Monocular Visual-IMU Odometry: A Comparative Evaluation of the Detector-Descriptor Based Methods

DOI:
10.1007/978-3-319-46604-0_6

Document Version
Accepted author manuscript

Published in:
Computer vision - ECCV 2016 workshops

Monocular Visual-IMU Odometry: A Comparative Evaluation of the Detector-Descriptor Based Methods

Xingshuai Dong [a], Xinghui Dong [b*], and Junyu Dong [a]

[a] Ocean University of China, Qingdao, 266071, China
[b] Centre for Imaging Sciences, University of Manchester, Manchester, M13 9PT, UK

Abstract. Visual odometry has been used in many fields, especially in robotics and intelligent vehicles. Since local descriptors are robust to background clutter, occlusion and other content variations, they have been receiving more and more attention in the application of the detector-descriptor based visual odometry. To our knowledge, however, there is no extensive, comparative evaluation investigating the performance of the detector-descriptor based methods in the scenario of monocular visual-IMU (Inertial Measurement Unit) odometry. In this paper, we therefore perform such an evaluation under a unified framework. We select five typical routes from the challenging KITTI dataset by taking into account the length and shape of routes, the impact of independent motions due to other vehicles and pedestrians. In terms of the five routes, we conduct five different experiments in order to assess the performance of different combinations of salient point detector and local descriptor in various road scenes, respectively. The results obtained in this study potentially provide a series of guidelines for the selection of salient point detectors and local descriptors.

Keywords: Monocular visual-IMU odometry, odometry, navigation, salient point detectors, local descriptors, evaluation

1 Introduction

Ego-motion estimation in real-world environments has been studied over the past decades. As one of the commonly-used methods for this problem, Visual Odometry (VO) estimates the pose of a vehicle by matching the consecutive images captured using the onboard camera [28]. According to the camera involved, visual odometry can be divided into two categories: monocular and stereo [28]. However, the architecture of stereo visual odometry systems is normally complex, which limits their practical applications. Stereo visual odometry also tends to degenerate to a monocular system when the distance between objects and the camera is large. On the other hand, monocular visual odometry systems are simple and can be easily used in practical applications. In addition, the joint use of the Inertial Measure Unit (IMU) and the camera (referred to as Visual-IMU Odometry) normally improves both the reliability and accuracy of motion estimation [19] because they are complementary [3]. Hence, the scope of this research is limited to the study of monocular visual-IMU odometry.

Considering local descriptors are insensitive to occlusion, background clutter and other changes [23], they have been extensively applied to visual odometry [26], visu-

*Corresponding author: Xinghui Dong. Email: xinghui.dong@manchester.ac.uk.

al-SLAM (Simultaneous Localization and Mapping) [5] and visual tracking [9]. Local descriptors are normally extracted at the salient points detected from images in order to accelerate the speed of feature matching. In this context, salient point detection and feature extraction are key to the detector-descriptor based visual odometry systems. As a result, an extensive evaluation of detectors and descriptors in a unified visual odometry framework is required in order to obtain guidelines for the choice of these.

To the authors' knowledge, however, there is no research which extensively assesses the performance of salient point detectors and local descriptors for the applications of monocular visual-IMU odometry. In this paper, we therefore conduct an extensive, comparative evaluation of different combinations of detector and descriptor in the scenario of monocular visual-IMU odometry. The contributions of this paper are: (1) we design a unified evaluation framework based on five typical routes containing different road scenes and a well-established monocular visual-IMU odometry system [15]; and (2) we survey five salient point detectors and eight local descriptors (in which HOG [4], LIOP [34], LM [18] and LSSD [32] have not been applied to visual odometry) and perform a comparative evaluation on different combinations of detector and descriptor, which produces a set of useful benchmarks and insights.

The remainder of this paper is organized as follows. Related work is reviewed in Section 2. In Section 3, the detail and implementation notes of the salient point detectors and local descriptors are described. The experiments are introduced in Section 4 and the results are reported in Section 5. Finally, conclusions are drawn in Section 6.

2 Related Work

In this section, we briefly review the existing work related to salient point detectors and local descriptors, the application and the evaluation studies of these methods.

2.1 Salient Point Detectors

Salient points are normally used to avoid the heavy computational cost of matching all the pixels in two images. Harris and Stephens [13] proposed a corner detector using the image gradient matrix. Based on this detector, Mikolajczky and Schmid [21] proposed the Harris-Laplace corner detector. The FAST (Features from Accelerated Segment Test) corner detector [27] was introduced based on a discretized circle of pixels surrounding the corner candidate point. Although corner points can be fast computed, they are less distinctive. In contrast, the points detected using blob detectors are more distinctive and redetected [28]. These detectors include the Difference of Gaussian (DoG) detector [20] and the Fast Hessian detector [1]. In addition, Geiger et al. [11] proposed a blob and corner detector in order to capture both types of points.

