HIGHLY SCALABLE 2D MODEL-BASED VIDEO CODING SCHEME USING WARPING MOTION COMPENSATED TEMPORAL FILTERING

Mingyou Hu, S. Worrall, A. H. Sadka, A. M. Kondoz

Centre for Communication Systems Research (CCSR),
University of Surrey, Guildford,
Surrey, United Kingdom

ABSTRACT

The development of multimedia technology and communication infrastructure makes it possible for universal multimedia access (UMA). UMA aims to enabling users to access the multimedia content freely by allowing adaptation of the media context to diverse network characteristics, end-terminal capability and user preference. Video representation and coding play an important role to achieve UMA, which should support several new functionalities, such as efficient scalable coding, and object manipulation. In this paper, a highly scalable 2D model-based video coding scheme is presented, in which content-adaptive scalable model and warping motion compensation is included. Motion compensated temporal filtering (MCTF) is conducted before scalable motion vector coding and wavelet-based texture coding. The performance of this coding system is evaluated through intensive experiments.

1. INTRODUCTION

Much research has been conducted for video coding in order to achieve UMA. Some recently standardised video coders, such as MPEG-4 [1] and H.264 [2], can achieve excellent compress performance. For example, H.264 can save about 60% bit rate when compared with MPEG-4. However, these standardised video coders can not provide highly scalable bit stream to diverse network characteristics, end-terminal capability and user preference.

Recently, scalable video coding techniques have been intensively studied and proposed to MPEG-21 SVC, such as [3-6]. Among these techniques, one promising technique is proposed by Microsoft Lab [4], which is called a Barbell lifting implementation of the wavelet transform. Within this technique, the lifted wavelet transform are adopted during temporal filtering. The motion compensation supports adaptive block size (similar to H.264/MPEG-4 AVC), overlapping block motion compensation (OBMC) and ¼ pel precision for motion vectors. The motion vectors are encoded in an embedded bitstream, in a coarse to fine fashion. The wavelet coefficients are encoded to provide SNR scalability. Good coding performance is reported for this technique.

Another promising technique is a scalable extension of the H.264/AVC video coding standard [6]. To achieve an efficient scalable bit-stream representation, the temporal dependencies between pictures are exploited by using an open-loop subband approach. In order to provide spatial scalability, a pyramid structure is employed. Although MCTF is independently applied in each spatial layer, a large degree of inter-layer prediction is incorporated. A remarkable feature of this hybrid scalable video coding scheme is that most components of H264/AVC are used. Experimental results indicate that this hybrid scalable coding method is capable of providing a coding efficiency nearly comparable to that of an original H264/AVC encoding [6].

Although efficient scalable coding techniques have been proposed, further research is required to improve the coding scalability, such as motion scalability and object-scalability. The objective of this paper is to develop an efficient scalable 2D model-based video coding scheme, which tries to achieve highly scalable video coding with high compression efficiency. Using model-based video coding can improve the visual performance of video coding and improve the motion compensation efficiency. The main differences of the proposed scheme to other approaches are:

- Motion compensation is conducted in object domain, instead of frame domain, which can improve the efficiency of motion compensation for the high motion video frames;
- Mesh-based motion estimation is conducted instead of block-based motion estimation to reduce the block artefact for very low bit rate coding;
- Rate-distortion optimised rate control among video objects, video frames, and among motion and textures.

The second section of the paper gives the detailed description of the proposed scheme. The third section

presents and discusses some experimental results obtained using the proposed schemes. The fourth section draws the conclusions and defines the future research direction

2. SCHEME DESCRIPTION

Figure 1 shows the general structure of the proposed system. In this system, it is assumed that the video frame has been segmented into several video objects with different motion patterns and the object shape has been encoded. The detailed descriptions of proposed scheme are included in the following sections.

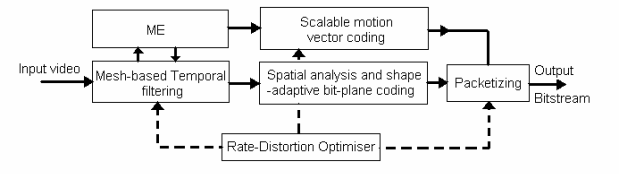


Figure 1 – General structure of the proposed 2D model-based video coding system

2.1. MC-based lifting scheme for temporal filtering

In the proposed scheme, lift-based 5/3-filtering structure is employed to achieve temporal filtering. With the temporal lifting structure, the frames go through the lifting stages step by step. The calculation process first upgrades the odd frames to the high pass wavelet coefficient frames, and upgrades the even frames to the low pass coefficient frames. For each pixel or patch in the odd frame, motion estimation (ME) is conducted to find the corresponding pixels and patches from its left or right even frame, or both frames.

