IMPROVED INTER PREDICTION BASED ON STRUCTURAL SIMILARITY IN H.264

Chun-Ling Yang[1], Hua-Xing Wang[1],Lai-Man Po[2]

[1]School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, Guangdong, 510641, China
[2] Department of Electronic Engineering, City University of Hong Kong,
83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China
eeclyang@scut.edu.cn, washing2002@163.com, eelmpo@cityu.edu.hk

ABSTRACT

H.264 achieves higher compression efficiency by employing multiple modes inter prediction, rate-distortion (RD) optimal mechanism and other new techniques. Distortion metric plays an important role in video compression performance. Structural similarity (SSIM) is a new image quality assessment method, which is more consistent with Human Vision Systems (HVS). So in this paper, we propose to adopt SSIM as the distortion metric in the inter prediction cost functions, named "improved inter prediction method based on SSIM" (IPBSS). It is an improved method of our previous work MEBSS. Simulation results indicate that the IPBSS can averagely save bit rate more than 13% while maintaining the almost same video quality with QP = 10, 20 and 30. That is a better result than our previous work MEBSS.

Index Terms—H.264; inter prediction; distortion metric; structural similarity (SSIM).

1. INTRODUCTION

The coding efficiency of H.264 is greatly improved compared with previous coding standards [1]. Multiple modes inter prediction in motion estimation (ME) is one of the new techniques, which contributes much to the improvements. The inter coding partition modes for luma macro block(MB) include 16×16, 16×8, 8×16 and 8×8, and the 8×8 partition can be further divided into 8×4, 4×8 or 4×4 sub-blocks according to the syntax element. Its encoder calculates the RD costs of all the possible modes and selects the one with minimum RD cost as the best mode. In ME process, sum of absolute difference (SAD) is adopted as the distortion metric for block matching, while sum of the square differences (SSD) is used to calculate the distortion in the inter prediction mode decision process. Both methods to assess distortion are error sensitivity-based. They have the similar physical meaning with mean square error (MSE) and peak signal and noise ratio (PSNR). These measures have the common shortcoming that large errors do not always result in large perceptual distortions. So a more reasonable measure can also improve compression efficiency.

Structural similarity (SSIM) [2,3] is the newly developed approach to assess image and video quality, it extracts structure information from two corresponding image blocks, which is much more consistent with HVS than MSE and PSNR. The prediction block having larger SSIM implicates that it's more similar to the original one, then the residual block will be a lower frequency signal, which can be highly compressed. Therefore, we propose to adopt SSIM as the distortion metric in H.264 inter prediction process in the hope of getting more desirable results.

Our previous work [4] about this idea has been published in ICASSP2005, in which the SSIM of each block is first computed within local 4×4 non-overlapped partitions, and then the whole block SSIM is obtained by averaging all the SSIM of 4×4 partitions. It is unreasonable to calculate SSIM in this way, and that is also the main reason why the simulation results are worse for large QP in [4]. So in this paper we improve the method of calculating SSIM in motion estimation, and import SSIM in the mode decision process too.

The paper is organized as follows. Section 2 gives the brief introduction of SSIM and H.264 inter prediction process; Section 3 describes the details of the improved inter prediction method based on structure similarity (IPBSS) for video coding. Experiments and analysis of the proposed algorithm are given in section 4. The paper's summary remarks are shown in section 5.

2. SSIM and H.264 Inter-prediction

2.1. Structure similarity (SSIM)

SSIM exhibits much more consistency with subjective measures compared with other image assessment measures. It is defined as follows:

$$SSIM = l(x,y) \cdot c(x,y) \cdot s(x,y) \qquad (1)$$

Where $l(x,y)$ is luma comparison, $c(x,y)$ is contrast comparison and $s(x,y)$ is structure comparison, which are defined as follows:

1-4244-1236-6/07/$25.00 © 2007 IEEE

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

$$c(x,y)=\frac{2\delta_x\delta_y+C_2}{\delta_x^2+\delta_y^2+C_2} \tag{3}$$

$$s(x,y)=\frac{\delta_{xy}+C_3}{\delta_x\delta_y+C_3} \tag{4}$$

x, *y* are two nonnegative image signals. $\mu_x\mu_y$ are the mean of image *x* and *y* respectively, $\delta_x\delta_y$ are the corresponding standard deviation of image x and y, δ_{xy} is the covariance of image x and y. C_1, C_2 and C_3 are small constants to avoid the denominator being zero, and taken the same values as in [2].

