IMPROVING IMAGE COMPRESSION- IS IT WORTH THE EFFORT ?

Ralf Schäfer, Guido Heising, Aljoscha Smolić
Heinrich-Hertz-Institut für Nachrichtentechnik Berlin GmbH
Einsteinufer 37, 10587 Berlin, GERMANY
schaefer@hhi.de, heising@hhi.de, smolic@hhi.de

ABSTRACT

This paper presents some arguments in favour of further research in video compression because of scarce spectrum resources for mobile applications and the lack of powerful tools for scalable coding. Finally some ideas and results for metadata based compression, improved prediction and scalable schemes are presented.

1. INTRODUCTION

Since the beginning of the 90th, image coding has developed from a mostly academic R&D field into a highly commercial business. Especially the standardization of MPEG-1 and MPEG-2 and of ITU-T H.26x has enabled a large variety of communication, multimedia and digital broadcast applications and services, which had not been possible without these technologies. Video streaming and teleconferencing with 20...n x 100 kbit/s, digital storage of video on CD-ROMS and DVDs with 1...5 Mbit/s and digital video broadcasting with 1.5...6 Mbit/s per program have created new businesses and especially revolutionised the whole broadcast industry.

After the establishment of MPEG-2 there were several voices stating, that it is not worth to put more effort into image coding, because the currently reached reduction to 1...2 % of the original data rate is considered as sufficient and the advances are becoming slower and slower. Already for the standardization of MPEG-4, more emphasis has been put on the enlargement of the functionalities than on increasing the compression rate. Only in the last phase of MPEG-4 this target has gained again more attention and with the definition of the ACE (Advanced Compression Efficiency) Profile the coding efficiency could be increased by 30 ... 50% compared to MPEG-2. Doubling compression rate is also the target of H.26L [3], the follow-up standard of H.263++ currently being developed within ITU-T SG16.

On the other hand we are facing an explosion of available bandwidth, both for storage and transmission. Every ten years the costs for digital storage are reduced by a factor of 100 or more and this trend will go on for the next 20 years. Already today, DVDs and harddiscs can store hours of video. Glass fibres can carry Terabits/sec and broadband Internet access via cable modems or xDSL is becoming widely available.

So why are we continuing R&D in image compression? The algorithms are becoming more and more complex and the advances are becoming marginal. What is still missing today and what are the applications asking for higher compression rates?

From the author's point of view there is one very important application field justifying this effort and one missing technology: The application field is mobile video services and the missing technology is scalability, which is most important for video over the Internet in general but over wireless channels in particular. Spectrum over the air is a scarce resource which cannot be enlarged. Even with the advent of UMTS/IMT 2000, the available data rate will be restricted to 144 kbit/s for mobile reception, higher rates up to 2 Mbit/s will only be available in microcells, i.e. in restricted areas like offices. 144 kbit/s is still a rather limited data rate for video transmission, if available coding schemes like MPEG-4 or H.263++ are used. On the other hand, mobile Internet traffic will increase at the same speed as Internet traffic in general, while the number of mobile terminals is exploding. Therefore it is foreseeable, that over the air bandwidth will remain a permanent bottleneck, which justifies all efforts to squeeze video even further. Although higher compression means more complex algorithms and more processing power, one should not be afraid about this, as Moore's law will solve these problems anyway.

This paper will describe some promising techniques, by which these objectives may be obtained, in less or more detail. Special emphasis is put on new motion models and on fine granularity scalability, which have recently been developed at HHI.

2. METADATA BASED CODING

A promising approach to increase the coding efficiency consists in the usage of metadata as currently defined in MPEG-7 for coding. This is a quite new approach and almost no literature on such techniques is available.

MPEG-7 will for instance provide the means for description of shot boundaries and corresponding transitions (cut, fade) within a video stream [1]. This information can be obtained directly from the production or can also be extracted automatically using cut- or fade-detection algorithms. If these metadata is available prior to encoding, it may help to improve the encoding process, through the design of a specific encoding strategy using the information of shot boundaries and their characteristics. Predictive coding can be switched off at scene cuts where no motion compensation model is of any use. Fade-in and fade-out can be coded separately and combined afterwards instead of using an irrelevant video sequence model for predictive coding. MPEG-7 will also enable the definition of regions or objects within a video stream. These metadata can also be available from production (blue screen, editing) or can be generated using segmentation algorithms. If

available, this information can be used for separate coding of moving objects, of the background, and of transparent layers. This overcomes the drawbacks of standard non-hinted compression algorithms near occluding object boundaries, where the motion compensation model cannot predict the progressive disclosure or disappearance of object parts. Such an approach could perfectly combine and benefit from metadata about objects (MPEG-7) and object-based coding (MPEG-4).

