A NOVEL MOTION ESTIMATION METHOD BASED ON STRUCTURAL SIMILARITY FOR H.264 INTER PREDICTION

Zhi-Yi Mai[1], Chun-Ling Yang[1], Kai-Zhi Kuang[1], Lai-Man Po[2]

1 School of Electronic and Information Engineering, South China University of Technology, Guangzhou, Guangdong, 510640, China
2 Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China
kathymaizy@yahoo.com.cn, eeclyang@scut.edu.cn, danfaul@163.com, eelmpo@cityu.edu.hk

ABSTRACT

In the motion estimation of H.264, the best matching blocks and the best prediction modes are chosen by Lagrange cost function whose distortion metric is the sum of absolute (transformed) differences [SA(T)D] which has similar meaning with MSE or PSNR. Recently a new image measurement called Structural Similarity (SSIM) based on the degradation of structural information was brought forward. It is proved that the SSIM can provide a better approximation for the perceived image distortion than the currently used PSNR (or MSE). In this paper, we propose an improved motion estimation method based on SSIM(MEBSS) for H.264 inter coding. Experiment results show that the MEBSS can reduce average 20% bit rate and 2% encoding time while maintaining the same perceptual video quality, and the maximum reduction in bitrate is more than 50%.

1. INTRODUCTION

The recently established H.264/AVC is the newest video coding standard whose main goals are enhancing compression performance and providing network video presentation [1]. Also based on hybrid coding framework, H.264 utilizes variable block sizes and quarter-sample accurate motion compensation with multiple reference frames as well as other advanced techniques, hence achieves much higher coding efficiency than previous video coding standards.

In hybrid coding, motion estimation is the most important part for exploiting the temporal redundancy between successive frames but also cost high computation. Lagrange cost, whose block distortion measure is the sum of absolute (transformed) differences [SA(T)D], is used as the matching metric in H.264 motion estimation [2]. The SA(T)D has the same meaning with MSE or PSNR, which are currently the most widely used objective metrics due to their low complexity and clear physical meaning. However, MSE and PSNR have been widely criticized for not correlating well with Human Visual System (HVS) for a long time [3]. In the past several decades, a great deal of effort has been made to develop new image quality assessment based on error sensitivity theory of HVS, but only limited success has been achieved by the reason that the HVS is rather complex and has not been well comprehended.

Recently a new philosophy for image quality measurement was proposed, based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field. It says that a measure of structural information change can provide a good approximation to perceived image distortion [4,5]. In that philosophy, an item called Structural Similarity (SSIM) index, which includes three comparisons, is introduced to measure the structural information change. Experiments showed that the SSIM index method, which is easy to be implemented, can better correspond with human perceived measurement than PSNR (or MSE). Therefore, in this paper we propose a novel motion estimation method based on Structural Similarity (MEBSS).

The remainder of this paper is organized as follows. In section 2, the P-frame coding of H.264 and the idea of SSIM index are summarized. The detail of our proposed methods MEBSS is given in section 3. Section 4 presents the experimental results to demonstrate the advantage of the MEBSS. Finally, section 5 draws the conclusion.

2. H.264 P-FRAME ENCODER AND SSIM

2.1. H.264 P-frame Encoder

In H.264 P-frame encoder, each picture is partitioned into fixed-size macroblocks (MB) that cover a rectangular area of 16×16 samples of the luma component and 8×8 samples of each chroma component. Each macroblock is motion compensated predicted from other previously decoded pictures. The prediction residual then is integrally transformed, quantized, entropy coded and transmitted together with the side information for indicating either

Intra-frame or Inter-frame prediction. Partitions with size of 16×16, 16×8, 8×16 and 8×8 for each luma component of MB are supported by the P-frame syntax in block matching motion estimation. The 8×8 partition can also be further subdivided into 8×4, 4×8 or 4×4 sub-block partitions according to the syntax element. In addition, quarter-pixel motion compensation is used in order to improve the accuracy of motion estimation.

