Motion Estimation in DCT Domain[†]

Mingzhou Song, Anni Cai, *Member, IEEE* and Jing-ao Sun, *Member, IEEE*
Radio Engineering Department, Beijing University of Posts and Telecommunications

Internal P. O. Box 57
Beijing Univ. of Posts and Telecommunications
Beijing, 100088, P. R. China

E-mail: jingsun@bupt.edu.cn

***Abstract* — In this paper, a new block matching criterion is proposed for motion estimation in video coding. The new criterion is based on a comparison between the discrete cosine transform (DCT) coefficients of two blocks to be matched. Since the spectrum of DCT statistically concentrates on the neighborhood of the *dc* component and the number of non-zero coefficients is quite small, only few coefficients need to be considered when matching blocks according to the new criterion. Almost in all the video coding standards, DCT is a necessary step for spatial redundancy reduction. Thus the generated DCT coefficients can be utilized in the new criterion. However, there are still some blocks whose DCT coefficients are not available. Several algorithms of calculating DCT coefficients of these blocks are given. When this new criterion is combined with the logarithm search, promising results are produced.**

I. Introduction

In video coding, motion estimation plays a dominant role in bit rate reduction. Block matching is the most popular motion estimation technique, which assigns each block in a current frame a vector to represent the relative motion to the reference frame. The reference frame can be a previous or a future frame. In both MPEG-I [1] and MPEG-II [2], the block matching method has been adopted. The motion vectors of all blocks have to be extracted for each frame with block matching and this is the most time-consuming task in the video encoder. So algorithms to accomplish the block matching with low computational complexity while keeping sufficient video quality are always desired. However, most of the researches have paid much more attention to finding better searching strategies than matching criteria, in fact the latter can also contribute greatly to matching efficiency.

In this paper, a new block matching criterion is proposed to judge how well two blocks match with each other based on the mean square error or the mean absolute difference of the DCT coefficients of the two blocks. As the non-zero DCT coefficients statistically concentrate on the neighborhood of *dc* component and the number of non-zero coefficients is small, only few coefficients need to be considered in the new criterion. This will ease the burden of matching computation.

In all the video coding standards, such as H.261, JPEG, MPEG-I and MPEG-II, motion estimation is performed before DCT. If the order of the two procedures is reversed, the DCT coefficients can be utilized into motion estimation, so the re-calculation of DCT coefficients can then be made ready for use in the proposed matching criterion. However, there do not exist ready-made DCT coefficients of all the candidate blocks in the reference frame. Only those blocks both horizontally and vertically located at integral multiple of the block size have ready-made DCT coefficients while the others do not. We call the former aligned blocks, the latter non-aligned blocks. For the non-aligned blocks, the DCT coefficients can be derived directly by DCT from samples in spatial domain or indirectly by the known DCT coefficients of the adjacent aligned blocks. Kou et al. and Chang et al. discussed the indirect approaches [3],[4], which will be reviewed in section III. Nevertheless, non-aligned blocks need extra computations which may scratch off the efficiency gained by using the new criterion. So searching strategy of motion estimation should try to work on more aligned blocks and less non-aligned blocks. Logarithm search [1] is one of the strategies that can be considered.

In section II, the new matching criterion in DCT domain is discussed. In section III, the algorithms of calculating DCT coefficients of non-aligned blocks are listed. In section IV, the case utilizing the new criterion in logarithm search is studied. In section V, performance and experimental results are analyzed. The conclusion is given in section VI.

II. Block Matching Criterion In DCT Domain

Two kinds of block matching criteria are widely used. One is MSE criterion, the other is MAD criterion. They are expressed by Eq. (1) and Eq. (2) [5], respectively.

$$\mathrm{MSE}(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[s_k(m,n)-s_{k-1}(m+i,n+j)]^2 \quad (1)$$

$$\mathrm{MAD}(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}|s_k(m,n)-s_{k-1}(m+i,n+j)| \quad (2)$$

where M is block width, N is block height, $s_k(m,n)$ is the value of the pixel located at (m,n) in a block of k-th frame, (i,j) is the relative displacement between a block in the k-th frame and a block in the (k-1)-th frame. We propose the following criteria in DCT domain,

[†] Work supported by National 863 Program on Communications.

$$D(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}[u_k(m,n)-u_{k-1}(m+i,n+j)]^2 \qquad (3)$$

or,

$$D(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|u_k(m,n)-u_{k-1}(m+i,n+j)\right| \qquad (4)$$

where $u_k(m,n)$ is value of the DCT coefficient located at (m,n) in a block of the k-th frame, M, N, (i,j) have the same meanings with those in (1), (2).

