國立交通大學

資訊科學系

碩 士 論 文

在H.264視訊編碼標準下基於全域運動估計之快速多重畫面選擇法

A fast method for multiple reference frame selection based on global motion estimation for H.264 video coding

研 究 生：黃合吉

指導教授：陳玲慧 教授

中華民國九十四年六月

在 H.264 視訊編碼標準下基於全域運動估計之快速多重畫面選擇法

A fast method for multiple reference frame selection based on global motion estimation for H.264 video coding

研 究 生：黃合吉　　Student：Ho-Chi Huang

指導教授：陳玲慧　　Advisor：Ling-Hwei Chen

國 立 交 通 大 學
資 訊 科 學 系
碩 士 論 文

A Thesis
Submitted to Department of Computer and Information Science
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer and Information Science

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

在 H.264 視訊編碼標準下基於全域運動估計之快速多重畫面選擇法

研究生：黃合吉　　　　　　　　　　　　　　　　指導教授：陳玲慧　博士

國立交通大學資訊科學研究所

中文摘要

最新的視訊壓縮標準 H.264/MPEG-4 Part10 提供新的功能使得在運動估計的錯誤能夠減少。其中一項是多重參考畫面。在 H.264 參考軟體 JM8.1 裡，因為對於運動估算採用暴力法,所以隨著參考畫面的張數增加整個編碼流程在運動估算部份所花費的時間會呈現線性增加。在此我們將提出一個新的演算法使得挑選參考畫面能夠更加迅速。由於在影片中相鄰兩畫面背景通常都是極為相似的，所以我們利用此特性對於欲編碼的畫面利用全域運動估計法找出兩畫面的背景部份，凡落在此背景上的區塊，我們直接取其前一畫面當作參考畫面。根據實驗結果顯示利用本方法在運動估算這一部份可以省掉約 13%到 72%的計算時間，卻能夠維持在 3%以下的畫面選擇失誤率。而且位元率僅增加 1%。

A fast method for multiple reference frame selection based on global motion estimation for H.264 video coding

Ho-Chi Huang and Ling-Hwei Chen

Department of Computer and Information Science, National Chiao Tung University
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.

Abstract

The new video coding standard, H.264/MPEG-4 Part10, provides some new features to reduce prediction error in the motion estimation. Using multiple reference frames is one of these features. In the reference software JM8.1, the exhaustive search is needed for multiple reference frame selection, it is very time consuming. In general, the backgrounds between two consecutive frames are almost the same, thus for each macroblock we can usually find suitable matching macroblock in the nearest neighboring frame. Based on this fact, we will propose a fast method to find the optimal reference frame. First, a global motion estimation technicality is provided to get the global motion vector of the current frame relative to the nearest neighboring frame. Then based on the global motion vector obtained, those macroblocks on the background of the current frame can be detected. And the nearest neighboring frame is considered directly as their reference frame. Simulation results show that the proposed algorithm can save about 13% to 70% searching time while keeping the average miss rate of optimal frame less than 3%. And the average bit-rate raises only less than 1%. Furthermore, we also compare the efficiency between our proposed algorithm and an exiting algorithm proposed by Huang et al. The experimental result shows that our algorithm is superior to Huang's.

致謝

首先對於我的指導教授 陳玲慧老師獻上最真誠的感謝，在她細心、耐心的教導之下，讓我能夠體驗到學習的樂趣與研究的精神。

此外，還要感謝自動化資訊處理實驗室的學長們，在我這兩年研究所的學涯上，給予我學習上的指導與協助，讓我能夠順利完成研究所的學業。其中一起畢業的同學，邵育睿、陳致生、翁崇芢、王朝君與李佳峰，大家在課業及日常生活上的互助，讓我能夠順利地完成研究所的課程。另外還要感謝在我身邊支持的人，特別是我的女朋友江妤璇總是在我失落低潮時為我加油打氣。

