BLOCK-MATCHING TRANSLATION AND ZOOM MOTION-COMPENSATED PREDICTION BY SUB-SAMPLING

Ka-Man Wong, Lai-Man Po, Kwok-Wai Cheung[#] and Ka-Ho Ng

kmwong@ee.cityu.edu.hk, eelmpo@cityu.edu.hk, kwcheung@chuhai.edu.hk, khng@ee.cityu.edu.hk
Department of Electronic Engineering, City University of Hong Kong, Hong Kong
Department of Computer Science, Chu Hai College of Higher Education, Hong Kong [#]

Abstract - In modern video coding standards, motion compensated prediction (MCP) plays a key role to achieve video compression efficiency. Most of them make use of block matching techniques and assume the motions are pure translational. Some attempts toward a more general motion model usually too complex to be practical in near future. In this paper, a new Block-Matching Translation and Zoom Motion-Compensated Prediction (BTZMP) is proposed to extend the pure translational model to a more general model with zooming in a practical way. It adopts the camera zooming and object motions that becomes zooming while projected on the video frames. The proposed BTZMP significantly improve motion compensated prediction. Experimental results show that BTZMP can give prediction gain up to 1.09dB compared to conventional sub-pixel block-matching MCP. In addition, BTZMP can be incorporated with Multiple Reference Frames (MRF) technique to give extra improvement, evidentially by the prediction gain ranging up to 2.08dB in the empirical simulations.

Index Terms – Video Coding, Motion Compensated Prediction, Translation and Zoom Motion.

I. Introduction

Motion compensated prediction is the core component in hybrid video codecs like H.26X series and MPEG series. Many techniques are proposed to improve the prediction accuracy thus the number of bits required to encode can be reduced. Most MCP techniques make use of a pure translation motion model as shown in fig. 1(a). It assumes the objects are rigid without translation and there is no depth change for the objects. In actual 3D environment the scene have depth thus object can move closer or away from the projection plane as shown in fig. 1(b). The objects with different depths will be projected with different size due to the perspective effect. Also, objects are not necessary to be rigid and without rotation. In current video coding standards, pure translation motion model is used due to its simplicity of implementation in software and hardware.

For hybrid video codecs using pure translation motion model, residue block are used to record the differences between the translated block and the actual content of the video source. In H.264 [1-4], many techniques are used to minimize the difference to achieve higher compression ration. Variable block size (VBS) [5] is used to improve motion compensated prediction by allowing using smaller sub blocks for prediction. Multiple reference frames (MRF) [6] is used to provide additional candidates for prediction for longer period of time. Sub-pixel MCP [7,8] is used to refine the accuracy of motion vectors by interpolated reference frames. Rate distortion optimization (RDO) [9] is used to optimize the tradeoff the prediction gain and the side bits introduced by these techniques.

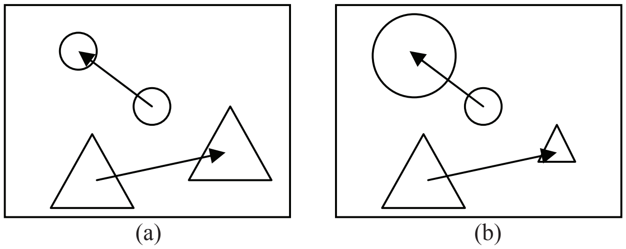

Fig. 1 2D Frame based motion model (a) Translation motion model (b) Translation and zoom motion model

Another approach is to use a more general affine model [10-12] involving translation, rotation and zoom motion to increase the prediction accuracy. In [10], affine parameter sets are estimated and multiple "wrapped" frames are generated based on the parameter sets as references and the affine parameters are transmitted if the block in a wrapped frame is selected as the best prediction candidate. In [11], two-stage motion compensation is carried out to find local and global motion vector (MV) involving affine parameter estimation both stages. In [12], motion vectors of other blocks are used to make an affine transformed block as a searching candidate. These methods require affine parameter estimation and the complexity limits the deployment of these motion models in practical applications.

