A Generic and Automatic Content-based Approach for Improved H.264/MPEG4-AVC Video Coding

Patrick Ndjiki-Nya, Tobias Hinz, Aljoscha Smolić, and Thomas Wiegand
Image Processing Department
FhG Heinrich-Hertz-Institut (HHI)
Berlin, Germany
{ndjiki/hinz/smolic/wiegand}@hhi.de

***Abstract*—A new content-based approach for improved H.264/MPEG4-AVC video coding is presented. The framework is generic because it is based on a closed-loop texture analysis by synthesis algorithm that can automatically identify and recover from video quality impairments through artifact detectors and appropriate countermeasures. The algorithm is flexible, for it can in principle be integrated into any standards-compliant video codec. The fundamental assumption of our approach is that many video scenes can be classified into subjectively relevant and irrelevant textures. The texture categorization is thereby done by a texture analyzer (encoder side), while the corresponding texture synthesizer performs the replacement of the subjectively irrelevant textures (decoder side), given the side information generated by the texture analyzer. When implementing the proposed approach into an H.264/MPEG4-AVC codec, bit rate savings of up to 33.3% compared to an H.264/MPEG4-AVC video codec without our approach are reported.**

Content-based coding; Video coding; Texture analysis; Texture synthesis; H.264/MPEG4-AVC; MPEG-7

I. INTRODUCTION

Content-based video coding aims to achieve bit rate reduction of compressed video sequences, while preserving high visual quality of decoded data. Content-based video coding approaches typically decompose the sequence into spatially, temporally, or spatio-temporally coherent regions. The coherence of a region is thereby measured based on motion, color, and/or texture features. These object attributes are typically described via compact representations given the video coding framework.

Content-based video coding schemes can be clustered into low-, mid- and high-level techniques based on the semantic significance of the objects they are tuned to identify. Systems with the capability of automatically capturing semantically meaningful objects can be seen as high-level approaches. They represent the most challenging types of content-based video coding and will probably be beyond reach for several years to come. Mid-level techniques reduce coding costs by processing different regions with similar motion, texture or color characteristics together, e.g. see [1],[3],[4]. For such algorithms, the semantic content of the identified objects is irrelevant. However, these objects must be described consistently in space and time. Low-level approaches can be seen as coding techniques that exclusively rely on spatial or temporal features for the detection of homogeneous regions and do not incorporate any inference mechanism concerning tracking and/or spatial consistency of identified regions [5],[6].

In this paper, a generic, closed-loop, mid-level content-based video coding scheme is proposed. It is assumed that many video scenes can be classified into detail-relevant and detail-irrelevant textures. Detail-irrelevant textures are highly texturized regions that are displayed with restricted spatial accuracy (e.g. flowers in the well-known test sequence "Flower Garden"), while the other textures are referred to as detail-relevant textures. It is further assumed that for detail-irrelevant textures, the viewer perceives the semantic meaning of the displayed texture rather than the specific details therein. Many detail-irrelevant texture regions are costly to code, when using the mean squared error (MSE) criterion as the coding distortion. Thus, in this paper, it is argued that MSE is not an adequate distortion measure for efficient coding of detail-irrelevant textures and it is claimed that global similarity measures, e.g. MPEG-7 descriptors [7], are better suited for assessing the distortion of such textures, as no MSE-accurate regeneration of these is requested. Often, the required bit rate for transmitting detail-irrelevant textures can be significantly reduced, if the number of bits needed for their description using the modified distortion measure is smaller than the number of bits for the description using MSE. The largest problem to solve within this approach is to perform the algorithm in an automatic way. Moreover, it also must be generic in that if a detail-irrelevant texture becomes detail-relevant (e.g. when zooming in on the detail), the algorithm must seamlessly switch between the two ways of coding. These challenges have been addressed in this work.

The remainder of the paper is organized as follows. In Section II, the principle of the automatic analysis-synthesis loop is introduced. The components of the loop, including texture analysis, texture synthesis, and video quality assessment among others, are presented in corresponding subsections. Finally, in Section III, the experimental results are shown.

