Descriptor Based Approach Towards Digital Video Motion Estimation

Ashwani Kumar Aggarwal
Electrical and Instrumentation Engineering Department
Sant Longowal Institute of Engineering and Technology,
Longowal, Sangrur, Punjab,
India 148106

Sunil Kumar
Electrical and Instrumentation Engineering department
Sant Longowal Institute of Engineering and Technology,
Longowal, Sangrur, Punjab,
India 148106

Vishav Mohan Goyal
ACSS, Sant Longowal Institute of Engineering and Technology,
Longowal, Sangrur, Punjab,
India 148106

***Abstract*— This paper presents a simple but effective method for video motion estimation in digital video. The method works by first extracting feature points in video frames and then computing descriptors around these extracted feature points. The descriptors are then matched using Euclidean distance calculation after removal of outliers. An average magnitude of motion vectors in two consecutive frames gives video motion estimation. The method has been tested on many different types of videos like sports video, digital broadcast video and street view image sequence. The method out performs many other methods for motion estimation and also found to be robust towards illumination changes.**

***Keywords*— Motion estimation, digital video processing**

I. INTRODUCTION

Motion of pixels in a digital video is important information the knowledge of which makes us capable of making many analytical decisions of video [1]. It finds application in image segmentation, recognition, scene understanding, image registration, stereo disparity and in other computer vision applications [2]. Video is a three dimensional array of pixels. Two dimensions are meant for spatial distribution of intensity values whereas third dimension depicts its temporal variation. A frame in digital video is a two dimensional array representing the spatial distribution of intensity values at particular time[3]. Video coding is done to take the advantage of spatial as well as temporal redundancy. A number of standards have been published by International Telecommunications Union (ITU) for video coding [4]. The moving Picture Experts Group (MPEG) is a working group of ISO/IEC charged with the development of video encoding standards [5]. MPEG is a standard for the generic coding of moving pictures and associated audio information. It is widely used format of digital television signals that are broadcast by terrestrial, cable and direct broadcast satellite TV systems [6]. This format is also used for distribution of movies on DVDs and CDs.

Motion in 3D is casted as motion vector in two consecutive frames of a video sequence. Among many methods, block matching method [7], pixel recursive algorithms [8], frequency domain methods [9] are widely used for motion vector estimation in digital video. Optical flow finds the displacement by computing apparent flow of pixels from one frame to another frame of is one of the computationally efficient methods used for calculation of motion vectors [10]. However all of these methods have one or the other limitation for their use in digital video motion estimation.

II. FEATURE EXTRACTION

A. Digital Video Sequence

Digital video frame is a RGB image with bit depth of 8-bits. A 256x256 digital frame is 256x256x8x3 bit image. A digital frame is obtained from a analog picture with the process of 2D sampling [11].

$$I(m,n) = A(x,y) \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(x - mX_s, y - nY_s) \quad (1)$$

Let frame f_i be i^{th} frame of a video consisting of N frames. The complete video sequence is represented as below.

$$F = f_i \quad \forall i = 1, 2, ..., N \quad (2)$$

B. Feature Extraction techniques

Feature extraction involves a process of reducing the large number of pixels in a digital video frame into smaller number of key points which are distinguishable, robust and repeatability characteristics [12]. Feature extraction methods can be classified as follows.

- Edge detectors
- Corner detectors
- Blob detectors
- Scale invariant feature transform (SIFT)
- SURF

Edge detectors work by finding pixels in image where there is brightness discontinuity. Corner detectors work by finding the intersection of two edges in an image. Blob detectors find blobs in an image within which some image properties are constants. SIFT feature point extraction method

is scale and rotation invariant method to extract feature points in a digital image. SURF feature point detection is robust to illumination changes and is fast as compared to SIFT. Among several methods, SIFT is used to extract features from video sequence. SIFT works by convolving image with Gaussian at different scales as follows.

$$D(x,y,\sigma)=L(x,y,k_i\sigma)-L(x,y,k_j\sigma) \quad (3)$$

and $L(x,y,k\sigma)=G(x,y,k\sigma)*I(x,y)$

Key points are taken as maxima and minima of difference of Gaussian. Each pixel in DOG is compared with 8 neighboring pixels. If the pixels is maxima or minima among all compared pixels then that pixel is taken as candidate key point.

III. DESCRIPTOR CALCULATION AND MATCHING

Around each feature point extracted from the image, a descriptor of size 128x1 is calculated. An orientation histogram of 8 bins is constructed around each key point in a neighborhood of 4x4 pixels. The descriptor represents the environment around the feature point.