2.2 Local Descriptors

Local descriptors have been widely applied in computer vision due to their powerful representation abilities. Local descriptors, for example, Scale-Invariant Feature Trans-

form (SIFT) [20] and Histogram of Orientation Gradient (HOG) [4], can be computed from local gradient histograms. As a faster alternative to SIFT, Bay et al. [1] introduced the Speeded-Up Robust Features (SURF) descriptor. Local descriptors can also be extracted in the form of filter responses [18] or image patches [33]. Besides, Shechtman and Irani [32] introduced a Local Self-Similarity Descriptor (LSSD) while Wang et al. [34] proposed a Local Intensity Order Pattern (LIOP) descriptor.

2.3 Detector-Descriptor Based Monocualr Visual (-IMU) Odometry

The application of local descriptors can be found in many visual odometry tasks. Nister et al. [26] applied the image patches extracted at the Harris corner points to monocular visual odometry, while Bloesch et al. [2] used the FAST detector and multi-level patches for monocular visual-inertial odometry. As one of the most famous local descriptors, SIFT [20] has been used in monocular visual-IMU odometry systems [15] [24]. Nilsson et al. [25] also proposed a monocular visual-aided inertial navigation system using SURF [1]. However, these descriptors are normally extracted from gray level images. In order to exploit richer image characteristics, Dong et al. [7] applied three sets of multi-channel image patch features to monocular visual-IMU odometry.

2.4 Comparative Evaluations of Salient Point Detectors and Local Descriptors

Many evaluation studies have been conducted for computer vision tasks. Schmid et al. [31] compared salient point detectors under different scale, viewpoint, lighting and noise conditions. Mikolajczyk and Schmid further assessed different affine-invariant detectors [22] and descriptors [23]. Recently, Gauglitz et al. [9] compared different salient point detectors and local descriptors for visual tracking. An evaluation study of local descriptors was also performed in the field of geographic image retrieval [35].

On the other hand, the similar comparative studies have also been performed for the visual odometry tasks in the indoor [30] and outdoor scenes [12], [16], [29]. However, only a small number of combinations of detector and descriptor were tested in these studies. In addition, the datasets used are not representative to road scenes. Therefore, we conduct a series of extensive (more detectors and descriptors) evaluation experiments based on a unified monocular visual-IMU odometry framework containing five particularly typical real-world routes. To our knowledge, this is the first extensive evaluation study in the scenario of monocular visual-IMU odometry.

3 Salient Point Detectors and Local Descriptors

We briefly review the salient point detectors and local descriptors tested in this study. The parameters used for these methods can be found in the supplementary material.

3.1 Salient Point Detectors

The five salient point detectors examined in this study are described as follows.

Blob and Corner (Blob&Corner) Geiger et al. [11] first convolved the blob and corner masks with an image. Then, non-maximum and non-minimum suppressions were applied to response images. Four types of points: "corner max", "corner min", "blob max", and "blob min" were derived.
Difference of Gaussian (DoG) Lowe [20] introduced a salient point detector by finding local extrema in an image. The convolution of the image and the DoG functions is performed at different scales. Thus, the salient points can be derived by convolving the extrema of the scale space in the DoG functions with the input image.
Fast Hessian As a scale-invariant salient point detector, Fast Hessian [1] was developed on the basis of the Hessian matrix. In order to reduce the computational cost, Bay et al. [1] used a set of box filters to approximate the Laplace of Gaussian functions. The salient points in an image can be obtained by detecting the local maximum of the determinate of the Hessian matrix over spatial locations and scales.
Features from Accelerated Segment Test (FAST) Rosten et al. [27] proposed the FAST detector. This detector operates on a circle of 16 pixels around the candidate corner point p. The point p is treated as a corner if there is a continuous arc of at least nine pixels that are darker than the pixel $I_p - s$ (s is a threshold) or brighter than the candidate pixel $I_p + s$. The FAST detector can be further accelerated by learning a decision tree in order to examine fewer pixels.
Harris-Laplace The Harris-Laplace detector [21] locates potential salient points in the scale space based on a multi-scale Harris corner detector. The key idea of the Harris-Laplace detector is to obtain the representative scale of a local pattern, which is the extremum of the Laplacian function across different scales. This scale is representative in the quantitative viewpoint because it measures the scale at which the maximal similarity between the detector and the local image pattern is reached.

3.2 Local Descriptors

In total, we tested eight different local descriptors in this study. We briefly introduce these below. For more details, please refer to the original publications.
Histogram of Oriented Gradients (HOG) The HOG descriptor computes the occurrence of the gradient orientation in the sub-regions of an image [4]. It first partitions the image into blocks which are further divided into cells. Then, a gradient orientation histogram is derived over each cell. The histograms obtained over each block are concatenated into a vector. We computed a 9-bin histogram from each 5×5 cell in the 15×15 block around a salient point in this study.
Image Patches (IMGP) The simplest description of a salient point is the image patch around this point. Extraction of image patches only requires cropping the image at a given point. The non-warped image patches retain the original image characteristics [33]. In our experiments, the size of image patches was set as 11×11 pixels.
Integral Channel Image Patches (ICIMGP) Dollár et al. [6] proposed a set of integral channels, including the gray level (or color) channel(s), the gradient magnitude channel and six gradient histogram channels. Dong et al. [7] first extracted the image patch around a point in each channel. Then, each patch was L_2 normalized separately.

All patches were combined into a single ICIMGP feature vector. In this study, the size of patches was set as 11×11 pixels.