2.2. H.264 Inter-prediction

In H.264, the coding efficiency is greatly improved by using RD optimum and variable block size for inter prediction. The inter prediction has two steps, the first step is to choose the best matching block of the current encoding MB for each inter prediction mode, the following Lagrange cost function is used as the matching metric.

$$MCOST(s,c)=SA(T)D(s,c)+\lambda_{MOTION}Bit(\Delta MV) \tag{5}$$

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 to integer pixel motion search while SATD is for sub-pixel [1]. λ_{MOTION} is the Lagrange multiplier for motion estimation. ΔMV is the difference between the predicted motion vector (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 matching block(s) for each prediction mode.

The second step is to choose the best inter prediction mode. RDcost is calculated after finding the best-matching block by the following formula:

$$J(s,c,MODE|QP)=D(s,c,MODE|QP)+\lambda_{MODE}R(s,c,MODE|QP) \tag{6}$$

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 used to encode the residue. The mode with the minimum RD cost will be chosen as the best prediction one for that MB.

3. Inter prediction based on structure similarity (IPBSS)

We adopt SSIM as the distortion metric in the H.264 inter prediction process, which is called inter prediction based on structure similarity (IPBSS). In the proposed method, SSIM rather than SAD or SSD is adopted as the distortion metric in the block matching and inter prediction mode decision. According to the theory of SSIM, the candidate block is more similar with the original one when its SSIM index is larger, while the SAD and SSD work in the other way. Therefore, the cost function for motion estimation for each inter prediction mode is defined as:

$$MCOST(s,c)=K_1(1-SSIM(s,c))+\lambda_{MOTION}Bit(\Delta MV) \tag{7}$$

And the cost function for mode(s) decision is defined as:

$$J(s,c,MODE|QP)=K_2(1-SSIM(s,c))+\lambda_{MODE}R(s,c,MODE|QP) \tag{8}$$

λ_{MOTION} and λ_{MODE} are the same as in formula (5) and (6). K_1 and K_2 are multipliers to enlarge (1-SSIM), which are obtained by intensive experiments. So the new Lagrange multipliers in the above two formulas correspond to λ_{MOTION}/K_1 and λ_{MODE}/K_2 .

The major steps of selecting the best inter prediction mode and the best matching block(s) for each MB are summarized as follows:

Step 1 Choose the best matching block(s) for each inter prediction mode:

Calculate the SSIM between the current block(s) and all the candidate ones in the current block size. Then find the best matching block(s) from all the candidate ones using formula (7). But for the P_SKIP mode, the *MCOST* is *1-SSIM(s,c)* because neither motion vector nor residual signal is transmitted in this mode.

Step 2 Choose the best prediction mode for the MB:

For each prediction mode and its best matching block(s), calculate the RDcost using formula (8), and the prediction mode with the minimum RDcost is selected as the best mode.

Compared IPBSS with MEBSS proposed in our previous work [4], there are two main improvements. The first improvement is that we modify the way to calculate SSIM in the ME process, which makes IPBSS can also achieve better result compared with MEBSS when QP is larger. In [4], the SSIM between the current block and the candidate ones is first computed within local 4×4 non-overlapped partitions, then SSIM is obtained by averaging all the partitions SSIM. This is not reasonable for larger block inter prediction mode, such as block with size of 16×16, 16×8, 8×16, 8×8, 8×4 or 4×8. Because SSIM is block-based rather than pixel-based, the mean SSIM of the 4×4 block partitions is not equal to that of the whole block. This makes the results deteriorate when QP is larger. In our proposed method, the SSIM are calculated in the current block's size directly, the result is promising especially for larger QP in this way. The second improvement is that we adopt SSIM instead of SSD as distortion metric in the mode decision process, as is shown in formula (8). The improved IPBSS can obtain better coding efficiency than MEBSS, which will be analyzed in next section.