0-7803-9134-9/05/$20.00

II. Principle Of The Closed-Loop Video Analysis-Synthesis Algorithm

Mid-level video coding schemes can be found in the literature. One of the early descriptions of this coding strategy was published by Wang and Adelson [2]. They assume that any video sequence can be represented as a set of overlapping layers, where a layer is a description of a coherent motion region. Ordering the layers in depth and applying the rules of compositing ideally yields the original video sequence. Dumitraş and Haskell proposed a content-based video coding method by texture replacement [4]. Replaceable textures are identified and removed from the corresponding regions of the original pictures. The resulting video sequence is encoded and the extracted parameters of the removed textures transmitted to the decoder. Absent or very slow global motion of removable textures as well as few scene objects with moderate motion are assumed, which are very strong constraints that confine the practical usability of this approach to a limited number of applications. Both algorithms [2],[4] are open-loop, i.e. there is no mechanism to identify and where necessary alleviate artifacts due to erroneous analysis or synthesis, which yields unregulated (subjective) video quality at the decoder output.

In this work, we have developed the closed-loop analysis-synthesis algorithm depicted in Fig. 1. The incoming video sequence is divided into overlapping groups of pictures (GoP). The first GoP consists of the first I picture of the sequence and the last picture of the GoP is the first P picture. Between these I and P pictures are B pictures. For example, when 3 B pictures are used, the first GoP has the structure $IBBBP_1$ in temporal order. The second GoP consists of the last picture (the P_1 picture) of the first GoP and the next P picture. In our example, the second GoP has the structure P_1BBBP_2. I and P pictures are so-called key pictures and coded using MSE distortion and an H.264/MPEG4-AVC encoder. B pictures (between the key pictures) are candidates for a possible partial texture synthesis and are also otherwise coded using MSE distortion and H.264/MPEG4-AVC.

Each GoP is analyzed by the texture analyzer (TA) and synthesized by the texture synthesizer (TS), given the (quantized) side information generated by the TA. The synthesized GoP is then submitted to the video quality assessment unit (VQA) for detection of possible spatial or temporal impairments in the reconstructed video.

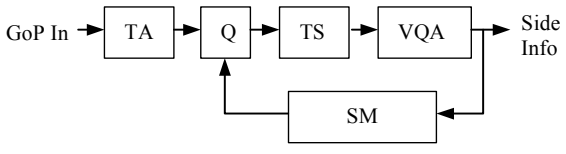

Fig. 1 - Principle of the closed-loop analysis-synthesis video coding approach

In the subsequent iterations, the degrees of freedom of the system are explored by a state machine (SM) in the quest of even better side information. Once all relevant system states have been visited for the given input GoP, a rate-distortion decision is made and the optimized side information is transmitted to the decoder. Detail-irrelevant textures for which no rate-distortion gains can be achieved are coded by the reference codec, which acts as fallback coding solution. Furthermore, the GoP structure used in our framework allows a seamless change from detail-irrelevant to detail-relevant coding, as the key pictures are coded based on MSE. In the following, the modules of our content-based video coding approach are explained in-depth.

A. Texture Analyzer (TA)

The TA identifies detail-irrelevant textures and generates the corresponding side information, which yields a spatio-temporal decomposition of the video sequence.

The principle of the TA is depicted in Fig. 2. It consists of an optional spatial texture analysis (STA) module, which creates initial picture partitions that typically provide the subsequent motion analysis with very helpful hints w.r.t. amount and size of regions. However, the STA module can be bypassed if desired. Motion analysis is done based on the estimated dense motion field between two consecutive pictures [8]. The motion field is split (MS module) into homogeneous motion regions based on a robust, iterative maximum-likelihood process called M-estimation [1]. The latter is a model-fitting approach that detects outliers within a dataset a posteriori and without any prior knowledge of outlier characteristics. The observations are a set of motion vectors in our specific framework, while outliers can be seen as motion vectors that reveal different motion properties than the inliers. Motion homogeneity is defined w.r.t. the perspective motion model [1], which was selected due to its ability to describe translation, rotation, and scaling of a planar patch in 3-D as we assume this geometry also for our synthesized textures. The M-estimator minimizes the influence of outliers on the model optimization by penalizing motion vectors yielding high modeling costs. The MSE-based cost function is thereby defined as the deviation between the observed [8] and the modeled dense motion field.

