International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Correlation Based Low Complex Video Frame Interpolation

Brahmadesam T. Madav , S. A. K. Jilani, S. Aruna Mastani

Abstract: ***Frame Interpolation is one of the main stages in video processing. Video coding standards skip some in-between frames for efficient compression and coding. At decoder the common approach to reconstruct the skipped frame using Motion Compensated Frame Interpolation (MCFI)methods. In MCFI, computational complexity is very high as calculation of Block Matching Algorithm, Motion Vectors (MV), Motion Estimation(ME) and Prediction logic of objects in motion between the frames, increases the complexity in MCFI method. A more feasible approach with minimum computational complexity using block level correlation is proposed in this paper. Errored MV at the decoder results in holes, occlusions, blurring and edge deformations in the interpolated frame. This proposal minimizes afore mentioned effects along with minimum complexity. The results are simulated in terms of peak-signal-to-noise-ratio (PSNR) and structural similarity index (SSIM).***

Keywords: Block Correlation, Frame Interpolation, MCFI, PSNR, SSIM

I. INTRODUCTION

Both in academia and consumer electronics, video Frame Interpolation (FI) has drawn greater attention for the last decade. To reduce jerkiness or remove blurriness at the receiver FI is needed between two video frames[1]. Due to increase in resolution of video frame capturing devices, various compression standards strive for efficient transmission. In bandwidth limited environment the video coding standards like AVC/HEVC [8] have opted to decrease frame rate by skipping in-between frames. At receiver decoder must interpolate the skipped frames for the receiver. In literature, many FI algorithms have been proposed [2]-[5]. These algorithms are broadly classified as Motion Compensated Frame Interpolation (MCFI) and non-MCFI algorithms. In MCFI, Motion Estimation (ME) algorithms and Motion Vector (MV) are used to obtain true motion using Block Matching Algorithm (BMA). If ME is not possible at the decoder, then MV processing techniques[6],[7] are used to obtain smoother motion. The MVs are identified as inliers and outliers based on smooth ME. Various algorithms are proposed to remove outliers.

There are many popular MCFI methods at block level such as BMA, Bidirectional ME and Optical Flow FI. These methods concentrate on generating MVs and utilizing them in FI which is a complex process. If MVs are lost or considered as outliers, hole appear in the interpolated frame. Filling up these holes is again another complex process[6] . If more than one MV points a single point in motion, selecting correct MV using methods like Vector Median Filtering (VMF), MV Variance is additional process. Also determining the reliability of MV is needed.

In this paper, we exploit the correlation at block level and blending at pixel level with minimum complexity. This proposed scheme effectively overcomes the complexity of MCFI algorithms, jerkiness and blurriness of non-MCFI algorithms.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Brahmadesam T Madav*, ECE department, MITS, JNTUA University, Madanapalle, India. Email: madhavbt@gmail.com

Dr. S.A.K Jilani, Research Supervisor,ECEdepartment, MITS Madanapalle, India. Email: drsakjilani@mits.ac.in

Dr. S. Aruna Mastani, ECE department, JNTUA, Ananthapuramu, India. Email: aruna_mastani@yahoo.com

Authors thank Madanapalle Institute of Technology and Science, Madanapalle for Laboratory and Financial support extended for this work.

II. BLOCKS AND CORRELATION

A. Frame Rate

In general video cameras have different frame rates like 24 fps, 25 fps, 30 fps, 60 fps, 120 fps and 240 fps based on subject or specific scene. A frame is considered as a picture or image from a group of pictures of a video sequence. Frame rate is defined as number of pictures taken per second by the camera. Let us understand clearly. If the camera captures 30 frames and we need 60 fps, then each frame is duplicated by frame repetition. If frame repetition is not done the video appears in fast moving objects. Let us consider another case. If camera captures 60 frames and the device operates at 30 fps, then alternate frames are dropped to accommodate 30 frames resulting in slow motion of the objects. If camera captures number of frames as needed fps, then it is known as “Project” or “Base” frame rate. Hence generally required frame rate from camera or recording is chosen greater than project frame rate.

In order to avoid motion blur, the ratio of recording frame rate to project frame rate must be even. Else the reproduction of video sequence appears that some frames are missing resulting in jerkiness. Thus, base frame rate is the highest quality frame rate of given camera rate. For human perception 24, 25 or 30 fps are best frame rate. When fps is greater or less than these rates, it’s a compromise between camera fps and resolution. If camera fps is greater than these rates, then camera is using a greater number of bits per second over a few frames and is loss of memory space. If camera fps is less than these rates, then camera is using greater number of bits per second for more frames than reproduction resulting more artifacts appear due to compression.