Since DCT can lead to spectrum concentration when the signal can be recognized as Markov process, we only need consider the concentrated part, i.e. the few DCT coefficients around *dc*. Thus, matching computation between two blocks can be greatly reduced.

Further considering that because human visual system (HVS) responses to DCT coefficients differently, we can add a visual weighting factor $\alpha(m,n)$ to each coefficient. $\{\alpha(m,n)\}$s should be determined by physiological experiments. The new criterion can then be modified as,

$$D(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\{\alpha(m,n)[u_k(m,n)-u_{k-1}(m+i,n+j)]\}^2 \qquad (5)$$

or

$$D(i,j)=\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|\alpha(m,n)[u_k(m,n)-u_{k-1}(m+i,n+j)]\right| \qquad (6)$$

In the following sections, we will refer the new criterion in DCT domain to (5) or (6) for generality.

III. Calculating Non-aligned DCT Coefficients

For the new criterion, the DCT coefficients of any two blocks to be matched should be calculated before matching. We call the block in current frame target block, the block in reference frame candidate block. A target block is supposed to be an aligned block, as show in Fig. 1. A candidate block can either be an aligned block or a non-aligned block inside the search range (see Fig. 2).

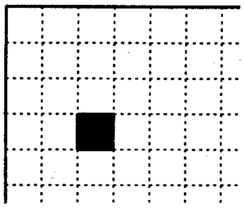


Fig. 1. Target block position in current frame

Since DCT is a necessary step almost in all current video coding standards, including H.261, motion JPEG, MPEG-I and MPEG-II, though the purpose is totally different with what we intend, this can be taken advantage of. If we reverse the motion estimation and DCT in conventional scheme, the DCT coefficients of target blocks and those of aligned candidate blocks are ready-made. The conventional scheme and the new scheme are illustrated in Fig. 3. The new scheme is different with the conventional one in a) DCT is done before motion estimation; b) there is no necessity to perform inverse DCT; c) there must be another function module to generate DCT coefficients for non-aligned blocks.

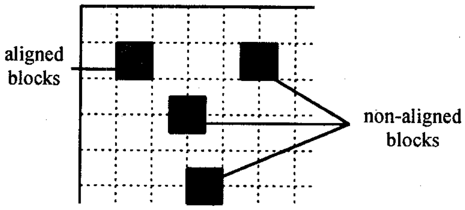
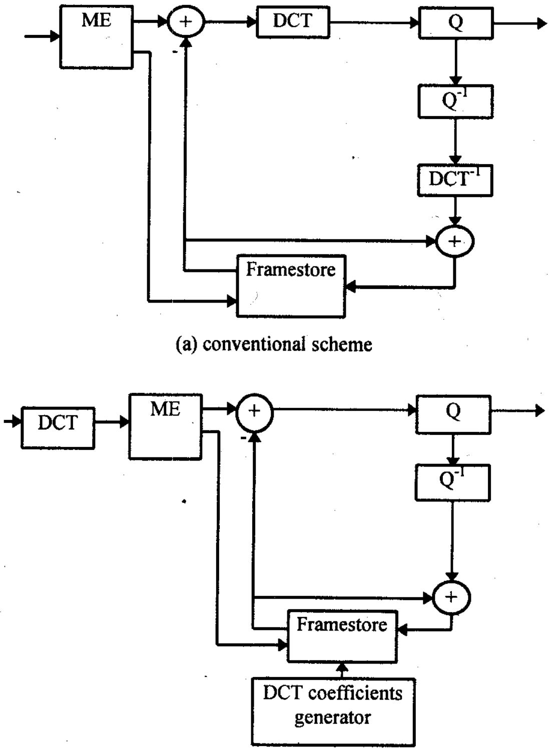


Fig. 2. Candidate blocks in reference frame

DCT coefficients of non-aligned blocks can be computed directly from spatial domain. In this way, the spatial pixel values must be kept in framestore, which will increase the burden of storage. Another way is indirect computation from the DCT coefficients of adjacent aligned blocks. Three approaches can be used for indirect computation, in which the spatial pixel values of all blocks are assumed unavailable.



(a) conventional scheme

(b) new scheme

Fig. 3. Illustration of the two different schemes performing ME and DCT. ME: Motion estimation. Q: Quantizer

A. Approach 1

This approach is quite straight forward. In Fig. 4, B_1 - B_4 are aligned candidate blocks, whose DCT coefficients are in framestore. B' is a non-aligned candidate block, whose DCT coefficients are to be calculated. We perform inverse DCTs of B_1 - B_4 to obtain spatial pixel values of these blocks. So DCT coefficients of B' and all the non-aligned blocks inside the four adjacent aligned blocks can then be calculated by forward DCT.

