ROBUST AND FAST GLOBAL MOTION ESTIMATION ORIENTED TO VIDEO OBJECT SEGMENTATION

Bin Qi *Aishy Amer*

Electrical and Computer Engineering, Concordia University
Montréal, Québec, Canada
{b_qi, amer}@ece.concordia.ca

ABSTRACT

Most global motion estimation (GME) methods are oriented to video coding while video object segmentation either assume no global motion (GM) or directly adopt a coding-oriented method to compensate GM. This paper proposes a hierarchical differential GME oriented to object segmentation. A combination of 3-step search and motion vector prediction is proposed for initial estimate. Two robust estimators are also proposed: to estimate global motion in the first frame and to reject outliers using object information. Subjective and objective results show that the proposed method is more robust and faster than the reference methods.

1. INTRODUCTION

The term *global motion* (GM) is used in this paper to describe the apparent 2D motion introduced by camera motion. Depending on applications, the objective of global motion estimation (GME) can be different. In video coding, the computed motion need not resemble the *true* motion as long as the bit rate is achieved for a given quality (e.g., [1]). In video object segmentation, the objective is to compensate the GM and retain the object motion (e.g., [2]) where accurate GME is needed. Note that in video coding, even if GM compensation fails, local motion compensation (LMC) can be used to maintain the coding quality. However, LMC is avoided in object segmentation to retain the objects. Most GME methods are designed for video coding while most object segmentation methods either assume no camera motion or directly adopt a coding-oriented GME method.

Computational complexity is a challenge in GME. More accuracy usually means extra computation. Some GME techniques have to sacrifice certain quality to gain speed.

This paper proposes a fast accurate GME for object segmentation. Section 2 introduces related GME principles. Section 3 proposes our method. Section 4 compares the proposed and reference methods. Section 5 is a conclusion.

This work was supported, in part, by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

2. RELATED WORK

2.1. Motion Model and Estimation Criteria

There are different parametric models for GME [9]. In this paper, 6-parameter affine model (Eq. 1) is used that can describe the projected motion of most camera motions [9].

$$\begin{aligned} x'_i &= a_0 + a_1 x_i + a_2 y_i \\ y'_i &= a_3 + a_4 x_i + a_5 y_i \end{aligned} \quad (1)$$

with (x_i, y_i), the i^{th} pixel in the current frame I_n at instant n, (x'_i, y'_i), the corresponding pixel in the previous frame I_{n-1}, and $\mathbf{a} = (a_0, a_1, a_2, a_3, a_4, a_5)$, the GM parameters.

This GM model is incorporated into an estimation criterion [9] to minimize the sum square differences (SSD), i.e., the sum square of the difference errors $\{e_i\}$ between the intensity values of the current I_n and the motion-compensated previous frame I'_{n-1} (Eq. 2 with N, the number of frame pixels).

$$SSD = \sum_{i=1}^{N} e_i^2, \qquad e_i = I'_{n-1}(x'_i, y'_i) - I_n(x_i, y_i) \quad (2)$$

2.2. GME Approaches

Broadly, GME can be classified into three categories: phase correlation approach [3], background matching approach [4], and hierarchical differential approach [5, 6]. The hierarchical differential approach is an efficient and effective tool for GME [1] with many advantages, such as its large search range and fast convergence. It consists of three steps: frame pyramid construction, initial motion estimate, iterative GM parameter optimization (e.g., using gradient descent method). The frame pyramid is built using spatial pre-filtering and sub-sampling. The computation starts at the top level with an initial estimation using a n-step search. Then, gradient descent method is performed iteratively to refine the estimation until a convergence criterion is met. The result is projected onto the lower level of the pyramid and the gradient descent is repeated. This loop is continued until the bottom of the pyramid is reached.

0-7803-9134-9/05/$20.00 ©2005 IEEE

3. PROPOSED GME METHOD

The proposed method (Fig. 1) is based on the hierarchical differential approach with 1) a fast initial estimate combining 3-step search and motion vector prediction (Sec. 3.1), 2) a robust estimate using residual information from previous frames (Sec. 3.2), and 3) a new robust estimate considering neighborhood to eliminate outliers (Sec. 3.3).

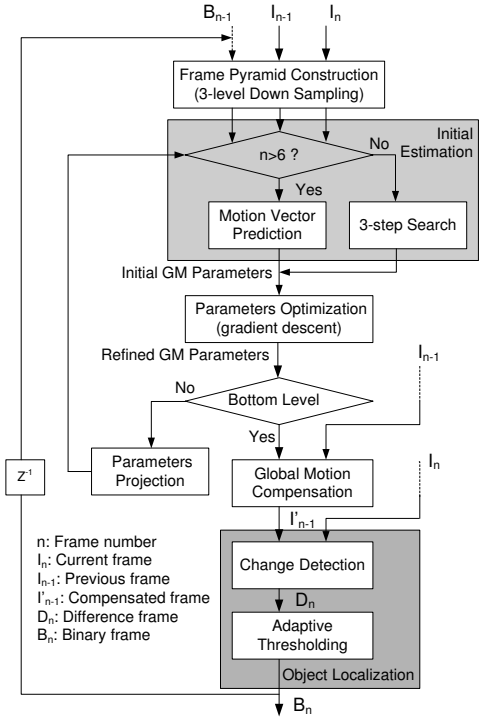

Fig. 1: Block diagram of the proposed method.

