IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.DOI

State of the Art in Vision-based Localization Techniques for Autonomous Navigation Systems

Yusra Alkendi[1], **Lakmal Seneviratne**[1] **and Yahya Zweiri**[1,2]

[1]Khalifa University Center for Autonomous Robotic Systems (KUCARS), Khalifa University of Science and Technology, Abu Dhabi, UAE, e-mail: {yusra.alkendi, lakmal.seneviratne, yahya.zweiri}@ku.ac.ae.

[2]Faculty of Science, Engineering and Computing, Kingston University London, London SW15 3DW, UK email: y.zweiri@kingston.ac.uk

Corresponding author: Yusra Alkendi (e-mail: yusra.alkendi@ku.ac.ae).

This publication is based upon work supported by the Khalifa University of Science and Technology under Award No. RC1-2018-KUCARS

ABSTRACT Vision-based localization systems, namely visual odometry (VO) and visual inertial odometry (VIO), have attracted great attention recently. They are regarded as critical modules for building fully autonomous systems. The simplicity of visual and inertial state estimators, along with their applicability in resource-constrained platforms motivated robotic community to research and develop novel approaches that maximize their robustness and reliability. In this paper, we surveyed state-of-the-art VO and VIO approaches. In addition, studies related to localization in visually degraded environments are also reviewed. The reviewed VO techniques and related studies have been analyzed in terms of key design aspects including appearance, feature, and learning based approaches. On the other hand, research studies related to VIO have been categorized based on the degree and type of fusion process into loosely-coupled, semi-tightly coupled, or tightly-coupled approaches and filtering or optimization-based paradigms. This paper provides an overview of the main components of visual localization, key design aspects highlighting the pros and cons of each approach, and compares the latest research works in this field. Finally, a detailed discussion of the challenges associated with the reviewed approaches and future research considerations are formulated.

INDEX TERMS Ego-motion Estimation, GNSS-denied, Self-localization, VIO, Visual Inertial Odometry, Visual Odometry, VO.

I. INTRODUCTION

UNMANNED aerial/ground vehicles (UAV/UGVs) have many advantages such as mobility which incorporates flexibility and strength. Thus, they have been employed in a wide range of applications such as for navigation [1], infrastructure inspection [2], [3], agriculture [4], [5], search and rescue [6], [7], and many other purposes. In addition, human-centered robots have become an important research field due to their ability to assist and support humans, i.e., in hospitals, restaurants, and service areas [8]. For safe and efficient autonomous navigation or path planning, a robot should accurately localize itself within the robot environment. Therefore, various studies have investigated the localization problem and many techniques have been proposed [9], [10].

There is an impressive progress in developing and investigating approaches on vision-based navigation systems in the robotics community [11]. This would allow autonomous vehicles to operate in global navigation satellite system GNSS–denied environments and feed the end user with useful information, i.e., real-time 3D reconstruction (map). Visual odometry (VO) is one type of vision-based navigation which estimates the robot's motion (rotation and translation)and to localize itself within the environment. The onboard vision system works by tracking visual landmarks to estimate motion parameters, rotation and translation, between two time instants.

The process of VO is defined as estimating the robot's ego-motion using the information obtained from single or multi-sensors onboard [12]. The gained information from the sensor(s) should represent a sufficient amount of meaningful data (i.e., shape, color, texture, ..etc) to aid the VO process of estimating the sensor's movement relative to the initial state in its surrounding environment [13]. Furthermore,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

Simultaneous Localization and Mapping (SLAM) means a process of robot's localization and simultaneously estimating a robot trajectory and building a map of the environment, thus VO is a subset of SLAM [9]. SLAM process is achieved, similar to VO, by utilizing the information gained from an onboard single or multi-sensors. The performance of VO is affected significantly by the environmental conditions such as illumination conditions and the image quality obtained by the sensor. Furthermore, inertial-based odometry is not affected by the surrounding conditions, however, the performance deteriorates with time. By Fusing the data obtained from visual sensors and the inertial measurements, resulting visual-inertial odometry (VIO) system, overcoming the limitations of both individual state estimators. Therefore, the use of IMU as a complementary sensor to visual-based localization enables obtaining a more robust and accurate pose estimation.

GNSS–denied and low-visibility environments are the main challenges in autonomous systems research since they affect the sensor input information and critically degrade the robot's action. An example of low visibility environment is low-light condition which could be solved by using onboard illumination [14], [15] or single to multi-sensor modalities such as LiDAR (light detection and ranging) and thermal imager [16]. Other low-visibility conditions are still very challenging, including those of smoke or fog-filled conditions. Normal standard cameras, Radars, or LiDARs are used in such harsh conditions for VO or SLAM, but they deliver ill–conditioned data, so consequently are not able to estimate a reliable robot pose, and therefore fail to construct the map of the environment.

This paper presents a survey on vision-based navigation paradigms, namely visual odometry and visual inertial odometry. Our review discusses each approach of the mentioned paradigms in detail in terms of the key design aspects in the main components, and the advantages and disadvantages of each category, where applicable. Localization techniques in low visibility conditions are also presented. Towards the end, the challenges associated with state-of-the-art techniques for self-localization are formulated.

The rest of the paper is structured as follows. Section II provides a brief review of self-localization schemes for navigation in GNSS-denied environments. Section III discusses the evolution of VO schemes under two broad paradigms, i.e., geometric and nongeometric approaches, and evaluates different state-of-the-art implementation choices. Section IV presents a review of recent works pertaining VIO from the literature, their design choices, and system performance. Section V provides the state-of-the-art studies related to localization techniques in visually degraded environments. Section VI presents an overview, discussion of the main aspects, and future research perspectives of visual localization. In Section VII, the outcomes of our review are highlighted and future research considerations in the area are identified.

II. GENERAL OVERVIEW OF LOCALIZATION TECHNIQUES

A main common challenge in autonomous navigation, path planning, object tracking, and obstacle avoidance platforms is to be able to continuously estimates the robot's ego-motion over time (position and orientation). Global Positioning System (GPS) is a conventional localization technique that has been used in various fields of autonomous systems. GPS is one type of Global Navigation Satellite System (GNSS). GPS provides any user, who has a GPS receiver, with positioning information with meter level accuracy [17], and has been employed as a self-localization source such as for drone security applications [18]. On the other hand, GPS suffers from a few limitations that makes it a less reliable alternative sensor for self-localization modules, with a few of these limitations being satellite signal blockage, high noisy data, multipath effects, low bandwidth, jamming, and inaccuracy [10], [19]. Although the rapid development of GPS technologies, i.e., RTK (real-time kinematic) and PPP (precise point positioning), are capable of providing positions with a decimeter or centimeter's level accuracy [20]. The strength of GPS satellite signals depends largely on the environmental conditions, it is effective in clear sky areas and not suitable for indoor navigation where it gets affected by the wall and objects. They are not a good candidate for precise localization which is the main autonomous navigation module.

In the last decade, many studies have investigated odometry techniques for SLAM applications [21]. In such systems, the robot's position and orientation are calculated based on the onboard sensor(s) information. As an opposite to GNSS, the self-contained odometry methods do not rely on external sources (i.e., radio signals from satellite in the case of GPS). Instead, they rely on the use of local sensory information for determining the robot's relative position and orientation with respect to its starting point. The main components of any SLAM technique are the map/trajectory initialization, data association, and loop closure [22]. Odometry algorithm is employed in SLAM system to localize the moving robot within the environment. Then, it is fed into the optimization algorithm for the developed global map to reduce the prediction's drift accumulated from previously estimated poses. Therefore, SLAM techniques are able to reduce the accumulated pose error when the robot returns to a previously observed scene using the history of robot poses in the global map. In addition, odometry algorithms implement local map optimization methods, such as windowed bundle adjustment, to optimize the local map only over the last poses, leading to local map consistency [22], [23]. SLAM aims at maintaining a global map consistency and odometry method is used partially during the SLAM first process which is followed by other steps [24], i.e., local or global map optimization.

Odometry techniques are highly dependent on sensor information which rely on vision, observation, or inertial measurements. Fusion of multiple types of sensing data helps increase the system reliability, robustness, resilience to failures, however, at the cost of the computational complexity effort.

Hence, the overall platform cost would be increased.

The proposed approaches to odometry techniques were surveyed by several researchers in the field and the existing solutions and open research problems were addressed [9], [10], [12], [19], [25], [26]. Figure 1 provides the general self-localization/Odometry techniques proposed in the literature [10]. Mohamed et al. [10] have recently reviewed the odometry methods for navigation and have categorized them based on two main approaches, i.e., GNSS-available and GNSS-denied approaches. They also have classified the GNSS-denied navigation techniques into single and hybrid-based frameworks. The five main categories of single-based approaches are wheel odometry, inertial odometry, radar odometry, visual odometry (VO), and laser-based odometry. Similarly, hybrid approaches can be categorized into visual-laser odometry, visual-radar odometry, visual-inertial odometry (VIO), and radar-inertial odometry techniques. A broader summary of each category was presented along with their advantages and weaknesses. A comparison between the different odometry techniques was also conducted in terms of performance, response time, energy efficiency, accuracy, and robustness. For more detailed information about odometry techniques, interested readers can refer to [10].

For VO, basic concepts and algorithms were described and state-of-the-art proposed techniques were compared by Scaramuzza et al., 2011 [12] and by Aqel et al., 2016 [19]. Poddar et al. [9] have recently reviewed the evolution of VO schemes over the previous few decades and discussed them under two main categories, geometric and non-geometric approaches. A general theoretical background of camera model, feature detection and matching, outlier ejection, and pose estimation frameworks was provided. Furthermore, a list of publicly available datasets for VO was provided. In 2015, VIO techniques have been reviewed in terms of filtering and optimization techniques [25]. Furthermore, for vision-based odometry, [26] have briefly provided a survey based on camera-based odometry for micro-aerial vehicle (MAV) applications in 2016. Their review focused on state-of-the-art studies and evaluation on monocular, RGB-D, and stereo-based odometry approaches.

A considerable body of research addressing the visual localization problem can be found in the literature. Based on the aforementioned surveys, an updated review reflecting the recent advances on VO and VIO is highly required for robotics research community. In this survey:

1) We provide a comprehensive review of the most recent works related to VO and VIO techniques, focusing on achievements made in the past five years (2016-2021).
2) We propose our understanding of the most important studies and successful works related to VO and VIO.
3) We conduct an overview of recent adopted approaches for localization in low-visibility environments. To the authors' knowledge, there is no review has addressed localization techniques in low-visibility environments that reflects the recent advances in the field.
4) We present a detailed discussion of vision-based self-localization systems, as shown in Fig. 13.

This article serves as a building block for researchers and developers to understand the basic concept, to compare and categorize existing applied paradigms, and to highlight open research problems to improve the recent self-localization techniques. In addition, it will provide key systematic points for the user on how to select an appropriate localization method for navigation based on the environmental conditions and application needs.

III. VISUAL ODOMETRY

VO is defined as the pose estimation process of a robot, human, or vehicle by evaluating a set of cues (variations) in a sequence of images of the environment obtained from a single to multiple cameras [9]. In short, VO means localizing the camera or sensor within the environment. VO is utilized in many applications such as navigation and control of robotics (i.e, aerial, underwater, and space robotics), automobile, wearable computing, industrial manufacturing, and etc [23], [27].

The concept of VO is similar to the wheel odometry incremental estimation of the vehicle's pose and motion by integrating the number of wheel turns over time. Equally, VO incrementally estimates the pose by evaluating the variations of motion induced on a set of images captured by on-board camera(s). VO is considered as a case of structure from motion (SfM) technique which is utilized to reconstruct a 3D scene of the environment and camera poses from a consecutive sequence of frames [12]. A 3D view is reconstructed by calculating the optical flow of key indicators, in which they are extracted from two consecutive frames using image feature detectors (i.e Moravec [28]) and corner detectors (i.e Harris [29]). Then, refinement/optimization of the constructed 3D structure is done by using the bundle adjustment method [30] or any other offline refinement technique. There are several ways to perform SfM depending upon many factors such as the number of on-boarded cameras, the number and order of images, and the camera calibration status. The last step in SfM is the refinement and global optimization of the structure and camera pose, it requires a high computation load, therefore it is performed offline. In contrast, VO is conducted in real time (online) to estimate the camera pose [31]. VO works effectively in conditions where the environment offers a sufficient illumination level, and a static scene with rich textures that are enough to aid observing and extract the apparent motion, and when enough scene is overlapped between consecutive frames.

A. MOTION ESTIMATION

The main pipeline of VO system is provided in Fig. 2. There are three standard VO motion estimation methods, which are segregated into 2D to 2D, 3D to 2D, and 3D to 3D motion estimation techniques. The methods are used to compute the transformation matrix between two consecutive images (the current and previous image). They depend on the captured

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

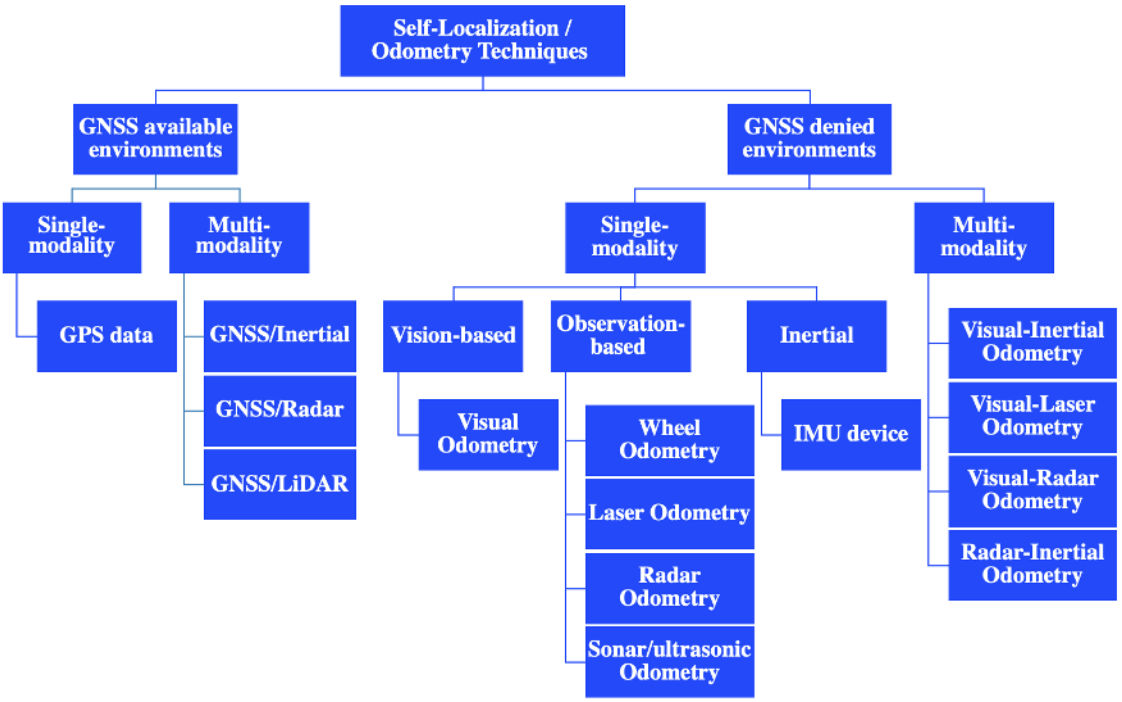

FIGURE 1. **Self-localization/odometry Techniques.**

features and their correspondences whether specified in 2D or 3D [1]. Appending these single estimated motions at a time would help in estimating the full robot trajectory. Lastly, bundle adjustment process is performed to iteratively refine the pose estimated over the last number of frames [12]. Figure 3 illustrates the VO scheme. At first, a relative pose, $T_{i,i+1}$, between cameras are determined by matching the location of the corresponding feature points of two consecutive 2D images. By using one of the mentioned VO motion estimation methods, the 3D point pose would be computed. Then the global camera poses, C_i, are computed using the concatenated relative transformations which are relative to an initial reference frame.

1) 3D to 3D algorithm

In this algorithm, the camera motion relative to an initial state is computed in the following steps. At first, match a set of 3D points extracted from a pair of successive images. Secondly, triangulate the 3D matched features between frames. The relative camera motion is estimated by the transformation of two consecutive frames that is computed based on minimizing the Euclidean distance between two corresponding 3D features [12].

2) 3D to 2D algorithm

Similar to the previous algorithm, the aim is to determine the transformation matrix that relies on minimizing the 2D reprojection error of its correspondence 3D feature points. The cost function is depicted by Eq. 1.

$$T_t^k = argmin_{T_t^k} \Sigma_i \| P_t^i - P_{t-1}^i \|^2 \tag{1}$$

where T_t^k is the transformation matrix to minimize the projection error between two consecutive frames t-1 and t. P_t is the 2D point image feature at the current frame whereas P_{t-1} is the 2D point reprojected from a 3D point feature into a previous image frame. This approach is also called the perspective-n-points (PnP) algorithm, as it estimates the camera pose using a k group of i number of 3D points into 2D. The minimum set of points required is determined by the number of constraints in the system. For instance, a minimal solution is called perspective-3-point (P3P) [32] utilizing a set of three 3D points into 2D to estimate the camera pose.

3) 2D to 2D algorithm

In this algorithm, there are three main steps that are used to estimate the motion. Firstly, the essential matrix (E) relates the geometric relation of two successive frames and it is defined by matching the 2D feature correspondences using the epipolar constraint, as shown in Fig 4. The essential matrix (E) and translation matrix ($\hat{t_k}$) are defined as Eq. 2 and Eq. 3, respectively.

$$E_k \simeq \hat{t_k} R_k \tag{2}$$

where t_k and R_k are the translation and rotation parts of camera motion parameters [12].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

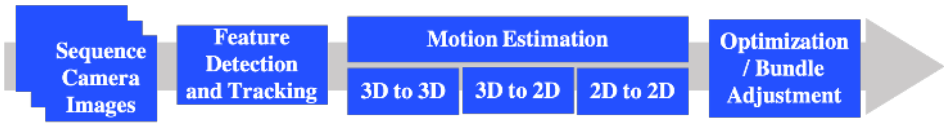

FIGURE 2. VO General Pipeline [12].