B. Mathematical Description

In our paper, we denote a high resolution video sequence as $\mathbf{S}(x,y,t) \in R^{m \times n \times k}$ where *x, y* are spatial coordinates, and *t* is temporal coordinate, such that $1 \leq x \leq m$, $1 \leq y \leq n$ and $1 \leq t \leq k$. The $R^{m \times n \times k}$ forms 3D space-time volume of *m* rows, *n* columns and *k* frames.

As per video coding standards, at *t=1,3,5, 7...,2k+1 i.e.*, odd numbered frames are received at the decoder while even numbered frames are to be interpolated. The input to the decoder being temporally down sampled video sequence is $\mathbf{X}(x, y, t) \in R^{m \times n \times k}$. the aim of FI is to estimate **S** from **X**.

C. Blocks

The frame resolution is *m* x *n*. Let a block size be *p* x *q*, then we get (*m* x *n*)/ (*p* x *q*) number of non-overlapping blocks per frame. If the block size is *'p'* pixel in *(x, y)* coordinates the number of blocks per frame are (*m* x *n*)/ p^2.

Each frame is divided into non-overlapping blocks of sizes 16 x 16, 32 x 32, 64 x 64, 128 x 128, 256 x 256, 512 x 512. The 256 x 256 and 512 x 512 block sizes are applied based on frame resolution [2]. The 256 x 256 block size is possible in Foreman video sequence as the resolution is 352 x 288. We get various number of blocks based on block size as shown in Table 1.

D. Correlation

Correlation is used in our proposed method to find similarity measure between the blocks. Corresponding blocks in previous and next frame are correlated. The correlation operation is simple, easy to implement and powerful operation that brings out similarity measure because of linearity and shift-invariance properties. Correlation is applied on every pixel in the corresponding blocks. Correlation will be high when the blocks are perfectly matched and will be low when blocks are mismatched. If the pixel intensity value is high, the correlation will also be high independent of pixel matching nature. This is disadvantage of simple correlation. Hence, we used normalized correlation given by

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{t-1}(i,j) f_{t+1}(i,j)}{\sqrt{\sum_{i=1}^{m} f_{t-1}(i)} \sqrt{\sum_{j=1}^{m} f_{t+1}(j)}} \quad (1)$$

Table 1: Block Sizes

Block Size	No. of Block Columns	No. of Block Rows	Total no. of blocks
16 x 16 Pixels	22	18	396
32 x 32 Pixels	11	9	99
64 x 64 Pixels	8	6	48
128 x 128 Pixels	4	3	12
256 x 256 Pixels	2	2	4

In our proposed correlation process, the similarity measure ranges between +1 and -1 indicating highly correlated and poorly correlated blocks respectively. All the block correlated values are stored. As a next step in our proposed method, the median amongst the correlated values is generated. If the number of correlated values is odd, after rearranging in sequential manner, the median amongst these correlated values is designated threshold. If

the number of correlated values is even, after rearranging in sequential manner, the average of two correlated values centered in the sequence is the median value and is designated as threshold.

III. VIDEO QUALITY METRICS

Video Quality Measures are of Objective and Subjective types. In this paper we considered objective type criterion which gives the measure of difference between the original and the reconstructed or processed.

A. Mean Squared Error(MSE)

MSE is the simple and basic quality measure which is given as

$$MSE = \frac{1}{N}\sum_{x,y}\sum_{t}[f_1(x,y,t_1) - f_2(x,y,t_2)]^2 \quad (2)$$

Where 'N' is number of pixels per frame

'x' is row dimension of frame, 'y' is column dimension of frame and 't' is time or temporal dimension.

For each color component MSE is computed separately.

B. Peak Signal to Noise Ratio (PSNR)

PSNR is one of the benchmarks for performance evaluation of objective video quality metrics. It is dependent on estimation of spatial alignment, temporal alignment, gain and level offset between interpolated frame and original frame.

$$PSNR = 10log_{10}\left[\frac{(Max.\ peak\ intensity\ value\ of\ video\ signal)^2}{Mean\ Squared\ Error}\right] \quad (3)$$

C. Structural Similarity Index Metric (SSIM)

In human beings, color sensation is attributed by Luminance and Chrominance. Chrominance is attributed by Hue and Saturation. Hue is defined as color tone which is dependent on peak wavelength of the light. Saturation is defined as purity of color which is dependent on bandwidth of light spectrum.