4. Experiment and results analysis

4.1. Experimental environment

All the experiments are carried out under the following conditions:

1. The proposed IPBSS is implemented by modifying the H.264/AVC reference software JM11.0 [5] and other experiments are based on this software too. Five reference frames and full search motion estimation are used for inter prediction; In order to compare the performance between the proposed IPBSS and the original H.264 in inter coding, intra mode coding is forbidden in inter frame coding in both algorithms;

2. Experiments were conducted for three quantization parameters QP = 10, 20 and 30. Each algorithm is tested with 10 video sequences. The sequence "waterfall" is common intermediate format (CIF) with size of 352×288. All the other test sequences are quarter common intermediate format (QCIF) with size of 176×144. For each sequence, 50 frames are encoded with the first frame as I-frame with QP=10, and the rest 49 frames as P-frame;

3. The results are performed on PD/2.8GHz personal computer with a 512×2M RAM and Microsoft Windows XP as the operation system.

4.2. Experimental Results

The Mean SSIM (MSSIM) is applied to assess the reconstructed video quality. It is measured frame by frame, and then average MSSIM of all frames as the whole sequence quality.

The MSSIM of each frame is obtained by averaging all the 8×8 sliding windows. The window starts from the top-left corner of the frame, moves pixel by pixel horizontally and vertically through all the rows and columns of the frame until the bottom-right corner is reached, we can assess the distortion more critically in this special way, especially for the sequences coded with larger QP, in which block effect is usually obvious. The SSIM of the sliding 8×8 window is calculated as follow:

$$SSIM = 0.6 \times SSIM_y + 0.2 \times SSIM_u + 0.2 \times SSIM_v \qquad (9)$$

where $SSIM_y$, $SSIM_u$, and $SSIM_v$ represent the SSIM of the component y, u and v of the current block respectively.

The coding performances are compared in terms of output Bit/Pic and MSSIM of the reconstructed videos. Bit/Pic represents the average bit number per picture, and is obtained by averaging all the P-frames' bit numbers. MSSIM is the average of all the P-frames' SSIM. The comparison results are tabulated in table 1, 2 and 3. In the tables, IPBSS represents the proposed method and H.264 represents the original method

As is clearly shown in the tables, the proposed IPBSS has the maximum reduction in Bit/Pic of 34.37%, and the average reductions are 13.49%, 14.97% and 17.98% for QP=10, 20 and 30 respectively, while the MSSIM degradations are negligible compared with H.264. In other words, we can't distinguish the differences by eyes.

The coding time for most sequences is increased, because SSIM in IPBSS increases the computational load. But for the sequences with little motion and large static areas, like "akiyo" and "grandma", the proposed IPBSS can achieved both time saving and better compression result with QP=10. Further research will be done on reducing the computational load of the algorithm.

It is obvious that the IPBSS (proposed in this paper) has better compression performance than MEBSS (Proposed in our previous work [4]). Because the simulation results of MEBSS are better than H.264 only with small *QP*, while the proposed IPBSS can maintain good coding results even with larger QP. On the other hand, SSIM is a new image distortion measurement method, which has little relation with PSNR, and PSNR is not consistent with HVS, so the PSNR results are not included in the tables either.

5. CONCLUSION

In this paper we propose a novel inter prediction method based on structural similarity (IPBSS), which is an improved method of our previous work MEBSS. In IPBSS, we adopted SSIM as the distortion metric in the cost function of inter prediction for H.264. Simulation results demonstrate that the IPBSS can averagely save bit rate 13.49%, 14.97% and 17.98% with QP=10, 20 and 30 respectively compared with H.264, and the video quality is maintained at the same time. But the calculation complexity is increased clearly for most sequences. Further research will be done on reducing the computational load of the proposed IPBSS.

6. ACKNOLEDGMENTS

The work described in this paper was substantially supported by research project from National Natural Science Foundation of China [Project No. 60402015], and research project from Guangdong Natural Science Foundation of China [project No. 06025642].

7. REFERENCE

[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] 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, April 2004.

[3] Z. Wang. L. Lu, and A.C. Bovik, "Video quality assessment using structural distortion measurement," Proceedings of the IEEE International Conference on Image Processing, pp. 65-68, September 2002.

[4] Z.Y. Mai, C.L. Yang, K.Z. Kuang and L.M. Po, "A Novel Motion Estimation Method Based on

Structural Similarity for H.264 Inter Prediction," 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), vol1.2, pp. 913-916, 2006.5.