最後，由衷地感謝我的父母與哥哥們多年來給我的關懷與栽培，使我得以專心致力於研究上，僅以我最誠摯的心意將此篇論文獻給我的父母與家人。

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

H.264 is the newest video coding that is instituted by Video Coding Experts Group (VCEG) and ISO/IEC Motion Picture Experts Group (MEPG). It provides better quality than MPEG-4 advanced simple profile (MPEG-4/ASP). The main improvement in H.264/AVC compared to MEPG-4/ASP is the prediction module [1], including variable block size, quarter-pixel motion search, in-loop deblocking filter, multiple reference frames and so on. Although the compression ratio is raised, it costs a lot of encoding time. Thus, some researchers developed several algorithms to speed up the encoding procedure. In this paper, we will focus on multiple reference frame searching. The motion estimation (ME) using multiple reference frames was introduced in [2] to extend the motion search range by using multiple decoded frames as the possible reference frames, instead of only using the nearest decoded frame. The use of multiple frames for ME always provides significantly improved predicated gain [2]. Fig. 1 from reference [3] shows that the motion vectors of different blocks in the current frame could be obtained from different prior-decoded frames [3]. This is due to that some blocks [see Fig. 2 from reference [4]], which contain moving objects, could not find a suitable motion vectors in the nearest reference frame. Multiple reference frame selection provides a way to solve this problem and reduces prediction error. Because the H.264 encoder reference software (JM8.1[5]) adopts exhaustive search for reference frame selection to get an optimal reference frame, the computational time is very expensive. Therefore, there are some algorithms [6-7] proposed to reduce the time complexity while keeping similar video quality.

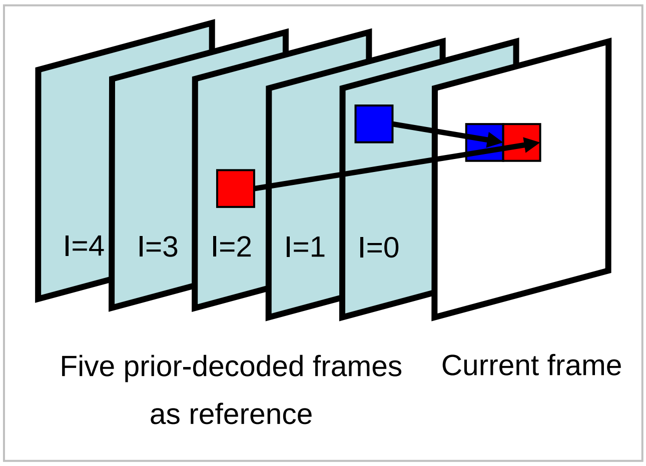


Fig. 1 Motion estimation using multiple reference frames. "I" is the index of the reference frame.

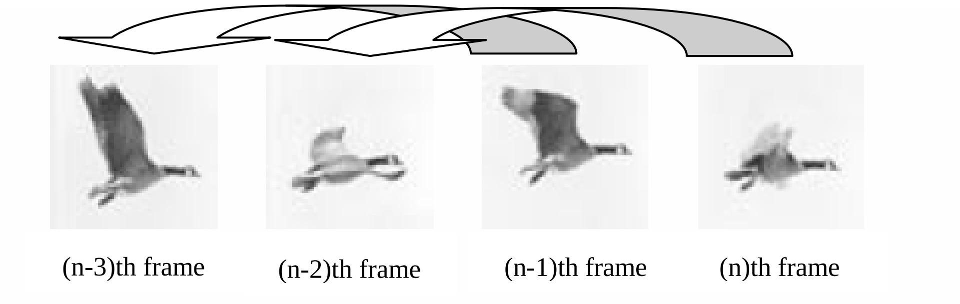


Fig. 2 An example for H.264 motion estimation using multiple reference frames.

Huang et al.[7] proposed a method to speed up the search time in H.264. Because H.264 in inter prediction provides variable-block-size modes which include 16x16, 16x8, 8x16, 8x8, 8x4, 4x8, and 4x4 mode, the proposed method was developed based on some information about the statistics of inter selected mode, motion vectors obtained by using the nearest frame as the reference one, and residues. The main concepts in [7] are as follows. First, for a macroblock (MB), if 16x16 block-size mode is selected after ME at the nearest reference frame, the optimal reference frame tends to be the nearest reference frame. Second, if inter-modes with smaller blocks are selected, searching more reference frames tend to be helpful. Third, if motion vectors of larger blocks are similar to motion vectors of smaller blocks, it is likely that no occlusion or uncovering occur in the MB, so referring to the nearest reference frame may be enough. Fourth, if motion vectors of larger blocks are more different from motion vectors of smaller blocks, the MB may contain moving object boundaries and thus requires more reference frames. Finally, if the texture of a MB is very complicated, it may require more reference frames. In [7], the average peak signal to noise ratio (PSNR) drop is 0.05 dB, the average miss detection rate is 3.9%, and the ME computational time can be saved 10%-67%.

