3D VIDEO FRAME INTERPOLATION VIA ADAPTIVE HYBRID MOTION ESTIMATION AND COMPENSATION

Xiaohui Yang[1,2]*, Zhiquan Feng*[1,2]

[1]School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022,China

ABSTRACT

This paper proposes a novel motion compensated frame interpolation (MCFI) method based on adaptive hybrid motion estimation and compensation (AHMEC) for 3D video. In our work, we deal with the problem of ghost artifacts around the foreground object boundaries by considering motion and depth information jointly. First, the motion vector field (MVF) of the interpolated frame is estimated using block-based method. We use depth and motion information to distinguish the occlusion areas in the interpolated frame. Then, an adaptive pixel-based motion estimation (ME) method is applied to detail the MVF in the covering and uncovering areas. Simulation results show that the proposed MCFI algorithm outperforms the conventional algorithms in terms of objective and subjective performances.

Index Terms— Motion compensated frame interpolation (MCFI), frame rate up-conversion (FRUC), 3D video, depth map, motion vector processing

1. INTRODUCTION

Motion compensated frame interpolation (MCFI), i.e., frame rate up-conversion (FRUC) is one of the fundamental technologies in conventional video processing field [1]. Briefly speaking, MCFI is utilized to increase video frame rate by interpolating extra frames, and hence motion jerkiness will be suppressed at the display and the visual quality will be improved. It has lots of applications, such as video format conversion, video compression and slow motion playback. Recently, 3D video (3DV) processing has received considerable attention because of its capability of providing realistic and immersive visual experience [2]. In 3D TV system the data amount is huge, and it is impractical to transmit the 3DV at high frame rate because of the bandwidth capacity. On the other hand, there are lots of conventional videos which need to be converted into 3DV to enable the compatibility of future 3D TV system, and these conventional videos are usually captured at low frame rate. However, when low frame rate 3DVs are displayed in LCD televisions, noticeable motion blur will be appeared because of its sample-and-hold nature. To reduce the motion blur and improve the visual effect in 3D TV system, MCFI is utilized to up-convert the original 3DV to the required rate.

Block-based algorithms are widely used for motion estimation (ME) as it is simple and easy to implement, and all the pixels within a block share the same motion vector (MV). The drawback is that it cannot provide accurate MVs when the block contains multiple motion layers, i.e., ghost artifacts will appear along foreground object boundaries. There are numerous literatures using the block-based ME method to conduct MCFI [3–6], and the interpolated frame can be obtained unidirectionally or bi-directionally as follows.

$$f_t\left(\mathbf{p}+\frac{\mathbf{v}}{2}\right)=\lambda_{t-1}\cdot f_{t-1}\left(\mathbf{p}+\mathbf{v}\right)+\lambda_{t+1}\cdot f_{t+1}\left(\mathbf{p}\right) \quad (1)$$

$$f_t\left(\mathbf{p}\right)=\lambda_{t-1}\cdot f_{t-1}\left(\mathbf{p}+\frac{\mathbf{v}}{2}\right)+\lambda_{t+1}\cdot f_{t+1}\left(\mathbf{p}-\frac{\mathbf{v}}{2}\right) \quad (2)$$

where f_t, f_{n-1} and f_{n+1} are the interpolated, previous and current frames respectively. $\mathbf{v}$ denotes the MVF, and vector $\mathbf{p}$ is the pixel location. λ_{t-1} and λ_{t+1} are set to 0.5 when f_t is located at the temporal middle of f_{t-1} and f_{t+1}.

In a different approach, Chen *et al.* [7] and Werlberger *et al.* [8] used pixel-based ME, i.e., optical flow, to do MCFI. However, the main issue of these pixel-based ME is the high computation complexity. To deal with the wrong MVs in occlusion areas, [9] proposed a FRUC algorithm based on variational image fusion. However, in order to obtain one interpolated frame, four optical flow based motion vector fields (MVFs) are needed, moreover, there is also a variational image fusion process, so the complexity is extremely high.

