Spatio-Temporal Scalability for MPEG Video Coding

Marek Domański, *Member, IEEE*, Adam Łuczak, and Sławomir Maćkowiak

***Abstract*—The existing and standardized solutions for spatial scalability are not satisfactory, therefore new approaches are very actively explored recently. The goal of this paper is to improve spatial scalability of MPEG-2 for progressive video. In order to avoid problems with too large bitstreams of the base layer produced by some of the hitherto proposed spatially scalable coders, spatio-temporal scalability is proposed for video compression systems. It is assumed that a coder produces two bitstreams, where the base-layer bitstream corresponds to pictures with reduced both spatial and temporal resolution while the enhancement layer bitstream is used to transmit the information needed to retrieve images with full spatial and temporal resolution. In the base layer, temporal resolution reduction is obtained by B-frame data partitioning, i.e., by placing each second frame (B-frame) in the enhancement layer. Subband (wavelet) analysis is used to provide spatial decomposition of the signal. Full compatibility with the MPEG-2 standard is ensured in the base layer. As compared to single-layer MPEG-2 encoding at bit rates below 6 Mbits/s, the bitrate overhead for scalability is less than 15% in most cases.**

***Index Terms*—MPEG-2, spatial scalability, subband analysis, temporal scalability, video coding.**

I. Introduction

SPATIALLY scalable or hierarchical video coders produce two bitstreams: a base layer bitstream, which represents low-resolution pictures, and an enhancement layer bitstream, which provides additional data needed for reproduction of pictures with full resolution. An important feature is that the base-layer bitstream can be decoded independently from an enhancement layer. Therefore, low-resolution terminals are able to decode only the base-layer bitstream in order to display low-resolution pictures. Such compression techniques are of great interest recently, because of development of communication networks with different transmission bit rates [1]–[13]. Moreover, scalable transmission is advantageous in error-prone environments where base-layer packets are well protected against transmission errors and losses, while the protection of the enhancement layer packets is lower. In such a system, a receiver is able to reproduce at least low-resolution pictures if quality of service decreases.

The MPEG-2 video-compression standard [14], [15] has established four types of scalability: spatial, temporal, SNR, and data partitioning. Among them, spatial scalability is of particular interest because of its prospective broad applications. Unfortunately, application of the MPEG-2 spatial scalability is mostly related to nonacceptably high bitrate overheads as compared to single-layer MPEG-2 encoding of video. This additional overhead for MPEG-2 spatial scalability is about 60%–70% of total bit rate [16]. By many test sequences, the total bitstream is not much smaller than sum of bitstreams obtained for simulcast transmission with two different resolutions.

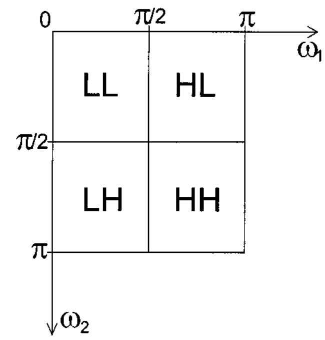


Fig. 1. Subband decomposition.

The goal of this paper is to propose alternative techniques which would provide spatial scalability with lower bitrate overheads. The assumption is that a high level of compatibility with the MPEG-2 video-coding standard would be ensured. In particular, it is assumed that the low-resolution base layer bitstream is fully compatible with the MPEG-2 standard. Moreover, a scalable codec should consists mostly of functional blocks present in a standard MPEG-2 codec.

In order to meet practical requirements, it is also assumed that the bitstream of the base layer does not exceed the bitstream of the enhancement layer.

Scalable compression of progressive video is considered in this paper because prospective applications of scalable encoders are related to emerging multimedia services where progressive format of video is gaining popularity related to its compatibility with computer display technology. Nevertheless, the approach proposed in this paper can be extended on the interlaced formats of video.

II. Spatio-Temporal Scalability

There were many attempts to improve spatially scalable coding of video. The proposed schemes were based on pyramid decomposition [5] or subband/wavelet decomposition [6]–[13]. Among various proposals, the latter approach should be considered very promising. The idea is to split each image into four spatial subbands. The subband of lowest frequencies constitutes a base layer, while the other three subbands are jointly transmitted in an enhancement layer (Fig. 1). Nevertheless, this approach often leads to allocation of much higher bit rates to a base layer than to an enhancement layer, which is disadvantageous for practical applications. Recently, Benzler

Manuscript received July 1999; revised April 2000. This work was supported by the Polish Committee for Scientific Research. This paper was recommended by Guest Editor Y.-Q. Zhang.

