Region-of-Unpredictable Determination for Accelerated Full-Frame Feature Generation in Video Sequences

Jia-Lin Chen [1], Chun-Chen Kuo [2], Liang-Gee Chen [3]

DSP/IC Design Lab, Graduate Institute of Electrical Engineering, National Taiwan University

[1] jocelyn@video.ee.ntu.edu.tw
[2] b99501010@ntu.edu.tw
[3] lgchen@ntu.edu.tw

***Abstract*—In this paper, we propose a novel concept of region-of-unpredictable (ROU) to accelerate full-frame feature generation in video sequences. Due to the high correlation between successive frames, there are only few regions in which the features could not be estimated accurately from the previous frame called region-of-unpredictable (ROU). We develop a scheme combining partial feature extraction in ROU with feature prediction from the previous frame. The full-frame features of the current frame can then be obtained to minimize information loss. Experimental results show that the ROU determination algorithm supports 95.71% detection rate. The full-frame feature generation scheme using ROU determination saves 79.38% computational time compared with the full-frame feature extraction.**

***Index Terms*—Region-of-unpredictable (ROU), video feature extraction, partial feature extraction, motion vectors, global motion estimation**

I. Introduction

Feature generation from a video is important for solving many computer vision tasks like object detection, recognition and tracking. They have variety applications ranging from activity recognition, automated surveillance, vehicle navigation, robot vision, and marker-less augmented reality. The progress of the high resolution video capture device, the wireless internet and the computation of cloud make the applications more and more real. In order to give the user a smooth update of object information, or to recognize new object has just appeared for immediately robot reacting, a fast feature generation algorithm is strongly demanded.

The good invariant properties of local feature make it possible to recognize objects in different scale, light condition and even occluded situation. However the computation of feature generation from a video increases significantly compared to from an image. In order to achieve real-time performance, two strategies are proposed in the previous work. One is to perform partial feature extraction only in the region-of-interest (ROI) instead of full-frame [1][2][3]. The computational effort is reduced by decreasing the area performed feature extraction. However information outside the ROIs is ignored. The other one is to reduce the feature information, like Bag-of-Word (BoW) [4] and CHoG [5]. The information of the features is concentrated to a histogram, and thus feature dimensionality is decreased significantly. These methods can work well in the mobile phone searching scenario, in which users need to input an image only containing the query object. It may not work when the input image containing many objects in clutter background like marker-less augmented reality scenario.

In this paper we propose a novel concept of region-of-unpredictable (ROU). Strong temporal correlation between successive frames in video sequences has been successfully exploited in standard video compression algorithms [6]. The high correlation between successive frames should be used to accelerate feature generation in video sequences. Since there is little difference between two successive frames, a majority of the features in the current frame could be predicted correctly from features in the previous frame. Feature extraction is only needed to perform partially in regions with new content and can not be predicted. Combining with features predicted from previous frame, the full-frame features in current frame can be obtained while computational effort can be reduced significantly.

The remainder of this paper is organized as follows. In the following section, we introduce the proposed accelerated full-frame feature generation scheme. Section III shows our experiment results and analysis. We conclude with discussion in Section IV.

II. Accelerated Full-frame Feature Generation

Our full-frame feature generation scheme is motivated by the fact that most features in current frame can be predicted from the features in the previous frame. The overall block diagram of the scheme is shown in Fig. 1. The motion vectors directly obtained from video encoder and the features from the previous frame are the input. ROU determination algorithm includes two steps, reliable motion vector selection and frame alignment. After ROU determination the global

