Motion Estimation for Content Adaptive Video Compression

Jiancong Luo, Ishfaq Ahmad, Yongfang Liang and Yu Sun
Department of Computer Science and Engineering
The University of Texas at Arlington, TX 76019
{jluo, iahmad, yliang, sunyu}@cse.uta.edu

Abstract

A multistage motion estimation scheme is proposed. The scheme extracts video characteristics by first performing an online video analysis separately for foreground and background regions. Motion parameters are extracted and passed to the next stage. The next stage includes a mathematical model for the block distortion surface (BDS) that enables the algorithm to accordingly adjust its search technique. The search is performed on a precise search area adaptive to the statistical property of the motion vector prediction error. Due to its self-tuning property, not only does the proposed scheme adapt to scenes by yielding better visual quality but it also yields a lower computational complexity, compared with the other predictive motion estimation algorithms on standard benchmark sequences.

1 Introduction

In the last two decades, several fast motion estimation (ME) algorithms have been proposed as substitutions for the exhaustive search algorithm. Examples include TSS [3], NTSS [4], DS [7], etc. Recently, ME algorithms based on the motion vector (MV) prediction techniques and flexible search patterns are proposed, such as MVFAST [2] and AMSED [1]. These algorithms enhance the search speeds while maintaining a close visual quality compared with the exhaustive search. However, the performance of a great number of previously introduced algorithms highly depends on the characteristics of the video contents.

In this paper, we introduce a ME scheme for a new paradigm of video compression that we call *content adaptive video compression*. The notion of content adaptive is that the encoder is aware of the content (e.g., objectivity and scene complexity) that it is encoding and the context in which it is being used (e.g., a certain performance measure). The aim is to provide adaptability and self-adjustment to the environment that the compression is being used for. The proposed scheme is called *content adaptive search technique* (CAST).

The paper is organized as follows. Section 2 introduces the proposed scheme. Section 3 presents the performance evaluation results. Concluding remarks are provided in section 4.

2 Content Adaptive Search Technique

Motion estimation is a combination of several techniques including MV prediction, search range, search pattern and termination criterion decision, etc. The performance of these techniques is highly related to the video contents. Therefore, an "intelligent" process that adapts to the video contents is required to maximize the performance. The proposed scheme analyzes and extracts the motion characteristics from the video contents and self-adjusts to adapt to the video contents. Shown in Fig. 1 is the block diagram of the proposed scheme.

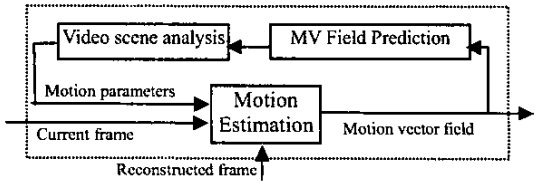


Fig. 1: The block diagram for CAST

CAST consists of three stages: MV field (MVF) prediction, video scene analysis, and motion estimation. They are described as follows:

Stage 1: MVF is predicted from previous frame using the proposed *weighted mean inertia* motion prediction technique.

Stage 2: Blocks are clustered into regions. Motion characteristics parameters of each region are extracted and passed to the next stage.

Stage 3: A ME algorithm combines various techniques that can be fine tuned by the motion parameters and the proposed BDS model.

2.1 Motion vector field prediction

MVF can be predicted based on the spatial/temporal correlation between MVs. [1] [5] proposed *Inertia* MV prediction method to predict the current motion by exploiting the motion inertia property.

Based on [1], we proposed an improved method to increase the accuracy and smoothness of the predicted MVF (PMVF). We call it *Weighted Mean Inertia* (WMI)

0-7803-8603-5/04/$20.00 ©2004 IEEE

MV prediction. Below is the description of WMI.