Leung-Malik (LM) Filter Bank The LM filter bank [18] contains 36 first- and second-order derivatives of Gaussian filters built at six orientations and three scales, eight Laplace of Gaussian filters, and four Gaussian filters. We applied the LM filter bank at each salient point in this study.

Local Intensity Order Pattern (LIOP) Given the image patch around a point, the LIOP descriptor [34] first partitions it into sub-regions using the overall ordinal data. Then, a LIOP is computed over the neighborhood of each pixel. The LIOPs contained in each sub-region are accumulated into an ordinal bin. The LIOP descriptor is obtained by combining different ordinal bins.

Local Self-Similarity Descriptor (LSSD) Given an image, LSSD [32] first computes a correlation surface for each pixel by comparing its local neighborhood with the neighbourhood of each pixel within a larger surrounding region. Then, the surface is partitioned into log-polar bins, which contains n_r radial bins and n_θ angular bins. Finally, the descriptor is obtained as the normalized bins by linearly stretching these bins into the range of [0, 1].

Scale-Invariant Feature Transform (SIFT) The SIFT descriptor [20] is extracted by computing a 128-bin histogram of local oriented gradient magnitudes and orientations in the neighborhood of a salient point.

Speeded-Up Robust Features (SURF) The SURF descriptor [1] first obtains an orientation from the disk around a salient point. Then, a square neighborhood that is parallel to this orientation is derived. The neighborhood is further divided into four 4×4 patches. The features computed from these patches are concatenated into a 64-D feature vector. Compared to SIFT features [20], the lower dimensionality boosts the computational and matching speed.

4 Evaluation Experiments

In this study, a monocular visual-IMU odometry system [15] was used and five experiments were conducted using different routes. In each experiment, we tested different combinations of salient point detector and local descriptor. The GPS/IMU navigation unit data [10] was used as ground-truth, while the pure inertial method (referred to as IMU, whose navigation data was obtained by integrating acceleration and angular velocity) was used as a baseline. The Euclidean distance and the rotation angle were used to compute the position and orientation errors respectively.

4.1 The Monocular Visual-IMU Odometry System

Hu and Chen [15] proposed a monocular visual-IMU odometry system (see Fig. 1 for pipeline) based on the multi-state constraint Kalman filter (MSCKF) [24]. In this system, the trifocal geometry relationship [14] between three consecutive frames is used as camera measurement. Hence, the estimation of the 3D position of feature points is avoided. Also, the trifocal tensor model [14] is used to map the matched

feature points between the first two frames into the third frame. A "bucketing" method [17] is further used to choose a subset of the matched points. Finally, the Random Sample Consensus (RANSAC) [8] method is applied in order to reject outlier points.

We used the modified version [7] of the system [15]. The feature matching and outlier rejection module was replaced with a self-adaptive scheme in order to prevent the system from exceptionally crashing when insufficient inliers were returned. The feature matching algorithm introduced by Lowe [20] was utilized in this study.

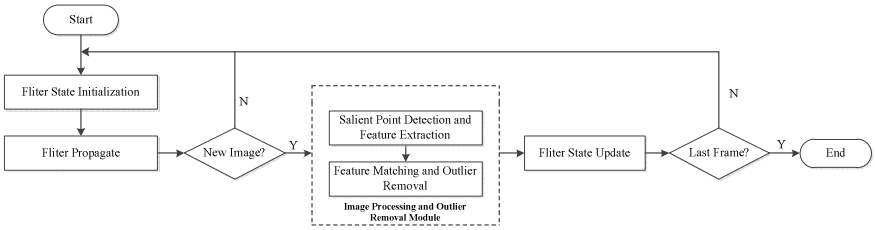

Fig. 1. The pipeline of the monocular visual-IMU odometry system [15] used in this study.

4.2 Dataset and Ground-Truth

In order to assess the detectors and descriptors fairly and explicitly, we selected five typical routes (see Figs. 2 and 3) from the KITTI dataset [10] according to the length and shape, the impact of the independent motion of other vehicles and pedestrians. (The configurations of the routes can be found in the supplementary material). The three factors are challenging for the existing visual odometry systems. Specifically, (1) Route 1 (Straight Line) and Route 2 (Quarter Turn) are on the urban road, in which other vehicles can be found; (2) Route 3 (Multiple Quarter Turns) and Route 4 (Multiple Curved Turns) are in the residential area, and are longer and more complicated; and (3) Route 5 (Loop Line) is also in the residential area and is a closed path. All images included in these routes are real-world driving sequences with the GPS/IMU ground-truth data. These images were captured at 10 fps using a recording platform equipped with multiple sensors [10]. We used the synchronized grayscale images in this study.

(a) Route 1 (b) Route 2 (c) Route 3 (d) Route 4 (e) Route 5

Fig. 2. Example images (corresponding images of the left color camera) of the five routes selected from the KITTI dataset [10].

4.3 Performance Measures

Since the measures based on the error of trajectory endpoints are usually misleading, we used the Root Mean Square Error (RMSE) measure computed from the position or orientation data. This measure has been extensively used for the navigation and autonomous driving systems. The RMSE measure is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]}{n}}, \quad (1)$$

where (x_i, y_i) means the ground-truth data while $(\hat{x}_i, \hat{y}_i)$ stands for the estimated data.