These examples above are related to structural metadata about video that can be provided by MPEG-7. It is also imaginable to exploit signal-based low-level descriptors to improve encoding. For instance knowledge about color and texture histograms or distribution could be used to adjust bit allocation or quantizer values. Another class of MPEG-7 descriptors carries information that is very promising for being exploited to improve encoding, the various motion descriptors [1]. The *Motion Trajectory* descriptor provides information about the 2D or 3D motion of objects within a scene, that could be exploited for motion estimation and compensation. The *Camera Motion* descriptor is directly related to the 3D camera operation in a scene. This information could also be recorded and made available while capturing the video, but it is also possible to estimate it from the video. Information about the 3D camera operation is also useful to improve motion estimation and compensation, especially for advanced prediction schemes like *Global Motion Compensation (GMC)* or *Sprite Coding* as defined in MPEG-4 [2]. In case of the *Parametric Motion* descriptor the MPEG-7 metadata is nearly identical with MPEG-4 encoding parameters of GMC and sprite coding, which allows direct use or even joined optimization of both in a single system. The motion of image regions (entire frame is a special case) is basically characterized by a set of motion parameters related to a certain motion model, like the affine or perspective model.

Another direct correspondence can be established between the *Mosaic* description scheme in MPEG-7 and sprite coding in MPEG-4. Mosaics will be defined in MPEG-7 for example for visualization and representation of the visual content of a whole shot by a single image. If a mosaic is available prior to encoding it can be used for very efficient compression of the shot.

3. IMPROVING PREDICTION

A second very promising research area for further improvement of video compression is the further reduction of temporal redundancies by new prediction methods. Recent work at the University of Hannover has shown, that alias compensation in the prediction loop is a powerful method, which makes it possible, to get a benefit from 1/8 pel accuracy in motion compensation. Recent results obtained at the University of Erlangen prove that long-term memory motion compensated prediction also provides an increased coding efficiency. Other promising approaches in this field include usage of 4x4 blocks or more complex motion models (affine) for block-based motion compensation. The aforementioned methods follow the classical hybrid coding approach, the reduction of spatial and temporal redundancy by exploration of statistical models of the video signal. They do not use any assumption about the video content and can therefore be used for any kind of material. This technology is currently under investigation for the new ITU-T standard H.26L [3].

In some application scenarios certain a priori knowledge about the video content is available and can be exploited for encoding. Such concepts combine multiple object and motion models in a layered coding scheme [4]. This allows the integration of knowledge-based coders with general purpose coders and thus provides maximum coding efficiency while being suitable for any kind of video material. The typical example is a videophone or -conferencing system, with a person captured by a static camera. The video can be subdivided into several regions corresponding to physical objects as shown in figure 1. Then each region can be encoded separately using a specifically adapted coding scheme that exploits a priori knowledge. A standard mode is always available as fallback for regions that can not be assigned any specific coder, or if the coding parameters can not be measured accurately. This ensures that theoretically a layered coder can never be outperformed by a standard general purpose coder.

Over the last years 3-D model-based coding of humans, especially faces, has been a very active field of research, with impressive results in terms of coding efficiency. The means for standardised transmission are already provided by MPEG-4 SNHC [2]. However, such technology is not yet mature enough to be included in real-world coding systems, mainly due to the difficulty of robust and reliable extraction of the facial animation parameters, which is crucial for a wide acceptance of such technology.

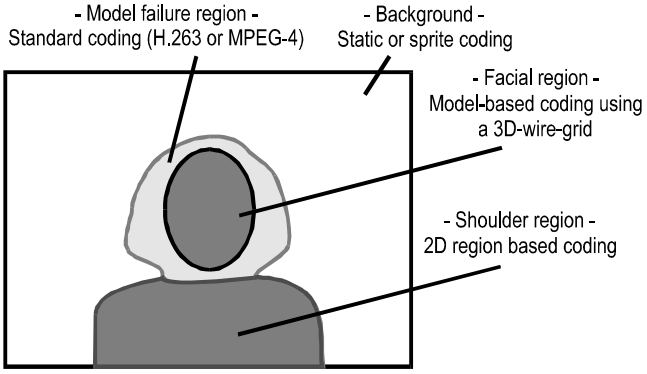

Fig.1: Separation of video into regions corresponding to physical objects in a layered coding scheme.

