Advances in the Scalable Amendment of H.264/AVC

Hsiang-Chun Huang, Wen-Hsiao Peng, and Tihao Chiang, National Chiao Tung University
Hsueh-Ming Hang, National Taipei University of Technology

ABSTRACT

To support clients with diverse capabilities, ISO/IEC MPEG and ITU-T form a Joint Video Team (JVT) to develop a scalable video coding (SVC) technology that uses single bitstream to provide multiple spatial, temporal, and quality (SNR) resolutions, thus satisfying low-complexity and low-delay constraints. It is an amendment of the emerging standard H.264/AVC and it provides an H.264/AVC-compatible base layer and a fully scalable enhancement layer, which can be truncated and extracted on-the-fly to obtain a preferred spatio-temporal and quality resolution. An overview of the adopted key technologies in the SVC and a comparison in coding efficiency with H.264/AVC are presented.

INTRODUCTION

To achieve flexible visual content adaptation for multimedia communications, the ISO/IEC MPEG and ITU-T VCEG form the Joint Video Team (JVT) to develop a scalable video coding (SVC) amendment for the H.264/AVC standard [1–3]. With worldwide industrial support, it is in the Committee Draft stage and will be elevated to Final Draft International Standard in January 2007. The SVC can be used for various applications such as multiresolution content analysis, content adaptation, complexity adaptation, and bandwidth adaptation. For example, when the video is transported over error-prone channels with fluctuated bandwidth for Internet or wireless visual communications, the clients, consisting of various devices, require different processing power and spatio-temporal resolutions. To serve diversified clients over heterogeneous networks, the SVC allows on-the-fly adaptation in the spatio-temporal and quality dimensions according to the network conditions and receiver capabilities. During transmission, the server or router truncates the bitstream to match the available bandwidth. Moreover, the client can skip parts of the received bitstream to match its capability in execution cycles and display dimension.

Figure 1 illustrates an application scenario for SVC. In Fig. 1a, the system contains three devices, including server, router, and wireless access point with different connection speeds. Multiple clients are connected to the networks. The SVC bitstream has:

- Two spatial resolutions: Common Intermediate Format (CIF, 352 × 288) and Four CIF (4CIF, 704 × 576)
- Three temporal resolutions: 60 frames/s, 30 frames/s, and 15 frames/s
- Three signal-to-noise ratio (SNR) layers for each spatial resolution

Figure 1b shows the bitstream structure for each connection. The bitstream consists of multiple pictures and each picture contains several spatial and quality resolutions. Initially, the video server retains only the first three SNR layers at the CIF resolution and the first and part of the second SNR layers at the 4CIF resolution to match the 4 Mb/s bandwidth between the video server and the router. To match the 3 Mb/s bandwidth between the router and the wireless access point, the router discards the bitstream for the second SNR layer at the 4CIF resolution and the additional temporal resolutions for 60 frames/s. Similarly, the two wireless clients of lower complexity and display resolution are supported with further truncation. The spatio-temporal pyramid is illustrated in Fig. 1c.

While SVC enjoys flexible bitstream adaptation, it comes with loss of coding efficiency. SVC addresses this issue with several new techniques:

- A hierarchical-B structure is used to support multilevel temporal scalability.
- Adaptive interlayer prediction techniques, including intratexture, motion, and residue predictions, are used to exploit correlations among spatial and SNR coding layers.
- The enhancement layer information is used in the prediction loops to exploit temporal redundancy while the leaky prediction technique can reduce the associated drifting error.
- The context adaptive entropy coding and the cyclic block coding result in improved coding efficiency and subjective quality.