Let MV denote the motion vector of a block. If the motion remains constantly, the block in the next frame will have a displacement $-MV$. The displaced block overlaps with one or more gridded blocks. Let $B_{i,t-1}$ denote a block i in frame t-1 and $D_{i,t-1}$ denote the displaced $B_{i,t-1}$. The corresponding motion vector and distortion (measured by the sum of absolute difference, abbr. SAD) are $MV_{i,t-1}$ and $SAD_{i,t-1}$. The overlap of $D_{i,t-1}$ and $B_{j,t}$ is denoted by $S_{i,j}$. The predicted MV and distortion of block j in frame t, $PMV_{j,t}$ and $PSAD_{j,t}$, are given below:

$$PMV_{j,t} = \sum_i MV_{i,t-1} S_{i,j} \Big/ \sum_{ii} S_{i,j} \quad (1)$$

$$PSAD_{j,t} = \sum_i SAD_{i,t-1} S_{i,j} \Big/ \sum_{ii} S_{i,j} \quad (2)$$

Table 1 shows the average correlation coefficients (ACC) between the PMVF and the true MVF . Compared with [1], WMI improves the prediction accuracy.

2.2 Video Scene Analysis

Motions in a frame may vary, but motions in a local region generally show similar properties. Therefore, we cluster the blocks into three regions, namely foreground, background and uncovered background. Uncovered background is a region that was covered by an object in the previous frame but is uncovered in the current frame due to the object movement. Usually, in natural video contents, motions in different regions have different characteristics. By extracting regions, we are able to use different techniques on different regions to achieve higher performance in both speed and quality.

2.2.1 Region extraction

The background MVF can be modeled by 6 affine transform parameters. Given the PMVF, we compute the background affine parameters a_1~a_6 using the approach introduced by [6]. The reconstructed background MVF is obtained by (3):

$$MV_{REC}(x,y) = \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} + \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3)$$

where (x,y) is the coordination of the block. E denotes the difference of PMVF and reconstructed background MVF:

$$E = \left| PMV(x,y) - MV_{REC}(x,y) \right| \quad (4)$$

where $PMV(x,y)$ is the predicted MV at (x,y). A threshold T is set to be the standard deviation of E. For each block $B_{j,t}$, if $\sum_i S_{i,j}$ is less than a half of the block size, we mark it as uncovered background. Otherwise, the type is determined as follows:

$$\text{block type} = \begin{cases} foreground & E(x,y) > T \\ background & E(x,y) \le T \end{cases} \quad (5)$$

2.2.2 Motion parameters

We define the following parameters to represent the regional motion characteristics. Motion velocity (M_{vel}) indicates the rapidity of motions in a certain region. Motion complexity (M_{comp}) indicates the degree of disorder of the motions. The definitions of M_{vel} and M_{comp} in background and foreground are below:

$$M_{Fvel} = \frac{1}{N''} \sum_{(x,y) \in Foreground} \left(|MV(x,y)| \right) \quad M_{Fcomp} = \frac{1}{N''} \sum_{(x,y) \in Foreground} \left(\left| MV(x,y) - \overline{MV_F} \right| \right)$$
$$M_{Bvel} = \frac{1}{N'} \sum_{(x,y) \in Background} \left(|MV(x,y)| \right) \quad M_{Bcomp} = \frac{1}{N'} \sum_{(x,y) \in Background} \left(\left| MV(x,y) - \overline{MV_B} \right| \right) \quad (6)$$

where N' and N'' are the numbers of background and foreground blocks. $\overline{MV_B}$ and $\overline{MV_F}$ are the average MVs of background and foreground respectively.

2.2.3 Scene Change Detection and Handling

It is essential to handle scene changes, since they will lead to false estimation of motion parameters. Denote the average pixel difference between previous frame and current frame as D_{pix}. We detect the scene change by comparing D_{pix} with a predetermined threshold T_{sc}:

$$\begin{cases} \text{scene change occurs} & if\ D_{pix} > T_{sc} \\ \text{no scene change} & if\ D_{pix} \le T_{sc} \end{cases} \quad (7)$$

Scene change detection is only performed at P frame. Once a scene change is detected, all blocks in that frame will be INTRA coded. The PMVF of the next P frame will be set to zeros.

