Content–based Coding of Videophone Sequences Using Automatic Face Detection

M. Wollborn, M. Kampmann, R. Mech
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung
Universität Hannover, Appelstraße 9A, D–30167 Hannover 1, Germany
wollborn/kampmann/mech@tnt.uni–hannover.de

ABSTRACT: *A content–based coding scheme for the transmission of videophone sequences at very low bit rates conforming to the MPEG–4 standard is presented. The goal is to improve the image quality in the facial area of a person, at the expense of a lower quality in the remaining image, which is subjectively less important for the communication partner. In a first step, the face of the talking person is detected automatically. Then, each image is coded and transmitted as two different video object planes (VOP): the face VOP is formed by the facial area, the residual VOP by the remaining image. Thus, a large amount of the available bit rate can be used for coding the face VOP in good quality, while the residual VOP is coded at a lower quality. Using typical videophone sequences and compared to a standard scheme which codes the whole image at the same quality, the proposed scheme shows significant improvements of the image quality in the facial area.*

1. INTRODUCTION

For video telephony at very low bit rates, e.g. for transmission over analogue telephone lines or mobile networks like GSM, the available bit rate is in general not sufficient to transmit the images without visible artifacts. However, for the normal scenario exhibiting head and shoulder of a person in front of a static background, the quality of the facial area of the person is much more important for the subjective impression of the image quality than the rest of the image, i.e. the body part and the background. Thus, the subjective impression of the image quality can be improved if a larger part of the bit rate is spend for coding the facial area of the person, while only a small portion is used for the rest of the image.

Standardized block–based hybrid codecs like ITU–T H.261 and H.263 code and transmit each image as a whole. Therefore, it is not possible to address the arbitrarily shaped facial area independently from the rest of the image. However, this kind of content–based functionality will be provided by the new standard ISO/MPEG–4, which is currently being developed. Here, so–called video object planes (VOP) of arbitrary shape and size are coded instead of rectangular images. Further, it is possible to code and transmit several of these VOPs at the same time, which are then composed at the receiver side and displayed together in one image.

In this paper, an algorithm for content–based coding of videophone sequences is proposed which uses an MPEG–4 conforming codec to transmit the facial areas of the image in a better quality compared to the remaining image. Since it is not known where the facial area is located in the image, an automatic face detection algorithm is applied.

2. PRINCIPLE BLOCK DIAGRAM

The principle block diagram of the proposed scheme is shown in Fig. 1.

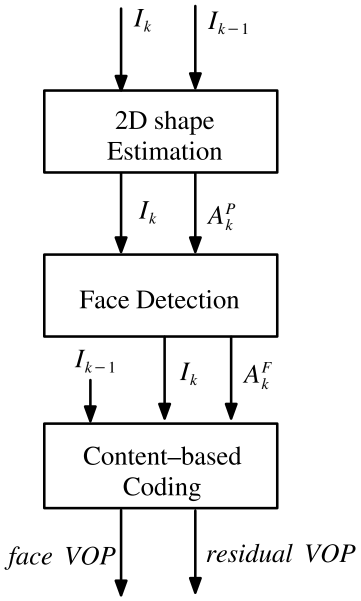


Figure 1 Principle block diagram of the proposed content–based coding scheme

By the first two blocks, each image is subdivided into one part denoting the facial area and another part denoting the remaining image. Therefore, first the 2D shape A_k^P describing the silhouette of a person in the scene is estimated. Then, the facial area within this silhouette is detected and described by its shape A_k^F.

By the last block the face VOP, formed by the facial areas, and the residual VOP, formed by the remaining image, are coded and transmitted separately. Therefore, the current MPEG–4 video verification model (VM) is used. These three blocks are described in the following paragraphs.