0163-6804/07/$20.00 © 2007 IEEE

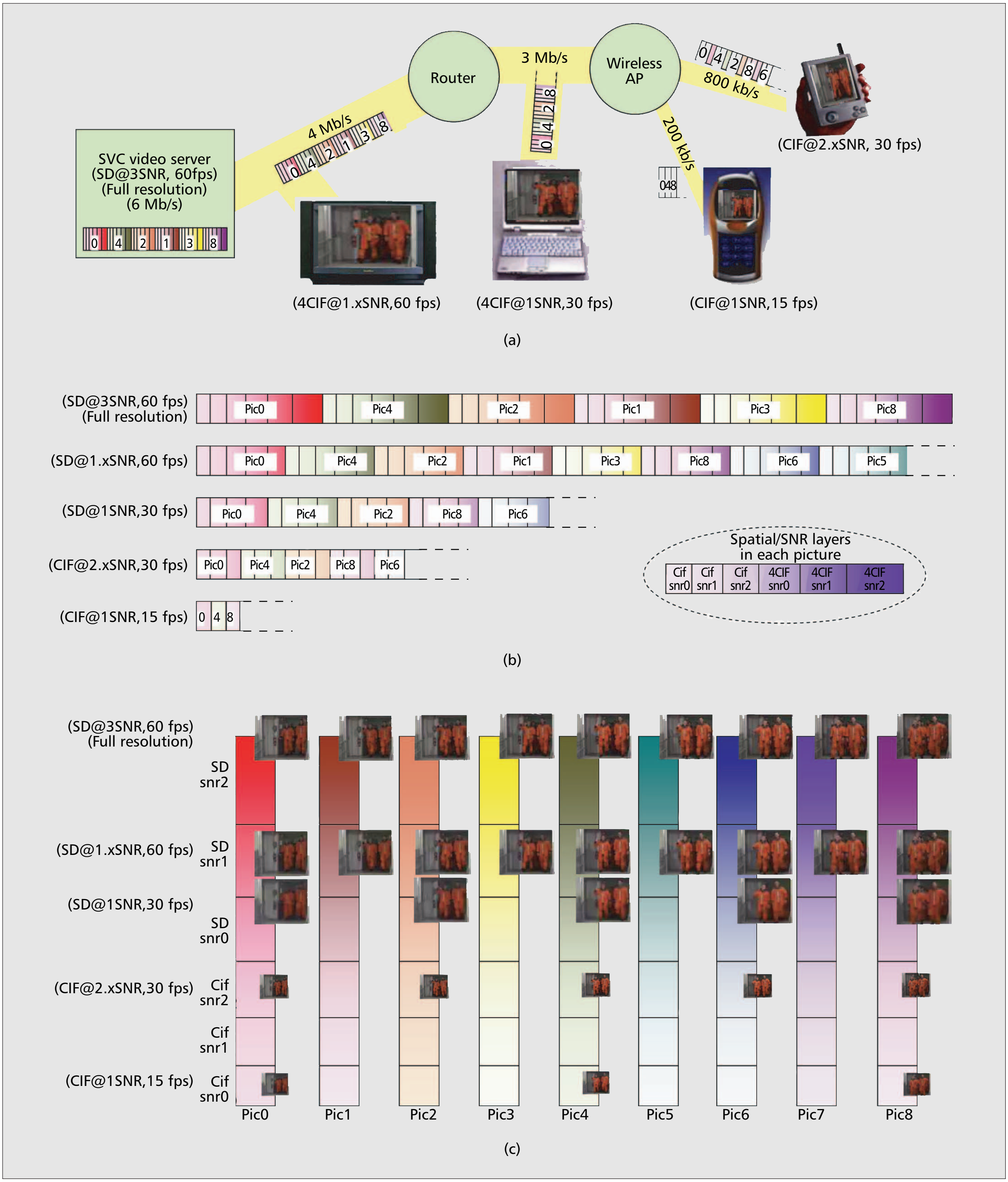

Figure 1. *An example of SVC: a) application scenario; b) bitstream extraction; c) the decoded video.*

- The embedded bit plane coding technique enables fine granularity scalability (FGS).

In this article we provide an overview of these technologies and a comparison of coding efficiency between H.264/AVC and SVC. The technical novelty as compared to the MPEG-2/4 standards is also described. The rest of this article is organized as follows. We describe the encoder structure of SVC. We then examine temporal, SNR, and spatial scalability. We then illustrate the ongoing interlaced representation and bit-stream adaptation. The coding efficiency between nonscalable H.264/AVC and SVC is compared, followed by the concluding remarks.

Temporal scalability is a technique that allows single bitstream to support multiple frame rates. It is typically supported with a pre-determined temporal prediction structure as defined by the standard. In MPEG-2/4, temporal scalability is achieved by the well-known "IBBP" prediction structure.

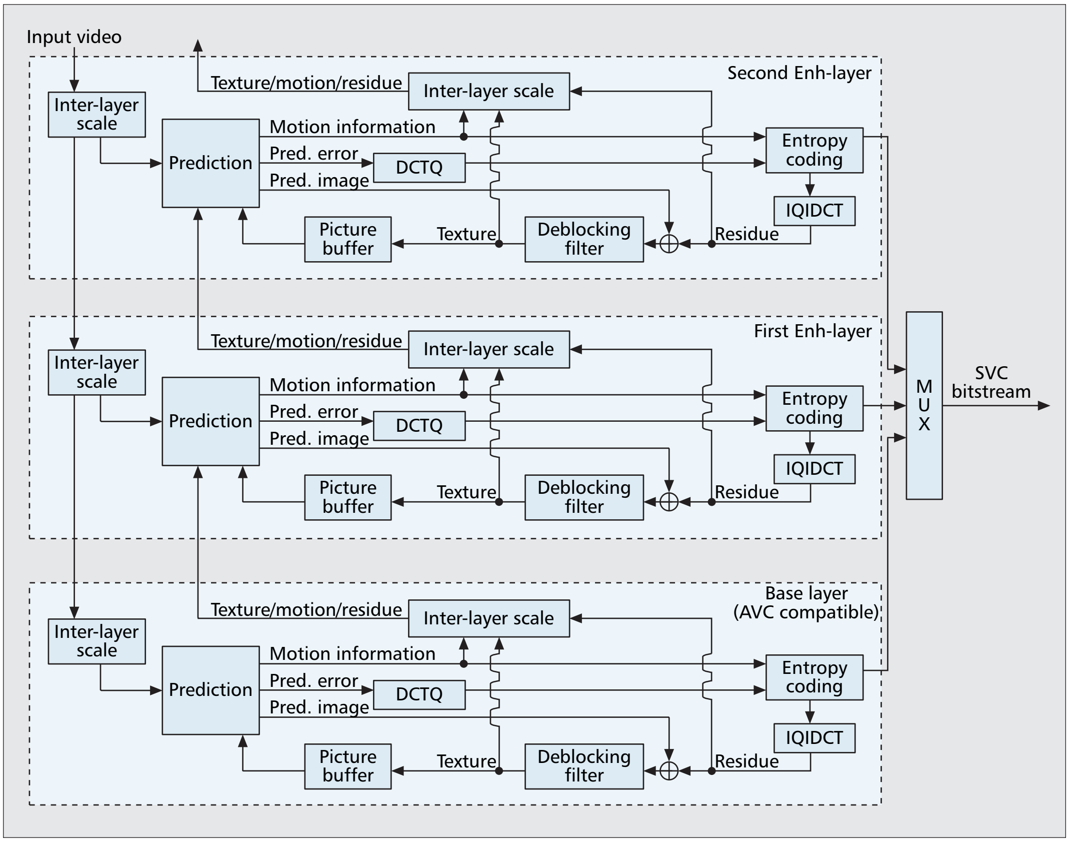

Figure 2. *SVC encoder structure with three spatial/SNR layers.*

Overall Encoder Structure

In this section we present an overview of the encoder structure of SVC. The SVC encodes the video into multiple spatial, temporal, and SNR layers[1] for combined scalability. Figure 2 shows the generic structure of an SVC encoder with three spatial layers (or SNR layers). Each layer is encoded with separated encoders, as shown in the dotted boxes of Fig. 2. The input video is spatially decimated to support multiple spatial resolutions.