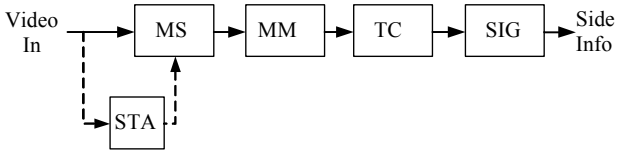

Fig. 2 - Principle of the spatio-temporal texture analyzer

In order to overcome the typically over-segmented result of the MS module, homogeneous texture segments that have similar motion parameters are fused to one single region by the motion merger (MM) module. The similarity criterion can be summarized as follows. Two homogeneous regions are merged when the modeling costs of the overall region do not exceed the single costs.

Temporally consistent label assignment to homogeneous textures is ensured by setting up a "texture catalog" (TC). Each identified texture is mapped to one of the indexed textures if similar or added to the texture catalog otherwise. Similarity is measured w.r.t. MPEG-7's scalable color descriptor [7], which is basically a color histogram in the HSV color space. MPEG-7 recommends corresponding distance measures for similarity

evaluation [7]. Two feature vectors are considered to be similar when the distance between them is smaller than a given threshold.

As a result of the texture analysis process, a mask sequence showing the detail-irrelevant textures of the considered GoP is generated as well as a perspective motion parameter set and a control parameter for each detail-irrelevant texture region, using the side information generator (SIG) module. The control parameter indicates whether the current texture region (residing in a B picture) is to be synthesized using the first or the last picture of the GoP, which are non-synthesized I or P key pictures, as mentioned above. The motion parameter set describes the texture mapping operation from the key picture towards the missing texture region in the corresponding B picture.

B. Quantization (Q)

The motion parameters generated by the texture analyzer are uniformly quantized and their quantization step size can be varied.

C. Texture Synthesizer (TS)

Two major texture types are considered at the texture synthesis (TS) stage; rigid (e.g. grass, flowers) and non-rigid textures (e.g. water, clouds). Non-rigid textures typically feature local motion activity, which is not the case for rigid video textures.

The texture synthesizer depicted in Fig. 3 is designed for rigid objects. The underlying hypothesis of this approach is that the picture-to-picture displacement of the objects can be described using the perspective motion model. The texture synthesizer warps the texture from the first or the last (key) picture of the considered GoP towards a synthesizable texture region identified by the texture analyzer as illustrated in Fig. 3, given a motion parameter set and a control parameter (cp. Section II.A.).

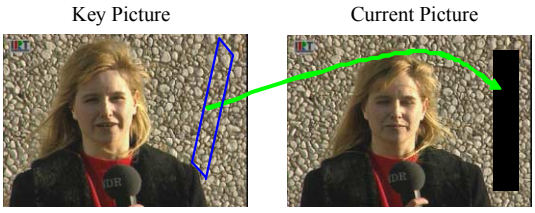

Fig. 3 - Texture synthesizer filling rigid texture region identified by texture analyzer using left key picture

The texture synthesizer for non-rigid textures is depicted in Fig. 4. In this approach, a texture is modeled based on Markov Random Field methods [9]. For that, each texture sample is assumed to be predictable from a small set of spatially neighboring samples and independent of the rest of the texture.