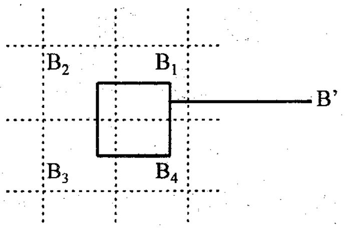


Fig. 4. B_1 - B_4 are aligned blocks, B' is non-aligned block

B. Approach 2

Kou and Fjällbrant presented this approach in [4]. But there is a prerequisite for their approach, i.e., the overlap length between the aligned block and the non-aligned block must be half of the block size. This prerequisite is so strict that the approach can be applied to merely a small number of non-aligned blocks efficiently. In addition, Kou gave the computation analysis in one dimensional case. Although this approach outperforms *Approach 1*, the reported improvement is fairly small.

C. Approach 3

Chang and Messerschmitt[3] proposed this approach to compose multiple video streams into a new one, with operations on each single video stream. From Fig. 5, we can see that the DCT of B' is,

$$DCT(B') = DCT(B_{13}) + DCT(B_{24}) + DCT(B_{31}) + DCT(B_{42}) \tag{7}$$

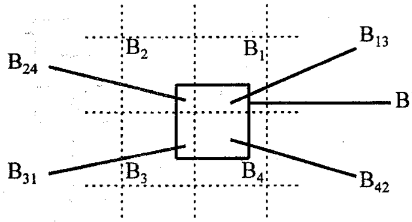


Fig. 5. B' is made up by sub-block B_{13}, B_{24}, $B_{31,}$ B_{42}. Each sub-block is a rectangle portion of an aligned block.

The sub-blocks composing B' can be represented by the following matrix form, here we only take B_{42} as illustration,

$$B_{42} = H_1 B_4 H_2 \tag{8}$$

where,

$$H_1 = \begin{bmatrix} 0 & 0 \\ I_h & 0 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 0 & I_w \\ 0 & 0 \end{bmatrix}$$

where I_h and I_w are identity matrices with size h × h and w × w, respectively. h and w are the number of rows and columns extracted. Since DCT is distributive to matrix multiplication,

$$DCT(B_{42}) = DCT(H_1B_4H_2) = DCT(H_1)DCT(B_4)DCT(H_2) \tag{9}$$

In the same manner, the DCT of B_{13} , B_{24} and B_{31} can be obtained. Substituting them into (7), we get,

$$DCT(B') = \sum_{i=1}^{4} DCT(H_{i1})DCT(B_i)DCT(H_{i2}) \tag{10}$$

The DCT of H_{i1} and H_{i2} can be pre-computed and stored in memory. (2N-2) matrices need to be stored. The required computation can be reduced by using sparse matrix multiplication techniques since many DCT coefficients are zero.

IV. Steps Of Logarithm Search And Improvement

As mentioned in section III, both the direct way and indirect way of calculating the DCT coefficients of the non-aligned blocks consume additional time and space, which will somewhat scratch off the gain obtained by introducing the new criterion. Under such circumstances, we should use a searching strategy which can minimize the frequency of matching target blocks to non-aligned candidate blocks.

The simplest search strategy is the full search. Within the chosen search range, all possible displacements are evaluated using a given block matching criterion. This procedure involves much more matching to non-aligned blocks than to aligned blocks. When using the new matching criterion, the overhead of computing DCT coefficients is considerably formidable.

A faster searching method is logarithm search. originally proposed by Koga et al[6]. In this search method, grids of 9 displacements are examined, and the search continues based on a smaller size by a factor of 3 at each step. Then the search is maximally efficient in the sense that any integer shift has a unique selection path to it. The search steps are shown in Fig. 6. To be more robust, the scaling factor can be adjusted to match the search ranges. Some possible grid spacings for various search ranges are given in TABLE I. The search ranges in this table are referred to MPEG [1].