Considering the merits of block-based and pixel-based MCFI methods, we proposed a MCFI method for 3DV in this paper. First, the MVF of the interpolated frame is estimated using block-based method. After that, we use the depth and motion information to distinguish the occlusion blocks in the interpolated frame. An adaptive pixel-based ME method is utilized to refine the MVF in the occlusion and disocclusion areas. Finally, the interpolated frame is compensated adaptively based on the pixel classification.

The rest of this paper is organized as follows. In Section 2, the details of adaptive hybrid motion estimation and compensation (AHMEC) is introduced. Section 3 presents the simu-

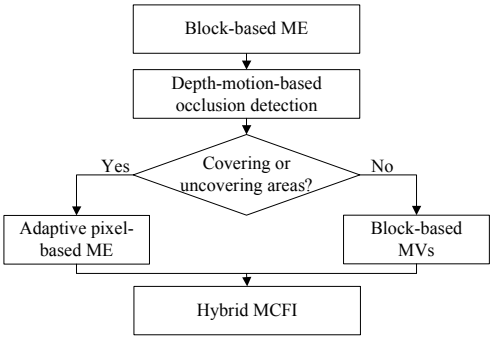

Fig. 1. Procedure of the proposed 3DV MCFI method.

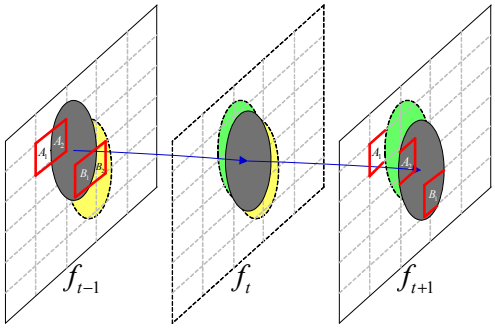

Fig. 2. Illustration of covering and uncovering areas.

lation results and discussion. Finally, the paper is concluded in Section 4.

2. THE PROPOSED 3DV MCFI METHOD

The procedure of our method is shown in Fig. 1. Details of our method are described in the following subsections.

2.1. Block-based ME

In our method, bi-directional ME is utilized to obtain the true motion vector (TMV). First, the interpolated frame is divided into non-overlapping blocks with the same size, and then TMV can be found by searching for the most similar blocks in the previous and current reference frames. The matching cost is calculated by the sum of the absolute differences (SAD) as follows.

$$\mathrm{SAD}\left(\mathbf{c}\right) = \sum_{\mathbf{p}\in B_{ij}} \left| f_{t-1}\left(\mathbf{p}+\mathbf{c}\right) - f_{t+1}\left(\mathbf{p}-\mathbf{c}\right)\right| \quad (3)$$

and,

$$\mathbf{v}' = \arg\min_{\mathbf{c}\in s}\left(\mathrm{SAD}\left(\mathbf{c}\right)\right) \quad (4)$$

where B_{ij} and $\mathbf{c}$ represent a block in the interpolated frame and its MV candidates, s denotes the search range, and $\mathbf{v}'$ is the calculated MV with the minimum SAD value.

2.2. Occlusion Detection

How to deal with the MVs in the occlusion areas is a crucial issue in MCFI, which can be classified as covering and uncovering areas. Normally, these areas occur at the transition regions of different layers. For example, in Fig. 2 the black ellipse is a foreground object with a motion from upper left to lower right, then the yellow region in f_t is the covering area, with its corresponding area only in f_{t-1}. Similarly, the green region in f_t is the uncovering area.

$$d_t\left(\mathbf{p}\right) = \frac{1}{2}\cdot d_{t-1}\left(\mathbf{p}+\frac{\mathbf{v}'}{2}\right) + \frac{1}{2}\cdot d_{t+1}\left(\mathbf{p}-\frac{\mathbf{v}'}{2}\right) \quad (5)$$

With the initial MVF from block-based ME, the coarse interpolated depth frame can be obtained in (5). Then, we use the variance of the corresponding depth blocks to measure the depth irregularity:

$$c_{ij} = \frac{1}{N}\sum_{\mathbf{r}\in D_{ij}} \left\| d_{\mathbf{r}} - \bar{d}\right\|_2 \quad (6)$$

where $d_{\mathbf{r}}$ is the depth value of a pixel in depth block D_{ij}, and $\bar{d}$ is the average depth value of D_{ij}. N is the pixel number in D_{ij}, and $\|\cdot\|_2$ represents 2-norm.