In this thesis, we propose an efficient approach to ME in the multiple reference frames selection. From our observation, we find the fact that for each block on the background, we can usually find a good motion vector in the nearest reference frame. However, for a block containing a moving object, a good match may not appear in the nearest reference frame. Based on the fact, we will develop a method to quickly decide the reference frame used for each block. The experimental results in our method show that the proposed method can save 15%-72% searching time of the ME while keeping close to the video quality obtained by the exhaustive method. And the average miss detection rate is less than 3%.

The paper is organized as follows. In Chapter 2, some observation and analysis will be discussed. The proposed method will be described in Chapter 3. Experimental results will be given in Chapter 4. Finally, the conclusion will be made in Chapter 5.

CHAPTER 2

Observation and Analysis

When we use camera to produce a video, we will find three facts. The first is that most part of a background will keep unchanged for at least several seconds. The second is that a background will be still when a camera is fixed (see Fig. 3(a) and Fig. 3(b)) and some backgrounds will be moving smoothly due to the camera motion (see Fig. 3(c) and Fig. 3(d)). The third is that a background usually occupies a large space in the video sequence.

From our observation, we also found two phenomena; one is that those blocks belonging to the background usually have the same motion vector as the global motion vector of the current frame referred to the nearest reference frame. The other is that the blocks belonging to the background usually refer to the nearest reference frame.

Here, we will do some experiments to verify the above-mentioned facts and phenomena that the global motion vector will dominate the frame motion and many blocks belonging to the background only need to refer to the nearest reference frame.

In this paper, our experiments are conducted with nine CIF format video sequences shown in Fig. 4. Some video sequences like “Container”, “Silent”, “Mother and Daughter”, and “Hall Monitor” contain the still background, some sequences like “Foreman”, “Coastguard”, and “Stefan” contain moving backgrounds due to camera motion of translation, and backgrounds in the other two sequences have zooming motion. The video format is CIF, and its dimension is 352x288. 100 frames are analyzed and experimented. And we use the H.264 reference software, JM8.1, to encode these video sequences.


(a) (b)

(c) (d)

Fig. 3 Two examples to show still and moving backgrounds (a) The 1st frame in "Silent" sequence, (b) the 84th frame in "Silent" sequence with the still background. (c) The 66th frame in "Stefan" sequence, (d) the 78th frame in "Stefan" sequence with the moving background.


Container


Silent


Mother and Daughter


Hall Monitor


Foreman


Coastguard

Stefan


Mobile Calendar

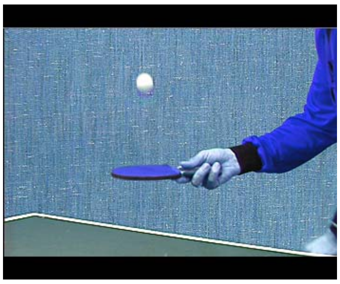


Table Tennis

Fig. 4 The video sequences used for analysis and simulation. (format : CIF 352x288, 100 frames)

2-1 The referred rate of each reference frame used as the best reference frame

All video sequences in Fig. 4 are tested under the following conditions: Quantization step size is 30, the number of encoding frames is 100, and the number of reference frames is 5. Table 1 shows the referred rate of each reference frame used as the best reference one.

We can see that for all video sequences, the most referred reference frame is the nearest reference frame (i.e. Ref0). The average referred rate in Ref0 is up to 80%, and the referred rate of the other reference frames is frequently below 20%. Especially those video sequences with the still background or camera translation motion like “Container”, “Hall Monitor”, “Mother and Daughter”, and “Silent” have higher referred rate in Ref0. Note that the referred rate in Ref0 of the sequence, Mobile Calendar, is 46% and lower than those of the other video sequences. This is due to that the sequence has the zooming motion and its texture is complicated.

Due to the above observation, our proposed method will adopt exhaustive search in the nearest reference frame. After searching the nearest reference frame, we will employ some available information to determine if it is helpful to further search more reference frames.

Table 1 The referred rate of each reference frame considered as the best reference frame. Ref0 is the nearest reference frame, Refi is the ith reference frame, and Intra stands for the intra prediction.