The authors are with Poznań University of Technology, Institute of Electronics and Telecommunication, 60- 965 Poznań, Poland.

Publisher Item Identifier S 1051-8215(00)08203-3.

[16] has proposed to avoid this problem by combining spatial and SNR scalability and neglecting the requirement of the full MPEG-compatibility in the base layer. Here, our goal is to use a fully MPEG-compatible coder in the base layer. For the required codecs, spatio-temporal scalability is proposed [17], [18]. Here, a base layer corresponds to the bitstream of the pictures with reduced both spatial and temporal resolutions. Therefore, in the base layer, the bit rate is decreased as compared to a encoder with spatial scalability only. Now, it is easy to get the base layer bit rate equal or even less than that of the enhancement layer. The enhancement layer is used to transmit the information needed for restoration of the full spatial and temporal resolution.

Embedding of subband decomposition into a motion-compensated encoder leads to in- or out-band motion compensation performed on individual subbands or on the whole image, respectively. The latter will be used here, because some experimental results show that it is more efficient [7], [8], [10].

Here, the term of spatio-temporal scalability is proposed for a functionality of video compression systems where the base layer corresponds to pictures with reduced both spatial and temporal resolution. An enhancement layer is used to transmit the information needed for restoration of the full spatial and temporal resolution.

The authors have already considered two basic approaches related to spatio-temporal scalability [17], [18]. The first approach exploits 3-D subband analysis while the second approach is based on B-frame data partitioning.

A. First Approach

The input video sequence is analyzed in a 3-D separable filter bank, i.e., there are three consecutive steps of analysis: temporal, horizontal, and vertical. For temporal analysis, very simple linear-phase two-tap filters are used similarly as in other papers on three-dimensional subband coding [19], [20]

$$H(z) = 0.5 \cdot (1 \pm z^{-1})$$

where "$+$" and "$-$" correspond to low- and high pass filters, respectively. This filter bank has a very simple implementation, needs to store one frame only and exhibits small group delay.

Temporal analysis results in two subbands L_t and H_t of low and high temporal frequencies, respectively. In both subbands, the temporal sampling frequency is reduced by factor two. Therefore, these two subbands correspond to two video sequences with reduced frame frequency. The two subbands are partitioned into four spatial subbands (LL, LH, HL, and HH) each. For spatial analysis, both horizontal and vertical, separable FIR filters are used. The 3-D analysis results in eight spatio-temporal subbands (Fig. 2). Three high-spatial-frequency subbands (LH, HL and HH) in the high-temporal-frequency subband H_t are discarded, as they correspond to the information being less relevant for the human visual system. According to the experimental authors' tests for 720×576 progressive 50-Hz test sequences, it reduces PSNR often to about 32–33 dB, and has negligible influence on subjective quality of the decoded video. Thus, five subbands are encoded:

- in a base layer—the spatial subband LL of the temporal subband L_t;

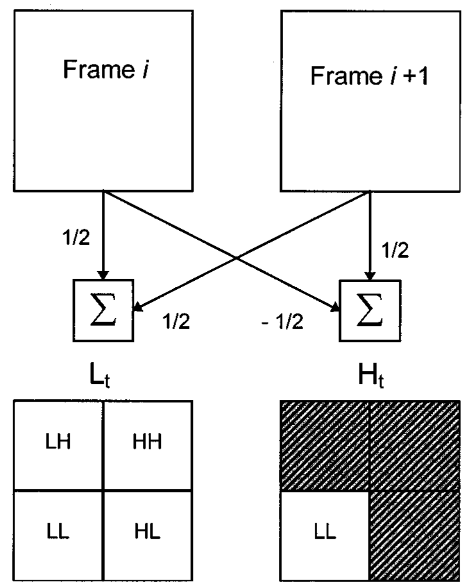


Fig. 2. 3-D subband analysis.

- the enhancement layer includes the spatial subbands LH, HL, and HH from the temporal subband L_t and the spatial subband LL of the temporal subband H_t.

The encoder structure is summarized in Table I.

B. Second Approach

In the second variant, the technique employs data structures already designed for standard MPEG-2 coding. Reduction of temporal resolution is obtained by removal of each second frame. It is assumed that groups of pictures (GOPs) consist of even number of frames. Moreover, it is assumed that each second frame is a B-frame, i.e., it can be removed from a sequence without affecting decodability of the remaining frames.

