Hierarchical Coding Structure for Video Coding and its Applicability in Scalable Video Coding

Gopal Thapa
Sikkim Manipal Institute of Technology, Majitar-Rangpo, Sikkim-737136

Kalpana Sharma
Sikkim Manipal Institute of Technology, Majitar-Rangpo, Sikkim-737136

M.K.Ghose
Sikkim Manipal Institute of Technology, Majitar-Rangpo, Sikkim-737136

ABSTRACT

With the advancement in the video compression and internet technology, the application of video streaming has increased. Because of different types of devices used for accessing the video application and the heterogeneous nature of the network the scalable video coding has become important in order to fulfill the needs of the user. Discrete wavelet transform (DWT) is best tool for scalable video coding because of its property of multiresolution. The coding and display order of pictures in H.264/MPEG4-AVC is completely decoupled and any picture can be marked as reference picture and is used for prediction of following picture independent of corresponding slice type. A wavelet based hierarchical coding structure has been proposed which selects best matched reference frame for the current frame to be encoded based on the mean square error (MSE) in the group of picture(GOP). The scalable encoding of hierarchical prediction has also been exploited. This algorithm has been compared with the performance of traditional multiresolution motion estimation technique based on peak signal to noise ratio. The proposed algorithm show an improvement of 13% in PSNR value over the traditional MRME technique.

Keywords

Discrete wavelet transform, video compression, hierarchical coding, scalable, PSNR.

1. INTRODUCTION

There is an advancement of video coding technology and standardization along with the rapid improvement of network infrastructure, storage capacity and computing power[1]. These developments are enabling increasing number of video application. Particularly, the streaming of video over internet has increased drastically in recent years. In response to the increasing demand on streaming video application over the internet, the coding objective for streaming video has changed to optimize the video quality for wide range of bit rates [2]. In streaming video application, the server normally has to serve a large amount of users with different screen resolutions and network bandwidth. Scalability has been a goal of video compression technology for addressing such a dynamic environment. Currently due to the adamant of different terminals available, the content provider would try to best serve as many devices as possible by generating scalable video coding stream with optimized decoding points for target video [2]. Discrete wavelet transform (DWT) has received considerable attention in the field of image and video processing because of it flexibility in representing nonstationary signals and its ability in adapting to human visual characteristics [3]. The DWT decomposes a nonstationary signal into a set of multiresolution wavelet coefficients which provide a natural means for exploiting their use in multiresolution motion estimation and scalable video coding. In a multiresolution motion estimation (MRME) technique, the motion estimation process is carried out in the wavelet-transform domain. It starts with estimating the motion vectors at the lowest resolution level, where most of the image energy resides. The motion vectors thus obtained are then used as predictions for higher resolutions, where they are either accepted as final motion vectors or further refined [4]. The low resolution output can be obtained by using only the approximation subimage at the decoder. The detailed subimages can be added to this approximated reconstructed subimages thus obtaining high resolution reconstruction video signal.

The paper has been organized as follows. In Section 2, the basic concepts of Scalable video coding and multiresolution representation of video signal using DWT has been summarized. Section 3 explains the architecture of proposed algorithm. Finally in Section 4, experimental results and discussion has been presented

2. BASICS OF SCALABLE VIDEO AND MRME TECHNIUQES.

2.1 Scalable video coding

A video is called scalable when parts of it can be extracted in such a way that the resulting sub-stream forms another decodable bit stream for the decoder, which represents the source content in a reduced reconstruction quality compared to the original bit stream [1]. The coding technique which helps in achieving the scalability in a video stream is called scalable video coding. Basically there are three basic type of scalability- temporal, spatial and quality. Temporally scalable bit stream represent the source content in reduced frame rate (temporal resolution) while those of a scalable bit stream is a reduced picture size (spatial resolution). With quality scalability, the substream provides the same spatial and temporal resolution as the global bit stream but in a lower fidelity or signal to noise ratio (SNR). For this reason, the quality scalability is often referred to as SNR scalability as well.

Individual scalability can be embedded in the bit stream but generally the combination of these scalabilities are used to provide multi-representation with different spatial –temporal resolution and bit rate balancing decoder complexity and coding efficiency [2].