5 Experimental Results

In this section, we report the position and orientation RMSE measures derived in the five experiments. (More figures are provided in the supplementary material).

5.1 Route 1: Straight Line

Since Route 1 was gathered on the express way, the average speed involved was high. Table 1 lists the overall position and orientation RMSE values computed between the estimated trajectories obtained using different methods and the ground-truth trajectory. Fig. 3(a) further shows the ground-truth trajectory and the estimated trajectories obtained using IMU and the best descriptor for each detector in this experiment.

		BLOB& CORNER [11]	DOG [20]	FAST [27]	FAST HESSIAN [1]	HARRIS LAPLACE [21]
Position RMSE (m)	IMU	19.7514 (Salient point detector is not applicable)				
	HOG [4]	11.8107	136.1566	**11.0005**	52.1558	16.7228
	ICIMGP [7]	**8.8652**	57.4174	17.9052	**5.2729**	**10.5965**
	IMGP [33]	152.6954	33.1861	36.6821	45.0751	42.7403
	LIOP [34]	110.3710	96.4385	254.7819	355.2067	162.6288
	LM [18]	30.4017	11.1896	18.7077	50.2143	40.8369
	LSSD [32]	48.7660	221.8634	13.6568	238.2317	198.0097
	SIFT [20]	78.9520	**5.4794**	17.7088	16.7066	44.4767
	SURF [1]	60.8664	8.8896	15.9851	15.3137	26.1256
Orientation RMSE (deg)	IMU	2.4215 (Salient point detector is not applicable)				
	HOG [4]	2.3897	9.9110	**2.2235**	2.9339	2.4759
	ICIMGP [7]	**1.9690**	3.9674	2.3187	**1.7411**	**2.1637**
	IMGP [33]	2.3714	2.5128	2.5516	2.5728	2.7986
	LIOP [34]	3.0980	2.7016	4.2099	4.1999	2.2773
	LM [18]	2.3281	2.6195	2.5136	2.3939	2.8105
	LSSD [32]	2.1357	7.1138	2.4031	9.1309	6.5336
	SIFT [20]	2.7823	**2.3661**	2.5850	2.6820	2.8603
	SURF [1]	2.5319	2.4400	2.4226	2.3399	2.8067

Table 1. The overall position and orientation RMSE values computed between the ground-truth trajectory and the trajectories obtained using different methods on Route 1.

It can be seen from Table 1 that: (1) the joint use of Fast Hessian [1] and ICIMGP [7] yields the best performance; (2) ICIMGP [7] can also achieve proper performance when used with other detectors, except the DoG detector [20]; (3) the HOG [4] and LSSD [32] descriptors perform properly when combined with FAST [27] while SIFT

[20], SURF [1] and LM [18] generates promising results when used with DoG [20]; (4) IMGP [33] and LIOP [34] do not provide good performance. Especially, LIOP performs worse than all its counterparts; and (5) the IMU method performs properly.

