Scalable Video Coding and Applications

N.S. Narkhede and Ramesh Prasad

***Abstract*—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 in H.264/AVC standard. SVC enables the transmission and decoding of partial bit streams to provide video services with lower temporal or spatial resolutions or reduced fidelity while retaining a reconstruction quality that is high relative to the rate of the partial bit streams. Hence, SVC provides functionalities such as graceful degradation in lossy transmission environments as well as bit rate, format, and power adaptation. These functionalities provide enhancements to transmission and storage applications.SVC has achieved significant improvements in coding efficiency with an increased degree of supported scalability relative to the scalable profiles of prior video coding standards.**

***Index Terms*—SVC, H.264, MPEG-4, AVC, JVT, standards, video.**

I. Introduction

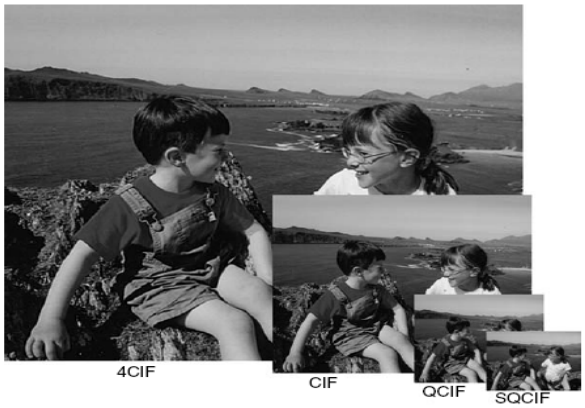
RAPID growth in communications infrastructure and increase in processing power of the communications devices, coupled with the advances in the video coding technology, is enabling deployment of new media centric applications. These applications must work in heterogeneous network environments with different capabilities. Applications like Multimedia Messaging (MMS), video on demand, video conferencing, TV broadcast etc should work with wireline as well as wireless networks. Considering the vast difference in the bit-rates, bit error rate (BER), packet loss probabilities of these two networks, and the capabilities of the respective communication devices used, the video stream must be adapted individually for each network type. Traditional video systems are characterized by a fixed spatio-temporal format of the video signal (SDTV or HDTV or CIF for H.320 video telephone). Their application behavior in such systems typically falls into one of the two categories: it works or it doesn't work. To support multiple video formats, multiple application infrastructure must be created. This entails overheads on the application infrastructure and hence increased cost. To avoid these overheads, scalability of the video bit-stream becomes a very attractive feature.

The term "scalability" in this paper refers to the removal of parts of the video bit stream in order to adapt it to the various needs or preferences of end users as well as to varying terminal capabilities or network conditions. The objective of the H.264 SVC standardization has been to enable the encoding of a high-quality video bit stream that contains one or more subset bit streams that can themselves be decoded with a complexity and reconstruction quality similar to that a achieved using the existing H.264/AVC design with the same quantity of data as in the subset bit stream.

N.S. Narkhede and Ramesh Prasad are with University of Mumbai, India

II. Types of Scalability

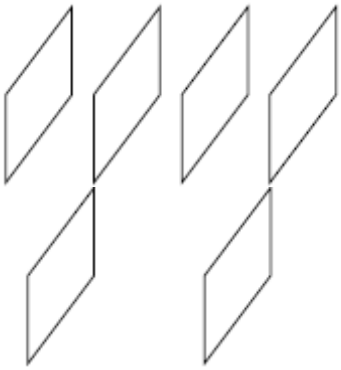
In general, a video bit stream is called scalable when parts of the stream can be removed in a way that the resulting substream forms another valid bit stream for some target decoder, and the sub-stream represents the source content with a reconstruction quality that is less than that of the complete original bit stream but is high when considering the lower quantity of remaining data. Bit streams that do not provide this property are referred to as single-layer bit streams. The usual modes of scalability are temporal, spatial, and quality scalability. Spatial scalability and temporal scalability describe cases in which subsets of the bit stream represent the source content with a reduced picture size (spatial resolution) or frame rate (temporal resolution), respectively. With quality scalability, the substream provides the same spatio-temporal resolution as the complete bit stream, but with a lower fidelity – where fidelity is often informally referred to as signal-to-noise ratio (SNR). Quality scalability is also commonly referred to as fidelity or SNR scalability. The different types of scalability can also be combined, so that a multitude of representations with different spatio-temporal resolutions and bit rates can be supported within a single scalable bit stream.