In SVC, the layer scheme is used to enable a full spatio-temporal and quality scalable codec. To adapt the data size to the changes in the bit rate, a unit in video such as a frame or macroblock is divided into small items. A measure of the number of such items comprising a unit is called granularity [5]. The first item of each unit contains the basic and coarsest part of data and the remaining items contain refinement to the base item. This scheme of gradual refining /increasing the granularity of a unit is called fine granular scalability (FGS). A gradual increase in the frame size, bit rate or frame rate is achieved through adapting the granularity of a stream to the required bit rate. The scalable video coding is an extension of widely adopted H.264/MPEG-4 Advance Video Coding, which extend the feature of its base specification by flexible scalability features in all directions (temporal, spatial and SNR) while maintaining high compression efficiency [1].

2.2 Discrete wavelet transform and Multiresolution Representation

The discrete wavelet transform has received considerable attention in the field of image and video processing. This is because of its flexibility in representing nonstationary signals and its ability in adapting to human visual characteristics [6]. It perfectly provides a multiresolution representation of a signal with localization in both time and frequency which is desired property in image and video coding [4]. A wavelet transform corresponds to two sets of analysis/synthesis digital filters, a high pass filter and low-pass filter which deco-relate the pixel values in video frame and result in frequency and spatial orientation separation. The two dimensional discrete wavelet transform of image f(x,y) with resolution depth M can be represented as a sequence of subimages

$$\{S_M f, [W_M^j f]_{j=1,2,3} \dots . [W_1^j f]_{j=1,2,3} - (1)$$

The sequence $\{S_m f : m = 1 \dots . M\}$ represent the approximation of a given video frame at different resolution and $[W_M^j f]_{j=1,2,3}$ represent three detail subimages at various orientation known as horizontal, vertical and diagonal subimages. The Fig. 1 depicts pyramid structure of 3 level wavelet decomposition of an image f(x,y). The pyramid consists of a total of 10 subimages with three subimages at each layer and one low pass subimages at the top. Each subimage of frames also represents the global motion structure of the particular video signal at different scales. The motion activities for particular subimages at different resolution may be different but are highly correlated. Therefore, the hierarchical search for the motion activities can be easily implemented. In this paper, the multiresolution motion estimation technique same as [8] has been used. The inverse wavelet transform is calculated in the reverse manner i.e. starting from the lowest resolution subimage, the higher resolution images are calculated recursively.

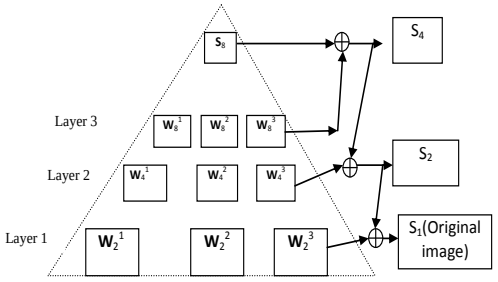

Fig 1: Pyramid Structure of 3 level decomposition of an image.

2.3 MRME Technique

In this section the various MRME techniques used for improving the selection process of motion vectors has been presented. In MRME technique, the motion vectors of the low resolution (low frequency band) subbands are rescaled and used as motion vectors for the corresponding high resolution subbands. Therefore, the motion vectors in the low frequency band must be estimated correctly in order to control the propagation of the error to the high frequency subbands which may ultimately leads to error in the reconstruction of video signal.

In [6], a Median Filtering Multiresolution Motion Estimation (MF-MRME) method was proposed to predict the motion vectors across the wavelet subbands of the video frame in order to overcome the problem of false MVs propagation. In this method, in order to predict the motion vector of a given block at a given resolution, firstly its parent block corresponding to the parent motion vector is identified. Then a parent window which consists of a set of motion vectors centered on this parent block is selected. The motion vector corresponding to the block in the parent window constitutes a candidate motion vector array. The x and y component of the MVs in the initial MV array are median filtered separately, and these two results are combined and appropriately scaled to form a trial child motion vector i.e. IMV, which is used as the starting point for further refinement to obtain FMV.