The block matching motion estimation is to seek the best matched block from the reference frames within a certain search range such as 16×16. Following Lagrange cost is used as the matching metric.

$$MCOST(s,c) = SA(T)D(s,c) + \lambda_{MOTION}\, Bit(\Delta MV) \quad (1)$$

In the above formula, *SA(T)D(s, c)* is the sum of absolute differences between original block *s* and candidate matching block *c*. SAD is applied for integer pixel motion estimation while SATD is for subpel [6]. λ_{MOTION} is the Lagrange multiplier for motion estimation. ΔMV is the difference between the predicted MV and the actual MV. *Bit(ΔMV)* is the number of bits representing the ΔMV.

The block(s) with the minimum *MCOST* will be chosen as the best matched block(s) for each prediction mode.

For each prediction mode, a rate-distortion cost is generated after finding the best-matched block by the following formula [7]:

$$J(s,c,MODE \mid QP) = D(s,c,MODE \mid QP) + \lambda_{MODE} R(s,c,MODE \mid QP) \quad (2)$$

where *MODE* is the prediction mode, *QP* is the quantization parameter. *D(s,c,MODE|QP)* is the sum of square differences (SSD) between original block *s* and reconstructed block *c*. *R(s,c,MODE|QP)* is the bit number of encoding the residue. The prediction mode with the minimum rate-distortion cost will be chosen as the best prediction for that MB.

2.2. Structural Similarity (SSIM)

The new idea of SSIM index is to introduce the measure of structural information degradation, which include three comparisons: luminance, contrast and structure [5]. It's defined as

$$SSIM(x,y) = l(x,y) \cdot c(x,y) \cdot s(x,y) \quad (3)$$

where $l(x,y)$ is Luma comparison, $c(x,y)$ is Contrast comparison and $s(x,y)$ is Structure comparison. They are defined as:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (5)$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (6)$$

where x and y are two nonnegative image signals to be compared, μ_x and μ_y are the mean intensity of image x and y respectively, σ_x and σ_y are the standard deviation of image x and y respectively, σ_{xy} is the covariance of image x and y. In fact, without C_3, the equation (6) is the correlation coefficient of image x and y. C_1, C_2 and C_3 are small constants to avoid the denominator being zero. It's recommended by [5]:

$$C_1 = (K_1L)^2,\ C_2 = (K_2L)^2,\ C_3 = C_2/2 \quad (7)$$

where $K_1, K_2 << 1$ and L is the dynamic range of the pixel values (255 for 8-bit grayscale images). In addition, the higher the value of $SSIM(x,y)$ is, the more similar the image x and y are.

3. NOVEL MOTION ESTIMATION BASED ON STRUCTURAL SIMILARITY (MEBSS) FOR P-FRAME

Variable block-size motion compensation is used in H.264 to improve the matching accuracy and achieve high compression efficiency. As the SSIM expresses the structural similarity of two images, the prediction block having larger SSIM implicates that it's more similar to the original one, and then lower frequency residual which can be easily encoded will be produced. And the best prediction modes in H.264 P-frame are determined after all the prediction residuals are transformed, quantized and entropy coded, which cost a great deal of computation complexity. According to the above analysis, we propose a novel motion estimation method based on Structural Similarity (MEBSS) for inter prediction.

In MEBSS, we use the SSIM rather than SAD as the distortion metric in the block matching motion estimation. According to the theory of SSIM, the candidate block is more similar with the original block when its SSIM index is greater while the SAD performs the other way. Therefore the distortion in our method is measured as:

$$D(s,c) = 1 - SSIM(s,c) \quad (8)$$

where *s* and *c* are the original and the prediction block respectively.

Due to the change of distortion measure, the Lagrangian multiplier should be modified correspondingly. In conformity to the relation between *SSIM(s, c)* and *Bit(ΔMV)* and motivated by the theory in [7] and [8], we obtain new Lagrangian multiplier from experiments. For example,

$\lambda'_{MOTION} = 294$ for Quantization Parameter is 10.