2.3 Motion Estimation

2.3.1 Modeling the block distortion surface (BDS)

BDS is a scalar field that consists of distortion values of all search points. A mathematical model for BDS is proposed here. We modeled the BDS as a function of motion, texture complexity and the distance away from the global minimum position (GSP).

Let the center of BDS be the GSP. The distortion value of a search point can be described by $D(r)$ where r is the chess-board distance from G. We observed that $D(r)$ is related to the scene texture complexity. Scene texture complexity can be represented by the minimum distortion $D(0)$. It is because when a scene contains complex textures, it is less possible to find a match for a block with similar texture structure. Therefore, the value of $D(0)$ is large if the texture is complex, and vice versa.

By evaluating the relation between $(D(r)-D(0))^2/D(0)^2$ and r, we observed that $(D(r)-D(0))^2/D(0)^2$ is close linearly related to r, as shown in Fig. 2. The solid lines in Fig. 2 are the linear regression fitting curves of the measured data. We formulated the relation as $(D(r)-D(0))^2/D(0)^2 = g \cdot r$, where g is a function of M_{vel}.

Moreover, experiment analysis shows that $g(M_{vel})$ can be modeled by a rational quadric polynomial, $g(M_{vel})=1/(a+bM_{vel}+cM_{vel}^2)$, where a, b and c are constants. Fig. 3 shows the measured data and the theoretical curve of $g(M_{vel})$. The deduction can be summarized as the following equation:

$$r = \left(\frac{D(r)-D(0)}{D(0)} \right)^2 (a + bM_{vel} + cM_{vel}^2) \quad (8)$$

Once we obtain $D(r)$, M_{vel} and $D(0)$, we are able to estimate the distance r. $D(0)$ is approximated by *PSAD* which is obtained by (2).

2.3.2 Tight Search Area

Our further investigation shows that the distribution of the MV prediction error is highly related to the motion characteristics. We found that in a scene with large M_{comp}, the prediction error has a diverse distribution, and vice versa. Therefore, we propose an adaptive tight search area. The region with a lower M_{comp} employs a tighter search area. Otherwise, a looser search area is applied.

We categorize M_{comp} into three levels, i.e., low, median and high. Extensive experiments have been done to find a proper search area radius for each level. A target probability p_{target} is set such that the accumulative probability of the prediction errors within the radius is no less than p_{target}, as shown in (9). p_i is the probability of the MV prediction error i. In our experiment, p_{target} is set to 99%.

$$p_{target} \le \sum_{i=0}^{radius} p_i \tag{9}$$

When M_{comp} level is high, we do not restrict the search area in order to increase the search accuracy. The tight search area is a diamond shape, covering the search points within the radius determined by (10).

$$radius = \begin{cases} 3 & M_{comp} < 0.1 \\ 7 & 0.1 \le M_{comp} < 0.6 \\ unrestricted & M_{comp} \ge 0.6 \end{cases} \tag{10}$$

2.3.3 Initial search point (ISP)

The proposed scheme maintains a predictor set (PSET). Elements of PSET are evaluated first. The element with the minimum SAD is the initial MV. ISP is the position pointed by the initial MV. The elements are selected from the set { MV_{left} , MV_{up} , $MV_{up\text{-}right}$, PMV_{co}, PMV_{low} , PMV_{right} , MV_{mean} } depending on the motion parameters, where MV_{left} , MV_{up} and $MV_{up\text{-}right}$ are MVs from the left, up and up-right coded block of current frame; PMV_{co}, PMV_{low} , PMV_{right} are MVs from collocated, lower and right block of PMVF, MV_{mean} is the mean of MV_{left} , MV_{up}, $MV_{up\text{-}right}$.

Define the region of support (ROS) as the left, upper and upper-right block of the coding block. Local motion activity (LMA) is defined as the maximum chess-board distance of MV_{left}, MV_{up} and $MV_{up\text{-}right}$. It represents the motion consistence in a small local area.