In this paper, a motion-compensated prediction model based on the translation and zoom motion model as fig. 1(b) will be proposed. By combining translation and zoom motion components, the proposed MCP can better suit the real motion than pure translation. In addition, by using block-based implementation, the new MCP model can be easily deployed into the existing video coding framework. It should be noted that the major difference between this work and the methods in [10-12] is that out method does not require any affine parameter estimation by performing an affine transform, also our method does not require additional memory requirement to store additional reference frames. Zoom parameters are indicated by the sub-sampling step thus no other side information about zoom scale will be transmitted. The paper is organized as follows. In section 2, block-matching model for MCP will be explained and the proposed model will be introduced in a way to utilize the zooming property with a zooming scheme that allows the zoomed frames can be acquired without additional computation. Section 3 provides experimental results for access achievable gain with the proposed technique. Finally, conclusion will be drawn in section 4.

II. Block-Matching Translation and Zoom Motion-Compensated Prediction

2.1. Translation motion based MCP

In block matching MCP with multiple reference frames (MRF)

$$\{F_{t-1}, F_{t-2}, F_{t-3}, \dots , F_{t-P}\} \ , \qquad (1)$$

F_{t-d} is the frame with delay d and $F(a,b)$ denotes the intensity of a

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

pixel located at *(a,b)*. The motion for a *NxN* block in the current frame F_t is represented by the motion vector

$$MV(u,v,d) \tag{2}$$

using the minimum sum of absolute difference (SAD) criterion

$$SAD(x,y,u,v,d) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|F_t(x+i,y+j) - F_{t-d}(x+i+u,y+j+v)\right| \tag{3}$$

between the block at *(x,y)* in the current frame F_t and the matched block at (u,v) in the reference frame F_{t-d} .

Since MRF use frames with different time delays, the limitation of single reference frame in case of temporary occlusions and periodic deformations can be resolved by selecting frames at other time that does not have such problems. Fig. 2 shows an example of MRF in case of temporary occlusion and periodic deformation.

2.2. Translation and Zoom MCP

To handle a more realistic motion model shown in Fig. 1(b), an additional zoom axis for zoomed frames is introduced and the set of reference frames becomes

$$\begin{Bmatrix} F_{t-1,-Z} & \cdots & F_{t-P,-Z} \\ \vdots & & \vdots \\ F_{t-1,0} & \ddots & F_{t-P,0} \\ \vdots & & \vdots \\ F_{t-1,+Z} & \cdots & F_{t-P,+Z} \end{Bmatrix} \tag{4}$$

where *Z* is the maximum number of zooming levels for zoom-in and zoom-out, the range (*-Z, +Z*) represents the zooming range used. The block matching criterion becomes

$$SAD(x,y,u,v,d,s) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|F_{t,0}(x+i,y+j) - F_{t-d,s}(x+i+u,y+j+v)\right| \tag{5}$$

where $F_{t-d,\,s}$ is the frame with delay *d* and with zoom level *s* that gives minimum SAD. The motion vector of a block is written as *MV*(*u, v, d, s*). Fig. 3 shows an example for BTZMP. The time axis shows the motions of two objects. The triangular object moves right and zooms in and out; and the circular object move left and zooms out. The zoom axis shows the zoomed frames. As the zooming is not periodic, MRF does not have the most appropriate frames for MCP. However, BTZMP zoom in frames and zoom out frames also provide suitable frames for MCP. In this case BTZMP gives a better predicted block for motion compensated prediction. Thus the prediction accuracy increased. Although providing zoomed frames for every zoomed object will make the motion prediction gain higher, it is not realistic to provide too many zoomed frames for the MCP. It will increase the memory requirement for storing such frames, and the computation for zooming and MCP will also be higher. So, a zooming scheme does not require addition memory and computation on generating zoomed frames will be proposed.