	Ref0	Ref1	Ref2	Ref3	Ref4	Intra
Container	92%	3%	2%	1%	1%	0%
Silent	94%	2%	1%	0%	1%	2%
Mother and Daughter	94%	2%	2%	1%	1%	0%
Hall Monitor	90%	4%	3%	1%	1%	1%
Foreman	79%	10%	6%	2%	2%	1%
Coastguard	88%	5%	3%	2%	1%	1%
Stefan	75%	10%	8%	3%	2%	2%
Mobile Calendar	46%	16%	17%	11%	9%	0%
Table Tennis	84%	6%	4%	2%	2%	4%
Average	83%	6%	5%	3%	2%	1%

2-2 The variances of the motion vectors

First, the encoder adopts exhaustive search under the following conditions: The reference frame number is 5 and the block size is 4x4. Then we divide all blocks in the current frame into two classes, A and B. Class A contains those blocks with their best motion vectors found in the nearest reference frame. Class B contains those blocks with the best motion vectors not found in the nearest reference frame. Finally, we define three variances : Var_A, Var_B, and Var_C as follows:

For each 4x4 block, let (x_i, y_i) be the best motion vector found in the nearest reference frame.

$$Var_A = \frac{\sum_{i \in A}\left(x_i - \overline{X_A}\right)^2 + \sum_{i \in A}\left(y_i - \overline{Y_A}\right)^2}{N_A}$$

$$Var_B = \frac{\sum_{i \in B}\left(x_i - \overline{X_B}\right)^2 + \sum_{i \in B}\left(y_i - \overline{Y_B}\right)^2}{N_B}$$

$$Var_C = \frac{\sum_{i \in A}\left(x_i - \overline{X}\right)^2 + \sum_{i \in A}\left(y_i - \overline{Y}\right)^2 + \sum_{i \in B}\left(x_i - \overline{X}\right)^2 + \sum_{i \in B}\left(y_i - \overline{Y}\right)^2}{N}, \text{ and}$$

$$\overline{X_A} = \frac{\sum_{i \in A} x_i}{N_A}, \overline{Y_A} = \frac{\sum_{i \in A} y_i}{N_A}$$

$$\overline{X_B} = \frac{\sum_{i \in B} x_i}{N_B}, \overline{Y_B} = \frac{\sum_{i \in B} y_i}{N_B}$$

$$\overline{X} = \frac{\sum_{i \in A} x_i + \sum_{i \in B} x_i}{N}, \overline{Y} = \frac{\sum_{i \in A} y_i + \sum_{i \in B} y_i}{N}$$

where N_A is the block number in Class A, N_B is the block number in Class B, and N is the sum of N_A and N_B. Here the resolution of the motion vectors in the analysis is set to be 1/4 pixel. 1/4 pixel is considered as a unit.

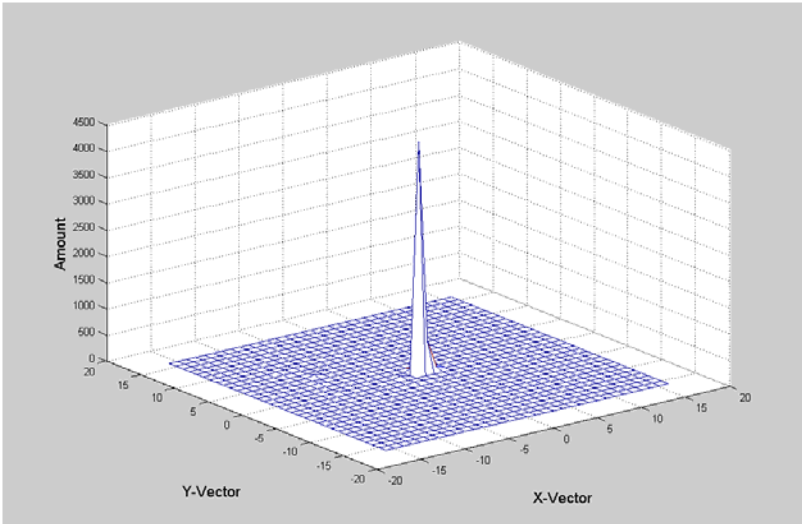
Table 2 shows the average variances of each kind in video sequences. We can see that Var_A is smaller than Var_B, and Var_C is smaller than Var_B. It means that motion vectors in Class A tend to be concentrated. However, the motion vectors in Class B tend to disperse. Based on the observation, we will further analyze the distribution of motion vectors in each frame.

Table 2 The average of three kinds of variances of three types in video sequences.