3. 2D SHAPE ESTIMATION

For estimating the 2D shape A_k^P describing the silhouette of a person in the scene, the segmentation algorithm described in [4] is used. There, the shape of moving objects in video sequences is estimated assuming a static camera. The algorithm can be subdivided into three steps (Fig. 2):

First, a change detection mask (CDM) between two successive frames I_k and I_{k-1} is estimated, in which all pels are marked where the corresponding luminance difference is caused by a moving object. For this purpose, a global thresholding of the luminance difference image is performed, followed by a local adaptive thresholding in order to increase the noise robustness of the algorithm [1]. In order to get temporally stable object shapes, a memory for the change detection masks is applied. The length of this memory adapts automatically to the sequence by evaluating the size and motion amplitudes of the moving objects. Furthermore, the previous object mask $OM_{(k-1)}$ is evaluated.

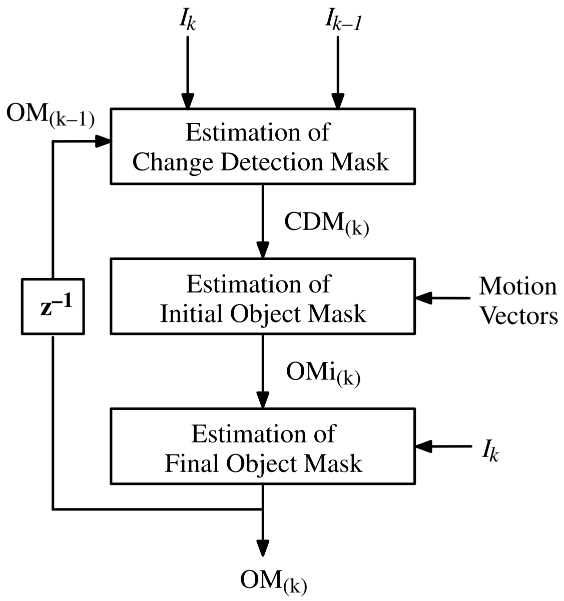


Figure 2 Principle block diagram of the algorithm for 2D shape estimation

In the second step, an initial mask of object shapes (OMi) is estimated by eliminating uncovered background from the CDM. Uncovered background is detected by evaluating an estimated displacement vector field inside the CDM, considering that a displacement vector of the moving object must point to pels inside the CDM.

In the third step, the mask OMi is refined using texture information of the current frame I_k, resulting in the final mask of object shapes (OM). The estimated 2D shape of a person is exemplary shown in Fig. 3.

An enhanced version of this segmentation algorithm, considering sequences with a moving camera, is described in [5], and is currently under investigation within a core experiment of the ISO/MPEG–4 standardization activities [3].

4. FACE DETECTION AND TRACKING

For automatic face detection, an extended version of the algorithm described in [2] is used which assumes that the 2D shape of the person consists of a top narrow area showing the person's head and a bottom wide area showing the body. First, eyes and mouth center positions are estimated and a 3D face model is adapted using these estimated positions. This is carried out only once at the beginning of the image sequence. Second, the face of the person is tracked by the face model in the following frames.

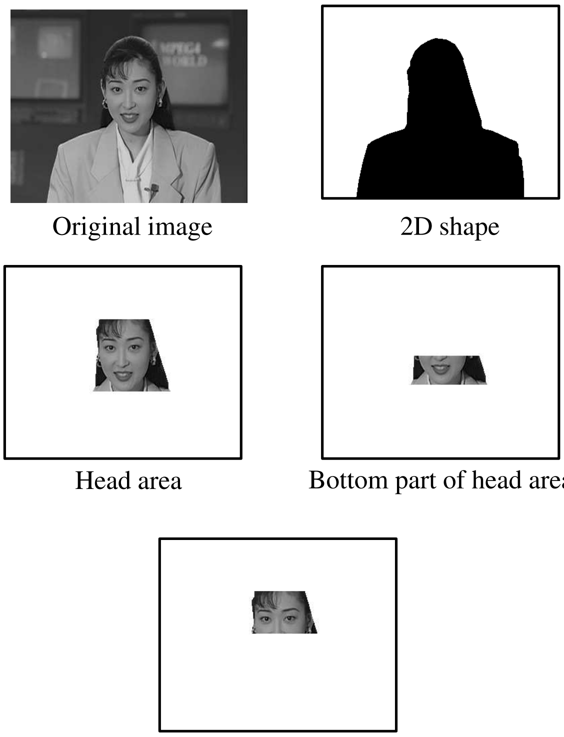
In the first part of automatic face detection, the 2D shape A_k^P of the person is evaluated and the head area is extracted (Fig. 2). For estimation of the mouth center position, horizontal contours are extracted in the bottom part of the head area (Fig. 2). Then, a template matching with a luminance mouth template is carried out. According to the area of the head, the size of the mouth template is roughly adapted to the real size of the person's mouth. The pel with the highest correlation between the luminance s_k and the mouth template is selected as potential mouth center position. For estimation of the eyes center positions, potential eyes areas are estimated first. Therefore, horizontal contours are extracted and a template matching with a luminance eye template is carried out. Similar to the mouth template, the size of the eye template is roughly adapted to the real size of the person's eyes by evaluating the area of the extracted head. Afterwards, the pupils as the eyes center positions are estimated. Therefore, a probability measure f_{eye} is evaluated inside the potential eyes areas