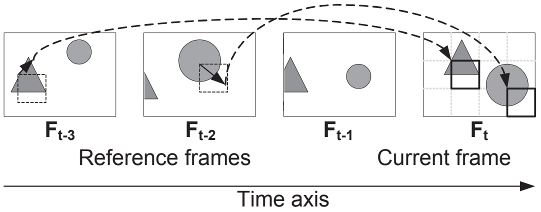

Fig. 2 An example of MRF where the triangular object moving out of the frame for two frames and back; and the circular object moving downward and zoom periodically.

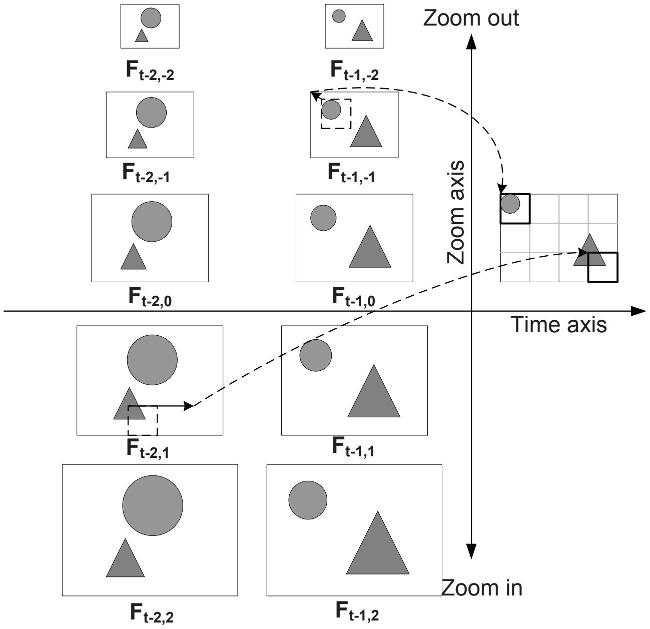

Fig. 3 An example of block MCP with MRF and BTZMP where the triangular object moving right and zooming in-out with time, and the circular object moving left and zooming out with time.

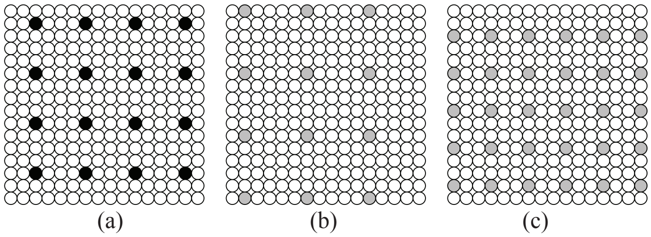

Fig. 4 (a) Original sub-sampling for conventional MCP; (b) Sub-sample by 5 to generate a zoom out frame; (c) Sub-sample by 3 to generate a zoom in frame.

2.3 The zooming scheme

The BTZMP uses several zoomed decoded frames with different zoom factors. It should be noted that the proposed method does not have the procedure to detect the zoom motion before the generation of zoomed frames like [9-11]. Although the BTZMP model involves multiple zoomed frames for MCP, it is not necessary to reserve memory space to store these frames, we can obtain them by sub-sampling. Since sub-pixel MV resolution is commonly used in modern video coding standards, interpolated reference frames will be available. We can sub-sample them by various sub-sampling factors to obtain zoomed frames with different zoom scales. Fig. 4 shows examples of sub-sampling. Assume 1/4 pixel MCP is used and a 4X interpolated reference frame is already available. Black dots in fig. 4(a) indicate the pixels for full pixel locations and the surrounding white dots are pixels for sub- pixel MCP. Grey dots in fig. 4(b) and 4(c) indicate sub-sampled full pixel locations in zoomed resolutions and its surrounding white dots can also be used for sub pixel MCP. For example if the image is sub-sampled by 5, it can have 1/5 pixel resolution for sub pixel MCP. The available scales for the zoomed frames can be given by:

$$scales = \frac{s}{i} \quad \text{where } i = \{1, 2, ...n\} \qquad (6)$$

where s is the inverse of base MV resolution, i is the sub-sample factors, and n is the number of zoom levels. If a video codec provides 1/4 pixel MCP, we can obtain zoomed frames for the following scales:

$$\{ 4, 2, 4/3, 1, 4/5, 4/6, 4/7, ..., 4/n\} \qquad (7)$$

Although this approach does not provide flexible zoom scales compared to generate the zoom frames directly. It does not require any computation to obtain the zoomed frames. Another advantage is that in modern video codecs like H.264, there will have a carefully selected interpolation filter for sub-pixel MCP thus it will be more easy to implement since it is not required to design another filter to interpolate the zoomed frames.

