Scalable motion vector coding

J. Barbarien, A. Munteanu, F. Verdicchio, Y. Andreopoulos, J. Cornelis and P. Schelkens

Scalable wavelet-based video codecs using spatial-domain motion compensated temporal filtering require a quality-scalable motion vector codec to support a large range of bit rates with optimal compression efficiency. Introduced is a new prediction-based architecture for quality-scalable motion vector coding that outperforms the state-of-the-art wavelet-based techniques previously proposed in the literature.

Introduction: Scalable spatial-domain motion compensated temporal filtering (SDMCTF)-based video codecs typically use non-scalable motion vector codecs (MVCs). Targeting low bit rates with such schemes requires generating less complex (but less accurate) motion fields, significantly affecting the overall codec performance at all rates. Using a quality-scalable MVC solves this problem. The minimum attainable bit rate is then no longer bound by the rate needed to losslessly code the motion information. Additionally, a scalable MVC allows for optimally dividing the available bit rate between motion and texture data. In comparison to a video codec using a non-scalable MVC, this yields a systematically better rate-distortion performance, especially at low bit rates [1].

Quality scalable motion vector coding techniques that perform an integer wavelet transform of the motion vector components followed by embedded coding of the resulting wavelet coefficients were recently proposed in [2] and later in [1]. Although providing scalability, the lossless compression performance of this type of scheme is significantly lower than that of non-scalable MVCs based on motion vector prediction [2]. In this Letter, a new architecture for scalable motion vector coding is introduced, combining the high compression efficiency obtained by using motion vector prediction with support for quality scalability.

Quality-scalable prediction-based motion vector coding: The proposed motion-vector coding architecture (see Fig. 1) is designed to compress motion information produced by multihypothesis block-based motion estimation (ME) using multiple block sizes and multiple reference frames [3, 4]. The algorithm generates a bit-stream consisting of a base-layer, which is always decoded losslessly, and a quality-scalable enhancement layer. The motion vectors are first quantised by discarding the information on the lowest bit-plane(s). Thereafter, the quantised motion vectors are compressed using a prediction-based coding technique and the macro-block splitting-information, hypothesis information and reference-frame indices (RFIs) are compressed using context-based adaptive arithmetic coding. The resulting encoded data forms the base-layer of the final bit-stream. The quantisation errors are coded using an embedded bit-plane coding algorithm. The resulting data forms the quality-scalable enhancement layer of the final bit-stream. In this way, the total motion vector bit rate can be chosen to lie anywhere between the rate needed to losslessly code the base-layer (which can be controlled by changing the quantisation step size) and the rate needed to losslessly reconstruct the motion information.

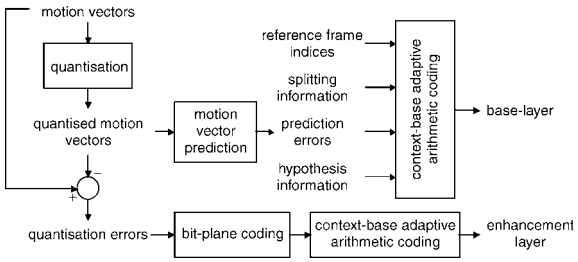

Fig. 1 *General architecture of proposed MVC*

Prediction-based coding of quantised motion vectors: The quantised motion vectors are encoded by performing motion vector prediction followed by lossless coding of the prediction errors. In both steps, the macro-blocks are visited in raster order. In a split macro-block, the sub-blocks are visited in depth-first quadtree scanning order. If multiple motion vectors are associated to the currently visited macro-block/sub-block B_c [4], they are treated sequentially before proceeding to the next block. The size of B_c is $S_c \times S_c$ pixels, and the top-left pixel of B_c has co-ordinates (x_c, y_c). Each quantised motion vector $\vec{\mathbf{v}}_c$ belonging to block B_c is predicted by taking the median of a set of quantised motion vectors U_p. The set U_p consists of vectors that (a) were predicted earlier (causality), and (b) have an RFI pointing to a frame lying at the same distance (in number of frames) from the predicted frame as the frame pointed to by the RFI of $\vec{\mathbf{v}}_c$. To construct U_p, a pixel-based motion field $\mathcal{P}$ is generated, by associating with each pixel in the predicted frame the motion information of the smallest block the pixel belongs to. All the quantised motion vectors in $\mathcal{P}$ at positions (x_c+j, y_c-1), $0 \le j < S_c$, (x_c-1, y_c+k), $0 \le k < S_c$, $(x_c+S_c, \quad y_c-1-l)$, $\quad 0 \le l < S_c/2$ and $\quad (x_c+S_c+m, \quad y_c-1)$, $0 \le m < S_c/2$ that satisfy conditions (a) and (b) form one sub-set of U_p. These are vectors belonging to the blocks in the spatial neighbourhood of B_c (e.g. Fig. 2). The second sub-set of U_p contains all the other quantised vectors associated with B_c that satisfy (a) and (b). These vectors are inserted S_c times into U_p to ensure that all the vectors in U_p have an impact on the median proportional to the size of the block they are associated with. If the RFI of one of the vectors involved in the prediction points to a future frame in the sequence, the signs of its components are inverted prior to insertion into U_p.