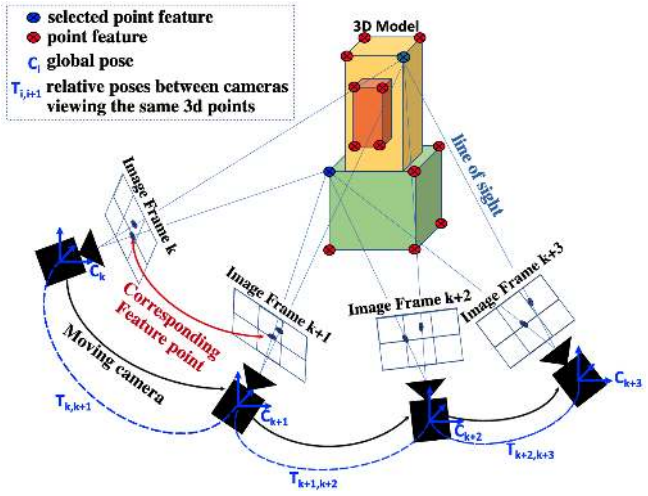

FIGURE 3. Illustration of VO Scheme.

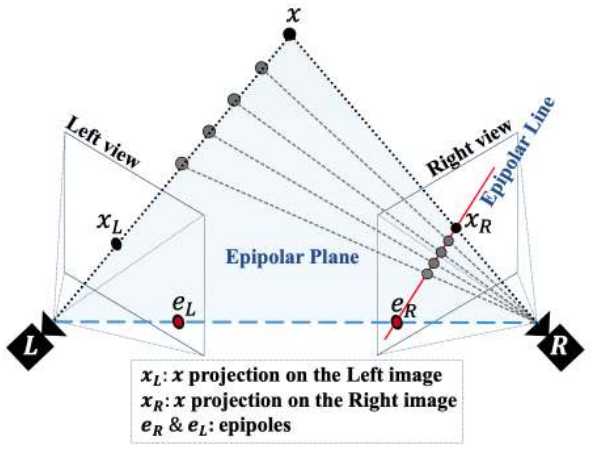

FIGURE 4. Illustration of Epipolar Geometry.

$$\hat{t_k} = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{3}$$

The translation vector (t_k) is defined as Eq. 4.

$$t_k = [t_x, t_y, t_z]^T \tag{4}$$

To compute E, a simple approach, the 8-points method is proposed by [33] which employs an 8 or less noncoplanar corresponding points of two successive images. Subsequently, a simple and widely used method is introduced by [34], called the Nister five-point algorithm which uses a set of five matched points to define the geometric relation of two sequential images. Then, decompose E into the rotation and translation information to form the transformation matrix, wherein finally, camera motion would relatively be estimated.

To conclude, 3D to 2D based motion estimation is faster in practice than the 2D to 2D algorithm [35]. In addition, it estimates the camera pose with higher accuracy than the 3D to 3D algorithm, since it relies on minimizing the reprojection error rather than 3D to 3D position error [12], [19], [21]. In what follows, a review of different implementation design choices of VO studies is presented (Subsection III-B).

B. KEY DESIGN CHOICES

The visual odometry scheme can be described as a platform made up of collecting sensor data and a processing architecture to provide an instant camera pose. Techniques to estimate camera pose can be classified into appearance or feature-based VO. Figure 5 provides a general classification of the VO systems, based on the visual module(s) used and the key approach selected. The used vision module consists of the type of visual sensor and the sensor placement and orientation on the robot, can be either facing forward or downward. The type of visual sensors could be monocular, stereo, RGB-D, omnidirectional, thermal, and event-based cameras. The first key design, i.e., the appearance-based VO method, estimates the motion by assessing the intensity value of the image pixel of two successive frames by, for example, the optical flow algorithm [36]. On the other hand, the feature-based VO works by tracking the detected points of interest through vectors that represent the tracked point's local region. This method purely depends on the image texture, hence it is not relevant for low feature-based conditions such as dark or sandy environments [37]. In addition to that, VO can be performed by a combination of the former methods utilizing hybrid information aiming at a more robust and efficient estimation. VO techniques can be also categorized into conventional and non-conventional approaches/techniques. The conventional approaches use camera geometrical relations to assess the motion. On the other hand, the non-conventional approaches are based on Machine Learning tools (i.e., a regression model) trained by VO parameters to estimate the motion [38], [39]. The advantage of employing the learning-based VO method over the others is that the initialization of camera parameters is not required as well as the process of correcting the scale of the estimated

trajectories, as in the case of monocular VO, is not needed [39]. The mentioned design choices of VO are elaborated in detail in the following subsections.

1) Conventional – Appearance-based VO

The appearance-based VO estimates the camera pose by analyzing the intensity of the captured image pixels based on minimizing the photometric error. Unlike the feature-based VO, this method uses all geometrical information of the captured camera frames, reducing aliasing issues related to similar pattern scenes and enhancing the pose estimate's accuracy and system robustness, especially when utilized for low textured and low visibility environments [40]. Figure 6 illustrates the main pipeline of appearance-based VO paradigms. The principle of appearance-based VO can be classified into the *region/template matching-based* and the *optical flow-based* methods.

For *regional-based method*, the motion is estimated by concatenating camera poses by performing an alignment process for two consecutive images. This technique has extended its implementation by measuring the invariant similarities of local areas and using global constraints. Vatani et al. [41] proposed a simple localized approach, relying on a constrained motion of a large vehicle. It used a modified correlation-based VO method with respect to the variation in size and location of the correlation mask based on the vehicle movement and fed a prior suggested prediction area in the mask for matching. Hence, its ability to reduce the computational time makes it more reliable for practical implementation. An extension of this work was proposed by Yu et al. [42] by utilizing a rotating template instead of a static template to find the translation and rotation between two consecutive images.

Furthermore, an adaptive template matching method was proposed by [43] utilizing a smaller mask size and by varying the template location with respect to vehicle acceleration. Several studies have incorporated visual compass with the template-matching based method for estimating the pixel displacement between images [40], image rotation for a more robust system with respect to accumulated camera calibration errors over time [37], image rotation and translation employing different cameras [36].

Studies on robust regional-based matching methods utilized for other purposes that could be implemented for VO problems are discussed next. Comport et al. [44] proposed a scheme of utilizing a pair of stereo images and matching its dense correspondences to estimate the 6-DoF pose. The process relies on the quadrifocal between the image pixel intensities that makes the system more robust under various conditions of occlusion, pixel-wise displacements, and illumination variations. In addition, Comport et al. [45] have expanded his work by adding a cost function to minimize the intensity errors of the whole image. Moreover, Lovegrove et al. [46] have assessed vehicle motion using image alignment techniques and aided by the features on road surfaces.

Other studies have also been performed on regional-based matching techniques by analyzing the motion parallax to compose 3-D translation and transformation of two successive images. Motion is estimated in Large Scale Direct-SLAM [47] by the image alignment method that relies on the depth map. The proposed framework consists of stereo and monocular cues and is able to compensate for brightness variations of image frames for a more accurate pose prediction. Furthermore, Engel et al. [48] have examined a direct sparse VO method based on optimizing photometric error, similar to the sparse bundle adjustment scheme, achieving a robust motion estimate by utilizing all image points, unlike featured-based VO which utilizes key geometrical points only.

For *Optical Flow-based method*, raw visual pixel data are imposed into the optical flow (OF) algorithm, wherein the pixel intensity change of two consecutive frames from the camera(s) is analyzed to estimate the motion [49]. As the illumination of a pixel varies, the camera motion would be defined by computing the 2D displacement vector of points projected on two frames. Works of Brox et al. [50] and [51] provide an example of a widely used OF methods that use motion constraints equations. Techniques of optical flow-based VO are also called direct methods since they utilize the whole image information and it is used for 2D/3D motion estimation paradigms. Kim et al. [52] proposed a method to handle problems of motion cease and changes of illumination conditions, by employing an integrated method of Black and Anandans [53] and Gennert and Negahdaripours [54], respectively, to estimate camera motion.

Campbell et al. [55] have employed the optical flow method to assess the robot ego-motion parameters. Rotation and translation are estimated by the far and nearby features of the images, respectively. For navigation in an unexplored environment, Hyslop and Humbert [56] have utilized an optical flow approach imposing a wide range of raw visual measurements to estimate a 6-DoF motion task. Grabe et al. [57] have estimated the continuous motion of a UAV by employing the optical flow method in a closed-loop operation instead of incremental estimating the motion in frame-to-frame way. Moreover, they have extended the work of [58] to improve velocity estimation by combining features in the optical flow technique. In addition, optical flow algorithms have been implemented to aid UAV navigation for other purposes such as object avoidance [59].

Some limitations of optical flow-based schemes are related to the strength of the environment texture as well as to the computational constraint. To overcome and minimize the computational energy consumed, RGB-D camera is utilized for VO problems and to estimate the motion by minimization of the photometric error in the dense map, such as in Kerl et al. [60]. Furthermore, the method proposed by Dryanovski et al. [61] has aligned 3D points on the global map by an iterative closest point (ICP) algorithm. In addition, a fast and low computed VO method was developed by Li and Lee [62] where the intensity values of selected key points were analyzed by ICP.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

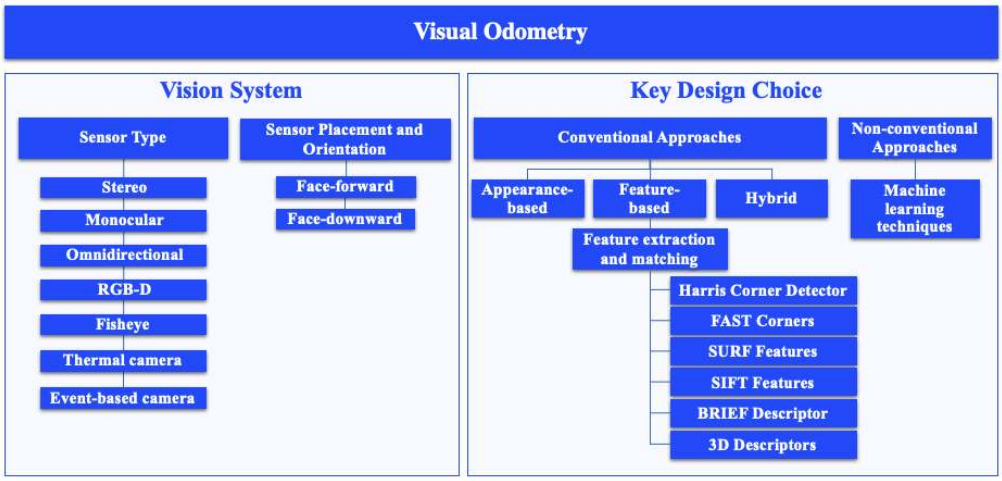

FIGURE 5. General Classification of VO Techniques Proposed in Literature.

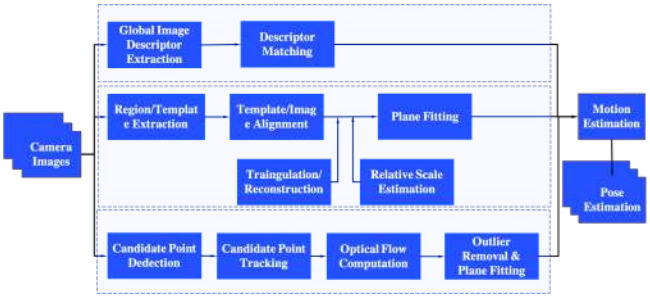

FIGURE 6. Main Pipelines of Conventional – Appearance-based VO Technique [9].

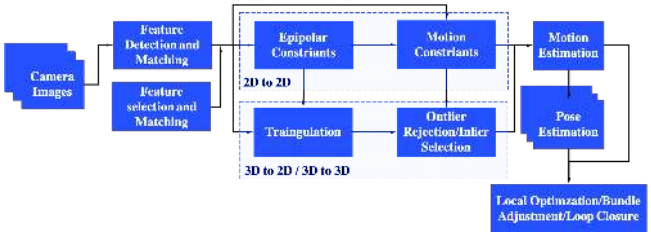

FIGURE 7. Main Pipelines of Conventional – Feature-based VO Technique [9].

2) Conventional – Feature-based VO

Featured based VO techniques start by targeting areas having key distinctive information such as lines, curves, edges, or corners, between successive image frames. Then, the matching and tracking of these features is performed by an optimization method to reduce the geometric error. Lastly, computing the transformation matrix is performed to estimate the motion [63]–[65]. Figure 7 illustrates the main pipeline of feature-based VO approach. Some of the feature detectors utilized in the literature include Harris detector [29], Shi–Tomasi corners [66], maximally stable extremal regions (MSER) [67], Laplacian of Gaussian detector [68], Difference of Gaussian [69], and features from accelerated segment test (FAST), adaptive and generic accelerated segment test (AGAST), and optimal accelerated segment test (OAST) [70]. In addition, some of the used feature descriptors are binary robust independent elementary features (BRIEF) [71], speeded up robust features (SURF) [72], scale-invariant feature transform (SIFT) [73], oriented FAST and rotated BRIEF (ORB) [74], and binary robust invariant scalable keypoints (BRISK) [75]. A comprehensive review on recent advances in feature detection and description algorithms is provided by [76].

Feature-based VO technique is robust when utilized in environments with high geometric distortions and it is independent of the illumination variations [77]. Furthermore, it may discard some of the valuable image data as it only extracts the key detected features. The post process of detecting features, which is to extract and match between frames, requires a high computational energy cost that is proportional to the number of features that have been extracted in the process. On the other hand, the higher the number of features extracted, the more accurate the pose estimation will be. To increase the feasibility of utilizing feature-based VO technique for resource-constrained platforms, i.e., UAV [78], certain key features are only maintained to be extracted. Kitt et al. [79] proposed a framework to optimize the results of VO method by uniformly distributing the location of extracted features within image frames as well as reducing the required computational load in the process. The approach was based on a bucketing technique to select an appropriate number of key features in which the image frame would segregate into grids and each such grid would have a certain number of key features that are used in the matching process. The authors [80] have then extended the method by adding a classification block following the bucketing stage, wherein the key features at each grid were sorted into moving or not moving features considering the randomized decision tree

model. Later, Cvisic and Petrovic [81] have also classified features of each bucket, however, into four different groups to enable the selection of good features for better pose estimation results.

Maeztu et al. [82] have assessed the complete feature-based VO framework utilizing the bucketing method for tracking and matching using feature descriptors in corresponding grids. This approach helped to improve the estimated motion by adding an external block. The purpose of the external block was to perform parallel computation (as in a multi-core framework) and reduce the outliers. Several studies have improved the results obtained from a feature-based VO system, however, not in the feature detection or tracking methods. For example, Badino et al. [83] have improved the accuracy of the estimated motion by averaging the key feature locations with respect to its all previous occurrences. Furthermore, Kreso and Segvic [84] have initially calibrated and corrected camera parameters by comparing and matching the corresponding points between frames employing ground truth motion. Cvisic and Petrovic [81] have utilized a five-point algorithm to estimate the camera rotation and translation that relied on minimizing the reprojection inconsistency for a combined stereo and monocular VO setup. Camera rotation was estimated by monocular case to overcome the error of an imperfect calibration, whereas camera translation was estimated by the stereo case to improve the results accuracy.

The design of the neuromorphic vision sensor, event-based camera, makes it an ideal alternative and indispensable for platforms that require accurate motion estimation and good tolerance in challenging illumination conditions. Event-based visual odometry (EVO) approach has been proposed by [85] to compute the camera pose estimation with high precision and obtain a semi-dense 3D map environment. Due to the event-based camera characteristics, the proposed pose estimation method was very efficient and feasible to be performed in real-time on a standard CPU.

3) Conventional – Hybrid-based VO

For low-textured scenarios, feature-based VO schemes are not considered as a robust scheme since only a few features are to be detected and tracked. On the other hand, the appearance-based VO schemes exploit all image information for detecting and matching process between frames, leading to a more efficient outcome at the cost of a considerable computational power. Thus, hybrid methods have been introduced to combine advantages of the two above-mentioned schemes. Scaramuzza and Siegwart [37] have utilized a hybrid VO framework wherein the translation of a ground vehicle was estimated by feature-based method and the rotation was obtained by the appearance-based method. In such a scheme, the vehicle pose would be estimated at a lower cost of the computational load compared to the feature-based ones.

Furthermore, a semi-direct VO framework was proposed by Forster et al. [86] in which the camera pose was estimated by two main phases: the relative camera pose to the prior frame (feature correspondences) was estimated by minimizing the photometric error (appearance-based scheme), whereas camera pose estimation relative to the structure was assessed by minimizing the reprojection error (feature-based scheme). Such a hybrid approach improves the estimation accuracy and eliminates the cost of feature extraction per frame. Silva et al. [87] utilized a dense appearance-based VO to estimate the vehicle-scaled rotation and translation incorporated with featured-based method to recover the scaling factor accurately. Moreover, Feng et al. [88] presented a localization system dependant upon the environmental conditions and consists of parallel direct (appearance-based) and indirect (feature-based) modules. Camera poses would be estimated by the direct method for low texture conditions and would be shifted to the indirect-based method if enough features were detected within the frame. Alismail et al. [89] proposed a hybrid framework wherein binary feature correspondences were aligned using the direct-based VO to increase system robustness, especially in low light scenarios.

4) Non-conventional – Machine Learning-based VO

With the development of Machine-learning tools, recent VO schemes have shifted towards learning- based approaches for more accurate motion estimation as well as for achieving faster processing speed of data. In addition, one of the advantages of utilizing VO based learning frameworks is that the results could be obtained without the need of a prior knowledge of camera parameters. Once a suitable training dataset is available, the developed regression or classification model would aid and improve ego-motion estimation. For example, it could be utilized for scale correction by estimating the translation and it is robust to deal with noises and outliers by which it is trained. Figure 8 provides a learning methodology of learning-based VO paradigm. The network is trained using sequence of successive frames as the input information to predict depth information, motion parameters, or pose estimation as the ground truth output data.