In [7] interpolation technique has been used. In block based motion estimation it is assumed that every block have an integer displacement which is in reality, not true. Therefore, to improve the motion estimation and to increase the accuracy of the prediction, a sub-pixel technique with a bilinear interpolation process can be used. This is done by interposing a line between each two lines and a column between each two columns of the image. Then ME, is applied to the new image. With this technique, a motion vector can point in a half or quarter of pixel location at a fraction of pixels will be better predicted.

A MRME with indexing has been proposed in [8]. Here sum of absolute difference (SAD) and sum of absolute values of the amplitudes (SAVA) has been used for selection of correct motion vectors. Firstly the SAD is calculated by using expression given below-

$$SAD(m,n) = \sum_{k=0}^{I-1}\sum_{l=0}^{J-1} |W_c(i+k, j+l) - W_r(i+k+m, j+l+n)| \dots (2)$$

Here I x J is the block size, $W_c(i+k, j+l)$ and $W_r(i+k, j+l)$ represent the wavelet coefficient at the location of the current and reference subbands respectively, m and n represent the relative displacements in the horizontal and vertical directions. To impose a selection criteria, SAVA of all of the wavelet coefficients within the current block is calculate as follows

$$SAVA(m,n) = \sum_{k=0}^{I-1}\sum_{l=0}^{J-1} |W_c(i+k, j+l)| \dots (3)$$

The SAD between the current block and the block pointed by this prediction motion vector in the reference subband is calculated and compared with the SAVA of the current wavelet block and if the former is less than the latter, this prediction motion vector is accepted otherwise, it is discarded and the current wavelet block is labeled as intracoded. Since this algorithm discriminates between good motion predictions hence the algorithm is able to improve the motion estimation performance substantially. However there is a small increase in the computational complexity.

All the above technique of improving the correctness of the motion vector leads to additional computational cost in terms of calculation of threshold value, median, interpolation and sum of absolute value of amplitude. These techniques do not explore the scalable video possibilities of MRME technique. The MRME technique can be used for the scalable video coding since it give the basis for the spatial, temporal and quality scalability.

In contrast to previous video coding standards, the coding and display order of pictures in H.264/MPEG4-AVC is completely decoupled and any picture can be marked as reference picture and is used for prediction of following picture independent of the corresponding slice types [9]. In general, hierarchical prediction structures for enabling temporal scalability can also be combined with the multiple reference picture concepts of H.264/AVC [5].

A typical hierarchical prediction structure with 4 dyadic hierarchy stages is depicted in Fig. 2.

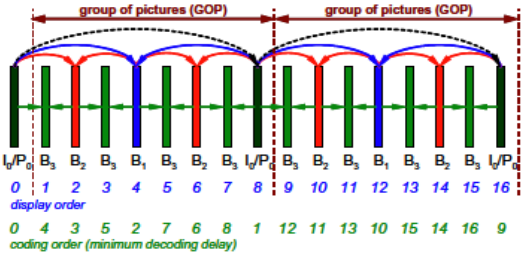

Fig. 2. Hierarchical coding structure with 4 temporal levels.

The first picture of the video sequence is intra-coded and is called the key picture (black in Fig.2). Key pictures are coded in regular interval [9]. A key picture and all pictures that are temporally located between the current key picture and previous key picture are considered to build a group of pictures (GOP). The key pictures are either intra-coded or inter-coded using previous (key) pictures as reference for

motion compensation and prediction (MCP). The remaining pictures of a GOP are hierarchically predicted as illustrated in Fig.2 and coded using the bi-predictive (B) slice syntax of H.264/MPEG4-AVC.

The computations complexity is high in this technique because of two factors. Firstly, the coding order has to be chosen in such a way that the reference picture is always coded before they can be applied for MCP. Hence an order procedure has to be followed for minimal decoding delay. Moreover, multiple reference system for the prediction of current picture is required which leads to the increase in computational complexity. Secondly, since multiple reference pictures are used as reference picture, all the reference pictures has to be stored in the buffer for their future use. This leads to additional buffer requirement.

Therefore, motivated by this observation, in this paper the possibility of scalable video coding using MRME technique has been explored. A new hierarchical coding structure algorithm based on the best matched reference frame has been proposed.