(a)

PSNR=30.6 db PSNR = 27.7db

(b)



(c)

Figure. 1 Tyapes of scalability : (a) Spatial Scalability, (b) SNR Scalability, (c) Temporal scalability

III. APPLICATIONS

This section discusses some of the applications in which SVC is useful.

A. Video Transmission

The hierarchical scalable video coding approach in general allows the split of the bit-stream into a base layer and various enhancement layers. That allows the transmission of these partial bit-streams via different channels or in different network streams. In IP networks the different quality classes can be assigned to one or even more consumed network streams, thus a possible billing depending on network streams is guaranteed. Such network types could be DVB-H or MBMS of 3GPP, with different terminal capability classes.

The layered coding approach has further benefits. If the network supports transmission of certain network streams to certain devices only, not even all streams have to be received by all terminals. In MBMS or DVB-H networks it should be possible to send out, e.g., the base layer of a scalable H.264/AVC stream to the low performance device only and guarantee that all layers of the stream reach the high performance terminal only. Such an example is shown in Figure 2.

By using the scalable H.264/AVC stream instead of simulcasting, the backbone could be drastically eased.

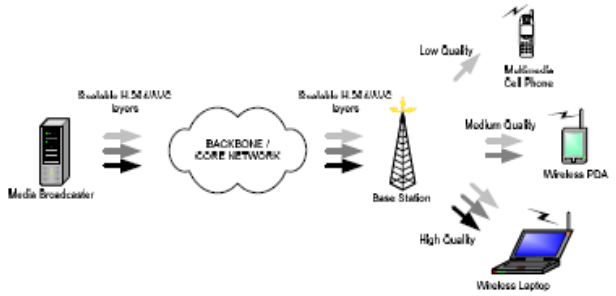


Figure 2: Scalable H.264/AVC video transmission in wireless broadcast networks.

Another benefit can be achieved, if the protection of the scalable H.264/AVC layers is treated in different ways. Forward error correction (FEC) is often used to protect data sent out via wireless broadcast channels. Unequal error or erasure protection (UEP/UXP) schemes can be used to ensure an error free transmission of important layers like, e.g., the base layer of a scalable H.264/AVC stream. UEP/UXP can be used on top of the already existing channel FEC. Such a scheme can guarantee a basic video quality over a large range of channel error rates. Such a scenario is shown in Figure 3.

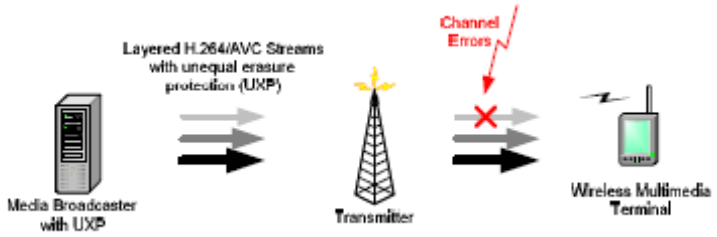


Figure 3: Unequal erasure protection with scalable H.264/AVC quality layers.

B. Video Surveillance

A very promising application of scalable coding is video surveillance. Typically videos from many cameras have to be stored and viewed on diverse displays, which may have different spatial and temporal resolutions (Figure 4). Examples are split screen display of many scenes on one monitor or viewing of scenes from dedicated cameras on mobile devices such as video phones or PDAs. For such applications scalable coding is most attractive, because no transcoding or format conversion is required. Decoding of the lower resolutions videos for split-screen display also saves computing powe therefore many videos can be decoded and displayed with the computing power required for one full resolution video.

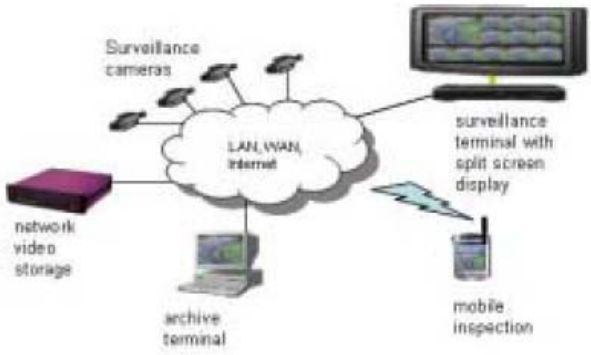


Figure 4: General networked surveillance scenario.