For each spatial layer (or SNR layer), the prediction comes from either spatially up-sampled lower layer picture or temporally neighboring pictures at the same layer. Since the information of different layers contains correlations, an interlayer prediction scheme reuses the texture, motion, and residue information of the lower layers to improve the coding efficiency at the enhancement layer. The prediction module needs to interpolate when a layer is up-sampled to different spatial resolution. SVC supports a nondyadic spatial resolution ratio among spatial layers. Temporal prediction utilizes the hierarchical-B structure [5] to support multilevel temporal scalability. The motion-compensated temporal filtering (MCTF) structure can be used as a preprocessing tool for better coding efficiency. The two prediction structures are illustrated in Fig. 3; more detail is described in the next section.

After the prediction module, the residues of each spatial layer (or SNR layer) are entropy encoded with either an embedded coder for FGS, or a nonscalable encoder for coarse granularity scalability (CGS). However, the entropy coding is restricted to nonscalable mode when it is the first SNR layer within a spatial layer. The bitstreams from all spatial or SNR layers are then combined to form the final SVC bitstream. The SVC bitstream can be stored in a server and adapted on-the-fly according to the network conditions or client capabilities, as shown in Fig. 1. In the following sections, we will describe the detail for temporal, SNR, and spatial scalability.

Temporal Scalability

Temporal scalability is a technique that allows single bitstream to support multiple frame rates. It is typically supported with a predetermined temporal prediction structure as defined by the standard. In MPEG-2/4, temporal scalability is achieved by the well-known "IBBP" prediction structure. Up to three frame rates are supported by decoding I-pictures only, both I- and P-pictures, or all of the I-, P-, and B-pictures, respectively. In the H.264/AVC and SVC, more levels are possible with hierarchical B-pictures, and MCTF can be used as a preprocessing tool for better coding efficiency.

Motion-Compensated Temporal Filtering

MCTF is a temporal decomposition technique that adaptively performs the wavelet decomposition and reconstruction along the motion trajectory using Haar and 5/3 wavelets, which can be implemented with lifting schemes with only one

[1] *In this article we use "SNR layer" instead of "quality layer" to indicate layers at the same resolution but with different quality. This is to prevent ambiguity with the quality layer technique used for bitstream adaptation, which is described subsequently.*

prediction/update step. Particularly, the lifting scheme of 5/3 wavelet is realized by traditional bidirectional prediction. In Fig. 3a, layer 3 contains full resolution and the 5/3 wavelet is used for most predictions. For temporal decomposition, the odd-indexed pictures are predicted from the adjacent even-indexed pictures to produce the high-pass pictures. The even-indexed pictures are updated to generate low-pass pictures using combination of the adjacent high-pass pictures.

When the Haar wavelet is selected, the unidirectional prediction is formed. As illustrated in Fig. 3a, the selected prediction and update paths of Picture 3 can be removed. The unidirectional prediction can be either forward or backward. The selection of uni/bidirectional prediction (i.e., the selection of Haar and 5/3 wavelet) is adaptive for each block. To remove the temporal redundancy, motion compensation is conducted before the prediction and update steps.

For temporal scalability of multiple levels, wavelet decomposition is recursively applied on the low-pass pictures of different layers. Using *n* decomposition stages, up to *n* levels of temporal scalability can be achieved. The video of lower frame rate consists of the low-pass pictures at lower layer. After the decomposition, the low-pass picture in layer 0 and the .high-pass pictures in the other layers are encoded in the bitstream.

The MCTF structure requires memory buffer and coding delay equal to the whole GOP size. To reduce complexity, some backward prediction/update path can be removed. As illustrated in Fig. 3, removal of the selected prediction/update paths reduces the memory requirement and coding delay to half (or a quarter) of the GOP size. More detailed discussion on MCTF is available in [6].

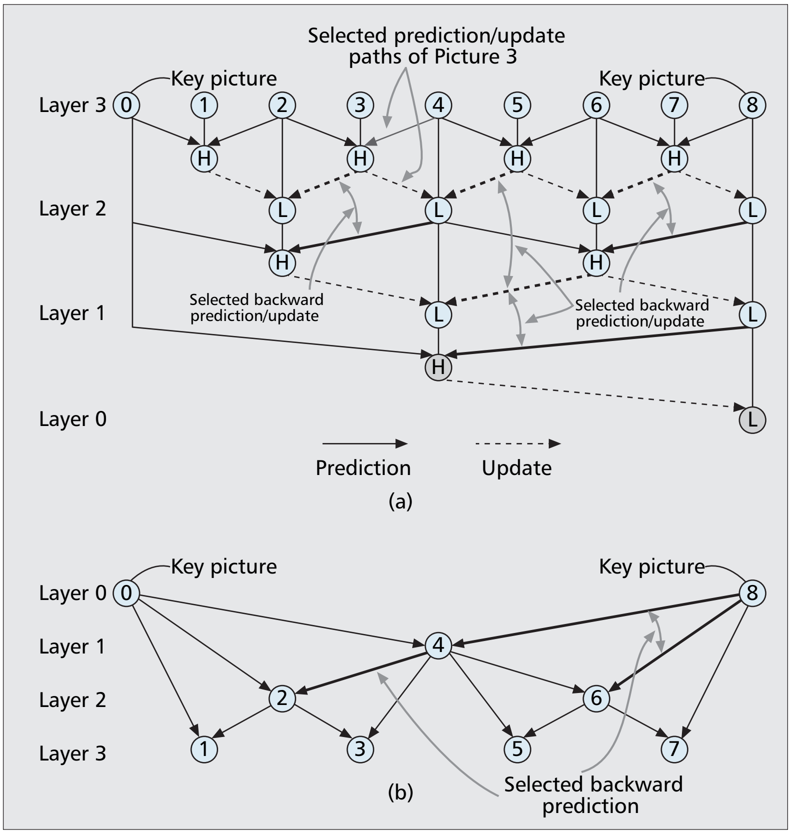

Figure 3. *Temporal decomposition: a) MCTP prediction structure; b) hiearchical-B prediction structure.*

Hierarchical-B Structure

In MCTF, the original pictures are employed for prediction leading to an open-loop control. With such a control, the encoder provides better prediction, since original pictures has higher quality. However, it causes mismatch error between the encoder and decoder due to quantization error. Furthermore, the update step doubles the complexity and increases memory requirement.