For the pixels in the even frame, motion compensation (MC) is conducted to find their corresponding matched pixels in their neighbouring odd frame or frames. Some criteria can be used to decide whether the pixels are predicted from one frame (forward or backward prediction) or from two frames (bi-directional prediction). For the pixels which have corresponding matched (one or two) pixels in its neighbouring frame, lifting step is employed to get the high-pass part. If no pixel is matching this pixel, intra-prediction is conducted. Following the lifting step, update step is conducted for the even frames. The pixels which are originally terminated in many-to-one mapping can continue the temporal filtering without being stopped, as shown in Figure 2. Within the elementary lifting operation, the original terminated pixel in $Frame_1$ can be upgraded using both its left and right matching pixels instead of being stopped at the right side. When the anchor pixel in $Frame_2$ is to be lifted, though many pixels in $Frame_1$ are pointing to it, it is only calculated with the first scanned one according to the motion scan order. For a non-referred pixel in an even frame, it is still linked on both sides using the motion vectors of the adjacent motion threads.

During the lifting process, the counterpart motion vectors are strictly kept with inverse direction. The quarter-pel operation resembles the similar method. Based on the elementary lifting operation, the reference frames in each lifting stage can be reproduced in the decoder side, thus the perfect reconstruction of the wavelet synthesis is guaranteed.

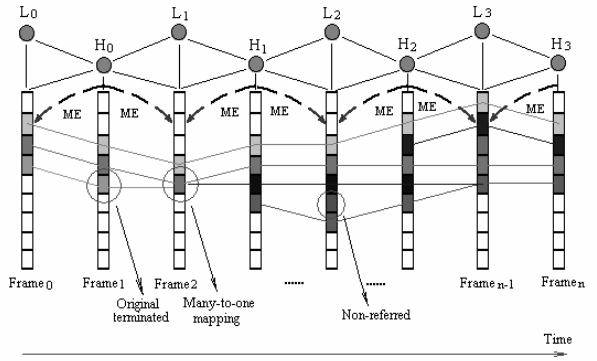


Figure 2 – Lifting-based temporal filtering based on bi-directional motion search

Many algorithms employ above lifting scheme [4] [7] [8]. The difference of our proposed scheme from these schemes are that warping motion compensation is applied instead of block-based motion compensation, and the temporal filtering of our scheme is conducted in object domain, instead of frame-domain.

2.2 Warping MC using scalable object mesh model

For the foreground video objects, the predefined 2-D scalable model, which has been discussed in detail in [9], is employed during the warping motion compensation. Before motion compensation, scalable model tracking is conducted to get the context-adaptive model, as shown in Figure 3, because only one frame is used to design the content-adaptive scalable model.

As there is an update process during the scalable model prediction, some pixels of the object have just one correspondence from either left or right frame. The object model after update process is not optimal. Therefore, a refinement process is needed. In order to refine the motion vectors of object model and select an optimal prediction method, rate-distortion optimised scheme is used during motion estimation. A Lagrangian multiplier λ^{estim} is used to choose the best control point motion vector considering the prediction error and the local motion vector variance between the candidature vector MV_j and the its K surrounding motion vectors MV_k of its neighbouring control points.

$$MV_i^{opt} = \arg\min_{MV_j}\left(MSE(MV_j) + \lambda^{estim} \cdot \sum_{k=1}^{K} (MV_k - MV_j)^2 \right) \quad (1)$$

After above refinement, the rate-distortion optimised motion vectors are achieved and they will be compressed by using the scalable coding method in Section 2.3.

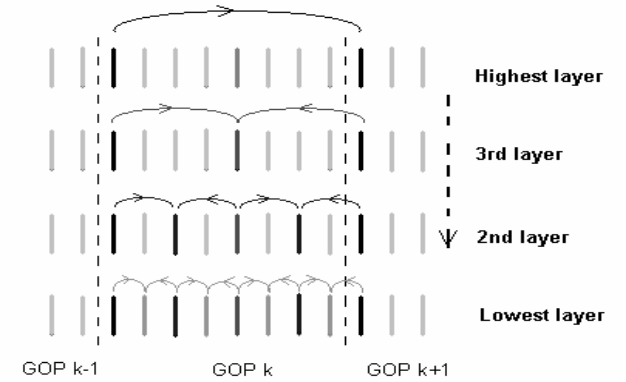


Figure 3 – Scalable model tracking along the video frames for video object

For the background, it is not efficient, in both computational complexity and compression, to building and employing its context-adaptive scalable model during video coding. In the proposed scheme, scalable quadrangular mesh model is applied for the background objects. As we know, quadrangular mesh model can not represent the object movements very precisely, especially along the objects boundaries with different motion mode. Even though for objective-based video coding, motion discontinuity is reduced. In order to improve its performance, overlapped block motion compensation (OBMC) and adaptive quadtree grid with variable density according to the varying motion activity are incorporated into the proposed scheme.