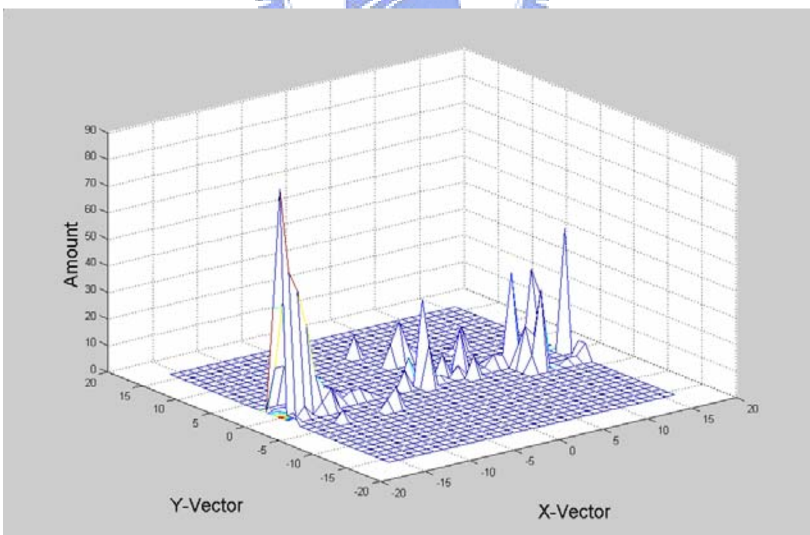
Sequences	Var_A	Var_B	Var_C
Container	6.5	6.5	8.0
Silent	49.9	234.1	54.3
Mother and daughter	18.8	43.0	20.6
Hall Monitor	12.5	65.8	20.6
Foreman	37.4	49.6	42.5
Coastguard	25.4	61.7	30.5
Stefan	180.9	237.1	192.8
Mobile calendar	41.6	108.3	81.6
Table Tennis	117.1	260.1	122.8
Average	54.5	118.5	63.7

2-3 The distribution of motion vectors

Here, we will consider the case about the distribution of the motion vectors. Fig. 5(a) shows the distribution of the motion vectors in Class A, and Fig. 5(b) shows the distribution of the motion vectors in Class B. In Fig. 5(a), we can see most motion vectors centralize locally and motion vectors not centralized are very rare relatively. However, in Fig. 5(b), the motion vectors of Class B which don't refer to the nearest reference frame don't centralize locally. This phenomenon is due to that most blocks in Class A belong to the frame background. And the centralized motion vector is the same as the global motion vector of the current frame that dominate the motion vectors of many blocks in the current frame. Based on the above observations, we can check if the best motion vector of a block in the nearest reference frame is similar to the global motion vector between the current frame and the nearest reference frame to determine if it is helpful to search other reference frames.



(a)



(b)

Fig. 5 (a) The 65th frame in “Silent” sequence. The distribution of motion vectors of Class A, (b) The 10th frame in “Stefan” sequence. The distribution of motion vectors of Class B.

CHAPTER 3

The proposed method

We have analyzed some information about the referred rate of each reference frame used as the best reference frame, the variances of motion vectors and the distribution of motion vectors. Let us summarize above analysis as follows. First, the nearest reference frame is frequently referred by the current frame. Second, the best motion vectors from Class A refer to the nearest reference frame tend to be concentrated. Third, there is a global motion vector that dominates the motion vectors of blocks in the current frame. According to the above analysis, we will propose a fast approach for multiple reference frame selection. Before describing the proposed method, we will introduce an efficient method to find global motion vector for the current frame.

3-1 The global motion vector finding algorithm

Here, we will describe how to obtain the global motion vector quickly for the current frame. First, if the global motion vector is obtained by the exhaustive search method, its complexity is heavy, thus, we will provide an efficient algorithm to obtain the global motion vector.

The algorithm mainly utilizes the multiresolution concept. The use of multiresolution representation for image processing was first introduced by Burt and Adelson [8]. Here, we will use pyramid structure. First, the current frame and the nearest reference frame will be downsampled using a low-pass filter, e.g. a 2x2 averaging filter.. The ratio and order of a downsampled frame is shown in Fig. 6.

When the search range is set to 16 pixels, we use 5 levels and let the 5th level be the original frame. Let the l th level frame of the current and nearest reference frames be represented by $\varphi_{t,l}(x)$, $x \in \Lambda_l$, t= 1, 2, 1 stands for the current frame and 2 for the nearest reference frame, where Λ_l is the set of pixels at level l. Denote the global motion vector of level $l-1$ as d_{l-1}. At the l th level, we use d_{l-1} to produce an initial motion estimate $\tilde{d}_l = 2(d_{l-1})$, where the number "2" represents the scalar between two successive levels. We then find the offset q_l in a given search window such that the estimation error [9]

$$\sum_{x \in \Lambda_l} | \varphi_{2,l}(x + \tilde{d}_l + q_l) - \varphi_{1,l} |$$

is minimized. And the search range in the above formula is set to 1 pixel. The new global motion vector obtained after this step is $d_l = q_l + \tilde{d}_l$. According to the above scheme, the global motion vector at the finest resolution is

$$d_5 = q_5 + 2(q_4 + 2(q_3 + 2(q_2 + 2(q_1 + d_0)))) .$$

The global motion vector of level 0 for this procedure is set to 0, i.e. $d_0 = 0$. By the above method, we can obtain the finest global motion vector between the current frame and the nearest reference frame.

