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

Xingshuai Dong, Xinghui Dong, Junyu Dong, and Huiyu Zhou

***Abstract*—Monocular visual-IMU (Inertial Measurement Unit) odometry has been widely used in various intelligent vehicles. As a popular technique, detector-descriptor based visual-IMU odometry is effective and efficient due to the fact that local descriptors are robust against occlusions, background clutter and abrupt content changes. However, to our knowledge, there is not a comprehensive and comparative evaluation study on the performance of different combinations of detectors and descriptors recently developed. In order to bridge this gap, we conduct such a comparative study in a unified framework. In particular, six typical routes with different lengths, shapes and road scenes are selected from the well-known *KITTI* dataset. Firstly, we evaluate the performance of different combinations of salient point detectors and local descriptors using the six routes. Finally, we tune the parameters of the best detector or descriptor obtained for each route, to achieve better results. This study provides not only comprehensive benchmarks for assessing various algorithms, but also instructive guidelines and insights for developing detectors and descriptors to handle different road scenes.**

***Index Terms*—Evaluation, odometry, monocular visual-IMU odometry, navigation, local descriptors, salient point detectors.**

## I. Introduction

AMONG the commonly used techniques for ego-motion estimation, Visual Odometry (VO) [1-5] plays important roles in the studies of computer vision, intelligent vehicles and robotics. By matching the consecutive video frames acquired by onboard cameras [5], VO incrementally estimates the pose of a vehicle. In this context, the information of the vehicle's motion state relative to the surrounding environment is essential for assessing the risk of collision in autonomous driving and advanced driver-assistance systems (ADAS).

In terms of the onboard camera, VO can be classified into two types: stereo and monocular [5]. A stereo VO system not only owns a complex architecture, but also generates a monocular system with the increase of the distance between the lens and the object. In contrast, the monocular VO system is simple and can be easily used in real systems. Moreover, the inertial measurements obtained from the IMU and visual data are complementary [6]. Thus, odometry system based on these measurements, i.e., Visual-IMU Odometry (VIO), are able to boost the accuracy and reliability of motion estimation. This advantage is of great significance to the development of intelligent vehicles. In this paper, we therefore focus on monocular visual-IMU odometry.

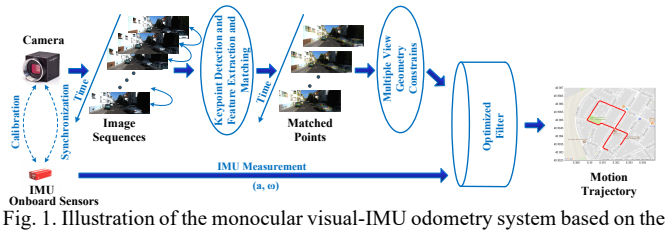



Fig. 1. Illustration of the monocular visual-IMU odometry system based on the detector-descriptor scheme.

Local descriptors can be defined as the description of the distinct patterns or structures contained in a local image region. As local descriptors are robust against occlusions, background clutter and other changes [7], they have been extensively used in various vision systems, such as visual tracking [8], visual odometry [9] and visual Simultaneous Localization and Mapping (visual-SLAM) [10]. To achieve efficient image matching, local descriptors are often extracted at salient point locations. In this case, both salient point detection and descriptor extraction play important roles in detector-descriptor based VIO systems. However, different detector-descriptor combinations may generate significantly different performance, especially, when different road scenes and route shapes are encountered. For practical applications, it is important to know which combination is superior to others under a certain circumstance. In this context, an extensive, comparative evaluation of various combinations of detectors and descriptors using a unified framework is necessary, to obtain useful benchmarks and guidelines for choosing or developing detectors and descriptors.

To our knowledge, nevertheless, there is not comprehensive evaluation research conducted to compare the performance of different detector-descriptor combinations in the scenario of monocular visual-IMU odometry. Compared with [11], our

J. Dong is supported by the National Natural Science Foundation of China (NSFC) (No. 61271405, 41576011). (Corresponding author: Xinghui Dong).

Xingshuai Dong was with the Department of Electronic Engineering, Ocean University of China, Qingdao, China (e-mail: dongxingshuai@gmail.com).

Xinghui Dong is with the Center for Imaging Sciences, the University of Manchester, M13 9PT, UK (e-mail: xinghui.dong@manchester.ac.uk).

J. Dong is with the Department of Computer Science, Ocean University of China, Qingdao, China (e-mail: dongjunyu@ouc.edu.cn).

H. Zhou is with Department of Informatics, University of Leicester, LE1 7RH, UK (e-mail: hz143@leicester.ac.uk).

current paper (1) provides more details of the related work and the investigated salient point detectors and descriptors; (2) utilizes more routes; (3) examines typical convolutional neural network (CNN) features; (4) investigates the impact of the parameters in the best salient point detector or local descriptor for each route; and (5) conducts more extensive analysis and discussion of experimental results.

The main contributions of this study can be summarized as follows. (1) Without loss of generality, we deliberately select six typical routes containing different lengths, shapes and road scenes from the known *KITTI* dataset [12] and use these data for monocular VIO. The routes can be used by the community for further research. (2) We review five salient point detectors and nine local descriptors. (3) We apply CNN features to monocular VIO. (4) We conduct a comprehensive evaluation study on various detector-descriptor combinations for monocular VIO. This study provides the community with a series of useful benchmarks and instructive guidelines.

The rest of this paper is organized as follows. The related work is reviewed in Section II. In Section III, the detectors and descriptors are surveyed and their implementation details are described. The experimental setup is introduced in Section IV. The results obtained using different detector-descriptor combinations are reported and discussed in Section V. In Section VI, the parameters of the best detector or descriptor for each route are tuned. Our conclusions are given in Section VII.

## II. Related Work

### A. Salient Point Detectors

To accelerate the computational speed of matching two images, many salient point detectors have been proposed. Using the image gradient matrix, Harris and Stephens [13] developed a corner detector. Inspired by the Harris detector, Shi and Tomasi [14] introduced a different corner detector that has a more principled feature selection criterion, namely, Good Features To Track Detector. The Features from Accelerated Segment Test (FAST) corner detector [15] were designed on the basis of a circle of pixels surrounding the candidate corner pixel. Mair *et al*. [16] proposed a derivation of the FAST detector, i.e., Adaptive and Generic Accelerated Segment Test (AGAST). This detector utilizes backward induction to create an optimal decision tree in order to improve the computational efficiency. On the basis of AGAST, Binary Robust Invariant Scalable Keypoints (BRISK) [17] were detected in a continuous scale space. Mikolajczky and Schmid [18] also designed the scale and affine invariant Harris-Laplace corner detector.

While these corner detectors have high efficiency, the distinctiveness of detected points is poor. On the other hand, blob detectors, e.g., the Difference of Gaussian (DoG) [19] and Fast Hessian [20] detectors, normally produce more distinctive points [5]. Especially, the latter uses box filters to approximate the Laplacian of Gaussian (LoG) functions, and achieves high detection speeds. Agrawal *et al*. [21] applied center-symmetric local binary patterns as an alternative to the orientation histogram approach of SIFT [19]. The Simple Blob Detector was implemented in the OpenCV [22] library. This detector extracts the connected regions from a binary image. Only the regions whose area lies in a given range are treated as blobs. Moreover, a hybrid detector was developed in order that both sorts of points can be detected [23]. Matas *et al*. [24] extracted regions using a watershed segmentation method.

All of the abovementioned detectors treat points or regions as isolated in the image. In contrast, Dense Feature Detector [25] is a hybrid approach using dense sampling on a regular grid and interest point detection. This detector first samples image patches using a regular grid, and then refines their positions and scales by an optimized interestingness measure. It generates a set of feature points on a semi-regular grid, densely covering the entire image similar to the case of dense sampling.

### B. Local Descriptors

Once salient points have been detected, a feature vector is usually extracted from the highlighted points for matching or tracking. Lowe [19] developed the Scale-Invariant Feature Transform (SIFT) descriptor using local gradient histograms. Similarly, Histogram of Orientation Gradient (HOG) [26] was designed based on locally normalized histograms of the gradient orientation data. Bay *et al*. [20] developed the Speeded-Up Robust Features (SURF) descriptor, providing the faster efficiency than SIFT. Sampling an image patch at a point is usually used as a feature descriptor [27]. In [28], Leung and Malik used the features computed using a filter bank for texture classification. The Local Self-Similarity Descriptor (LSSD) was proposed in [29]. In addition, the Local Intensity Order Pattern (LIOP) descriptor was introduced by Wang *et al*. [30]. Recently, Hariharan *et al*. [31] applied the hyper-column features computed at multiple convolutional layers of a pre-trained convolutional neural network (CNN) to object segmentation, and obtained state-of-the-art results.

Binary descriptors are different from the aforementioned real-valued descriptors by directly constructing strings via the pixel-level comparison. In this context, the BRIEF descriptor [32] utilizes a set of binary intensity tests in order to extract strings from an image patch. However, BRIEF is sensitive to the in-plane rotation and scale changes. To solve this problem, Leutenegger *et al*. [17] developed the BRISK descriptor. Similarly, Rublee *et al*. [33] proposed an improved BRIEF descriptor. Since this descriptor is usually combined with a multi-scale FAST detector, they are named Oriented FAST and Rotated BRIEF (ORB).

### C. Detector-Descriptor Based Monocular Visual (-IMU) Odometry

It has been proved that local descriptors are robust against occlusions, background clutter and other changes. As a result, they were widely used in visual (-IMU) odometry. Nister *et al*. [9] used $11 \times 11$ image patches cropped around the Harris corner points for the VO tasks. Leutenegger *et al*. [34] tightly integrated inertial measurements into the keyframe-based visual odometry using a customized multi-scale SSE-optimized Harris corner detector and BRISK descriptor. Bloesch *et al*. [35] implemented a monocular VIO system based on the FAST

detector and multi-level patch features. Compared with the above corner points, blob points are more distinctive and redetected in the VO applications [5]. Of the most successful local descriptors, SIFT was widely utilized in the application of monocular VIO [36], [37]. In [38], a monocular visual-aided inertial navigation system was developed based on SURF [20]. Nevertheless, only gray level images were used for extraction of the aforementioned descriptors. For purposes of using richer image characteristics, three different multi-channel descriptors were adapted for monocular VIO [39].

### *D. Comparative Evaluations of Salient Point Detectors and Local Descriptors*

Salient point detectors and local descriptors have been evaluated in several publications. By changing the conditions of camera noise, image orientation, illumination, scale and viewpoint, Schmid *et al*. [40] assessed several detectors. Mikolajczyk and Schmid further evaluated different affine invariant detectors [41] and descriptors [7]. In [42], Heinly *et al*. compared three types of binary descriptors: BRIEF [32], ORB [33] and BRISK [17] with two baselines of the SIFT [19] and SURF [20] descriptors. It was found that SIFT was more robust to changes in affine images and perspectives than its counterparts. Zhang *et al*. [43] compared local descriptors for texture and object classification. The performance of twelve detectors was compared by Bostanci *et al*. [44] using the analysis of variance. Yang and Newsam [45] also compared local descriptors for image retrieval. Besides, evaluation studies were conducted for visual tracking [46], [47], [48].