To investigate the performance of loop control and justify the complexity increase of the update step, several studies have shown that the closed-loop structure without update step outperforms the open-loop MCTF structure in most testing conditions [5]. The update step can be replaced by a simpler preprocessed noise reduction filter and it can be disabled at the decoder side without significant subjective quality degradation. However, the update step at the encoder side does reduce the quality variation of decoded pictures. After these studies, a closed-loop control at encoder side replaces the open-loop control and the update step is now removed from the normative parts of SVC. This new temporal decomposition structure is known as "hierarchical-B" or "pyramid-B" prediction structure, as shown in Fig. 3b. To support closed-loop encoding, the pictures at lower layers are encoded first such that the pictures at higher layers can refer to the reconstructed pictures at lower layers. Another advantage is that such a prediction scheme is already supported by the syntax of H.264/AVC [1]. Comparing to the "IBBP" structure, the hierarchical-B structure has better coding efficiency using more efficient frame level bit allocation, especially for sequences with fine texture and regular motion. To reduce the memory requirement and coding delay, similar concept used in MCTF can be applied to hierarchical-B structure.

SNR Scalability

SNR scalability consists of CGS and FGS. The former encodes the transform coefficients in a nonscalable way while the latter can be truncated at any location.

Coarse Grain Scalability

The CGS layer data can only be decoded as an integral part. Similar technique exists in the MPEG-2 SNR Scalable Profile. In MPEG-2, the decoder contains only one prediction loop and one motion vector set, both the base and enhancement layer information are used for prediction. The encoder can use either both layers or only base layer in the prediction loop. The former approach enjoys high coding efficiency when both layers are received, but it suffers from drift when only the base layer is received. The latter approach has better performance when only base layer is received.

In SVC, there are several new techniques to

The transform coefficients are represented by significance and refinement symbols in zigzag order. Each symbol is assigned with a scanning position according to its location in zigzag order. Then, symbols from different blocks are coded in a cyclical manner based on their scanning positions.

address these issues found in MPEG-2. For example, each CGS layer has separate motion vectors and temporal prediction mode. It solves the drift problem and allows individual optimization for each layer. As discussed below, the interlayer prediction exploits redundancy from lower layers. Spatial interpolation is unnecessary as all layers have identical resolution.

Fine Grain Scalability

The FGS layer arranges the transform coefficients as an embedded bitstream enabling truncation at any arbitrary point. The FGS technique was first standardized in MPEG-4. However, the enhancement layer is intracoded to prevent drifting error should the enhancement layer be corrupted. The enhancement layer is encoded with Huffman code, while both context adaptive method and arithmetic coding are not considered.

In SVC FGS, the enhancement layer information is used to improve the temporal prediction. The drift problem is alleviated with leaky prediction and the hierarchical prediction structure, as discussed above. For the FGS encoding, there are three cyclical techniques, including normal, vector, and group modes, to achieve embedded representation and improve visual quality. The transform coefficients are represented by significance and refinement symbols in zigzag order. Each symbol is assigned with a scanning position according to its location in zigzag order. Then, symbols from different blocks are coded in a cyclical manner based on their scanning positions.

The significance symbol records the significance and insignificance of each coefficient. Each significance symbol contains an end-of-block and a "significance run" followed by a significant coefficient. The end-of-block signals whether the last significant coefficient of a block is reached or not. Accordingly, in zigzag order, the significance run indicates the insignificant coefficients between two significant ones in the current layer. For a significance symbol, its scanning position is the zigzag index where the significance run starts. The refinement symbol denotes the refinement magnitude of –1 to +1 for coefficient that was significant in the subordinate layers. Similarly, for a refinement symbol, its scanning position is the zigzag index where the significant coefficient is refined.

In cyclical coding, different types of symbols are jointly coded in multiple cycles. In the normal mode, the symbols from different blocks with scanning positions set to the cycle number are coded in a cycle. However, in vector and group modes, the symbols coded in a block must reach a specified scanning index before the next block is enabled for encoding. In the vector mode, the scanning indices to be reached for different cycles are coded by a vector in the picture parameter set. In the group mode, the syntax *groupingSizeMinus1* defines the number of scanning positions in a coding cycle. When the enhancement layer is truncated, the normal mode provides more uniform quality for different blocks. Both the vector and group modes reduce memory access. Each symbol is coded by Context Adaptive Binary Arithmetic Coding (CABAC) or Context Adaptive Variable Length Code (CAVLC).

Besides using different cyclical modes and entropy coders, each FGS slice provides *motion refinement flag* to select prediction process. When this flag is set to 0, the motion information is not refined and the FGS layer reuses the motion of the previous SNR layer and successively refines the residue of the previous SNR layer. When the flag is set to 1, it has its own motion and the residue is adaptively predicted from the previous SNR layer. The motion refinement provides up to 1 dB gain, which enables FGS to provide similar performance as CGS.