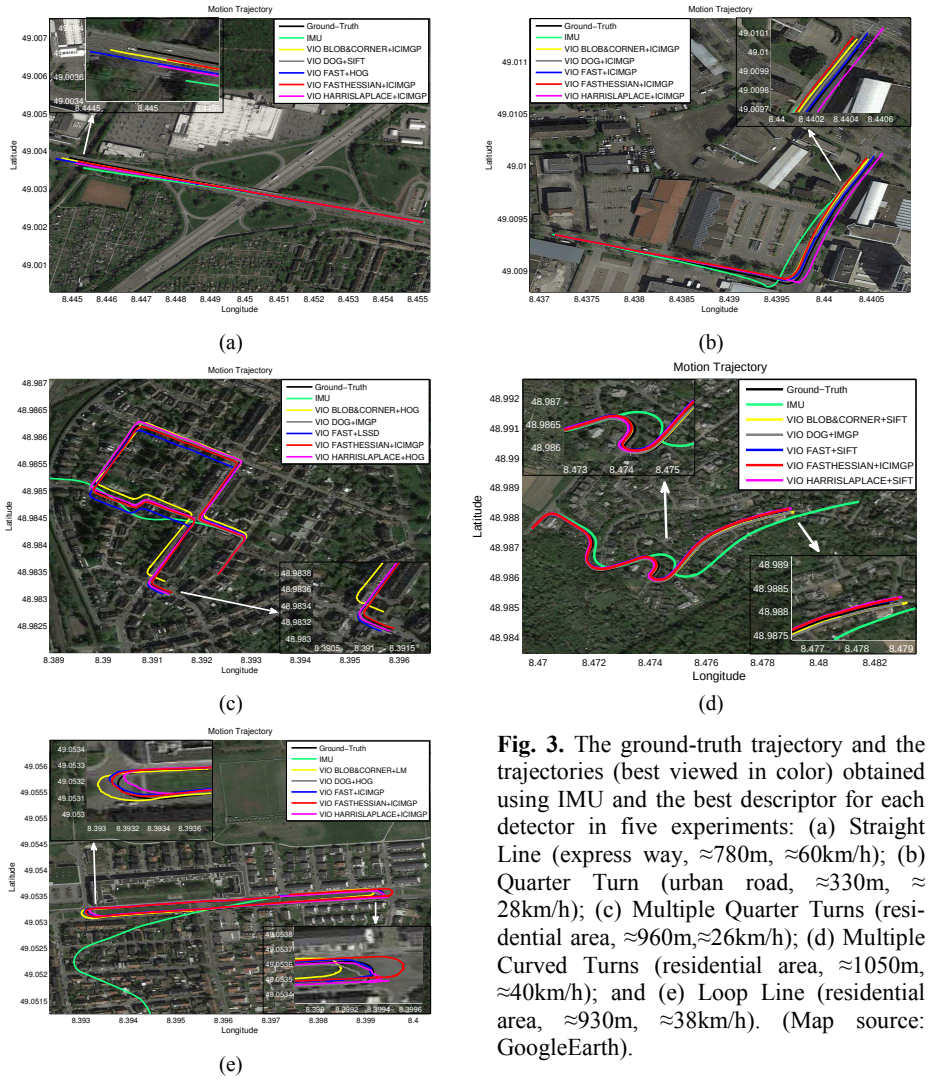

Fig. 3. The ground-truth trajectory and the trajectories (best viewed in color) obtained using IMU and the best descriptor for each detector in five experiments: (a) Straight Line (express way, ≈780m, ≈60km/h); (b) Quarter Turn (urban road, ≈330m, ≈ 28km/h); (c) Multiple Quarter Turns (residential area, ≈960m,≈26km/h); (d) Multiple Curved Turns (residential area, ≈1050m, ≈40km/h); and (e) Loop Line (residential area, ≈930m, ≈38km/h). (Map source: GoogleEarth).

5.2 Route 2: Quarter Turn

The route used in this experiment is a simple quarter turn. Table 2 reports the overall position and orientation RMSE values derived using IMU and different combinations of salient point detector and local descriptor. As can be seen, (1) the ICIMGP descriptor [7] performs the best, especially, when combined with the FAST detector [27]; (2) the combination of HOG [4] and FAST [27] is comparable to this result; (3)

SIFT [20] and SURF [1] yield proper performances; (4) the performance of LIOP [34] is better than that it obtained in Section 5.1 but is still worse than those of the other descriptors in most cases; (5) LM [18] and LSSD [32] perform well when combined with the Blob&Corner detector [11]; (6) IMGP [33] performs properly when used with the FAST [27] or DoG [20] detectors; and (7) the performance of the IMU method is proper. In addition, the ground-truth trajectory and the trajectories obtained using IMU and the best descriptor for each salient point detector are shown in Fig. 3(b).

		BLOB& CORNER [11]	DOG [20]	FAST [27]	FAST HESSIAN [1]	HARRIS LAPLACE [21]
Position RMSE (m)	IMU	19.2885 (Salient point detector is not applicable)				
	HOG [4]	5.1639	8.0799	4.1790	13.9958	8.9684
	ICIMGP [7]	**4.2406**	**5.8100**	**3.5711**	**4.9501**	**8.4941**
	IMGP [33]	25.5154	14.4854	10.8527	22.5591	21.9990
	LIOP [34]	22.6505	45.1004	35.8729	35.1246	18.4742
	LM [18]	8.6962	13.6695	14.1090	14.9695	24.6382
	LSSD [32]	7.4026	7.7328	7.4474	62.5453	15.2898
	SIFT [20]	13.1014	13.9049	13.5340	19.0370	14.2213
	SURF [1]	22.5823	9.0183	14.5184	16.0266	18.1139
Orientation RMSE (deg)	IMU	3.9190 (Salient point detector is not applicable)				
	HOG [4]	1.6664	1.8306	1.6076	1.7836	1.6327
	ICIMGP [7]	**1.4627**	**1.5223**	**1.4032**	**1.4785**	**1.5449**
	IMGP [33]	4.1985	2.0223	1.8827	3.9470	4.2121
	LIOP [34]	4.3496	4.2128	4.5956	4.1954	4.2518
	LM [18]	1.4296	1.6533	1.6976	1.7509	4.1327
	LSSD [32]	1.5367	1.4339	1.5818	9.8677	2.2354
	SIFT [20]	1.6483	1.5586	1.5179	2.9828	2.6253
	SURF [1]	3.9259	1.5555	1.6213	2.3997	2.5474

Table 2. The overall position and orientation RMSE values computed between the ground-truth trajectory and the trajectories obtained using different methods on Route 2.

5.3 Route 3: Multiple Quarter Turns

The route used in this experiment was captured in the residential area. Compared to Routes 1 and 2, this route is longer and more complicated. Table 3 lists the overall position and orientation RMSE values obtained using different methods. It can be observed that: (1) the best result is produced by the combination of Fast Hessian [1] and ICIMGP [7]; (2) HOG [4] also performs well, especially, when used with the Harris Laplace detector [21]; (3) the performance of IMGP [33] is even comparable to the best result when combined with DoG [20] and is proper when used with the other detectors; (4) LM [18] performs properly and yields its best performance when combined with the Harris Laplace detector [21] while SIFT [20] and SURF [1] perform properly in most cases; (5) LIOP [34] produces better results than it did on Routes 1 and 2, and yields its best performance when used with Harris Laplace [21]; (6) LSSD [32] provides proper performance when combined with Blob&Corner [11], DoG [20] or FAST [27]; and (7) the performance of the IMU method is worse than those of all the descriptors. Besides, the ground-truth trajectory and the trajectories obtained using IMU and the best descriptor for each salient point detector are shown in Fig. 3(c).

