio-visual quality into account our video conferencing system aims to reduce the required bandwidth of the system while the audio-visual quality remains constant. To achieve this, we use on the one hand the concept of region of interest encoding and on the other hand adaptive scene composition. A region of interest encoded video stream saves bitrate by encoding interesting parts of a video in a better quality then less interesting parts. This may result in visual coding artifacts in the less interesting areas, which in turn is controversial to our aim of high visual quality. We therefore introduce a scene composition concept, where we only display the last n active speakers, which could be located at different terminals. The transmission of a thumbnail to the server becomes thus superfluous and further bandwidth can be saved. Note that as long as our proposed scene composition is used, the complete background could be omitted leading to an even lower data rate. However, to ensure interoperability with other video conferencing systems that do not include a scene composition, at least a low quality background has to be transmitted.

Therefore, in contrast to a classical video conferencing system, we need active video and audio analysis to detect participants at each terminal and separate individual speakers. Fig. 2 gives an overview of the system and shows the interaction of video analysis, audio analysis, video encoding and scene composition in the sending and receiving client. The video analysis consists of a face detection and an object tracking part. It steers the region of interest encoder with information about the position of all conferees and provides the audio analysis with information about their angle in the room. The audio analysis can then separate each speaker and create an activity index for each individual speaker. This information together with the ROI encoded video will be transmitted to a media mixer that decides which individual person is visible at which client. Finally, the receiving client decodes the video streams and displays the last n active speakers on the screen.

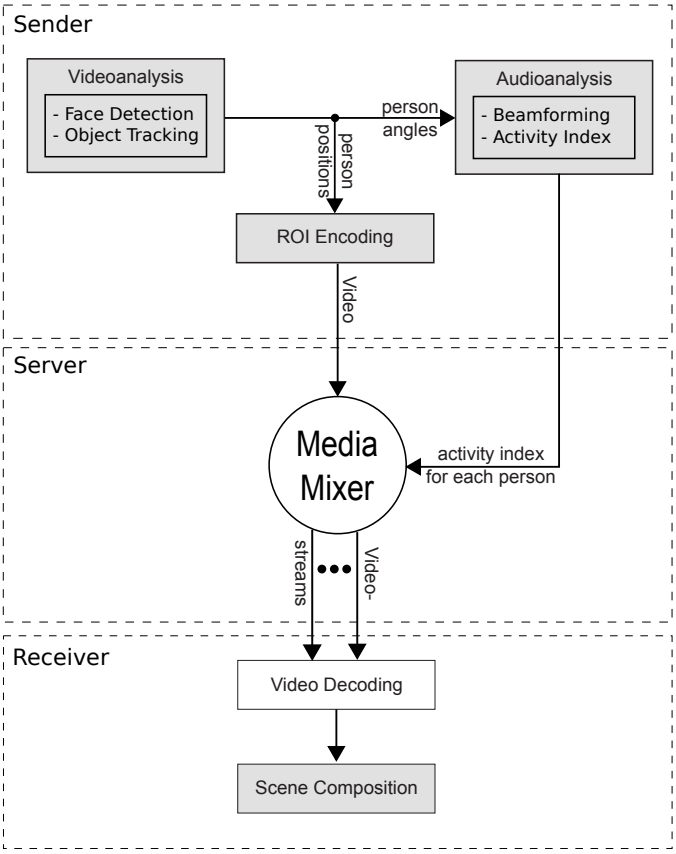


Fig. 2. System overview. Interaction of videoanalysis, audioanalysis, video encoding and scene composition in sending and receiving client.

The following subsections will explain each component in detail. We start with an explanation of the video analysis part in Section II-A. The modification on the video encoder will be presented in Section II-B and Section II-C describes the audio analysis part. Finally, Section II-D explains the CoVR scene composition.

A. Video Analysis

In order to feed the encoder with information about the position of all conferees and to provide information about their angle in the room for the audio analysis, an analysis of the video data is performed. The conferees will be detected with a

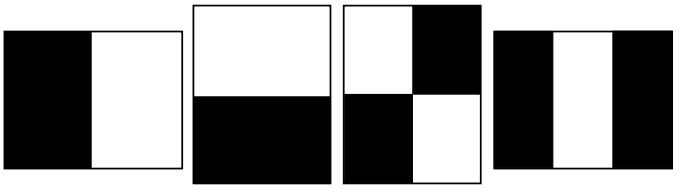


Fig. 3. Rectangle features used for face detection. Left to right: horizontal and vertical two-rectangle features, diagonal four-rectangle feature and horizontal three-rectangle feature.

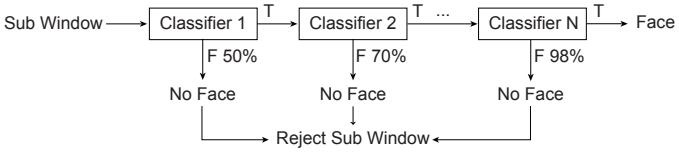


Fig. 4. Cascaded classifier structure. Simple classifier reject many negative sub-windows while complex classifiers reduce the false positive rate.

Viola Jones object detector [6], that has been trained to detect frontal views of faces. Once a face of a conferee has been detected a Mean Shift [7] tracker will be initialized to track it. The tracker is necessary for two reasons: The face detection algorithm may not provide a result in every frame, however, the encoder expects a result for each frame. Tracking of the detected persons across consecutive frames will provide the encoder with the necessary information in those frames. A second motivation for the use of a tracker is given by the fact that persons may not look at the camera all the time. In this case, the face detector would not be able to detect these persons which might finally lead to a classification of these areas as not being of interest.

In the following subsections, we will explain the used face detection and tracking algorithms in detail.