In the videophone scenario the background is modelled as a static region that has to be encoded only once, possibly with updates of initially occluded parts. This concept is easily extended to moving camera sequences by incorporation of a global motion model. Such technology is also already available in MPEG-4. GMC can be used for predictive coding between consecutive frames (neglecting B-frames) as an alternative to local motion compensation.

Significant improvements in terms of subjective image quality can be achieved with sprite coding techniques. But although syntactically possible, the static sprite coding scheme in MPEG-4 is not designed for on-line applications.

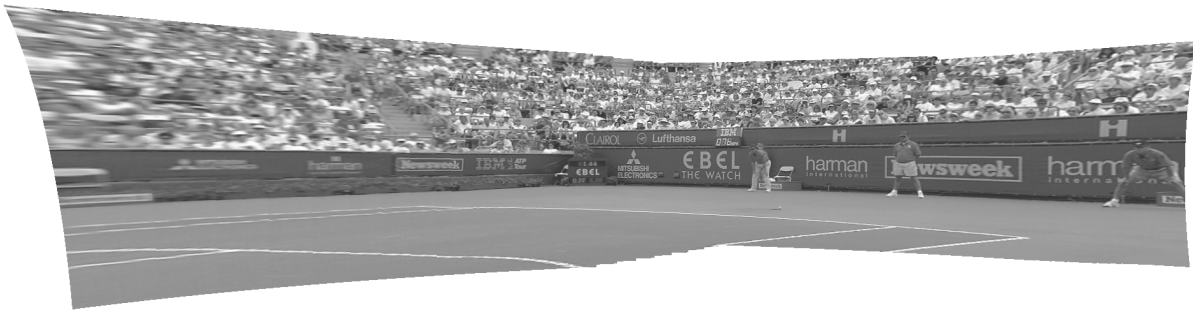

Fig.2: Background mosaic (sprite) for MPEG-4 test sequence 'Stefan' over 250 frames.

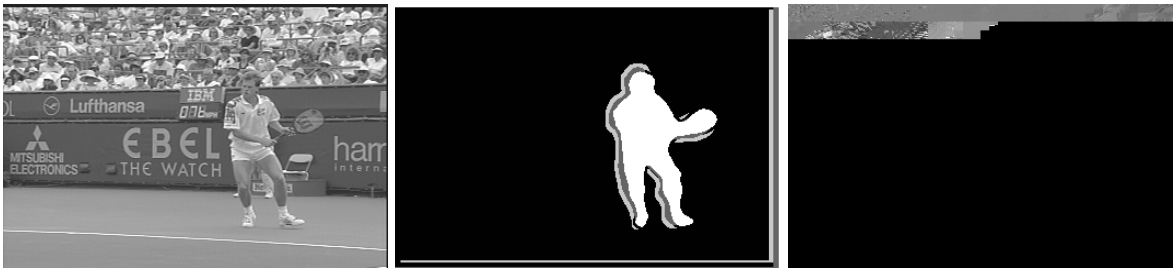

Fig.3: Coded content for 2nd frame of sequence 'Stefan', left: original frame, middle: new appearing content, right: rescanned and reorganized pixels to be encoded.

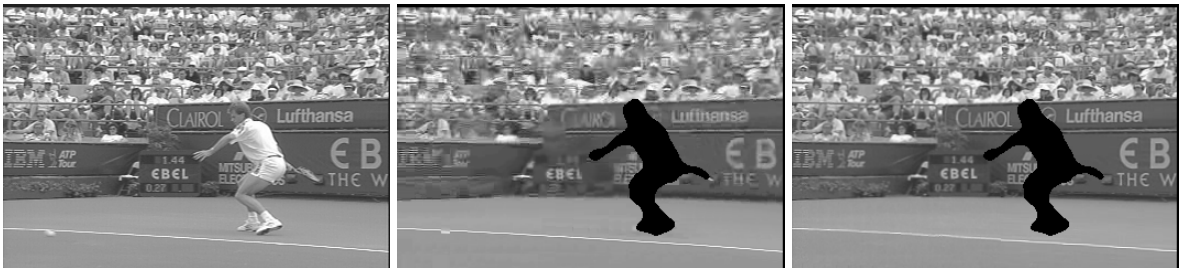

Fig.4: Coding results for sequence 'Stefan' (352x240, 30 Hz, 250 frames), frame 100, 305 kbit/s, left: original, middle: MPEG-4 arbitrary shape coder, right: proposed on-line sprite coding scheme.