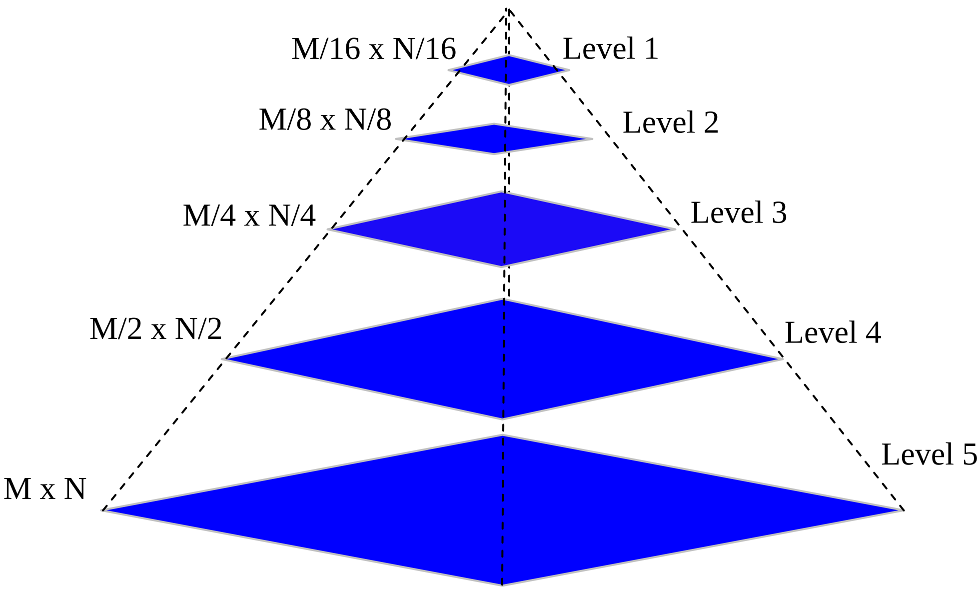


Fig. 6 The pyramid structure of the multiresolution ME. “M” is the width of an original frame, and “N” is the height of an original frame.

3-2 The fast reference-frame selection method

After obtained the finest global motion vector, we propose an algorithm to reduce searching time for multiple reference frame selection. The steps of the algorithm are described in Fig. 8. Here, $Diff(MV_{block})$ means $| MV_{block} - GMV_{i\text{th frame}} |$. Note that MV_{block} is the motion vector of the encoding block using the nearest reference frame as reference one, and $GMV_{i\text{th frame}}$ is the global motion vector of the ith frame referred to the nearest reference frame. Here, TH_{gmv} and TH_{inter} are thresholds which depends on the quantization step size. And interSATD is the sum of absolute transform difference (SATD) of the encoding block with the inter motion estimation.

First, the algorithm will take a block with a certain inter-mode as input data. Second, the motion vector of the block will be obtained by applying ME on the nearest reference frame. Third, if the difference between the global motion vector and the obtained motion vector in the above step is small enough, we will not search the rest reference frames. However, due to the zooming motion in the video sequences we may get a wrong global motion vector. To solve this problem, we will restrict the SATD to refine our algorithm. That is the restriction about the global motion vector is satisfied, and the SATD is small enough, we determine not to search the rest reference frames.