1) Face detection: Our face detection algorithm is based on the Viola-Jones object detection framework [6]. It has three key components. 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. To speed up the classification process the single classifiers will be combined in a cascade.

Fig. 3 depicts exemplarily the features that were used in the object detection system. The response of each feature is the sum of all pixels inside the black area subtracted form the sum of all pixels 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 pixels 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 extend. However, a combination of these weak classifiers can yield a strong classifier.

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 has to be selected from all possible features which is done by AdaBoost [8].

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:

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

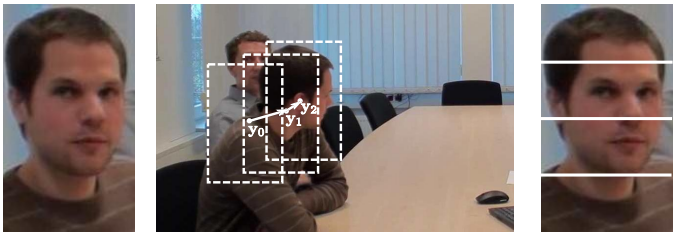
with J denoting the total number of weak classifier and w_j a specific weight for each weak classifier. More information on the determination of the weights can be found in [6].

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. 4. 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 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.

For example, a single rectangle feature classifier at the beginning of the cascade can be adjusted to detect 100% of the faces while rejecting 50% of all negative sub-windows. This simple classifier can thus significantly reduce the number of sub-windows for subsequent classification stages if subsequent stages are evaluated just in case of a positive result from the previous stage. A negative result in any of the classification stages causes the sub-window to be rejected.

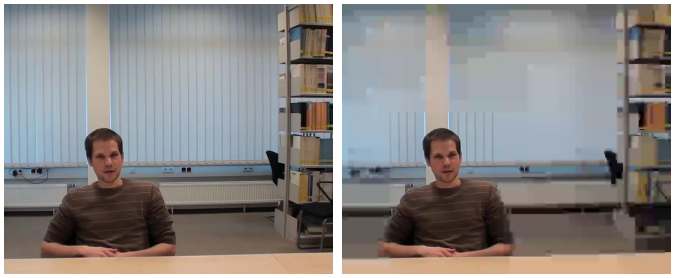
2) Mean Shift Tracking: Since the face detection does not provide a detection result for each frame, a tracking of the face positions across consecutive frames is necessary. In the general case, given the object location and its representation in frame t we want to estimate the object location in frame $t+1$. 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 $\{\mathbf{x}_n\}$ [7]. 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) holistic target representation (b) target candidates and their position estimations (c) multi-part target representation

Fig. 5. Target representation and new location estimation by iterative mean shift updates.



(a) QP 26 in ROI and background (b) QP 26 in ROI and 51 in background

Fig. 6. Comparison of image qualities within and outside of region of interest.

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 pixels 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 [7].

Because the appearance of the target may change over time (e.g., 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. 5 shows an example of the iterative Mean Shift procedure in a possible conference scenario. The target is depicted in Fig. 5a, the target candidates and the estimated locations as well as the final object location in Fig. 5b.

In order to get a more distinct object representation and thus an improved and robust tracking result, we divide our object representations according to [9] into parts which will be tracked separately. Fig. 5c shows an example of such a multi-part object representation. In contrast to the holistic representation illustrated in Fig. 5a, a multi-part representation provides information about the distribution of features for each subregion of the object. Further information especially about the influence of the object representation on the performance of the stability of the tracker is given in [10].

B. ROI Video encoding

Implementing a region of interest algorithm strongly alters the behavior of encoders and creates vastly 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 the viewer's 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 viewer's perception of the videos quality. Fig. 6a 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 that uses the same quantization parameter (QP) for the person and the bookshelves, thus showing both 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 compression from areas with little interest to areas with high interest. Fig. 6b shows a very extreme case of ROI encoding, where the bookshelf and the background outside the ROI is now encoded in a much lower quality (higher QP) than the face of the person.

A region of interest in its simplest form is 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

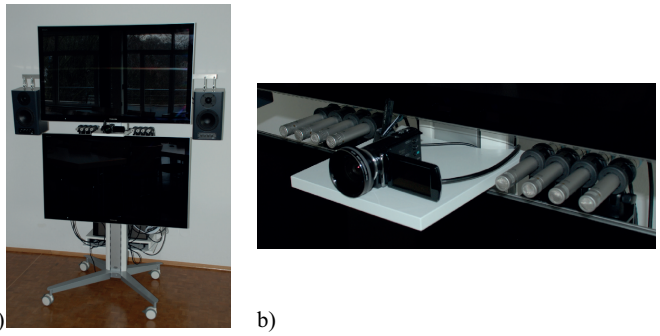
well as in the distribution of weights within the region.

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 stick to the current H.264/AVC video coding standard. 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 decoder.

Taking all these conditions into account, we chose the modification of the quantization parameters for each individual macro-block (MB), similar to the approach by Ferreira et al. [11]. 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 [12]; 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. [11] 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 are not 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 room to 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 chose 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.



a) b)

Fig. 7. Hardware setup of the video conferencing system:
a) complete system
b) detail view of the microphone array and the camera.

The benefit of this approach is a flexible region of interest, implemented into the H.264/AVC encoder without breaking standard compliance. This way any client with a compliant H.264/AVC decoder can decode the video. 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.

C. Audio Analysis

The information from the video analysis stage about the detected participants of the video conference as described in Section II-A is not only the basis for the ROI encoding but also for the audio analysis. The information about the position is exploited by a beamforming algorithm that allows to flexibly target different areas in front of the video conferencing system. With this system, the signals of the participants can be separated, which allows to quantify the current activity of each participant.