Consequently, the new cost function can be written as:

$$MCOST(s,c) = 1 - SSIM(s,c) + \lambda'_{MOTION}\, Bit(\Delta MV) \quad (9)$$

The major steps for each macroblock selecting the best matched block(s) and the best inter prediction mode are summarized as follow:

Step 1: For a whole MB, find the best matching block from all the candidate blocks using formula (9).

Step 2: Divide the MB into two 16×8 non-overlapped blocks. For each 16×8 block, find the best-matched block from all the reference frames using formula (9).

Then calculate the sum MCOST of these two 16×8 blocks.

Step 3: Divide the MB into two 8×16 non-overlapped blocks. Then do similarly as step 2.

Step 4: Divide the MB into four 8×8 non-overlapped blocks. For each 8×8 block, find the best-matched block from all the reference frames using formula (9). Then calculate the total MCOST of these four 8×8 blocks.

Step 5: If further sub-partition is allowed, find the best-matched blocks similarly for type 8×4, 4×8 and 4×4 respectively. Otherwise go step 6 directly.

Step 6: Find the prediction block of P_SKIP mode. The MCOST of it is 1-SSIM(s,c) because neither a motion vector nor a reference index parameter is transmitted in this mode.

Step 7: The prediction mode with the minimum MCOST is chosen as the best inter prediction mode of the MB. The residual of this best mode is transformed, quantized and entropy coded.

From the above motion estimation process description, it is clear that the *RDcost* which is shown in equation (2) is not used in MEBSS. In that way, MEBSS can reduce several coding processes for a MB, which leads to the decrement of coding computation load. However, since the computation load of SSIM is larger than SAD, we can't hope a larger reduction in the whole coding computation load. The simulation results in the next section will prove that clearly.

4. EXPERIMENTS

4.1. Experimental environment

Experiments are carried out using several color(Y,U,V) video sequences containing 50 frames. They are Carphone, Foreman, Grandma and News with size of 176×144, Hall_monitor and Mobile with size of 352×288.

All of our experiments are based on the JVT reference software JM92 program [9]. The results are performed on a P4/2.4GHz personal computer with 256MB RAM and Microsoft Windows XP as the operation system.

4.2. Experimental Results

The SSIM of each block is first computed within local 4×4 non-overlapped windows and then all the local SSIMs are averaged to a mean SSIM during motion estimation. The SSIM of the whole reconstructed image for each component is computed alike but using a 16×16 slide window instead. The MSSIM for the whole frame is generated by formula (10) according to the character of Human Visual System [10].

$$MSSIM = 0.7 * SSIM_Y + 0.15 * SSIM_U + 0.15 * SSIM_V \quad (10)$$

Furthermore, the following parameter settings is used in the SSIM measure: K_1=0.01, K_2=0.03, L=255. All our experiments use full search motion estimation, only one I frame and none B frame coded and allow using 8×8 transform additionally. The frame rate is 30Hz.

Results in terms of bit rate, MSSIM of the whole reconstructed image, total time and the comparison between MEBSS and H.264 are listed in table 1. These results are generated with QP=10.

We can obtain from Table1 that while maintaining almost the same MSSIM, our proposed MEBSS can reduce average 20% bit rate and 2.5% encoding time, and the maximum reduction in bit rate is more than 50% which is very great. The reason is that SSIM used as the distortion in MEBSS can choose the best matched block(s) whose residual images will be mostly low frequency signals and can be easily encoded at low bitrate.

On the other hand, we find the coding time has an average 2.5% reduction in our MEBSS. That's because only the $MCOST(s,c) = DISTORTION + \lambda_{MOTION} Bit(\Delta MV)$ is used, where *DISTORTION* is produced between the original image block and the prediction one. Only the residual signal of the best matched macroblock are transformed and encoded. Although the SSIM has a larger computation load than the SAD, MEBSS throws away several residual signals' transforming and encoding process. Therefore, the MEBSS has a little computation time reduction. When our MEBSS is used in H.264 fast motion estimation with fewer searching blocks, its advantage in computation load will be more obvious. And this will be approved in detail in our future work.