Obviously, the variance value of most blocks in the depth frame are small. This is because the depth values are homogeneous inside the same layer. However, if a block locates at the boundary of different layers, the variance value will be large. Therefore, a threshold T_c is defined empirically to identify these boundary blocks. Let D_{ij}^t be a boundary block in the interpolated depth frame corresponding to a initial MV $\mathbf{v}'$, and vector $\mathbf{d}$ indicates the depth distribution of D_{ij}^t. The initial point and terminal point of $\mathbf{d}$ correspond to the mass center and geometric center of D_{ij}^t, respectively. Then the angle θ between $\mathbf{d}$ and $\mathbf{v}'$ is calculated, and if $0 \le \theta < \pi/2$, D_{ij}^t is an covering block. Otherwise, if $\pi/2 < \theta \le \pi$, D_{ij}^t is an uncovering block.

Suppose D_{ij}^t is a covering or uncovering block, $D_{i'j'}^t$ is one of its surrounding blocks. $\bar{d}$ is the average depth value of D_{ij}^t, $\bar{d}'$ and $c_{i'j'}$ are the average depth value and variance of $D_{i'j'}^t$, respectively. If $\left(\bar{d}' < \bar{d}\right) \cap \left(c_{i'j'} < T_c\right)$, this depth homogeneous block will be re-marked as a covering or uncovering block.

2.3. Adaptive Pixel-based ME

It is obvious that block-based MVs of the occlusion blocks are unreliable. In this subsection we will introduce an adaptive pixel-based ME method to refine the initial MVF. In our scheme, optical flow is adopted for pixel-based ME [10].

As illustrated in Fig. 2, two blocks containing both foreground and background pixels are marked by red rectangles in f_{t-1}, which are denoted as block A (uncovering) and block

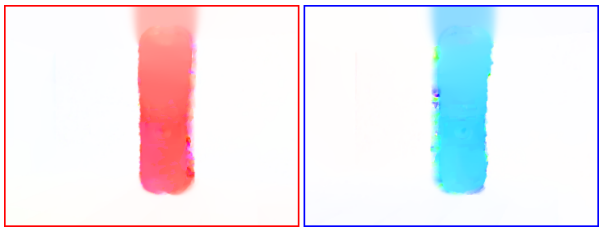

Fig. 3. Optical flow fields for one frame of Mobile sequence. Left: the result using forward optical flow estimation; Right: the result using backward optical flow estimation.

B (covering). A_1 and A_2 are the foreground and background parts for block A, similarly, B_1 and B_2 for block B. Suppose the foreground object has a different motion with the background, therefore, occlusion occurs in f_{t+1}. Apparently, we cannot find true matching blocks for A and B in current frame f_{t+1}. Nevertheless, in pixel level every pixel in block A can find the true corresponding pixel in f_{t+1}. On contrary, for block B only the pixels of B_1 have the true corresponding pixels. Thereby the MVs obtained using pixel-based ME for block A is more accurate than the MVs for block B. As shown in Fig. 3, a foreground object moves from left to right, and the left and right figures of Fig. 3 are the forward and backward optical flow estimation results for F_{t-1} with F_{t+1} and F_{t-3} as the reference frames respectively. We can see that forward optical flow estimation result on the left side of the foreground object is more accurate, but the quality of the optical flow on the right side deteriorates obviously. Contrarily, in the right figure the optical flow estimation at the right side outperforms that at the left side.