The sprite is supposed to be generated off-line prior to encoding.

A new approach to sprite coding is proposed in [5], that is suitable for on-line (real-time, low-delay) applications, while featuring the advantages of MPEG-4 static sprites in terms of visual quality. This method is derived from the process of sprite generation also referenced as mosaicking. The sprite is built up from the constructing video sequence, by estimating the global motion for each frame with respect to a certain higher order motion model (affine, perspective, polynomial) and a fixed spatiotemporal reference. Each frame is warped towards the common reference controlled by the estimated motion parameters. Those parts of the background region, that are not yet registered in the sprite are updated from each image. The result is a knowledge-based long-term description of motion and visual content, an example is shown in figure 2.

New content typically appears at the image borders and by uncovering behind foreground objects, as shown in figure 3, middle. Only the updates have to be transmitted to the decoder. The location of these pixels can be calculated at the decoder too, if the segmentation masks and motion parameters are transmitted. Therefore the pixels can be rescanned and reorganised into blocks and macroblocks as shown in figure 3, right, that can be coded very efficiently using a standard DCT method. Figure 4 compares coding results obtained using the proposed method and a standard MPEG-4 arbitrary shape coder. The visual quality of the presented sprite coding scheme is significantly better at the same bitrate. Drawbacks of the proposed scheme are that errors of the global motion estimation and segmentation result in annoying artifacts. Also a significant change of already coded background information can not be handled. These drawbacks can be overcome by sending additional texture updates.

4. SCALABLE CODING

As far as scalable coding is concerned, new forms of wavelet decomposition have the potential for more efficient scalability compared to what is available in the DCT based coding schemes of MPEG. In its original form, wavelet analysis is a linear tool, which leads to various artefacts when coding image or video sequences containing sharp edges. However new wavelets, based on non-linear transforms, are able to preserve significant structures inside scenes such as edges, textures etc. A general and flexible framework for the wavelet construction is provided by the lifting scheme, which enables the modification of existing

wavelet decompositions, and the inclusion of nonlinearities and/or data-dependencies. On the other hand, nonlinear (e.g., morphological, adaptive) filtering offers the great advantage that it allows to preserve pertinent details even at low-resolution levels. Moreover, these details may be present in the approximation subband, contrary to the case of linear wavelet decompositions, where such details are always located in the detail subbands which then also need to be transmitted. Therefore, nonlinear decompositions may give rise to a substantial decrease of bit rate.

Furthermore recent developments in the context of wavelet-based still image coding, like the emerging JPEG-2000 standard [6], proposes the usage of bit plane coding in combination with context based arithmetic coding to achieve scalability in terms of quality and spatial resolution. Currently, we are investigating how to combine these efficient coding strategies with hybrid predictive video coding systems. With this approach we are targeting at video streaming applications over IP based heterogeneous networks where we encounter the problem of possible packet losses or varying bandwidth over time. To overcome this problem the bitstream should be adapted to the network quality of service by incorporating error resilience tools and different types of scalability, such as temporal, spatial and quality scalability.

Our approach follows the ideas which have been considered in MPEG-4 fine granular scalability (FGS) [7] where a base layer is coded using a hybrid predictive loop and where an additional enhancement layer delivers the progressively encoded residue between the reconstructed base layer and the original frame. This is done by coding the bit planes of the DCT transformed image domain residue. Starting from the most significant bit plane a Huffman table based run length coding is employed. In contrast to the MPEG-4 FGS scheme we reorder the residue in subbands as shown in figure 5 for the case of 4x4 transform blocks [8]. This enables us to better exploit the remaining intra band correlations. In addition, by coding a bit plane starting from the DC up to the highest AC band using a global zig-zag scan we can assure a more or less spatially equal distribution of the increase of reconstruction quality. For encoding the subbands are split into different subsources, namely the significance bits, indicating the bit plane of their first appearance, sign bits and refinement bits to describe their change in amplitude. The first two binary sources are coded by a context based binary arithmetic entropy coder similar to the one used in [9] to achieve a higher coding gain. For the refinements bits a simple binary arithmetic coder is used. Our new approach provides the functionality of fine granular quality scalability. However, for all such FGS schemes known to the authors the loss of coding efficiency compared to a single layer approach is still high. Therefore we are currently investigating the incorporation of drift prediction in the enhancement layers and the combination with spatial and temporal scalability.