		BLOB& CORNER [11]	DOG [20]	FAST [27]	FAST HESSIAN [1]	HARRIS LAPLACE [21]
Position RMSE (m)	IMU	1540 (Salient point detector is not applicable)				
	HOG [4]	**9.0217**	17.1248	15.1831	9.2064	**4.9899**
	ICIMGP [7]	12.1206	7.7111	16.9728	**4.4340**	6.2895
	IMGP [33]	18.3862	**4.8790**	19.2680	20.4940	10.0586
	LIOP [34]	15.1372	39.5007	14.8697	15.4606	9.7781
	LM [18]	12.9940	14.2923	22.6583	14.3077	9.6142
	LSSD [32]	16.0244	11.8442	**12.4355**	37.3118	34.9911
	SIFT [20]	13.4720	9.2277	24.5953	10.8567	6.8956
	SURF [1]	32.1153	7.0529	25.2071	8.5964	6.3545
Orientation RMSE (deg)	IMU	11.2301 (Salient point detector is not applicable)				
	HOG [4]	**2.6607**	4.3870	2.5740	1.8785	**1.4192**
	ICIMGP [7]	2.7049	2.9338	2.6542	**1.3711**	1.7284
	IMGP [33]	2.8369	**2.7518**	2.8495	2.8208	1.7913
	LIOP [34]	2.7034	2.7309	2.6419	3.2402	1.7431
	LM [18]	2.7518	2.9264	2.6755	2.9101	1.7195
	LSSD [32]	2.7423	2.9747	**2.3698**	4.4367	3.2621
	SIFT [20]	2.6880	2.7993	2.7558	1.8593	1.7595
	SURF [1]	2.5756	2.8834	2.8303	1.8531	1.7349

Table 3. The overall position and orientation RMSE values computed between the ground-truth trajectory and the trajectories obtained using different methods on Route 3.

5.4 Route 4: Multiple Curved Turns

The route used in this experiment contains several curved turns. Table 4(a) lists the overall position and orientation RMSE values derived using different methods. It can be seen that: (1) the joint use of Fast Hessian [1] and ICIMGP [7] achieves the best result; (2) SIFT [20] yields the comparable performance to this result when used with FAST [27] and performs better than it did on Routes 1, 2 and 3; (3) HOG [4], SURF [1] and LM [18] perform properly while LSSD [32] only produces proper performance when used with Blob&Corner [11], DoG [20] or FAST [27]; (4) IMGP [33] yields its best performance when combined with DoG [20] and also performs properly when used with the other detectors; (5) the trajectories obtained using LIOP [34] suffer from the drift issue except when used with Harris Laplace [21] and are even worse than that obtained using IMU. Fig. 3(d) also shows the ground-truth trajectory and the trajectories derived using IMU and the best descriptor for each detector.

5.5 Route 5: Loop Line

A closed route is used in this experiment. Table 4(b) reports the overall position and orientation RMSE values computed between the trajectories obtained using different methods and the ground-truth data. As can be seen, (1) the combination of FAST [27] and ICIMGP [7] performs the best; (2) HOG [4] yields promising results except when it is used with Blob&Corner [11]; (3) LM [18], IMGP [33], SIFT [20] and SURF [1] generate proper performance while LSSD [32] only yields proper performance when used with DoG [20] or FAST [27]; (4) LIOP [34] performs properly when combined with the Blob&Corner [11], DoG [20] or Harris Laplace [21] detectors; and (5) the performance of IMU is the worst while it can be improved by being jointly used with local descriptors. In addition, Fig. 3(e) shows the ground-truth trajectory and the trajectories obtained using IMU and the best descriptor for each salient point detector.

		BLOB& CORNER [11]	DOG [20]	FAST [27]	FAST HESSIAN [1]	HARRIS LAPLACE [21]
Position RMSE (m)	IMU	86.6306 (Salient point detector is not applicable)				
	HOG [4]	28.1719	10.8526	8.0639	16.6777	18.0063
	ICIMGP [7]	14.5597	9.1290	15.3423	**6.5293**	14.3628
	IMGP [33]	18.0994	**8.1375**	15.3857	24.7288	28.3987
	LIOP [34]	133.3808	250.2128	164.4068	190.7820	33.4962
	LM [18]	12.1501	16.7098	8.3151	18.9811	18.6905
	LSSD [32]	13.3530	22.0602	11.6544	42.8612	50.3251
	SIFT [20]	**10.1651**	11.4728	**6.9466**	7.4098	**10.8258**
	SURF [1]	12.2692	8.2212	10.2553	12.2840	19.8987
Orientation RMSE (deg)	IMU	3.6691 (Salient point detector is not applicable)				
	HOG [4]	3.9045	2.6484	2.6698	2.7958	2.8306
	ICIMGP [7]	2.9066	2.6197	2.8062	**2.5820**	2.8042
	IMGP [33]	3.0646	**2.5938**	2.8138	3.4242	2.8374
	LIOP [34]	6.7869	8.6987	8.7326	7.9277	5.5268
	LM [18]	3.0063	2.7083	2.7701	2.7169	2.8940
	LSSD [32]	3.0072	2.8973	3.4745	6.7956	4.9426
	SIFT [20]	**2.7768**	2.6617	**1.5408**	2.6173	**2.7813**
	SURF [1]	2.8159	2.6166	2.8096	2.5472	2.8429