As an example of the earliest work on learning-based VO, Roberts et al. [90] divided each image into blocks. Then, they developed a k-Nearest Neighbor (KNN) regression model that was trained to compute the optimal flow for each block. Motion was then estimated by a voting system between distinct blocks. Moreover, Roberts et al. [91] proposed another learning-based method to estimate the optical flow by a linear subspace if there were considerable depth regularities relative to the robot motion in the environment. The expectation–maximization EM algorithm has been utilized to enhance the learning of subspace properties.

Similarly, Guizlini and Ramos [92], [93] have developed Coupled Gaussian Processes (CGP) as a regression model to obtain optical flow feature parameters. This work was later extended in [94] whereby they introduced a CGP for the VO problem. The CGP has enhanced the multitask capability of the VO system to exploit the correlation between the permitted multitasks through the coupled covariance functions. Furthermore, to enhance the system performance, they modified

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

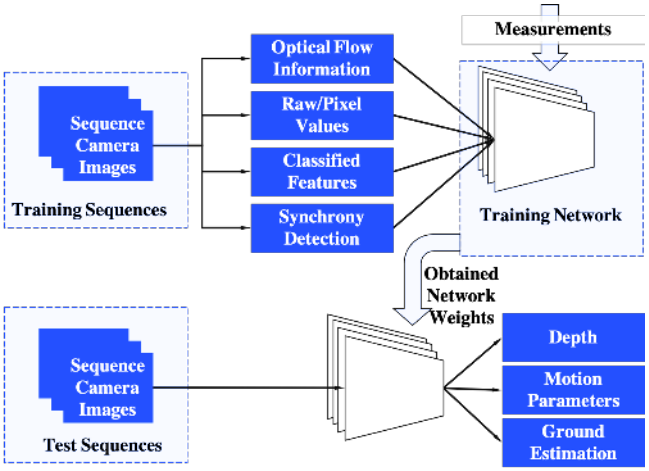

FIGURE 8. **Main pipeline of Non-conventional – Learning-based VO Technique [9].**

the common zero mean assumption of GP by using a standard geometric model of the camera. This would provide an initial estimate which is then refined by the non-parametric model. This fusion step permitted the inclusion of prior knowledge in a simple direct way. In a similar manner, the hypothesis with the other mean of sensor fusion is also possible like the Kalman Filtering [95], [96].

The utilization of a convolutional neural network (CNN) based approach was presented by Konda and Memisevic [97] to jointly estimate the depth and motion information obtained across image pairs. Later, in [98], the study was expanded utilizing CNN model for VO problems to estimate changes in local velocity and its direction. In the work of Mohanty et al. [99], a deep CNN model was developed to extract useful feature sets between two time series data streams for estimating transformation. The recurrent CNN was developed for pose estimation in an end-to-end manner that was trained by a series of geometrical features. In addition, CNN model was developed using monocular vision to estimate the vehicle's position in a true scale. Furthermore, Peretroukhin et al. [98] have combined a recurrent CNN with a Bayesian CNN to infer the sun direction to improve VO ego-motion estimation. For a more accurate and robust VO system, Clement and Kelly [100] have incorporated deep neural networks (DNN) trained by image canonical appearance to enhance pose estimation.

To achieve high monocular VO accuracy, Jiao et al. [101] utilized a learning framework by combining CNN and Bi-LSTM to leverage the feature properties of image pairs and to permit understanding of the relationship between the features of successive images, respectively. Another approach to VO problem is the development of RCNN proposed by Liu et al. [102], which is a learning-based model in an end-to-end training manner employing RGB-D sensors. The inclusion of depth information along with monocular imaging helped in injecting an image scaling factor, and thus, inferring an accurate pose from monocular images. Recently, Wang et al. [103] have exploited a new framework based on two models, the first one called Deep Siamese convolutional neural network (DSCNN) and the second one called DL-based Monocular VO which depends on the first network. For recovering the camera trajectory, DSCNN model was trained by the geometrical relationship of consecutive images to find a 6-DoF camera pose, and the developed model was validated through experiments. Table 1 provides an overview of recent VO-based studies with a summary of their implementation frameworks.

IV. VISUAL INERTIAL ODOMETRY

Inertia-based navigation systems have been traditionally used in autonomous vehicles to measure motion in GNSS-denied environments (i.e., urban environments and indoor operations). Such systems rely on an onboarded 6-axis inertial measurement unit (IMU) which measures the vehicle's local linear acceleration and angular velocity. Recently, with the advancements in hardware measurement technologies, micro-electro-mechanical IMUs emerged as a convenient inertial sensing device that fits any mobile platform (such as micro aerial vehicles) and comes in a compact and light weight module, and at low cost with high accuracy levels.

Although it has been emerged in a wide range of real-time augmented reality applications in mobile devices [142], unfortunately, the high rate of IMU data are corrupted with noise and biases. Therefore, the performance of inertial-based odometry methods deteriorate with time and are unreliable for long-term pose estimation. Integrating the VO pipeline with an inertial-based localization method would overcome the limitations of each individually-based approach, yielding visual inertial odometry (VIO) systems. Images capture features of the scene, while the IMU data provide accurate pose estimation in a very short time at high frequency in alleviating the impact of moving objects on the visual sensor estimation. Thus, the use of IMU as a complementary sensor to visual-based localization enables obtaining a more robust and accurate pose estimation.

Figure. 9 presents the general categories of the existing solutions of VIO systems. They can be categorized based on the processing stage where sensor fusion occurs into three categories, namely loosely-coupled, semi-tightly coupled, and tightly-coupled approaches. In addition to this, VIO systems can also be classified based on the type of data fusion between the visual and IMU data into filtering-based and optimization-based approaches. Moreover, similar to VO-based methods, the type of visual sensor used and how key information are selected for pose estimation process are analyzed next. Visual sensors could be one of the following kinds, i.e., monocular, stereo, RGB-D, omnidirectional, thermal, and event-based cameras. Table 2 reviews recent state-of-the-art VIO studies.

A. DEGREE OF DATA FUSION

TABLE 1. Review of state-of-the-art VO approaches

Ref	Author(s) Year	Vision System		Key Design Choice	Feature detectors / descriptors	Estimation/Solver	Comments
		Type / No. of camera(s)	Camera Pose Direction				
[64]	Saurer et al. 2016	Stereo	Facing down	FB-VO	SURF	• Points or features were assumed to be on a known plane. • Estimation was based on 2-3 point algorithm.	• Bundle Adjustment was required.
[104]	Escalera et al. 2016	Stereo	Bird View	FB-VO	SIFT	• Features were selected from the ground surface (only static features were detected and tracked). • Estimation was based on PnP algorithm.	• A Kalman filter was used to improve odometry estimation. • Bundle Adjustment was not required.
[105]	Zhou et al. 2018	Monocular	Facing forward	FB-VO	ORB	• Random fern classifier was used for matching features between images. • Estimation was based on 8 points normalization algorithm.	• Results were experimentally validated over 2400 images. • Results were way better than the traditional VO method in terms of lower computational time and errors computed over distance and angles.
[106]	Borges & Vidas 2016	Monocular (Thermal camera)	Facing forward	FB-VO	SURF	• Features were detected from the segmented ground surface/plane. Ground plane was segmented using gradient-based watershed method [107] for better scale estimation. • The periodic nonuniformity correction (NUC) process of thermal cameras was solved by proposing an efficient NUC management module which was automatically performed based on the current and predicted camera rotations.	• Five different experimental setups were evaluated. • The Euclidean distance error of the trajectory as well as the median and variance of the errors were considered for evaluation analysis. • The experimental results showed similar performance to VO-based on visible spectrum camera with the advantage of being able to perform effectively at nighttime.
[108]	Kueng et al. 2016	Event-based camera	Facing forward	FB-VO	Edges (Canny's method) and corners (Harris detector [29])	•Features were detected in greyscale image frames (traditional method) and then asynchronously being tracked using high dynamic range event data. • Two cooperative methods were used to track detected features that correspond to event data using events' spatial histograms for short- and long-term tracking, respectively.	•Tracking errors as a function of time, position error (Euclidean distance), and the orientation error (geodesic distance) were computed for evaluation against ground truth (motion capture device). •The experimental results more accurately estimated 6DoF camera ego-motion (event-based) compared to SVO-method (frame based) [86] and motion capture (ground truth) when tested on natural scenes and rich in brightness changes of different magnitudes.
[109]	Liu et al. 2017	Fisheye (two-stereo fisheye cameras)	Facing forward	FB-VO	AGAST corner detector [70].	• Plane-sweeping stereo algorithm was used for fisheye stereo matching and depth initialization. • Semi-dense direct image alignment algorithm was used for camera ego-motion estimation.	• Position and orientation errors averaged over all possible trajectory lengths were computed for evaluation. • Experimental results are proved to achieve significantly accurate motion estimates as well as a high-quality point cloud, compared to results of the semi-direct fisheye stereo VO method of Heng and Choi [110].
[111]	Kottath et al. 2017	Stereo	Facing forward	FB-VO	SURF	• Inertial measurements were considered to restrict feature selection. • 3D-3D motion estimation method was considered.	• Results were validated using KITTI vision dataset. • Results were way better than the normal VO method (without IMU consideration).
[65]	Guan et al. 2018	Monocular	Facing forward	FB-VO	SIFT	• Vertical/gravity vector was known. • Ground plane was known. • Estimation was based on 1.5 points algorithm.	• A Kalman filter was used to improve odometry estimation. • Bundle Adjustment was not required.
[112]	Almalioglu et al. 2018	Monocular	Facing forward	ML-VO	NA	• Adversarial and recurrent unsupervised learning framework was used to jointly estimate the 6DoF camera pose and depth map. • Convolution part of Pose Regressor network extracted features which then were passed through the recurrent neural network (LSTM) to estimate the camera pose. • The framework did not require any global optimization techniques (i.e., loop closure detection or bundle adjustment).	• Developed framework was validated using KITTI [113] dataset of 11 driving scenarios. • Results were way better than the monocular ORB SLAM and other learning-based state-of-the-art VO methods [114]–[116], in terms of absolute trajectory error (ATE) when compared to ground truth.

(continued)

TABLE 1. Review of state-of-the-art VO approaches (continued)

Ref	Author(s) Year	Vision System		Key Design Choice	Feature detectors / descriptors	Estimation/Solver	Comments
		Type / No. of camera(s)	Camera Pose Direction				
[117]	Mahjourian et al. 2018	Monocular	Facing forward	ML-VO	NA	• Enhanced depth and ego-motion estimation were used using an unsupervised learning technique, by involving differential 3D loss functions to distinguish geometrical consistency within adjacent frames. • No need to have a calibrated sequence of images/video. The model could be tested/trained on a low-quality uncalibrated video.	• The developed framework was validated using KITTI [113] • The model results were way better than the supervised method, Monocular ORB SLAM, and another unsupervised learning-based VO method [114], in terms of absolute trajectory error (ATE) when compared to the ground truth.
[118]	Valada et al. 2018	Monocular	Facing forward	ML-VO	NA	• End-to-end supervised learning framework which took two consecutive images and simultaneously estimates 6-DoF global pose odometry. • Geometric Consistency Loss were utilized to elevate the network learning curve effectiveness for accurate pose estimation.	• The proposed framework was validated using Microsoft 7-Scenes [119], consists of 7 indoor scenes, and Cambridge Landmarks datasets [120], consists of 5 different outdoors scenes). • The model results were way better than three other state-of-the-art CNN models, DeepVO [121], cnnBspp [122] and LBO [123], in terms of transnational and rotational errors as a function of sequence length by 27.0% and 16.67%, respectively.
[102]	Liu et al. 2019	Monocular	Facing forward	ML-VO	NA	• Unsupervised end-to-end training framework. • Depth information along with monocular images were used to train the network to include scale information. • The convolution layer of the recurrent network was used to extract feature maps from images. • No postprocessing process was required to recover the trajectory scale nor optimize pose predictions.	• Root Mean Square Error (RMSE) was adopted for evaluation. • Experimental validation was based on KITTI VO [124] dataset (22 stereo sequences). • The pose predictions provided competitive estimations compared to other state-of- the-art VO systems, (I) unsupervised model, SfMLearner [114], and (II) FB-VO methods, VISO2-Mono and VISO2-Stereo [125].
[126]	Jaramillo et al. (2019)	Omni – single camera	Facing forward	FB-VO	ORB	• Pose was estimated by using a single camera. • 3D-2D motion estimation method was considered (P3P algorithm). • No motion assumptions were considered. • The framework did not apply any graph optimization or loop closure techniques.	• The relative pose error (RPE) and the absolute trajectory error (ATE) were adopted for evaluation. • The indoor experimental results, performed under practical sensing range, achieved comparable pose estimations compared to the RGB-D based VO method. • Single-Omni camera-based VO demonstrated its ego-motion capabilities in practical conditions for real time aerial navigation system.
[127]	Wang et al. (2019)	Omni – single camera	Facing forward	ML-VO	NA	• End-to-end supervised learning framework which takes Omni image sequence for camera pose estimation using trained CNN. • The influence of 6 different FOVs of Omni image on CNN-based VO model were examined.	• Average errors in position and rotation were considered for evaluation. • The use of a perspective mosaic image representation (PMIR) as Omni image representation has provided the best pose predictions compared to an equirectangular (EIR) or multichannel perspective (MCPIR) image representation.
[128]	Dai. et al. (2019)	Stereo setup (Normal and thermal cameras)	Facing forward	AB-VO	NA	• The proposed framework was based on Multi-spectral visual odometry using the modified direct sparse odometry (DSO) technique [48]. • Explicit stereo matching was not required to process multispectral data of two sensors (normal and thermal camera)	• Translational Root Mean Square Error (RMSE) of Absolute Trajectory Error (ATE) was used for evaluation. • Seven indoor sequences were recorded in different environments to prepare a multispectral dataset that has a motion capture system as the ground truth. • The experimental results have achieved comparable and more robust pose estimations compared to the ORB-SLAM2 [129] and DSO [48].
[130]	Li et al. (2020)	Monocular	Facing forward	ML-VO	NA	• Online meta-learning algorithm – unsupervised model. • End-to End framework. • ConvLSTM network was considered to enable the adaptation of previous spatial-temporal scene information for better estimation in the current frame.	• Results were validated using KITTI vision dataset [124] (11 driving scenarios), as outdoor data, and TUM-RGBD dataset as indoor data. • Evaluation criteria were based on RMSE. • Results were outperforming three state-of-the-art self-supervised VO methods.

(continued)

TABLE 1. Review of state-of-the-art VO approaches (continued)

Ref	Author(s) Year	Vision System		Key Design Choice	Feature detectors / descriptors	Estimation/Solver	Comments
		Type / No. of camera(s)	Camera Pose Direction				
[131]	Zhan et al. (2020)	Monocular	Facing forward	FB & ML-VO	NA	• Hybrid approach of deep learning method with epipolar geometry and PnP method for better estimating the depth and optical flow. • Deep neural network was used to establish 2D-2D/3D-2D correspondences for pose estimation. • Framework did not suffer from scale-drift issue.	• Extensive experiments were evaluated using KITTI vision dataset [124] by measuring Absolute trajectory error (ATE) and Relative Pose Error (RPE) • Hybrid approaches, by integrating deep predictions with geometry-based methods, have proved to provide better estimates compared to end-to-end learning-based or convention geometry-based VO systems.
[132]	Zhai et al. (2020)	Monocular	Facing forward	ML-VO	NA	•End-to-end supervised learning framework that takes a pair of images to estimate its corresponding ego-motion. • Data augmentation techniques were utilized to avoid overfitting while training the neural network model.	• Evaluation was based on KITTI VO [124] and Malaga 2013 [133] datasets. • The experiment results provided a competitive pose prediction compared to state-of-the-art monocular geometric [134], [135] and learning methods [121], [136]. Thus, they encouraged researchers to further explore and investigate learning-based VO methods.
[137]	Huang et al. (2020)	Stereo	Facing forward	FB & ML-VO	ORB	• At first, incoming frames were passed through YOLO object detection network to detect and generate semantic bounding boxes. In addition, ORB algorithm was used to extract features and establish matched correspondences across stereo frames. Then, the observed landmarks in the current frame were clustered into different objects by Heterogeneous CRF module. Pose was estimated by performing a sliding window technique that employed an optimized and a novel double-track frame management design. Lastly, the estimations were passed into the optimization module before being updated in the static maps and clusters.	• Standard Bundle Adjustment was required. • The employed metrics for evaluation were Root Mean Square Error (RMSE) of the Absolute Trajectory Error, the Rotational and Translational Relative Pose Error. • Evaluation was based on Oxford Multi-motion (OMD) [138] and KITTI VO [124] datasets. • The experimental results provided competitive pose estimations compared to state-of-the-art systems including ORB-SLAM2 [129], DynSLAM [139], Li et al. [140] and ClusterSLAM [141].

Omni:Omnidirectional; *AB*: Appearance-based; *FB*; Feature-based; *VO*: Visual odometry

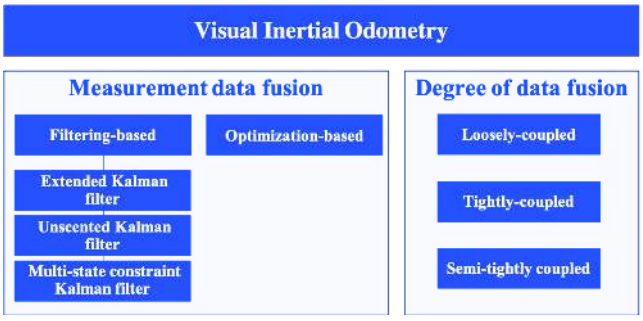

FIGURE 9. General Classification of VIO Techniques Proposed in Literature.