This section begins with a short introduction of the hardware setup that forms the basis of the audio analysis which is then described in two parts. First, the beamforming algorithm is presented followed by the determination of the speaker activity.

1) Microphone Array Design: The acoustic analysis of speaker activity requires the use of multiple microphones. A microphone array was specifically designed for a near field beamforming algorithm in a video conference szenario. The algorithm is utilized to extract separate acoustic signals for all speakers that were detected by the video analysis stage as described in Section II-A.

The design of a microphone array for video conferences in general has to consider the constraints that are given by the application. In the CoVR project, the main constraint is the

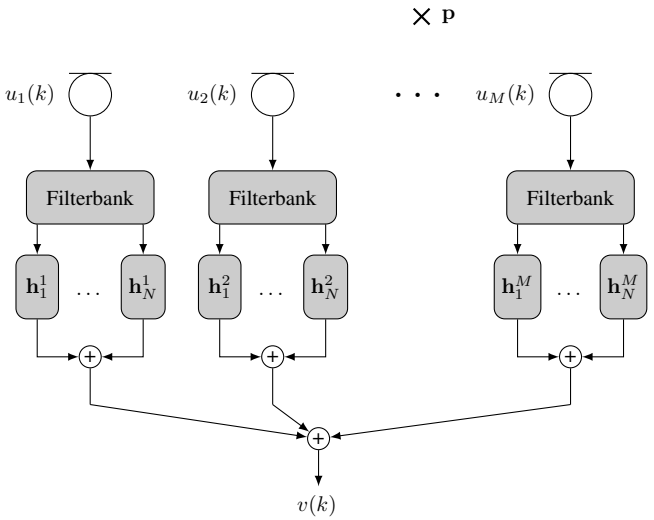


Fig. 8. Filter-and-sum beamformer with M Microphones and N non-uniform sub-bands.

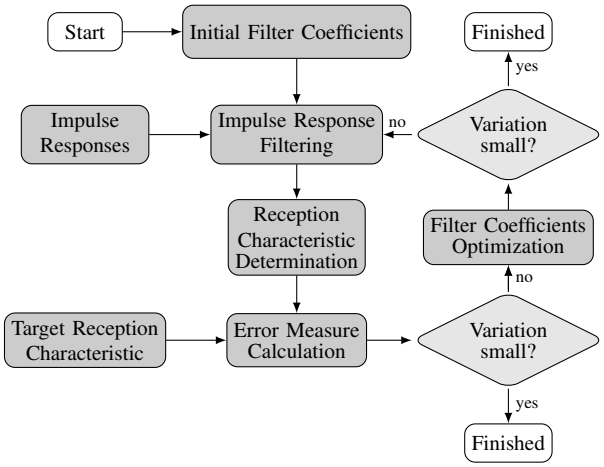


Fig. 9. Block diagram of the optimization process for each frequency sub-band n.

integration of the different components that are necessary for the visual aspects of a business video conference: two video screens and a video camera. The chosen solution is depicted in Fig. 7-a). The most convenient place to integrate a microphone array that allows to extract spatial properties of the acoustic environment is inbetween the displays. Due to the position of the video camera, two groups of four microphones each on both sides of the camera proved to be the best solution. The final setup is depicted in Fig. 7-b).

2) Beamforming Algorithm:

a) System Overview: The developed beamforming algorithm belongs to the class of filter-and-sum beamformers. The most important aspect when designing such a beamformer algorithm is the optimization of the filter coefficients. The target within the video conferencing system is to quantify the speaker's activities. Since the participants are usually located fairly close to the conferencing system, a numerical optimization procedure for the filter coefficients was developed [13], [14], [15] which exploits the acoustic properties of the room including the near field.

During the optimization procedure a predefined reception characteristic is approximated which can be chosen according to the application, e.g., extracting a specific speaker. A simplified block diagram of the proposed microphone array system is depicted in Fig. 8. It consists of a filterbank with N sub-bands followed by different filter-and-sum units represented by the impulse responses $\mathbf{h}_n^m$ $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots, N\}$ with n denoting the sub-band index and m the microphone index at all M microphones. The samples $u_m(k)$ are obtained by analog-digital conversion with a sampling frequency of $f_s = 48\,\mathrm{kHz}$, where k is the discrete time index.

To obtain an almost uniform broadband reception characteristic independently of the operating frequency, a non uniform filterbank [16] is applied to subdivide each microphone signal into N frequency sub-bands. The optimization of the filter-and-sum units is carried out in these frequency bands. Thereby the degrees of freedom for the filter coefficients determination are increased.

A point source $s(k)$ is assumed to be at position $\mathbf{p}$ on an appropriately chosen spatial grid (e.g., in a two-dimensional Cartesian coordinate system: $\mathbf{p} = (x\ y)^T$). With the impulse responses $h_{\mathbf{p}m}(k)$ from the point source to each microphone the microphone signal $u_m(k)$ can be expressed as:

$$u_m(k) = h_{\mathbf{p}m}(k) * s(k). \tag{9}$$

The output $v(k)$ depends on the source location $\mathbf{p}$ and can be calculated according to:

$$v(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * \left(h_n^{\mathrm{FB}}(k) * u_m(k)\right), \tag{10}$$

where $h_n^{\mathrm{FB}}(k)$ represents the filterbank and $h_n^m(k)$ the FIR sub-band filters of length L_h.