Reduction of spatial resolution is obtained by use of subband decomposition. Proper design of the filter bank results in negligible spatial aliasing in the LL subband, which constitutes the base layer. Unfortunately the technique does not provide any means to suppress temporal aliasing. The effects of temporal aliasing are similar as those related to frame skipping in hybrid encoders.

A standard order of frames in the base and enhancement layers is as shown in Table II.

The base-layer data are used to produce low-quality images; therefore, it is reasonable to perform more rough quantization here than in the enhancement layer. On the other hand, quality of the subband LL is strongly related to the quality of the full-sized picture. The low quality of the LL subband restricts the full-sized picture quality to a relatively low level, despite of the amount of information in the remaining subbands. Therefore, it is important to transmit additional information ΔLL in the enhancement layer. This information is used to improve quality of the subband LL when used to synthesize full-sized images in the enhancement layer.

III. Scalable Encoder

The second variant based on B-frame data partitioning is described in more detail below. The fundamental assumption

TABLE I
STRUCTURE OF THE SYSTEM BASED ON 3-D ANALYSIS

Layer	Temporal subband	Spatial subband	Encoding scheme
Base	L_t	LL	Encoded using a standard MPEG-2 encoder for reduced spatial resolution and frame frequency
Enhancement		LH	Encoded using motion vectors calculated for full-size frames
		HL	
		HH	
	H_t	LL	
Discarded		LH	Not encoded
		HL	
		HH	

TABLE II
GOP STRUCTURE IN BOTH LAYERS

Base layer (only subband LL)	Enhancement layer
I	I (without LL subband)
skipped	B
B	B (without LL subband)
skipped	B
P	P (without LL subband)
skipped	B
B	B (without LL subband)
skipped	B
P	P (without LL subband)
skipped	B
B	B (without LL subband)
skipped	B
P	P (without LL subband)
skipped	B
B	B (without LL subband)
skipped	B

which restricts the structure of such a encoder is that a least the base layer encoder has to be MPEG-2-compatible.

The structure of the encoder is shown in Fig. 3.

The base-layer encoder is implemented as a standard motion-compensated hybrid MPEG-2 encoder. This encoder supplies the enhancement layer encoder with three data streams:

1) DCT coefficients from LL subband;
2) quantized DCT coefficients from LL subband;
3) motion vectors.

In the enhancement-layer encoder, motion is estimated for full-resolution images, and full-frame motion compensation is performed. Therefore, all subbands have to be synthesized into full-resolution frames before motion-compensated prediction. After motion compensation, spatial subbands are produced again. The prediction errors are calculated and encoded for three subbands (HL, LH, and HH). Therefore, there are two subband analysis stages and one subband synthesis stage in the encoder.

In the enhancement-layer encoder, the subband LL used for frame synthesis is more finely quantized than this transmitted in the base layer. It corresponds to a sum of information contained in the base layer and in the bitstream ΔLL transmitted in the enhancement layer.

The bitstream ΔLL contains additional least significant bits which are used for correction of the transform coefficients transmitted in the base layer.

Motion vectors MV_b are estimated for the base layer. Other motion vectors MV_e are estimated for the enhancement layer, i.e., these four MV_e vectors that correspond to a MV_b vector. In the enhancement layer, difference values $(MV_e - MV_b)$ are transmitted.

The motion-compensated predictor employed in the enhancement layer uses two kinds of reference macroblocks:

1) motion-compensated macroblocks from neighboring frames;
2) interpolated blocks obtained by decoding of the base layer data (when applicable).

Actually, the second type of reference macroblocks is the same as already standardized in the MPEG-2 standard for the enhancement layer by spatial scalability. Therefore, for the encoder proposed, prediction in the enhancement layer switches between spatial interpolation (like in the standard MPEG-2 scalable encoder) and temporal interpolation as show in Fig. 4.

A small extension to the MPEG-2 compression technique is that those B-frames which correspond to the B-frames from the base layer, can be used as reference frames for other B-frames in the enhancement layer.

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The purpose of the experiments was to examine the properties of the codec proposed. Therefore, easy-to-modify software was written in the C++ language. The most important feature is its flexibility ,allowing tests of different variants of coding algorithms. Currently, the program includes about 14 000 lines of code. It includes software implementation of an MPEG-2 MP@ML encoder for the base layer. The software runs on Sun 20 workstations under the Solaris operational system.