Based on this observation, an adaptive pixel-based ME is proposed to refine the initial block-based MVF. Let $B_t^{\mathbf{v}}$ and $B_t^{\mathbf{u}}$ denote a covering block and an uncovering block in the interpolated frame f_t, respectively. Next, we find the coarse matching block for $B_t^{\mathbf{v}}$ in f_{t+1} based on the initial MVF, and denote it as $B_t^{\mathbf{v}'}$. Similarly, $B_t^{\mathbf{u}'}$ is the coarse matching block for $B_t^{\mathbf{u}}$ in f_{t-1}. Assuming $\mathcal{O}_1$ is the set of the coarse matching blocks in f_{t+1} for all covering blocks in f_t, and $\mathcal{O}_2$ is the coarse matching block set in f_{t-1} for all uncovering blocks in f_t. Algorithm 1 outlines the pseudocode of adaptive pixel-based ME for initial MVF refinement.

In Algorithm 1, $\mathrm{warp}\,(\cdot)$ denotes MV warping using the MVs via forward or backward optical flow estimation. $\mathbf{V'}_t$ is the initial dense MVF refined based on optical flow estimation, and $\mathbf{V''}$ is the dense MVF after hole filling. $\mathrm{MVfill}\,(\cdot)$ is applied to fill the holes in $\mathbf{V'}_t$ referring to the depth information [11].

Assume a pixel $f_{t+1}(\mathbf{p}_1) \in \mathcal{O}_1$ with $\mathbf{v}_1$ as its MV, and this MV will warp to a pixel $f_t(\mathbf{p}_m)$ in the interpolated frame,

$$\mathbf{p}_m = \mathbf{p}_1 + \frac{1}{2}\mathbf{v}_1 \tag{7}$$

Algorithm 1 Initial MVF refinement based on optical flow estimation.

Input: The initial MVF; Previous frame f_{t-1} and d_{t-1}; Current frame f_{t+1} and d_{t+1}; $\mathcal{O}_1$; $\mathcal{O}_2$.

Output: Dense MVF of covering and uncovering blocks on f_t.

for each pixel $p_{ij} \in \mathcal{O}_1$ on f_{t+1} **do**
  Compute the MV $\hat{\mathbf{v}}_{t+1}$ for p_{ij} using backward optical flow estimation, f_{t-1} is the reference frame.
  $\mathbf{v}_t^b = \mathrm{warp}\,(\hat{\mathbf{v}}_{t+1})$.
  $d_t^b = \mathrm{warp}\,(d_{t+1})$.
end for
for each pixel $p_{ij} \in \mathcal{O}_2$ on f_{t-1} **do**
  Compute the MV $\hat{\mathbf{v}}_{t-1}$ for p_{ij} using forward optical flow estimation, f_{t+1} is the reference frame.
  $\mathbf{v}_t^f = \mathrm{warp}\,(\hat{\mathbf{v}}_{t-1})$.
  $d_t^f = \mathrm{warp}\,(d_{t-1})$.
end for
$\mathbf{V''}_t = \mathrm{MVfill}\,(\mathbf{V'}_t)$

with its MV as $\frac{1}{2}\mathbf{v}_1$. Similarly, as for a pixel $f_{t-1}(\mathbf{p}_2) \in O_2$, the position and MV of its corresponding pixel in the interpolated frame are $\mathbf{p}_m = \mathbf{p}_2 + \frac{1}{2}\mathbf{v}_2$ and $\frac{1}{2}\mathbf{v}_2$. In our MV warping stage, if there are multiple source MVs warp to one target pixel, then the MV with the largest depth value is chosen as the matching MV. If there is no source MV warping to a pixel in the interpolated frame, then the pixel will be labeled as a hole pixel. And the MVs of the hole pixels will be filled via hole filling method proposed in [11].