2.3 Scalable motion vector coding

The scalable motion vector coding process consists of two steps: motion vector prediction and differential MV coding. For the motion vector prediction, different prediction method is used for different model layers. For layer 0, the MV of vertex is predicted from its preceding vertex based on the pre-decided connectivity. As the connectivity of vertices has been compressed, the motion vector of its preceding vertex can be determined correctly. For other layers, the MV of vertex is predicted from its two neighbouring vertices of current layer and the coded layers, according to the connectivity information.

After getting the differential MV components, they are encoded using variable length codec (VLC). In the proposed scheme, context-adaptive binary arithmetic codec (CABAC), which is similar to that used for H.264 [2], is employed to compress the differential MVs. This method consists of two steps: in the binarization step, each motion vector symbol is represented by a unique binary pattern. A binary arithmetic coding engine follows this and allows encoding different bins with different models. Context-adaptive models are employed to compress different bins. Please refer to [10] for the detailed list of the designed context model.

The sign of the current prediction residual is encoded using a separate context model. The motion information to be transmitted of each layer consists of the prediction model (Bidirectional, uni-left and uni-right frame model) and motion vectors for each triangle for foreground and each rectangular for background.

2.4 Rate-distortion optimised bit allocation

Rate-distortion optimised bit allocation scheme is employed to truncate the encoded bitstream. During the motion compensation, the coding rate of motion vectors R_{mv} and the corresponding compensation error D_{mv} of motion vectors are recoded. At the same time, for the spatial bit-plane coding, each bitplane is scanned twice, resulting in two passes: sorting pass and refinement pass. During the bitplane coding, the coding rate $R_{spatial}$ and distortion $D_{spatial}$ of each pass is recorded at the end of each bit plane. All of the recorded rate and distortion data of the video object is used in the bitstream assembler to achieve optimal bitstream truncation. The optimisation algorithm is the same as that used in JPEG2000 [11]. As both the motion and spatial coding processes are included during the optimisation, more optimal bit allocation can be achieved.

Since each object and each frame in the GOP are encoded independently, the bitstream of each object is separable. The decoder can easily extract only a few video objects or frames and decode these objects and frames. So the implementation of temporal, quality and object scalabilities is natural and easy. If we are more interested in some special video objects, it can be achieved by assigning different Lagrange multipliers to different video objects according to their importance during R-D optimization for multiple video objects.

As the bit number and slope of each truncation point are included in the header of the bitstream, the final bitstream can be rearranged to meet other requirements. This property makes the final bitstream very flexible to be reused for all sorts of applications without re-encoding again.

3. EXPERIMENTAL RESULTS

Intensive experiments have been conducted to evaluate the performance of the proposed scalable 2D model-based video coding scheme, and compare it with MPEG-4 and H 264 standard. The test video clips Coastguard, News, and Motr_dhtr in QCIF resolution (30fps) were used in our experiments. Figure 4 illustrates the PSNR performance of Y-component for different encoding bit rates. Readers are reminded that the experimental results presented here for each sequence are decoded from an embedded

bitstream for the proposed encoding method and from different bitstreams corresponding to individual target coding rates for MPEG-4 and H. 264. From the results, it is found that the proposed method has 1 - 4 dB superior performance to MPEG-4 coder for a wide range of bit rate. However, it is inferior to H. 264 for medium and high bit rates. For some video sequence, our proposed method can achieve better compression performance than H. 264 for very low-bit rate coding due to scalable MV coding. When the target bit rate is very low, which is not enough to encode the full MV information, only the first layers of MV are encoded and more bits are saved for encoding the first frame of GOP.