The structural information of an image is modeled by SSIM quality metric. The structural information change in SSIM defines the image degradation. Luminance, Contrast and Structure comparison are the steps in similarity measurement. Symmetry, Boundedness and Unique maximum are properties of SSIM. SSIM is defined as

$$SSIM(x,y) = f(l(x,y).c(x,y).s(x,y)) \quad (4)$$

where *l(x,y)* is the luminance at (x,y) location, *c(x,y)* is the contrast at (x,y) location and *s(x,y)* is the structural comparison at (x,y) location.

IV. EXPERIMENTAL RESULTS AN DISCUSSION

To evaluate the performance of proposed method, extensive experiments are conducted on Foreman Video sequence in CIF resolution (352 x 288) at 30 fps. Consecutive first 50 frames are used for experiment. All even frames amongst 50 frames are discarded for generality and are interpolated. The H.265 (HEVC) / H.264 (AVC) video coding standards generate bit streams at 15 fps[8]. Hence even numbered frames are discarded. All programs are executed on Intel Core i5 CPU @ 2.50GHz, 8GB Memory, 64-Bit Windows 10 Pro Operating System, in MATLAB R2015b version.

A. Step by step processing

Second frame of Foreman video sequence is used as reference for interpolation in this proposed scheme. In order to estimate and interpolate second frame, first and third frames also called as previous and next frame respectively are used. The previous and next frames are divided into non-overlapped blocks as first stage of FI[6]. These blocks range from 16 x 16 to 256 x 256-pixel size. The block structure of size 64 x 64 size of frame 1 is shown in the figure1. Identified that the blocks of face and collar had variations in 1st and 3rd frames. The other blocks had little variations. Cross correlation between corresponding blocks shown in figure 2.

Corresponding blocks of 1st and 3rd frames which are in marked square are less correlated than the other blocks. The blocks which contain the high object motion, are less correlated than the blocks with low object motion. In this Fig.2, the less correlated blocks are around face and neck of the Foreman.

B. Thresholding and Blending

The median value of the correlation values is chosen as the threshold for next pixel level stage operation. The cross-correlation values that are less than threshold (LTT) are used in this stage. Blocks in LTT range in 1st and 3rd frames are combined by blending corresponding pixel intensities. For blending of blocks with correlation LTT, blending object that combines two images is used. The blocks are overlayed one on other highlighting the selected pixels. These blended blocks of previous and next frames are corresponding blocks in to be interpolated frame. The blocks in 1st and 3rd frames with LTT are identified with red line around the block as shown in figure2. The other blocks have cross correlation greater than threshold (GTT). These blocks are in background, building, etc. Thus, blended blocks along with non-blended blocks are concatenated to reconstruct the interpolated frame. Fig.7 shows the cross correlation between 1st and 3rd frames with different blocks sizes. The correlation increased with increase in block size as the motion of the object i.e., Foreman's head and neck predominates when the block size was very small. As the block size increases the predominance decreases. Similar interpretation is done between 13th and 15th frames as shown in Fig.8.