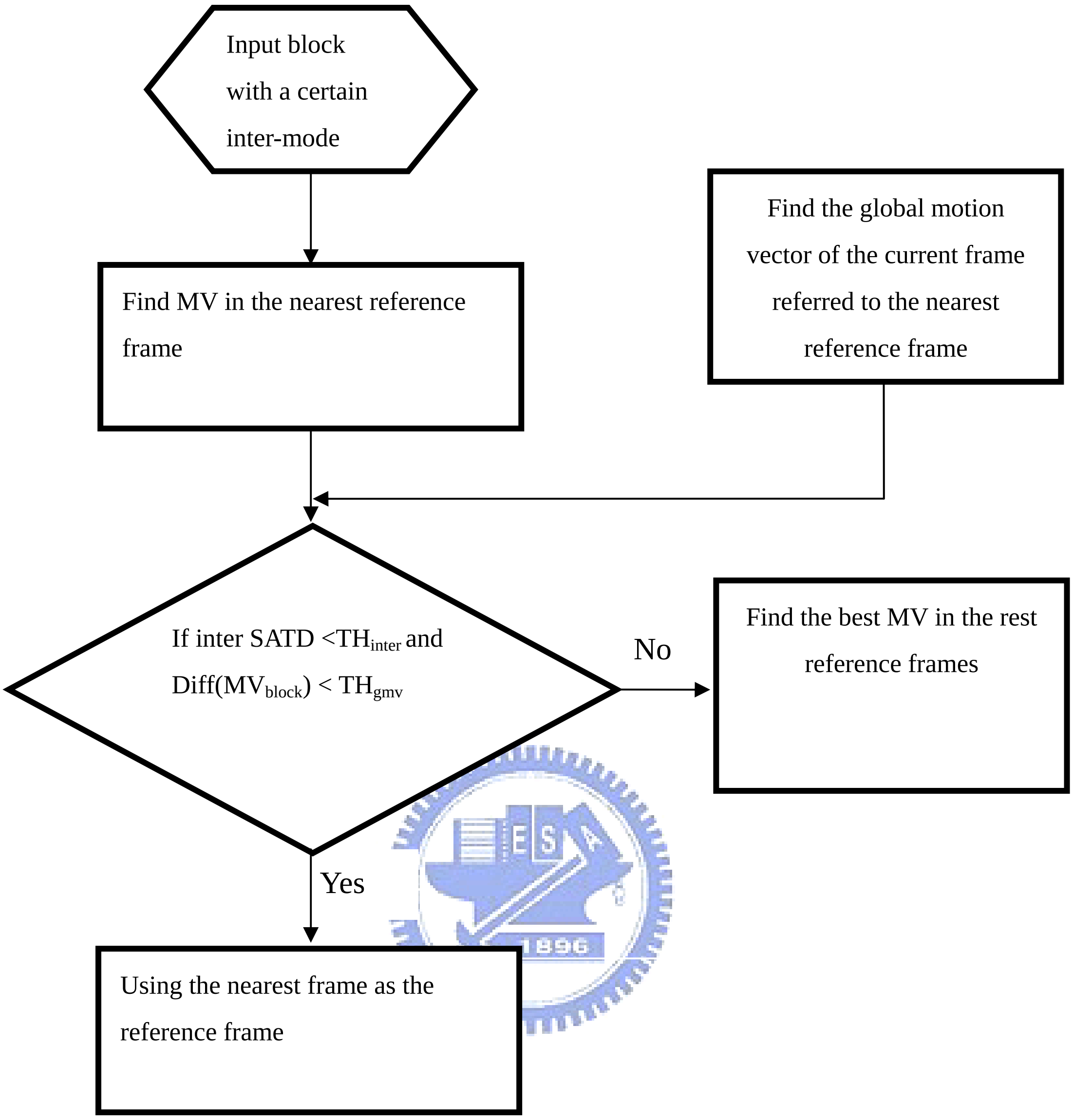


Fig. 7 The flowchart of our proposed method.

CHAPTER 4

Experimental Results

Our method is tested on nine video sequences which are listed in Fig. 4. Here, we encode these video sequences using JM8.1 whose environments are set in Table 3. The results of our method compared with the exhaustive method are shown in Table 4. We can find that the maximum peak PSNR drop is 0.13dB, and the average PSNR drop is 0.05dB. The ME time reduction which is dependent on video sequences is 15%-72%. However, the average bit-rate only increases 1%. It is clear that $GMV_{frame\ ith}$ and TH_{inter} provide the trade-off among the encoding time, the quality, and the bit-rate of the encoded image sequences.

A miss detection means that the selected reference frame of a block using the proposed method is different from the selected one using the exhaustive method. The miss detection rates using our proposed method and those of the method proposed by Huang et al. are shown in Table 5. We can find that our proposed method have lower miss detection rate. Note that the video sequence," Mobile and Calendar", has higher miss detection rate than other video sequences, this is due to that the video sequence has zooming effect and the complicated texture. The proposed method applied to this kind of video sequences will have higher miss detection rate, since we only consider translation motion.