$978-1-4799-6139-9/14/$*31.00 2014 IEEE*

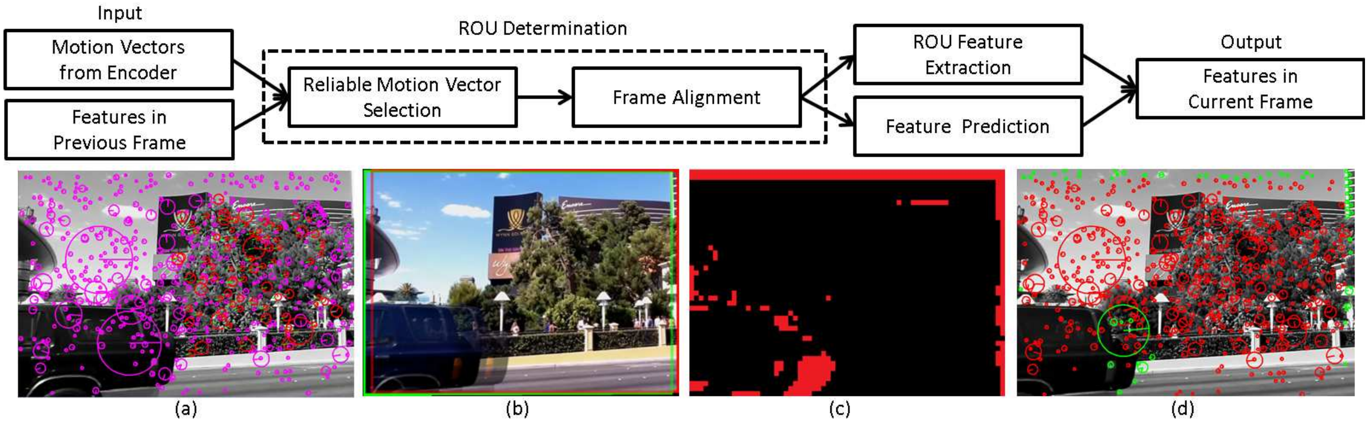
Input
Motion Vectors from Encoder
Features in Previous Frame
ROU Determination
Reliable Motion Vector Selection
Frame Alignment
ROU Feature Extraction
Feature Prediction
Output
Features in Current Frame
(a)
(b)
(c)
(d)

Fig. 1. Proposed accelerate full-frame feature generation scheme: (a) result of reliable motion vector selection, (b) result of frame alignment, (c) result of ROU determination, and (d) output of full-frame feature generation.

motion model and ROU are obtained. Feature generation is performed including partial feature extraction in ROU and feature prediction by the global motion model. Full-frame features in the current frame are obtained as output. The details of each step are described below.

A. Analysis of Motion Vectors

In this work, the motion vectors between each pair of successive frames are considered as input directly obtained from the encoder. The challenges of using motion vectors are discussed [7]. First, motion vectors point backwards in time to macroblocks in previous frames, while the feature points to be tracked are known in the previous frame and must be propagated forward in time. A reverse map is then needed to check the reverse motion vectors from location in previous frames. Second, motion vectors from video encoder are optimized for compression using a rate-distortion Lagrangian metric and do not represent true motion. However work by Takacs et al. [7] have shown that features still can be tracked with a high level of accuracy even though the motion vectors are highly rate-distortion optimized. Third, not all pixels have motion vectors associated with them. For example, the macroblocks intra-coded or the B frames used future frames as reference frames do not have motion vectors from previous frames. Since standardized video encoders are widely implemented in video capture devices and the state-of-the-art video software encoder supports resolutions up to 7680x4320 and frame rates up to 120 fps [8], it is reasonable to assume that the motion vectors from the previous frames can be produced by the dedicated hardware with little additional computation.

Fig. 2 (a) shows a frame from video sequence Pan-Left overlaid with 16x16 macroblock motion vectors. Some motion vectors are estimated incorrectly because of the aperture problem or the new content. The aperture problem shows that if an image has oriented structure, then only the component of motion that is perpendicular to this oriented structure can be measured. Thus the motion vectors of blocks in flat or edge regions are not reliable. For example, the motion vectors in the regions of sky and building with parallel lines are often estimated incorrectly. The regions with new content do not exist in previous frames, therefore the motion vectors are meaningless for feature tracking. The motion vectors include global motion vectors and local motion vectors. The former results from motion of a camera and the latter results from displacements of individual objects composing the scene, such as the car moving left. The goal is to select global motion vectors estimated correctly to build global motion model. The analysis of motion vectors is organized as Fig. 2 (b).