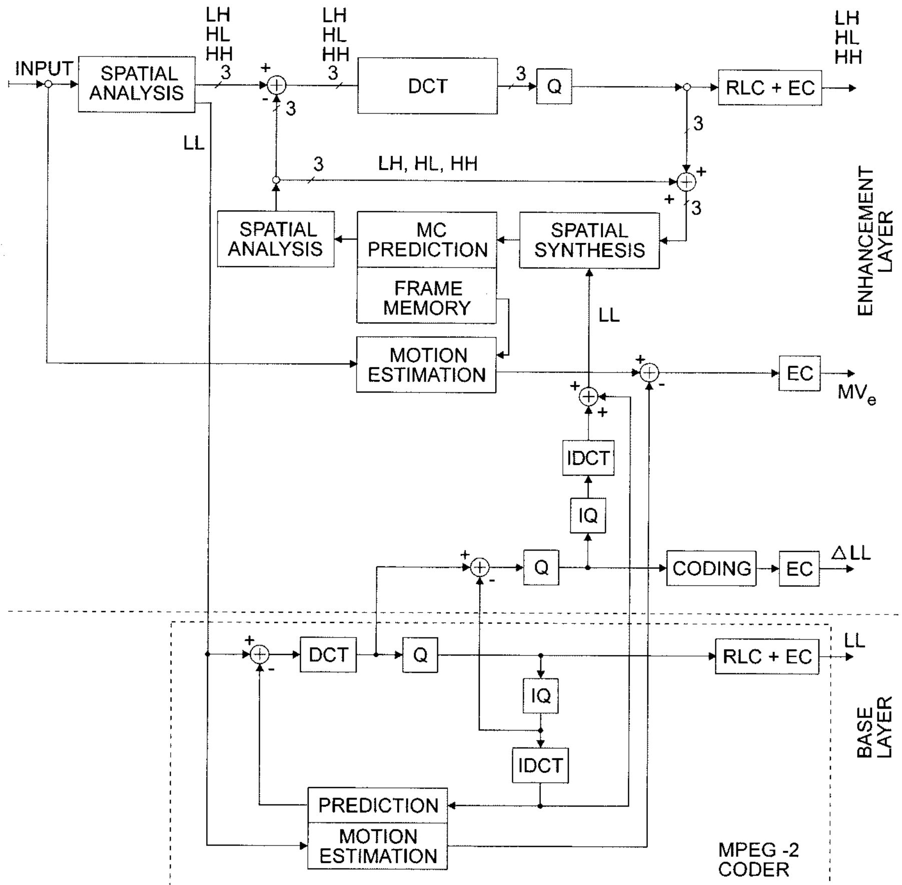


Fig. 3. Basic block diagram of a coder with the functionality of spatio-temporal scalability.

TABLE III
EXPERIMENTAL RESULTS.

	Test sequence						
	Flower Garden		Mobile Calendar	Basketball		Funfair	
Single layer (MPEG-2) bitstream [Mbps]	4.90	7.57	7.89	4.67	6.99	4.67	6.15
Average luminance PSNR [dB]	29.6	32.6	30.8	29.9	32.1	31.3	32.9
Scalable bitstream [Mbps]	5.22	8.66	10.70	5.02	9.26	4.97	8.45
Average luminance PSNR [dB]	29.6	32.4	29.9	29.6	31.1	31.0	32.3
Base layer bitstream [Mbps]	2.13	3.01	3.35	2.07	3.12	2.02	3.17
Base layer bitstream [%]	40.8	34.8	31.3	40.3	33.7	40.6	37.5
Bitstream overhead [%]	6.5	14.4	35.6	7.5	32.5	6.4	37.4

The encoder is aimed at processing of progressive 720×576 50-Hz test sequences. Therefore, the base layer is in the SIF format. The experiments have been done with 4 : 2 : 0 sequences.

The analysis and synthesis filter banks were Johnston 12th-order FIR QMF's [23] and Daubechies (9,3) wavelets [24]. Neither has performed much better than the other.

The experimental results for two progressive 720×576 50-Hz test sequences ("*Flower garden*" and "*Funfair*") are given in Table III for the encoder with B-frame data partitioning.

The first goal, i.e., base-layer bitstream not exceeding that of enhancement layer, has been reached for all bit rates and video-test sequences tested where the base-layer bit rate was between 31%–41% of the total bit rate.