$$f_{eye}(x, y) = w_1 \frac{255 - s_k(x, y)}{255} + w_2 c_k^{eye}(x, y). \qquad (1)$$

Because the pupil of an eye is darker than the rest of the eye, the first term assigns a high value to dark pels (x,y) of the luminance s_k. The correlation $c_k^{eye}(x, y)$ is computed between the luminance s_k and the eye template in a window centered at (x,y). The weighting factors w_1 and w_2 have been empirically selected to w_1=1 and w_2=1. Those two pels with the highest values of f_{eye} are selected as potential eye center positions.

After estimation of the potential eyes and mouth center positions, a verification of these positions is carried out. For a successful verification, the potential positions must fulfill the following conditions. Eyes and mouth must span an isosceles triangle. Furthermore, the vertical distance between the eyes and the mouth has to be between one and two times the eyes distance. After successful verification, the 3D face model *Candide* [7] is adapted to the person's face using the estimated positions of eyes and mouth.

In the second part of automatic face detection, the 3D face model is motion compensated throughout the image sequence. Therefore, 3D rotation and translation parameters of the face model are estimated using a gradient method [2]. By projection of the motion compensated face model onto the image plane, the facial area described by its shape A_k^F is detected continuously throughout the image sequence.



Original image 2D shape

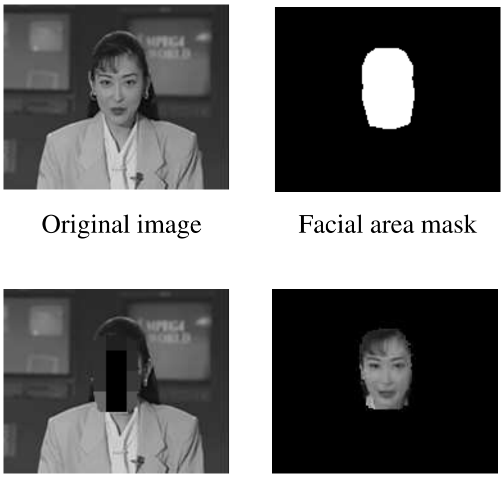
Head area Bottom part of head area

Top part of head area

Figure 3 Information used for face detection

5. CONTENT–BASED CODING

In order to efficiently code the face and the remaining image, the current MPEG–4 video VM [6] is used. For the face VOP, shape, motion and texture parameters are coded and transmitted, for the residual VOP only motion and texture parameters. Instead of coding the corresponding shape parameters for this VOP, the facial area is filled with arbitrary data using a lowpass extrapolation padding technique [6]. Then, the whole image (including padded facial area) is coded and transmitted as residual VOP. At the receiver side, both VOPs are decoded and composed, putting the face VOP in the foreground and thus covering the padded area of the residual VOP. By using this padding technique, a less overall bit rate is required since coding of the padded area is less expensive in terms of bit rate than coding the shape information of the residual VOP. An example showing the original image, the facial area mask and both VOPs to be coded is given in Fig. 4.