On the other hand, ego-motion estimation tasks [49], [50], [51] have been used to assess detectors and descriptors. Benseddik *et al*. [52] assessed SIFT [19] and SURF [20] for monocular VO. Scaramuzza *et al*. [53] examined the performance of SIFT, Harris and Kanade-Lucas-Tomasi (KLT) for monocular VO. Jiang *et al*. [54] benchmarked the performance of several detectors and descriptors for stereo VO. For the application of visual-SLAM, Gil *et al*. [55] investigated the repeatability of a set of detectors and the invariance and distinctiveness of different local descriptors under various perceptual conditions. Moreover, Diosi *et al*. [3] evaluated three salient point detectors, including DoG [19], Harris-Laplace [18] and Maximally Stable Extremal Regions (MSER) [24], for the task of visual path following in outdoor urban surroundings.

However, only a small number of detectors and/or descriptors were evaluated in the above studies. Also, the involved datasets were smaller and not representative to various road scenes and route shapes for the VO task. To address these problems, we therefore perform a comprehensive (using more detectors and descriptors) evaluation based on a well-established monocular visual-IMU odometry system together with six representative real world routes. To the authors' knowledge, this study is the first attempt to comprehensively evaluate different detector-descriptor combinations for the task of monocular visual-IMU odometry.

## III. Salient Point Detectors and Local Descriptors

It is known that local descriptors are insensitive to occlusions, background clutter and other changes. Therefore, they have been widely used in visual (IMU) odometry systems [9], [35-39]. Salient points are normally detected before local descriptors are extracted, to reduce the number of the points required for matching two images. Popular salient point detection methods include corner and blob detectors. Although corner detectors can be efficiently computed, they own low distinctiveness. Comparably, blob detectors are more distinctive while the detection speed is slower.

Image sequences captured by the onboard camera can be utilized as the input data of VIO. These images usually contain illumination changes, motion blurring, perspective transformation and independently moving objects. In order to diminish the influence of moving objects, the majority of the detected features should fall in the static background or distribute across the whole image. However, the detected features using the MSER detector tend to distribute in parts of an image [56]. Heinly *et al*. [42] found that the matching precision, matching score and recall values of BRIEF [32], ORB [33] and BRISK [17] are not satisfying and even far from the baseline under the conditions of affine and perspective transforms. According to [47], the Harris [13], CenSure [21] and Good Features To Track [14] detectors cannot properly cope with scale changes, illumination changes and motion blur, which occur in many natural images. In addition, Simple Blob Detector [22] is absolutely controlled by parameters and hence it may not guarantee the number of detected features. However, Dense Feature Detector [25] requires intensive computation.

In this work, we select two blob detectors: Difference of Gaussian (DoG) [19] and Fast Hessian (FH) [20] and two corner detectors: Features from Accelerated Segment Test (FAST) [15] and Harris-Laplace [18] for our evaluation study. Besides, we select a hybrid blob and corner detector [23]. Regarding local descriptors, nine methods are selected because they are commonly used in computer vision and visual odometry. We will review these detectors and descriptors and describe their implementation details as follows.

### *A. Salient Point Detectors*

In this subsection, we briefly describe the five salient point detectors.

#### *1) Difference of Gaussian (DoG)*

Lowe [19] used the scale-space extrema of the Difference of Gaussian (DoG) function to detect salient points. For the sake of obtaining invariance against scale changes, an image is convolved with the DoG function at multiple scales. The scale space of the image is expressed using the formula:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (1)$$

where $G(x, y, \sigma)$ is a Gaussian function with the scale of $\sigma$, $I(x, y)$ is the image, and $*$ denotes the convolution operation.

The DoG function can be calculated based on the difference between two neighboring scales split by a constant value:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \quad (2)$$

where $k$ represents the sampling interval in the scale space. If $D(x, y, \sigma)$ is a local maximum or minimum, $(x, y, \sigma)$ is treated as a feature region surrounding the salient point $(x, y)$.

*2) Fast Hessian (FH)*

Using the determinant of the Hessian matrix, the Fast Hessian detector was developed [20]. The $\sigma$ scale Hessian matrix can be computed using the following formula:

$$H(x,y,\sigma)=\begin{bmatrix}\frac{\partial^2}{\partial x^2}G(\sigma)*I(x,y) & \frac{\partial}{\partial x}\frac{\partial}{\partial y}G(\sigma)*I(x,y)\\ \frac{\partial}{\partial x}\frac{\partial}{\partial y}G(\sigma)*I(x,y) & \frac{\partial^2}{\partial y^2}G(\sigma)*I(x,y)\end{bmatrix}, \quad (3)$$

where $I(x,y)$ is an image, and $\frac{\partial}{\partial x}G(\sigma)$ and $\frac{\partial^2}{\partial x^2}G(\sigma)$ are the 1st- and 2nd-order Gaussian derivative functions respectively.

To accelerate the computational speed, box filters are used to approximate the Laplace of Gaussian functions. The approximated determinant of the Hessian matrix is defined as:

$$det\left(H_{approx}\right)=D_{xx}D_{yy}-(0.9D_{xy})^2, \quad (4)$$

where $D_{xx}$, $D_{yy}$ and $D_{xy}$ are the box filter approximations in the directions of $x$, $y$ and $xy$ respectively. Salient points can be produced from an image by detecting the local maximum of $det\left(H_{approx}\right)$ across different locations and scales.

*3) Features from Accelerated Segment Test (FAST)*

The FAST detector that Rosten *et al*. [15] introduced works on a discretized circle which consists of 16 pixels and centers at the candidate corner point $p$. When a continuous arc containing at least nine pixels that are brighter than the candidate pixel $I_p+s$ ($s$ is a threshold), or darker than the candidate pixel $I_p-s$, exists, the point $p$ is regarded as a corner point. A decision tree is further trained in order that fewer candidate pixels are examined and higher efficiency is achieved.

Non-maximal suppression is not applicable because FAST does not utilize corner response functions. Rosten *et al*. [15] produced a score for each potential salient point using

$$V=max(\sum_{x\epsilon S_{bright}}|I_{p\to x}-I_p|-t,\sum_{x\epsilon S_{dark}}|I_p-I_{p\to x}|-t), \quad (5)$$

where $S_{bright}$ and $S_{dark}$ are the subsets of the pixels on the circles that are smaller and larger than $p$ by $t$ respectively.

*4) Harris-Laplace (H-L)*

The Harris-Laplace detector [18] uses the scale-adapted Harris function to identify salient points in the scale space. The typical scale of a local image pattern is derived via finding the extremum of the Laplacian function across multiple scales. A scale is typical in the quantitative standpoint as it measures the scale where the maximal similarity between the local pattern and the detector is achieved. The Harris-Laplace detector consists of two steps: multi-scale point detection and an iterative algorithm implemented for identifying the location and scale of the salient points.

*5) Blob and Corner (B&C)*

In [23], Geiger *et al*. proposed a hybrid salient point detector. An image was first convolved with the blob and corner masks in order to identify stable salient points. Then, the convolved maps were processed using non-maximum and non-minimum suppressions. In total, four sets of points were obtained, i.e., "blob min", "blob max", "corner min" and "corner max".

### *B. Local Descriptors*

The nine local descriptors are briefly reviewed below.

*1) Histogram of Oriented Gradients (HOG)*

The HOG descriptor [26] is extracted from the normalized gradient orientation histograms computed in a dense grid. It first divides an image into small connected spatial blocks. Each block is further partitioned into cells. A gradient orientation histogram is computed within each cell. Finally, the histograms computed from each block are collapsed into a feature vector.

*2) Hyper-column CNN Features (HC-CNN)*

The hyper-column features that Hariharan *et al*. [31] proposed are extracted at multiple layers of a pre-trained CNN in terms of a pixel location. Given an image, a series of feature maps are derived at each layer after this image is fed to the CNN. The feature maps are upsampled into the original resolution of the image using interpolation. For a pixel location, the features at the corresponding locations in all the feature maps are combined into a hyper-column feature vector. In comparison to the other CNN features, e.g., the fully-connected layer features, hyper-column features encode the characteristics of pixels and contain more precise localization information.

*3) Image Patches (IMGP)*

A simple representation of a pixel is the image patch sampled around this pixel. Especially, the speed of image patch sampling is faster than that of extraction of other descriptors, such as SIFT [19] and SURF [20]. Also, image patches indeed preserve original image characteristics and encode the detailed image information [27]. Compared with the full images, local patches often encounter less distortion. Therefore, it is easier to define the similarity between two local patches.

*4) Integral Channel Image Patches (ICIMGP)*

Dollár *et al*. [57] introduced a series of integral channels. These channels comprise color or gray level channel, the gradient magnitude channel and six different gradient histogram channels. In comparison with the pure gray level or color channels, the integral channels are more diverse but heterogeneous. On the basis of these channels, Dong *et al*. [39] adapted integral channel image patch features. First, the image patch was sampled around a pixel in each channel. Second, each patch was individually $L_2$ normalized. Finally, all the patches were concatenated into a single feature vector.

*5) Leung-Malik (LM) Filter Bank*

In total, 36 1st- and 2nd-order Gaussian derivative filters built at six orientations and three scales, eight Laplace of Gaussian filters and four Gaussian filters are included in the LM filter bank [28]. By convolving with an image, a 48-dimensional feature vector is produced with regard to each pixel.

*6) Local Intensity Order Pattern (LIOP)*

In [30], Wang *et al*. introduced the LIOP descriptor. This descriptor captures not only the local ordinal data of each pixel but also the overall ordinal data. The LIOP descriptor was developed by assuming that the relative order of the pixel's gray levels is constant when the gray level monotonically varies. In terms of each image patch sampled around a pixel, LIOP first divides it into sub-regions according to the overall ordinal data. Then, LIOP is calculated over the patch corresponding to each pixel. An ordinal bin is derived from the LIOPs calculated in each sub-region. Finally, all the bins are collapsed into a LIOP feature vector.

*7) Local Self-Similarity Descriptor (LSSD)*

The LSSD was developed by Shechtman and Irani in [29].

TABLE I
THE PARAMETER VALUES OF FIVE SALIENT POINT DETECTORS.

| Detector | Parameters | Values |
|---|---|---|
| Difference of Gaussian [19] | Octaves | 4 |
| | Levels Per Octave | 3 |
| | $\sigma_0$ | 1.6 |
| Fast Hessian [20] | Octaves | 4 |
| | Threshold $T$ | 0.2 |
| | Sampling Step $n$ | 2 |
| FAST [15] | Threshold $T$ | 30 |
| Harris-Laplace [18] | Peak Threshold | 0.000002 |
| | Edge Threshold | 10 |
| Blob and Corner [23] | Window Size | $13 \times 13$ |

Given a pixel location $q$, an image patch centered at this location is cropped. Then, the patches around the pixels within a larger image region around $q$ are sampled and compared with the patch based on the sum of square differences (SSDs) between patch colors/intensities. The SSDs are normalized and transformed into a correlation surface, which is defined as:

$$S_q(x,y) = exp\left(-\frac{SSD_q(x,y)}{max(var_{noise}, var_{auto}(q))}\right), \quad (7)$$

where $var_{noise}$ denotes the acceptable photometric variation, and $var_{auto}(q)$ is the maximal variance of the differences between the patches sampled based on the region size. The correlation surface is converted into $n_r$ radial bins and $n_\theta$ angular bins in the log-polar space. Finally, they are linearly stretched into $[0, 1]$, which are comprised of the descriptor.