3. ARCHITECTURE OF HIERARCHICAL CODING STRUCTRE FOR SCALABLE VIDEO CODING

A high level block diagram of the hierarchical coding structure using wavelet is shown in fig.3. The encoder module consists of Wavelet filter decomposition and the motion estimation and compensation and variable length coding. The video signal decomposition is achieved by using the appropriate wavelet filter. The decomposed frames (all the approximate and the detail frames) are stored in the frame buffer for further processing and transmission. Multiresolution motion estimation technique used in [8] for the motion estimation and compensation has been used in our proposed method. If a video frame is decomposed up to three levels, resulting in a total of ten subimages with three subimages at each of the first two levels and four at the top level, including the subimages S_8 (Fig:1) which represents the lowest frequency band. S_8 contains a major percentage of the total energy present in the original frame, though it is only 1/64 of its size. The motion vectors are calculated for all four lowest resolution frequency bands, i.e., S_8 and $\{W_8^j : i = 1,2,3\}$. The motion vectors of $\{W_8^j : i = 1,2,3\}$ are appropriately scaled and used as motion vectors for all corresponding higher-resolution subbands [4].

In general for three level decomposition, if $V_8^j(x,y), j = 1,2,3$ represent the motion vector of the third level coarse subimages then the motion vectors of its corresponding higher resolution subimages $W_4^j : i = 1,2,3$ *and* $W_2^j : i = 1,2,3$ are obtained by rescaling and refining $V_8^j(x,y)$. This technique not only reduces the computational complexity as there is no need to calculate the motion vector for all the subimages but also decrease the number of bits to be transferred to the decoder in order to reconstruct the video signal. From Fig.5, it is clear that low resolution output can be obtained by using only the approximation subimage at the decoder. The detailed subimages can be added to this approximated reconstructed subimages thus obtaining high resolution reconstruction video signal. Similarly the temporal scalability can be obtained by limiting the no of reconstructed frame. The quality scalability can be obtained by using the vector quantization technique.

Instead of adopting various corrective measures as discussed in section 2 we propose an algorithm which select best matched reference frame for the current frame to be encoded based on the mean square error (MSE) in the group of picture. Input video signal is decomposed into sequence of frames. These sequences of images are grouped into group of pictures (GOP). Then the MSE of these images are calculated and are maintained in a matrix. The average value of each column is stored in the array. The frame with the smallest average error is taken as a reference frame.

This algorithm eliminates the storage of multiple reference frames in the buffer and there is no requirement for the predictions of a current frame from the multiple frames. Since best match frame is used as the reference frame, the quality of the reconstructed image will be as efficient as the hierarchical B prediction. The codec for the proposed algorithm is shown in Fig. 4.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In order to test the effectness of the proposed algorithm, simulation studies have been carried on a video signals namely "Kapildev" of size 128 x 128 with 120 frame sequence. The 7/9 biorthogonal wavelet has been employed in the study as it is one of the best wavelet for image coding [10]. Three level of decomposition of the image sequence has been applied using 7/9 biorthogonal filter. For motion estimation and compensation, the conventional MRME technique used in [8] and the proposed technique has been used. The performance comparison of two techniques has been presented in tabular form.

Mean absolute difference is used as the measure for motion estimation performance. If the size of a video frame is X x Y, the MAD in wavelet domain transform is given by

$$MAD = \frac{1}{XY}\sum_{i=0}^{X-1}\sum_{j=0}^{Y-1}\left[W(i,j) - \widehat{W}(i,j)\right] \dots (4)$$

Where W(i,j) and $\widehat{W}(i,j)$ respectively represent the wavelet coefficient of the original frame and the wavelet coefficients resulting from an MRME Technique. Peak signal to noise ratio (PSNR) is being used to measure the quality of the proposed algorithm. Denoting pixels of the original image by P_i and the pixels of reconstructed signal by Q_i, the mean square error (MSE) between the two images is given by

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(P_i - Q_i)^2 \dots (5)$$

Where is n is the total number of pixels. Now the root mean square error(RMSE) is defined as the square root MSE and PSNR is defined as

$$PSNR = 10log_{10}\frac{max_i|P_i|}{RMSE} \dots . (6)$$

4.1 Results and Discussion

The proposed algorithm includes automatic determination of the reference frame which is the frame with the least mean square error (MSE) in comparison with all the other frames in a Group of Pictures (GOP). Table 1. shows the MSE calculated for five sequences of frames (say for example a GOP with five frames) from video signal "Kapildev". The first frame is intracoded and for selecting the reference frame for the other frames in the sequence their corresponding MSE value with respect to the other frames are calculated. The MSE is maintained in a matrix. The frame with the smallest error average value is taken as a reference frame.

