Visual–Inertial Fusion-Based Human Pose Estimation: A Review

Tong Li, *Member, IEEE*, and Haoyong Yu, *Senior Member, IEEE*

***Abstract*— Human pose estimation provides valuable information for biomedical research on human movement and applications, such as entertainment and physical exercise. The fusion of visual and inertial data has been increasingly studied in the past two decades to take advantage of these two naturally complementary sensing modalities. In this article, we systematically reviewed the advances in visual–inertial fusion-based human pose estimation with a thorough search for related studies in five mainstream literature databases. A total of 54 studies were identified and included by screening 4586 records retrieved in the review process. The estimation targets, hardware design, fusion methods, evaluation metrics, and system accuracy of these included studies were summarized and categorized for analysis. From these state-of-the-art studies, challenges in terms of mobility, calibration, real-time estimation, and evaluation methods are further discussed in depth and possible directions to overcome these issues are recommended. We expect that this systematic review can provide researchers and engineers with a thorough idea of the progress and performance in visual–inertial fusion-based human pose estimation. We also hope that the discussions on challenges and possible future directions can facilitate future work to improve such systems and promote their applications in real life.**

***Index Terms*— Computer vision, motion capture, pose estimation, sensor fusion, visual–inertial.**

I. Introduction

HUMAN pose estimation is the process of using sensors and algorithms to identify the position and orientation of body segments and joint angles. The estimated body pose can provide important information for many application scenarios, such as entertainment [1], [2], industry [3], [4], rehabilitation [5], and sports [6]. Therefore, many efforts have been taken to develop advanced sensor systems and computation methods to estimate body pose. Pose estimation methods based on machine vision [7], [8], inertial sensors [9], [10], optical encoders [11], and textile sensors [12] have been proposed to capture the pose of the full body or part of the body. Of these methods, vision- and inertial sensor-based systems have been widely studied and applied.

Vision-based systems are most widely used in human body pose estimation. The optical motion capture (OMC) system that includes multiple high-speed infrared (IR) cameras around a predefined capture volume has been widely used as a gold-standard method for capturing human body pose [13]. Spherical reflective markers are commonly used to reflect IR light, which can be captured by cameras. The spatial trajectories of the markers can be calculated with stereo vision theory when they are captured by at least two cameras. Such systems can achieve sub-millimeter accuracy [14] with a large capture volume and thus are usually applied as the ground truth in benchmarking newly proposed systems. Estimating body pose from multiple ($\geq$2) red–green–blue (RGB) cameras without using markers has been extensively studied [15]. The pose of the human body may be built with generative, discriminative, or hybrid approaches. Fusing a monocular RGB camera with a depth camera can also be used to estimate the body pose and a representative example is the commercial Kinect camera [1]. Using multiple RGB-depth (RGB-D) cameras extends the capture volume and enhances the accuracy when fusing results from multiple cameras [16]. A single RGB camera can also be used to estimate 3-D body pose [17], [18], [19]. However, the problem is indeterminate and the accuracy is usually lower than multiview vision-based methods [20]. Vision-based methods capture abundant information on human body pose as images but suffer from several notable issues, such as blur during fast body motion and occlusion caused by the human body itself or environmental objects. The vision-based systems are rarely wearable [21], [22] as cameras need to be placed off the body surrounding the capture volume to ensure capturing the full body during motion.

Inertial sensors have been proposed as a wearable alternative for body pose estimation and they can overcome the occlusion and blur issues that existed in vision systems [23], [24]. The inertial measurement units (IMUs) usually consist of a triaxial accelerometer, a triaxial gyroscope, and optionally a triaxial magnetometer. With reference to the gravitational vector and the Earth's magnetic field, the pose of the IMU in the global coordinate frame can be determined by fusing measurements from these three sensors inside [25], [26], [27]. IMUs can be attached to body segments to estimate their pose and joint angles between the adjacent segments. Full-body or partial-body pose can be estimated depending on the number of sensor units used. The IMU sensors based on micro-electromechanical system (MEMS) technology require little power and are affordable for applications in daily life. However, several issues also restrain the applications of IMUs. Magnetic or ferromagnetic objects could distort the measurement of the Earth's magnetic field, which is rather weak [28], [29]. The sensor-to-segment transform needs to

Manuscript received 7 January 2023; revised 16 April 2023; accepted 1 June 2023. Date of publication 14 June 2023; date of current version 29 June 2023. This work was supported by the Agency for Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program, with A*STAR Science and Engineering Research Council (SERC) under Grant 192 25 00045. The Associate Editor coordinating the review process was Dr. Xun Chen. *(Corresponding author: Haoyong Yu.)*

The authors are with the Department of Biomedical Engineering, National University of Singapore, Singapore 117583 (e-mail: zjulitong@zju.edu.cn; bieyhy@nus.edu.sg).