III. Experimental results and analysis

The motion compensated prediction gain of BTZMP will be presented via several experiments. Firstly, prediction gains of sub-pixel MCP without BTZMP will be provided to verify the properties of conventional sub-pixel MCP. Secondly, prediction gain of BTZMP will be shown to provide a picture of the improvement. And then, prediction gain of various number of zoom levels will be investigated and the effective number of zoom levels will be selected. Finally, additional prediction gain of combining BTZMP and MRF will also be given by experimental results.

3.1. Experiment setup

In the experiments, sequences *akiyo*, *foreman*, *mobile* and *stefan* are used, they are at CIF resolution and 300 frames. The macroblock size is fixed at 16x16 and the search windows size is set at ±16 and exhaustive search is used within the search window. Various sets of zoom scales from different base MV resolutions will be tested. For the zoomed frame, sub pixel locations will also be searched. To interpolate the image for sub-pixel MCP, the filter [1 -5 20 20 -5 1] in H.264 is used. Sum of absolute difference (SAD) is used for the block distortion analysis. Since these tests are used to explain the potential of BTZMP improves the motion compensated prediction, original frames and zoomed original frames are used for the BTZMP. It is different from a real video codec that use reconstructed frames for motion compensated prediction, so the bit rate is not available at this stage.

3.2. Sub-pixel MCP

Fig. 5 shows the PSNR improvement of conventional sub-pixel MCP with different MV resolution, it can be seen that the improvement of 1/4 pixel MCP is significant and the further improvements become small while 1/8 and 1/16 pixel MCP are used. It matched the results found in [7-8]. This is one of the reasons that 1/8 and 1/16 MV resolutions are not used in practical video codecs.

3.3. Prediction accuracy improvement of BTZMP

To show the improvement of BTZMP, we test several sets of zoom scales and the detailed zoom scales used are tabulated in TABLE 1. Fig. 6 shows the PSNR improvement of BTZMP. The labels 1/2, 1/4, 1/8, 1/16 relate to the base MV resolution and the zoom scale set used. The improvement is relative to the conventional sub-pixel MCP with the base MV resolution. It can be observed that BTZMP significantly improved the estimation accuracy, and also the improvement keep growing while higher base MV resolution is used.

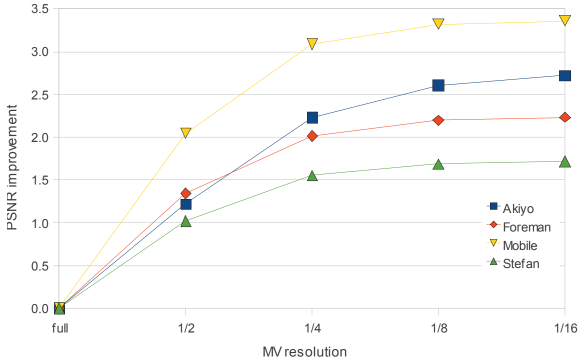

Fig. 5 PSNR improvement of conventional sub-pixel MCP with various MV resolutions compared to full pixel MCP

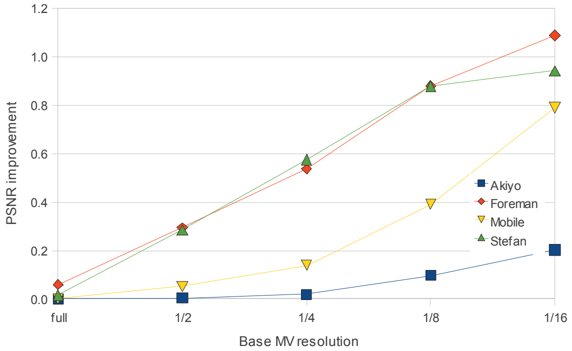

Fig. 6 PSNR improvement of BTZMP with various base MV resolutions in addition to conventional sub-pixel MCP