In order to compare video perceptual quality between our MEBSS and H.264 full search, the fiftieth reconstructed frame of Grandma is shown in figure 1, from which we can see that our MEBSS has the same visual quality with the H.264 full search while reducing the bit rate greatly.

5. CONCLUSION

In this paper we propose a novel motion estimation method based on Structural Similarity. Experiments show that MEBSS can reduce average 20% bit rate and 2.5% coding time while maintaining the same reconstructed video quality when quantization parameter is small (QP=10). The maximum reduction in bitrate is more than 50% which is very obvious. Although the time saving is not very large, the computation complexity analysis indicates that better results may be obtained when our MEBSS is used in H.264 fast motion estimation with fewer searching blocks for each prediction mode.

6. ACKNOWLEDGMENTS

The work described in this paper was substantially supported by research projects from National Natural

Table 1. Results with QP=10

Image	H.264 algorithm			MEBSS			Comparison (%)		
	Bit Rate (kbits/s)	MSSIM	Total Time (s)	Bit Rate (kbits/s)	MSSIM	Total Time (s)	Bit Rate inc.	MSSIM inc.	Time inc.
Carphone	1203.1	0.9970	301.1	1052.2	0.9951	303.1	-12.54	-0.19	0.68
Foreman	1567.9	0.9972	315.7	1299.8	0.9957	293.7	-17.10	-0.16	-6.96
Grandma	851.3	0.9966	302.0	414.1	0.9942	295.0	-51.36	-0.24	-2.31
News	562.17	0.9979	305.10	415.54	0.9971	284.33	-26.08	-0.08	-6.81
Hall_ monitor	7865.1	0.9955	1210.8	6952.8	0.9911	1214.5	-11.60	-0.44	0.31
Moblie	10587.2	0.9985	1409.4	9961.5	0.9975	1189.9	-5.91	-0.10	-15.57
Average	-	-	-	-	-	-	-20.77	-0.20	-2.53

(a) Grandma (original) (b) Encoded by H.264 with QP=10 MSSIM=0.9966 (c) Encoded by FMDBS with QP=10 MSSIM=0.9943

Figure 1. The 50th reconstructed frame of Grandma by H.264 and MEBSS

Science Foundation of China. [Project No. 60402015, No.60325310]

7. REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video coding Standard," *IEEE Trans. on CAS for video Technology*, no.7, Vol. 13, pp.560-576, July 2003.

[2] X.Q. Yi, J. Zhang, N. Ling and W.J. Shang, "Improved and simplified fast motion estimation for JM," presented at 16th JVT meeting (JVT-P021), Poznan, Poland, 24-29 July, 2005.

[3] J.L. Mannos, J.D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *In IEEE Trans. Information Theory,* no.4, pp. 525-536, 1974.

[4] Z. Wang, A.C. Bovik, and L.Lu, "Why is image quality assessment so difficult," in *Proc. IEEE Int. Conf. Acoustics, speech, and Signal Processing*, vol. 4, Orlando, FL, pp.313–3316, May 2002.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no.4, pp. 600–612, Apr. 2004.

[6] M. Wien, "Variable Block-Size Transforms for H.264/AVC," *IEEE Trans. on CAS for Video Technology*, vol.13, no.7, pp. 604-613, July 2003.

[7] T. Wiegand and B. Girod, "Lagrangian multiplier selection in hybrid video coder control," *in Proc. ICIP 2001*, Thessaloniki, Greece, Oct. 2001.

[8] G. J. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression", *IEEE Signal Processing Magazine,* vol. 15, no. 6, pp. 74-90, Nov. 1999.

[9] http://bs.hhi.de/~suehring/tml/download

[10] Z. Wang, L.G. Lu, A.C. Bovik, "Video Quality Assessment Using Structural Distortion Measurement," *in Proc. ICIP 2002*, Rochester, NY, USA, Sep. 22-25, 2002