Adaptive Reference FGS

In the hierarchical-B structure described above, the key pictures get temporal prediction only from the base layer of the previously coded key pictures, but the nonkey pictures include both the base and SNR enhancement layers for temporal prediction. Since the base layer has low bit rate and thus poor quality, the key pictures generally have poor prediction efficiency. To improve coding efficiency, the prediction of key pictures should incorporate the SNR enhancement layers. However, drift occurs when the enhancement layer is truncated. The same problem also exists in the nonkey pictures, but the hierarchical-B structure significantly constrains the length of the prediction path and propagation of drift. The drift problem of key pictures was also extensively discussed during the development of MPEG-4 FGS [7]. In MPEG-4 FGS, the enhancement layer is only predicted from the base layer with poor quality, leading to poor coding efficiency. Several works employ the enhancement layer for prediction with various drift control mechanism [8, 9]. In particular, robust FGS (RFGS) [9] uses leaky prediction to improve coding efficiency while constraining drifting errors. The prediction from the enhancement layer is multiplied by a leaky factor, which is smaller than one, in each prediction loop. When the predicted data from the enhancement layer are truncated, the drift is decayed by the leaky factor in each prediction loop leading to 3 to 4 dB improvement [9]. The stack robust FGS (SRFGS) further incorporates multiple prediction loops to improve R-D performance over a wide range of bit rates [10].

In SVC the adaptive reference FGS (ARFGS) approach adaptively selects the leaky factor at transform coefficient level for improving the coding efficiency of key pictures. The ARFGS prediction process is performed in the transform domain. For each coefficient at the enhancement layer, the ARFGS reference coefficient is constructed from both the co-located coefficient at the base layer and the predicted coefficient at the enhancement layer from the previous frame. Depending on whether the co-located residue at the base layer is zero or not, the ARFGS reference coefficient is set as a weighted average of the two sources. After generating the ARFGS reference coefficients, they are inversely transformed back to spatial domain to obtain the ARFGS reference block. If all the co-located residues in the base layer are zeros, the derivation of ARFGS reference block is simplified to the weighted average of the two sources in the spatial domain, and the transform domain prediction process is skipped. In addition, the multiloop prediction in SRFGS is also implemented in SVC. A single enhancement layer loop decoding method can be used to reduce complexity with some degradation of the coding efficiency improvement of multiloop prediction.

Spatial Scalability

Similar to the MPEG-2/4 approach, spatial scalability is achieved by decomposing the original video into a spatial pyramid. As shown in Fig. 2, each spatial layer is encoded independently while the motion and temporal prediction are derived from the reference pictures at the same layer. To remove the redundancy among layers, in MPEG-2/4 the interlayer prediction comes from only the reconstructed picture of the most recent layer. However, in SVC such texture prediction can come from any lower layers. Furthermore, in SVC the motion and residue information of the lower layers are reused. In the following sections, we first describe the flexible interlayer prediction structures in SVC, followed by the three interlayer prediction techniques: intra texture, motion, and residue prediction.

Interlayer Prediction Structure

Interlayer prediction is dependent on the types of layers used. The spatial and CGS layers can flexibly select the reference layer from any lower layers while the FGS layer must be predicted from the previous SNR layer at the same resolution.

As demonstrated by an example in Fig. 4, the three columns represent three spatial resolutions: QCIF, CIF, and 4CIF. Each spatial resolution contains several SNR layers. In the first QCIF column, the QCIF_L0 is the lowest layer that is compatible with H.264/AVC. On top of the QCIF_L0, QCIF_L1 and QCIF_L2 are encoded as CGS layers, which are predicted from QCIF_L0 and QCIF_L1, respectively. In the second CIF column, CIF_L0 is the base layer of the second spatial layer. With flexible selection of the reference layer, CIF_L0 can refer to QCIF_L1 instead of QCIF_L2, while CIF_L1 can refer to QCIF_L2 instead of CIF_L0. In this example, CIF_L1 is decodable even when CIF_L0 is corrupted. The rule for the FGS layer is different for CGS and spatial layer. The FGS layer can only refer to previous SNR layer with the same resolution. With the configuration shown in Fig. 4, the decoding of certain layer may not need all the layers at lower resolution. For instance, the QCIF_L2 is not necessary for decoding CIF_L0. Similarly, CIF_L0 is not necessary for decoding CIF_L1. Such flexibility enables rate-distortion performance optimization or error resilience.

Intra Texture Prediction

Intratexture prediction comes from a reconstructed block in the reference layer. Motion compensation is necessary when such a block is either an inter block or an intra block predicted from its neighboring inter blocks. When multiple spatial layers are coded, such a process may be invoked multiple times leading to significant complexity.

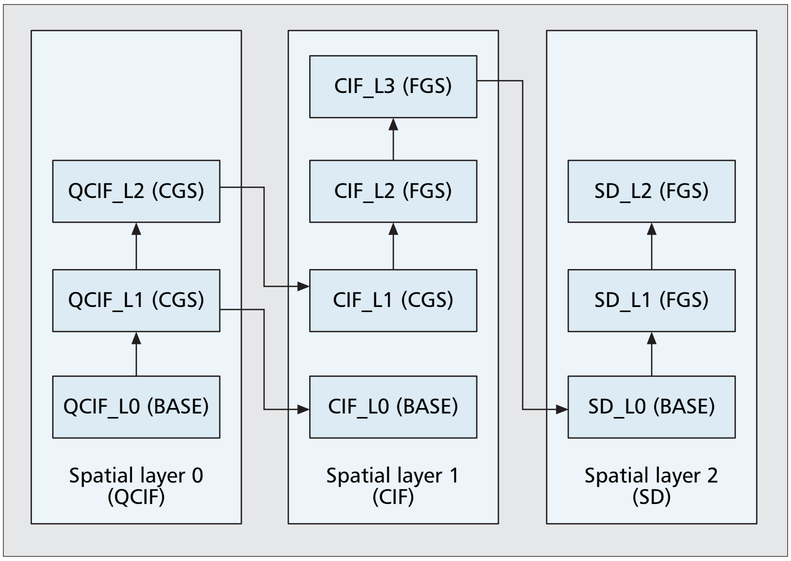

Figure 4. *Configuration of interlayer prediction*

To reduce the complexity, constrained interlayer prediction is used to allow only intra texture prediction from an intra block at the reference layer. Moreover, the referred intra block can only be predicted from another intra block (i.e., the reference layer reuse of "constrained intra prediction" in H.264/AVC). In this way, the motion compensation is invoked only at the highest layer. Such a constraint is also referred to as "single loop decoding."