*8) Scale-Invariant Feature Transform (SIFT)*

In order to achieve invariance to image rotation, SIFT [19] first assigns an orientation to each detected salient point. Then, features are extracted for normalized image patches based on the 3-D histogram of the gradient location and orientation. The gradient location is quantized into a $4 \times 4$ location grid and the orientation is quantized into eight bins. In total, a 128-D feature vector is generated at each pixel location.

*9) Speeded-Up Robust Features (SURF)*

The SURF descriptor [20] first obtains an orientation for each salient point. A square region that is parallel to this orientation is derived around the point. Furthermore, the region is split into $4 \times 4$ sub-regions. All the features extracted from these sub-regions are concatenated into a 64-dimensional feature vector. Compared with the SIFT descriptor, the lower dimensionality boosts the computational and matching speeds.

### C. Implementation Details

We use the source code published together with the original literature if only this is applicable; otherwise, we implement the algorithm according to the literature. For the parameters of the five salient point detectors and the nine local descriptors, we use the optimal values reported in the original literature if there are not specific statements in the rest of this paper, to derive unprejudiced evaluation results. The values of the parameters of these detectors and descriptors are shown in Tables I and II.

## IV. EXPERIMENTAL SETUP

To assess the performance of different detector-descriptor combinations, we conduct an evaluation study. The adapted version [39] of an existing monocular VIO system [37] is used. Especially, six representative routes covering different lengths,

TABLE II
THE PARAMETER VALUES OF NINE LOCAL DESCRIPTORS

| Descriptor | Parameters | Values |
|---|---|---|
| HOG [26] | Block Size | $15 \times 15$ |
| | Cell Size | $5 \times 5$ |
| | Number of Orientations | 9 |
| HC-CNN [31] | Model | VGG16[58]-places205[59] |
| | Layers | Conv-1, 2, 3, 4 & 5 |
| IMGP [27] | Patch Size | $11 \times 11$ |
| ICIMGP [39] | Number of Channels | 8 |
| | Patch Size | $11 \times 11$ |
| LM [28] | $\sigma_x$ | $\sqrt{2}$, 2, $2\sqrt{2}$ |
| LIOP [30] | Number of Spatial Bins | 6 |
| | Number of Neighbors | 4 |
| | Sampling Radius $R$ | 6 |
| LSSD [29] | Patch Size | $5 \times 5$ |
| | Region Radius | 40 |
| | $n_r$ | 12 |
| | $n_\theta$ | 3 |
| SIFT [19] | Sub-window Size | $4 \times 4$ |
| | Number of Bins Per Window | 8 |
| SURF [20] | Window Size | $4 \times 4$ |

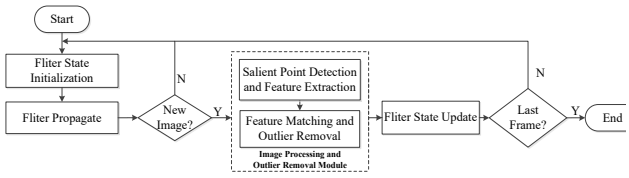



Fig. 2. Pipeline of the monocular visual-IMU odometry system [37].

shapes and road scenes are selected and used in this study. The GPS/IMU localization unit data is utilized as the ground-truth data. The IMU data derived via fusing the acceleration and angular velocity with time is employed as the benchmark. In addition to the popular Root Mean Square Error (RMSE) metric, we also utilize a Segment Based End Point Error (SEPE) [60] metric and the *Hausdorff* distance [61] as performance measures.

### A. The Monocular Visual-IMU Odometry System

Using the Multi-state Constraint Kalman Filter (MSCKF) [36], a moving-window monocular VIO system (see Fig. 2 for the pipeline) was developed by Hu and Chen [37]. Specifically, the camera measurement is obtained based on the trifocal geometry relationship [62] between three consecutive frames. As a result, it is not necessary to estimate the 3-D position of the feature points. The matched feature points between the first two frames are mapped into the third frame using a trifocal tensor model [62]. Furthermore, a subset is selected from these points using the "bucketing" approach [63]. Finally, outlier points are filtered through Random Sample Consensus (RANSAC) [64].

In this study, we utilize the adapted version [39] of the aforementioned system [37]. Compared with the original system, different detectors and descriptors can be incorporated into this version. The feature matching and outlier rejection unit was replaced by a self-adaptive module. This module is able to prevent the system from crashing in the case that inadequate inliers are obtained. In addition, we utilize the feature matching method that Lowe [19] proposed.

### B. Dataset and Ground-Truth

For the purpose of fairly and explicitly assessing the detectors

(a) Route 1 (b) Route 2 (c) Route 3 (d) Route 4 (e) Route 5 (f) Route 6

Fig. 3. Example images (captured using the left camera) of the six routes selected from the *KITTI* dataset [12]: (a) a straight line on the express way, (b) a straight line in the residential area, (c) a quarter turn on the urban road, (d) multiple quarter turns in the residential area, (e) multiple curved turns in the residential area, and (f) a close loop in the residential area.

TABLE III
DETAILS OF THE SIX ROUTES USED IN THIS STUDY.

| Route No. | Sequence No. in the KITTI Dataset [12] | No. of the Start Point Image (The Left Color Camera Image) | No. of the End Point Image (The Left Color Camera Image) |
|---|---|---|---|
| 1 | 2011_09_26_drive_0101_sync | 0000000361 | 0000000816 |
| 2 | 2011_09_26_drive_0023_sync | 0000000090 | 0000000472 |
| 3 | 2011_09_26_drive_0009_sync | 0000000000 | 0000000405 |
| 4 | 2011_10_03_drive_0027_sync | 0000000301 | 0000001610 |
| 5 | 2011_10_03_drive_0034_sync | 0000000000 | 0000000899 |
| 6 | 2011_09_30_drive_0020_sync | 0000000000 | 0000000834 |

and descriptors, we select six representative routes from the known *KITTI* dataset [12] by considering the length, shape, road scene (including the environment and the impact of the independent motion of pedestrians and other vehicles) and the speed of the recording platform. These aspects are challenging for the established VO systems. Fig. 3 shows three example images for each route. Furthermore, the number of the sequence and the numbers of the start and end point images provided by the *KITTI* dataset [12] in terms of the six routes are detailed in Table III. Specifically, (1) Routes 1 and 2 are the straight lines on the express way and in the residential area respectively. Compared to Route 1 (≈780m, ≈60km/h), the average speed on Route 2 (≈357m, ≈33km/h) is much lower. Besides, there are few independent motion vehicles and pedestrians on Route 2; (2) Route 3 (≈330m, ≈ 28km/h) is a quarter turn on the urban road where other vehicles may be encountered; (3) Route 4 (≈ 960m, ≈26km/h) includes multiple quarter turns while Route 5 (≈1050m, ≈40km/h) contains multiple curved turns. Both the routes locate in the residential area. They have more complicated shapes and are longer than Routes 1, 2 and 3; and (4) Route 6 (≈930m, ≈38km/h) is a close loop and also locates in the residential area.

All the frames with regard to the six routes were acquired at 10 fps using a driving car. A gathering setup which consisted of several sensors [12] was equipped on the car. Those frames have the resolution of $1240 \times 375$ pixels. At the same time, the GPS and IMU data were recorded. The fusion of the two sets of data is used as the ground-truth data. Only synchronized gray level images are used except that color images are used for the hyper-column CNN (HC-CNN) descriptor [31].

### C. Performance Measures

The Root Mean Square Error (RMSE) and end-point error are popular performance metrics used for the VO research. We utilize the RMSE measure calculated from the orientation or position data as a performance metric. The RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]}{n}}, \quad (8)$$

where $(x_i, y_i)$ and $(\hat{x}_i, \hat{y}_i)$ are the ground-truth and estimated data respectively.

Instead of using the end position error of the whole trajectory, we adapt the Segment-Based End Point Error (SEPE) [60] measure. Given a start point, we take a segment of the ground-truth trajectory over the length of 100m. The images and GPS/IMU data of the KITTI dataset [12] are synchronized at the same rate (i.e., each image corresponds to a set of GPS/IMU data). The VIO system used a fixed camera height to derive absolute scale. The estimated trajectory has been transformed into the same reference system as the ground-truth. Thus, we can obtain the corresponding end points on the estimated trajectory. The position error over the traversed distance is derived by computing the *Euclidean* distance between the two end points. Along with the start position is shifted by 2m each time, this process is repeated until the end of the path is reached. The mean, median and standard deviation values of the errors are reported in this paper.

In addition, the modified *Hausdorff* distance [61] is used as a similarity measure between the estimated and ground-truth trajectories. The *Hausdorff* distance between two point sets: $A$ and $B$ is computed according to:

$$h(A,B) = \frac{1}{N_a}\sum_{a\in A}\{\min_{b\in B}\{\, d(a,b)\}\}, \quad (9)$$

where $a$ and $b$ are the points in $A$ and $B$ respectively, and $d(a,b)$ is the direct distance (which is usually computed using the *Euclidian* distance) between $a$ and $b$. As shown in Equation (9), the *Hausdorff* distance is not symmetric, i.e., $h(A,B) \neq h(B,A)$. Therefore, a more general form of the *Hausdorff* distance defined below is normally used.

$$H(A,B) = max\{h(A,B), h(B,A)\}. \quad (10)$$

## V. EVALUATION ON DIFFERENT COMBINATIONS OF SALIENT POINT DETECTORS AND LOCAL DESCRIPTORS

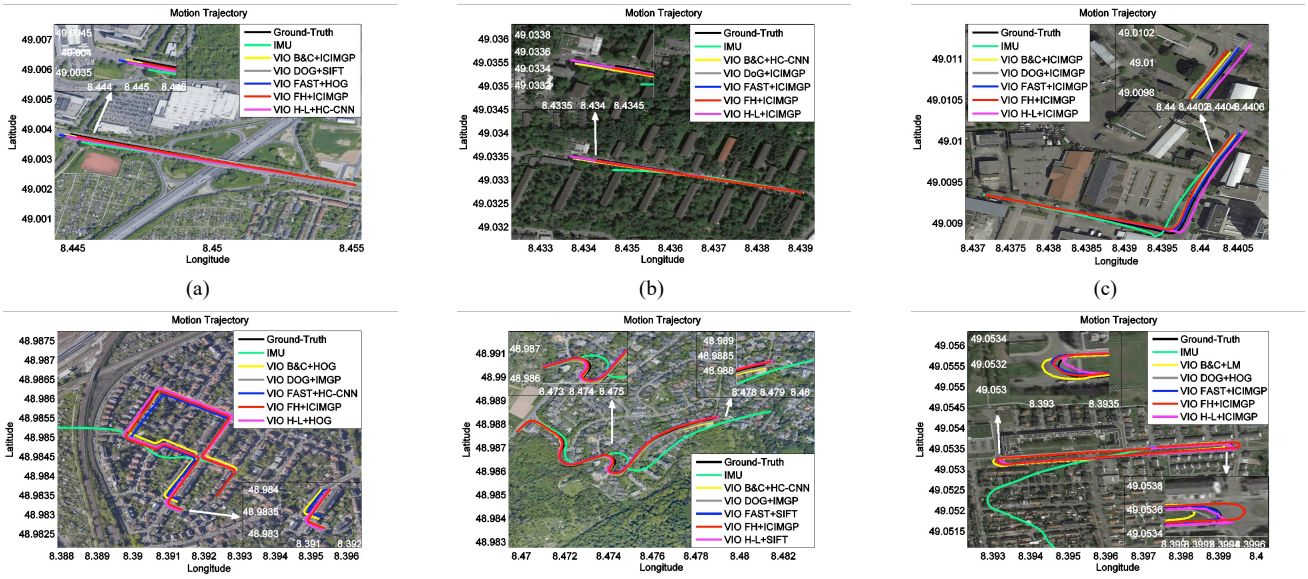



Fig. 4. The ground-truth trajectory and trajectories (best viewed in color) obtained using IMU and the best descriptor for each detector with six routes: (a) a straight line on the express way, (b) a straight line in the residential area, (c) a quarter turn on the urban road, (d) multiple quarter turns in the residential area, (e) multiple curved turns in the residential area, and (f) a loop in the residential area. (Map source: GoogleEarth). More figures can be found in the supplementary material.