(a)

		BLOB& CORNER [11]	DOG [20]	FAST [27]	FAST HESSIAN [1]	HARRIS LAPLACE [21]
Position RMSE (m)	IMU	314.4739 (Salient point detector is not applicable)				
	HOG [4]	61.4032	**6.4047**	12.5497	29.8080	18.2804
	ICIMGP [7]	18.0249	8.9149	**4.5590**	**9.1430**	**6.9035**
	IMGP [33]	14.4585	15.2747	13.3215	35.9567	19.8309
	LIOP [34]	16.2077	14.4689	35.1219	51.8146	9.0775
	LM [18]	**12.2945**	8.6872	18.3456	13.8273	21.9467
	LSSD [32]	64.8867	17.0492	20.7358	64.2704	57.0051
	SIFT [20]	22.6322	6.7678	14.7502	14.0927	8.0781
	SURF [1]	27.8902	11.8585	11.1392	11.5467	35.0271
Orientation RMSE (deg)	IMU	9.7546 (Salient point detector is not applicable)				
	HOG [4]	3.8551	**2.4639**	3.5458	3.9671	3.5270
	ICIMGP [7]	3.5059	3.5002	**2.3885**	**2.6783**	**2.6362**
	IMGP [33]	3.2276	3.6311	3.3802	3.5993	3.4252
	LIOP [34]	3.2975	3.3220	3.4657	6.1507	2.7783
	LM [18]	**3.1867**	3.3307	3.2962	3.2579	3.3887
	LSSD [32]	3.7195	3.7232	3.4236	7.0683	3.6475
	SIFT [20]	3.3112	3.3340	3.3997	3.4244	2.7540
	SURF [1]	3.5203	3.4688	3.2930	3.5448	3.6012

(b)

Table 4. The overall position and orientation RMSE values computed between the ground-truth trajectory and the trajectories obtained using different methods on (a) Route 4 and (b) Route 5.

5.6 Summary

The performance of the descriptors varies when they are used with different detectors or on different routes. To summarize, a set of insights can be obtained as follows:

(1) In the five experiments, the best result is always produced by ICIMGP [7], especially, when it is used with the FAST [27] or Fast Hessian [1] detectors. It suggests that ICIMGP [7] is suitable for monocular visual-IMU odometry. Those promising results should be attributed to the fact that ICIMGP [7] encodes richer image characteristics than its counterparts that are normally extracted from gray level images;

(2) The HOG [4] and LSSD [32] descriptors perform properly when they are used with FAST [27]. However, their performance varies when used with other detectors;

(3) The DoG detector [20] is the best choice for IMGP [33]. In this case, IMGP [33] performs better than ICIMGP [7] on Routes 3 and 4. However, it does not yield promising results on the straight express way (Route 1). The similar finding can be

obtained for LIOP [34] when it is used with Harris Laplace [21]. These results show that gray level image patches are not sufficient for the use on the straight express way and probably need to be combined with other image characteristics (see ICIMGP);

(4) The LM [18], SIFT [20] and SURF [1] descriptors produce promising results when used with DoG [20] while their performances are not stable when used with the other detectors. Surprisingly, SURF [1] normally performs better when combined with DoG [20] than Fast Hessian [1] even if the latter was proposed for it; and

(5) According to the average position RMSE, Route 3 is the easiest (15.0±8.8) but Route 1 is the most difficult (65.2±79.2) for the detectors and descriptors tested here.

The above insights provide the meaningful guidelines for choosing the salient point detector and local descriptor in the monocular visual-IMU odometry applications.

We did not compare the computational speed of different detectors and descriptors because they were implemented in different programming languages. However, the time cost of feature matching depends on the dimensionality of the local descriptors extracted at the same salient points. Table 5 lists the dimensionality of the eight local descriptors. It can be seen that the dimensionality of the ICIMGP [7] descriptor is the highest while it did produce the best results in this study.

Descriptor	HOG [4]	ICIMGP [7]	IMGP [33]	LIOP [34]	LM [18]	LSSD [32]	SIFT [20]	SURF [1]
Dim.	279	968	121	144	48	36	128	64

Table 5. The dimensionality of the local descriptors examined in this paper.

6 Conclusions and Future Work

In this paper, we first reviewed five salient point detectors and eight local descriptors. Then, we conducted a comparative evaluation study on different combinations of detector and descriptor using a unified monocular visual-IMU odometry framework and five typical routes [10]. To our knowledge, this is the first extensive comparative evaluation on salient point detectors and local descriptors for monocular visual-IMU odometry by using these explicit types of routes. The experimental results can be used as a set of baselines in the further research. The analysis of these results also provides a set of useful insights to the community, which could be used as guidelines for the selection of the detector-descriptor combinations.