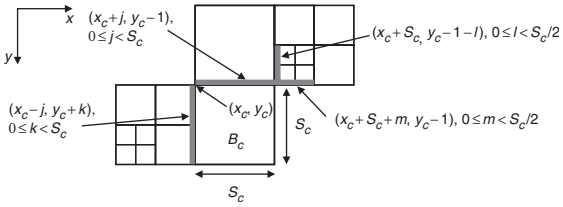

Fig. 2 *Motion-vector prediction*

Context-based adaptive arithmetic coding is used to encode the prediction errors. The horizontal and vertical components of the prediction-error vectors are coded separately. The interval of possible component values is split into a number of sub-intervals, as follows [2, 5]:

$$S_i = \begin{cases} \{0\} & \text{if } i = 0 \\ [-(2^i - 1), -(2^{i-1})] \cup [2^{i-1}, 2^i - 1] & \text{if } i > 0 \end{cases} \quad (1)$$

For each component value, a symbol representing the sub-interval S_i it belongs to is coded first. Thereafter, the offset within the sub-interval is coded [2, 5].

Embedded coding of quantisation errors: The horizontal and vertical components of the quantisation-error vectors are coded in a bit-plane by bit-plane fashion. Each bit-plane is coded in two passes, i.e. a significance pass and a refinement pass. With each bit-plane i, a threshold $T_i = 2^i$ is associated. A quantisation-error component c_e is said to be significant for a threshold T_i if $|c_e| \ge T_i$. In the significance pass, the significance of all previously non-significant components is encoded. The quantisation-error vectors are visited in the same order as the quantised motion-vectors in their coding process. When a component becomes significant for the first time and its corresponding quantised motion-vector component is 0, its sign is also coded. In the refinement pass, already significant components are refined by coding the binary value corresponding to the current bit-plane. The significance, refinement and sign information is compressed using context-based adaptive arithmetic coding.

Experimental results: First, the lossless compression performance of the proposed MVC is compared to that of the wavelet-based quality-scalable motion vector coding technique of [2]. The wavelet-based algorithms proposed in [1, 2] separately code the motion vector components by performing a 5/3 integer wavelet transform followed by quality scalable coding of the resulting coefficients. In [1] the encoding is performed using the JPEG2000 codec, while the QT-L codec of [6] is used in [2]. As shown in [6], the performance of QT-L is on par with that of JPEG2000, hence both scalable MVCs are expected to yield similar performance.

The motion information used in this comparison is generated by a multi-hypothesis motion estimation algorithm embedded in an SDMCTF codec [7], using two reference frames, quarter-pel accuracy,

but no intramode and no macro-block splitting. This type of motion data is chosen to avoid difficulties in adapting the wavelet-based MVCs [1, 2] to more complex motion information. Three CIF-resolution sequences, 'Football', 'Canoa' and 'Container', respectively 260, 220 and 300 frames long, were used in the experiment. All sequences have a frame-rate of 30 Hz. Table 1 shows the average number of bytes spent on the motion vectors of a predicted frame for the wavelet-based MVC and for the proposed MVC using different quantisation steps ($Q = k$ means the lowest k bit-planes are discarded in the quantisation).