TABLE 1 List of zooming scales in zoom step analysis

Base MV resolution	Zoom scales of zoomed frames
1/2	{2/1, 2/2, 2/3, …., 2/14, 2/15}
1/4	{4/1, 4/2, 4/3, …, 4/14, 4/15}
1/8	{8/1, 8/2, 8/3, …., 8/14, 8/15}
1/16	{16/9, 16/10, 16/11, …, 16/23, 16/23}

We can also see that improvement is up to 0.79-1.09dB (except *akiyo*) for these sequences. The improvement for *akiyo* is small since the background is static and the object does not have significant depth changes throughout the sequence thus the improvement is limited. Among the base MV resolutions, 1/16 gives the highest PSNR improvement that indicates it cover more zoom motion than other tested zoom scale sets. For CIF resolution, 1/16 base MV resolution can be selected for sub-sampling the zoomed frames for BTZMP.

3.4. Number of zoom levels

Although the estimation can be further improved by adding more zoom levels and using higher base MV resolution, it enlarges the memory requirement and increases the computation complexity, an effective number of zoom levels should be found to obtain significant improvement. PSNR improvement of various number of zoom levels will be tested. Fig. 7 shows an analysis with 1-15 zoom levels, it can be seen that 7 levels can give improvement close to the improvement with 15 zoom levels.

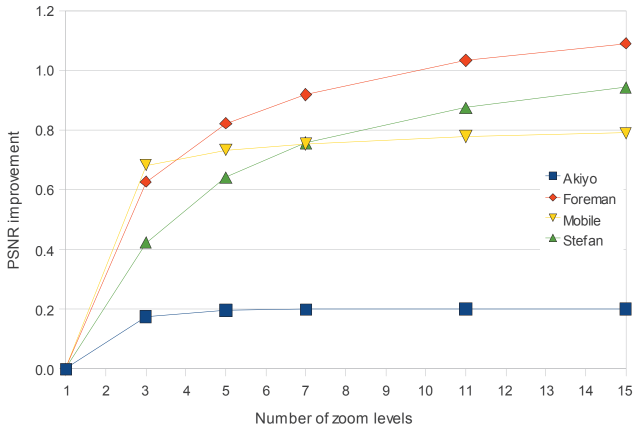

Fig. 7 PSNR improvement of BTZMP with various number of zoom levels

3.5. Combine with multiple reference frames

As mentioned in introduction, BTZMP is to provide a new aspect to existing techniques for improving prediction accuracy. In this part, analysis on BTZMP compared to and combined with MRF is provided. TABLE 2 tabulates the PSNR results of these techniques. It is clear that BTZMP is very competitive with multiple reference frames. It can also be seen that BTZMP+MRF always give the highest prediction improvement. BTZMP+MRF with 5 zoom levels and 5 reference frames can give a prediction gain up to 1.62dB compared to conventional method, and it further go up to 2.08dB while 11 zoom levels and 11 reference frame are used. The combination of BTZMP and MRF gives about additional improvement about 0.16 – 0.95dB in sequences compare to either MRF or BTZMP. From these observations it can be seen that these techniques can be combined very well and make an extra gain in prediction accuracy improvement.

IV. Conclusions

In this study, a new technique for motion estimation that exploits the zooming effect in video sequences is presented. The conventional pure translation motion model is extended by a zooming axis that covers zooming effect from real world motions projected on the video frame.

BTZMP is proposed for this translation and zoom motion model by using multiple zoomed reference frames. The proposed method does not require detection of zooming effect between current frame and reference frame. It generates the zoomed frame by sub sampling the interpolated reference frames for sub-pixel MCP, it does not require addition computation for interpolation.