In this section, we assess the five salient point detectors and nine local descriptors reviewed in Section III. The experimental setup introduced in Section IV is used. In particular, we investigate the ability of different combinations of detectors and descriptors for the task of monocular visual-IMU odometry on six typical routes containing different lengths, shapes and road scenes. The parameters used for these methods are shown in Tables I and II. We report the overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE computed between the ground-truth and estimated trajectories as follows (Only the performance of the best descriptor for each detector in terms of each route is reported. Please refer to the supplementary material for complete results).

### *A. Route 1: Straight Line on the Express Way*

Route 1 is a straight line captured on the express way. The average speed of the vehicle was high. The overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE computed between the estimated trajectories obtained using IMU and the best descriptor for each detector and the ground-truth trajectory are reported in Table IV (a). Figure 4(a) further displays the ground-truth trajectory and the estimated trajectories.

It can be seen from Table IV (a) that: (1) the combination of the FH detector [20] and the ICIMGP descriptor [39] yields the best RMSE performance; (2) ICIMGP [39] also performs properly when used with the other detectors, except the DoG detector [19]; (3) the HOG [26] and LSSD [29] descriptors perform properly when combined with FAST [15], while SIFT [19], SURF [20] and LM [28] generates promising results when used with DoG [19] and FAST [15]; (4) IMGP [27] and LIOP [30] do not provide good performance. Particularly, the performance of LIOP is worse than all of its counterparts; (5) the state-of-the-art HC-CNN descriptor [31] only performs well when combined with the H-L detector [18]; and (6) the IMU method performs properly on this route.

### *B. Route 2: Straight Line in the Residential Area*

Route 2 is also a straight line and was acquired in the residential area. Compared with Rout 1, the average speed of the vehicle used for this route was lower. Besides, there are rare independent motions of other vehicles and pedestrians. Table IV (b) lists the overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE computed between the estimated and ground-truth trajectories. Figure 4(b) further shows the ground-truth trajectory and the estimated trajectories obtained using IMU and the best descriptor for each detector.

As shown in Table IV (b), (1) the joint use of the FH detector [20] and ICIMGP [39] produces the best performance; (2) the performance of ICIMGP is comparable to the best result when combined with the DoG detector [19]; (3) the performance of LSSD [29] is inferior. This is particularly true when it is used with the FH [20] or H-L detectors [18]; (4) the HC-CNN [31], HOG [26], IMGP [27], LIOP [30] and SIFT [19] descriptors perform properly when combined with the DoG detector [19]. It is noteworthy that DoG [19] is the best choice for HC-CNN [31]; (5) in terms of LM [28] and SURF [20], the best choices of the detector are DoG [19] and FAST [15] respectively.

### *C. Route 3: Quarter Turn*

Route 3 is a simple quarter turn on the urban road. Table IV (c) reports the overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE derived using IMU and the best descriptor for each detector. As can be seen, (1) the ICIMGP descriptor [39] often outperforms its counterparts, especially, when combined with the FAST detector [15]; (2) the joint use of the FAST detector and the HOG descriptor [26] is comparable to this result; (3) both SIFT [19] and SURF [20]

TABLE IV
RESULTS COMPUTED BETWEEN THE GROUND-TRUTH TRAJECTORY AND THE TRAJECTORIES OBTAINED USING IMU AND THE BEST DESCRIPTOR FOR EACH SALIENT POINT DETECTOR ON ROUTES 1, 2 AND 3. MORE RESULTS CAN BE FOUND IN THE SUPPLEMENTARY MATERIAL.

| | IMU | B&C+ICIMGP | DOG+SIFT | FAST+HOG | FH+ICIMGP | H-L+HC-CNN |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 19.75 | 8.87 | 5.48 | 11.00 | 5.27 | 9.70 |
| Ori. RMSE(deg) | 2.42 | 1.97 | 2.37 | 2.22 | 1.74 | 1.97 |
| Hausd. Dist. (m) | 7.17 | 2.21 | 3.12 | 6.70 | 2.50 | 5.70 |
| Mean(m) | 18.59 | 8.27 | 4.95 | 9.80 | 4.63 | 9.55 |
| Median(m) | 16.84 | 8.41 | 5.20 | 7.96 | 5.45 | 8.49 |
| Std. Dev. (m) | 10.85 | 4.64 | 3.29 | 6.63 | 4.37 | 4.20 |

**(a) Route 1**

| | IMU | B&C+HC-CNN | DOG+ICIMGP | FAST+ICIMGP | FH+ICIMGP | H-L+ICIMGP |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 15.41 | 20.21 | 4.98 | 17.74 | 3.84 | 21.71 |
| Ori. RMSE(deg) | 2.05 | 3.47 | 1.32 | 2.19 | 1.05 | 2.85 |
| Hausd. Dist. (m) | 5.40 | 4.14 | 1.29 | 3.23 | 1.55 | 3.60 |
| Mean(m) | 15.74 | 20.39 | 4.09 | 15.36 | 3.34 | 21.94 |
| Median(m) | 13.97 | 19.93 | 4.48 | 10.05 | 2.99 | 18.44 |
| Std. Dev. (m) | 8.65 | 12.06 | 1.48 | 13.85 | 0.92 | 12.81 |

**(b) Route 2**

| | IMU | B&C+ICIMGP | DOG+ICIMGP | FAST+ICIMGP | FH+ICIMGP | H-L+ICIMGP |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 19.29 | 4.24 | 5.81 | 3.57 | 4.95 | 8.49 |
| Ori. RMSE(deg) | 3.92 | 1.46 | 1.52 | 1.40 | 1.48 | 1.54 |
| Hausd. Dist. (m) | 8.25 | 3.20 | 3.04 | 2.10 | 3.81 | 3.65 |
| Mean(m) | 19.51 | 4.48 | 4.45 | 3.20 | 5.17 | 7.45 |
| Median(m) | 20.70 | 4.46 | 3.65 | 3.11 | 5.12 | 6.27 |
| Std. Dev. (m) | 8.06 | 0.44 | 2.30 | 0.90 | 0.53 | 3.33 |

**(c) Route 3**

TABLE V
RESULTS COMPUTED BETWEEN THE GROUND-TRUTH TRAJECTORY AND THE TRAJECTORIES OBTAINED USING IMU AND THE BEST DESCRIPTOR FOR EACH SALIENT POINT DETECTOR ON ROUTES 4, 5 AND 6. MORE RESULTS CAN BE FOUND IN THE SUPPLEMENTARY MATERIAL.

| | IMU | B&C+HOG | DOG+IMGP | FAST+HC-CNN | FH+ICIMGP | H-L+HOG |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 1540 | 9.02 | 4.88 | 12.18 | 4.43 | 4.99 |
| Ori. RMSE(deg) | 11.23 | 2.66 | 2.75 | 2.54 | 1.37 | 1.42 |
| Hausd. Dist. (m) | 1014 | 6.84 | 2.43 | 6.76 | 1.98 | 2.79 |
| Mean(m) | 1280.66 | 12.45 | 4.52 | 11.68 | 2.31 | 4.73 |
| Median(m) | 946.41 | 12.09 | 4.51 | 10.56 | 3.10 | 4.53 |
| Std. Dev. (m) | 1095 | 4.77 | 2.47 | 6.25 | 2.09 | 2.25 |

**(a) Route 4**

| | IMU | B&C+HC-CNN | DOG+IMGP | FAST+SIFT | FH+ICIMGP | H-L+SIFT |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 86.63 | 9.08 | 8.14 | 6.95 | 6.53 | 10.83 |
| Ori. RMSE(deg) | 3.67 | 2.46 | 2.59 | 1.54 | 2.58 | 2.78 |
| Hausd. Dist. (m) | 35.08 | 5.07 | 3.79 | 4.65 | 3.69 | 5.46 |
| Mean(m) | 72.18 | 9.23 | 7.06 | 6.66 | 3.87 | 9.64 |
| Median(m) | 56.25 | 9.70 | 6.93 | 6.82 | 5.20 | 8.56 |
| Std. Dev. (m) | 56.72 | 2.44 | 4.37 | 2.54 | 2.93 | 5.39 |

**(b) Route 5**

| | IMU | B&C+LM | DOG+HOG | FAST+ICIMGP | FH+ICIMGP | H-L+ICIMGP |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 314.48 | 12.29 | 6.40 | 4.56 | 9.14 | 6.90 |
| Ori. RMSE(deg) | 9.76 | 3.19 | 2.46 | 2.39 | 2.68 | 2.64 |
| Hausd. Dist. (m) | 208.90 | 2.52 | 1.67 | 1.95 | 2.66 | 1.64 |
| Mean(m) | 282.00 | 11.92 | 5.58 | 3.46 | 8.68 | 6.77 |
| Median(m) | 262.95 | 12.96 | 3.98 | 4.09 | 7.62 | 6.21 |
| Std. Dev. (m) | 199.37 | 5.71 | 4.40 | 1.72 | 4.40 | 3.42 |

**(c) Route 6**

yield proper performance; (4) the performance of LIOP [30] is superior to that it obtains on Route 1 but is still worse than that of its counterparts in most cases; (5) when combined with the B&C detector [23], LM [28] and LSSD [29] perform properly on this route; (6) the proper performance is produced by IMGP [27] along with the FAST [15] or DoG [19] detectors; (7) for both the RMSE and SEPE metrics, the DoG detector is the best choice for HC-CNN [31]; and (8) IMU performs properly. In addition, Figure 4(c) shows the ground-truth trajectory and the trajectories obtained using IMU and the best descriptor for each salient point detector.