For the determination of appropriate impulse responses $\mathbf{h}_n^m$ of the filter-and-sum units an iterative numerical optimization process was developed. The basic concept is depicted in Fig. 9. The optimization method is based on measured or simulated impulse responses $h_{\mathbf{p}m}$. By exciting the system with dirac impulses, the reception characteristic, i.e., the spatial distribution of damped and amplified areas, is determined. The procedure consists of an iterative minimization of an error measure, which is the summed level difference between the predefined target reception characteristic and the calculated one, based on the current state of the filter coefficients $\mathbf{h}_n^m$.

The current reception characteristic can be computed by the following three steps:

- simulating or measuring impulse responses between points on an appropriately chosen spatial grid in the near field and all microphones,
- processing these impulse responses with the sub-band filter-and-sum beamformer (see Fig. 8) to get an overall filter for every point in the near field, and
- determining the amplification and damping for every point from these overall filters.

Therefor the output signal $v(k)$ has to be calculated for each source location $\mathbf{p}$ on the spatial grid which can be expressed

as a filtered version of the source signal:

$$v(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * h_n^{\mathrm{FB}}(k) * h_{\mathbf{p}m}(k) * s(k), \quad (11)$$

the overall filter $g_{\mathbf{p}}(k)$ is thus obtained as:

$$g_{\mathbf{p}}(k) = \sum_{m=1}^{M} \sum_{n=1}^{N} h_n^m(k) * h_n^{\mathrm{FB}}(k) * h_{\mathbf{p}m}(k). \quad (12)$$

The frequency transform of the overall filter $g_{\mathbf{p}}(k)$ results in:

$$G_{\mathbf{p}}(f) = \mathcal{F}\left\{g_{\mathbf{p}}(k)\right\} . \quad (13)$$

Finally the reception characteristic $S_{\mathbf{p}}(f)$ in dB can be obtained at frequency f for every point $\mathbf{p}$ in the vicinity of the microphone array by:

$$S_{\mathbf{p}}(f) = 20 \cdot \log_{10} \left|G_{\mathbf{p}}(f)\right| . \quad (14)$$

This calculated reception characteristic is compared with a predefined target reception characteristic $\hat{S}_{\mathbf{p}}(f)$. The target reception characteristic is specified as a spatial distribution of areas with defined amplification $\mathbb{P}_{\mathrm{high}}$ (target level S_{high}) or damping $\mathbb{P}_{\mathrm{low}}$ (target level S_{low}) in front of the microphone array. It can be defined individually for all frequencies but a frequency-independent target is suitable for many applications:

$$\hat{S}_{\mathbf{p}}(f) = \hat{S}_{\mathbf{p}} = \begin{cases} S_{\mathrm{high}} & \text{for } \mathbf{p} \in \mathbb{P}_{\mathrm{high}} \\ S_{\mathrm{low}} & \text{for } \mathbf{p} \in \mathbb{P}_{\mathrm{low}}. \end{cases} \quad (15)$$

The precise choice of the areas and levels depends on *a priori* knowledge from the application, e.g., in a conferencing scenario, where the target speaker activity should be calculated, the reception characteristic is defined by the number and position of possible speakers. This information is provided by the video analysis.

A quadratic error measure Δ_S between the two reception characteristics is determined as the summed level difference for all points where $\hat{S}_{\mathbf{p}}(f)$ is set according to (15) and over all frequencies f_i ($i \in \{i_{\min}, \ldots, i_{\max}\}$):

$$\Delta_S(n) = \sum_{i=i_{\min}}^{i_{\max}} \sum_{\mathbf{p} \in (\mathbb{P}_{\mathrm{high}} \cup \mathbb{P}_{\mathrm{low}})} \hat{S}_{\mathbf{p}}(f_i) - S_{\mathbf{p}}(f_i), \quad (16)$$

where $f_{i_{min}}$ and $f_{i_{max}}$ denote the lower and upper edge frequencies of sub-band n.

Based on the error measure $\Delta_S(n)$ the optimum filter coefficients for each sub-band n are determined in a minimum mean square error (MMSE) sense by:

$$\left[\mathbf{h}_n^1, \ldots, \mathbf{h}_n^M\right]_{\mathrm{opt}} = \arg\min_{\mathbf{h}} \Delta_S(n)^2 . \quad (17)$$

The optimization is carried out by an iterative interior-point algorithm [17]. The algorithm checks whether the filter coefficients change between each iteration. If the change is sufficiently small, the algorithm is terminated. The optimization of the filter coefficients can take place on the basis of generated and measured impulse responses.

3) Determination of Speaker Activity: The beamforming stage of the audio processing system allows to extract one audio signal $v_s(k)$ per participant s of the video conference. The determination of the activity $a(\lambda)$ of this participant in a 20 ms signal frame λ ($L = 960\ samples$) is done based on the corresponding extracted audio signal. In a first step, a short term energy of the signal of participant s is calculated by

$$V_s(\lambda) = \sum_{i=0}^{L-1} v_s^2(\lambda \cdot L + i) . \quad (18)$$

This energy fluctuates quite strongly on such a short timeframe so a smoothing of the energy is introduced according to

$$\bar{V}_s(\lambda) = \alpha \cdot \bar{V}_s(\lambda - 1) + (1 - \alpha) \cdot V_s(\lambda) . \quad (19)$$

The smoothing factor α is chosen as 0.9 to be able to adapt quickly to changes while the larger fluctuations are leveled out.

This smoothed energy could directly be used as the indicator of activity for the scene composition. However, it can be observed that the smoothed energy values increase steeply between situations with no activity and those with a lot of activity. Hence, as additional step, a mapping to a target scale range from 0 to 100 with 0 indicating no activity and 100 indicating a lot of activity is suitable. The activity index $a_s(\lambda)$ which is finally used for the scene composition is calculated by applying a sigmoid function

$$a_s(\lambda) = \frac{100}{1 + \mathrm{e}^{-\beta \cdot (\bar{V}_s(\lambda) - \gamma)}} . \quad (20)$$

The parameters of the sigmoid function are set as $\beta = 110$ and $\gamma = 0.05$. The resulting activity index is then rounded to the nearest integer and sent to the video encoding and scene composition modules of the video conferencing system.