The elements of PSET are selected as follows:

Uncovered background region:

PSET is composed of the MVs in ROS and MVs of the lower and right blocks in PMVF, i.e. PSET = { MV_{left}, MV_{up}, $MV_{up\text{-}right}$, PMV_{low}, PMV_{right}}.

Other regions:

If all the blocks of ROS are in the same region:

1) If PMV_{co} equals to MV_{mean},

   Foreground region: PSET={ PMV_{co} }

   Background region: If all MVs in ROS are identical and MV_{Bvel} is small, the motion of current block is slow and stable. We skip the search and use *PMV* as the current MV. Otherwise PSET={ PMV_{co} }

2) Or else, if the LMA<5, PSET={ PMV_{co}, MV_{mean}}, otherwise, PSET={ PMV_{co}, MV_{left}, MV_{up}, $MV_{up\text{-}right}$}.

If at less one of the blocks of ROS is from different region: PSET={ PMV_{co}, MV_{left}, MV_{up}, $MV_{up\text{-}right}$}.

2.3.4 Exponential Expand Search (E-search)

We propose an *exponential expand search* which starts from the ISP and is bounded within the tight search area. Two search patterns, cross pattern and exponential expand pattern, are used.

Cross pattern is used to determine the gradient descent direction in BDS and verify the minimum position. A cross (+) shape is chosen because of the fact that most motions in real-world video are along horizontal or vertical direction due to camera panning and tilting. A cross pattern consists of 4 check points, (S,0), (0,-S), (-S,0), (0,S). S is adjustable. The initial S is obtained by:

$$S = \begin{cases} 1 & r < 4 \\ 2 & r \ge 4 \end{cases} \tag{11}$$

Exponential expand pattern aims at fast approaching the global minimum along the gradient decent direction by increasing the step size exponentially.

The search procedure is described as follows:

Step 1: Mark P as ISP.

Step 2: Evaluate the cross pattern centered at P. The point with minimum SAD is marked as the current best point (*CBP*). If P=*CBP*, go to step 3. Otherwise, go to step 4.

Step 3: If S=1, stop. Otherwise, S = 1, go to step 2.

Step 4: Switch to exponential expand pattern: Let V= *CBP*-P. Compute expanded search point *ESP*=2V+P. If *ESP* has the minimum SAD, mark *ESP* as the new *CBP*, repeat step 4. Otherwise, mark the *CBP* as P and go to step 2.

3 Performance Evaluation

This section presents the performance comparison between CAST and two ME algorithms based on MV prediction technique, MVFAST and AMSED. Microsoft MPEG-4 VM encoder has been used for the simulation. 17 most popular test sequences are included, from low motion to fast motion.

Table 2 is the average visual quality and speed comparisons. Visual quality is measured by the increase of peak signal to noise ratio (PSNR gain) compared with MVFAST. Speed up is measured by the ratio of *number of search point* (NSP) compared with MVFAST. As shown in table 2, CAST generally achieves better visual qualities, and outperforms both MVFAST and AMSED in terms of speed-up.

Fig. 5 illustrates per frame comparison of the PSNR

and NSP at scene change occasion. Comparing the visual quality, it can be observed that CAST adapts to the new scene much faster and more linearly than the other two algorithms. Concerning the computational complexity, CAST adapts to the new scene without a burst of NSP at the scene change point.

4 Conclusions

We have proposed a multistage content adaptive motion estimation scheme. The proposed scheme outscores the MVFAST and AMSED in terms of visual quality and computational cost, while showing a better adaptability to various types of scenes and abrupt scene change occasions. The proposed scheme has the best overall performance among the compared algorithms after considering the overhead introduced by the video analysis process. Simplified realization of the proposed scheme is another goal in further study.

Reference

[1] Zheng, W. and Ahmad, I., and Liu, M. "Adaptive Motion Search with Elastic Diamond for MPEG-4 Video Coding," in Detection and Tracking of motion, *ICIP2001*.