Table 1: Average number of bytes spent per predicted frame for proposed MVC using different quantisation step sizes and for wavelet-based quality-scalable MVC [2]

	Proposed MVC					
Sequence	$Q=0$	$Q=1$	$Q=2$	$Q=3$	$Q=4$	Wavelet-based MVC
Football	652	662	673	682	689	722
Canoa	642	650	656	664	677	709
Container	161	162	158	154	150	211

Table 2: Average PSNR obtained per frame for video codec equipped with scalable and non-scalable motion vector coding

Target rates (texture + motion) (kbit/s)		128	192	256	512	1024
Football	Non-scalable MVC	–	22.99	25.99	28.24	30.84
	Scalable MVC	21.76	24.98	26.04	28.23	30.83
Canoa	Non-scalable MVC	–	–	23.16	26.10	28.58
	Scalable MVC	20.16	22.30	23.55	26.05	28.55

In the second experiment, the compression performance of the SDMCTF video codec is assessed when using scalable and non-scalable motion vector coding (the latter corresponding to the proposed MVC with no quantisation and no enhancement layer). The codec employs multi-hypothesis motion estimation using two reference frames, two block sizes and quarterpel accuracy [3]. The distribution of the rate between texture data and motion data is performed using a heuristic technique, but a rate-distortion optimal approach can be used, similar to [1]. For each target bit rate, the motion vectors are decoded at five different rates, evenly spread out between the base-layer rate and the rate needed to losslessly reconstruct the motion vectors. The base-layer rate is kept below 96 kbit/s by appropriately selecting the quantisation step size for each predicted frame. The combination of motion and texture rates yielding the best quality (PSNR) is retained. The results of the experiment are shown in Table 2 for the 'Football' and 'Canoa' sequences. We report the average PSNR (in dB) per frame when decoding to different target bit rates.

Conclusions: The proposed MVC yields state-of-the-art results, outperforming wavelet-based quality-scalable motion vector coding [1, 2]. The benefits of integrating the designed quality-scalable MVC into an SDMCTF-based video codec are experimentally demonstrated. Lower bit rates can be attained without sacrificing motion estimation efficiency and the overall coding performance at low rates is improved by better distribution of the available rate between texture and motion information.

Acknowledgments: This work was supported by the Flemish Institute for the Promotion of Innovation by Science and Technology (PhD bursary Joeri Barbarien). P. Schelkens has a post-doctoral fellowship with the Fund for Scientific Research—Flanders (FWO), Egmontstraat 5, B-1000 Brussels, Belgium.

30 March 2004
Electronics Letters online no: 20040490
doi: 10.1049/el:20040490

J. Barbarien, A. Munteanu, F. Verdicchio, Y. Andreopoulos, J. Cornelis and P. Schelkens (*Department of Electronics and Information Processing, Vrije Universiteit Brussel, Pleinlaan 2, 1050, Brussels, Belgium*)

References

1 Taubman, D., and Secker, A.: 'Highly scalable video compression with scalable motion coding'. IEEE Int. Conf. Image Processing (ICIP 2003), Barcelona, Spain, September 2003, Vol. 3, pp. 273–276
2 Barbarien, J., *et al.*: 'Coding of motion vectors produced by wavelet-domain motion estimation'. ISO/IEC JTC1/SC29/WG11, M9249, Awaji Island, Japan, December 2002
3 Andreopoulos, Y., *et al.*: 'Response to the call for evidence on scalable video coding advances'. ISO/IEC JTC1/SC29/WG11, M9911, Trondheim, Norway, July 2003
4 Flierl, M., Wiegand, T., and Girod, B.: 'Rate-constrained multihypothesis prediction for motion-compensated video compression', *IEEE Trans. Circuits Syst. Video Technol.*, 2002, **12**, (11), pp. 957–969
5 Barbarien, J., *et al.*: 'Motion vector coding for in-band motion compensated temporal filtering'. IEEE Int. Conf. Image Processing (ICIP 2003), Barcelona, Spain, September 2003, Vol. 2, pp. 783–786
6 Schelkens, P., *et al.*: 'Wavelet coding of volumetric medical datasets', *IEEE Trans Med. Imaging*, 2003, **22**, (3), pp. 441–458
7 Andreopoulos, Y., *et al.*: 'Open-loop, in-band, motion-compensated temporal filtering for objective full-scalability in wavelet video coding'. ISO/IEC JTC1/SC29/WG11, M9026, Shanghai, China, October 2002