Motion Prediction

Motion prediction is used to remove the redundancy of motion information, including macroblock partition, reference picture index, and motion vector, among layers. In addition to the macroblock modes available in H.264/AVC, SVC creates an additional mode, namely, the *base-layer mode*, for the interlayer motion prediction. The base-layer mode reuses the motion information of the reference layer without spending extra bits. If this mode is not selected, independent motion is encoded. Note that the motion vectors and macroblock partition of the reference layer may be interpolated before the prediction.

Residue Prediction

Residue prediction is used to reduce the energy of residues after temporal prediction. A similar idea was proposed in PFGS [8], where the DCT coefficients of the enhancement layer are predicted from those of the base layer. In SVC, the residue prediction is performed in the spatial domain. Due to the interlayer motion prediction, consecutive spatial layers may have similar motion information. Thus, the residues of consecutive layers may exhibit strong correlations. However, it is also possible that consecutive layers have independent motion and thus residues of two consecutive layers become uncorrelated. Therefore, the residue prediction in SVC is done adaptively at macroblock level.

The concept of the quality layer is to add side information in the NAL units that encapsulates FGS layers so as to provide better bitstream adaptation. The quality layer id is sent as side information with each NAL unit to signal the importance of each unit.

Interlaced Coding

While the SVC has considered progressive format so far, the interlaced coding tools are necessary when applying the scalability among several common video formats. In interlaced coding, the main issue is interlayer prediction, since two successive layers may be coded by different modes. Some proposals utilize a "two-step" approach: one step deals with the interlayer prediction between different modes (frame or field), but with the same resolution; another step handles the interlayer prediction between different resolutions, but with the same mode. The first step is applied on the base layer to generate a "virtual layer" while the second step is applied further on the "virtual layer" to produce the final interlayer prediction.

Bitstream Extraction and Adaptation

The SVC bitstream contains a set of predefined spatio-temporal and quality resolutions. An extractor can be used to extract the bitstream for the prescribed resolution. There are two extraction methods, namely, simple truncation and quality layers extraction.

Simple Truncation

For simple truncation [3], the extractor determines all the reference layers required for decoding the base layer of the requested spatio-temporal resolutions. Because of the sequential encoding process, the lower layers have higher priority in the extraction process. The higher layer is excluded first if the requested bit rate only allows partial layers to be transmitted. If more bandwidth is available, the SNR layers of the requested spatio-temporal resolutions are then transmitted. If CGS is used for SNR scalability, the bitstream needs to be truncated at the layer boundary. If FGS is used, every picture is equally truncated according to the target bit rate.

Quality Layer Adaptation

The concept of the quality layer is to add side information in the NAL units that encapsulates FGS layers so as to provide better bitstream adaptation. The *quality layer id* is sent as side information with each NAL unit to signal the importance of each unit. The extractor can drop a packet according to the quality layer id, that is, the packet of least importance will be dropped first.

Bitstream extraction, similar to the simple truncation method, keeps the required reference layers from the lower layers to the higher layers until the base layer of the requested spatio-temporal resolution is reached. At the requested spatio-temporal resolution, the extractor firstly computes the bit rate of each quality layer and then removes the NAL units according to the quality layer id. If the target bit rate cannot cover all the NAL units of a quality layer, all the NAL units with this quality layer id will be equally truncated. From the simulation results, the concept of quality layer provides up to 0.5dB PSNR improvement vs. simple truncation. An bitstream extraction technique for FGS is described in [4].

Performance Comparison between H.264/AVC and SVC

Here we compare the coding efficiency of H.264/AVC and SVC. For the simulation, we encode the sequence Crew using the H.264/AVC reference software, JM (Joint Model), with version 10.1, and the SVC reference software, JSVM (Joint Scalable Video Model), with the tag JSVM_6_8_1. Both H.264/AVC and SVC have the same GOP size of 32 and all the key pictures are intra coded. Without any particular statements, the other configurations are identical to those in [5].

We firstly demonstrate the limitation of H.264/AVC while being sent through network with fluctuated bandwidth. We then demonstrate the loss of coding efficiency due to scalability. The comparison contains three parts: SVC with spatial scalability only, SVC with SNR scalability only, and SVC with combined scalability (i.e., simultaneously enable spatial, temporal, and SNR scalability). Temporal scalability is not compared separately because it is already supported in H.264/AVC by the hierarchical-B structure.

SVC with Combined Scalability vs. AVC with Picture Skipping

In this experiment, the limitation of H.264/AVC is demonstrated for transmission over networks with fluctuated bandwidth. There are three H.264/AVC bitstreams for the three resolutions: 4CIF, CIF, and QCIF. The frame rate and GOP structure are the same as those mentioned below. When bandwidth is reduced, bit rate adaptation of H.264/AVC is achieved with skipped frames. To compute the PSNR, the skipped picture is concealed with the temporal direct mode in H.264/AVC. The performance is compared against the SVC with combined scalability described below. As shown in Fig. 5a, when half of the pictures are skipped for H.264/AVC, the PSNR is worse than SVC by 1.5 to 2.0 dB. When more pictures are skipped, the performance becomes even worse. It demonstrates that SVC outperforms H.264/AVC for transmission over networks with fluctuated bandwidth.

SVC with Spatial Scalability Only

In this comparison, the bitstream contains three spatial layers: QCIF, CIF, and 4 CIF. The SNR scalability is disabled and the distortion of each bit rate is generated by multiple encoding all at 60 frames/s. As shown in Fig. 5b, the QCIF layer, which is H.264/AVC compatible, has identical performance as the H.264/AVC. At the CIF layer, there is 0.5 dB loss compared with H.264/AVC. At the 4 CIF layer, the loss is up to 1.0 dB at low bit rate and around 0.3 dB at high bit rate. As expected, scalability is gained at minor loss of coding efficiency.