1) Loosely-coupled VIO

A loosely coupled approach for VIO system processes considers the visual and inertial information as independent entities, so each unit estimates the vehicle pose as two pose estimator modules. Then, with the consideration of the vehicle motion constraints, the estimated poses from VO and IMU modules are fused and processed to refine the vehicle ego-motion estimation in the delayed fusion stage. In other words, each pose estimator (VO and IMU units), is processed to estimate the vehicle position and orientation as independent frameworks. Therefore, one of the main drawbacks is the information loss, in terms of accuracy, which may be encountered during the fusion process of decoupled pose estimation. On the other hand, this approach is simple, computationally efficient, and easily for expanding and integrating it with other sensor modalities. The most common sensory data fusion technique is by utilizing the Kalman filter (KF). In addition, nonlinear optimization methods can be used to couple sensory data for better, accurate, and robust vehicle pose estimation [25], however, at the cost of additional computational load, making them impractical for resource-constrained platforms (i.e., UAV [153]). The general pipeline of loosely coupled VIO is illustrated in Fig. 10.

Solutions based on loosely coupled VIO methods can be classified into two categories based on how the data are processed for prediction and observation in the fusion/filter stage. In the first category, the IMU measurements are used for state estimation in the kinematics model, whereas the estimations of VO units are used to update the KF (as observation data), such as [153], [171]–[173]. As the IMU measurements comprise linear acceleration data, this approach would provide high-rate accurate linear velocity estimation, making it suitable for robotic ego-motion estimation that maneuvers at variable speeds. The main problem of this approach lies in the model's sensitivity to IMU data's biases and drifts. Therefore, the cross-coupling of these IMU measurements in the kinematics model could cause severe accumulation errors in the pose estimation results when integrated with VO

TABLE 2. Review of state-of-the-art VIO approaches.

Ref	Author(s) Year	Type of visual sensor						Key design choice		Measurement data fusion		Degree of data fusion			Comments
		Monocular	Stereo	RGBD	Fisheye	Thermal	Event-based	Appearance-based	Feature-based	Filtering-based	Optimization-based	Loosely-coupled	Semi-tightly coupled	Tightly-coupled	
[143]	Mourikis & Roumeliotis 2007	✓							✓	✓				✓	• The reprojection error of 3D landmarks is used as the measurement updates in an IMU-driven filtering framework. • The computational load is linear to the number of features being processed • Extensive number of experiments within an urban environment have been conducted to validate the model.
[144]	Bloesch et al. 2015		✓					✓		✓				✓	• No initialization step was required. • The proposed framework could operate in real time on a UAV to accurately estimate the camera pose when 50 features per frame were observed and integrated in the filter state. • The performance significantly deteriorated at features below 20 features per frame.
[145]	Leutenegger et al. 2015		✓						✓		✓			✓	• A Framework that tightly integrates inertial measurements into keyframe-based visual SLAM were proposed. • Real-time operation was demonstrated due to employ keyframes marginalization to remove old states and maintain a bounded-sized optimization window • Results achieved have experimentally confirmed the benefits of the proposed model over VO and loosely-coupled VIO models.
[146]	Usenko et al. 2016		✓					✓			✓			✓	• Results have evaluated using EuRoC [147] and Malaga [133]. • Model can operate in real-time on a CPU and robust to rapid motion and significant illumination changes. • Results achieved have outperformed VO and loosly-coupled VIO models.
[148]	Forster et al. 2017	✓							✓		✓			✓	• A Framework that integrates the preintegrated IMU model and VIO approach is proposed. • Structureless model is used for visual measurements to avoid optimization over a 3D landmarks. • Lower drift errors in yaw direction than OKVIS [145] and MSCKF [143] were achieved.
[149]	Schwaab et al. 2017		✓					✓		✓				✓	• State estimation is based on fusion of direct VO and inertial measurement using an iterated EKF. • Experimental validation is based on EuRoC MAV dataset [147]. • Results achieved were comparable to the ground truth data.
[150]	Zheng et al. 2017	✓						✓		✓				✓	• Extensive evaluation using dataset of approximately 1.5 hours of localization data. • The use of photometric residuals results in increased pose estimation accuracy with 23% lowering estimation errors.
[151]	Bloesch et al. 2017		✓					✓		✓				✓	• No initialization step was required. • The use of full-state refinement per landmark processes have elevated the model pose prediction's accuracy and robustness. • Results achieved were comparable to other VIO models and were operated in real-time.
[152]	Caruso et al. 2017			✓					✓		✓			✓	• A framework that fuses information from a monocular camera with the Magneto-inertial Dead-reckoning (MIDR) technique was proposed. • An inverse square root filter inspired by the MSCKF [143] was used. • Results achieved have demonstrated higher robustness compared to other VIO, specially at condition where vision is non informative.
[14]	Papachristos et al. 2017					✓			✓	✓				✓	• A framework that fuses information from thermal camera and VIO module using iterative closest point scene registration for localization and mapping • Results achieved have demonstrated the ability of the proposed framework to work robustly under significantly challenging conditions such as visually–degraded real-world.

(continued)

TABLE 2. Review of state-of-the-art VIO approaches (continued)

Ref	Author(s) Year	Type of visual sensor						Key design choice		Measurement data fusion		Degree of data fusion			Comments
		Monocular	Stereo	RGBD	Fisheye	Thermal	Event-based	Appearance-based	Feature-based	Filtering-based	Optimization-based	Loosely-coupled	Semi-tightly coupled	Tightly-coupled	
[153]	HaoChih & Francois 2017		✓						✓	✓		✓			• An indirect error state estimation due to the high dynamic rate of IMU data were employed to avoid information loss using ESKF. • A keyframe concept was adopted which reduces IMU drits increased the system stability and performance.
[154]	Ling et al. 2018		✓						✓		✓		✓		• The vision pose estimator was based on edge alignment and data fusion was based on a sliding window optimization scheme at the back-end block. • Utilization of an efficient IMU preintegration and two-way marginalization scheme was proposed for smooth and accurate pose estimation and appropriate for resource-constrained platforms. • Framework can operate in real-time state estimation for aggressive quadrotor motions.
[155]	He et al. 2018	✓							✓		✓			✓	• A sliding window optimization framework was proposed where the state was optimized by minimizing a cost function which uses the pre-integrated IMU error term alone with the point and line re-projection error. • Experimental validation was based on EuRoC MAV [147] and PennCOSYVIO [156] datasets. • Results achieved have comparable performance predictions to ROVIO [151], OKVIS [145], and VINS-Mono [157].
[157]	Qin et al. 2018	✓							✓		✓			✓	• A novel and robust monocular tightly coupled VIO framework was proposed that incorporates IMU preintegration, estimator initialization, online extrinsic calibration, relocalization, and efficient global optimization. • Experimental results have demonstrated real-time operation using a single camera and an IMU to estimate the vehicle pose. • Experimental validation was based on EuRoC MAV [147] and showed superior performance over OKVIS [145].
[158]	Mur-Artal et al. 2017	✓							✓		✓			✓	• IMU initialization process within a few seconds. • Experimental validation was based on EuRoC MAV [147]. • Remarkable results were achieved by the proposed framework compared to [146].
[159]	Song et al. 2018				✓				✓	✓				✓	• Results achieved high-precision estimation of the pose and velocity of UAV when compared to motion capture system - Opti-track. • Online calibration of the IMU bias and the extrinsic parameters was performed during the robot motion.
[160]	Von et al. 2018	✓						✓			✓			✓	• A dynamic marginalization technique was proposed to adaptively employ marginalization strategies even in cases where certain variables undergo drastic changes. • Experimental validation was based on EuRoC MAV [147] and showed superior performance over ROVIO [144] and DSO [48].
[161]	Khattak et al. 2019					✓		✓			✓			✓	• A framework that fuses a thermal camera with inertial measurements to extend the robotic capabilities to navigate in GNSS-denied and visually degraded environments was proposed. • Results achieved have comparable performance predictions to ROVIO [151], OKVIS [145], and DSO [48].
[162]	Ma et al. 2019		✓						✓		✓			✓	• FAST feature detector and KLT sparse optical algorithm for feature tracking were used which reduce the computational cost. • Experimental validation was based on EuRoC MAV [147]. • Results showed superior performance over OKVIS [145], VINS-MONO [157], and S-MSCKF [163].
[164]	Yang et al. 2019		✓						✓		✓			✓	• Experimental validation was based on EuRoC MAV [147]. • Results showed comparable performance to OKVIS [145][14] and VINS-MONO [157]
[165]	Chen et al. 2019		✓						✓		✓			✓	• Results achieved lower relative pose estimation error compared to ORB-SLAM2 [138] and OKVIS [145]. • Results achieved have outperformed ORB-SLAM2 [138] and OKVIS [145] in terms of root mean square error (RMSE), mean error, and standard deviation (STD).

(continued)

TABLE 2. Review of state-of-the-art VIO approaches (continued)

Ref	Author(s) Year	Type of visual sensor						Key design choice		Measurement data fusion		Degree of data fusion			Comments
		Monocular	Stereo	RGBD	Fisheye	Thermal	Event-based	Appearance-based	Feature-based	Filtering-based	Optimization-based	Loosely-coupled	Semi-tightly coupled	Tightly-coupled	
[166]	Jiang et al. 2020	✓						✓			✓			✓	• Experimental validation was based on EuRoC MAV [147]. • Results achieved have comparable performance predictions to VINS-MONO [157] and ROVIO [151] in terms of both the accuracy and the robustness.
[167]	Zhang 2020		✓						✓	✓				✓	• Experimental validation was based on EuRoC MAV [147] and TUM-VI dataset [168] • Results achieved have comparable and good performance predictions compared to S-MSCKF [163].
[169]	Zhong & Chirarattananon 2020	✓						✓		✓				✓	• Results achieved have comparable and good performance predictions compared to VINS-MONO [157] and ROVIO [151]. • Execution time due to the single plane assumption in the proposed estimator was faster by 15-30 times faster than the two benchmark models, VINS-MONO [157] and ROVIO [151].
[170]	Sun et al. 2021	✓					✓	✓			✓			✓	• A novel state-estimation framework that integrates IMU measurements, a range sensor, and a vision sensor (standard camera or an event camera). • Experimental results showed that the use of event camera in low-light environments provided an advantageous over the standard camera as the sensor does not suffer from motion blur.

observations [174].

In the second category, the VO is used to estimate the model states, while the IMU data is integrated as the observations to update the KF. This approach is able to provide long-term attitude estimations which are accurate, robust, stable, and drift-free. However, such an approach, as opposite to the first category, is mostly not based on IMUs for pose prediction. Thus, linear velocity estimations are not that accurate. Therefore, an orientation filter should be considered while using such an approach. Another drawback of these frameworks is that the pose estimation module is mostly dependant on the VO estimations. Once VO either fails or stops estimating poses, the position and orientation of VIO system would not be available. An example of this approach is proposed by Konolige et al. [175].

To overcome the challenge of the fusion interval mismatch in both loosely coupled VIO estimation methods, Liu et al. [174] proposed an approach to make full use of both camera and IMU information. They proposed the use of separated attitude filter into orientation and position filters to combine the advantages of the first and second category, respectively. This stereo VIO approach has proved to suppress the drawbacks and achieved accurate pose estimation even when low-precise IMU devices were used. Lin and Defay [153] have adopted loosely coupled stereo VIO based on the error-state kalman filter (ESKF). To avoid information loss, an indirect error state estimation due to the high dynamic rate of IMU data were employed. Two state estimates were produced; the first one being a nominal-state estimate where noises were not considered; and the other one being an error-state estimate where accumulated errors were collected. The error

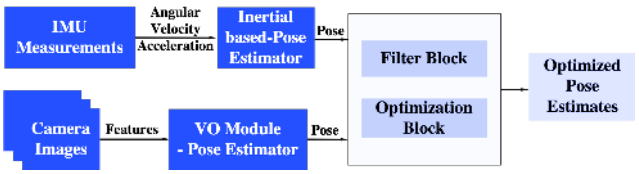

FIGURE 10. General Pipeline of Loosely-coupled VIO.

states could be easily estimated by computing the error states' Jacobian matrix. In addition to that, to reduce the effect of IMU drift, a keyframe concept was adopted which increased the system stability and performance.

The Extended Kalman Filter (EKF) or Unscented Kalman Filter (UKF) techniques are applied to fuse the VO results with inertial measurements to improve VIO performance. These methods are explained in detail in the filtering-based VIO method in Subsection IV-B1. A typical study that employed EKF method to enhance the VIO performance and incorporate system robustness was proposed by [176]. They proposed a loosely coupled indirect feedback Kalman filter integration using the error propagation model and by considering various characteristics of individual sensors. Moreover, Gopaul et al. [177] proposed a loosely coupled VIO using a discrete EKF and took into account processing pairwise time-correlated measurements. Their results were way better than VIO solutions using standard Kalman filter and the Kalman filter with the conventional shaping filter by 18% and 7%, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

2) Semi-tightly coupled VIO

A semi-tightly coupled approach for VIO system processes the visual pose estimation with the IMU sensory data while maintaining a balance between robustness and computational complexity. This approach aids real-time robotic navigation, which is able to cope with big latency between visual image data with the IMU measurements and performs with limited computational resources. An example of semi-tightly coupled VIO approach is presented in [178] for Micro Aerial Vehicle (MAV) platform equipped with a single camera and an IMU. Data fusing was based on EKF and visual pose estimation is based on eight-point algorithm. The framework has demonstrated its capability to estimate the 6-DoF vehicle pose in real-time operation.

Moreover, another semi-tightly coupled VIO framework was developed by [154], to tackle real-time state estimation of aggressive quadrotor motions, as presented in Fig. 11. The vision pose estimator was based on edge alignment and data fusion is based on a sliding window optimization scheme at the back-end block. For smooth and accurate pose estimation, they utilized an efficient IMU preintegration and two-way marginalization scheme, which are appropriate for resource-constrained platforms.

3) Tightly-coupled VIO

A tightly coupled VIO system processes the key information and IMU measurements together with the motion and observation models for vehicle state estimation. With the advancements in computer and software technologies, most VIO studies are focused on employing a tightly coupled framework in their application as shown in Table 2. In tightly coupled approaches, as opposed to loosely coupled methods, all sensor measurements are jointly optimized, thereby producing higher accuracy state estimation. The general pipeline of tightly coupled VIO is illustrated in Fig. 12.

Tightly-coupled approaches can be categorized into two classes, i.e., filtering-based and optimization-based VIO methods, which are to be discussed in the following subsection IV-B. The classical tightly- EKF-based approach and well known in the VIO research area is the multi-state constraint Kalman filter (MSCKF) which were developed in [143]. In this work, multi-geometrical constraints were derived in the measurement model from multi-continuous camera poses, that arose when the same feature was observed in the motion scenes. The computational load of this framework was in the order of one and was a function of the number of detected features in the frames. The experimental results showed that this approach was able to provide high accurate pose estimation using a monocular camera and IMU when performed in real time and in large-scale environments.

In addition, ROVIO [144] is another tightly coupled approach based on EKF using a monocular camera and an IMU. In this work, the pixel intensity errors of image patches were used to formulate the observation equation in EKF. This approach did not require any initialization stage since it utilized the inverse-distance landmark positions which quickly constructed points in the map and started predicting the vehicle pose accurately. This work was later extended in [151] by inherently dealing with the tracked landmarks using iterated-EKF algorithm. Therefore, this tight-fusion approach of visual and IMU data and full-state refinement per landmark processes have elevated the model pose prediction's accuracy and robustness.

Another tightly coupled VIO approach was proposed by [145]. An IMU error term and the landmark reprojection error were integrated in a single nonlinear cost function, thereby marginalizing the previous states and reducing the computation loads. Therefore, the number of states in the sliding window optimization stage has been bounded to ensure real-time system feasibility. Experimental results have demonstrated real-time operation using a stereo camera and an IMU to estimate the vehicle pose. Results obtained were more accurate and robust compared to both vision-based and loosely coupled visual inertial approaches. Later, the same framework was adopted by [157], albeit using a single camera setup. More tightly coupled VIO studies are provided in Table 2.

B. TYPE OF DATA FUSION

Existing VIO studies, especially tightly coupled approaches can be generally categorized based on the type of data fusion into filtering-based and optimization-based paradigms. This section provides a detailed description of each approach and existing solutions based on each approach.

1) Filtering-based VIO

Filtering-based VIO processes data in two stages, i.e., it integrates IMU data to process the state estimation and then updates the state estimation of the vision-based estimator. In addition, filtering-based VIO approaches can be formulated as a maximum a posteriori probability (MAP) estimator [25], where IMU measurements from proprioceptive sensors are used to construct the platform pose prior distribution as the internal state of MAP. In addition, the visuals from exteroceptive sensors are used to compute the platform pose likelihood distribution as the external state of MAP. In other words, the IMU linear acceleration and angular velocities are used to drive the vehicle dynamic model to estimate the vehicle pose. This model is used later to update the vehicle state using the key information obtained from the visual data for ego-motion estimations.

To date, majority of the proposed filter-based solutions can be divided into four frameworks, i.e., algorithms based on Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF), Multi-State Constraint Kalman Filter (MSCKF), and particle filter (PF). Existing solutions based on these frameworks are provided in the following subsections.

- Extended and Unscented Kalman Filters

Autonomous vehicles or robots are considered as examples of nonlinear models. Data associated with nonlinear and dynamic models can be fused using any nonlinear filter such

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

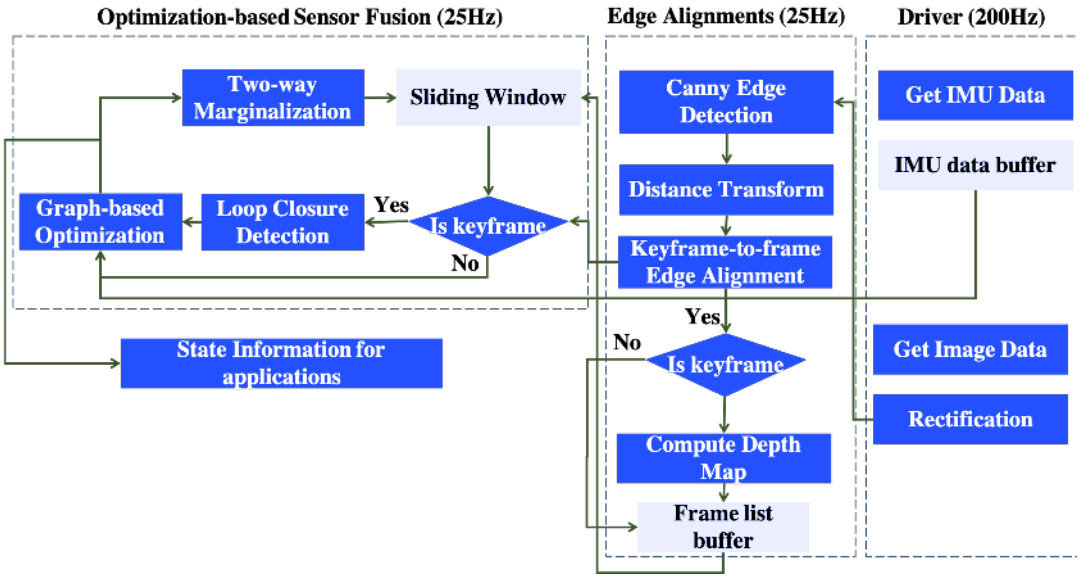

FIGURE 11. The Framework of Semi-tightly Coupled VIO based on Edge Alignments Developed by [154].