In general, the requirements for surveillance applications can be summarized as follows:

- Simultaneous transmission and storage of different spatial and temporal resolutions in one bit stream, in order to feed different displays
- Fine granularity scalability for feeding different transmission links of varying capacity
- Decoding of lower resolutions should be less complex for split screen display
- Multiple adaptation of the bit stream should be possible for erosion storage

The latter requirement arises from the necessity to store a huge amount of data delivered by the surveillance cameras. By using scalable coding it becomes possible to delete higher resolution layers of stored scenes after certain expiration times and to keep just a lower resolution copy for the archive. This allows a much more flexible usage of storage capacity without the necessity for re-encoding and copying. Typically the full resolution video is kept for 1-3 days, a medium quality (reduced temporal or spatial resolution) video is kept for one week and a low quality (reduced temporal and spatial resolution) video is kept for long time archiving. This functionality results in the following requirements for a file format, which are mostly fulfilled by the MPEG-4 file format as:

- random access to the different layers must be possible without parsing the bit stream
- an appropriate hierarchical storage structure
- a hinting track (e.g. RTP) to support streaming applications

C. Movie on Chip

Nowadays the use of external memory card has increased a lot due to falling prices and support by large number of mobile handsets. This creates a Business opportunity for content providers to reach out to a large number of customers for distribution of content preloaded onto a Memory card. Content like audio and video, which have large file sizes, have restriction on distribution via traditional means like Internet etc. This restricts the reach of this type of content. This restriction is overcome by preloading the content onto the memory card and selling these cards via retail outlets just like CD/DVDs. A consumer spending money for buying a movie would want to use it for PC and home theater as well. But a movie created for mobile doesn't give the same user experience as a DVD. This problem can be solved by using scalable video coding. The video on the memory chip would contain video at different scales for viewing on mobile handset as well as PC/Home Theater, thus making it more attractive for the consumer.

IV. OVERVIEW OF H.264

The H.264/ AVC video coding standard was defined by the Joint Video Team (JVT), which was formed jointly by the ISO/IEC Moving Picture Expert Group (MPEG) and the ITU-T Video Coding Expert Group(VCEG).H.264 concentrates specifically on efficient compression of video frames. Key features of the standard include compression efficiency (providing significantly better compression than any previous standard), transmission efficiency (with a number of built-in features to support reliable, robust transmission over a range of channels and networks) and a focus on popular applications of video compression.

A. Basics of Video Coding

A digitized video signal consists of a periodical sequence of images called frame. Each frame consists of a two dimensional array of pixels. Each pixel consists of three color components, R, G and B. Usually, pixel data is converted from RGB to another color space called YUV in which U and V components can be sub-sampled. A block-based coding approach divides a frame into macro blocks (MBs) each consisting of say 16x16 pixels. In a 4:2:0 format, each MB consists of 16x16 = 256 Y components and 8x8 = 64 U and 64 V components. Each of three components of an MB is processed separately. To compress a MB, a hybrid of three techniques is used: prediction, transformation & quantization and entropy coding. The procedure works on a frame of video.

1) Prediction

Prediction tries to find a reference MB that is similar to the current MB under processing so that, instead of the whole current MB, only their (hopefully small) difference needs to be coded. Depending on where the reference MB comes from, prediction is classified into inter-frame prediction and intra-frame prediction. In an inter-predict (P or B) mode, the reference MB is somewhere in a frame before or after the current frame, where the current MB resides. It could also be some weighted function of MBs from multiple frames. In an intra-predict (I) mode, the reference MB is usually calculated with mathematical functions of neighboring pixels of the current MB.The difference between the current MB and its prediction is called residual error data (residual). It is transformed from spatial domain to frequency domain by means of discrete cosine transform. Because human visual system is more sensitive to low frequency image and less sensitive to high frequency ones, quantization is applied such that more low frequency information is retained while more high frequency information discarded. The third and final type of compression is entropy coding. A variable-length coding gives shorter codes to more probable symbols and longer codes to less probable ones such that the total bit count is minimized. After this phase, the output bit stream is ready for transmission or storage.