D. Scene Composition

The proposed region of interest concept combined with the joint audio and video analysis offers the possibility to compose a video based on the detected persons at the receiving client. Inspired by the idea of telepresence video conference systems, which create 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 on the participating persons, an alternative video composition could be achieved by showing only the detected persons. Each person is scaled and placed side by side at the receiving client. This concept can be extended in that way, that only the last n most active speakers will be displayed at the receiving client. Determining the active speaker has been discussed in Section II-C3 and 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 media mixing component that compares the activity indices of all participants.

Fig. 10 shows an example of the scene composition with three parties and three active participants rendered at each party. Of course, due to the fact that no client will render conferees from its own party, different clients may have different scene compositions.

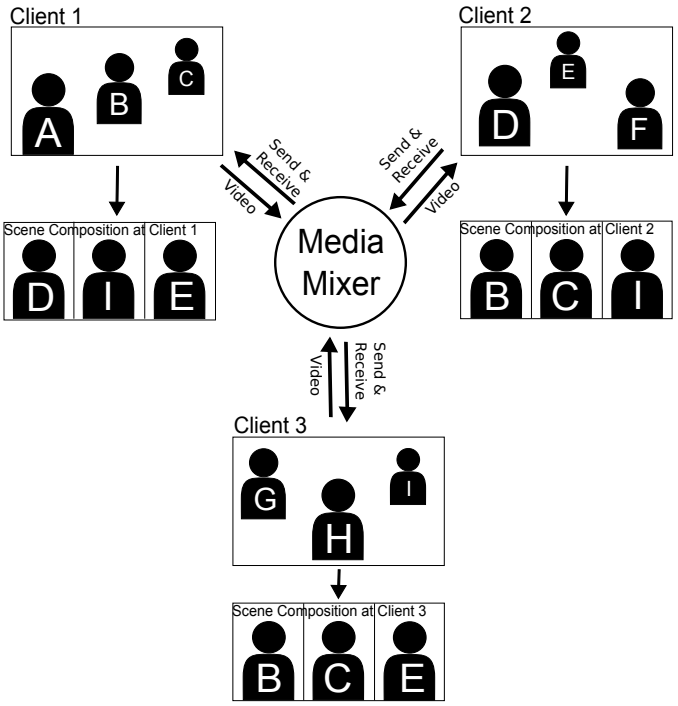


Fig. 10. Exemplary scene composition of three clients with three persons at each client. Different scenes are rendered at each client based on the decisions in the media mixer.

In addition to the advantage that our proposed scene composition depicts only relevant and active conference participants, the roughly quantized background gets discarded and the visual quantization artifacts depicted in Fig. 6 can be neglected. This kind of scene composition thus allows a very coarse quantization of the background what in turn yields large bandwidth savings (cf. Section III-A2).

III. Evaluation

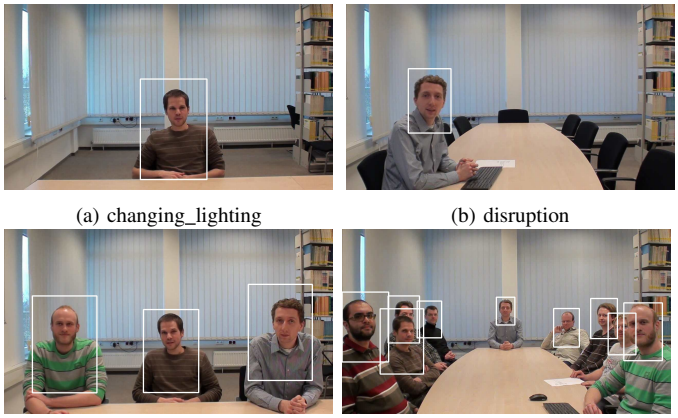
The system that is proposed in Section II has been implemented in a real time video conferencing prototype, which has also been demonstrated publicly.[2] However, due to the complexity of the system a detailed evaluation of the entire system is not feasible. Therefore, we focus our evaluation to some specific aspects of the system, such as the bitrate savings that can be achieved with region of interest encoding and the directivity improvement of the beamforming algorithm.

Due to the complexity of the task and the unavailability of reliable ground truth data, a detailed evaluation of the tracking algorithm and of the activity index is not included in this paper.

A. Region of Interest encoding

Our investigations focus on the bitrate savings achievable through region of interest (ROI) encoding in a video-conference. We thereby assume that the result of the detection and tracking algorithm is reliable. 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 encoding 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.

[2]Demonstration of video conferencing prototype at:
International Workshop on Acoustic Signal Enhancement (IWAENC'12), 2012, Aachen, Germany
Centrum für Büroautomation, Informationstechnologie und Telekommunikation (CeBIT'13), 2013, Hannover, Germany



(a) changing_lighting (b) disruption

(c) next_to_each_other (d) individual_spakers

Fig. 11. Sample Images of our test sequences with detected ROIs.

1) 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. 11 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.

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 video-conferences: bi-prediction disabled, base profile, one intra frame every 300 frames, constant quantization instead of rate-control, and deblocking turned on. The periodic intra frame allows for a re-entry into the decoding process, but does not allow frequent joining of a conference; whenever a new user joins the video-conference, a new intra frame is requested. Deblocking helps improve the visual quality for

highly compressed areas so it has been turned on for the whole test set.

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.