2.4. Hybrid Motion Compensated Frame Interpolation

The pixels in the interpolated frame have been classified into three types: pixels of depth homogeneous blocks, pixels of covering blocks and pixels of uncovering blocks. The MVs of the pixels in depth homogeneous blocks are obtained via block-based motion estimation, and these pixels can be interpolated using the average of the previous frame and current frame, which is the same as traditional MCFI methods. The MVs of the pixels in the covering and uncovering blocks are obtained by pixel-based motion estimation, i.e., backward or forward optical flow. If the pixel belongs to the covering blocks, only the previous frame is referred to. Otherwise, if the pixel belongs to the uncovering block, current frame is selected as the reference frame.

3. SIMULATIONS

In this section, simulation results are demonstrated to evaluated the performance of AHMEC. Four depth plus video sequences, i.e., *BeerGarden*, *BookArrival*, *Cafe* and *Newspaper*, are used in the experiments. All of these 3D sequences are provided by the MPEG. The spatial resolution of *Beer-*

Table 1. Simulation results in average PSNR and SSIM.

Sequences		FullSearch [12]	FullSearch+AHMEC	TriFilter [3]	TriFilter+AHMEC	MSEA [4]	MSEA+AHMEC
BeerGarden	PSNR	33.2151	**36.9310**	36.7978	**37.6155**	36.1466	**36.4495**
	SSIM	0.9785	**0.9860**	**0.9874**	0.9869	0.9866	**0.9872**
BookArrival	PSNR	30.3804	**32.7084**	32.1357	**32.9487**	32.8953	**33.0340**
	SSIM	0.9456	**0.9609**	0.9579	**0.9635**	0.9528	**0.9541**
Cafe	PSNR	34.2504	**36.4354**	36.0739	**36.8294**	36.4564	**36.5710**
	SSIM	0.9513	**0.9563**	**0.9640**	0.9636	0.9543	**0.9552**
Newspaper	PSNR	31.4424	**36.0471**	35.5153	**36.4887**	36.8465	**37.3277**
	SSIM	0.9730	**0.9849**	0.9848	**0.9865**	0.9851	**0.9867**

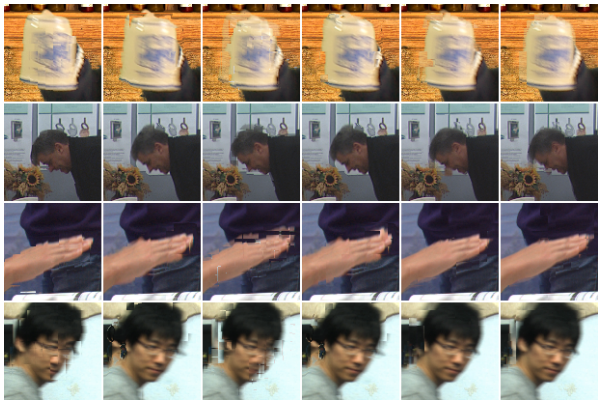

Fig. 4. Zoom-in results of the interpolated frames using the methods of FullSearch (1st column), FullSearch+AHMEC (2nd column), TriFilter (3rd column), TriFilter+AHMEC (4rd column), MSEA (5rd column) and MSEA+AHMEC (6rd column). From top to bottom: *BeerGarden*, *BookArrival*, *Cafe*, *Newspaper*.

Garden and *Cafe* is 960×540, and the spatial resolution of *BookArrival* and *Newspaper* is 512×384. In our experiments, the sequences are down-sampled by skipping the even frames, and then these even frames are interpolated by various MCFI methods. The block size in motion estimation and compensation is 8×8, and the search range is 13×13.

Three block-based benchmark MCFI methods are adopted: full search motion estimation algorithm (FullSearch) [12], FRUC using trilateral filtering (TriFilter) [3] and multi-level successive eliminate algorithm (MSEA) [4]. The proposed AHMEC is integrated into these benchmark methods to test its performance. Please note that TriFilter algorithm adopts unidirectionally block-based ME. In order to integrate the proposed AHMEC to TriFilter, we allocate a MV for each block in the interpolated frame based on the forward and backward MVFs by considering the spatial correlation. The quality of the interpolated frames are evaluated by Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Table 1 shows the average PSNR and SSIM of the test sequences, which contains 51 original frames. We can observed that the proposed AHMEC method improves the qualities of the interpolated frames for all the three block-based benchmark methods (FullSearch, TriFilter and MSEA).