There is also a decoding path in the encoder. One has to use a reconstructed frame as the reference for prediction since in the decoder side only the reconstructed frame instead of the original frame is available. The restored residual data is obtained by performing inverse quantization and then inverse transformation. Adding the restored residual to the predicted MB, the reconstructed MB is obtained, that is then inserted to the reconstructed frame. Now, the reconstructed frame can be referred to by either the current I-type compression or future P-type or B-type prediction. Prediction exploits the spatial or the temporal redundancy of a video sequence so that only the difference between actual and predict instead of the whole image data need to be encoded. There are two types of prediction: intra prediction for I-type frame and inter prediction for P-type (Predictive) and B-type (Bidirectional Predictive) frame.

(1) Intra Prediction

There exists high similarity among neighboring blocks in a video frame. Consequently, a block can be predicted from its

neighboring pixels of already coded and reconstructed blocks. The prediction is carried out by means of a set of mathematical functions. In H.264/AVC, an I-type 16x16, 4:2:0 MB has its luminance component (one 16x16) and chrominance components (two 8x8 blocks) separately predicted. There are many ways to predict a macro block. The luminance component may be intra-predicted as one single INTRA 16x16 block or 16 INTRA 4x4 blocks. When using the INTRA 4x4 case, each 4x4 block utilizes one of nine prediction modes (one DC prediction mode and eight directional prediction modes). When using the INTRA16x16 case, which is well suited for smooth image area, a uniform prediction is performed for the whole luminance component of a macro block. Four prediction modes are defined. Each chrominance component is predicted as a single 8x8 block using one of four modes.

(2) Inter Prediction (Motion Estimation)

High quality video sequences usually have high frame rate at 30 or 60 frames per second (fps). Therefore, two successive frames in a video sequence are very likely to be similar.

The goal of inter prediction is to utilize this temporal redundancy to reduce data needed to be encoded. When encoding frame t, we only need to encode the difference between frame t-1 and frame t, instead of the whole frame t. This is called motion estimated inter-frame prediction. In most video coding standards, the block-based motion estimation (BME) is used to estimate for movement of a rectangular block from the current frame. For each M xN-pixel current block in the current frame, BME compares it with some or all of possible M x N candidate blocks in the search area in the reference frame for the best match, as shown in Figure 5.

The reference frame may be a previous frame or a next frame in P-type coding, or both in B-type coding. A popular matching criterion is to measure the residual calculated by subtracting the current block from the candidate block, so that the candidate block that minimizes the residual is chosen as the best match. The cost function is called sum of absolute difference (SAD), which is the sum of pixel by pixel absolute difference between predicted and actual image.

There are three new features of motion estimation in H.264: variable block-size, multiple reference frames and quarter-pixel accuracy.

Variable block-size – Block size determines tradeoff between the residual error and the number of motion vectors transmitted. Fixed block-size motion estimation (FBSME) spends the same efforts when estimating the motion of moving objects and background (no motion). This method causes low coding efficiency. Variable block-size motion estimation (VBSME) uses smaller block size for moving objects and larger block size for background, to increase the video quality and the coding efficiency. H.264 specifies multiple block sizes starting from 16x16 going down to 4x4.

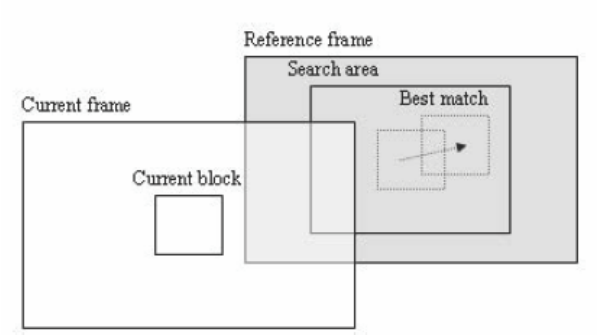


Figure 5. Block-based motion estimation

Multiple reference frames -- In previous video coding standards, there is only one reference frame for motion estimation. In H.264, the number of reference frames increases to 5 for P frame and to 10 (5 previous frames and 5 next frames) for B frame. More reference frames result in smaller residual data and, therefore, lower bit rate. Nevertheless, it requires more computation and more memory traffic.