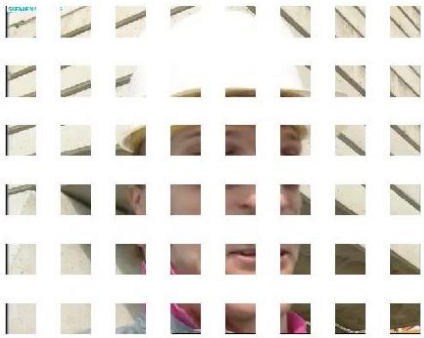

Figure 1: 64 x 64 Pixel blocks

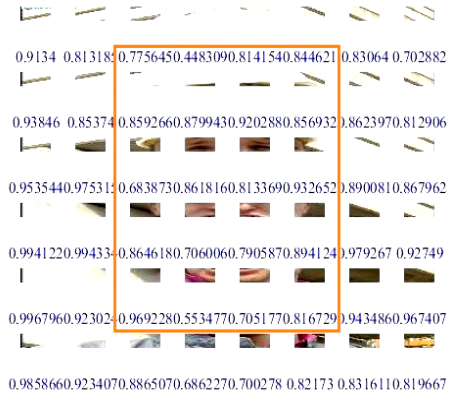

Figure 2: Block Correlation Values

Figure 3: Original & Reconstructed Frame2

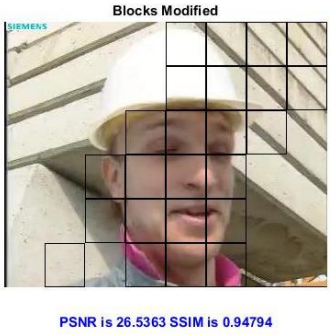

Figure 4: Blocks Interpolated

Figure 5: Original & Reconstructed Frame14

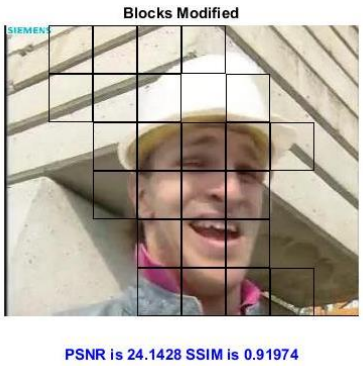

Figure 6: Blocks Interpolated

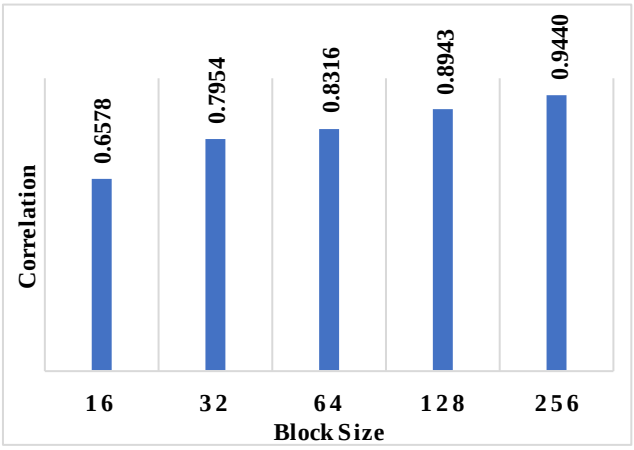

Figure 7: Average Cross Correlation of 1 and 3 Frames

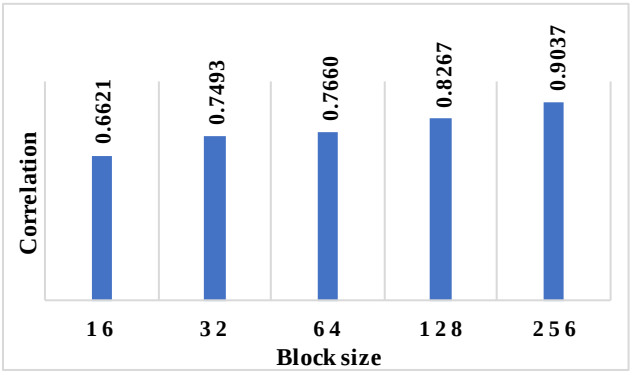

Figure 8: Average Cross Correlation of 13 and 15 Frames

C. Concatenation

As shown in Fig.2, in our proposed method, the blocks in the squared box are blended at pixel level. The blended blocks can be observed in Fig.4 and Fig.6. the blending at pixel level is done by using system object in MATLAB. This resulted in smooth blending without blurriness in the motion object and no ghost effects at the edges which were normally visible in many of the MCFI schemes [7] with low complexity.

In Fig.3, the original frame which was to be interpolated is show first as, original frame. The next is the interpolated frame by our proposed method. Smoothness can be observed throughout the frame. Similar conclusions can be drawn from Fig.5. Our proposed method was experimented on first 50 frames. The average correlation of 16 to 256-pixel block sizes for all the odd numbered frames are shown in Fig.9. The 5^{th} and 7^{th} frame have high correlation. Also frames 19 & 21, 27 &29, 29 & 31, 43 & 45 and 45 & 47 had very high correlation when compared with other pair of frames and correlation between 27 & 29 been the highest. The correlation between the 23 & 25 is the lowest. The lowest correlation resulted because of high motion content in the frames.

D. Statistical Analysis

The PSNR and SSIM are the popular quality metrics used to measure closeness of the interpolated frame with original frame. In our proposed method for 50 frames we achieved maximum average PSNR of 29.5898 and SSIM of 0.9638 at 32- and 64-pixel block sizes as shown in Fig.10.