### *D. Route 4: Multiple Quarter Turns*

Route 4 was captured in the residential area and contains multiple quarter turns. Compared with Routes 1, 2 and 3, this route is longer and has more complicated shape. Table V (a) lists the overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE obtained using different methods. It can be observed that: (1) the joint use of the FH detector [20] and ICIMGP [39] yields the best RMSE and SEPE performance; (2) HOG [26] generates the proper result, especially, when combined with the H-L detector [18]; (3) the performance of IMGP [27] is even comparable to the best result when combined with DoG [19] and performs properly together with the other detectors; (4) LM [28] performs properly and yields its best performance when combined with the H-L detector [18]; (5) both SIFT [19] and SURF [20] perform properly in most cases; (6) LIOP [30] produces better results than that it performs on Routes 1, 2 and 3, and yields its best performance when used with H-L [18]; (7) LSSD [29] provides proper performance when combined with B&C [23], DoG [19] or FAST [15]; (8) HC-CNN [31] generates proper performance along with the H-L, DoG or FAST detectors; and (9) the performance of the IMU method is worse than those of all the detector-descriptor combinations. Furthermore, the ground-truth trajectory and the trajectories obtained using IMU and the best descriptor for each salient point detector are presented in Fig. 4(d).

### *E. Route 5: Multiple Curved Turns*

Multiple continuous curved turns are included in Route 5 which was acquired in the residential area. The overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE computed between the estimated trajectories obtained using IMU and the best descriptor for each detector and the ground-truth trajectory are reported in Table V (b). It can be seen that: (1) the combination of FH [20] and ICIMGP [39] generates promising RMSE and SEPE performance; (2) SIFT [19] yields the comparable performance to these results when used with FAST [15] and performs better than that it does on Routes 1, 2 and 3; (3) HOG [26], SURF [20] and LM [28] perform properly while LSSD [29] only produces proper performance when used with B&C [23], DoG [19] or FAST [15]; (4) IMGP [27] yields its best performance when combined with DoG [19] and also performs properly when used with the other detectors; (5) the results obtained using LIOP [30] severely suffer from the drift issue except when used with H-L [18] and are even worse than that obtained using IMU; (6) the overall performance of HC-CNN [31] is better than that it yields on the previous four routes, and it performs properly when combined with DoG [19]. Figure 4(e) further shows the ground-truth trajectory and the trajectories derived using the IMU method and the best descriptor for each detector.

### *F. Route 6: Loop Line*

Route 6 is a closed loop captured in the residential area. In Table V (c), we report the overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE computed between the trajectories obtained using different methods and the ground-truth trajectory. As can be seen, (1) the ICIMGP descriptor [39] generates the best RMSE performance when

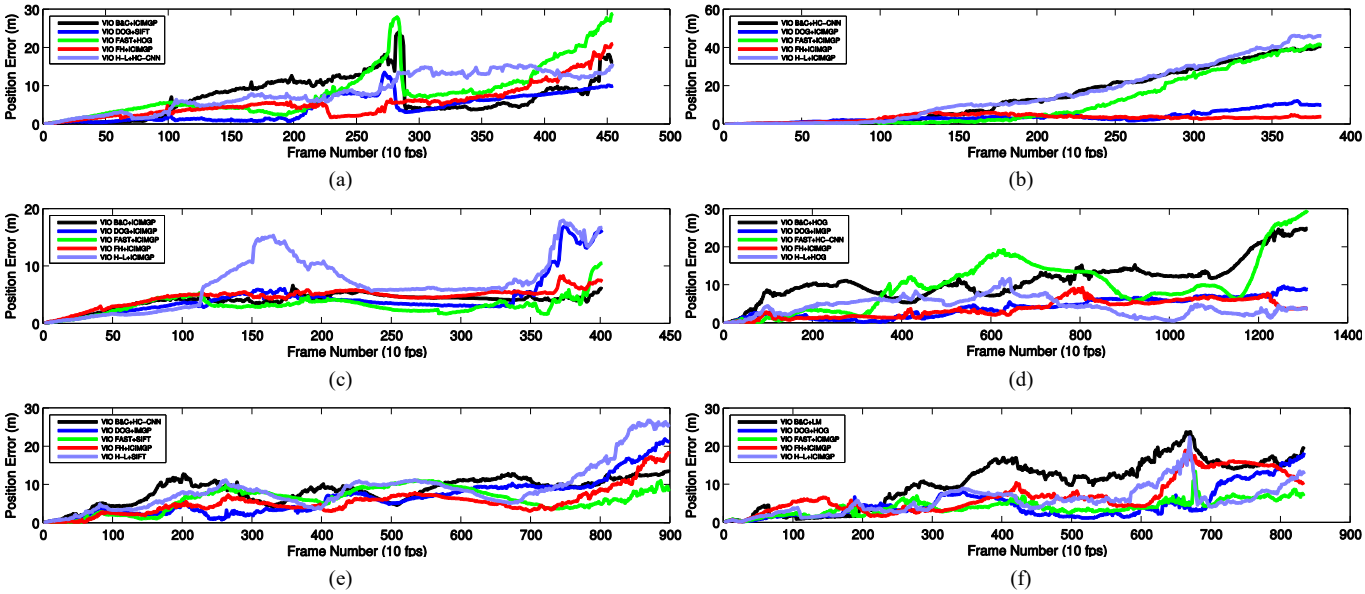



Fig. 5. Position errors (best viewed in color) obtained using the best descriptor for each detector in terms of each video frame along with the six routes: (a) a straight line on the express way, (b) a straight line in the residential area, (c) a quarter turn on the urban road, (d) multiple quarter turns in the residential area, (e) multiple curved turns in the residential area, and (f) a loop in the residential area.

combined with the FAST detector [15]. It also performs well when used with the other detectors; (2) HOG [26] yields promising results except when it is used with B&C [23]. Regarding the median value of the SEPE metric, "DOG+HOG" and "FAST+ICIMGP" shows similar performance. However, the worst case error (Mean + 3 Standard Deviation) [60] values produced by both combinations are 18.78m and 8.62m respectively; (3) the proper results are derived using HC-CNN [31]. In particular, the best detector for it is DoG [19]; (4) LM [28], IMGP [27], SIFT [19] and SURF [20] generate proper performance while LSSD [29] only yields proper performance when used with DoG [19] or FAST [15]; (5) LIOP [30] properly performs when combined with the B&C [23], DoG [19] or H-L [18] detectors; and (6) the performance of IMU is the worst while it can be improved by being jointly used with the detector-descriptor methods. Figure 4(f) also shows the ground-truth trajectory and the trajectories obtained using IMU and the best descriptor for each salient point detector.

### G. *Position Errors Per Video Frame*

Given each frame is considered, we compute the position errors generated by the best descriptor for each detector along with the six routes and present these results in Fig. 5.

According to Fig. 5(a), the position error of the "B&C+ICIMGP" combination greatly increases and arrives at a peak with 24.03m from Frame 93 to Frame 283. The similar phenomenon can be observed for the case of "FAST+HOG" where a peak with 28.06m is reached to at Frame 283. Since Route 1 was gathered on the express way with high speed, there were a lot of cars moving in the camera view. In this situation, the moving vehicles (please refer to Fig. 3 (a) for examples) in front of the recording platform made this route challenging. In contrast, the combinations of "DoG+SIFT" and "H-L+ICIMGP" perform better in that interval. However, the combination of FH and ICIMGP properly perform between those frames.

Route 2 was acquired in the residential area with cyclists and cars moving in front of the camera in the second half of the route (see Fig. 3(b) for examples). As a result, the position error produced by "B&C+HC-CNN", "FAST+ICIMGP" and "H-L+ICIMGP" apparently rises (see Fig. 5(b)).

As shown in Fig. 5(c), the performance of the combinations of "B&C+ICIMGP", "FAST+ICIMGP" and "FH+ICIMGP" is satisfactory. From Frames 110 to 165, there were cars in front of the camera, pedestrians on the right-hand sidewalk and a left turn (see Fig. 3(c) for examples). These complex road scenes should account for the high position errors that "H-L+ICIMGP" suffers in the interval. In addition, there were many cars drove across the intersection after Frame 352. This situation results in the increased position error with "DoG+ICIMGP" and "H-L+ICIMGP".

Route 4 contains multiple quarter turns and was captured in the residential area with rectilinear motions, left and right quarter turns, and moving cyclists and pedestrians (see Fig. 3(d)). The most significant challenge with this route should be due to the turns and independent moving objects. Over the full sequence, "DoG+IMGP", "FH+ICIMGP" and "H-L+HOG" produce low position errors (which are normally below 4m). In the turn frames, position errors that "B&C+HOG" and "FAST+HC-CNN" yield encounter large, continuous fluctuations.

Multiple curved turns in a residential area are included in Route 5. There are three continuous semicircle turns between Frames 150 and 600 (see Fig. 3(e) for examples). As displayed in Fig. 5(e), the position error starts to increase and meets three peaks during this period. Also, the long distance curved turn results in the increased position error after Frame 700.

The sixth route is a loop recorded in the residential area with rectilinear motions and semicircle turns. However, fewer moving objects are encountered in this route (see Fig. 3(f) for examples). The position error that the straight line motion generates is relatively low except for "B&C+LM"; while the

semicircle turns yield two apparent position error peaks.

### *H. Useful Insights*

The results show that the performance of the nine descriptors varies when they are used with different detectors or routes. To summarize, seven insights can be derived as follows.

(1) The best result is normally produced by ICIMGP [39], especially, when it is combined with the FAST [15] or FH [20] detectors. It is suggested that ICIMGP [39] is suitable for monocular visual-IMU odometry. The promising results should be attributed to the fact that ICIMGP [39] encodes richer image characteristics than its counterparts that are usually extracted from gray level images. When combined with the FAST [15] and H-L [18] detectors, ICIMGP [39] generates promising results on Routes 3 and 6. This observation should be attributed to the fact that the majority of the images captured on Routes 3 and 6 are asphalt pavement or concrete building surfaces. These surfaces contain corner-like structures and therefore can be easily extracted using the corner detector. In [65], it has been demonstrated that the performance of the FAST [15] detector is far better than that of H-L on the asphalt and concrete texture surroundings. This finding should be account for the fact that "FAST+ICIMGP" usually outperforms "H-L+ICIMGP".

(2) The HOG [26] and LSSD [29] descriptors perform properly when they are used with the DoG [19] or FAST detectors [15]. However, their performance varies when used with other detectors.

(3) Regarding the IMGP descriptor [27], the most suitable detector could be DoG [19] or FAST [15]. When it is combined with the DoG detector [19], it even outperforms ICIMGP [39] on Routes 4 and 5. However, IMGP does not yield promising results on the straight line express way route (Route 1). The similar finding can be observed for LIOP [30] when it is used with the H-L detector [18]. These results suggest that gray level image patches are not sufficient for the high speed motion road scene or the straight line route and probably need to be used with other image feature channels (please refer to ICIMGP).

(4) The LM [28], SIFT [19] and SURF [20] descriptors yield promising results when they are jointly used with DoG [19]. But their performance is not stable when used with the other detectors. Surprisingly, SURF [20] normally performs better when combined with DoG [19] than that it does along with FH [20].