Original image Facial area mask

Residual VOP (with padded facial area) Face VOP

Figure 4 Original image, facial area mask and video object planes to be coded

In order to control the different qualities of the face VOP and the residual VOP, either the respective quantization parameters or the frame rate can be chosen differently. A combination of both parameter variations is alternatively realized. Up to now, no automatic rate control has been developed which allows to control these parameters. Thus, they have been set manually, taking into account the desired overall bit rate.

6. EXPERIMENTAL RESULTS

The proposed scheme has been compared to the current MPEG–4 video VM in frame based mode, i.e. coding each image as one rectangular VOP. Results are presented for two different kinds of test sequences with typical videophone contents. On the one hand, test sequence *Claire* and MPEG–4 test sequence *Akiyo* are taken where most significant motion is located in the face of the person, while in the remaining image the motion is very low. On the other hand, test sequence *Salesman* is used, where the amount of motion in the face and the remaining image is similar.

All sequences have been coded in QCIF resolution and at bit rates between 9 and 24 kbit/s. The bit rate allocation between the two VOPs was realized by setting the respective quantization parameter or the respective frame rate differently. For the presented results, all sequences except for *Akiyo* were coded at a frame rate of 10 Hz. In case of *Akiyo*, the residual VOP was coded at a reduced frame rate of 5 Hz.

Claire (PSNR[dB]: face / remaining image)		
Bit rate (kbit/s)	Reference scheme (10Hz / 10Hz)	Proposed scheme (10Hz / 10Hz)
9.0	28.6/34.8	28.6/30.7
12.0	30.3/36.6	30.4/31.2
18.0	32.0/38.3	33.4/31.5

Table 1: PSNR of face and remaining image for proposed scheme (content–based MPEG–4) and reference scheme (frame–based MPEG–4)

As can be seen in Tab. 1 for the sequence *Claire*, especially at low bit rates of about 9 to 12 kbit/s, nearly no improvements for the facial area are achieved, while the quality of the remaining image is significantly decreased. This is due to the fact that for this kind of sequences the most significant motion is located in the face of the person, while in the remaining image the motion is very low. Thus, also a coder without face detection uses the main part of the bit rate for coding the face. On the other hand, the content–based scheme has a higher overhead for transmitting two VOPs and additional shape information. This can lead to an even decreased quality in the remaining image, while the quality of the face is the same.

Akiyo (PSNR[dB]: face / remaining image)		
Bit rate (kbit/s)	Reference scheme (10Hz / 10Hz)	Proposed scheme (10Hz / 5Hz)
9.0	28.8/33.7	29.8/30.3
14.0	31.0/35.8	32.6/31.4
24.0	33.5/38.4	36.3/31.4

Table 2: PSNR of face and remaining image for proposed scheme (content–based MPEG–4) and reference scheme (frame–based MPEG–4)

However, as can be seen in Tab. 2 for the sequence *Akiyo*, also for this kind of sequences improvements for the facial area can be achieved. Therefore, not only the quality but also the frame rate for the remaining image has to be reduced. By this, the bit rate for transmitting the remaining image can further be reduced, since less VOPs have to be transmitted. Since the motion in the remaining image is very low, no annoying artefacts are introduced by reducing the frame rate.

Salesman (PSNR[dB]: face / remaining image)		
Bit rate (kbit/s)	Reference scheme (10Hz / 10Hz)	Proposed scheme (10Hz / 10Hz)
9.5	27.4/28.2	27.8/27.4
12.0	28.0/28.7	28.4/27.5
16.0	29.1/30.0	30.5/28.3

Table 3: PSNR of face and remaining image for proposed scheme (content–based MPEG–4) and reference scheme (frame–based MPEG–4)

Results for the second kind of sequences, i.e. where the amount of motion is similar in the facial area and the remaining image, are shown in Tab. 3 for the sequence *Salesman*. As can be seen, significant improvements are achieved for the image quality in the facial area, however again at the expense of the quality in the remaining image. Compared to the first kind of sequences, the possible improvements are higher since the motion is not only located in the facial area but equally distributed over the image.