5. CONCLUSIONS

Research in video coding and further significant progress in compression efficiency are absolutely required

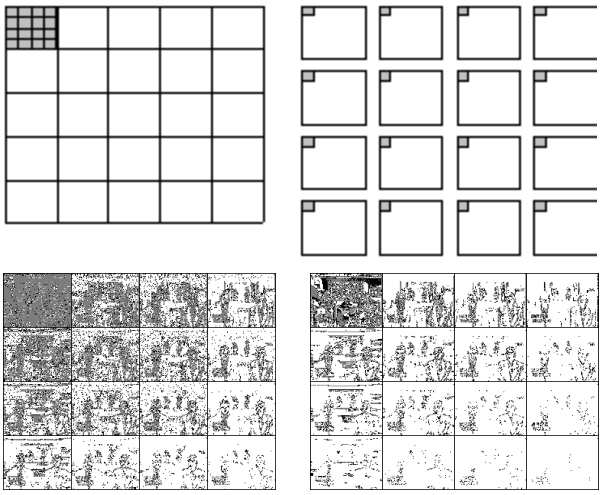

Fig.5: Building subbands by reordering transform coefficients of 4x4 blocks (top left) in 16 subbands (top right). In the example at the bottom the significance bits of the subbands are marked in black and the refinements bits in gray for the first (left) and the third (right) bit plane.

due to the lack of spectrum for future mobile video services. Furthermore there are no efficient methods for scalable coding, which would be a very useful tool for video applications over any channel with varying data rate or unpredictable transmission quality like the (mobile) Internet. A promising approach to reach higher compression rates could be in the usage of metadata and some first ideas on this topic have been presented. The other most promising method is to further improve prediction by using better motion models. Finally a novel approach for scalable coding and an outlook to further improvements are presented.

REFERENCES

[1] ISO/IEC/JTC1/SC29/WG11, "MPEG-7 Overview", Doc. N3349, Noordwijkerhout, Netherlands, March 2000.

[2] ISO/IEC/JTC1/SC29/WG11, "MPEG-4 Video VM 16.0", Doc. N3312, Noordwijkerhout, Netherlands, March 2000.

[3] ITU SG 16 Q.15, ”H.26L Test Model Long Term Number 3 (TML-3)”, Doc. Q15J08, Osaka, Japan, May 2000.

[4] A. Smolić and T. Sikora, ”Coding of Image Sequences Using a Layered 2-D/3-D Model-Based Coding Approach”, Proc. PCS’97, Picture Coding Symposium, Berlin, Germany, September 1997, pp. 541-546.

[5] A. Smolić, T. Sikora and J.-R. Ohm, "Long-Term Global Motion Estimation and its Application for Sprite Coding, Content Description and Segmentation", IEEE Trans. on CSVT, Vol. 9, No.8, pp. 1227-1242, December 1999.

[6] ISO/IEC CD 15444-1; JPEG-2000 Image Coding System, Commitee Draft, Version 1.0, Dec. 1999.

[7] ISO/IEC/JTC1/SC29/WG11, "VM of ISO/IEC 14496-2 MPEG-4 Video FGS (v4.0)", Doc. N3317, Noordwijkerhout, Netherlands, March 2000.

[8] G. Blättermann, G. Heising, D. Marpe, ”Proposal for a quality scalable mode in H.26L”, submitted to ITU SG 16 Q.15, Doc. Q15J24, Osaka, Japan, May 2000.

[9] G. Heising, D. Marpe, and H. L. Cycon, “A Wavelet-Based Video Coding Scheme Using Image Warping“, IEEE Int. Conf. on Image Proc., Chicago, IL, USA, Oct. 4-7, 1998.