2) Results: In Fig. 12 the encoder performance for different quantization values for the ROI and the non ROI region are shown for all four sequences in the test set. 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 peak signal-to-noise ratio (PSNR) measure only takes the area 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) in the sequence *changing_lighting* 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%.

In Table I, the Bjøntegaard delta rate (BD-rate [18]) savings are shown for the test set ar different QP differences between ROI and non ROI regions. Table II shows the average BD-

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

QP Difference	Y	U	V
6	-52.75%	-59.45%	-58.86%
12	-57.91%	-64.98%	-64.08%
18	-54.99%	-63.22%	-63.89%

(a) *disruption*

QP Difference	Y	U	V
6	-53.52%	-60.26%	-59.87%
12	-55.49%	-63.94%	-64.00%
18	-57.27%	-61.12%	-64.25%

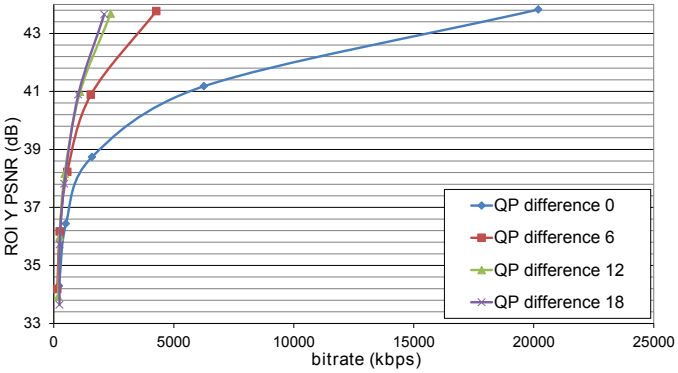
(b) *changing_lighting*

QP Difference	Y	U	V
6	-25.28%	-31.11%	-29.93%
12	-25.48%	-33.81%	-32.33%
18	-20.98%	-33.07%	-32.49%

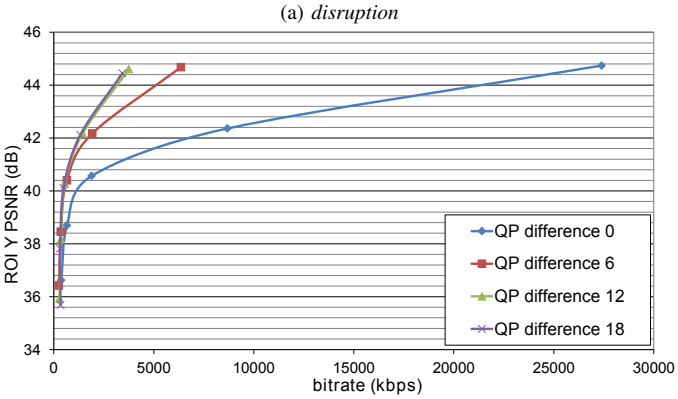
(c) *next_to_each_other*

QP Difference	Y	U	V
6	-42.47%	-44.17%	-45.12%
12	-46.75%	-48.08%	-48.42%
18	-46.35%	-47.59%	-48.50%

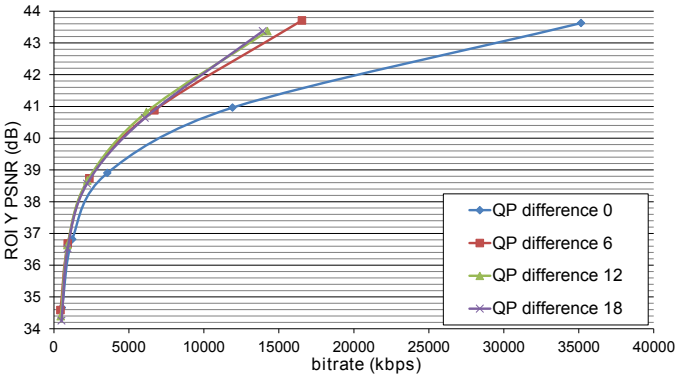
(d) *individual_speakers*



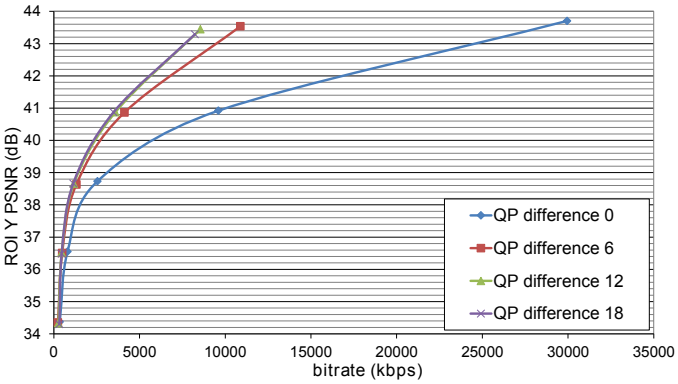
(a) *disruption*



(b) *changing_lighting*



(c) *next_to_each_other*



(d) *individual_speakers*

Fig. 12. ROI Y-PSNR vs Bitrate for each test sequence and different differences relations between the QP inside and outside of the ROI.

rate savings. In the table as well as in Fig. 12 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.

TABLE II. 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%

B. Audio analysis

The performance of the beamformer design will be demonstrated by a comparison with the well known *Minimum Variance Distortionless Response* (MVDR) beamformer [19]. For an identical setup the reception characteristic of both is depicted in Fig. 13 and Fig. 14 at two different frequencies (500Hz and 2000Hz). For better comparison the simulation were carried out under free field conditions for the two algorithms.