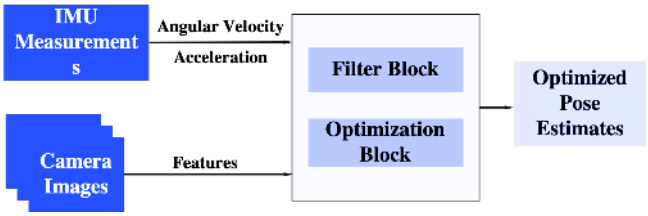

FIGURE 12. General Pipeline of Tightly-coupled VIO.

as the extended Kalman filter (EKF) and particle filter (PF). PFs offer advantages over the EKF as it can deal with Non-Gaussian models for nonlinear systems, however, at the cost of expensive computation [10], [25]. Thus, EKFs have been used as preferable nonlinear filters in the robotics community.

Extended Kalman Filter (EKF) is similar to Kalman filter (KF), however, it can deal with nonlinear models. Basically, it linearizes all nonlinear model parameters at each time step via the first-order Taylor expansion so that a conventional linear KF can be applied. In EKF-based VIO system, the vehicle pose is determined by fusing the propagated state from IMU noisy raw data and the extracted key information obtained from visual data captured from single or multivisual onboarded sensors [179].

Bloesch et al. [144] have proposed an EKF-based VIO, ROVIO, using a monocular camera. In this work, the state was updated (Kalman update) by an innovation term which encompassed the intensity errors of multilevel patch features (linear constraint). For accurate tracking performance, a purely robocentric representation was used, therefore, no initialization step was required, and the effects of system nonlinearities were significantly reduced [180]. Limited number of tracked features were used in the filter state and a heuristic management method was adopted for the decision process regarding the preservation of a particular feature. Hence, features captured within the last frames were evaluated based on a global quality score and were removed if the score fell below a certain threshold. This threshold was based on the number of features captured per frame. The proposed framework could operate in real time on a UAV to accurately estimate the camera pose when 50 features per frame were observed and integrated in the filter state, whereas the performance significantly deteriorated at features below 20 features per frame. However, increasing the number of features would increase the system complexity and computational cost. This work was later improved in [151] by using an iterated EKF framework with fully robocentric representation and by incorporating a photometric error model. A VIO approach based on an iterated EKF (iEKF) framework with a fully robocentric formulation and photometric error model was proposed. In the iEKF framework, a full state refinement was carried out simply by using an iterative process to update the perlandmark, providing simultaneous landmark tracking and considering visual and inertial information.

Furthermore, EKF-based VIO approach was proposed by Zheng et al. [179] using a stereo camera. The framework was based on line feature detectors, in addition to point feature detectors, improving the system robustness under challenging conditions such as low texture environments or under illumination variation. A lightweight filtering approach was adopted to reduce the effect of accumulated drifts for long-term estimation, without the need of a back-end optimization stage. In this work, the closing of the EKF update was formulated by relocating the current sliding window into the past key frames, reducing the accumulated drifts and errors.

Recently, Stremayr and Weiss [181] proposed an EKF-based monocular VIO with a self-calibration property. This

approach used all image information which enabled VIO to be performed at challenging environmental conditions such as in low-textured and low-gradient areas. A higher order covariance propagation (depth) and pixel (intensity) forward-backward propagation was employed, which enabled more precise motion estimation in different light condition environments An alternative and an extension of EKF is the UKF in which a Bayesian filter is used that updates the systems' states via a group of sigma points. The prior distribution is used to derive the weighted sigma points. Furthermore, the mean and covariance contours are computed by using the weighted sigma points using a nonlinear method. In [182], the authors proposed a UKF-based VIO system designed directly on the 3D Special Euclidean Group, SE(3). A matrix Lie group G is a group equipped with a smooth manifold structure such that the group multiplication and inversion are smoothly operated. Processing rotation in the kinematic model is considered as the main nonlinearity contributor. Typically, Euler angles [183] and Quaternions [184] are used to represent the model orientation. In this framework, the kinematics of rotation is modeled on the SE(3) space and by processing the visual and inertial information in the filter, a unique and global 6DoF pose is estimated. Inertial measurements are used to control the inputs, while the visual data are processed to update the state. Detailed analysis of UKF-based on Lie group algorithm is provided in [185]. Furthermore, to improve the performance of pose estimation of UKF in presence of dynamic model errors, many adaptive UKF filtering methods have been addressed in literature [186]–[188]. Once the dynamic model errors are identified, the UKF estimation is corrected.

- Multi-State Constraint Kalman Filter

One of the main drawbacks of EKFs approaches is the requirement of high computational load, which may not be suitable for resource-constrained platforms (i.e., UAV). On the other hand, structure-less approaches such as MSCKF framework are considered a better version in terms of accuracy and consistency because they do not rely on strict probabilistic assumptions or delayed linearization [189]. In addition to that the MSCKF [143] framework has complexity which is linear as a function of landmarks due to marginalization of 3D feature points.

Recently, a novel IMU initialization approach has been proposed by [167], which could estimate the model's main parameters within a few seconds. This approach was decoupled with the stereo-based MSCKF framework [163] to deal with system inherited nonlinearities and measurement or observation noises. This noise-adaptive state estimator enhances pose prediction accuracy and overall model robustness. The results provided have outperformed the results of state-of-the-art VIO method of [163].

2) Optimization-based VIO

Optimization-based VIO processes state estimation by solving the least square nonlinear problem over the IMU measurement and the visual data for optimal prediction. Therefore, in the optimization-based, VIO enables state vector linearization of various points for more precise state estimations than the ones provided by filtering-based methods [190]. In such approaches, IMU measurement constraints are calculated by integrating inertial data between two frames. Whereas, in the conventional IMU integration technique, the IMU body state initialization is computed at the initial captured images. Lupton and Sukkarieh [191] proposed an IMU preintegration technique to avoid such duplicated integrations. IMU preintegration module has been widely adopted in optimization-based VIO studies such as [145], [158], [192].

IMU preintegration process was reformulated in Forster et al. [190] by using the rotation group which was computed by a manifold rather than by Euler angles. Furthermore, a continuous preintegration technique was adopted in the optimization-based VIO framework of Liu et al. [193]. Precise localization was achieved by using optimization-based approaches, however, at the cost of extra computational load, which is due to the higher number of landmarks required in the optimization module. Therefore, optimization-based VIO approaches might not be applicable for resource-constrained platforms. To address this issue, solutions have been proposed in the literature that aim at achieving a constant processing time, such as algorithms that marginalize partial past states and measurements to maintain a bounded-sized optimization window [145], [158], [192], [194].

In OKVIS [145], a group of nonsequential old camera poses, new sequential inertial states and measurements were evolved in the nonlinear optimization module for a refined and precise pose estimation. In addition, Qin et al. [157] have proposed an optimization-based VIO approach using a monocular camera incorporating loop closure modules that ran concurrently in multithread mode to ensure reliability and to guarantee real-time operation. Another VIO approach was proposed by [195], however they efficiently utilized loop closures that ran in a single thread, thus it had a linear computational complexity.

Furthermore, Rebecq et al. [196] proposed an event-based VIO algorithm using a nonlinear optimization for pose estimation. The generated asynchronouse events, which have a microsecond resolution, are accumulated into a frame per spatiotemporal windows size. Features are then detected and tracked using FAST corner detector and the Lucas-Kanade tracker, respectively. Then, the 3D matched features are used to triangulate between frames in order to estimate the relative camera motion between frames. The estimated camera poses and 3D landmark positions are periodically refined by minimizing the reprojection error and the inertial measurement error for effective fusion process (visual and IMU measurements). The performance of the model was evaluated on a large scale and an extremely high-speed dataset. This evaluation demonstrated the accuracy and robustness of the model.

The work of Mueggler et al. [197] proposed a continuous-

time framework using event camera to perform VIO. In their framework, a direct integration of the asynchronous events at micro-second resolution and the high rate of IMU measurements. Cubic splines were used to approximate the trajectory of event camera by a smooth curve in the space of rigid body motions. Their model was evaluated on real time using extensive scenes against a ground truth obtained from a motion-capture system with a remarkable accuracy (position and orientation errors are less 1%).

V. LOCALIZATION TECHNIQUES IN LOW-VISIBILITY ENVIRONMENTS

For navigation through visually degraded environments, new vision-based localization techniques have to be explored by expanding the model work capability even beyond the visible band. As opposed to standard visible cameras, infrared cameras are more robust against illumination changes. With the recent advancements in thermal sensors in terms of size, weight, resolution, load, and cost, thermal-inertial odometry (TIO) is now considered a promising technique for autonomous UAV and UGV systems that works in low visibility conditions without relying on GNSS data or any other costly sensor such as LiDARs. The working principle of thermal cameras is capturing the temperature profile in the scene; thus, it can be used in low visibility environments (i.e., low light, night) without the need of any additional source of light.

However, the disadvantages of thermal sensors include providing low textured, featured and image resolution as well as having quite low signal-to-noise ratios [198]. Therefore, in this case, several computer vision algorithms would not be effective and hence, would need further developments. In literature, very limited studies have been proposed related to TIO, such as [14], [199]–[201], and could be further investigated to overcome the limitations of thermal imagers. Moreover, [202] provides an approach employing LiDAR for a better visualization under different degraded environments. For SLAM applications, Shim and Kim [16] proposed a direct thermal-infrared SLAM platform which is optimized and tightly coupled with LiDAR measurements. This multi-modality system was selected to overcome the photometric consistency problem of thermal images due to accumulated sensor noise over time. The first step was to rescale 14-bit raw radiometric data into grey-scale for feature extraction, and the photometric consistency of thermal images was then resolved by tracking the depth information of LiDAR measurements.

For night-time visual systems, researchers have investigated advanced night imaging systems utilizing onboared low light level cameras with the aid of computer vision algorithms and the use of artificial intelligence (AI) based frameworks. These cameras reflect the thermal energy of different objects in the scene into a visible image under a wide range of spectra: visible (0.4–0.7μm) to near-infrared (0.7–1.0μm) light or to long-wave infrared (8–14μm). Such systems, therefore, are illumination dependant and incompatible to fulfill scene understanding at conditions where visibility is insufficient, such as during low light or night missions.

The wide range of spectral band images and the development of rendered visible/thermal fused imagery and enhancement algorithms are able to aid the platform observation tasks such as to operate and recognize several aspects of a scene and detect and localize targets. The objective of providing such fused imagery is to present a more informative content than the individual image (i.e., visible or thermal image), easy and clear to recognize, and robust under visual degraded environments. For example, during low light conditions, color remapping could improve target detection ability by enhancing image contrast and the use of highlighting color [203] which lead to a faster scene recognition [204]. Various image fusion methods have been investigated in the literature, such as integrating visual and near-infrared context information [205]–[209] and enhance image contrast [210]. Recently, AI-based approaches have been investigated for color mapping of gray-scale thermal image [211]–[213] to enhance image intensity and retain high level scene information, thus leading for a better visualization.

Recently, two localization techniques in low visibility environments were proposed by Mandischer et al. [214], a novel radar-based SLAM and another radar-based localization strategy employing laser maps. These approaches are evaluated in indoor environments with heavy dust formulation to emulate vision scenarios of the grinding process. In the first approach, scan-to-map technique was developed based on probabilistic iterative correspondence (pIC) SLAM. While in the second approach, they utilized environmental information prior to the grinding process. This data set is used to generate a laser map that aids the localization process of radar-based SLAM. They provided a strategy to improve localization using laser maps with line fitting on Radar-based SLAM.

The performance of VO is negatively affected in challenging illumination conditions and high dynamic range (HDR) environments due to brightness inconsistency. Therefore, Gomez-Ojeda et al. [215] have proposed a learning-based method to enhance the image representation of the sequences for VO. They have adopted long short-term memory (LSTM) layers to maintain temporal consistency of the image sequences, thanks to LSTM internal memory cell. The trained network has been implemented in two state-of-the-art algorithms of VO methods (ORB-SLAM [134] and DSO [48]) and tested in challenging environments. Pose estimation results using the enhanced image representation were compared to the VO results using normal image sequences and proved the network's benefits that enhance localization especially in challenging conditions.

Alismail et al. [216] have proposed the use of binary feature descriptors in the direct VO framework, which enhances visual state estimation and increases system robustness under illumination variation. This approach is invariant to monotonic changes of the image intensity. In addition,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

Park et al. [217] have performed a systematic evaluation of the performance of various direct image alignment methods in terms of accuracy and robustness under significant illumination changes. Kim et al. [218] have proposed a stereo VO algorithm which employs affine illumination model in each image patch to cope with abrupt illumination variation in direct state estimation model. The proposed approach has demonstrated a real-time system capability to accurately localize the aerial robot while maneuvering under significant illumination changes. In addition, a multi-sensor fusion pose estimation technique based on a factor graph framework was proposed by [219] to navigate in visually degraded environments. Four different sensors were used including IMU, stereo camera with 2 LED lights, active infrared (IR) camera, and 2D LiDAR which have been employed on a UGV and tested in totally dark environments.

The high dynamic range property of dynamic vision sensor could leverage visual localization models to operate in challenging illumination conditions, such as low light room. Hence, Vidal et al. [220] have proposed a hybrid framework that fuses event data, visual data from standard images, and IMU measurements for a more accurate and robust pose estimation. The model was integrated with a resource-constrained platform (quadrotor UAV) and evaluated extensively with different flight scenarios such as hovering mode, flying in fast circles, and different lighting conditions. Their model outperformed the pose estimation obtained from standard frame based VIO by 85%.

VI. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

Recent researches and technologies have proven the capability of autonomous vehicles to navigate in GNSS–denied and low-visibility environments. Such platforms can feed the end user with useful information that rely on vision-based systems. Based on the application need, an appropriate UAV or UGV navigation system would be adopted. Vision-based localization is one of the promising research directions related to computer vision and deep learning (DL), and aims to estimate the robot's ego-motion within the environments using a set of subsequent measurements. Researchers have investigated novel approaches to enhance vehicle self-localization (position and orientation) accuracy, robustness, reliability, and adaptability while maneuvering.

This survey provides a comprehensive overview of most of the state-of-the-art visual-based localization solutions. These techniques employ visual sensory data and other(s) to localize the robot in GNSS-denied environments and in low-visibility conditions. Two main vision-based navigation paradigms have been reviewed, visual odometry and visual inertial odometry, and discussed in terms of the key design aspects, advantages, and limitations of each paradigm, where applicable.

Key design choices of VO schemes can be classified based on the used visual sensor(s) and the selected processing modules, into geometric and nongeometric approaches. In the first VO method, camera geometrical relations are identified to estimate the ego-motion such as the intensity value of image pixels (appearance-based VO [36], [37], [43], [44], [48]–[50], [128]) and the image texture (feature-based VO [64], [65], [104]–[106], [108], [109], [111], [126], [137]). This method could provide precise state estimation only if enough features within the environment are observed in good lighting conditions. On the other hand, the nongeometric approach, learning-based VO [102], [112], [117], [118], [127], [130], [132], does not require the initialization step for camera parameters and the process of scale correction of the estimated trajectory such as the case of monocular VO. VO scheme could be a good candidate for precise localization in GNSS-denied and textured environments at good illumination conditions. Table 1 provides a summary of the recent literature in the VO field highlighting key design choices and evaluation criteria.

Inertial-based odometry approaches use the high rate of IMU data (linear acceleration and angular velocity) to estimate the vehicle pose. This approach is unreliable for long-term state estimation as the IMU data are corrupted with noise over time. Hence, solutions based on VIO are proposed to overcome the limitation of visual odometry and inertial odometry techniques. VIO techniques are classified based on the processing stage where sensor fusion (visual data + IMU) occurs into loosely coupled [153], [176], [177] and tightly coupled models [158]–[162], [166], [167], [169], [170], [199]. The loosely coupled VIO processes the visual and inertial information independently and each module will estimate camera pose. Then, at a delayed stage, the poses estimated from IMU and VO state estimators are fused to produce a refined pose. Such an approach is simple and easy to be integrated with other sensor modality frameworks. However, in terms of pose estimation accuracy and robustness, its lower than tightly coupled VIO techniques where all sensor measurements (visual + IMU) are jointly processed and optimized for pose estimation.

The VIO can be further classified based on type of data fusion into filtering-based [153], [159], [162], [167], [181] and optimization-based [145], [158], [190]–[193] solutions. In general, performing state estimation using Filtering-based VIO is processed in two stages, (i) estimate the vehicle pose using the IMU linear acceleration and angular velocities that drive the vehicle dynamic model and (ii) update the vehicle pose using the key information of the visual data that estimated the vehicle ego-motion. Existing filtering-based VIO solutions use the nonlinear filter framework (Kalman filter) where errors are linearized producing accurate pose estimation. These solutions can be categorized based on the filtering frameworks into EKF [144], [151], [179], [181], UKF [182] and MSCKF [143], [163], [167].

The utilization of the EKF approach can be used to linearize the data associated with the nonlinear and dynamic model parameters, thus providing good pose estimation. On the other hand, this comes at the cost of computational power which increases quadratically relative to the number of fea-

tures tracked per frame. Moreover, to deal with highly nonlinear models, an approach based on an UKF are proposed, which is an extension of EKF framework, and achieve higher accuracy at the cost of computational load [182]. To deal with the computational load constraints, a MSCKF framework is proposed and provides better accuracy and consistency. In addition to that, the computational load required is linearly proportional to the number of detected landmarks.