Finally, we will compare the ME time reduction, PSNR drop, and the bit-rate raising between our proposed method and the Huang's method. The comparisons are listed in Table 6. It can be seen that the ME time reduction using our proposed method is higher than the Huang's method. And the average PSNR drops are the same.

Furthermore, the peak PSNR drop using our proposed method is lower than that using Huang's method. In summary, due to that the average miss detection rate using our proposed method is very low, this makes the video quality keep high and bit-rate keep stable.

Table 3 The environments in JM8.1.

Frames To Be Encoded	100
Search Range	16
Number of Reference Frames	5
Video Sequence Format	IPPP

Table 4 The comparison between our proposed method and the exhaustive method.

	ME time reduction (%)	PSNR drop (dB)	Bit-rate raise (%)
Container	59.72	0.04	2.01
Silent	56.95	0.01	-0.42
Mother and Daughter	71.57	0.06	0.79
Hall Monitor	68.38	0.11	4.78
Foreman	47.31	0.13	0.84
Coastguard	23.12	0.02	0.05
Stefan	14.55	0.03	0.37
Mobile and Calendar	19.90	0.07	1.70
Table Tennis	28.61	0.01	-0.08
Average	43.33	0.05	1.0

Table 5 The comparison of the miss detection rate between our method and the method proposed by Huang et al..

	Our method (%)	Huang's method (%)
Container	1.65	1.90
Silent	1.01	3.73
Mother and Daughter	1.58	2.67
Hall Monitor	3.45	5.37
Foreman	5.17	2.57
Coastguard	1.92	6.21
Stefan	2.02	5.62
Mobile and Calendar	7.24	5.86
Table Tennis	0.63	2.41
Average	2.74	4.04

Table 6 The comparison between our proposed method and the Huang's method.

	Our method	Reference: method
ME time reduction	15%-72%	10%-67%
Average PSNR drop	0.05dB	0.05dB
Peak PSNR drop	0.13dB	0.2dB
Miss detection	2.74%	4.04%
Bit-rate	1%	Non-available

CHAPTER 5

Conclusions

In this paper, a fast algorithm is proposed to speed up the multiple reference frame selection in H.264. The algorithm is developed based on the property of the background and the property of the camera motion. The global motion vector between the current frame is needed and the nearest reference frame to determine if it is necessary to search the other reference frames. The experimental results show that the proposed method has up to 97% hit rate. And it can reduce about 15%-72% ME time while keeping the image sequence quality close to exhaustive search method.

Reference

[1] Committee Draft of Joint Video Specification (ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC), July, 2002.

[2] Wiegand T., Xiaozheng Zhang, and Girod B.. “Long-term memory motion-compensated prediction“, Circuits and Systems for Video Technology, IEEE Transactions on Volume 9, Issue 1, Page(s):70 – 84, Feb. 1999.

[3] R. Schafer, T. Wiegan and H. Schwarz. “The emerging H.264/AVC standard”, EBU TECHNICAL REVIEW, January 2003.

[4] Wiegand T., Sullivan G.J., Bjntegaard G.., and Luthra A.. “Overview of the H.264/AVC video coding standard”, Circuits and Systems for Video Technology, IEEE Transactions on Volume 13, Issue 7, Page(s):560 – 576, July 2003.

[5] Joint Video Team(JVT) software JM8.1, July 2004, http://bs.hhi.de/~suehring/tml/download/

[6] Chang A, Au O.C., and Yeung Y.M.. ”A novel approach to fast multi-frame selection for H.264 video coding”, IEEE Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on Volume 2, 25-28 Page(s):II-704 - II-707, May 2003.

[7] Y.W. Huang, Bing-Yu Hsieh, Tu-Chih Wang, Shao-Yi Chen, Shyh-Tih Ma, Chun-Fu Shen, and Liang-Gee Chen. “Analysis and reduction of reference frames for motion estimation in MPEG-4 AVC/JVT/H.264”, IEEE Multimedia and Expo, 2003. ICME ’03. Proceedings. 2003 International Conference on Volume 2, 6-9 Page(s):II – 809 – 12, July 2003.

[8] Peter J. Burt and Edward H. Adelson. ” The Laplacian pyramid as a compact image code”, Communication, IEEE Transactions on [legacy, pre – 1988] Volume

31, Issue 4, Page(s):532-540, April 1983.

[9] Yao Wang, Jorn Qstermann, and Ya-Qin Zhang. “Video Processing and Communications“, Prentice Hall, 1st edition, September 27, 2001.