Moreover, coding efficiency measured by the bit rate overhead for scalable coding as compared to MPEG-2 nonscalable encoder was low. The experiments with bitrates mostly below 6 Mbits/s resulted in the overhead of about 6%–15%. Therefore, this new encoder strongly outperforms the spatially scal-

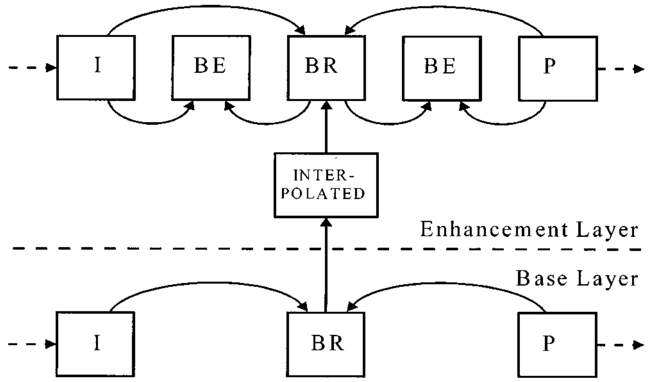


Fig. 4. Modified prediction of B-frames: arrows define possible directions of usage of macroblocks for prediction.

able MPEG-2 encoder, which implies an overhead for scalability of about 50%–70 % of the total single-layer bitstream. Similar results have been obtained for the encoder with 3-D subband analysis.

V. Conclusion

Very promising results have been obtained for the mixed spatio-temporal scalability. First of all, the bit-rate overhead of the spatio-temporal scalability is much lower than by standard spatial scalability. On the other hand, simultaneous reduction of both spatial and temporal resolutions seems to be reasonable in many applications. The bit rate of the base layer is lower than that in the enhancement layer. This is a substantial advantage over solutions based only on spatial subband decomposition.

It is worthy to mention that the complexity of the new encoder defined by the number of operations necessary to implement it is only about 30% higher than that for a single-layer MPEG-2 encoder.

Many further improvements are possible for the compression algorithm. Optimization of the encoding of high-frequency spatial subbands (LH, HL, HH), as well as better encoding of the signal ΔLL, could decrease enhancement-layer subbands significantly.

Acknowledgment

The authors wish to thank the anonymous reviewer for his valuable suggestions. The authors express their sincere thanks to Prof. H. G. Musmann and U. Benzler from the University of Hannover for their advice and fruitful discussions.

References

[1] T. Hanamura, W. Kameyama, and H. Tominaga, “Hierarchical coding scheme of video signal with scalability and compatibility,” *Signal Processing: Image Commun.*, vol. 5, pp. 159–184, February 1993.

[2] G. Morrison and I. Parke, “A spatially layered hierarchical approach to video coding,” *Signal Processing: Image Commun.*, vol. 5, pp. 445–462, December 1993.

[3] T. Chiang and D. Anastassiou, “Hierarchical HDTV/SDTV compatible coding using Kalman statistical filtering,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 424–437, Apr. 1999.

[4] A. Nosratina and M. Orchard, “Multi-resolution backward video coding,” *Proc. Int. Conf. Image Processing*, vol. 2, pp. 563–566, 1995.

[5] G. Lilienfield and J. Woods, “Scalable high definition video coding,” in *Proc. SPIE Visual Communication and Image Processing*, San Jose, CA, 1998, pp. 158–169.

[6] Y.-Q. Zhang and S. Zafar, “Motion-compensated wavelet transform coding for color video compression,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, pp. 285–296, Sept. 1992.

[7] K. Shen and E. Delp, “Wavelet based rate scalable video compression,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 109–122, Feb. 1999.

[8] T. Tsunashima, J. Stampleman, and V. Bove, “A scalable motion-compensated subband image coder,” *IEEE Trans. Commun.*, vol. 42, pp. 1894–1901, 1994.

[9] P.-Y. Cheng, J. Li, and C.-C. Kuo, “Multiscale video compression using wavelet transform and motion compensation,” *Proc. Int. Conf. Image Processing*, vol. 1, pp. 606–609, 1995.

[10] U. Benzler, “Scalable multiresolution video coding using sub-band decomposition,” in *Proc. 1st Int. Workshop Wireless Image/Video Communication*, Loughborough, U.K., 1996, pp. 109–114.

[11] P.-C. Chang and T.-T. Lu, “A scalable video compression technique based on wavelet and MPEG coding,” in *Proc. Int. Conf. Consumer Electronics*, 1999, pp. 372–373.