Analysis found that while using the base MV resolution of 1/16 pixel, BTZMP gives prediction improvement up to 1.08dB with CIF video sequences. Analysis also found that using 7 zoom levels can give prediction gain close to the maximum improvement. These figures show that the prediction gain achievable with BTZMP is significant. As BTZMP give significant improvement with 1/8 and 1/16 pixel base MV resolutions, it resolved a drawback for not using higher sub-pixel MV resolution.

BTZMP is also found very compatible with MRF. It improves motion prediction in a different aspect, combination of BTZMP and MRF take the prediction improvement to another stage. It can be concluded that BTZMP will be a promising and also achievable technique for future multimedia standards.

TABLE 2 PSNR comparison of BTZMP and multiple reference frames

	Akiyo	Foreman	Mobile	Stefan
Full pixel MCP	43.1343	32.2988	24.7680	24.9241
1/16 pixel MCP	45.8548	34.5324	28.1314	26.6275
MRF, 5 references	46.0518	35.3555	28.8652	27.2688
BTZMP, 5 zoom levels	45.9566	35.3328	28.7368	27.0548
BTZMP+MRF, 5 zoom levels, 5 references	46.1284	36.1545	29.5005	27.6998
MRF, 11 references	46.0569	35.5661	28.9096	27.5028
BTZMP, 11 zoom levels	46.0701	35.6632	28.8547	27.2586
BTZMP+MRF, 11 zoom levels, 11 references	46.2339	36.6180	29.6603	28.1110

ACKNOWLEDGMENT

The work described in this paper was substantially supported by a grant from Applied R & D Funding (ARD) of City University of Hong Kong [Project No. 9667014] and also partially supported by a contrast research project for "Future Multimedia Standard" from ASTRI (Hong Kong Applied Science and Technology Research Institute Company Limited).

References

[1] A. Luthra, G. J. Sullivan, and T. Wiegand, "Introduction to the special issue on the H.264/AVC video coding standard," IEEE Trans. CSVT., vol. 13, no. 7, pp. 557–559, Jul. 2003.

[2] I. E. G. Richardson, H.264 and MPEG-4. New York: Wiley, 2003.

[3] T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. CSVT., vol. 13, no. 7, pp. 560–576, Jul. 2003.

[4] "Coding of Audio-Visual Objects – Part 2: Visual," MPEG-4 Visual Version 1, ISO/IEC 14 496–2, 1999.

[5] G. J. Sullivan and R. L. Baker, "Rate-distortion optimized motion compensation for video compression using fixed or variable size blocks," in Proc. GLOBECOM'91, Phoenix, AZ, Dec. 1991, pp. 85–90.

[6] T. Wiegand, X. Zang, and B. Girod, "Long-term memory motion-compensated prediction," IEEE Trans. CSVT., vol. 9, no. 1, pp. 70–84, Feb. 1999.

[7] B. Girod, "Motion-Compensating Prediction with Fractional-Pel Accuracy", IEEE Trans. Communications, vol. 4, no.4, pp. 604 – 612, Apr. 1993.

[8] T. Wedi and H. G. Musmann, "Motion- and Aliasing-Compensated Prediction for Hybrid Video Coding", IEEE Trans. CSVT., vol. 13, no. 7, pp. 577–586, Jul. 2003.

[9] G. J. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," IEEE Signal Processing Magazine, Vol. 15, No. 6, pp. 74-90, Nov. 1998.

[10] T. Wiegand, E. Steinbach, and B. Girod, "Affine Multi-Picture Motion-Compensated Prediction," IEEE Trans. CSVT, Vol. 15, No. 2, pp. 197-209, Feb. 2005.

[11] H. Jozawa, K. Kamikura, A. Sagata, H. Kotera, H. Watanabe, "Tow-Stage Motion Compensation Using Adaptive Global MC and Local MC", IEEE. Trans. CSVT, vol.7, no. 1, pp. 75-85, Feb. 1997.

[12] R. C. Kordasiewicz, M. D. Gallant, S. Shirani, "Affine Motion Prediction Based on Translational Motion Vectors," IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 10, pp. 1388 - 1394, Oct. 2007.