(5) The state-of-the-art HC-CNN [31] descriptor is inferior or comparable to ICIMGP [39]. We attribute this result to the following two reasons. First, we utilized the Places205-VGG model [59] to extract HC-CNN features. Places205-VGG was trained using the VGG-VD-16 CNN [58] on the 205 scene categories of the Places dataset [59] rather than the KITTI dataset [12]. However, the difference between the two datasets is apparent. The Places dataset was collected from the Internet, and the images contained were taken at different time and places independently. Nevertheless, the KITTI dataset was captured during car driving whilst recording various driving situations, e.g., driving straight on different speeds and making the left or right turn. Although the Places dataset contains more images than the KITTI dataset, it does not contain these driving situations. Since the KITTI dataset does not provide the annotated data for road scene classification, we cannot fine-tune the Places205-VGG model using the KITTI dataset. Hence, it is likely that the Places205-VGG model [59] cannot represent road scenes well.

Second, the Places205-VGG model was trained for the task of scene classification. Nonetheless, the size of feature maps becomes smaller and smaller with the layers become deeper because of the downsampling operations. To extract HC-CNN features at each pixel location, the feature maps are up-sampled into the original image resolution using interpolation. This process loses much localization information and may lead to feature matching errors and inaccurate ego-motion estimation results.

(6) It is difficult to decide the best salient point detector as the performance of the detectors varies with different routes. Also, the performance of a detector depends on which descriptor is combined with it. Generally, the DoG [19] detector provides relatively stable and good results. For a practical VO system, road scene classification can be used as a pre-processing before salient points are detected. A detector is selected and applied according to the class of the image frame. This selection process makes the detector adaptive to different road scenes.

(7) According to the average position RMSE, the ranking in the ascending order of the difficulty is: (Routes) 4, 3, 6, 5, 2, and 1. To be exact, the average position RMSE values: 15.34m and 16.08m are derived on Routes 4 and 3 respectively. These values are much lower than the values: 46.65m and 65.75m generated on Routes 2 and 1. It is shown that straight line routes are challenging for the detector-descriptor methods tested here. In contrast, quarter turns are easier for these methods.

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

### *I. Discussion*

In this subsection, we make discussion on the obtained results in terms of four different aspects.

#### *1) Multi-channel vs. Single-Channel Descriptors*

Multi-channel images usually contain the complementary information among different channels. As a result, they encode richer characteristics and are able to provide the more complete description of the related content than any of the single channels (e.g., the gray level image). We tested a multi-channel image patch descriptor, i.e., ICIMGP [39], which was extracted from multiple channels of an image, including the gray level channel and seven other channels (please refer to Section III-B-4 for more details). In contrast, the single channel image patch descriptor: IMGP [27] was directly extracted from gray level images. The superior performance of ICIMGP to IMGP and the other descriptors tested in this study has been observed in our experiments. The average position errors (m) generated by ICIMGP on the six routes are 20.01, 16.54, 5.41, 9.51, 11.98 and 9.51 respectively; while the corresponding six values produced by IMGP are 62.08, 45.02, 19.08, 14.62, 18.95 and 19.77 respectively. We attribute this result to the fact that

TABLE VI
THE DIMENSIONALITY OF THE NINE LOCAL DESCRIPTORS.

| **Descriptor** | HOG | HC-CNN | ICIMGP | IMGP | LIOP |
|---|---|---|---|---|---|
| **Dim** | 279 | 1475 | 968 | 121 | 144 |
| **Descriptor** | LM | LSSD | SIFT | SURF | |
| **Dim** | 48 | 36 | 128 | 64 | |

ICIMGP incorporates additional feature channels, such as gradient magnitude and gradient histogram.

*2) Blob vs. Corner Detectors*

We examined two blob detectors: DoG [19] and FH [20] and two corner detectors: FAST [15] and H-L [18]. In the two challenging straight line and the loop line routes, most of the descriptors generated good results when combined with the DoG [19] detector. Regarding the FH [20] detector, however, its performance showed a large variance across the six routes when combined with different descriptors. On the other hand, the FAST detector [15] produced promising results on Routes 1, 3, 5, and 6, except when combined with the LIOP [30] descriptor. Although the H-L detector produced high position RMSE values on Routes 1 and 2, they yielded good results on Route 3. Therefore, it is not practical to determine which of the blob and corner detectors is better than the other.

*3) Impact of the Road Scene and Route Shape*

All the six routes were gathered at different times of the day and a variety of locations. As a result, different road scenes, including lighting conditions, shadow presence and numbers of vehicles, pedestrians and cyclists, and different route shapes are contained in these routes. In terms of the road scene, Routes 1 and 2 have the same shape (i.e., straight line) but quite different road scenes. Route 1 was captured on the express way where many moving vehicles and cyclists occurred in the view of the camera. In contrast, Route 2 was acquired in the residential area, which is a low dynamic surrounding. Therefore, Route 1 is more challenging than Route 2 as the independent motion of vehicles, pedestrians and cyclists results in the mismatching of features. The difference in road scenes should account for the difference in the average position RMSE (65.75m vs. 46.65m), the median (53.02m vs. 40.47m) and standard derivation (49.05m vs. 28.39m) of the SEPE generated on Routes 1 and 2.

On the other hand, the multiple quarter turns (Route 4), multiple curved turns (Route 5) and close loop (Route 6) routes were gathered in the residential area. Although the road scenes of these routes are similar, the shapes of these are obviously different. However, the average position RMSE values: 15.34m, 30.91m and 21.18m and the average SEPE median values: 14.93m, 22.98m and 18.90m are produced on Routes 4, 5 and 6. These results suggest that the impact of the route shape on the VO performance is also significant.

*4) Comparison of the Feature Dimensionality*

We do not compare the computational speed of different detectors and descriptors because they are 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 VI lists the dimensionality of the nine local descriptors. It can be seen that the dimensionality of the ICIMGP [39] descriptor is high while it indeed produces the best results in this study.

## VI. EFFECT OF PARAMETER VALUES ON THE PERFORMANCE OF THE BEST DETECTOR OR DESCRIPTOR FOR EACH ROUTE

In Section V, we tested different combinations of detectors and descriptors using fixed parameter values. In this section, we intend to investigate the effect of different parameter values on the performance of these combinations. However, it is not practical to tune the parameters of all detectors and descriptors. Alternatively, we tune the parameters of the best detector or descriptor obtained with each route in Section V in order to further augment the obtained results. According to the position RMSE and the SEPE (supplemented by the worst case error), the best detector-descriptor combination performed on Routes 1, 2, 4 and 5 is the FH detector [20] and the ICIMGP descriptor [39]; while the combination of FAST [15] and ICIMGP [39] performs best on Routes 3 and 6. We will examine the effect of the parameters of these methods on the visual-IMU odometry performance in this section.

### *A. Tuning Parameters of the Best Detector for Each Route*

We here tune the threshold $T$ for the FH [20] and FAST [15] detectors. For the FH detector, $T$ is set to 0.2 (which has been used in Section V), 0.4, 0.6, 0.8 and 1.0. Meanwhile, the values of $T$ are assigned with 20, 25, 30 (which has been used in Section V), 35 and 40 for the FAST detector. For simplicity, the size of image patches that ICIMGP [39] uses is kept as $11 \times 11$ pixels.

Table VII reports the overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE calculated between the ground-truth trajectory and the estimated trajectories on the six routes. As can be seen, the performance of the FH detector [20] varies when the value of $T$ is changed. It can be seen that on Routes 1, 4 and 5, the FH detector produces the better results when the value of $T$ is set to 0.2 than it performs when $T$ is assigned with the other values. In contrast, the FH detector yields similar results when $T$ is set to 0.2 and 0.6 on Route 2. However, these results are better than those obtained using FH with the other $T$ values.

On the other hand, the results derived using the FAST detector [15] when $T$ is set to 25 are superior to those that it produces when $T$ is assigned with the other values on Routes 3 and 6. Nevertheless, the influence of the $T$ value to the FH [20] and FAST [15] detectors is not significant when Routes 2 and 3 are considered respectively. In addition, the best results obtained using FH [20] or FAST [15] and ICIMGP [39] on different routes are relatively stable. Compared with the results provided by the IMU method, FH [20] or FAST [15] and ICIMGP [39] always yield the better performance.

### *B. Tuning Parameters of the Best Descriptor for Each Route*

Since the ICIMGP descriptor [39] normally generates the best results on the six routes, we alter the neighborhood size $N$ of the image patches used by this descriptor. Specifically, the values of $N$ are set to 7, 9, 11, 13 and 15. Regarding the detector, FH [20] ($T = 0.2$) is used for Routes 1, 2, 4 and 5 while FAST [15] ($T = 30$) is used for Routes 3 and 6.

TABLE VII
RESULTS COMPUTED BETWEEN THE GROUND-TRUTH TRAJECTORY AND THE TRAJECTORIES OBTAINED USING THE FH [20] OR FAST [15] DETECTOR WITH VARIED $T$ VALUES AND THE ICIMGP DESCRIPTOR [39] ON DIFFERENT ROUTES