In Table 1., best average value is 69.7909 (for frame C) therefore in this case frame C will be the reference frames for all the remaining frames in the GOP. The peak signal to noise ratio obtained by using the proposed algorithm is shown in Table 2.

Table 3 shows the PSNR for the same video signal using the traditional MRME technique. The various frames reconstructed using the proposed algorithm is shown in Fig. 5.

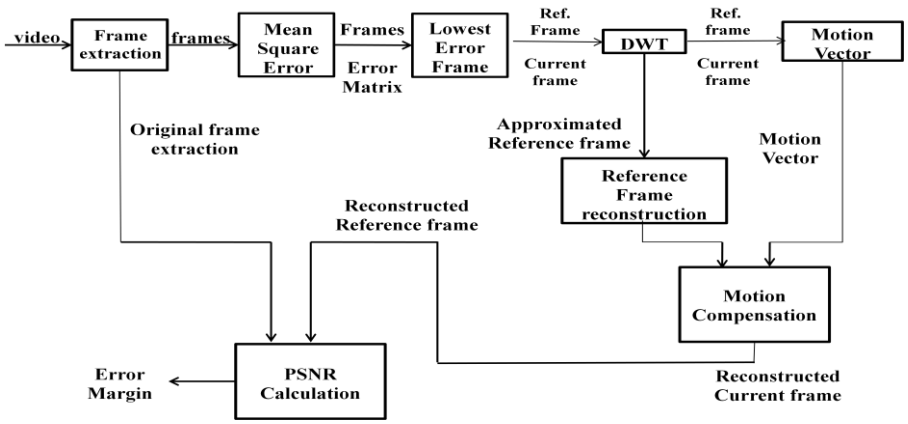

Fig. 4: Codec for automatically selecting the best reference frame.

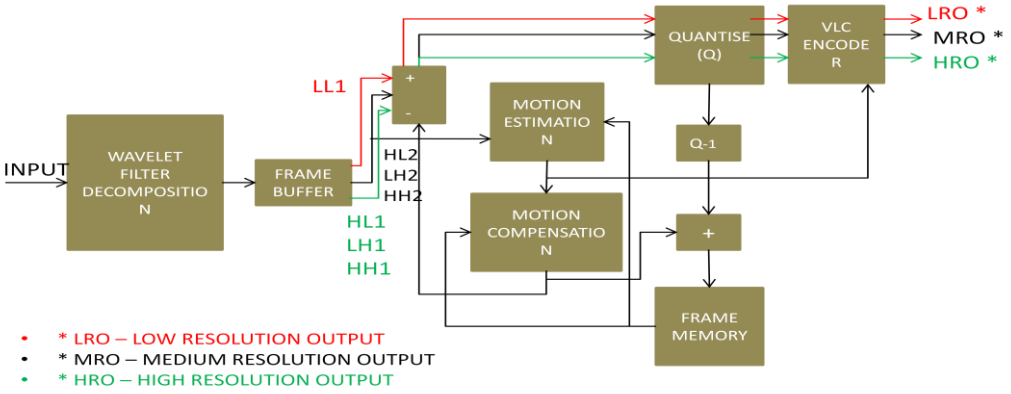

Fig 5:Architecture of Hierarchical coding structure for scalable video coding

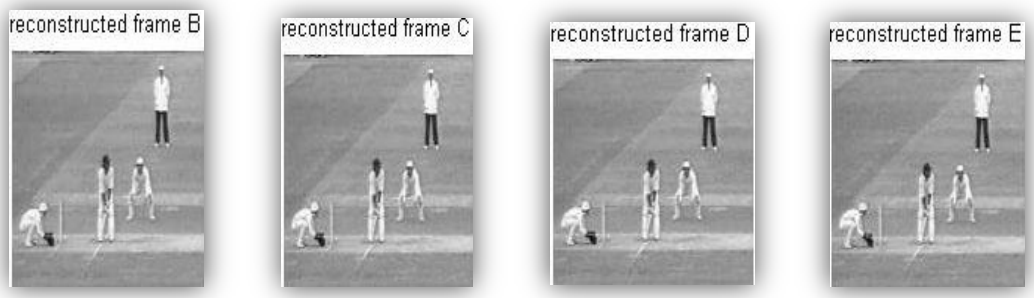

Fig. 6: Reconstructed Video frames.