In MPEG[1], the DCT block size is 8 by 8. Note that all the 9 matchings in one step will be performed on aligned blocks if the grid spacing is integral multiple of block size and the center of the grid is set at the aligned block. This property is very suitable for the new criterion in DCT domain. From TABLE I we can see that, the larger the search range is, the larger proportion of aligned blocks there will be, and the larger computational improvement we can get by using the new criterion.

```
* * * * * * * * * * * * * * *
* * * * * * * * * * * * * * *
* * * * * * * * * * * * * * *
* * * 1 * * * 1 * * * 1 * * *
* * * * * * * * * * * * * * *
* * * * * * * * * * * * * * *
* * * * * * * * * * * * * * *
* * * 1 * * * 1 * * * 1 * * *
* * * * * * * * * * * * * * *
* * * * * 2 * 2 * 2 * * * * *
* * * * * * * * * * * * * * *
* * * 1 * 2 * 1 * 2 * 1 * * *
* * * * * * * * 3 3 3 * * * *
* * * * * 2 * 2 3 2 3 * * * *
* * * * * * * * 3 3 3 * * * *
```

Fig. 6. Illustration of logarithm search. First, the blocks at the positions marked by 1 are matched to the target block. The best one, assumed at lower middle 1, is determined. Then reducing the grid spacing by half, we continue to match the blocks at the positions marked by 2 to target block, the best match at lower right 2 is obtained. At last the eight blocks at positions marked by 3 and the block at lower right 2 are matched. The best match block of this time is considered to be the final match block of the target block.

TABLE I

Search range	steps	grid spacings
± 7.5	4	4, 2, 1, 1/2
± 15.5	5	8, 4, 2, 1, 1/2
± 31.5	6	16, 8, 4, 2, 1, 1/2
± 63.5	7	32, 16, 8, 4, 2, 1, 1/2
± 127.5	8	64, 32, 16, 8, 4, 2, 1, 1/2
± 255.5	9	128, 64, 32, 16, 8, 4, 2, 1, 1/2
± 511.5	10	256, 128, 64, 32, 16, 8, 4, 2, 1, 1/2
± 1023	10	512, 256, 128, 64, 32, 16, 8, 4, 2, 1

In MPEG, the size of motion estimation block is 16 by 16, which is called macroblock. When using the new criterion, as shown in Fig. 7, we divide a macroblock into four blocks — each one is a DCT block. The matching criterion for two macroblocks is defined as the summation of differences of the four pairs of corresponding DCT blocks. So the problem of matching two macroblocks is converted to the problem of matching DCT blocks. Our previous discussion still holds.

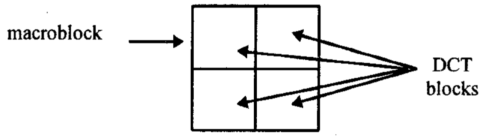


Fig. 7. DCT block and macroblock

V. Performance Analysis

A. Computational Complexity

The computation of motion estimation in DCT domain depends on, a) how many blocks are compared for matching? b) how many non-aligned blocks are there in the total blocks compared? c) how many operations are there to get DCT coefficients of non-aligned blocks? d) how many DCT coefficients are used in one matching of two blocks?

a) is determined by the searching strategy used. Assume search range is X by X and motion vector has a precision of one pixel. Total X^2 blocks should be compared for full search, only $9\log_2 X$ blocks for logarithm search.

b) also relies on the search strategy. In full search, only $(X/8)^2$ blocks are aligned blocks and $X^2 - (X/8)^2$ blocks are non-aligned blocks. When we use logarithm search, there are $9(\log_2 X - 3)$ aligned blocks and 9×3 non-aligned blocks. If $X \geq 16$, the proportion of the number of aligned blocks to non-aligned blocks is larger than 1:3.

c) is tightly related to the way of generating DCT coefficients of non-aligned blocks. The number of arithmetic operations of different approaches in section III are shown in TABLE II. We assume the block size is 8 by 8. The fast DCT algorithm is based on that of Chen et al. [7].

TABLE II The number of arithmetic operations of different ways to generate DCT coefficients of a non-aligned block. Approach 2 is ignored because of its limited applicable range.

	direct way	indirect way Approach 1	indirect way Approach 3 [3]
multiplication	88	176	$1024(1/\beta + 1/\sqrt{\beta})$
addition	232	464	$1024(1/\beta + 1/\sqrt{\beta})$ $+192$

where β is the ratio between the total number of the DCT coefficients and the number of non-zero DCT coefficients. β can be 64/k, (k: 1,2, ..., 64), or ∞ (when all DCT coefficients are zero).

The direct way has the least computational complexity. Approach 3 depends heavily on β. Only when β is larger than 32, the performance of Approach 3 will become near Approach 1.

d) is determined by the criterion. In conventional criteria, all pixel values of a block in spatial domain must be considered, because they are of equal importance. But in DCT domain, the energy concentrates around *dc* and HVS is not sensitive to variations of DCT coefficients far from *dc*, thus only K < MN coefficients need to be calculated in the new criterion. The time gain is MN/K.