| | FH-0.2 | FH-0.4 | FH-0.6 | FH-0.8 | FH-1.0 | IMU |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | **5.27** | 12.17 | 16.78 | 9.48 | 8.52 | 19.75 |
| Ori. RMSE(deg) | **1.74** | 1.80 | 1.91 | 1.82 | 1.79 | 2.42 |
| Hausd. Dist. (m) | **2.50** | 3.47 | 3.29 | 3.50 | 3.63 | 7.17 |
| Mean(m) | **4.63** | 7.86 | 8.96 | 6.80 | 6.01 | 18.59 |
| Median(m) | **5.45** | 6.48 | 10.44 | 6.85 | 6.47 | 16.84 |
| Std. Dev. (m) | **4.37** | 7.53 | 10.40 | 5.32 | 4.57 | 10.85 |
| **(a) Route 1** | | | | | | |
| | FH-0.2 | FH-0.4 | FH-0.6 | FH-0.8 | FH-1.0 | IMU |
| Pos. RMSE(m) | 3.84 | 4.03 | **3.62** | 5.92 | 5.90 | 15.41 |
| Ori. RMSE(deg) | **1.05** | 1.35 | 1.21 | 1.41 | 1.40 | 3.05 |
| Hausd. Dist. (m) | 1.55 | 1.69 | **1.49** | 1.51 | 1.72 | 5.40 |
| Mean(m) | 3.34 | 3.41 | **3.09** | 4.80 | 4.81 | 15.74 |
| Median(m) | 2.99 | 3.67 | **2.91** | 5.98 | 3.88 | 13.97 |
| Std. Dev. (m) | **0.92** | 1.53 | 0.98 | 1.14 | 0.84 | 8.65 |
| **(b) Route 2** | | | | | | |
| | FAST-20 | FAST-25 | FAST-30 | FAST-35 | FAST-40 | IMU |
| Pos. RMSE(m) | 3.07 | **3.01** | 3.57 | 4.11 | 5.91 | 19.29 |
| Ori. RMSE(deg) | 1.35 | **1.30** | 1.40 | 1.51 | 1.66 | 3.92 |
| Hausd. Dist. (m) | 1.93 | **1.62** | 2.10 | 2.27 | 3.76 | 8.25 |
| Mean(m) | 2.90 | **2.59** | 3.20 | 3.64 | 5.87 | 19.51 |
| Median(m) | 2.75 | **2.41** | 3.11 | 3.16 | 6.30 | 20.70 |
| Std. Dev. (m) | **0.87** | 1.21 | 0.90 | 1.51 | 1.71 | 8.06 |
| **(c) Route 3** | | | | | | |
| | FH-0.2 | FH-0.4 | FH-0.6 | FH-0.8 | FH-1.0 | IMU |
| Pos. RMSE(m) | **4.43** | 5.04 | 6.98 | 4.61 | 11.93 | 1540 |
| Ori. RMSE(deg) | 1.37 | 1.42 | 1.45 | **1.37** | 2.31 | 11.23 |
| Hausd. Dist. (m) | **1.98** | 2.43 | 3.75 | 2.42 | 5.48 | 1014 |
| Mean(m) | **2.31** | 3.00 | 4.68 | 3.52 | 6.25 | 1280.66 |
| Median(m) | **3.10** | 4.33 | 4.16 | 4.18 | 5.01 | 946.41 |
| Std. Dev. (m) | 2.09 | 2.05 | 3.40 | **1.75** | 6.41 | 1095 |
| **(d) Route 4** | | | | | | |
| | FH-0.2 | FH-0.4 | FH-0.6 | FH-0.8 | FH-1.0 | IMU |
| Pos. RMSE(m) | **6.53** | 9.45 | 9.25 | 10.65 | 11.00 | 83.63 |
| Ori. RMSE(deg) | **2.58** | 2.77 | 2.71 | 2.96 | 3.02 | 3.67 |
| Hausd. Dist. (m) | **3.69** | 5.9123 | 5.93 | 6.58 | 6.63 | 35.08 |
| Mean(m) | **3.87** | 6.51 | 6.38 | 7.89 | 8.30 | 72.18 |
| Median(m) | **5.20** | 8.06 | 9.60 | 8.98 | 7.36 | 56.25 |
| Std. Dev. (m) | 2.93 | 2.30 | **2.01** | 2.41 | 4.33 | 56.72 |
| **(e) Route 5** | | | | | | |
| | FAST-20 | FAST-25 | FAST-30 | FAST-35 | FAST-40 | IMU |
| Pos. RMSE(m) | 8.99 | **4.20** | 4.56 | 10.22 | 5.58 | 314.48 |
| Ori. RMSE(deg) | 3.42 | **2.21** | 2.39 | 3.69 | 2.46 | 9.76 |
| Hausd. Dist. (m) | 3.84 | 1.97 | **1.95** | 2.34 | 2.25 | 208.90 |
| Mean(m) | 6.82 | **2.67** | 3.46 | 7.39 | 3.44 | 282.00 |
| Median(m) | 5.45 | **1.55** | 4.09 | 6.91 | 4.64 | 262.95 |
| Std. Dev. (m) | 2.00 | **1.64** | 1.72 | 2.84 | 2.57 | 199.37 |
| **(f) Route 6** | | | | | | |

FH-T or FAST-T denotes Fast Hessian or FAST with the threshold of $T$.

The overall position and orientation RMSE values, the *Hausdorff* distance, and the mean, median and standard deviation values of the SEPE computed between the ground-truth trajectory and the estimated trajectories on the six routes are shown in Table VIII. The most obvious observation is that ICIMGP [39] normally produces the best result when the value of $N$ is set to 11. It is worth to note that $N$ = 15 generates comparable RMSE and mean SEPE values to $N$ = 11, while the median and standard derivation values produced by the former are inferior to those generated by the latter. That is to say, the performance of ICIMGP does not enhance when the size of image patches is larger than $11 \times 11$ pixels. This finding is consistent with that Dong *et al*. [39] and Gauglitz *et al*. [47] observed. The performance of the ICIMGP descriptor using $11 \times 11$ patches is relatively steady no matter what route is used. In contrast, the performance of IMU is normally inferior.

TABLE VIII
RESULTS COMPUTED BETWEEN THE GROUND-TRUTH TRAJECTORY AND THE TRAJECTORIES OBTAINED USING THE FH [20] ($T$=0.2) OR FAST [15] ($T$=30) DETECTOR AND THE ICIMGP DESCRIPTOR [39] WITH VARIED NEIGHBORHOOD SIZES ($N$) ON DIFFERENT ROUTES

| | $N$ = 7 | $N$ = 9 | $N$ = 11 | $N$ = 13 | $N$ = 15 | IMU |
|---|---|---|---|---|---|---|
| Pos. RMSE(m) | 12.44 | 49.00 | **5.27** | 10.18 | 13.25 | 19.75 |
| Ori. RMSE(deg) | 2.22 | 4.26 | **1.74** | 2.03 | 2.36 | 2.42 |
| Hausd. Dist. (m) | 3.37 | 10.00 | **2.50** | 3.056 | 3.39 | 7.17 |
| Mean(m) | 7.83 | 22.90 | **4.63** | 7.27 | 9.40 | 18.59 |
| Median(m) | 9.53 | 7.18 | **5.45** | 6.58 | 10.60 | 16.84 |
| Std. Dev. (m) | 6.34 | 41.98 | **4.37** | 7.24 | 10.90 | 10.85 |
| **(a) Route 1** | | | | | | |
| | $N$ = 7 | $N$ = 9 | $N$ = 11 | $N$ = 13 | $N$ = 15 | IMU |
| Pos. RMSE(m) | 5.41 | 4.04 | 3.84 | 4.31 | **3.76** | 15.41 |
| Ori. RMSE(deg) | 1.43 | 1.33 | **1.05** | 1.40 | 1.28 | 3.05 |
| Hausd. Dist. (m) | **1.47** | 1.69 | 1.55 | 1.75 | 1.64 | 5.40 |
| Mean(m) | 4.09 | 3.64 | 3.34 | 3.60 | **3.18** | 15.74 |
| Median(m) | 5.61 | 3.00 | **2.99** | 3.37 | 4.01 | 13.97 |
| Std. Dev. (m) | 1.77 | 2.10 | **0.92** | 1.98 | 1.16 | 8.65 |
| **(b) Route 2** | | | | | | |
| | $N$ = 7 | $N$ = 9 | $N$ = 11 | $N$ = 13 | $N$ = 15 | IMU |
| Pos. RMSE(m) | 5.57 | 4.44 | **3.57** | 5.69 | 3.73 | 19.29 |
| Ori. RMSE(deg) | 1.79 | 1.62 | **1.40** | 1.70 | 1.51 | 3.92 |
| Hausd. Dist. (m) | 3.37 | **2.09** | 2.10 | 2.68 | 2.67 | 8.25 |
| Mean(m) | 3.76 | 3.44 | **3.20** | 4.44 | 3.65 | 19.51 |
| Median(m) | 4.13 | **2.95** | 3.11 | 3.39 | 3.22 | 20.70 |
| Std. Dev. (m) | 1.24 | 1.86 | **0.90** | 2.89 | 0.92 | 8.06 |
| **(c) Route 3** | | | | | | |
| | $N$ = 7 | $N$ = 9 | $N$ = 11 | $N$ = 13 | $N$ = 15 | IMU |
| Pos. RMSE(m) | 22.76 | 13.50 | **4.43** | 11.09 | 30.36 | 1540 |
| Ori. RMSE(deg) | 3.25 | 2.33 | **1.37** | 3.01 | 4.25 | 11.23 |
| Hausd. Dist. (m) | 11.16 | 6.86 | **1.98** | 5.66 | 12.65 | 1014 |
| Mean(m) | 13.20 | 7.27 | **2.31** | 8.85 | 19.60 | 1280.66 |
| Median(m) | 15.36 | 10.17 | **3.10** | 7.04 | 21.09 | 946.41 |
| Std. Dev. (m) | 8.27 | 5.71 | **2.09** | 5.03 | 15.26 | 1095 |
| **(d) Route 4** | | | | | | |
| | $N$ = 7 | $N$ = 9 | $N$ = 11 | $N$ = 13 | $N$ = 15 | IMU |
| Pos. RMSE(m) | 12.23 | 8.04 | **6.53** | 10.84 | 9.99 | 83.63 |
| Ori. RMSE(deg) | 2.95 | 2.75 | **2.58** | 2.90 | 2.80 | 3.67 |
| Hausd. Dist. (m) | 6.07 | 4.81 | **3.69** | 5.51 | 5.82 | 35.08 |
| Mean(m) | 9.71 | 6.57 | **3.87** | 7.31 | 6.46 | 72.18 |
| Median(m) | 10.72 | 7.17 | **5.20** | 6.85 | 7.19 | 56.25 |
| Std. Dev. (m) | 6.15 | 3.20 | **2.93** | 4.24 | 9.79 | 56.72 |
| **(e) Route 5** | | | | | | |
| | $N$ = 7 | $N$ = 9 | $N$ = 11 | $N$ = 13 | $N$ = 15 | IMU |
| Pos. RMSE(m) | 10.22 | 6.68 | **4.56** | 9.98 | 4.66 | 314.48 |
| Ori. RMSE(deg) | 3.23 | 2.79 | **2.39** | 3.11 | 2.42 | 9.76 |
| Hausd. Dist. (m) | 2.15 | **1.88** | 1.95 | 2.60 | 2.02 | 208.90 |
| Mean(m) | 8.00 | 4.07 | **3.46** | 7.71 | 3.97 | 282.00 |
| Median(m) | 8.02 | 4.58 | **4.09** | 5.96 | 4.14 | 262.95 |
| Std. Dev. (m) | 3.81 | 3.83 | **1.72** | 4.36 | 2.03 | 199.37 |
| **(f) Route 6** | | | | | | |

To summarize, the combinations of FH [20] ($T$ = 0.2) or FAST [15] ($T$ = 30) and ICIMGP [39] are promising for the monocular visual-IMU odometry task.

## VII. CONCLUSIONS AND FUTURE WORK

In this paper, we first reviewed five salient point detectors and nine local descriptors. Then, we deliberately selected six typical routes from the known *KITTI* dataset [12] by taking into account lengths, shapes and road scenes, to rigorously assess the detectors and descriptors. Using the adapted version [39] of an established monocular visual-IMU odometry system [37] and these routes, we performed an extensive evaluation study on different combinations of the detectors and descriptors. To our knowledge, this is the first extensive comparative evaluation on salient point detectors and local descriptors for monocular visual-IMU odometry using the representative routes with different lengths, shapes and road scenes.

We examined the detectors and descriptors with the fixed

parameters in order to find out the most promising detector-descriptor combination. The results of the experiments provide the community with a set of benchmark data for future research. More importantly, the analysis of the results can be used as guidelines for developing new algorithms or implementing practical odometry systems. To further augment the experimental results, we also tuned the parameters of the best detector or descriptor for each route. To be specific, the FH [20] and FAST [15] detectors and the ICIMGP [39] descriptor were examined in this experiment. It was found that the joint use of FH [20] ($T$ = 0.2) or FAST [15] ($T$ = 30) and ICIMGP [39] ($N$ = 11) produced relatively stable and promising results across the six routes for the monocular visual-IMU odometry application.