Table 1. MSE for selection of best reference frame.

	A	B	C	D	E
A	0	43.406	44.736	172.643	219.202
B	43.406	0	3.644	123.638	181.496
C	44.736	3.664	0	121.539	179.035
D	172.643	123.638	121.539	0	84.564
E	219.202	179.035	179.035	84.565	0
Avg	95.997	70.457	69.791	100.477	132.879

Table 2. Calculation of PSNR for the reconstructed frame using the proposed algorithm.

Frame	PSNR with ref. frame C
A	38.90
B	39.35
C	39.11
D	39.24
E	39.26
Total PSNR	195.86
Average PSNR	39.172

Table 3. Calculation of PSNR for the reconstructed frame using the Convectional MRME Technique.

Frame	PSNR with ref. frame A
A(reference frame)	39.78
B	38.74
C	38.87
D	38.92
E	38.91
Total PSNR	195.22
Average PSNR	39.04

Comparing Table 2 and Table 3, there is an improvement of 13% in PSNR value of the proposed algorithm.

5. CONCLUSION

In this paper the MRME technique has been exploited for its use in scalable video coding. Instead of using multiple frames as the reference frame, the new algorithm based on the best match criteria among a group of pictures has been proposed. The new algorithm not only reduces the computational complexity of multiple reference frames, it also reduces the buffer requirement for the storage of reference frame. The mean square error has been used as the criteria for selection of reference frame. The proposed algorithm shows an improvement of 13% in PSNR of the reconstructed video signal over the conventional MRME technique.

6. REFERENCES

[1] Heike Schwarz, Detlev Marpe, Thomas Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard" IEEE Transactions on Circuits and Systems for Video Technology. Vol. 17, No.9, September 2007.

[2] Peixoto Eduardo, Toni Zgaljic, Ebroul Izquierdo, "Transcoding from Hybrid Nonscalable to Wavelet Based Scalable Video Codecs", IEEE Transactions on Circuits and Systems for Video Technology. Vol.22 No.4 April 2012.

[3] Hyun-Wood, Park, Hyung-Sun Kim, "Motion Estimation Using Low-Band-Shift Method for Wavelet-based Moving Picture Coding", IEEE Transactions on Image processing, Vol.9 No.4, April 2000.

[4] Sohail Zafar, Ya-Zin Zhang, "Multiscale Video Representation using Multiresolution Motion Compensation and Wavelet Decomposition", IEEE Journal on Selected areas in Communication. Vol. 11, January 1993.

[5] Li Weiping, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard", I*EEE* Transactions on Circuits and Systems for Video Technology, Vol. 11, No.3, 2001.

[6] Jinwen Zan, M. Omair Ahmad, M.N.S Swamy, " Wavelet Based Multiresolution Motion estimation through Median filtering", Acoustics, Speech, and Signal Processing, Vol.4, 2002.

[7] Najib Ben Aoun, Maher EL'ARBI, Choki Ben Amar, "Multiresolution motion estimation and compensation for video coding" Signal Processing (ICSP), 2010.

[8] Jinwen Zan, M. Omair Ahmad, M.N.S Swami, "A Multiresolution Motion Estimation Technique with indexing", IEEE Transactions on Circuits and System for Video Technology, Vol. 16, No. 2, 2006.

[9] Heiko Schwarz, Detlev Marpe, Thomas Wiegand, "Analysis of Hierarchical B Pictures and MCTF", Multimedia and Expo, 2006.

[10] Jinwen Zan, M. Omair Ahmad, M.N.S Swamy, "Comparision of Wavelets for Multiresolution Motion Estimation. IEEE Transactions on Circuits and System for Video Technology. Vol.16, No.3, March 2006.