For both kinds of sequences it can be seen, that with increasing bit rate the possible improvements are higher. This is due to the fact that at higher bit rates the overhead for coding of two VOPs and additional shape information has less impact.

7. CONCLUSIONS

An algorithm for content–based coding of videophone sequences conforming to the MPEG–4 standard has been proposed. With this algorithm, the facial areas of each image are coded and transmitted at a higher quality, at the expense of the image quality in the remaining image. Since in the case of video telephony the image quality of the face is very important for the communication partner, this leads to an improved subjective impression of the overall image quality, compared to a standard block–based hybrid coder which codes the image as a whole and thus can not address specific image areas separately.

In order to locate the face in each image, an automatic face detection algorithm has been applied, which subdivides the image into a facial part and a remaining image part. These two parts form two separate video object planes, which are then coded with the current MPEG–4 video verification model. In order to allocate different qualities to the two VOPs, either the quantization parameters or the frame frequency are chosen differently, or a combination of both is realized. Since up to now no rate control has been developed which allows to control these parameters automatically, they have been set manually.

The proposed algorithm has been compared to the current MPEG–4 video VM in frame based mode, i.e. coding each image as one rectangular VOP. Typical videophone sequences like *Claire, Salesman* and the MPEG–4 test sequence *Akiyo* in QCIF resolution have been coded at bit rates between 9 and 24 kbit/s. The results show significant improvements of the image quality in the facial area for those sequences which show a large amount of motion not only in the face but also in the remaining image. If most of the motion is already located in the facial areas, not only the quantization parameter but also the frame rate for the residual image has to be reduced in order to achieve improvements. Finally, it is shown that the improvements raise with increasing bit rate, since the overhead of coding two VOPs and the additional shape information has less impact.

Future work in this area will deal with developing a rate control which allows an automatic allocation of quantization step size and frame rate. Further, a scalable algorithm could be applied which codes the whole image at low quality, while for the facial area additional information is transmitted. However, this is not possible with the current MPEG–4 video VM, since the required mode for quality scalabilty is not specified yet.

8. ACKNOWLEDGEMENT

Parts of this work were carried out in the framework of the EU–ACTS project *MoMuSys*. For all simulations, the MPEG–4 video VM software provided by the *MoMuSys* project was used.

9. REFERENCES

[1] T. Aach, A. Kaup, R. Mester, "Statistical model–based change detection in moving video", *Signal Processing*, Vol. 31, No. 2, March 1993, pp. 165–180.

[2] M. Kampmann, J. Ostermann, "Automatic Adaptation of a Face Model in a Layered Coder with an Object–based Analysis–Synthesis Layer and a Knowledge–based Layer", *Signal Processing: Image Communications*, Vol. 9, No. 3, March 1997, pp. 201–220.

[3] R. Mech, P. Gerken, "Automatic segmentation of moving objects (Partial results of core experiment N2)", Doc. ISO/IEC JTC1/SC29/WG11 MPEG97/1949, Bristol, UK, April 1997.

[4] R. Mech, M. Wollborn, "A Noise Robust Method for Segmentation of Moving Objects in Video Sequences", ICASSP 97, Munich, Germany, April 1997.

[5] R. Mech, M. Wollborn, "A Noise Robust Method for 2D Shape Estimation of Moving Objects in Video Sequences Considering a Moving Camera", WIAMIS 97, Louvain–la–Neuve, Belgium, June 1997.

[6] MPEG–4 Video Group, "MPEG–4 Video Verification Model Version 6.0", Doc. ISO/IEC JTC1/SC29/WG11 N1582, Sevilla, Spain, February 1997.

[7] R. Rydfalk, "CANDIDE, A parameterised face", Internal Report Lith–ISY–I–0866, Linköping University, Linköping, Sweden, 1987.