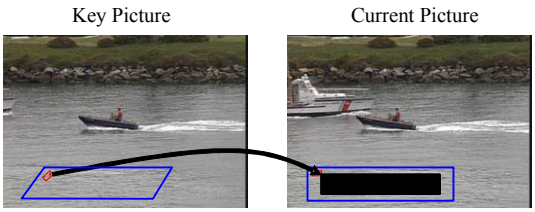

Fig. 4 - Texture synthesizer filling non-rigid texture region identified by texture analyzer using left key picture

The initial step of warping a texture region from the key to the current picture is the same as in Fig. 3. The causal neighborhood of each sample of the missing texture is compared to the neighborhoods of the warped samples within a restricted area (typically 3x3 samples) in order to capture local motion characteristics of the given texture. Hence, motion estimation has to be done at the decoder with a very limited search range. The warped sample with the most similar neighborhood (MSE) is finally copied to the location of the corresponding sample of the missing texture.

D. Video Quality Assessment (VQA)

Video texture synthesis can potentially yield annoying spatial and/or temporal artifacts. Objective measures indicating impairments are presented in the following. The purpose of our methods is to infer the impact of texture synthesis from specific features that can be automatically extracted from the video.

It is obvious that, given a correct texture analysis within the selected detail-irrelevant regions, most spatial impairments will occur at the transitions from synthesized to original textures in the form of spurious edges (cp. Fig. 3). Thus, the spatial quality assessor consists of a relatively simple linear anisotropic edge detector, the Kirsch detector [10]. The reason for selecting the Kirsch detector over other edge detectors is described in [11]. Only the vertical and horizontal directionalities of the detector are used, as the synthesizable texture regions are composed of square macroblocks of H.264/MPEG4-AVC [12]. The spatial quality assessor compares the ratio of the edge samples found in the synthesized and original pictures with a critical threshold [11]. In case the threshold is exceeded, the synthetic picture is classified as erroneous.

Temporal VQA consists in evaluating the motion properties of the synthesized textures w.r.t. the original reference. The basis of our temporal impairment detector is the estimated dense motion field in the relevant regions. It is analyzed (on a macroblock basis) whether the motion vectors of the original and corresponding synthetic texture come from the same distribution. The distance between two distributions can thereby be determined with any adequate distance measure (e.g. l_1 norm). This approach allows for small deviations between original and synthetic motion, which is in line with the fundamental assumptions of our content-based video coding approach.

E. State Machine (SM)

The state machine explores relevant system states. The quantizer resolution is varied, which allows examination of the sensitivity of synthesis results to the accuracy degradation of the motion parameters. The impact of motion description complexity on the texture synthesizer is also explored, i.e. the motion of detail-irrelevant regions is successively described with 8 (perspective model), 6 (affine model) and 2 (translational model) of the 8 perspective motion parameters and the corresponding synthetic GoPs are evaluated with the VQA module.

III. Experimental Results

We have integrated our approach into an H.264/MPEG4-AVC codec. The test sequences "Concrete", "City", "Preakness", and "Coastguard" are used to demonstrate that an approximate representation of some rigid and non-rigid textures can be done without subjectively noticeable loss of quality.