3.1. Initial Motion Estimation

After the frame pyramid has been built, We propose to use the 3-step search for initial motion estimate in the first *six* frames. From the *seventh* frame on, we use the motion vector (MV) prediction method [6]. In [6], six MVs are used. To reduce computational complexity, we propose to use only four candidates (Eq. 3) where the MV with the minimum sum absolute difference is selected.

$$\begin{array}{l}\text{zero MV} : \vec{v}_{zero} = 0 \\ \text{past MV} : \vec{v}_{past} = \vec{v}_{n-1} \\ \text{acceleration MV} : \vec{v}_{acceleration} = 2\vec{v}_{n-1} - \vec{v}_{n-2} \\ \text{long-term average MV} : \vec{v}_{average} = \frac{1}{5}\sum_{i=1}^{5}\vec{v}_{n-i}\end{array} \quad (3)$$

Note that we need six frames to get the five previous MVs.

MV prediction is used instead of 3-step search since: 1) six GM parameters are predicted instead of two resulting in more accurate estimates; 2) MVs are predicted using history data because GM gradually changes; and 3) it is faster.

3.2. Using Residual Information for Robust Estimation

One conflict in GME is that there is only one GM model applied to the whole frame, but not necessarily all the pixels experience the same GM. Therefore, those pixels which have local motion will cause big SSD and bias the estimate of GM parameters. Robust estimation aims at solving this problem. The basic idea of robust estimation is to identify the pixels that are not undergoing the GM as outliers, and the remaining pixels as inliers [9]. Then the outliers will be eliminated from the next iteration. Since in this paper, GME is a pre-process for object segmentation, binary residual frames B_n can be used to eliminate outliers.

Assuming that the GM is successfully compensated, the residual information between I_n and I'_{n-1} should contain the objects and the newly appeared background. This object information is derived by applying a binarization method [7] which consists of change detection to obtain the difference frame D_n between I_n and I'_{n-1} and thresholding of D_n to obtain B_n. Then B_n is used to eliminate outliers when estimating GM parameters of the next frame (See Fig. 1).

To prevent outlier misclassification, the pixels of B_n are grouped into blocks. First, the top 30% of the blocks with the most object pixels are selected as outlier candidates. Second, if a candidate block is a boundary block, it is labeled as an outlier block. If a block is not at boundary and has more than three candidate blocks in its 8-neighborhood, it is labeled as an outlier block. If after the previous steps, a candidate block has at least one outlier block in its 8-neighborhood, it is labeled as an outlier block. Then the SSD in Eq. 2 is modified to:

$$SSD = \sum_{i=1}^{N}\rho(e_i), \rho(e_i) = \begin{cases} e_i^2 : B_{n-1}(x_i, y_i) = 0 \\ 0 : B_{n-1}(x_i, y_i) = 1 \end{cases} \quad (4)$$

where $B_{n-1}(x_i, y_i)$ is the i^{th} pixel in the binary B_{n-1}.

To not propagate estimate errors if the GME of the previous frame fails (e.g., the total number of the outlier blocks changes drastically), the residual information from the last successful GM compensation is used instead as follows:

$$\begin{array}{l} O_d = |P_n - P_{n-1}|/P_{n-1} \\ If \quad (O_d > t_O) \\ \quad B_n = B_{n-1}; \quad \mathbf{a_n} = \mathbf{a_{n-1}}; \\ \quad P_n = 0.3P_n + 0.7P_{n-1} \end{array} \quad (5)$$

with O_d, the outlier difference, P_n (P_{n-1}), the number of the outlier blocks in B_n (B_{n-1}), $\mathbf{a_n}$ ($\mathbf{a_{n-1}}$), the GM parameters of I_n (I_{n-1}), and $0.4 < t_O < 1$.

Two advantages to use residual information are: 1) it contains pixels that do not undergo GM and thus is more accurate in eliminating outliers than a statistical estimate, and 2) no extra computation is involved since the residual information comes from the segmentation.