[5] http://iphome.hhi.de/suehring/tml/download/

Table 1. Result comparisons with parameter QP=10

Sequence	H.264			IPBSS			Results comparisons		
	Bit/Pic (bit)	MSSIM	Time (ms)	Bit/Pic (bit)	MSSIM	Time (ms)	△Bit/Pic (%)	△MSSIM(%)	△Time (%)
grandma	27012.73	0.9948	5136.12	24067.59	0.9941	4477.43	-10.90	-0.07	-12.83
carphone	39140.73	0.9951	6683.43	35928.82	0.9946	9300.94	-8.21	-0.05	54.84
coastguard	61746.45	0.9961	8396.41	59399.84	0.9958	32810.06	-3.80	-0.03	290.78
forman	51110.37	0.9954	6071.08	45490.94	0.9937	14571.51	-10.99	-0.17	140.03
apple	58116.41	0.9935	7133.55	57150.37	0.9934	12509.31	-1.72	-0.01	75.34
news	17034.78	0.9962	4063.80	12211.43	0.9951	5262.33	-28.32	-0.11	23.41
silent	20762.78	0.9962	4251.90	16992.33	0.9956	8606.82	-18.16	-0.07	115.71
trevor	39077.06	0.9966	4877.10	34685.71	0.9959	12653.10	-12.66	-0.07	159.44
akiyo	10358.20	0.9965	3279.39	6797.70	0.996	2988.55	-34.37	-0.05	-8.87
waterfall	214879.50	0.9965	24136.47	202541.60	0.9961	79464.57	-5.74	-0.04	229.23
Average							-13.49	-0.07	106.71

Table 2. Result comparisons with parameter QP=20

Sequence	H.264			IPBSS			Results comparisons		
	Bit/Pic (bit)	MSSIM	Time (ms)	Bit/Pic (bit)	MSSIM	Time (ms)	△Bit/Pic (%)	△MSSIM(%)	△Time (%)
grandma	3166.20	0.9881	3405.98	2420.08	0.9872	3510.90	-23.56	-0.09	3.08
carphone	8694.53	0.9848	5140.37	7249.31	0.9838	7038.25	-16.63	-0.10	36.93
coastguard	21590.53	0.9845	7860.69	19768.98	0.9821	32017.51	-8.44	-0.24	307.30
forman	11627.10	0.9816	6067.51	9463.84	0.9776	12158.43	-18.60	-0.41	100.37
apple	6466.61	0.9685	5275.18	6969.31	0.9689	6474.78	7.74	0.04	22.74
news	4852.57	0.9909	3556.16	3971.92	0.9897	4834.29	-18.15	-0.12	35.94
silent	5845.06	0.9902	3485.00	5249.63	0.9896	8231.86	-10.18	-0.06	136.21
trevor	12692.57	0.9879	4631.67	10781.71	0.9856	11586.08	-15.06	-0.23	150.15
akiyo	2418.78	0.9942	2801.00	1612.57	0.9933	2718.10	-33.32	-0.09	-0.03
waterfall	39083.76	0.9795	21617.33	33826.12	0.9787	49718.96	-13.45	-0.08	130.00
Average							-14.97	-0.14	92.27

Table 3. Result comparisons with parameter QP=30

Sequence	H.264			IPBSS			Results comparisons		
	Bit/Pic (bit)	MSSIM	Time (ms)	Bit/Pic (bit)	MSSIM	Time (ms)	△Bit/Pic (%)	△MSSIM(%)	△Time (%)
grandma	529.14	0.9818	2010.82	405.55	0.9804	4269.18	-23.44	-0.14	112.38
carphone	1932.25	0.9669	3717.80	1325.39	0.9625	9438.31	-31.42	-0.46	153.85
coastguard	4356.90	0.9364	6386.10	3889.80	0.9303	37393.45	-10.70	-0.65	485.54
forman	2055.35	0.9599	4325.90	1578.61	0.9531	16022.37	-23.16	-0.71	270.37
apple	936.82	0.9509	3186.78	929.63	0.9504	7602.29	-0.77	-0.05	138.53
news	1346.61	0.9798	2186.00	1149.55	0.9762	5846.29	-14.63	-0.37	167.44
silent	1691.59	0.9754	2586.10	1536.00	0.9742	9645.18	-9.17	-0.12	272.87
trevor	3439.84	0.9631	4058.08	2658.61	0.9529	16011.43	-21.93	-1.06	294.55
akiyo	498.29	0.9902	1784.67	352.00	0.9888	3096.27	-29.82	-0.14	73.49
waterfall	5140.74	0.9496	17134.29	4383.02	0.9470	67240.78	-14.74	-0.27	292.44
Average							-17.98	-0.40	226.15