1) Test environment: A desired reception characteristic can be defined in front of the microphone array (including the near field) for the numerical optimization of the filter coefficients. Both systems were designed in such way that acoustic sources on the left side ($-0.5\,\mathrm{m} \leq x < 0\,\mathrm{m} \wedge 0.2\,\mathrm{m} < y \leq 0.8\,\mathrm{m}$) are amplified while sources on the right side ($0\,\mathrm{m} < x \leq 0.5\,\mathrm{m} \wedge 0.2\,\mathrm{m} < y \leq 0.8\,\mathrm{m}$) are damped. In Figs. 13 and 14 those areas are marked by white edged boxes. The resolution of the spatial grid for each space dimension (x, y) was set to $1\,\mathrm{cm}$, which results in 3000 points for each area $\mathbb{P}_{\mathrm{high}}$ and $\mathbb{P}_{\mathrm{low}}$. A level difference of 40 dB between the amplified $\mathbb{P}_{\mathrm{high}}$ and damped $\mathbb{P}_{\mathrm{low}}$ areas was chosen.

The microphone array consists according to Fig. 7-b of $M = 8$ sensors with a spacing of [3, 3, 3, 30, 3, 3, 3] cm. Spatial alias can be expected for frequencies greater than approx. 5600 Hz. Thus, the behavior above this frequency can not be clearly controlled. For the developed algorithm a non uniform filterbank [16] is used, which consists of $N = 6$ sub-bands. The frequency range of the sub-bands are given in Table III. For simplicity all sub-band filters have been realized as FIR filters. The length of the impulse responses of the filter-and-sum units $\mathbf{h}_n^m$ was set to $L_h = 8$. The filter length of the MVDR beamformer was set to 96 for the complete frequency range and per microphone with susceptibility K_0 set to 3 [20]. This makes it twice as long as the effective filter length of the new system ($N \cdot L_h = 48$).

2) Results: Fig. 13 depicts the reception characteristics of the MVDR system at $500\,\mathrm{Hz}$ and $2000\,\mathrm{Hz}$. At $500\,\mathrm{Hz}$ the MVDR beamfomer is only able to achieve a low directivity. For the configuration at $2000\,\mathrm{Hz}$ a noticeable level difference between $\mathbb{P}_{\mathrm{high}}$ and $\mathbb{P}_{\mathrm{low}}$ is observable. However, in this case the system has a very inhomogeneous behavior in the stop-band.

TABLE III. FILTERBANK SUB-BANDS

Band	Frequency range [Hz]		Band	Frequency range [Hz]	
1	1	268	4	1549	2614
2	268	839	5	2614	4731
3	839	1549	6	4731	12049

The reception characteristic of the proposed beamformer (see Fig. 14) has a significantly improved directivity characteristic. A significant level difference is recognizable for both operating frequencies between both areas $\mathbb{P}_{\mathrm{high}}$ and $\mathbb{P}_{\mathrm{low}}$. Especially at the edge regions the predefined reception characteristic is well approximated.

In comparison with the widely used MVDR beamformer the new numerical optimization procedure is able to achieve a much better separation of (in this case two) speakers across a wide frequency range.

IV. Conclusion and future work

In this paper, we presented a system that combines video and audio analysis elements to provide information for region of interest based encoding and scene composition to improve the users video conferencing experience. The video analysis stage consists of a face detection with a suitably trained Viola Jones detector and a continuous tracking of found participants by a Mean Shift tracker. The combination of these two stages provides very stable and robust results even in adverse conditions. The information about the position of the participants is utilized as the foundation of both the ROI encoding and the audio analysis stage. The audio analysis is based on a microphone array setup that was specifically designed for the video conferencing system and uses a novel beamforming algorithm which is capable of directly using the information from the video analysis. The output of the beamformer is then used to quantify the activity of the participant. Hence, the combination of the video and the audio analysis allows to simultaneously achieve robust results regarding the position and the activity of every participant of the video conference.

It was shown that the region of interest based encoding allows to either save a significant amount of bandwidth or increase the quality of the video inside the ROI by choosing a coarser quantization for the non ROI regions. 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.

The future work will focus on improving the accuracy of the face detection and tracking to provide reliable information also in difficult environments. Additionally, 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. The output of the beamformer is so far only utilized for the determination of the activity of every participant, the separated signals of the participants could also be used for an enhancement to the scene composition by matching the spatial properties of the audio signals to the spatial position of the participant on the display.

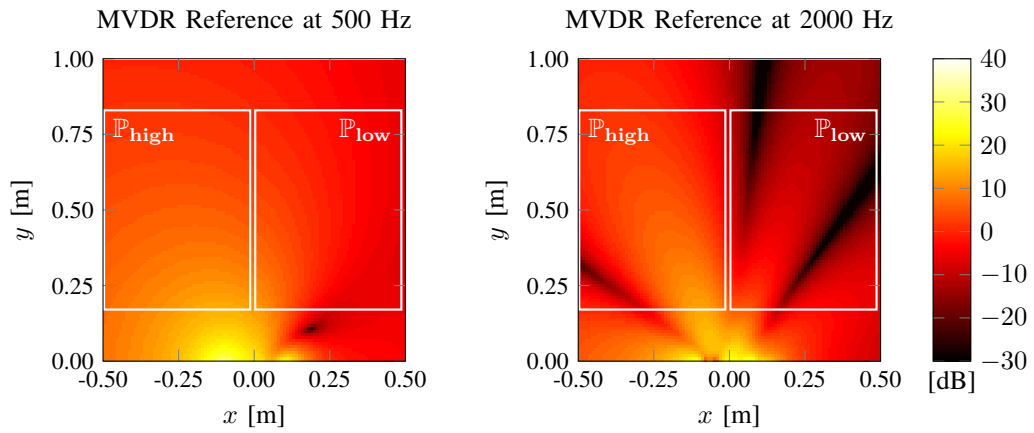


Fig. 13. Reception characteristics of the reference MVDR beamformer.