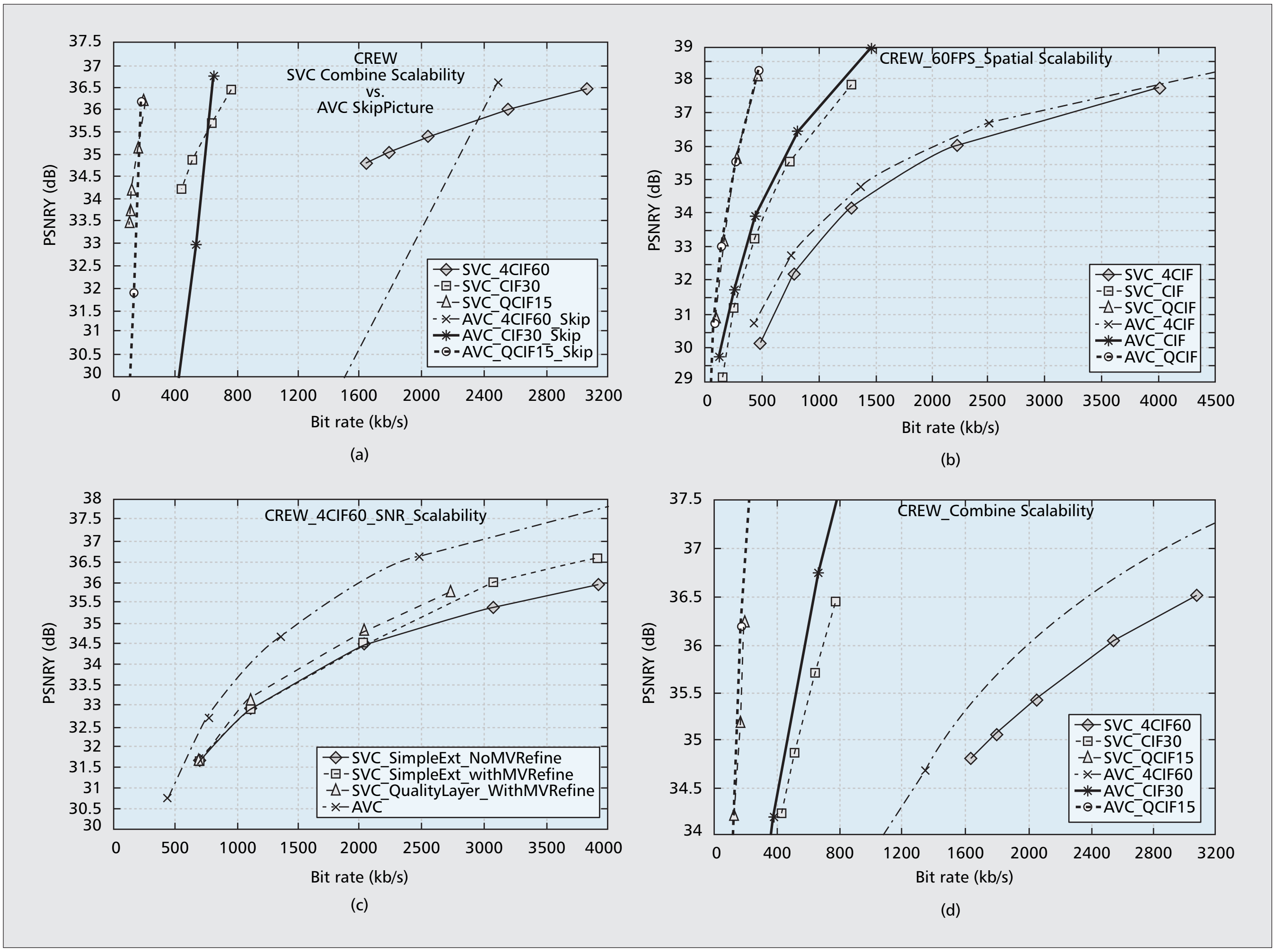

Figure 5. *Performance comparison betweeen H.264/AVC and SVC: a) SVC with combined scalability vs. AVC with picture skipping; b) SVC with spatial scalability only; c) SVC with SNR scalability only; d) SVC with combined scalability..*

SVC WITH SNR SCALABILITY ONLY

In this comparison, the bitstream supports SNR scalabilities with FGS. Both the simple extraction and the quality layer methods are tested. The performance of motion refinement is also tested. Note that quality layer has some problems in the JSVM_6_8_1 so the results at high bit rate are not shown. The 4CIF is encoded at 60 frames/s. As shown in Fig. 5c, the SVC with motion refinement offers 0.6 dB improvement at high bit rate. Furthermore, quality layer truncation has 0.3 dB improvement compared with the simple extraction. However, as compared to H.264/AVC, SVC still has up to 1.2 dB PSNR loss.

SVC WITH COMBINED SCALABILITY

In this comparison, the bitstream supports spatial, temporal, and SNR scalabilities. For the SNR scalability, we use FGS with motion refinement and quality layer truncation. Both the H.264/AVC and SVC is encoded with 60 frames/s at 4 CIF, 30 frames/s at CIF, and 15 frames/s at QCIF. The GOP size is 32/16/8 for 4CIF/CIF/QCIF, respectively. As shown in Fig. 5d, SVC has PSNR loss from 0.5 dB to 0.9 dB, as compared to H.264/AVC.

CONCLUSIONS

As an amendment of H.264/AVC, SVC provides an H.264/AVC compatible base layer and a fully scalable enhancement layer that supports spatial, temporal, and SNR scalability. For spatial scalability, the pyramid structure is used with improved interlayer prediction. For temporal scalability, the hierarchical-B structure is adopted and may improve the coding efficiency. For SNR scalability, both CGS and FGS are supported with successive quantization. To assist the bitstream adaptation process, priority information can be embedded in the NAL units. As expected, scalability is gained with loss of coding efficiency. As compared to H.264/AVC, SVC has 0.3 to 1.2 dB PSNR loss. Thus, coding efficiency is still an issue for SVC.