B. Reliable Motion Vector Selection

Instead of to random sample motion vectors, the location of the features in previous frames are treated as reference for reliable motion vector selection. The regions within feature scales are mostly textured, since low contrast points and

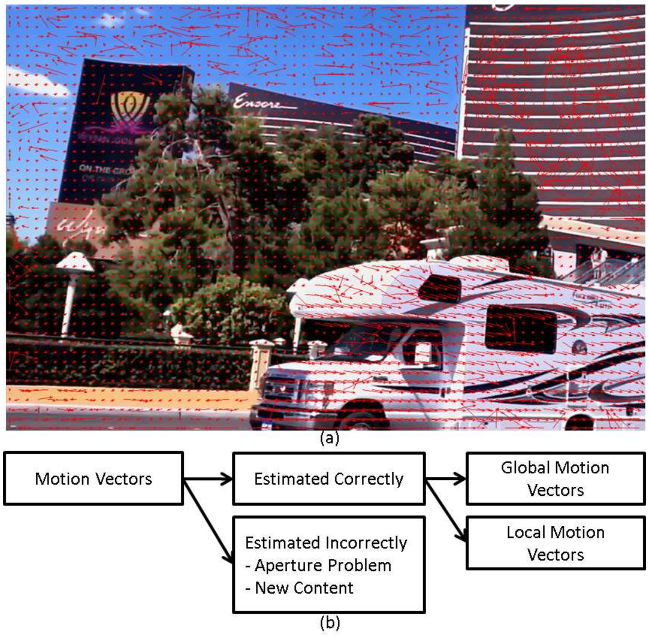

Fig. 2. Example of motion vectors: (a) the current frame overlaid with 16x16 macroblock motion vectors, and (b) the analysis of motion vectors.

edge response points along an edge were screened out by the constraint of feature extraction algorithm [9]. The motion vector estimation is performed with one motion vector per macroblock. More than one motion vectors may be estimated within a feature scale and the consistency of those is checked to select reliable motion vectors. Fig. 1 (a) shows an example result of reliable motion vector selection. All circles indicate the regions within feature scales in the previous frame and red circles indicate the regions within feature scales in which reliable motion vectors detected. The reliable motion vectors are indicated by green arrows.

There are some examples of the motion vector selection illustrated in Fig. 3. Partial previous frame and current frame are shown in the left column and right column respectively. In each previous frame, red circles indicate the region within feature scale. There are five boxes represent five macroblocks locate at the central and four corners within the feature scale. In each current frame, the boxes with the same color corresponding to the previous frame indicate the macroblocks transformed by the motion vectors. Fig. 3 (a) shows two examples without consistent motion vectors. The first row shows that the motion vector in edge regions are not reliable because of the aperture problem. The second row shows the new content or the occluded region would not have consistent motion vectors. Fig. 3 (b) shows two examples with consistent motion vectors which are estimated correctly and reliable. However it is observed that the consistency motion vectors may results from global motion and local motion.

C. Frame Alignment

Under the presupposition that motion vectors result from global motion are majority, RANSAC(RANdom Sample Consensus) algorithm [10] is adopted to select global motion vectors and build the global motion model. The idea is to repeatedly guess parameters of the global motion model using small subsets of motion vectors that are drawn randomly from the input reliable motion vector set. With a large number of draws, there is a high probability to draw a subset of motion vectors that are part of the same model. After each subset draw, the model parameters for this subset are determined and the number of correspondences in that are consistent with these parameters is counted. The set of model parameters with the largest support is considered to be the correct parameters of the global motion transformation matrix **H**.

We then apply image warping by the global motion transformation matrix **H** on the previous frame. The previous frame is warped with the spatial transform to gain the frame aligning with the current frame. Fig. 1 (b) shows an example of frame alignment. The current frame is bounded by red box while the previous frame warped is bounded by green box. The difference between the previous frame warped and the current frame is the reference of ROU determination, the larger difference the more unpredictable. ROU is finally determined in a fixed area ratio to make computation apportioned equally in each frame. Fig. 1 (c) shows the result of ROU determination and ROU is indicated by red color.

