A Low Complexity, Descriptor-Less SIFT Feature Tracking System

Brian Fransioli, 이혁재
서울대학교
terranpro@capp.sun.ac.kr, hjlee@capp.snu.ac.kr

Brian Fransioli, Hyuk-Jae Lee
Seoul National University

Abstract

Features which exhibit scale and rotation invariance, such as SIFT, are notorious for expensive computation time, and often overlooked for real-time tracking scenarios. This paper proposes a descriptor-less matching algorithm based on motion vectors between consecutive frames to find the geometrically closest candidate to each tracked reference feature in the database. Descriptor-less matching forgoes expensive SIFT descriptor extraction without loss of matching accuracy and exhibits dramatic speed-up compared to traditional, naive matching based trackers. Descriptor-less SIFT tracking runs in real-time on an Intel dual core machine at an average of 24 frames per second.

1. Introduction

Local invariant features such as SIFT [1], while popularized for image registration, have also been successfully applied to the realm of object detection. While they are generally regarded as powerful due to scale and rotation invariance, SIFT features are notorious for high computational complexity. As such, they are often overlooked for real-time applications such as tracking. In contributing towards real-time tracking, [2] has developed a binary representation of the SIFT descriptor based on the medians of the dimensions of descriptor space. This has reduced problems seen with nearest neighbor matching in high dimensional spaces, and reduced the distance computation to a Hamming distance that can be performed efficiently using XOR instructions and bit counting. [3] has proposed a low complexity homography matrix based SIFT tracking system which uses a subset of the SIFT keypoints available for matching, reducing the number of descriptors generated and the number of distance calculations performed. They also use a [constant] window region around the previous object location as a seed for search and extraction in subsequent frames. Still, the above improvements fall short of real-time performance. [4] suggests a SIFT matching process which then involves calculating motion vectors for video stabilization.

Continuing the exploitation of temporal locality and further reduction of expensive descriptor computations, this paper proposes a descriptor-less, motion vector based nearest neighbor matching scheme. SIFT keypoints are extracted from an adaptively sized window each frame, then tracked and matched with the closest reference feature in the database. Outlier estimation of the match set is performed using a standard RANSAC implementation, and localization of the object in the current frame along with trajectory updates of the reference features are done.

2. Descriptor-Less Feature Tracking

In exchange for SIFT descriptor extraction and distance computations between features, the proposed algorithm uses a block based motion vector which will approximate the distance between features $F_{i,n-1}$ and $F_{j,n}$ of subsequent frames. The choice of motion vector is not limiting, as we use a block based, H.264 encoder generated motion vector to prove features can be accurately matched even with approximate motion vectors.

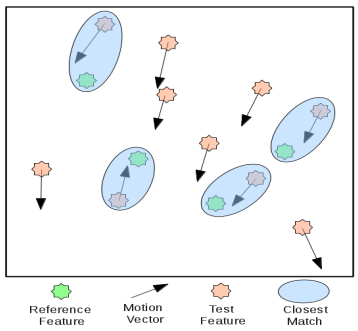

Figure 1. Test features are projected using their motion vectors to their predicted, previous location, and the closest candidate is matched to a reference feature.

In each frame, an adaptive detection window is created by calculating the dominant motion vectors of the object as it was located in the previous frame. This approximates the location of the object in the frame, and creates a window for SIFT feature extraction. After keypoint extraction, motion vectors for each feature are used to calculate the nearest

neighbor (in 2D space) to a reference feature in the database. The smallest Euclidean squared distance between a reference feature and a test feature in the current frame constitutes a matching pair of features. Figure 1 illustrates the descriptor-less matching process. A maximum distance threshold for matching features is established, adaptively, using the dominant motion vector.

This set of matching features will contain both correctly and incorrectly matched features. As such, outlier determination must be performed to generate an accurate model for homography estimation. The RANSAC algorithm is used to localize the tracked object in the current frame; in addition to object localization, the homography matrix between successive frames is used to accurately update the location of reference features in the database, matched or unmatched in the current frame.

3. Results

The descriptor-less tracking system is compared to a naive SIFT tracking implementation which generates matches based on the nearest neighbor metric first described by Lowe in [1]. It uses no motion vectors or concepts of temporal locality. Binary SIFT (BSIFT) descriptors for tracking [2] are also used, which use naive matching metrics but have a lower cost for distance computations.