Quarter-pixel accuracy -- In previous video coding standards, motion vector accuracy is half-pixel at most. In H.264, motion vector accuracy is down to quarter-pixel and results in smaller residual data.

2) Compensation

Corresponding to prediction, there is also two kinds of compensation, intra compensation for I-type frame and inter compensation for P-type and B-type frame.

3) Transformation and Quantization

Discrete Cosine Transform (DCT) is a popular block-based transform for image and video compression. It transforms the residual data from time domain representation to frequency domain representation.

Since most image and video are low frequency data, DCT can centralize the coding information. The main functionality of quantization is to scale down the transformed coefficients and to reduce the coding information. The H.264 standard employs a 4x4 integer DCT.

4) In-loop filter

One of the disadvantages of block-based video coding is that discontinuity is likely to appear at the block edge. In order to reduce this effect, the H.264 standard employs the deblocking filter to eliminate blocking artifact and thus generates a smooth picture.

5) Entropy Coding

There are two popular entropy coding methods; variable length coding and arithmetic coding. The H.264 standard defines two entropy coding methods: context adaptive variable length coding (CAVLC) and context based adaptive arithmetic coding (CABAC) .For baseline profile, only CAVLC is employed.

B. Profiles and Levels

Profiles and levels specify conformance points. These conformance points are designed to facilitate interoperability between various applications of the standard that have similar functional requirements. A profile defines a set of coding

tools or algorithms that can be used in generating a conforming bit-stream, whereas a level places constraints on certain key parameters of the bit stream. In H.264/AVC, three profiles are defined, which are Baseline, Main, and Extended Profile. The Baseline profile supports all features in H.264/AVC except the following two feature sets:

- Set 1: B slices, weighted prediction, CABAC, field coding, and picture or macro block adaptive switching between frame and field coding.
- Set 2: SP/SI slices, and slice data partitioning.

V. OVERVIEW OF SCALABLE EXTENSION IN H.264

This section describes the scalable extension of the H.264

A. Temporal Scalability

Temporal scalability is based on the concept of Hierarchical B frames and reference picture management of H.264 which are described briefly below-

1) Hierarchical B Frames

A typical hierarchical prediction structure with 4 dyadic hierarchy stages is depicted in Figure. 6. The first picture of a video sequence is intra-coded as IDR (instantaneous decoder refresh) picture; so-called key pictures are coded in regular (or even irregular) intervals. A picture is called key picture, when all previously coded pictures precede the picture in display order. As illustrated in Figure. 6, a key picture and all pictures that are temporally located between the current key picture and the previous key picture are considered to build a group of pictures (GOP). The key pictures are either intra-coded (e.g. in order to enable random access) or inter-coded using previous (key) pictures as references for motion-compensated prediction (MCP). The remaining pictures of a GOP are hierarchically predicted as illustrated in Figure. 6 and coded using the bi-predictive (B) slice syntax of H.264/MPEG4-AVC.

Hierarchical prediction structures can also be used for supporting several levels of temporal scalability. For this purpose it has to be ensured that only pictures of a coarser or the same temporal level are employed as references for MCP. Then, the sequence of key pictures represents the coarsest supported temporal resolution, and this temporal resolution can be refined by adding the temporal refinement pictures of finer temporal levels.

2) Reference Picture Management

Each macroblock partition in an inter coded macroblock in a B slice may be predicted from one or two reference pictures, before or after the current picture in temporal order. Depending on the reference pictures stored in the encoder and decoder (see the next section), this gives many options for choosing the prediction references for macroblock partitions in a B macroblock type. Figure 7 shows three examples: (a) one past and one future reference (similar to B-picture prediction in earlier MPEG video standards), (b) two past references and (c) two future references.