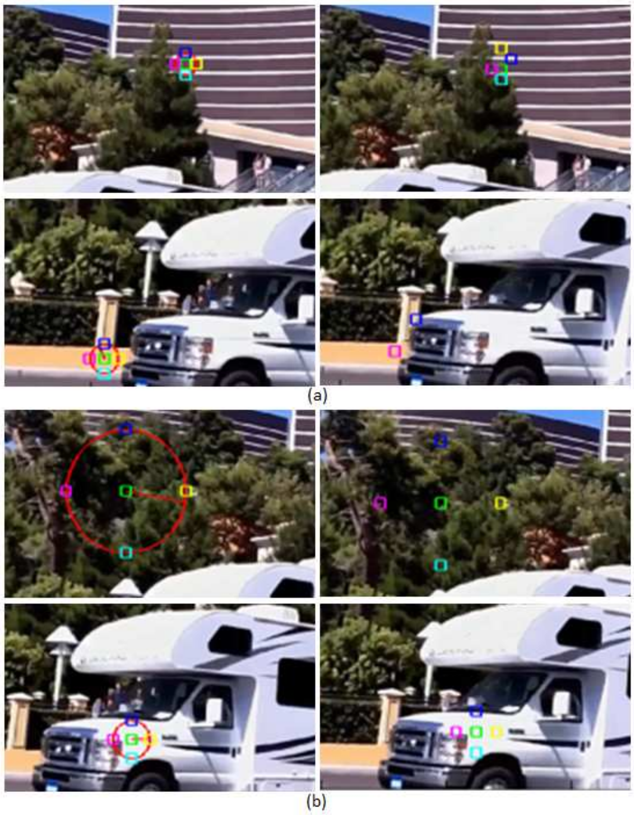

Fig. 3. Example of motion vectors: (a) motion vectors without consistence, and (b) motion vectors with consistence.

D. Full-frame Feature Generation

Full-frame features generation is achieved by feature prediction and partial feature extraction. The features from the previous frame are transformed by **H**. If the features transformed are not located in ROU, they are remained as the features in current frame. Feature extraction is only applied in ROU. Since the hardware feature extraction is usually implemented by tile based architecture, the computational time of partial feature extraction can be considered proportional to the ratio of ROU. Full-frame features in current frame are then obtained by combining the transformed features from the previous frame and the new extracted features in ROU. Fig. 1 (d) shows an example result. The red circles indicates the transformed features, while the green circles indicates the new features generated by partial feature extraction. Comparing with the features in the previous frame shown in Fig. 1 (a), it is observed that the features transformed are located at the proper location and the new features are extracted from the new content.

III. Experiments

We consider six sequences with global camera motions including Pan-Left, Pan-Right, Tilt-Up, Tilt-Down, Zoom-In and Zoom-Out. All the video sequences were collected from the famous video-sharing website Youtube [11]. These sequence are real-world video captured as travel souvenir, thus global motions and local motions both exist in the sequence.

The primary objective of ROU determination is to extract the regions in which the features can not be estimated correctly from the previous frame. The feature extraction is only needed to apply in ROU instead of in whole frame. To generate the ground truth data we perform feature extraction on every frame at first. Feature matching is then carried out between each pair of successive frames. The regions with solitary features are considered as ground truth of ROU.

Precision is the fraction of extracted region that is unpredictable, while recall is the fraction of unpredictable region that is extracted. For hardware design friendly, the ROU determination algorithm extract a fixed ratio of ROU to apportion computation equally in each frame. Hence, precision is not important in our experiments.

After computing the ground truth of ROU, we take recall as the detection rate to evaluate our ROU determination algorithm. Recall is defined as the area of ROU detected correctly by our algorithm divided by the area of ROU ground truth. In Fig. 4 we first compare the recall versus ROU ratio for each video sequence. It is noticeable that only extract 20% area ratio can achieve over 95% recall in average. In other words, feature extraction is only applied in 20% area and over 95% unpredictable features can be detected.