B. Matching Accuracy

We analyzed the performance of the new criterion by computer simulation. The searching strategy used is logarithm search and is within a ±64 searching range. The accuracy of motion vector is one pixel. The searching steps and spacing of grid are in light of TABLE I. In the first three steps of searching, following the order of zigzag[1], a designated number of DCT coefficients are taken into account in the new criterion. In the last three steps of searching, the MSE criterion is still used in order to avoid the overhead of computing the DCT coefficients of non-aligned blocks. After all the motion vectors of the blocks in the image are determined, the residue image is calculated by subtract the original image by the motion-compensated prediction image. Then DCT is performed on the residual image (This computation can be done directly by using the scheme shown in Fig. 3 (b)). The DCT coefficients are entropy coded with adaptive arithmetic coding in this instance.

Reconstructed image is obtained by decoding the compressed bit stream. Then the MSE and Peak Signal to Noise Ratio (PSNR) of the reconstructed image to the original image are calculated.

In TABLE III, the results are given when the number of considered DCT coefficients varies in the new criterion. The total bytes of the compressed residual image are also listed. The experiments are performed on the 12-th, 13-th luminance images of Susan sequence. The image dimension is 720x576x8.

From Table III we see that performance in the worst case of using only one DCT coefficient degrades 0.6dB from the best case when all 64 coefficients are used. During subject observation, the difference between these two cases can be perceived only by very careful comparison. When 43 coefficients are used, both PSNR and the total bytes of compressed image are in the proximity of the best case. During the subject observation, the fine shade can hardly be discriminated. Note that the result in the best case should be the same with that when using the conventional MSE criterion (Practically there may be subtle variance because of effect of limited word length).

So, the new criterion has strong practicability.

TABLE III Results of experimentation

the number of DCT coefficients used in the new criterion	MSE	PSNR (dB)	total bytes of compressed residual image (bytes)
1	12.74	37.05	12093
3	12.30	37.20	11281
6	12.11	37.26	10965
10	11.97	37.31	10778
15	11.90	37.34	10591
21	11.77	37.39	10486
28	11.57	37.46	10395
36	11.27	37.58	10120
43	11.10	37.64	10062
49	11.13	37.63	10025
54	11.11	37.63	10034
58	11.09	37.65	10002
61	11.08	37.65	10018
63	11.07	37.65	10001
64	11.08	37.65	9998

VI. Conclusion

This paper addresses the feasibility of changing conventional MSE or MAD criteria to a new proposed DCT domain criterion for block matching method in motion estimation. In this approach, the generation of DCT coefficients of non-aligned DCT blocks is a critical task. When using the aforementioned indirect approaches in section III, we can make a compromise between time and space. The logarithm search is studied as an example into which this new criterion can be applied. The performance shows a considerable improvement in matching.

Although we have only studied the new criterion in DCT domain, it is theoretically possible to explore it in other transform domains.

Acknowledgment

The authors greatly appreciate Prof. Rongshu Gong's invaluable advice.

References

[1] Draft International Standard MPEG-I, ISO/IEC JTC 1/SC 29 N 071, CD11172, Dec. 1991.

[2] Draft International Standard MPEG-II, ISO/IEC DIS 13818, 1994.

[3] Shih-Fu Chang and David G. Messerschmitt, "Manipulation and compositing of MC-DCT compressed video", IEEE J. Select. Areas Commun., Vol. 13, no. 1, pp. 1-11, Jan. 1995.

[4] Weidong Kou and Tore Fjällbrant, "A direct computation of DCT coefficients for a single block take from two adjacent blocks", IEEE Trans. on Signal Processing, Vol. 39, no. 7, pp. 1692 - 1695, July 1991.

[5] H. G. Musmann, P. Pirsch and H-J. Grallert, "Advances in picture coding", Proceedings of the IEEE, Vol. 73, no. 4, pp. 523 - 548, April 1985.

[6] T. Koga, K. Iinuma, A. Hirano, Y. Iijima and T. Ishiguro, "Motion-compensated interframe coding for video conferencing", in NTC 81 proc., pp. G5.3.1 - G5.3.5, 1981.

[7] W.-H. Chen, C. H. Smith and S. C. Fralick, "A fast algorithm for the discrete cosine transform", IEEE Trans. on Commun., vol. COM-25, pp. 1004 - 1009, Sept. 1977.