In a digital image, several feature points are extracted and their corresponding feature descriptors ate computed. The number of feature points extracted in an image depends upon image content, resolution of image and many threshold values used.

Once feature points and feature descriptors are calculated in consecutive frames of a video sequence, similarity between the two frames is estimated using various distance metrics as follows.

A. *Euclidean distance*

This distance is most widely used distance metric. Euclidean distance finds the distance between vectors of feature vectors of two images [9].

$$d=\sum_{i=0}^{n}\left|I_i^1-I_i^2\right|^2 \quad (4)$$

B. *Canberra distance*

The feature vector distance is normalized by dividing the distance with sum of feature vectors magnitudes.

$$d=\sum_{i=0}^{n}\frac{\left|I_i^1-I_i^2\right|}{\left|I_i^1\right|+\left|I_i^2\right|} \quad (5)$$

C. *Sum of Squared absolute distance (SSAD)*

This distance is sum of squares of difference between magnitudes of feature vectors of two images.

$$d=\sum_{i=0}^{n}\left(\left|I_i^1\right|-\left|I_i^2\right|\right)^2 \quad (6)$$

D. *Sum of absolute distance(SAD)*

This distance calculates sum of difference of absolute value of feature vectors of two images.

$$d=\sum_{i=0}^{n}\left|I_i^1\right|-\left|I_i^2\right| \quad (7)$$

E. *Maximum value distance*

This distance is used to calculate the largest value of distance between feature vectors of two images.

$$d=\max\left(\left|I_1^1-I_1^2\right|,\left|I_2^1-I_2^2\right|,...,\left|I_n^1-I_n^2\right|\right) \quad (8)$$

A nearest neighbor candidate of a feature point of one image in the second image is calculated by finding the minimum Euclidean distance among their feature vectors. To remove outliers, second nearest neighbor is also estimated. If the ratio of distance of first nearest neighbor to second nearest neighbor is not less than 0.7, then the match is neglected.

IV. RESULTS AND DISCUSSION

A video sequence in outdoor environment is taken for experiments. The video sequence is captured with camera giving panoramic images. Figure 1 shows consecutive frames of a video sequence taken in outdoor environment.

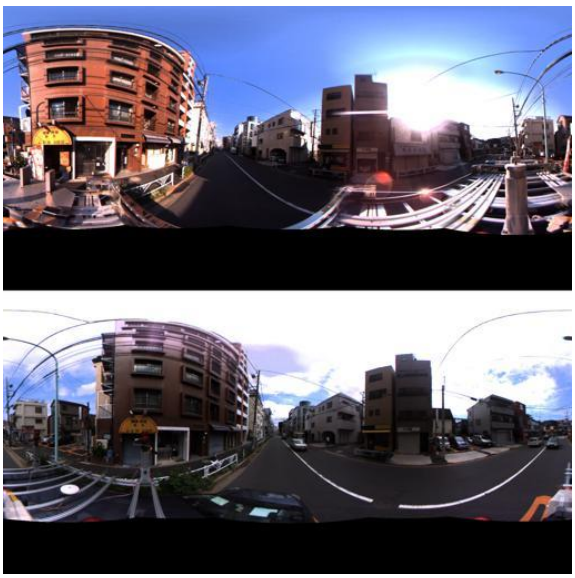

Fig.1: Video frame sequence

The image resolution is 512x1024. Feature points are extracted from the images using SIFT feature point method. Fig.2. Shows extracted feature points on one of the frames of the video sequences. The feature points are overlapped on the actual image by yellow dotted markers. It is noticeable that many feature points are extracted on the building where a few feature points are extracted on the sky. This is due to that the fact that sky portion has relatively constant texture and brightness region as compared to that of buildings.

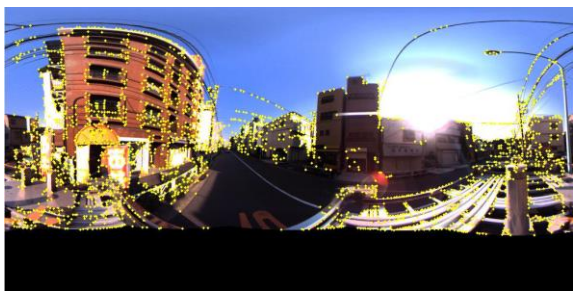

Fig.2: Feature points on video frame

SIFT descriptors of two consecutive frames are matched using Euclidean distance and are shown in Fig.3.