All experiments were performed in MATLAB. Detection rate and computational time are compared in Table I. ROU ratio 100% means applying full-frame feature extraction, hence there is no computational time needed for ROU determination. Experiment results show that the ROU determination algorithm proposed achieves 95.71% detection rate and only needs 0.0197 sec. Since the computational effort is relatively small, the total time is mainly determined by feature extraction. There is a trade-off between detected rate and saved time under which vary with ROU ratio. The full-frame feature generation scheme using 20% ROU determination saves 79.38% computational time compared with the full-frame feature extraction.

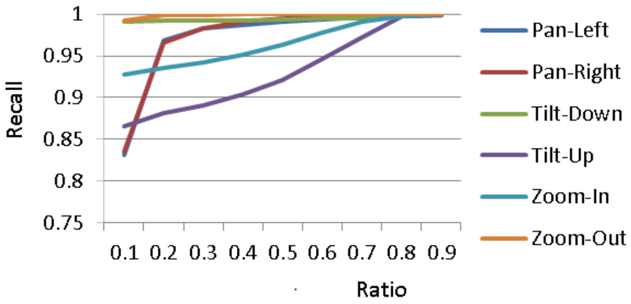

Fig. 4. Recall versus ROU ratio for each video sequence.

IV. Conclusion

In this work we propose a novel concept of partial feature extraction in Region-of-Unpredictable (ROU), which makes best use of the high correlation between successive frames. It is also friendly to hardware design due to the computation apportioned equally in each frame. Besides, we also develop a robust and fast ROU determination algorithm using both motion vectors obtained from video encoder and the features in the previous frame. Experiment results show that the algorithm proposed achieves 95.71% detection rate with little computational effort. The accelerated full-frame feature generation scheme based on ROU determination is accelerated significantly and saves 79.38% computation time compared with the full-frame feature extraction.[12][13]

TABLE I
Detection Rate and Computational Time

ROU Ratio (%)	Detection Rate (%)	Saved Time (%)	Computational Times (sec)	
			ROU Determination	Feature Extraction
100	100	0	0	3.2013
20	95.72	79.38	0.0197	0.643

References

[1] Y.-C. Su, K.-Y. Huang, T.-W. Chen, Y.-M. Tsai, S.-Y. Chien, and L.-G. Chen, "A 52 mw full hd 160-degree object viewpoint recognition soc with visual vocabulary processor for wearable vision applications," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 4, pp. 797–809, 2012.

[2] J. Oh, G. Kim, J. Park, I. Hong, S. Lee, and H.-J. Yoo, "A 320mw 342gops real-time moving object recognition processor for hd 720p video streams," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*. IEEE, 2012, pp. 220–222.

[3] Y.-M. Tsai, T.-J. Yang, C.-C. Tsai, K.-Y. Huang, and L.-G. Chen, "A 69mw 140-meter/60fps and 60-meter/300fps intelligent vision soc for versatile automotive applications," in *VLSI Circuits (VLSIC), 2012 Symposium on*. IEEE, 2012, pp. 152–153.

[4] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 427–434.

[5] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, "Compressed histogram of gradients: A low-bitrate descriptor," *International journal of computer vision*, vol. 96, no. 3, pp. 384–399, 2012.

[6] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[7] G. Takacs, V. Chandrasekhar, B. Girod, and R. Grzeszczuk, "Feature tracking for mobile augmented reality using video coder motion vectors," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 141–144.

[8] (2013) The nippon telegraph and telephone corp. website. [Online]. Available: http://www.ntt.co.jp/news2013/1308e/130808a.html

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[11] (2005) Youtube. [Online]. Available: https://www.youtube.com/

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] Y.-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A new method for camera motion parameter estimation," in *Image Processing, 1995. Proceedings., International Conference on*, vol. 1. IEEE, 1995, pp. 406–409.