In optimization-based VIO framework, pose estimation is processed by solving the least square nonlinear problem for the IMU and visual information. Such approaches outperform the pose prediction obtained from filtering-based VIO due to their capability to linearize the state vector of various points, producing more precise predictions at the cost of extra computational loads. To tackle this issue and make the approach suitable to be deployed in resource constraints platforms, i.e., drone solutions have been proposed to utilize constant processing time via a framework based on a bounded-sized optimization window and marginalize past states [145], [158], [192], [194]. In other words, few states are updated via the nonlinear optimization solver, which reduce the computation load and make it more feasible for real-time operation. Table 2 highlights the design choices of the latest studies in the VIO field.

State-of-the-art solutions for self-localization in low-visibility environments can be categorized as follows: single modality or multi-modality frameworks. In a single modality, the state estimation was performed based on data obtained from a single sensor i.e., stereo-based VO [218], thermal-based VO [106], or event-based [85]. Thermal sensors provide images at low resolution with low textured and features. To address this issue, many image fusion techniques with the visible image have been proposed based on computer vision algorithm [205]–[209] or machine learning tools [211]–[213]. This is to retain high level scene information for better visualization and scene understanding.

To enhance robustness to difficult illumination conditions and high dynamic range (HDR) environments, enhanced VO frameworks are proposed in the literature, such as by using binary descriptors [216], affine illumination model [218], and learning-based methods to enhance image representation [215]. Moreover, a multi-modality framework has proposed to cope with difficulties in percepting the environment around the vehicle at low visibility. In such frameworks, the robot's pose was estimated by using a multi- sensory data fusion technique, i.e., thermal imager with IMU [14], [201], event-based camera with IMU [196], [197], [221], thermal imager and LiDAR measurements [16], Radar and LiDAR [214], and more than two sensory data [199], [200], [219].

Based on the reviewed research studies, various research components have been considered when developing visual-based localization approaches, such as: sensor modality, type of environment, type of platform (ground or aerial vehicle) and available computation resources, and the dimension of pose estimation (2D or 3D). Performing visual localization can be processed in three main modules: preprocessing, state estimation process, and postprocessing modules. The use of these processes will affect the performance, prediction accuracy, and system power and energy efficiency. Therefore, based on the application needs, a suitable localization approach should be investigated or researched for optimal performance. Environmental texture properties and lighting conditions affect the main source of perception to self-localize the robot within the workplace (scene understanding). The main evaluation metrics considered in the literature are performance, accuracy, power and energy efficiency, and system robustness. The different parts of performing visual localization are summarized in Fig. 13.

According to the literature reviewed, the following challenges hinder the progression of effective self-localization systems.

1) **Robustness:** In the presence of illumination variation such as lighting or weather conditions, VO and VIO approaches based on standard camera are poorly performing localization due to lack of features detected within the environment. There are post processing techniques available to reduce the effect of outliers and enhance the performance, however, at the cost of additional computational load. To that end, post-processing vision-based state estimators by deploying deep learning (DL) approaches may significantly improve the pose results, and hence, results in a more robust visual localization system, such as in [222]. DL approaches have the ability to adapt with inherited system nonlinearities as well as the variation in the environment. In addition to that, very limited studies have employed thermal camera for odometry estimation and overcoming its low feature resolution [14], [199]–[201].
2) **Applicability:** Some platforms are limited with the power and computational capabilities. Therefore, real-time operating sensory data approaches should utilize fully learning-based or hybridized learning and conventional-based paradigms. Such systems are considered as application dependent models, however, with the advances in machine learning tools, their capability can be extended over time via fine tuning or transfer learning techniques.
3) **Reliability:** Real time operation requires the system to have the above mentioned criteria, robustness, and applicability. Based on the reviewed visual-based localization approaches, they are application dependent models. Online-based state estimators require the robot to have self-awareness ability regarding the surrounded environment and based on the situation, the best suitable odometry technique should be operated. Having this intelligent decision (ID) platform wherein based on the condition, best suited approach is operated for robot' ego-motion estimation. Having such ID platform would improve state estimation performance, increase system adaptability, reliability and robustness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3082778, IEEE Access

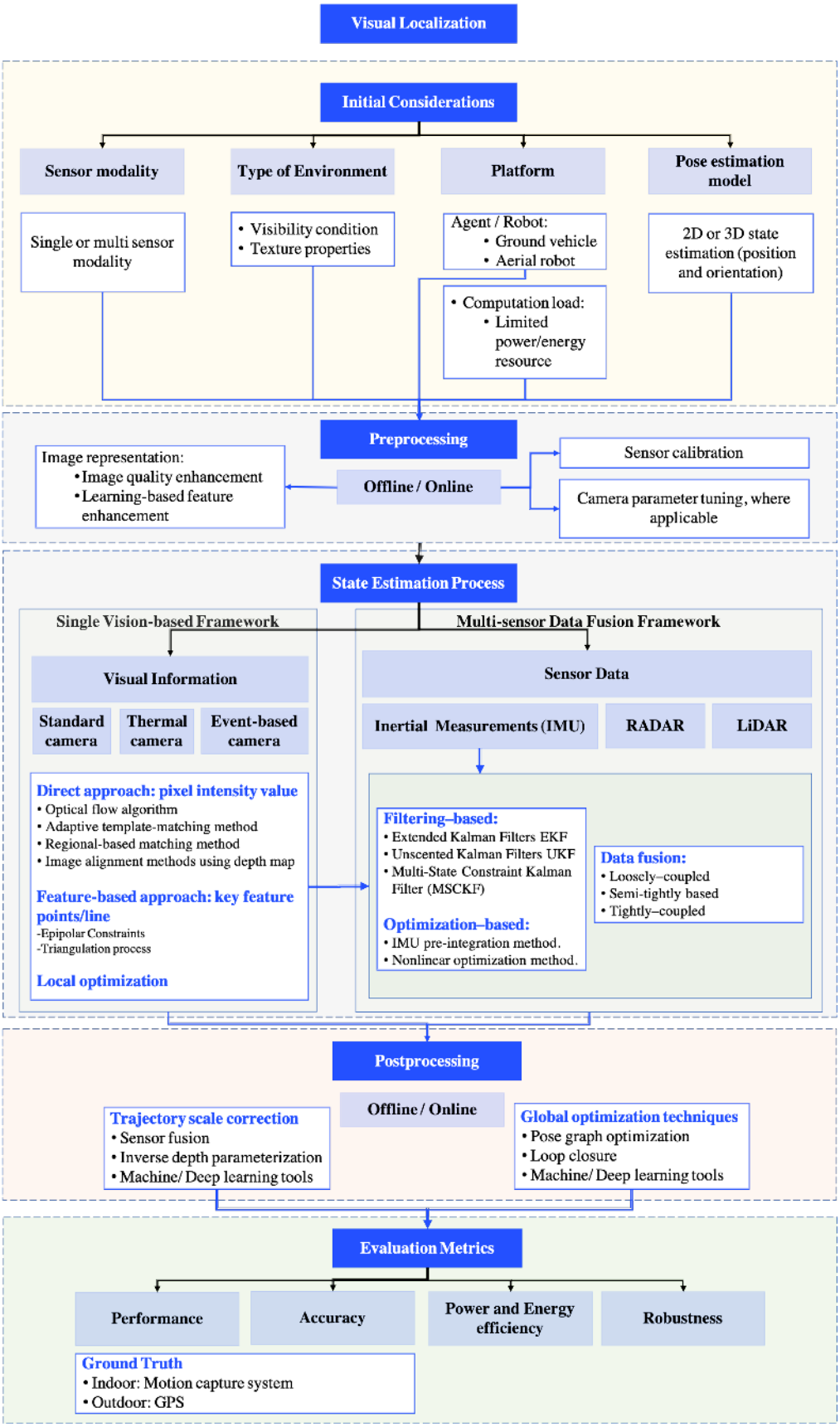

FIGURE 13. **The Main Components of the Visual-localization.**

4) **Adaptability:** Persistence pose estimation to adapt with the changes in the environmental texture properties, features, and illumination conditions over time. Enhance image representation at the visual localization preprocessing module using deep learning approaches would play a role for better perception and extraction of meaningful information, such as the work proposed by [215].
Dynamic vision sensor (DVS) has the capability, compared to visible cameras, providing sufficient data under low light conditions at low latency, high dynamic range, no motion blur [223]. Event based observation and navigation approaches have been widely proposed in literature [223], however, their key potential to be utilized for real time application under low light conditions has not yet largely been investigated.
For low visibility conditions, with the mentioned capability of event-based camera, fusion techniques of thermal and event-based sensory data with an IMU might lead to more robust self-localization scheme. Such schemes could be investigated with the integration of DL tools to learn and capture the inherited nonlinearities error patterns associated with overall visual odometery estimation. This would enhance the end-to-end framework or preprocessing the raw data and hence increase system efficiency and reliability when navigating at reduced visibility conditions.
The generated event data from event-based sensors are noisy depending on the illumination condition and very sensitive to camera parameters. In low light conditions, the features or edges of moving objects, even when tuning the camera parameters to their optimal values, are highly scattered and very noisy. Therefore, the need for an approach that could reject these noises and sharpen the real event data is essential for a better extraction of meaningful information under normal, low light, and/or variation of lighting conditions. Yet, event denoising methods based on conventional spatio-temporal correlation or learning approaches are still largely unexplored [224]–[230].

VII. CONCLUSION

In this article, we have surveyed most of the state-of-the-art studies related to visual-based localization solutions, namely VO and VIO, to aid autonomous navigation in GNSS-denied environments. In addition, we have conducted a comprehensive review on self-localization techniques for autonomous navigation in visually degraded environments. The main components of performing visual localization were identified and discussed.

Studies related to VO have been classified based on key design choices into conventional approaches (appearance, feature and hybrid-based methods) and non-conventional approaches (learning-based methods). An overview of the key design aspects of each category was provided, and the challenges associated with each approach were highlighted, where applicable. In addition, VIO-related studies have been categorized based on the type of the sensory data that were fused and the stage at which this fusion takes place. VIO techniques can be categorized into filtering or optimization-based paradigms, which include loosely, semi-tightly, and tightly coupled approaches. Key design characteristics, strengths, and weaknesses of each type were discussed. For lighting conditions challenges, pose estimation is processed by frameworks that enhance image representation and feature extraction modules. Furthermore, data fusion of multi-sensors is also examined to cope with difficulties in percepting the environment such as thermal imagers, event-based camera, IMU, LiDAR, and RADAR measurements.

Advances in computer vision algorithms, machine learning tools, and both software and hardware technologies should be directed towards developing an efficient self-localization system. Such systems should have an environment-awareness capability, be resilient to outliers, adapt to environmental challenges, and provide reliable, robust, and accurate estimations in real-time. Based on the surveyed papers, the main future self-localization direction includes pose estimation in GNSS-denied, complex, and visually degraded environments. The main future research trends in this topic are robustness, applicability, reliability, and adaptability.

REFERENCES

[1] G. Balamurugan, J. Valarmathi, and V. P. Naidu, "Survey on UAV Navigation in GPS Denied Environments," International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings, pp. 198–204, 2017.
[2] A. Bircher et al., "Structural Inspection Path Planning Via Iterative Viewpoint Resampling With Application to Aerial Robotics," in 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 6423–6430.
[3] C. Papachristos, S. Khattak, and K. Alexis, "Uncertainty-aware Receding Horizon Exploration and Mapping Using Aerial Robots," Proceedings - IEEE International Conference on Robotics and Automation, pp. 4568–4575, 2017.
[4] D. Zermas et al., "Automation Solutions for the Evaluation of Plant Health in Corn Fields," IEEE International Conference on Intelligent Robots and Systems, vol. 2015-Decem, pp. 6521–6527, 2015.
[5] D. Zermas et al., "Estimating the Leaf Area Index of Crops Through the Evaluation of 3D Models," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 6155–6162.
[6] J. G. Mooney and E. N. Johnson, "Integrated Data Management for a Fleet of Search-and-rescue Robots," Journal of Field Robotics, vol. 33, no. 1, pp. 1–17, 2014. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/rob.21514/abstract
[7] C. Papachristos, D. Tzoumanikas, and A. Tzes, "Aerial Robotic Tracking of a Generalized Mobile Target Employing Visual and Spatio-temporal Dynamic Subject Perception," IEEE International Conference on Intelligent Robots and Systems, vol. 2015-Decem, pp. 4319–4324, 2015.
[8] W. He, Z. Li, and C. L. P. Chen, "A Survey of Human-centered Intelligent Robots: Issues and Challenges," IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 602–609, 2017.
[9] S. Poddar, R. Kottath, and V. Karar, "Motion Estimation Made Easy: Evolution and Trends in Visual Odometry," in Recent Advances in Computer Vision. Springer, 2019, pp. 305–331.
[10] S. A. Mohamed et al., "A Survey on Odometry for Autonomous Navigation Systems," IEEE Access, vol. 7, pp. 97 466–97 486, 2019.
[11] Y. D. V. Yasuda, L. E. G. Martins, and F. A. M. Cappabianco, "Autonomous Visual Navigation for Mobile Robots: A Systematic Literature Review," ACM Comput. Surv., vol. 53, no. 1, Feb. 2020. [Online]. Available: https://doi-org.libconnect.ku.ac.ae/10.1145/3368961

[12] D. Scaramuzza and F. Fraundorfer, "Visual odometry Part I: The First 30 Years and Fundamentals," IEEE Robotics and Automation Magazine, vol. 18, no. 4, pp. 80–92, 2011.

[13] W. Rone and P. Ben-Tzvi, "Mapping, Localization and Motion Planning in Mobile Multi-robotic Systems," Robotica, vol. 31, no. 1, pp. 1–23, 2013.

[14] C. Papachristos, S. Khattak, and K. Alexis, "Autonomous Exploration of Visually-degraded Environments Using Aerial Robots," 2017 International Conference on Unmanned Aircraft Systems, ICUAS 2017, pp. 775–780, 2017.

[15] A. Djuricic and B. Jutzi, "Supporting Uavs in Low Visibility Conditions By Multiple-Pulse Laser Scanning Devices," ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-1/W1, no. May, pp. 93–98, 2013.

[16] Y. Shin and A. Kim, "Sparse Depth Enhanced Direct Thermal-Infrared SLAM Beyond the Visible Spectrum," IEEE Robotics and Automation Letters, vol. 4, no. 3, pp. 2918–2925, 2019.

[17] A. Saha, A. Kumar, and A. K. Sahu, "FPV Drone with GPS Used for Surveillance in Remote Areas," Proceedings - 2017 3rd IEEE International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2017, vol. 2017-Decem, pp. 62–67, 2017.

[18] H. N. Viet et al., "Implementation of GPS Signal Simulation for Drone Security Using Matlab/Simulink," in 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 2017, pp. 1–4.

[19] M. O. Aqel et al., "Review of Visual Odometry: Types, Approaches, Challenges, and Applications," SpringerPlus, vol. 5, no. 1, 2016.

[20] D. Liu et al., "A Low-Cost Method of Improving the GNSS/SINS Integrated Navigation System Using Multiple Receivers," Electronics, vol. 9, no. 7, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/7/1079

[21] K. Yousif, A. Bab-Hadiashar, and R. Hoseinnezhad, "An Overview to Visual Odometry and Visual SLAM: Applications to Mobile Robotics," Intelligent Industrial Systems, vol. 1, no. 4, pp. 289–311, 2015.

[22] R. Azzam et al., "Feature-based Visual Simultaneous Localization and Mapping: A Survey," SN Applied Sciences, vol. 2, no. 2, pp. 1–24, 2020. [Online]. Available: https://doi.org/10.1007/s42452-020-2001-3

[23] F. Fraundorfer and D. Scaramuzza, "Visual Odometry : Part II: Matching, Robustness, Optimization, and Applications," IEEE Robotics Automation Magazine, vol. 19, no. 2, pp. 78–90, 2012.

[24] G. Huang, "Visual-Inertial Navigation: A Concise Review," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 9572–9582.

[25] J. Gui et al., "A Review of Visual Inertial Odometry from Filtering and Optimisation Perspectives," Advanced Robotics, vol. 29, no. 20, pp. 1289–1301, 2015.

[26] M. Shan et al., "A Brief Survey of Visual Odometry for Micro Aerial Vehicles," IECON Proceedings (Industrial Electronics Conference), pp. 6049–6054, 2016.

[27] D. Scaramuzza and F. Fraundorfer, "Visual Odometry [Tutorial]," IEEE robotics & automation magazine, vol. 18, no. 4, pp. 80–92, 2011.

[28] H. P. Morevec, "Towards Automatic Visual Obstacle Avoidance," in Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2, ser. IJCAI'77. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1977, p. 584.

[29] C. G. Harris and J. Pike, "3D Positional Integration from Image Sequences," Image and Vision Computing, vol. 6, no. 2, pp. 87–90, 1987.

[30] J. F. Hoelscher et al., "Bundle Adjustment —A Modern Synthesis Bill," Conference Record of the IEEE Photovoltaic Specialists Conference, vol. 34099, pp. 943–946, 2000.

[31] R. Munguia and A. Grau, "Monocular SLAM for Visual Odometry," Parallax, 2007.

[32] L. Kneip, D. Scaramuzza, and R. Siegwart, "A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2969–2976, 2011.

[33] H. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections," pp. 61–62, 1981.

[34] D. Nistér, "An Efficient Solution to the Five-point Relative Pose Problem," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 6, pp. 756–770, 2004.

[35] D. Nistér, O. Naroditsky, and J. Bergen, "Visual Odometry for Ground Vehicle Applications," Journal of Field Robotics, vol. 23, no. 1, pp. 3–20, 2006.

[36] R. Gonzalez et al., "Combined Visual Odometry and Visual Compass for Off-road Mobile Robots Localization," Robotica, vol. 30, no. 6, pp. 865–878, 2012.