However, the experiments presented in this paper are not exhaustive and only investigate different combinations of detector and descriptor using a monocular visual-IMU odometry system. In the next stage of this study, we will tune the parameters of the detectors and descriptors and also test a different monocular visual odometry system in order to augment the results reported in this paper.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (NSFC) (No. 61271405).

References

1. Bay, H., Ess, A., Tuytelaars, T., Van G, L.: Speeded-up robust features (SURF). Computer vision and image understanding, 110 (3): 346-359 (2008)
2. Bloesch, M., Omari, S., Hutter, M., Siegwart, R.: Robust visual inertial odometry using a direct EKF-based approach. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 298-304 (2015)
3. Corke, P., Lobo, J., Dias, J.: An introduction to inertial and visual sensing. The International Journal of Robotics Research, 26(6): 519-535 (2007)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893 (2005)
5. Davison, A., Reid, I., Molton, N., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6): 1052-1067 (2007)
6. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features, In: British Machine Vision Conf (2009)
7. Dong, X., He, B., Dong, X., Dong, J.: Monocular visual-IMU odometry using multi-channel image patch exemplars. Multimedia Tools and Applications (submitted)
8. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6): 381-395 (1981)
9. Gauglitz, S., Höllerer, T., Turk, M.: Evaluation of interest point detectors and feature descriptors for visual tracking. International Journal of Computer Vision, 94(3): 335-360 (2011)
10. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. The International Journal of Robotics Research, pp. 1229-1235 (2013)
11. Geiger, A., Ziegler, J., Stiller, C.: StereoScan: Dense 3d reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium, pp. 963-968 (2011)
12. Govender, N.: Evaluation of feature detection algorithms for structure from motion. Council for Scientific and Industrial Research, Technical Report (2009)
13. Harris, C., and Stephens, M.: A combined corner and edge detector. In: Alvey vision conference (1988)
14. Hartley, R., and Zisserman, A.: Multiple view geometry in computer vision, 2nd ed. Cambridge University Press, 2008.
15. Hu, J., and Chen M.: A sliding-window visual-IMU odometer based on tri-focal tensor geometry. In: IEEE International Conference on Robotics and Automation, pp. 3963-3968 (2014)
16. Jiang, Y., Xu, Y., Liu, Y.: Performance evaluation of feature detection and matching in stereo visual odometry. Neurocomputing, 2013, 120: 380-390
17. Kitt, B., Geiger, A., and Lategahn, H.: Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In: IEEE Intelligent Vehicles Symposium, pp. 486-492 (2010)
18. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision, 43(1): 29-44 (2001)
19. Li, M., Mourikis A I.: Improving the accuracy of EKF-based visual-inertial odometry. In: IEEE International Conference on Robotics and Automation, pp. 828-835 (2012)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91-110 (2004)

21. Mikolajcyk, K., Schmid, C.: An affine invariant interest point detector. In: International Conference on Computer Vision, pp. 128-142 (2002)
22. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision, 60(1): 63-86 (2004)
23. Mikolajczyk, K., and Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Machine Intell, 27(10), 1615-1630 (2005)
24. Mourikis, A., and Roumeliotis, S.: A multi-state constraint Kalman filter for vision-aided inertial navigation. In: IEEE International Conference on Robotics and Automation, pp. 3565-3572 (2007)
25. Nilsson, J. O., Zachariah, D., Jansson, M., and Handel, P.: Realtime implementation of visual-aided inertial navigation using epipolar constraints. In: IEEE Position Location and Navigation Symposium (PLANS), pp. 711-718 (2012)
26. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 652-659 (2004)
27. Rosten, E., Porter, R., Drummond, T.: Faster and better: A machine learning approach to corner detection. IEEE Trans. Pattern Anal. Mach. Intell, 32(1): 105-119 (2010)
28. Scaramuzza, D., Fraundorfer, F.: Visual odometry [Tutorial]. IEEE Robotics & Automation Magazine, 18 (4), 80-92 (2011)
29. Scaramuzza, D., Fraundorfer, F., Siegwart, R.: Real-time monocular visual odometry for on-road vehicles with 1-point ransac. In: IEEE International Conference on Robotics and Automation, pp. 4293-4299 (2009)
30. Schmidt, A., Kraft M., and Kasiński A.: An evaluation of image feature detectors and descriptors for robot navigation. Computer Vision and Graphics, pp. 251-259 (2010)
31. Schmid, C., Mohr, R., and Bauckhage, C.: Evaluation of interest point detectors. International Journal of computer vision, 37(2): 151-172 (2000)
32. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8 (2007)
33. Varma. M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. IEEE Trans. Pattern Anal. Mach. Intell, 31, 2032-2047 (2009)
34. Wang, Z., Fan, B., Wu, F.: Local intensity order pattern for feature description. In: International Conference on Computer Vision, pp. 603-610 (2011)
35. Yang, Y., Newsam, S.: Geographic image retrieval using local invariant features. IEEE Transactions on Geoscience and Remote Sensing, 51(2), 818-832 (2013)