Figure 4 illustrates the zoom-in results of the interpolated. It can be observed that in the interpolated frames of the benchmark methods, ghost artifacts exist at the foreground objects boundaries. This is mainly because block-based ME is not accurate if the block contains multiple motion layers. However, by using the proposed AHMEC, the MVs of these boundaries blocks are detailed to pixel-wise, furthermore, these pixels can be interpolated adaptively according to the relationship of depth and motion. Therefore, these ghost artifacts can be suppressed significantly. Take the *Newspaper* sequence for example, in Fig.4, the boundaries of the person's head are more clear by applying the proposed AHMEC to the benchmark methods.

4. CONCLUSIONS

A MCFI method based on adaptive hybrid motion estimation and compensation for 3D video is proposed in this paper. We focus the problem of ghost artifacts caused by occlusion in the interpolated frame. First, the MVF of the interpolated frame is estimated via traditional block-based ME method. Then, the covering and uncovering areas in the interpolated frame is distinguished by the depth and motion information. After that, an adaptive pixel-based ME method is applied to refine the MVF in the covering and uncovering areas. Experimental results show that, by applying the proposed AHMEC, the qualities of the frame interpolation are improved in terms of average PSNR and SSIM, and the visual results of the interpolated frames are better.

5. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Nos. 61173079, 61472163 and 61501204), the Science and technology project of Shandong Province (No. 2015GGX101025) and the Doctor Fund of University of Jinan.

6. REFERENCES

[1] B. D. Choi, J. W. Han, C. S. Kim, and S. J. Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 407–416, Apr. 2007.

[2] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Processing Mag.*, vol. 24, no. 7, pp. 10–21, Nov. 2007.

[3] C. Wang, L. Zhang, Y. He, and Y.-P. Tan, "Frame rate up-conversion using trilateral filtering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 886–893, June 2010.

[4] Y.-L. Lee and T. Nguyen, "Fast one-pass motion compensated frame interpolation in high-definition video processing," *In Proc. ICIP*, pp. 369–372, Nov. 2009.

[5] Y. Cho, H. Lee, and D. Park, "Temperal frame interpolation based on multiple feature trajectory," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2105–2115, Dec. 2013.

[6] D. Wang, A. Vincent, P. Blanchfield, and R. Klepko, "Motion-compensated frame rate up-conversion Part II: New algorithms for frame interpolation," *IEEE Trans. Broadcast.*, vol. 56, no. 2, pp. 142–149, June 2010.

[7] K. Chen and D. A. Lorenz, "Image sequence interpolation using optimal control," *J. Math. Imag. Vision*, vol. 41, no. 3, pp. 222–238, 2011.

[8] M. Werlberger, T. Pock, M. Unger, and H. Bischof, "Optical flow guided TV-L1 video interpolation and restoration," *In Proc. Energy Minimization Methods Comput. Vision Pattern Recognit.*, vol. 6819, pp. 273–286, 2011.

[9] W. Lee, K. Choi, and J. Ra, "Frame rate up conversion based on vatiational image fusion," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 399–412, Jan. 2014.

[10] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," *Doctoral Thesis*, Massachusetts Institute of Technology, May 2009.

[11] X. Yang, J. Liu, J. Sun, X. Li, and W. Liu, "DIBR based view synthesis for free-viewpoint television," *In Proc. IEEE 3DTV-Conf.*, pp. 1–4, May 2011.

[12] M.-J. Chen, L.-G. Chen, and T.-D Chiueh, "One-dimensional full search motion estimation algorithm for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, no. 5, pp. 504–509, Jan. 1994.