[37] D. Scaramuzza and R. Siegwart, "Appearance-guided Monocular Omnidirectional Visual Odometry for Outdoor Ground Vehicles," IEEE Transactions on Robotics, vol. 24, no. 5, pp. 1015–1026, 2008.

[38] T. A. Ciarfuglia et al., "Evaluation of Non-geometric Methods for Visual Odometry," Robotics and Autonomous Systems, vol. 62, no. 12, pp. 1717–1730, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2014.08.001

[39] V. Guizilini and F. Ramos, "Visual Odometry Learning for Unmanned Aerial Vehicles," Proceedings - IEEE International Conference on Robotics and Automation, no. November 2016, pp. 6213–6220, 2011.

[40] L. Frédéric, "The Visual Compass: Performance and Limitations of an Appearance-Based Method," Journal of Field Robotics, vol. 33, no. 1, pp. 1–17, 2006. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/rob.21514/abstract

[41] N. Nourani-Vatani, J. Roberts, and M. V. Srinivasan, "Practical Visual Odometry for Car-like Vehicles," Proceedings - IEEE International Conference on Robotics and Automation, pp. 3551–3557, 2009.

[42] Y. Yu, C. Pradalier, and G. Zong, "Appearance-based Monocular Visual Odometry for Ground Vehicles," IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM, pp. 862–867, 2011.

[43] M. O. Aqel et al., "Adaptive-search Template Matching Technique Based on Vehicle Acceleration for Monocular Visual Odometry System," IEEJ Transactions on Electrical and Electronic Engineering, vol. 11, no. 6, pp. 739–752, 2016.

[44] A. I. Comport, E. Malis, and P. Rives, "Accurate Quadrifocal Tracking for Robust 3D Visual Odometry," Proceedings - IEEE International Conference on Robotics and Automation, no. April, pp. 40–45, 2007.

[45] A. I. Comport, E. Malis, and P. Rives, "Real-time Quadrifocal Visual Odometry," The International Journal of Robotics Research, vol. 29, no. 2-3, pp. 245–266, 2010.

[46] S. Lovegrove, A. J. Davison, and J. Ibañez-Guzmán, "Accurate Visual Odometry from a Rear Parking Camera," IEEE Intelligent Vehicles Symposium, Proceedings, pp. 788–793, 2011.

[47] D. Caruso, J. Engel, and D. Cremers, "Large-scale Direct SLAM for Omnidirectional Cameras," IEEE International Conference on Intelligent Robots and Systems, vol. 2015-Decem, pp. 141–148, 2015.

[48] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 3, pp. 611–625, 2017.

[49] D. Valiente García et al., "Visual Odometry through Appearance- and Feature-Based Method with Omnidirectional Images," Journal of Robotics, vol. 2012, pp. 1–13, 2012.

[50] T. Brox et al., "High Accuracy Optical Flow Estimation based on a Theory for Warping," in European conference on computer vision. Springer, 2004, pp. 25–36.

[51] A. Bruhn and J. Weickert, "Towards Ultimate Motion Estimation: Combining Highest Accuracy with Real-time Performance," Proceedings of the IEEE International Conference on Computer Vision, vol. I, pp. 749–755, 2005.

[52] Y. H. Kim, A. M. Martínez, and A. C. Kak, "Robust Motion Estimation Under Varying Illumination," Image and Vision Computing, vol. 23, no. 4, pp. 365–375, 2005.

[53] M. J. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-smooth Flow Fields," Computer Vision and Image Understanding, vol. 63, no. 1, pp. 75–104, 1996.

[54] E. J. Corey and W.-g. Su, "Relaxing the Brightness Constancy Assumption in Computing Optical Flow," Tetrahedron Letters, vol. 28, no. 44, pp. 5241–5244, 1987.

[55] J. Campbell et al., "A Robust Visual Odometry and Precipice Detection System using Consumer-grade Monocular Vision," Proceedings - IEEE International Conference on Robotics and Automation, vol. 2005, no. April, pp. 3421–3427, 2005.

[56] A. M. Hyslop and J. S. Humbert, "Autonomous Navigation in Three-dimensional Urban Environments using Wide-field Integration of Optic Flow," Journal of Guidance, Control, and Dynamics, vol. 33, no. 1, pp. 147–159, 2010.

[57] V. Grabe, H. H. Bülthoff, and P. R. Giordano, "On-board Velocity Estimation and Closed-loop Control of a Quadrotor UAV Based on Optical

Flow," Proceedings - IEEE International Conference on Robotics and Automation, pp. 491–497, 2012.

[58] V. Grabe, H. H. Bulthoff, and P. Robuffo Giordano, "Robust Optical-flow Based Self-motion Estimation for a Quadrotor UAV," IEEE International Conference on Intelligent Robots and Systems, pp. 2153–2159, 2012.

[59] T. Low and G. Wyeth, "Obstacle Detection Using Optical Flow," Proceedings of the 2005 Australasian Conference on Robotics and Automation, ACRA 2005, 2005.

[60] C. Kerl, J. Sturm, and D. Cremers, "Robust Odometry Estimation for RGB-D Cameras," Revista Gestão, Inovação e Tecnologias, vol. 3, no. 5, pp. 427–436, 2014.

[61] I. Dryanovski, R. G. Valenti, and Jizhong Xiao, "Fast Visual Odometry and Mapping from RGB-D Data," in 2013 IEEE International Conference on Robotics and Automation, 2013, pp. 2305–2310.

[62] S. Li and D. Lee, "Fast Visual Odometry Using Intensity-Assisted Iterative Closest Point," IEEE Robotics and Automation Letters, vol. 1, no. 2, pp. 992–999, 2016.

[63] A. De La Escalera et al., "Stereo Visual Odometry in Urban Environments based on Detecting Ground Features," Robotics and Autonomous Systems, vol. 80, pp. 1–10, 2016.

[64] O. Saurer et al., "Homography Based Egomotion Estimation with a Common Direction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 2, pp. 327–341, 2017.

[65] B. Guan et al., "Visual Odometry Using a Homography Formulation with Decoupled Rotation and Translation Estimation Using Minimal Solutions," in 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 2320–2327.

[66] J. Shi and C. Tomasi, "Good Features," Image (Rochester, N.Y.), pp. 593–600, 1994.

[67] J. Matas et al., "Robust Wide-baseline Stereo from Maximally Stable Extremal Regions," Image and Vision Computing, vol. 22, no. 10 SPEC. ISS., pp. 761–767, 2004.

[68] T. Lindeberg, "Feature Detection with Automatic Scale Selection," International Journal of Computer Vision, vol. 30, no. 2, pp. 79–116, 1998.

[69] D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[70] E. Mair et al., "Adaptive and Generic Corner Detection Based on the Accelerated Segment Test," in European conference on Computer vision. Springer, 2010, pp. 183–196.

[71] M. Calonder et al., "BRIEF: Computing a Local Binary Descriptor Very Fast," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 7, pp. 1281–1298, 2012.

[72] H. Bay et al., "Speeded-Up Robust Features (SURF)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, 2008.

[73] D. G. Lowe, "Object Recognition from Local Scale-invariant Features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 1150–1157 vol.2.

[74] E. Rublee et al., "ORB: An Efficient Alternative to SIFT or SURF," Proceedings of the IEEE International Conference on Computer Vision, pp. 2564–2571, 2011.

[75] R. Y. S. Leutenegger Stefan, Margarita Chli, ""BRISK: Binary Robust Invariant Scalable Keypoints." Computer Vision (ICCV)," Iccv, pp. 2548–2555, 2011.

[76] E. Salahat and M. Qasaimeh, "Recent Advances in Features Extraction and Description Algorithms: A Comprehensive Survey," in 2017 IEEE International Conference on Industrial Technology (ICIT), 2017, pp. 1059–1063.

[77] L. P. Morency and R. Gupta, "Robust Real-time Egomotion from Stereo Images," IEEE International Conference on Image Processing, vol. 2, pp. 719–722, 2003.

[78] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion," Fourth IEEE International Conference on Computer Vision Systems (ICVS'06), pp. 45–45, 2006.

[79] B. Kitt, F. Moosmann, and C. Stiller, "Moving on to Dynamic Environments: Visual Odometry Using Feature Classification," IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings, pp. 5551–5556, 2010.

[80] B. Kitt, A. Geiger, and H. Lategahn, "Visual Odometry Based on Stereo Image Sequences with RANSAC-based Outlier Rejection Scheme," IEEE Intelligent Vehicles Symposium, Proceedings, pp. 486–492, 2010.

[81] I. Cvišić and I. Petrović, "Stereo Odometry Based on Careful Feature Selection and Tracking," 2015 European Conference on Mobile Robots, ECMR 2015 - Proceedings, pp. 0–5, 2015.

[82] L. De-Maeztu et al., "A Temporally Consistent Grid-based Visual Odometry Framework for Multi-core Architectures," Journal of Real-Time Image Processing, vol. 10, no. 4, pp. 759–769, 2015.

[83] H. Badino, A. Yamamoto, and T. Kanade, "Visual Odometry by Multi-frame Feature Integration," Proceedings of the IEEE International Conference on Computer Vision, pp. 222–229, 2013.

[84] I. Krešo and S. Šegvić, "Improving the Egomotion Estimation by Correcting the Calibration Bias," VISAPP 2015 - 10th International Conference on Computer Vision Theory and Applications; VISIGRAPP, Proceedings, vol. 3, pp. 347–356, 2015.

[85] H. Rebecq et al., "EVO: A Geometric Approach to Event-Based 6-DOF Parallel Tracking and Mapping in Real Time," IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 593–600, 2017.

[86] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast Semi-direct Monocular Visual Odometry," in 2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014, pp. 15–22.

[87] H. Silva, A. Bernardino, and E. Silva, "Probabilistic Egomotion for Stereo Visual Odometry," Journal of Intelligent and Robotic Systems: Theory and Applications, vol. 77, no. 2, pp. 265–280, 2014.

[88] J. Feng et al., "A Fusion Algorithm of Visual Odometry Based on Feature-based Method and Direct Method," in 2017 Chinese Automation Congress (CAC). IEEE, 2017, pp. 1854–1859.

[89] H. Alismail et al., "Direct Visual Odometry in Low Light Using Binary Descriptors," IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 444–451, 2017.

[90] R. Roberts et al., "Memory-based Learning for Visual Odometry," Proceedings - IEEE International Conference on Robotics and Automation, pp. 47–52, 2008.

[91] R. Roberts, C. Potthast, and F. Dellaert, "Learning General Optical Flow Subspaces for Egomotion Estimation and Detection of Motion Anomalies," 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, vol. 2009 IEEE, pp. 57–64, 2009.

[92] G. Vitor and R. Fabio, "Learning Visual Odometry for Unmanned Aerial Vehicles," IEEE International Conference on Robotics and Automation, vol. 2011-Janua, pp. 316–320, 2011.

[93] V. Guizilini and F. Ramos, "Semi-parametric Models for Visual Odometry," Proceedings - IEEE International Conference on Robotics and Automation, pp. 3482–3489, 2012.

[94] V. Guizilini and F. Ramos, "Semi-parametric Learning for Visual Odometry," The International Journal of Robotics Research, vol. 32, no. 5, pp. 526–546, 2013. [Online]. Available: https://doi.org/10.1177/0278364912472245

[95] P. Gemeiner, P. Einramhof, and M. Vincze, "Simultaneous Motion and Structure Estimation by Fusion of Inertial and Vision Data," International Journal of Robotics Research, vol. 26, no. 6, pp. 591–605, 2007.

[96] L. Porzi et al., "Visual-inertial Tracking on Android for Augmented Reality Applications," 2012 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems, EESMS 2012 - Proceedings, pp. 35–41, 2012.

[97] K. Konda and R. Memisevic, "Unsupervised Learning of Depth and Motion," arXiv preprint arXiv:1312.3429, 2013.

[98] V. Peretroukhin, L. Clement, and J. Kelly, "Inferring Sun Direction to Improve Visual Odometry: A Deep Learning Approach," The International Journal of Robotics Research, vol. 37, no. 9, pp. 996–1016, 2018.

[99] V. Mohanty et al., "DeepVO: A Deep Learning Approach for Monocular Visual Odometry," arXiv preprint arXiv:1611.06069, 2016.

[100] L. Clement and J. Kelly, "How to Train a CAT: Learning Canonical Appearance Transformations for Direct Visual Localization under Illumination Change," IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 2447–2454, 2018.

[101] J. Jiao et al., "MagicVO: An End-to-End Hybrid CNN and Bi-LSTM Method for Monocular Visual Odometry," IEEE Access, vol. 7, pp. 94 118–94 127, 2019.

[102] Q. Liu et al., "Using Unsupervised Deep Learning Technique for Monocular Visual Odometry," IEEE Access, vol. 7, pp. 18 076–18 088, 2019.

[103] H. Wang et al., "Monocular VO Based on Deep Siamese Convolutional Neural Network," Complexity, vol. 2020, 2020.

[104] A. de la Escalera et al., "Stereo Visual Odometry in Urban Environments Based on Detecting Ground Features," Robotics and

Autonomous Systems, vol. 80, pp. 1 – 10, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889015303183

[105] W. Zhou, H. Fu, and X. An, "A Classification-Based Visual Odometry Approach," in 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 02, 2016, pp. 85–89.

[106] P. V. K. Borges and S. Vidas, "Practical infrared visual odometry," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 8, pp. 2205–2213, 2016.

[107] I. Vanhamel, I. Pratikakis, and H. Sahli, "Multiscale Gradient Watersheds of Color Images," IEEE transactions on Image Processing, vol. 12, no. 6, pp. 617–626, 2003.

[108] B. Kueng et al., "Low-latency Visual Odometry Using Event-based Feature Tracks," in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 16–23.

[109] P. Liu et al., "Direct Visual Odometry for a Fisheye-stereo Camera," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 1746–1752.

[110] L. Heng and B. Choi, "Semi-direct Visual Odometry for a Fisheye-stereo Camera," in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016, pp. 4077–4084.

[111] R. Kottath et al., "Inertia Constrained Visual Odometry for Navigational Applications," in 2017 Fourth International Conference on Image Information Processing (ICIIP), 2017, pp. 1–4.

[112] Y. Almalioglu et al., "GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 5474–5480.

[113] M. Menze and A. Geiger, "Object Scene Flow for Autonomous Vehicles," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061–3070.

[114] T. Zhou et al., "Unsupervised Learning of Depth and Ego-motion from Video," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.

[115] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in CVPR, 2017.

[116] Z. Yin and J. Shi, "GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1983–1992.

[117] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints," in CVPR, 2018.

[118] A. Valada, N. Radwan, and W. Burgard, "Deep Auxiliary Learning for Visual Localization and Odometry," in International Conference on Robotics and Automation (ICRA 2018). IEEE, 2018.

[119] J. Shotton et al., "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2930–2937.

[120] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2938–2946.

[121] S. Wang et al., "DeepVO: Towards End-to-end Visual Odometry with Deep Recurrent Convolutional Neural Networks," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 2043–2050.

[122] I. Melekhov et al., "Relative Camera Pose Estimation Using Convolutional Neural Networks," 2017.

[123] A. Nicolai et al., "Deep Learning for Laser Based Odometry Estimation," in RSS workshop Limits and Potentials of Deep Learning in Robotics, vol. 184, 2016.

[124] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[125] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D Reconstruction in Real-time," in 2011 IEEE Intelligent Vehicles Symposium (IV), 2011, pp. 963–968.

[126] C. Jaramillo et al., "Visual Odometry with a Single-camera Stereo Omnidirectional System," Machine Vision and Applications, vol. 30, pp. 1145 – 1155, 2019.

[127] L. Wang et al., "Estimating Pose of Omnidirectional Camera by Convolutional Neural Network," in 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), 2019, pp. 201–202.

[128] W. Dai et al., "Multi-Spectral Visual Odometry without Explicit Stereo Matching," 2019 International Conference on 3D Vision (3DV), pp. 443–452, 2019.

[129] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," IEEE Transactions on Robotics, vol. 33, pp. 1255–1262, 2017.

[130] S. Li et al., "Self-Supervised Deep Visual Odometry With Online Adaptation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6338–6347, 2020.

[131] H. Zhan et al., "Visual Odometry Revisited: What Should Be Learnt?" 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 4203–4210, 2020.

[132] G. Zhai et al., "PoseConvGRU: A Monocular Approach for Visual Ego-Motion Estimation by Learning," Pattern Recogn., vol. 102, no. C, Jun. 2020. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.107187

[133] J.-L. Blanco-Claraco, F.-Á. Moreno-Dueñas, and J. González-Jiménez, "The Málaga Urban Dataset: High-rate Stereo and LiDAR in a Realistic Urban Scenario," The International Journal of Robotics Research, vol. 33, no. 2, pp. 207–214, 2014.

[134] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147–1163, 2015.

[135] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3D Reconstruction in Real-time," in 2011 IEEE intelligent vehicles symposium (IV). Ieee, 2011, pp. 963–968.

[136] M. R. U. Saputra et al., "Learning Monocular Visual Odometry through Geometry-Aware Curriculum Learning," 2019 International Conference on Robotics and Automation (ICRA), pp. 3549–3555, 2019.

[137] J. Huang et al., "ClusterVO: Clustering Moving Instances and Estimating Visual Odometry for Self and Surroundings," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2165–2174.

[138] K. M. Judd and J. D. Gammell, "The Oxford Multimotion Dataset: Multiple SE(3) Motions With Ground Truth," IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 800–807, 2019.

[139] I. A. Barsan et al., "Robust Dense Mapping for Large-Scale Dynamic Environments," in 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 7510–7517.

[140] P. Li, T. Qin et al., "Stereo Vision-based Semantic 3D Object and Ego-motion Tracking for Autonomous Driving," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 646–661.

[141] J. Huang et al., "ClusterSLAM: A SLAM Backend for Simultaneous Rigid Body Clustering and Motion Estimation," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5874–5883.

[142] J.-C. Piao and S.-D. Kim, "Adaptive Monocular Visual–Inertial SLAM for Real-Time Augmented Reality Applications in Mobile Devices," Sensors, vol. 17, no. 11, p. 2567, 2017.