[2] P.I. Hosur and K.K. Ma, "Motion Vector Field Adaptive Fast Motion Estimation," *Second International Conference on Information, Communications and Signal Processing (icics '99)*, Singapore, 7-10 Dec. 1999

[3] J.R. Jain and A.K. Jain, "Displacement Measurement and Its Application in Interframe Image Coding," *IEEE Transactions on Communication*, Vol. COM-29, pp. 1799-1808, Dec. 1981.

[4] R. Li, B. Zeng, and M. Liou, "A New Three-Step Search Algorithm for Block Motion Estimation," *IEEE Trans. Circuit and Systems for Video Technology*, vol. 4, NO. 4, pp. 438-442, Aug. 1994

[5] Tieyan Liu, Kwok-Tung Lo, Jian Feng and Xudong Zhang, "Frame interpolation scheme using inertia motion prediction," *Signal Processing: Image Communication*, Vol. 18, 2003

[6] Demin Wong and Limin Wang, "Global Motion parameters estimation using a fast and robust algorithm," *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 7, No. 5, Oct. 1997

[7] S. Zhu and K.K. Ma, "A New Diamond Search Algorithm for Fast Block-Matching Motion Estimation," *IEEE Trans. Image Processing*, vol. 9, NO. 2, pp. 287-290, Feb. 2000

Table 1: ACC between the PMVF and the true MVF

Frame	WMI		Inertia	
	X	Y	X	Y
0	0.668	0.314	0.634	0.320
1	0.634	0.320	0.601	0.287
2	0.448	0.262	0.429	0.223

Table 2: Average speed up and PSNR gains.

	Speed Up			PSNR Gain		
	CIF	QCIF	CCIR	CIF	QCIF	CCIR
MVFAST	1.00	1.00	1.00	0	0	0
AMSED	1.44	1.44	1.37	-0.024	-0.014	+0.070
CAST	3.05	3.83	2.23	+0.025	+0.011	+0.046

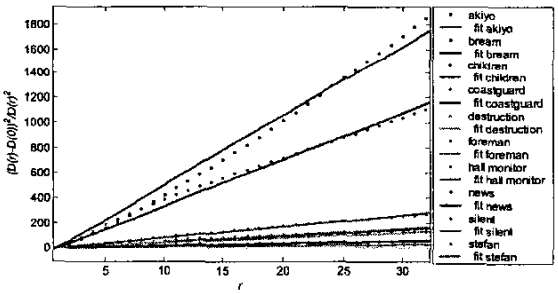


Fig. 2: Measured data of $(D(r)-D(0))^2/D(0)^2$ vs. r and linear fitting curves of various test sequence.

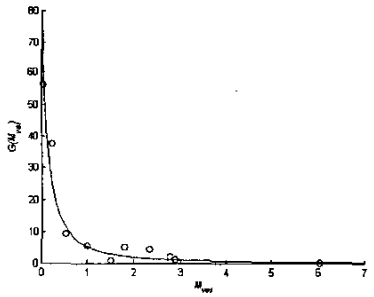


Fig. 3: The measured data and the theoretical curve of function $g(M_{vel})$. ($a = 0.013$, $b = 0.1$, $c = 0.081$)

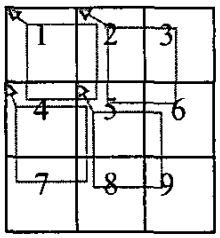
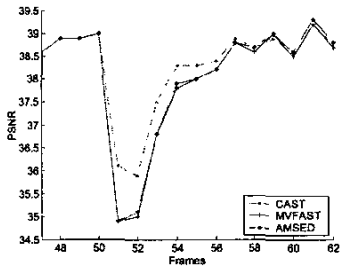
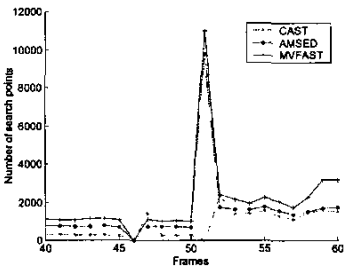


Fig. 4: weighted mean inertia motion vector prediction



a) PSNR



b) NSP

Fig. 5: Comparison on Scene change: *Hall Monitor* concatenated with *Foreman*.