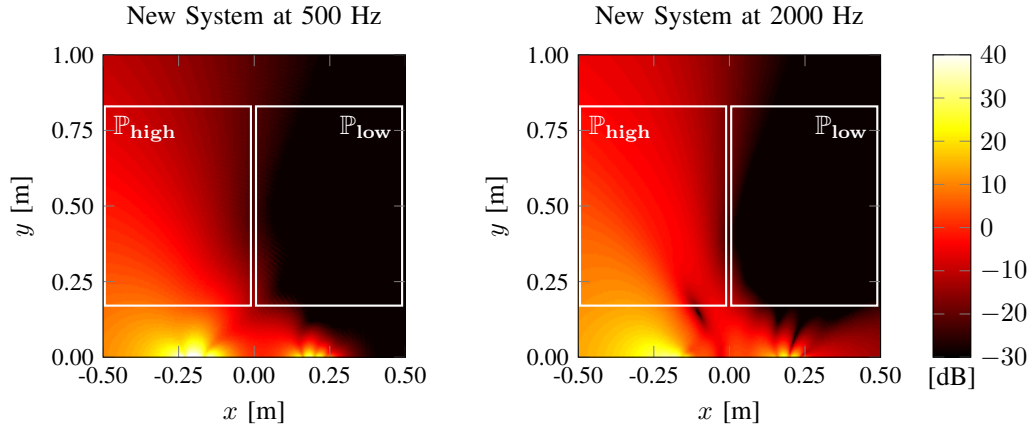


Fig. 14. Reception characteristics of the new beamforming algorithm.

REFERENCES

[1] C. Bulla, C. Feldmann, and M. Schink, "Region of Interest Encoding in Video Conference Systems," in *Proc. of International Conferences on Advances in Multimedia (MMEDIA)*, 2013, pp. 119–124.

[2] "Connected Visual Reality (CoVR) - High Quality Visual Communication in Heterogeneous Networks", 2013, joint project of Ericsson GmbH, MainConcept GmbH, part of Rovi, as well as two institutes of the RWTH Aachen: Institut für Nachrichtentechnik and Institute for Communication Systems and Data Processing, http://www.covr.rwth-aachen.de (last access: 15 Dez. 2013).

[3] C. Feldmann, M. Wien, J. Hsu, and F. Jäger, "Single-loop SNR scalability using Binary Residual Refinement Coding," 12th Meeting, Geneva, Switzerland, Tech. Rep. JCTVC-L0154, Jan. 2013.

[4] C. Feldmann, C. Bulla, and B. Cellarius, "Efficient Stream-Reassembling for Video Conferencing Applications using Tiles in HEVC," in *Proc. of International Conferences on Advances in Multimedia (MMEDIA)*, Venice, Italy, Apr. 2013, pp. 130–135.

[5] T. Schlien, F. Heese, M. Schäfer, C. Antweiler, and P. Vary, "Audiosignalverarbeitung für Videokonferenzsysteme," in *Workshop Audiosignal- und Sprachverarbeitung (WASP)*. Gesellschaft für Informatik, Sep. 2013, workshop im Rahmen der 43. Jahrestagung der Gesellschaft für Informatik.

[6] P. Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137 –145, 2004.

[7] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 790–799, 1995.

[8] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119–139, 1997.

[9] D. Caulfield and K. Dawson-Howe, "Evaluation of Multi-Part Models for Mean-Shift Tracking," in *Proc. of International Machine Vision and Image Processing Conference*, 2008, pp. 77–82.

[10] 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 (MMEDIA)*, 2013, pp. 1–6.

[11] L. Ferreira, L. Cruz, and P. Assunção, "H. 264/SVC ROI Encoding with Spatial Scalability," in *Proc. of International Conference on Signal Processing and Multimedia Applications*, 2008, pp. 212–215.

[12] T. Wiegand, G. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H. 264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, 2003.

[13] M. Schäfer, F. Heese, J. Wernerus, and P. Vary, "Numerical Near Field Optimization of Weighted Delay-and-Sum Microphone Arrays," in *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2012.

[14] F. Heese, M. Schäfer, P. Vary, E. Hadad, S. M. Golan, and S. Gannot, "Comparison of Supervised and Semi-supervised Beamformers Using Real Audio Recordings," in *Proceedings of IEEE 27-th Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, Nov. 2012.

[15] F. Heese, M. Schäfer, J. Wernerus, and P. Vary, "Numerical Near Field Optimization of a Non-Uniform Sub-band Filter-and-Sum Beamformer," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[16] H. W. Löllmann, "Allpass-Based Analysis-Synthesis Filter-Banks: Design and Application," Ph.D. dissertation, IND, RWTH Aachen University, Nov. 2011.

[17] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming," *Mathematical Programming*, vol. 89, no. 1, pp. 149–185, 2000.

[18] G. Bjøntegaard, "Calculation of Average PSNR Differences between RD Curves," document VCEG-M33, ITU-T Q6/16, Austin TX, USA, Tech. Rep., Apr. 2001.

[19] P. Vary and R. Martin, *Digital Speech Transmission - Enhancement, Coding & Error Concealment*. John Wiley & Sons, Ltd., Jan. 2006.

[20] M. Dörbecker, "Mehrkanalige Signalverarbeitung zur Verbesserung akustisch gestörter Sprachsignale am Beispiel elektronischer Hörhilfen," Ph.D. dissertation, IND, RWTH Aachen University, Jul. 1998.