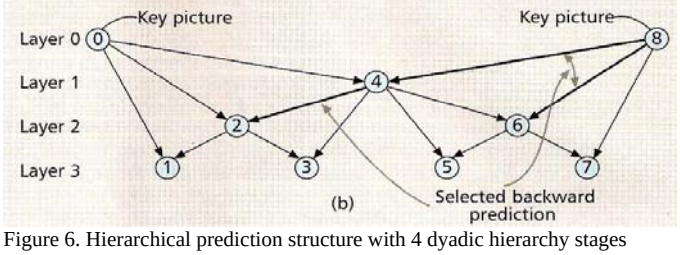


Figure 6. Hierarchical prediction structure with 4 dyadic hierarchy stages

B slices use two lists of previously-coded reference pictures, list 0 and list 1, containing short term and long term pictures. By default, an encoded picture is reconstructed by the encoder and marked as a short term picture, a recently-coded picture that is available for prediction. Short term pictures are identified by their frame number. Long term pictures are (typically) older pictures that may also be used for prediction and are identified by a variable LongTermPicNum. These two lists can each contain past and/or future coded pictures (pictures before or after the current picture in display order). The long term pictures in each list behave in a similar way to the description. The short term pictures may be past and/or future coded pictures and the default index order of these pictures is as follows:

List 0: The closest past picture (based on picture order count) is assigned index 0, followed by any other past pictures (increasing in picture order count), followed by any future pictures (in increasing picture order count from the current picture).

List 1: The closest future picture is assigned index 0, followed by any other future picture (in increasing picture order count), followed by any past picture (in increasing picture order count)[6].

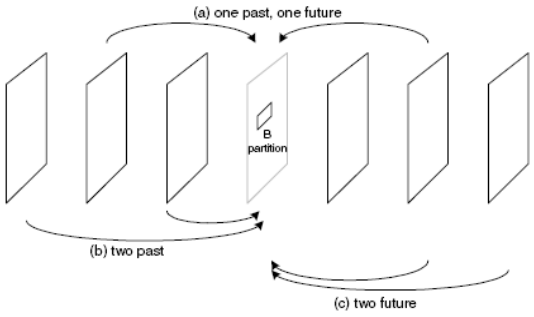


Figure 7. Partition prediction examples in a B macroblock type: (a) past/future, (b) past, (c) future

Temporal scalability with dyadic temporal enhancement layers can be very efficiently provided with the concept of hierarchical B pictures as illustrated in Figure.8(a). The enhancement layer pictures are typically coded as *B* pictures, where the reference picture lists 0 and 1 are restricted to the temporally preceding and succeeding picture, respectively, with a temporal layer identifier less than the temporal layer identifier of the predicted picture. Each set of temporal layers {T0,…,T*k*} can be decoded independently of all layers with a temporal layer identifier $T > k$. In the following, the set of pictures between two successive pictures of the temporal base layer together with the succeeding base layer picture is referred to as a *group of pictures* (GOP).

Hierarchical prediction structures for enabling temporal

scalability can always be combined with the multiple reference picture concept of H.264/AVC. This means that the reference picture lists can be constructed by using more than one reference picture, and they can also include pictures with the same temporal level as the picture to be predicted. Furthermore, hierarchical prediction structures are not restricted to the dyadic case. As an example, Figure8(b). illustrates a non-dyadic hierarchical prediction structure, which provides 2 independently decodable sub-sequences with 1/9-th and 1/3-rd of the full frame rate.

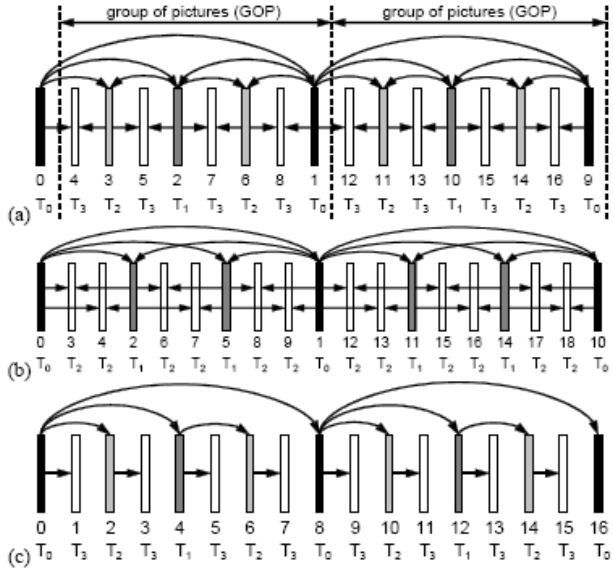


Figure. 8. Hierarchical prediction structures for enabling temporal scalability: (a)coding with hierarchical B pictures, (b) non-dyadic hierarchical prediction structure, (c) hierarchical prediction structure with a structural encoder/ decoder delay of zero. The numbers directly below the pictures specify the coding order, the symbols T*k* specify the temporal layers with *k* representing the corresponding temporal layer identifier.