Results for the descriptor-less tracking system's operation times and accuracy are shown in Figure 2 and Table 1. *Match* shows the amount of time (in seconds) to make a decision about a best match, or no match, for all reference features. For the naive matcher, this consists of selection of the closest two matching features' descriptors among the entire test feature set with a nearest neighbors ratio metric as described by Lowe. BSIFT chooses the best match based on similarity of bit count [2]. For descriptor-less matching, matching consists of a 2D nearest neighbor search among test features shifted by their motion vectors. Homography calculation (RANSAC) times show the time taken for outlier estimation to converge to a solution. The *1-Precision* is also shown which is given by

$$\frac{FP}{FP+TP} \equiv \frac{Outliers}{Matches}$$

where *FP* and *TP* are the number of false and true positives, respectively, as determined by the outlier count from RANSAC. It is expressed as a percentage from *0-1*.

Results show dramatic improvements in matching time, and moderate improvements in outlier estimation. During challenging frames, where the outlier ratio is relatively high, the time taken for RANSAC to converge to a solution with naive matching is longer. The descriptor-less matching system maintains an overall lower ratio of outliers, evident of the advantages of exploiting temporal locality, and that naive matching undoubtedly incurs erroneous matches of similar descriptors. Table 1 shows system performance comparison for several test sequences. Descriptor-less matching forgoes SIFT descriptor computations, which results in a direct time savings without loss of matching accuracy or increased outlier count. *1-Prec* is the average *1-Precision* percentage from *0-1*. *ErrorD* represents average spatial deviation in pixels of the object location from ground truth.

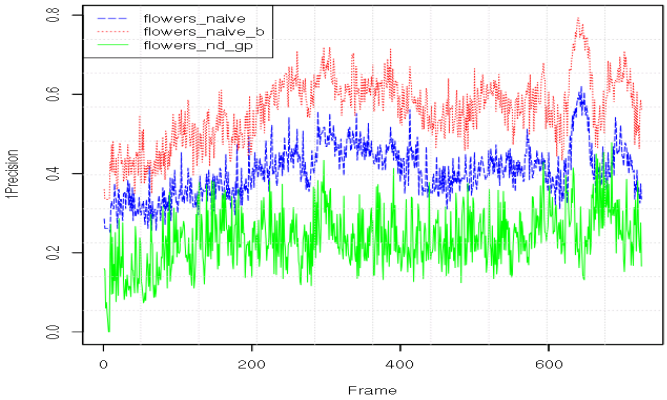

Figure 2. 1-Precision for the *flowers* test sequence (from top to bottom) BSIFT (dot,red), Naive (dash,blue), and D-Less (solid,green) systems.

		Descript	Match	RANSAC	FPS	1-Prec	ErrorD
Bottle	Naive	48.50	108.00	5.00	5.02	0.58	3.98
	BSIFT	51.00	4.01	16.00	9.43	0.67	4.29
	DLess	0	0.45	2.00	27.03	0.34	4.24
Firesign	Naive	38.00	100.00	5.00	5.68	0.35	1.64
	BSIFT	39.00	4.00	5.00	11.24	0.56	1.64
	DLess	0	1.80	2.00	25.64	0.24	1.66
Flowers	Naive	55.00	390.00	2.00	2.07	0.41	2.02
	BSIFT	58.00	16.00	6.00	8.62	0.57	1.49
	DLess	0	4.00	2.00	23.81	0.31	1.45
Beatles	Naive	38.00	166.00	5.00	4.11	0.55	1.56
	BSIFT	56.00	19.00	19.64	7.94	0.55	2.18
	DLess	0	2.00	7.50	26.31	0.46	1.29

Table 1. Comparisons for a variety of test sequences; times in *ms*, FPS (frames/sec), 1-Prec outlier percentage, and ErrorD is distance error to ground truth in pixels. Sequences use 720x480 resolution.

4. Conclusion

As we have shown, descriptor-less matching forgoes SIFT descriptor computations, which results in a direct time savings without loss of matching accuracy. Motion vectors used can be as simple as block-based (e.g. H.264), or even as complex as optical flow (e.g. LKT). Optical flow was tested to run on VGA and 720x480 sized test sequences on average in 3~4 ms per frame, and does not contribute significantly to the overall time. The descriptor-less SIFT tracking system achieves real-time performance, running at an average of 24 FPS.

Acknowledgment

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST) (2011-0027502).

References

[1] D.G. Lowe. "Distinctive image features from scale-invariant keypoints," *Intl Journal of Computer Vision* 60.2, 2004.

[2] M. Stommel et al. "A fast, robust and low bit-rate representation for SIFT and SURF features." *in Proc. IEEE Intl Symp on Safety, Security, and Rescue Robotics*, 2011.

[3] Lu et al. "Low Complexity Homography Matrix Based SIFT for Real-time 2D Rigid Object Tracking," *in Proc. 6th Intl Conf on Wireless Communications Networking and Mobile Computing*, 2010.

[4] S. Battiato et al. "SIFT Features Tracking for Video Stabilization" *in Proc. 14th Intl Conf on Image Analysis and Proc.* 2007.