4. CONCLUSION

This paper mainly discusses our proposed scalable 2D model-based video coding scheme after reviewing state-of-art scalable video coding techniques. First, video sequences are represented by semantic foreground objects and background objects. Then scalable content-adaptive model is constructed for each foreground object. After motion compensated temporal filtering, wavelet-based bit-plane coding algorithm is used to generate scalable bitstream. The motion vectors of scalable models are encoded through CABAC coder. Rate-optimised bit truncation algorithm is used to decide the optimal truncation point given the target bit rate. Experimental results show that the proposed scheme can achieve 1-4 dB gain when compared with MPEG-4 coder for a wide range of bit rates. It can also be competitive to H.264 at very low bit rates even though it is inferior to H264 at high bit rates. Future improvement can be made to the performance of this scalable 2D model-based video coding system in order to make it comparable to that of H.264 at all target rates.

ACKNOWLEDGEMENT: The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

5. REFERENCES

[1] ISO/IEC 14496-2: 2001, "Information Technology – Coding of audio-visual objects- Part 2: Visual", 2001

[2] ISO/IEC JTC1/SC29/WG11, ISO/IEC FDIS 14496-10 (AVC): Information Technology-Coding of Audio-Visual Objects-Part 10: Advanced Video Coding, 2003

[3] ISO/IEC TC1 / SC29 / WG11 MPEG2003 / N6193, "*Call for Proposals on Scalable Video Coding Technology*", Waikoloa, December 2003

[4] Xu J., Xiong R., Feng B., Sullivan G., Lee M.-C., Wu F., Li S., "3D sub-band video coding using barbell lifting", ISO/IEC JTC1/SC29/WG11 M10569/S05, Munich MPEG meeting, Germany, March 2004.

[5] Wien M., Rusert T., Hanke K., "RWTH proposal for scalable video coding technology", ISO/IEC JTC1/SC29/WG11 M10569/S16, Munich MPEG meeting, Germany, March 2004

[6] Schwarz H., Marpe D., and Wiegand T., "Scalable Extension of H.264/AVC", ISO/IEC JTC1/SC29/WG11, Doc. M10569/S03, Munich, Germany, Mar. 2004

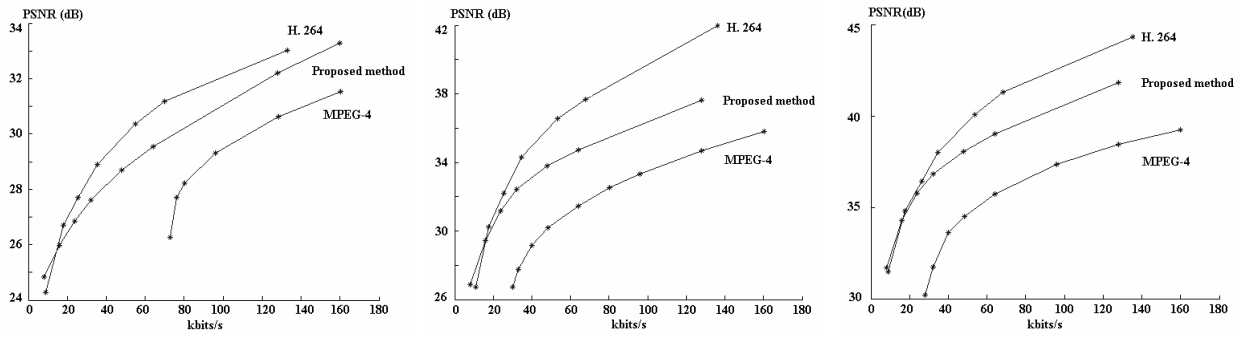
[7] Secker A. and Taubman D., "Motion-compensated highly scalable video compression using an adaptive 3-D wavelet transform based on lifting", in proceeding of ICIP, 2001, Thessalonica, GR, Vol.2, Oct. 2001, pp. 1029-1032

[8] Luo L., Wu F., Li S., and Zhang Z., "Advanced lifting-based motion-threading (MTh) techniques for 3D wavelet video coding", Proc. of SPIE Vol. 5150, VCIP2003, Lugano, Switzerland, July 2003, pp. 707-718

[9] M. Hu, S. Worrall, A.H. Sadka and A.M. Kondoz, "Model design for scalable two-dimensional model-based video coding", IEE Electronics Letters, Vol.38, No.24, Nov. 2002, pp.1513-1515

[10] M. Hu, "Highly scalable 2-D model-based video coding", PhD Thesis, University of Surrey, Dec. 2004

[11] ISO / IEC JTC1/SC29/WG1.FCD 15444-1: Infor-mation technology – JPEG2000 image coding system, March 2000



(a) (b) (c)

Figure 4 – PSNR performance comparison for Y-component of (a) Coastguard; (b) News; and (c) Motr_dhtr sequence