[12] F. Bosveld, “Hierarchical video compression using SBC,” Ph.D. dissertation, Delft Univ. of Technology, Delft, The Netherlands, 1996.

[13] H. Gharavi and W. Y. Ng, “H.263 Compatible Video Coding and Transmission,” in *Proc. 1st Int. Workshop Wireless Image/Video Communication*, Loughborough, U.K., 1996, pp. 115–120.

[14] *Information Technology—Generic Coding of Moving Pictures and Associated Audio Information*, ISO/IEC Int. Standard 13818.

[15] B. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2*. London, U.K.: Chapman & Hall, 1997.

[16] U. Benzler, “Scalable multi-resolution video coding using a combined subband-DCT approach,” in *Proc. Picture Coding Symp.*, Portland, OR, 1999, pp. 17–20.

[17] M. Domański, A. Łuczak, S. Maćkowiak, and R. Świerczyński, “Hybrid coding of video with spatio-temporal scalability using subband decomposition,” in *Proc. Signal Processing IX: Theories and Applications*, Rhodes, Greece, Sept. 1998, pp. 53–56.

[18] ——, “Hybrid coding of video with spatio-temporal scalability using subband decomposition,” in *Proc. SPIE* San Jose, CA, 1999, vol. 3653, pp. 1018–1025.

[19] R. Ohm, “Three-dimensional subband coding with motion-compensation,” *IEEE Trans. Image Processing*, vol. 3, pp. 559–571, Sept. 1994.

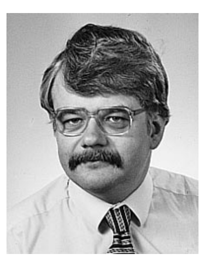
[20] Ch. Podilchuk, N. Jayant, and N. Farvardin, “Three-dimensional subband coding of video,” *IEEE Trans. Image Processing*, vol. 4, pp. 125–139, Feb. 1995.

[21] vA. Puri, L. Yan, and B. Haskell, “Temporal resolution scalable video coding,” *Proc. Int. Conf. Image Processing*, vol. 2, pp. 947–951, 1994.

[22] G. Conklin and S. Hemami, “A comparison of temporal scalability techniques,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 909–919, Sept. 1999.

[23] J. Johnston, “A filter family designed for use in quadrature mirror filter banks,” in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, 1980, pp. 291–294.

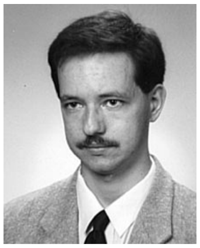
[24] I. Daubechies, “Ten lectures on wavelets,” *SIAM*, 1992.



Marek Domański (M’90) was born in 1954. He received the M.S., Sc.D., and Habilitation degrees from Poznań University of Technology, Poznań, Poland, in 1978, 1983, and 1990, respectively.

In 1977, he joined Poznań University of Technology as a member of staff. He was with Ruhr University, Bochum, Germany, during 1986–87 and 1990–1991 as both a DAAD and Alexander von Humboldt Fellow, respectively. Currently, he is a Professor at Poznań University of Technology, and Head of the Multimedia Communication and Radioelectronics Division. He is an author or co-author of more than 100 papers and conference contributions, mostly on image processing, image and video coding, color image processing, digital filters, and multidimensional digital systems. He is author of the book *Advanced Techniques for Image and Video Compression* (Poznań, Poland: Poznań University of Technology, 1998), and co-author of a chapter in the book *Colour Image Processing* (London, U.K.: Chapman & Hall, 1988). He headed several research and development projects on image and video compression, image and video enhancement and restoration, multidimensional digital filters, and telemedicine.

Dr. Domański is a member of EURASIP.



Adam Łuczak was born in Poznań, Poland, in 1972. He received the M.S. degree in electronics and telecommunications from Poznań University of Technology, Poznań, Poland, in 1997.

He joined Poznań University of Technology as an Assistant in 1997. His research interests include control schemes of hybrid video encoders.

Sławomir Maćkowiak was born in Poznań, Poland, in 1972. He received the M.S. degree in electronics engineering from Poznań University of Technology, Poznań, Poland, in 1997.

He joined the Multimedia Communications and Radioelectronics Division, Institute of Electronics and Communications, Poznań University of Technology, in 1997, where he is currently a Research Assistant. His main research interests include scalable video compression and multimedia systems.