3.3. Robust Estimation for the First Compensated Frame

The robust estimator in Sec. 3.2 cannot be applied for the first compensated frame I'_1 since no previous residual frame exists. Robust estimate in I'_1 is, however, of significant importance for algorithm convergence. We propose the following scheme to reject outliers in I'_1 (see Eq. 6):

1. Sort $\{|e_i|, 1 \leq i \leq N\}$ of I'_1 in descending order.
2. Exclude the top $p\%$ of the sorted $|e_i|$s ($5 < p < 15$).
3. Classify a pixel i as an inlier only if:
   a) $|e_i| \leq e_p$, and b) it has m_i neighbors ($m_i > 6$) in its 8-neighborhood $W_8(i)$ with $|e_j| \leq e_p$, $j \in W_8(i)$.

$$SSD = \sum_{i=1}^{N} \rho(e_i), \rho(e_i) = \begin{cases} e_i^2 : |e_i| \leq e_p \wedge m_i > 6 \\ 0 \ : \text{otherwise} \end{cases} \quad (6)$$

4. RESULTS AND COMPARISON

To evaluate the performance of the proposed method, we compared it to the reference method [5] which is used in the MPEG-4 verification model V.18.0 [1]. (Sample comparison results to the methods in [6] and [4] are also given.) Simulations were carried out using the standard test sequences with global and object motion: *BBCcar*, *tennis*, *marble*, *ferrari*, *stefan*, and *coastguard*.

Fig. 2 shows selected output frames of each test sequence. Change-detected D_n followed by binary B_n are given to show the affect of using different GME methods on object segmentation. As can be seen, the proposed method achieves better subjective results.

4.1. Objective Results

To objectively compare the proposed method and the reference method [5], Figs. 3-5 shows the mean absolute error (MAE) for test sequences used. As can be seen, the proposed method has significantly less MAE than [5].

The size of objects changes, in general, gradually between frames of a video sequence. Fig. 6 shows a comparison of the percentage of white (object) pixels in the output frame B_n. As can be seen, the proposed method shows more stable object regions than the reference method [5].

Furthermore, we have integrated the proposed and [5] method into an object segmentation method [2]. Then we evaluated the segmentation output using both GME methods following the measures in [8]. Fig 7 shows that using our method significantly lower and more stable temporal histogram difference is achieved than using the reference method [5]. The same figure shows that the proposed method outperforms also the methods in [4] and [6].

Finally, the proposed method is about 1.6 time faster than [5] and 1.2 time faster than its faster version [6].

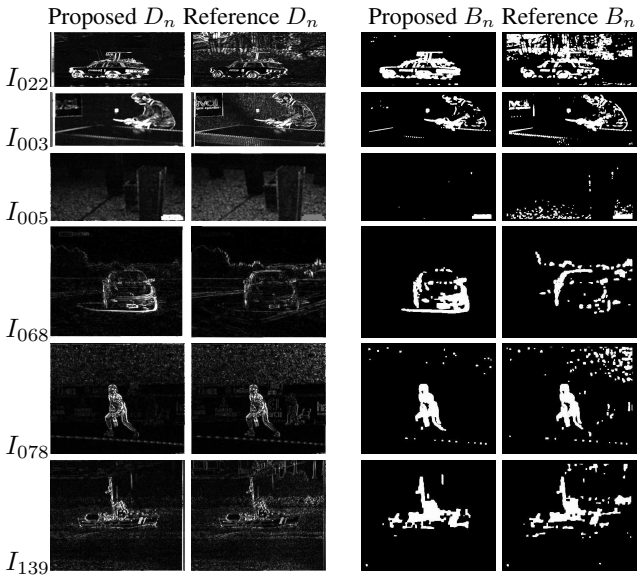

Fig. 2: Comparison: The 1^{st} and 2^{nd} column show results of D_n using the proposed and the reference method [5]. The 3^{rd} and 4^{th} column show B_n of the 1^{st} and 2^{nd} column.

5. CONCLUSIONS AND FUTURE WORK

A robust fast GME method for video object segmentation is proposed with 1) a fast initial estimate using a combination of 3-step search and MV prediction, 2) a robust estimate using object information, and 3) a robust estimate considering neighborhood to eliminate outliers. Both subjective and objective results show that the proposed method is more robust, faster, and more suitable for object segmentation than the reference methods. Future work includes optimizing the interface between the GME and object segmentation.

References

[1] MPEG, "MPEG-4 Video Verification Model Version 18.0", ISO/IEC *JTC1/SC29/WG11 MPEG2001/N3908,* Pisa, January, 2001.

[2] E. Izquierdo, J. H. Xia, R. Mech, "A generic video analysis and segmentation system", *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* Vol. 4, May 2002.

[3] S. Kumar, M. Biswas, T. Q Nguyen, "Global Motion Estimation in Frequency and Spatial Domain", *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* Vol. 3, May 2004.

[4] J.S.Lee, K.Y.Rhee, S.D.Kim, "Moving Target Tracking Algorithm Based on the Confidence Measure of Motion

Vector" *ICIP '01, Proc. of IEEE Int. Conf. on Image Processing,* Vol.1, pp.369-372, October 2001.