B. Spatial Scalability

Spatial scalability is achieved by decomposing the original video into spatial pyramid. Each spatial layer is encoded independently. To remove the redundancy among the layers, texture prediction can come from any lower layers. Motion and residue information of the lower layers are reused for temporal prediction. Three types of inter-layer prediction are used-

1) 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 predicted from its neighboring inter blocks. When multiple spatial layers are coded, such a process may be invoked multiple times leading to significant complexity. To reduce the complexity, constrained inter layer 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. In this way, the motion compensation is invoked only at the highest layer.

2) 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.

3) Residue Prediction

Residue prediction is used to reduce the energy of residues after temporal prediction. 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 the 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.

C. SNR Scalability

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

The CGS layer data can only be decoded as an integral part. Each CGS layer has separate motion vectors and temporal prediction mode. The interlayer prediction exploits redundancy from lower layers.

The FGS layer arranges the transform coefficients as an embedded bitstream enabling truncation at arbitrary point. The enhancement layer information is used to improve the temporal prediction.

VI. ARCHITECTURE

The figure 9, shows the architecture for the encoder.

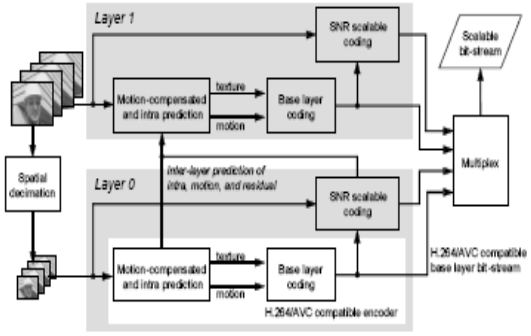


Figure 9. SVC encoder structure example

The input frame is spatially decimated to generate a lower resolution frame for Spatial scalability. The spatially decimated frame is coded using AVC to generate the base layer. The SNR scalable coding block uses this base layer information to generate the Quality scalable stream. This constitutes the Layer 0. Similarly Layer 1 is formed, but it uses the Layer 0 information for Interlayer prediction as described above. All the layers are multiplexed to generate a single scalable bitstream.

VII. CONCLUSION

The scalable extension of the AVC enables many new applications as well as optimizes the existing one. It obviates the need for generating and transmitting multiple streams. A single stream can be used for many different applications and device profiles.

VIII. REFERENCES

Periodicals:

[1] D. Marpe, T. Wiegand, and G. J. Sullivan, "The H.264 / MPEG4 Advanced Video Coding standard and its applications", IEEE Communications Magazine, vol. 44, no. 8, pp. 134-144, Aug. 2006.

[2] G. J. Sullivan, H. Yu, S. Sekiguchi, H. Sun, T. Wedi, S. Wittmann, Y.L. Lee, A. Segall, and T. Suzuki, "New standardized extensions of MPEG4-AVC/H.264 for professional-quality video applications", Proceedings of ICIP'07, San Antonio, TX, USA, Sep. 2007.

[3] Heiko Schwarz, Detlev Marpe, Member, IEEE, and Thomas Wiegand, “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard”. IEEE Transactions on Circuits and Systems for Video Technology, September 2007.

[4] T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, and M. Wien, eds.,"Joint Draft 11 of SVC Amendment," Joint Video Team, doc. JVTX201,Geneva, Switzerland, July 2007.

[5] J. Reichel, H. Schwarz, M. Wien, eds., "Joint scalable video model 1 (JSVM 11)," Joint Video Team, doc. JVT-X202, Geneva, Switzerland, July 2007.

Books:

[6] Iain E. G. Richardson“H.264 and MPEG-4 Video Compression”, Video Coding for Next-generation Multimedia. John Wiley & Sons Ltd, PO19 8SQ, England 2003