[143] A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation," in Proceedings 2007 IEEE International Conference on Robotics and Automation, 2007, pp. 3565–3572.

[144] M. Bloesch et al., "Robust Visual Inertial Odometry Using a Direct EKF-based Approach," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 298–304.

[145] S. Leutenegger et al., "Keyframe-based Visual–inertial Odometry Using Nonlinear Optimization," The International Journal of Robotics Research, vol. 34, no. 3, pp. 314–334, 2015. [Online]. Available: https://doi.org/10.1177/0278364914554813

[146] V. Usenko et al., "Direct Visual-inertial Odometry with Stereo Cameras," in 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 1885–1892.

[147] M. Burri et al., "The EuRoC Micro Aerial Vehicle Datasets," The International Journal of Robotics Research, vol. 35, no. 10, pp. 1157–1163, 2016.

[148] C. Forster et al., "On-Manifold Preintegration for Real-Time Visual-Inertial Odometry," IEEE Transactions on Robotics, vol. 33, no. 1, pp. 1–21, 2017.

[149] M. Schwaab et al., "Tightly Coupled Fusion of Direct Stereo Visual Odometry and Inertial Sensor Measurements Using an Iterated Information Filter," in 2017 DGON Inertial Sensors and Systems (ISS), 2017, pp. 1–20.

[150] X. Zheng et al., "Photometric Patch-based Visual-inertial Odometry," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 3264–3271.

[151] M. Bloesch et al., "Iterated Extended Kalman Filter Based Visual-inertial Odometry using Direct Photometric Feedback," The International Journal of Robotics Research, vol. 36, no. 10, pp. 1053–1072, 2017. [Online]. Available: https://doi.org/10.1177/0278364917728574

[152] D. Caruso et al., "A Robust Indoor/Outdoor Navigation Filter Fusing Data from Vision and Magneto-Inertial Measurement Unit," Sensors, vol. 17, no. 12, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/12/2795

[153] L. HaoChih and D. Francois, "Loosely Coupled Stereo Inertial Odometry on Low-cost System," 2017.

[154] Y. Ling, M. Kuse, and S. Shen, "Edge Alignment-Based Visual—Inertial Fusion for Tracking of Aggressive Motions," Auton. Robots, vol. 42, no. 3, p. 513–528, Mar. 2018. [Online]. Available: https://doi.org/10.1007/s10514-017-9642-0

[155] Y. He et al., "PL-VIO: Tightly-coupled Monocular Visual-inertial Odometry using Point and Line Features," Sensors, vol. 18, no. 4, p. 1159, 2018.

[156] B. Pfrommer et al., "PennCOSYVIO: A Challenging Visual Inertial Odometry Benchmark," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 3847–3854.

[157] T. Qin, P. Li, and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," Trans. Rob., vol. 34, no. 4, p. 1004–1020, Aug. 2018. [Online]. Available: https://doi.org/10.1109/TRO.2018.2853729

[158] R. Mur-Artal and J. D. Tardós, "Visual-inertial Monocular SLAM with Map Reuse," IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 796–803, 2017.

[159] J. Song et al., "Tightly Coupled Visual Inertial Odometry based on Artificial Landmarks," in 2018 IEEE International Conference on Information and Automation (ICIA). IEEE, 2018, pp. 63–70.

[160] L. Von Stumberg, V. Usenko, and D. Cremers, "Direct Sparse Visual-inertial Odometry Using Dynamic Marginalization," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 2510–2517.

[161] S. Khattak, C. Papachristos, and K. Alexis, "Keyframe-based Direct Thermal-inertial Odometry," in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 3563–3569.

[162] S. Ma et al., "Robust Stereo Visual-Inertial Odometry Using Nonlinear Optimization," Sensors, vol. 19, no. 17, p. 3747, 2019.

[163] K. Sun et al., "Robust Stereo Visual Inertial Odometry for Fast Autonomous Flight," IEEE Robotics and Automation Letters, vol. 3, pp. 965–972, 2018.

[164] G. Yang et al., "Optimization-based, Simplified Stereo Visual-inertial Odometry with High-accuracy Initialization," IEEE Access, vol. 7, pp. 39 054–39 068, 2019.

[165] C. Chen et al., "A Stereo Visual-inertial SLAM Approach for Indoor Mobile Robots in Unknown Environments Without Occlusions," IEEE Access, vol. 7, pp. 185 408–185 421, 2019.

[166] J. Jiang et al., "DVIO: An Optimization-based Tightly Coupled Direct Visual-Inertial Odometry," IEEE Transactions on Industrial Electronics, 2020.

[167] Z. Zhang et al., "Improving S-MSCKF With Variational Bayesian Adaptive Nonlinear Filter," IEEE Sensors Journal, vol. 20, no. 16, pp. 9437–9448, 2020.

[168] D. Schubert et al., "The TUM VI Benchmark for Evaluating Visual-Inertial Odometry," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1680–1687.

[169] S. Zhong and P. Chirarattananon, "An Efficient Iterated EKF-based Direct Visual-Inertial Odometry for MAVs Using a Single Plane Primitive," IEEE Robotics and Automation Letters, 2020.

[170] S. Sun et al., "Autonomous Quadrotor Flight Despite Rotor Failure With Onboard Vision Sensors: Frames vs. Events," IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 580–587, 2021.

[171] J. P. Tardif et al., "A New Approach to Vision-Aided Inertial Navigation," in Proceedings of (IROS) IEEE/RSJ International Conference on Intelligent Robots and Systems, October 2010, pp. 4161 – 4168.

[172] S. Sirtkaya, B. Seymen, and A. A. Alatan, "Loosely coupled Kalman filtering for fusion of Visual Odometry and inertial navigation," in Proceedings of the 16th International Conference on Information Fusion, 2013, pp. 219–226.

[173] D. Scaramuzza et al., "Vision-Controlled Micro Flying Robots: From System Design to Autonomous Navigation and Mapping in GPS-Denied Environments," IEEE Robotics Automation Magazine, vol. 21, no. 3, pp. 26–40, 2014.

[174] Y. Liu et al., "Stereo Visual-Inertial Odometry With Multiple Kalman Filters Ensemble," IEEE Transactions on Industrial Electronics, vol. 63, no. 10, pp. 6205–6216, 2016.

[175] K. Konolige, M. Agrawal, and J. Sola, "Large-scale Visual Odometry for Rough Terrain," in Robotics research. Springer, 2010, pp. 201–212.

[176] S. Sirtkaya, B. Seymen, and A. A. Alatan, "Loosely coupled Kalman filtering for fusion of Visual Odometry and inertial navigation," in Proceedings of the 16th International Conference on Information Fusion, 2013, pp. 219–226.

[177] N. S. Gopaul, J. Wang, and B. Hu, "Loosely Coupled Visual Odometry Aided Inertial Navigation System Using Discrete Extended Kalman Filter with Pairwise Time Correlated Measurements," in 2017 Forum on Cooperative Positioning and Service (CPGPS), 2017, pp. 283–288.

[178] S. Weiss et al., "Real-time Onboard Visual-inertial State Estimation and Self-calibration of MAVs in Unknown Environments," in 2012 IEEE International Conference on Robotics and Automation, 2012, pp. 957–964.

[179] F. Zheng et al., "Trifo-VIO: Robust and Efficient Stereo Visual Inertial Odometry Using Points and Lines," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 3686–3693.

[180] J. Civera et al., "1-point RANSAC for EKF-based Structure from Motion," in 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, pp. 3498–3504.

[181] A. Hardt-Stremayr and S. Weiss, "Monocular visual-inertial odometry in low-textured environments with smooth gradients: A fully dense direct filtering approach," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 7837–7843.

[182] G. Loianno, M. Watterson, and V. Kumar, "Visual Inertial Odometry for Quadrotors on SE(3)," in 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 1544–1551.

[183] S. Shen et al., "Vision-based state estimation and trajectory control towards high-speed flight with a quadrotor," in Proceedings of Robotics: Science and Systems (RSS '13), June 2013.

[184] J. Kelly and G. S. Sukhatme, "Visual-inertial Simultaneous Localization, Mapping and Sensor-to-Sensor Self-calibration," in 2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation - (CIRA), 2009, pp. 360–368.

[185] M. Brossard, S. Bonnabel, and J. Condomines, "Unscented Kalman Filtering on Lie Groups," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 2485–2491.

[186] G. Hu et al., "A new direct filtering approach to INS/GNSS integration," Aerospace Science and Technology, vol. 77, pp. 755–764, 2018.

[187] ——, "Model Predictive Based Unscented Kalman Filter for Hypersonic Vehicle Navigation with INS/GNSS Integration," IEEE Access, vol. 8, pp. 4814–4823, 2019.

[188] C. Shen et al., "Dual-optimization for a MEMS-INS/GPS System During GPS Outages Based on the Cubature Kalman Filter and Neural Networks," Mechanical Systems and Signal Processing, vol. 133, p. 106222, 2019.

[189] M. Li and A. I. Mourikis, "High-precision, Consistent EKF-based Visual-inertial Odometry," The International Journal of Robotics Research, vol. 32, no. 6, pp. 690–711, 2013.

[190] C. Forster et al., "On-Manifold Preintegration for Real-Time Visual–Inertial Odometry," IEEE Transactions on Robotics, vol. 33, no. 1, pp. 1–21, 2016.

[191] T. Lupton and S. Sukkarieh, "Visual-Inertial-Aided Navigation for High-Dynamic Motion in Built Environments Without Initial Conditions," IEEE Transactions on Robotics, vol. 28, no. 1, pp. 61–76, 2012.

[192] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled Monocular Visual-inertial Fusion for Autonomous Flight of Rotorcraft MAVs," in 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015, pp. 5303–5310.

[193] Y. Liu et al., "Monocular Visual-Inertial SLAM: Continuous Preintegration and Reliable Initialization," Sensors, vol. 17, no. 11, 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/11/2613

[194] Z. Yang and S. Shen, "Monocular Visual–Inertial State Estimation With Online Initialization and Camera–IMU Extrinsic Calibration," IEEE Transactions on Automation Science and Engineering, vol. 14, no. 1, pp. 39–51, 2017.

[195] P. Geneva, K. Eckenhoff, and G. Huang, "A linear-complexity EKF for visual-inertial navigation with loop closures," in 2019 International

Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 3535–3541.

[196] H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Real-time Visual-inertial Odometry for Event Cameras Using Keyframe-based Nonlinear Optimization," 2017.

[197] E. Mueggler et al., "Continuous-time Visual-inertial Odometry for Event Cameras," IEEE Transactions on Robotics, vol. 34, no. 6, pp. 1425–1440, 2018.

[198] T. Mouats et al., "Thermal Stereo Odometry for UAVs," IEEE Sensors Journal, vol. 15, no. 11, pp. 6335–6347, 2015.

[199] K. Alexis, "Resilient Autonomous Exploration and Mapping of Underground Mines using Aerial Robots," in 2019 19th International Conference on Advanced Robotics (ICAR), 2019, pp. 1–8.

[200] C. Papachristos et al., "Autonomous Navigation and Mapping in Underground Mines Using Aerial Robots," in 2019 IEEE Aerospace Conference, 2019, pp. 1–8.

[201] J. Delaune et al., "Thermal-Inertial Odometry for Autonomous Flight Throughout the Night," in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, pp. 1122–1128.

[202] A. Djuricic and B. Jutzi, "Supporting Uavs in Low Visibility Conditions by Multiple-Pulse Laser Scanning Devices," The international archives of photogrammetry, remote sensing and spatial information sciences, vol. 40, no. 1W1, pp. 93–98, 2013.

[203] M. A. Hogervorst and A. Toet, "Evaluation of a Color Fused Dual-band NVG," in 2009 12th International Conference on Information Fusion, 2009, pp. 1432–1438.

[204] A. Toet et al., "Perceptual Evaluation of Color Transformed Multispectral Imagery," Optical Engineering, vol. 53, no. 4, pp. 1–13, 2014. [Online]. Available: https://doi.org/10.1117/1.OE.53.4.043101

[205] D. P. Bavirisetti and R. Dhuli, "Two-scale Image Fusion of Visible and Infrared Images Using Saliency Detection," Infrared Physics & Technology, vol. 76, pp. 52–64, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350449515300955

[206] D. P. Bavirisetti and R. Dhuli, "Fusion of Infrared and Visible Sensor Images Based on Anisotropic Diffusion and Karhunen-Loeve Transform," IEEE Sensors Journal, vol. 16, no. 1, pp. 203–209, 2016.

[207] H. Li et al., "Infrared and Visible Image Fusion Scheme Based on NSCT and Low-level Visual Features," Infrared Physics & Technology, vol. 76, pp. 174–184, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350449515301146

[208] J. Ma et al., "Infrared and Visible Image Fusion via Gradient Transfer and Total Variation Minimization," Information Fusion, vol. 31, pp. 100–109, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S156625351630001X

[209] Z. Zhou et al., "Perceptual Fusion of Infrared and Visible Images Through a Hybrid Multi-scale Decomposition with Gaussian and Bilateral Filters," Information Fusion, vol. 30, pp. 15–26, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253515001013

[210] T. Shibata, M. Tanaka, and M. Okutomi, "Versatile Visible and Near-infrared Image Fusion Based on High Visibility Area Selection," Journal of Electronic Imaging, vol. 25, no. 1, pp. 1–16, 2016. [Online]. Available: https://doi.org/10.1117/1.JEI.25.1.013016

[211] N. Bhat et al., "Generating Visible Spectrum Images from Thermal Infrared using Conditional Generative Adversarial Networks," in 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 1390–1394.

[212] J. Ma et al., "FusionGAN: A Generative Adversarial Network for Infrared and Visible Image Fusion," Information Fusion, vol. 48, pp. 11–26, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253518301143

[213] H. Li and X. Wu, "DenseFuse: A Fusion Approach to Infrared and Visible Images," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2614–2623, 2019.

[214] N. Mandischer et al., "Bots2ReC: Radar Localization in Low Visibility Indoor Environments," in 2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), 2019, pp. 158–163.

[215] R. Gomez-Ojeda et al., "Learning-based Image Enhancement for Visual Odometry in Challenging HDR Environments," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 805–811.

[216] H. Alismail et al., "Direct Visual Odometry in Low Light Using Binary Descriptors," IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 444–451, 2017.

[217] S. Park, T. Schöps, and M. Pollefeys, "Illumination Change Robustness in Direct Visual SLAM," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 4523–4530.

[218] P. Kim, H. Lee, and H. J. Kim, "Autonomous Flight with Robust Visual Odometry Under Dynamic Lighting Conditions," Autonomous Robots, vol. 43, no. 6, pp. 1605–1622, 2019.

[219] M. Sizintsev et al., "Multi-Sensor Fusion for Motion Estimation in Visually-Degraded Environments," in 2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), 2019, pp. 7–14.

[220] A. R. Vidal et al., "Ultimate SLAM? Combining Events, Images, and IMU for Robust Visual SLAM in HDR and High-speed Scenarios," IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 994–1001, 2018.

[221] A. Zihao Zhu, N. Atanasov, and K. Daniilidis, "Event-based Visual Inertial Odometry," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5391–5399.

[222] R. Azzam et al., "A Stacked LSTM-Based Approach for Reducing Semantic Pose Estimation Error," IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1–14, 2021.

[223] G. Gallego et al., "Event-based Vision: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2020.

[224] S. Guo et al., "A Noise Filter for Dynamic Vision Sensors using Self-adjusting Threshold," 2020.

[225] D. Czech and G. Orchard, "Evaluating Noise Filtering for Event-based Asynchronous Change Detection Image Sensors," Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics, vol. 2016-July, pp. 19–24, 2016.

[226] Y. Feng et al., "Event Density Based Denoising Method for Dynamic Vision Sensor," Applied Sciences (Switzerland), vol. 10, no. 6, 2020.

[227] A. Khodamoradi and R. Kastner, "O(N)-Space Spatiotemporal Filter for Reducing Noise in Neuromorphic Vision Sensors," IEEE Transactions on Emerging Topics in Computing, vol. XX, no. X, pp. 1–8, 2017.

[228] J. Wu et al., "Probabilistic Undirected Graph Based Denoising Method for Dynamic Vision Sensor," IEEE Transactions on Multimedia, vol. 9210, no. c, pp. 1–13, 2020.

[229] S. Guo et al., "SeqXFilter: A Memory-efficient Denoising Filter for Dynamic Vision Sensors," 2020. [Online]. Available: http://arxiv.org/abs/2006.01687

[230] R. W. Baldwin et al., "Event Probability Mask (EPM) and Event Denoising Convolutional Neural Network (EDnCNN) for Neuromorphic Cameras," pp. 1698–1707, 2020.

YUSRA ALKENDI received the M.Sc. degree in mechanical engineering from Khalifa University, Abu Dhabi, United Arab Emirates, in 2019, where she is currently pursuing the Ph.D. degree in aerospace engineering with a focus on robotics with the Khalifa University Center for Autonomous Robotics Systems (KUCARS). Her current research is focused on the application of artificial intelligence (AI) in the fields of dynamic vision for perception and navigation.

LAKMAL SENEVIRATNE received B.Sc.(Eng.) and Ph.D. degrees in Mechanical Engineering from King's College London (KCL), London,U.K. He is currently a Professor in Mechanical Engineering and the Director of the Robotic Institute at Khalifa University. He is also an Emeritus Professor at King's College London. His research interests are focused on robotics and autonomous systems. He has published over 300 refereed research papers related to these topics.

YAHYA ZWEIRI is the School Director of Research & Enterprise, Kingston University London, UK, and he is currently visiting associate professor in Robotics Institute, Khalifa University, UAE. He received his PhD from King's College London in 2003. He was involved in defence & security research projects in the last twenty years at Defence Science and Technology Laboratory (Dstl, Uk), King's College London and King Abdullah II Design and Development Bureau (KADDB),Jordan.

Dr. Zweiri's central research focus is interaction dynamics between unmanned systems and unknown environments by means of deep learning, machine intelligence, constrained optimization and advanced control. He has published over 85 refereed journal and conference papers; and filed five patents in USA and GB in unmanned systems field.