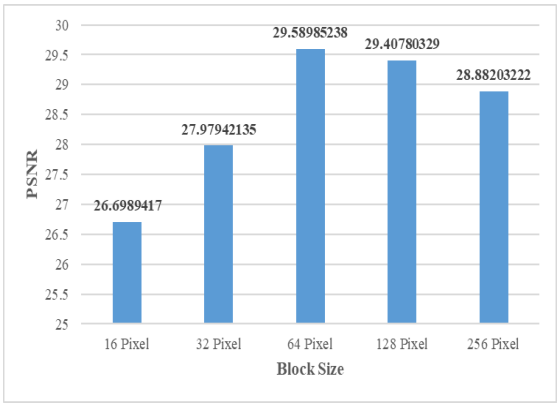

Figure 9: PSNR from Original and Interpolated Frame2

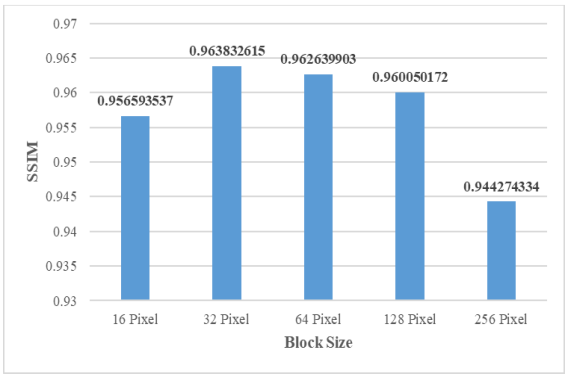

Figure 10: SSIM from Original and Interpolated Frame2

V. CONCLUSIONS

In this paper a novel block correlation-based Frame Interpolation technique is implemented. In MCFI the ME using BMA is a complex process. MVs are to be predicted using search window. At decoder these MVs are to be estimated. Finding the correct MV is a complex uncertain task. Correlation between previous and next frames have been proposed for various block sizes and has been used for FI. The 256-pixel block size is best correlated. The interpolated frame has the best similarity when the block size is of 64-pixels. This technique has the reduction in computational complexity and edge deformation. With respect to MCFI, correlation proves to be low complex. Therefore, according to the obtained results, the complexity of MCFI can be reduced.

REFERENCES

1. A.M.Tekalp, “Digital Video Processing”, Englewood Cliffs, NJ,USA, Printice-Hall,1995
2. S.Fujiwara and A.Taguchi, “Motion-compensated frame rate up-conversion version based on block matching algorithm with multi-size blocks,” *Proc.ISPACS*,Hong Kong, Dec.2005, pp. 353–356.
3. A.M.Huang and T,Nguyen, “A novel motion compensated frame interpolation based on block merging and residual energy,”in *Proc.Multimedia Signal Processing Workshop,* Victoria,BC,Canada,Sep,2006,vol.4,pp.353-356.
4. A.M.Huang and T,Nguyen, “A multistage motion vector processing method for motion-compensated frame interpolation,”*IEEE Trans. Image Process*.,vol.17,no.5,pp.694-708,May 2008
5. A.M.Huang and T,Nguyen, “Correlation-based motion vector procesing with adaptive interpolation scheme for motion compensated frame interpolation,” *IEEE Trasn. Image Processing*,vol.18,no.4,pp.740-752,Apr.2009.

6. Z.Yu,H.Li,Z.Wang,Z.Hu and C.W.Chen, "Mutli-level video frame interpolation: exploiting the interaction among different levels,"*IEEE Trans. on Circuits and Systems for Video Technology*.,vol.23,no.7,Jul.2013.
7. M.Ebdelli,O.Le Meur and C.Guillemot, "Video inpainting with short-trem windows: application to object removal and error concealment," *IEEE Trans. on Image Processing*, vol.24,no.10,Oct.2015.
8. H.265(02/2018),ITU-T, Telecommunication Standardization Sector of ITU, Series H:Audiovisual and Multimedia Systems, 2018.

AUTHORS PROFILE

B.T.Madav is pursuing Full-time research, Ph.D in JNTUA, Ananthapuramu at MITS, Madanapalle, India as research centre. He is graduated from IETE(117384), New Delhi, Post-graduated from JNTUK, Kakinada, India. He has industrial experience for 7 years and teaching experience of 13 years. His areas of interest include, Image Processing, Video Processing and Computer Vision.

S.A.K Jilani is currently Professor in Electronics and Communication Engineering (ECE), Madanapalle Institute of Technology and Science(MITS), Madanapalle, India. He received his Ph.D degree from Sri Krishnadevaraya University, Anantapur, India in 2002. His areas of interest are Image and Video Processing, Embedded Systems, Material Science. He is currently Principal Investigator of a DST project on Cognitive Science. He contributed many research papers to International, National Journals, Conferences and authored textbooks.

S. Aruna Mastani is working as Assistant Professor in JNTU College of Engineering, Anantapuramu. She received B.E in Electronics and Communication Engineering from JNTU College of Engineering Anantapur and M.Tech in Digital Systems and Computer Electronics from JNTU College of Engineering, Anantapur. She received Ph.D degree in the area of Digital Image Processing from JNTUA, Anantapur. She worked as academic assistant in JNTUCE, Anantapur and worked as Assistant Professor and Associate Professor in Intell Engineering College Anantapur, She Published several papers in national and international journals and conferences in Image Processing, Signal Processing, and VLSI