REFERENCES

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

[2] ITU-T and ISO/IEC JTC1, JVT-T201r2, "Joint Draft 7 of SVC Amendment (Revision 2)," July 2006.

[3] ITU-T and ISO/IEC JTC1, JVT-S202, "Joint Scalable Video Model JSVM-6," Apr. 2006.

[4] X. M. Zhang *et al.*, "Constant Quality Constrained Rate Allocation for FGS-Coded Videos," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 13, no. 2, Feb. 2003, pp. 121–30.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Comparison of MCTF and Closed-Loop Hierarchical B Pictures," ITU-T and ISO/IEC JTC1, JVT-P059, July 2005.

[6] J. R. Ohm, "Advances in Scalable Video Coding," *Proc. IEEE*, vol. 93, no. 1, Jan. 2005, pp. 42–56.

[7] ISO/IEC JTC1/SC29/WG11/N3904, "Streaming Video Profile— Final Draft Amendment (FDAM 4)," Jan. 2001.

[8] F. Wu, S. Li, and Y. Q. Zhang, "A Framework for Efficient Progressive Fine Granularity Scalable Video Coding," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 11, no. 3, Mar. 2001, pp. 332–44.

[9] H. C. Huang, C. N. Wang, and T. Chiang, "A Robust Fine Granularity Scalability Using Trellis Based Predictive Leak," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 12, no. 6, June 2002, pp. 372–85.

[10] H. C. Huang and T. Chiang, "Stack Robust Fine Granularity Scalability," *IEEE Int'l. Symp. Circuits Syst.*, vol. 3, 2004, pp. III-829–32.

Biographies

TIHAO CHIANG (tchiang@mail.nctu.edu.tw) received a Ph.D. degree in electrical engineering from Columbia University in 1995. Then he joined the David Sarnoff Research Center as a member of technical staff and was later promoted to program manager. In 1999 he joined the faculty at National Chiao-Tung University (NCTU), Taiwan, R.O.C. On his sabbatical leave in 2004, he worked with Ambarella USA and initiated its R&D operation in Taiwan.

HSUEH-MING HANG [F] (hmhang@mail.nctu.edu.tw) was with AT&T Bell Laboratories, Holmdel, NJ, from 1984 to 1991. He joined NCTU in December 1991, and is currently Dean of the Electrical Engineering and Computer Science College of National Taipei University of Technology. He has been involved in the international video standards since 1984. He is a recipient of the IEEE Third Millennium Medal.

HSIANG-CHUN HUANG (sleeping.ee89g@nctu.edu.tw) received B.S. and Ph.D. degrees in electronics engineering from NCTU in 2000 and 2006, respectively. He is currently a member of technical staff at Ambarella Taiwan Ltd., Hsinchu, Taiwan, R.O.C. His research interests are scalable video coding and video encoder architecture optimization.

WEN-HSIAO PENG (pawn@mail.si2lab.org) received B.S. and M.S. degrees with highest distinction in electronics engineering from NCTU in 1997 and 1999, respectively. In 2000 he joined Intel Microprocessor Research Laboratory, Santa Clara, CA, where he developed the first real-time MPEG-4 FGS codec and demonstrated its application in 3D peer-to-peer videoconferencing. In 2002 he joined the Institute of Electronics of NCTU as a Ph.D. candidate. He received his Ph.D. degree in 2005 with his dissertation, "Scalable Video Coding — Advanced Fine Granularity Scalability." Since 2003 he has actively participated in ISO's Moving Picture Expert Group (MPEG) digital video coding standardization process and contributed to the development of the MPEG-4 Part 10 AVC Amd.3 scalable video coding standard. His major research interests include scalable video coding, video codec optimization, and platform-based architecture design for video compression. He has published more than 30 technical papers in the field of video and signal processing.

IEEE Communications Magazine
Call for Papers
Network & Service Management Series

IEEE Communications Magazine announces the creation of a new series on Network and Service Management. The series will be published twice a year, in April and October, with the first issue planned for October 2005. It intends to provide articles on the latest developments in this well-established and thriving discipline. Published articles are expected to highlight recent research achievements in this field and provide insight into theoretical and practical issues related to the evolution of network and service management from different perspectives. The series will provide a forum for the publication of both academic and industrial research, addressing the state of the art, theory and practice in network and service management. Both original research and review papers are welcome, in the style expected for IEEE Communications Magazine. Articles should be of tutorial nature, written in a style comprehensible to readers outside the speciality of Network and Service Management. This series therefore complements the newly established IEEE Electronic Transactions on Network & Service Management (eTNSM). General areas include but are not limited to:

* Management models, architectures and frameworks
* Service provisioning, reliability and quality assurance
* Management functions
* Management standards, technologies and platforms
* Management policies
* Applications, case studies and experiences

The above list is not exhaustive, with submissions related to interesting ideas broadly related to network and service management encouraged.

IEEE Communications Magazine is read by tens of thousands of readers from both academia and industry. The magazine has also been ranked the number one telecommunications journal according to the ISI citation database for year 2000, and the number three for year 2001. The published papers will also be available on-line through Communications Magazine Interactive, the WWW edition of the magazine. Details about IEEE Communications Magazine can be found at http://www.comsoc.org/ci/.

Schedule for the First Issue

Manuscripts due: March 30, 2005
Acceptance notification: June 30, 2005
Manuscripts to publisher: July 30, 2005
Publication date: October 2005

Series Editors

Prof. George Pavlou
Center for Communication Systems Research
Dept. of Electronic Engineering
University of Surrey
Guilford, Surrey GU2 7XH, UK.
e-mail: G.Pavlou@eim.surrey.ac.uk

Dr. Aiko Pras
Center for Telematics and Information Technology
Dept. of Electrical Engineering, Mathematics and Computer Science
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands.
e-mail: a.pras@utwente.nl