In our future work, we intend to explore deep learning methods when a large road scene dataset which contains various scenes is available. We may also analyze the results using spatial statistics of image features [44].

## ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments and suggestions to improve this manuscript.

## REFERENCES

[1] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera and R. Toledo, "Multispectral stereo odometry," *IEEE Trans. Intell. Tran. Syst.*, vol. 16, no. 3, pp. 1210-1224, 2015.

[2] I. P. Alonso, D. F. Llorca, M. Gavilán, S. A. Pardo, M. A. Garcia-Garrido, L. Vlacic, and M. Á. Sotelo, "Accurate global localization using visual odometry and digital maps on urban environments," *IEEE Trans. Intell. Tran. Syst.*, vol. 13, no. 4, pp. 1535-1545, 2012.

[3] A. Diosi, S. Segvic, A. Remazeilles, and F. Chaumette, "Experimental evaluation of autonomous driving based on visual memory and image-based visual servoing," *IEEE Trans. Intell. Tran. Syst.*, vol. 12, no. 3, pp. 870-883, 2011.

[4] P. V. K. Borges and S. Vidas, "Practical Infrared Visual Odometry," *IEEE Trans. Intell. Tran. Syst.*, vol. 17, no. 8, pp. 2205-2213, 2016.

[5] D. Scaramuzza and F. Fraundorfer, "Visual odometry [Tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80-92, 2011.

[6] P. Corke, J. Lobo and J. Dias, "An introduction to inertial and visual sensing," *Int. J. Robot. Res.*, vol. 26, no. 6, pp. 519-535, 2007.

[7] K. Mikolajczyk and C. Schmid: "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615-1630, 2005.

[8] H. Zhou, Y. Yuan and C. Shi, "Object tracking using SIFT features and mean shift," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 345-352, 2009

[9] D. Nistér, O. Naroditsky and J. Bergen, "Visual odometry," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recog.*, pp. 652-659, 2004.

[10] A. Davison, I. Reid, N. Molton and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 29, no. 6, pp. 1052-1067, 2007.

[11] X. Dong, X. Dong and J. Dong, "Monocular visual-IMU Odometry: A Comparative Evaluation of the Detector-Descriptor Based Methods," in *Proc. Eur. Conf. on Comput. Vis. Workshops*, Part I, pp. 81-95, 2016.

[12] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: the KITTI dataset," *Int. J. Robot. Res.*, pp. 1229-1235, 2013.

[13] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988.

[14] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 593–600, 1994.

[15] E. Rosten, R. Porter and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105-119, 2010.

[16] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *Proc. Eur. Conf. on Comput. Vis.,* pp. 183-196, 2010.

[17] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2548-2555, 2011.

[18] K. Mikolajcyk and C. Schmid, "An affine invariant interest point detector," in *Proc. Eur. Conf. on Comput. Vis.*, pp. 128-142, 2002.

[19] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91-110, 2004.

[20] H. Bay, A. Ess, T. Tuytelaars and L. Van G, "Speeded-up robust features," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346-359, 2008.

[21] M. Agrawal, K. Konolige and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in Proc. Eur. Conf. on Comput. Vis., pp.102-115, 2008.

[22] G. Bradski, and A. Kaehler, "Learning OpenCV: Computer vision with the OpenCV library, O'Reilly Media, Inc., 2008.

[23] A. Geiger, J. Ziegler and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," in *Proc. IEEE Intell. Veh. Symp.*, pp. 963-968, 2011.

[24] J. Matas, O. Chum,M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," in Proc. Brit. Mach. Vis. Conf., vol. 1, pp. 384-393, 2002.

[25] T. Tuytelaars, "Dense interest points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2281-2288, 2010.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 886-893, 2005.

[27] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 2032-2047, 2009.

[28] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29-44, 2001.

[29] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1-8, 2007.

[30] Z. Wang, B. Fan and F. Wu, "Local intensity order pattern for feature description," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 603-610, 2011.

[31] B. Hariharan, P. Arbeláez, R. Girshick and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 447-456, 2015.

[32] M. Calonder, V. Lepetit, C. Strecha and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *Proc. Eur. Conf. on Comput. Vis.*, pp. 778-792, 2010.

[33] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2564-2571, 2011.

[34] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314-334, 2015.

[35] M. Bloesch, S. Omari, M. Hutter and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, pp. 298-304, 2015.

[36] A. I. Mourikis and S. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 3565-3572, 2007.

[37] J. Hu and M. Chen, "A sliding-window visual-IMU odometer based on tri-focal tensor geometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 3963-3968, 2014.

[38] J. O. Nilsson, D. Zachariah, M. Jansson and P. Handel, "Realtime implementation of visual-aided inertial navigation using epipolar constraints," in *Proc. IEEE Pos. Loc. Navi. Symp.*, pp. 711-718, 2012.

[39] X. Dong, B. He, X. Dong and J. Dong, "Monocular visual-IMU odometry using multi-channel image patch exemplars," *Multimedia Tools Applicat.*, vol. 76, no. 9, pp. 11975-12003, 2017.

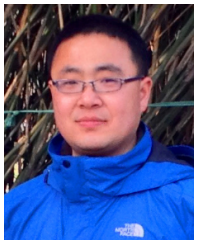
[40] C. Schmid, R. Mohr and C. Bauckhage, "Evaluation of interest point detectors" Int. J. Comput. Vis., vol. 37, no. 2, pp. 151-172, 2000.

[41] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis*., vol. 60, no. 1, pp. 63-86, 2004.

[42] J. Heinly, E. Dunn and J. M. Frahm, "Comparative evaluation of binary features," in *Proc. Eur. Conf. on Comput. Vis.,* pp. 759-773, 2012.

[43] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object, Categories: A Comprehensive Study," *Int'l J. Computer Vision*, vol. 73, no. 2, pp. 213-238, 2007.

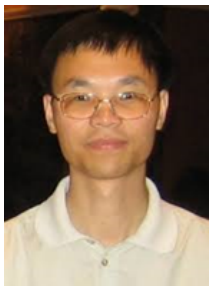
[44] E. Bostanci, N. Kanwal and A. F. Clark, "Spatial statistics of image features for performance comparison," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 153-162, 2014.
[45] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818-832, 2013.
[46] A. Pieropan, M. Björkman, N. Bergström and K. Danica, "Feature Descriptors for Tracking by Detection: a Benchmark," *arXiv preprint* arXiv:1607.06178, 2016.
[47] S. Gauglitz, T. Höllerer and M. Turk, "Evaluation of interest point detectors and feature descriptors for visual tracking," *Int. J. Comput. Vis.*, vol. 94, no. 3, pp. 335-360, 2011.
[48] L. Čehovin, A . Leonardis and M. Kristan, "Visual object tracking performance measures revisited," *IEEE Trans. on Image Process.*, vol. 25, no. 3, pp. 1261-1274, 2016.
[49] N. Govender, "Evaluation of feature detection algorithms for structure from motion," *Council. Sci. Ind. Res.*, *Technical Report*, 2009.
[50] H. Chien, C. Chuang, C. Chen and R. Klette, "When to use what feature? SIFT, SURF, ORB, or A-KAZE features for monocular visual odometry," in *Proc. Int. Conf. Image Vis. Computing*, pp. 1-6, 2016.
[51] A. Schmidt, M. Kraft and A. Kasiński, "An evaluation of image feature detectors and descriptors for robot navigation," *Comput. Vis. Graph.*, pp. 251-259, 2010.
[52] H. E. Benseddik, O. Djekoune and M. Belhocine, "SIFT and SURF Performance evaluation for mobile robot-monocular visual odometry," *J. Image Graph.*, vol. 2, no. 1, pp. 70-76, 2014.
[53] D. Scaramuzza, F. Fraundorfer and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point ransac," in *Proc. IEEE Int. Conf. Robot. Autom.,* pp. 4293-4299, 2009.
[54] Y. Jiang, Y. Xu and Y. Liu, "Performance evaluation of feature detection and matching in stereo visual odometry," *Neurocomput.*, vol. 120, pp. 380-390, 2013.
[55] A. Gil, O M. Mozos, M. Ballesta and O. Reinoso, "A comparative evaluation of interest point detectors and local descriptors for visual SLAM," *Mach. Vision Applicat.*, vol. 21, no. 6, pp. 905-920, 2010.
[56] C. L. Zitnick, and K. Ramnath, "Edge foci interest points," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 359-366, 2011.
[57] P. Dollár, Z. Tu, P. Perona and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint* arXiv:1409.1556, 2014.
[59] B. Zhou, A. Lapedriza, J. Xiao, T. Antonio and O. Aude, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 487-495, 2014.
[60] A. E. Johnson, S. B. Goldberg, Y. Cheng and L. H. Matthies, "Robust and efficient stereo feature tracking for visual odometry," in Proc. IEEE Int. Conf. Robot. Autom., pp. 39-46, 2008.
[61] M. P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. Int. Conf. Pattern Recog.*, vol. 1, pp 566-568, 1994.
[62] R. Hartley, and A. Zisserman, "Multiple view geometry in computer vision," 2nd ed. Cambridge University Press, 2008.
[63] B. Kitt, A. Geiger and H. Lategahn, "Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme," in *Proc. IEEE Intell. Veh. Symp.*, pp. 486-492, 2010.
[64] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communicat. ACM*, vol. 24, no. 6, pp. 381-395, 1981.
[65] L. Zhang and S. Rusinkiewicz, "Learning to Detect Features in Texture Images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6325-6333, 2018.



**Xingshuai Dong** received his MSc degree from the Department of Electronic Engineering, Ocean University of China. He will be a PhD student at the University of New South Wales, Canberra, Australia. His research interests include road scene understanding and vision-aided inertial navigation.

**Xinghui Dong** received his PhD degree from Heriot-Watt University, UK, in 2014. He is currently working as a Research Associate in the Centre for Imaging Sciences, The University of Manchester, UK. His research interests include automatic defect detection, image representation, texture analysis and visual perception.

**Junyu Dong** received his BSc and MSc from the Department of Applied Mathematics at Ocean University of China in 1993 and 1999 respectively. From 2000 and 2003, he studied in the UK and received his PhD in Image Processing in November 2003, from the Department of Computer Science at Heriot-Watt University. Dr. Dong joined Ocean University of China in 2004 and he is currently a professor and the Head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision and underwater image processing.



**Huiyu Zhou** received a Bachelor of Engineering degree in Radio Technology from the Huazhong University of Science and Technology of China, and a Master of Science degree in Biomedical Engineering from the University of Dundee of United Kingdom, respectively. He was then awarded a Doctor of Philosophy degree in Computer Vision from the Heriot-Watt University, Edinburgh, United Kingdom. Dr. Zhou is Reader at Department of Informatics, University of Leicester, United Kingdom. He has published over 200 peer-reviewed papers in the field. His research work has been or is being supported by UK EPSRC, EU, Royal Society, Leverhulme Trust, Puffin Trust, Invest NI and industry.