V. CONCLUSION

It has been observed that feature points are matched with their corresponding pixels in the two frames. To remove outliers in the matching process, distance ratio of first nearest neighbor to the second nearest neighbor is calculated. If the ratio is less than 0.7, then it is accepted as a match otherwise it is neglected in match set. RANSAC is also used to remove outliers in the matching process.

The experiment is carried out with many image sequences under different conditions, viz. different illumination conditions and different image resolution etc. It has been observed that motion estimation calculation is computationally fast as compared to other methods and the method works under different illumination conditions.

Various performance indices are also used to compare the performance of the method with other methods. Results show that the method outperforms other methods for motion estimation in digital video.

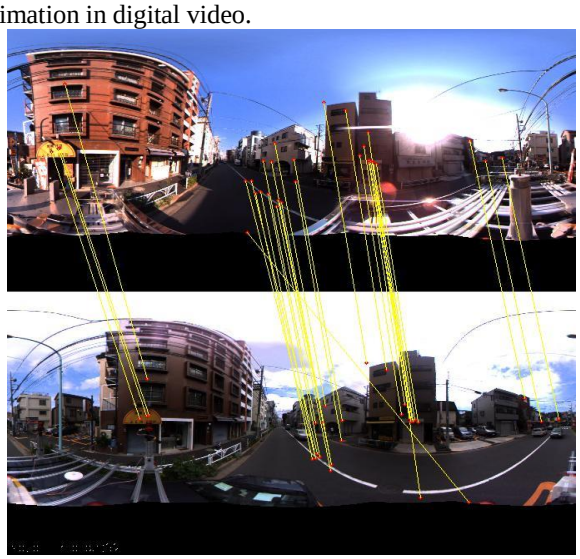

Fig.3: Descriptor matching of video frames

VI. ACKNOWLEDGEMENTS

The author is thankful to Director Sant Longowal Institute of Engineering and Technology, Longowal for providing facilities for the work. The authors are also thankful to their colleague who helped in proof reading this paper.

REFERENCES

[1] Z. Pan, Y. Zhang and S. Kwong, "Efficient Motion and Disparity Estimation Optimization for Low Complexity Multiview Video Coding", in IEEE Transactions on Broadcasting, vol. 61, no. 2, pp. 166-176, June 2015.

[2] Anil. K Jain, Fundamentals of Digital Image Processing, Prentice Hall of India Private Limited, New Delhi.

[3] Rafael C Gonzalez, Richard E Woods, Steven A Eddins, Digital Image Processing using MATLAB, Pearson Education (Singapore) Pte. Ltd, Indian Branch, Delhi, India

[4] Dufaux, F., and Konrad, J., "Efficient, robust, and fast global motion estimation for video coding", in IEEE Transactions on Image Processing, vol. 9, issue 3, pp. 497-501, Mar 2000.

[5] D. Fleet and Y. Weiss, Handbook of Mathematical Models in Computer Vision, Springer Verlag, pp. 237-257, 2006.

[6] Kamp, S., Evertz, M., Wien, M., "Decoder side motion vector derivation for inter frame video coding", in Proceedings IEEE International Conference on Image Processing, San Diego, CA, 2008.

[7] Tok, M.; Glantz, A.; Arvanitidou, M.G.; Krutz, A.; Sikora, T. Compressed Domain Global Motion Estimation using the Helmholtz Tradeoff Estimator, in IEEE Proceedings International Conference on Image Processing (ICIP), Hong Kong, 2010.

[8] Zhu, C., Lin, X., Chau, L.P., "Hexagon-based search pattern for fast block motion estimation", IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 5, pp.349-355, 2002.

[9] Zimmer, H.; Bruhn, A.; Weickert, J.; Valgaerts, B.R.L.; Salgado, A.; Seidel, H.-P. Complementary Optic Flow, in Proceedings International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, 2009.

[10] Zhu, S.; Ma, K.-K. A new diamond search algorithm for fast block-matching motion estimation, IEEE Transactions on Image Processing, vol. 9, no. 2, pp. 287 -290, 2000.

[11] Sheikh, H.R.; Bovik, A.C. "Image information and visual quality", IEEE Transactions on Image Processing, vol.15, no. 2, pp. 430-444, 2006.

[12] Seitz, S.; Baker, S. Filter flow, in Proc. IEEE International Conference on Computer Vision, 2009.