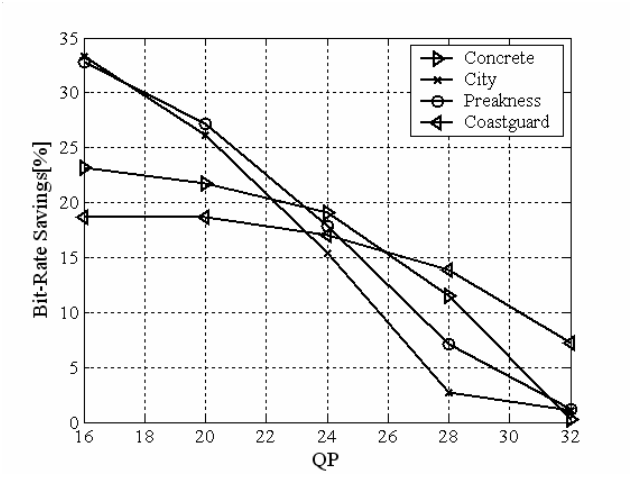

Fig. 5 - Bit rate savings w.r.t. quantization accuracy

The following set-up was used for the H.264/MPEG4-AVC codec. Three B pictures, one reference picture for each P picture, CABAC (entropy coding method), rate distortion optimization, 30 Hz progressive video at CIF resolution. The quantization parameter QP was set to 16, 20, 24, 28 and 32. Fig. 5 depicts the bit rate savings obtained for each of the test sequences. Here we have assumed and verified through visual inspection that the MSE coded and synthesized textures cannot be distinguished. It can be seen that the highest savings are measured for the highest quantization accuracy considered. The most substantial bit rate savings (33.3%) are measured for the "City" sequence. The bit rate savings decrease with the quantization accuracy due to the fact that the volume of the side information remains constant over the different QP settings. All results are derived from decoding bit-streams and the encoder is run automatically for each sequence. Sequences for subjective evaluation can be downloaded from http://ip.hhi.de/imagecom_G1/closed_loop.htm.

IV. Conclusions And Future Work

An automatic, closed-loop, content-based approach for improved H.264/MPEG4-AVC video coding was presented. The fundamental hypothesis of our approach is that textures in a video sequence can be classified into two classes: detail-relevant and detail-irrelevant. Our experiments show that our algorithm yields bit rate savings of up to 33.3% compared to an H.264/MPEG4-AVC video codec without our approach.

The complexity of the analysis-synthesis loop will be addressed in future implementations. The state machine will, for instance, be provided with a memory to avoid considering all discrete system states for each GoP. This will allow setting the initial system's state for the current GoP to the best state, in the rate-distortion sense, selected for the previous GoP. Synthesis of textures with local motion (e.g. water) will be improved.

References

[1] J.-R. Ohm, "Multimedia Communication Technology", ISBN 3-540-01249-4, Springer, Berlin Heidelberg New York, 2004.

[2] J. Y. A. Wang and E. H. Adelson, "Representing Moving Images with Layers", IEEE Trans. on IP, Special Issue on Image Sequence Compression, Vol. 3, No. 5, pp. 625-638, September 1994**.**

[3] P. Salembier, L. Torres, F. Meyer, and C. Gu, "Region-based Video Coding using Mathematical Morphology**"**, Proc. of the IEEE , Vol. 83, No. 6, pp. 843–857, June 1995.

[4] A. Dumitraş and B. G. Haskell, "An Encoder-Decoder Texture Replacement Method with Application to Content-based Movie Coding", IEEE Trans. on CSVT, Vol. 14, No. 6, pp. 825-840, June 2004.

[5] E. Steinbach, T. Wiegand, and B. Girod, "Using Multiple Global Motion Models for Improved Block-based Video Coding", Proc. ICIP 1999, Vol. 2, pp. 56-60, Kobe, Japan, October 1999.

[6] A. Smolić, Y. Vatis, H. Schwarz, and T. Wiegand, "Improved H.264/AVC Coding using Long-term Global Motion Compensation", Proc. VCIP 2004, SPIE Visual Comm. & IP, pp. 343-354, San Jose, CA, USA, January 2004.

[7] ISO/IEC JTC1/SC29/WG11/N4358, "Text of ISO/IEC 15938-3/FDIS Information technology – Multimedia Content Description Interface – Part 3 Visual", Sydney, Australia, July 2001.

[8] M. J. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-smooth Flow Fields", Computer Vision and Image Understanding, Vol. 63, No. 1, pp. 75-104, January 1996.

[9] L.-Y. Wei and M. Levoy, "Fast Texture Synthesis using Tree-structured Vector Quantization", Proc. of SIGGRAPH 2000, New Orleans, Louisana, USA, July 2000.

[10] J. R. Parker, "Algorithms for Image Processing and Computer Vision", Wiley Computer Publishing, USA, 1997.

[11] P. Ndjiki-Nya, M. Kootz, and T. Wiegand, "Automatic Detection of Video Synthesis Related Artifacts", Proc. ICASSP 2004, Montreal, Canada, May 2004.

[12] ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC: "Advanced Video Coding for Generic Audiovisual Services", 2003.