[5] F. Dufaux, J. Konrad, "Efficient, Robust and Fast Global Motion Estimation for Video Coding", *IEEE Trans. on Image Processing,* Vol. 9, No. 3, March 2000.

[6] W. C. Chan, O. C. Au, M. F. Fu "Improved Global Motion Estimation Using Prediction and Early Termination", *Proc. of IEEE Int. Conf. on Image Processing,* Vol. 2, Sept. 2002.

[7] A. Amer, " Memory-based spatio-temporal real-time object segmentation", *Proc. SPIE Int. Symposium on Electronic Imaging, Conf. on Real-Time Imaging (RTI),* Santa Clara, USA, vol. 5012, Jan. 2003.

[8] C. Erdem, B. Sankur, and A. Tekalp "Performance Measures for Video Object Segmentation and Tracking" *IEEE Trans. on Image Processing,* Vol.13, No.7, July 2004.

[9] Y. Wang, J. Ostermann, Y. Q. Zhang, "Video Processing and Communications", Prentice Hall, 2001.

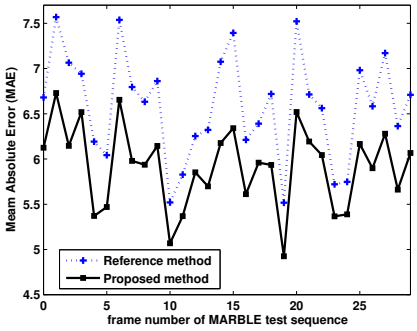

Fig. 3: MAE comparison for *marble*. (Due to space constraints, 25 to 50 frames are selected in the results figures.)

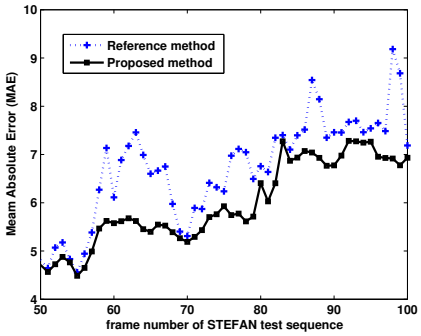

Fig. 4: MAE comparison for *stefan*.

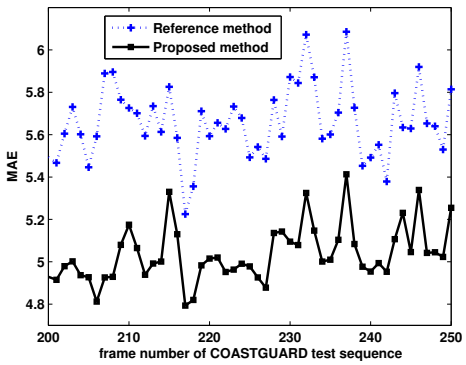

Fig. 5: MAE comparison for *coastguard*.

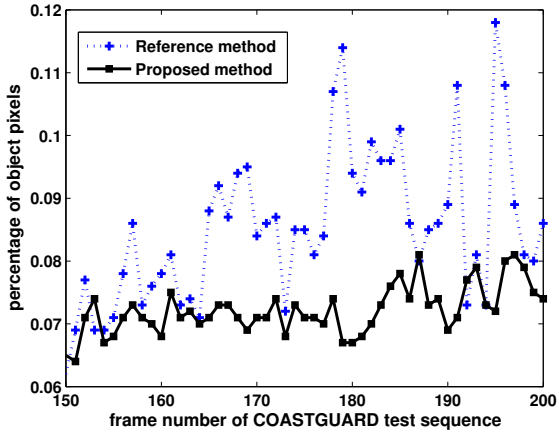

Fig. 6: Percentage of white pixels for *coastguard*.

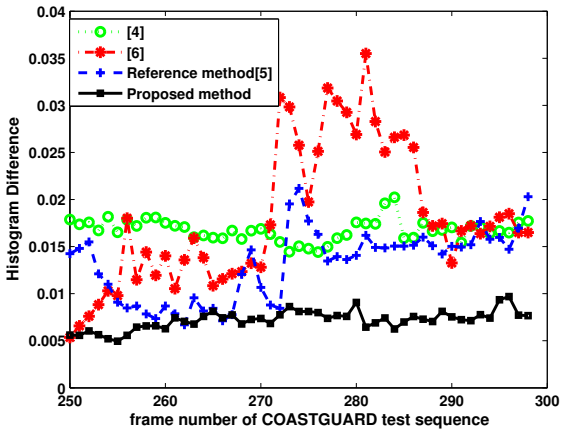

Fig. 7: Comparison of temporal histogram difference for *coastguard* of the proposed and the methods in [5, 4, 6].