Digital Object Identifier 10.1109/TIM.2023.3286000

be calibrated for each use and could be affected by the movements of soft tissues such as muscles underneath the body [30], [31], [32]. Global position information is difficult to obtain from IMUs due to the severe drifting issue in the integration of acceleration.

Visual–inertial fusion is a promising method for accurate and stable pose estimation by integrating the merits of both visual and inertial sensors [33], [34], [35]. The occlusion, low-frequency, and blur issues in the vision system can be complemented by the inertial sensors, while the visual information is free from magnetic distortion and can be used to compute a global driftless position. Thus, visual–inertial fusion has been extensively studied in the fields of mobile robots [36], [37], virtual/augmented reality [38], [39], and human activity monitoring/classification [40], [41]. In terms of human pose estimation, diverse hardware systems and fusion algorithms have been proposed in recent years [42], [43], [44], [45], [46]. These systems enable simpler calibration methods or better performance, making them promising to be widely applied in next-generation systems for entertainment, human–robot interaction, and biomedical research.

A systematic review of the rapid advancement in visual–inertial fusion-based human pose estimation is still lacking. Closely related, some previous studies surveyed pose estimation purely based on computer vision [47] or inertial sensors [48]. Multimodal sensor fusion-based pose estimation for rehabilitation was reviewed in [49], while visual–inertial sensor fusion was only briefly mentioned. Visual–inertial fusion for the navigation of mobile robots [50], [51] or human activities monitoring [52], [53] was also reviewed. However, the complexity of human body structure and accuracy requirements make the human pose estimation problem different from the navigation or activity classification problems. In this work, we systematically review the sensor systems and methods for human body pose estimation based on visual–inertial fusion. We aim to provide a thorough survey of the hardware systems, data processing methods, and fusion algorithms that have been proposed in recent studies. The evaluation metrics and accuracy of the systems are also summarized to discuss the performance of these systems. Challenges and potential directions to improve accuracy are also discussed to provide insights for future studies.

The remainder of this review is organized as follows. Section II presents the search strategies and screening procedures applied in the review process. The detailed results of the review process are listed in Section III. Section IV covers an in-depth discussion of the challenges and possible future directions, followed by the conclusions conducted in Section V.

II. Methods

This review was conducted following the guidelines for Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [54] as shown in Fig. 1. We searched for related studies in five widely used literature databases, including Web of Science, Engineering Village, IEEE Xplore, Scopus, and PubMed in October 2022. The settings and keywords used in the databases are listed in Table I. The keywords for searching were slightly different due to the setting differences across these search engines. The records returned by the search engines were collected and reviewed in the Endnote software (version X9.3.3). Duplicates were first removed manually by checking the article metadata (e.g., title, authors, and publication name). An initial screening process was first conducted and the full texts of the remaining records were retrieved and then assessed for eligibility. The criteria applied in the screening process include the following.

TABLE I
Search Strategy in Databases

Database	Search Field	Keywords
Web of Science	Topic	(camera* OR vision OR visual) **AND** (IMU OR IMUs OR "inertial measurement unit*" OR "inertial sensor*" OR MARG) **AND** ("human pose" OR "body pose" OR "hand pose" OR "pose estimation" OR posture OR gesture OR "motion capture" OR "motion tracking" OR "skeletal tracking")
Scopus	TITLE-ABS-KEY	
PubMed	All Fields	
Engineering Village	All Fields	
IEEE Xplore	All Metadata	

C1: The included paper should present a sensing system that includes both cameras and IMUs. The data from cameras and IMUs should be utilized in the proposed methods rather than for evaluation. Studies without using data from cameras or IMUs were excluded.

C2: The included study should present the pose estimation for the full body or a certain part of the human body. Studies on other topics, such as human activity classifications or pose estimations for nonhuman objects, were excluded.

C3: For studies with similar setups and methods from the same authors or research group, e.g., journal papers extended from conference papers, only one paper was included.

The language of the included studies was restricted to English during both the search (using settings in the search engines if available) and the screening process. Only peer-reviewed publications, including journals and conference papers, were included. To avoid missing studies due to the selection of searching keywords, the references from the retrieved full texts were also checked to include studies that did not appear in the search process.

III. Results

In the review process (see Fig. 1), the search engines returned 4586 records in total and 2452 duplicates were removed. The remaining records were screened based on the criteria by checking their titles and abstracts. The full texts of 132 records were sought for retrieval. By assessing the full texts, 84 records were excluded, and six papers were added from the references. Therefore, a total of 54 studies were included in the final review. The included studies presented diverse characteristics in the aspects of body segments, hardware design, fusion methods, and performances. We reviewed

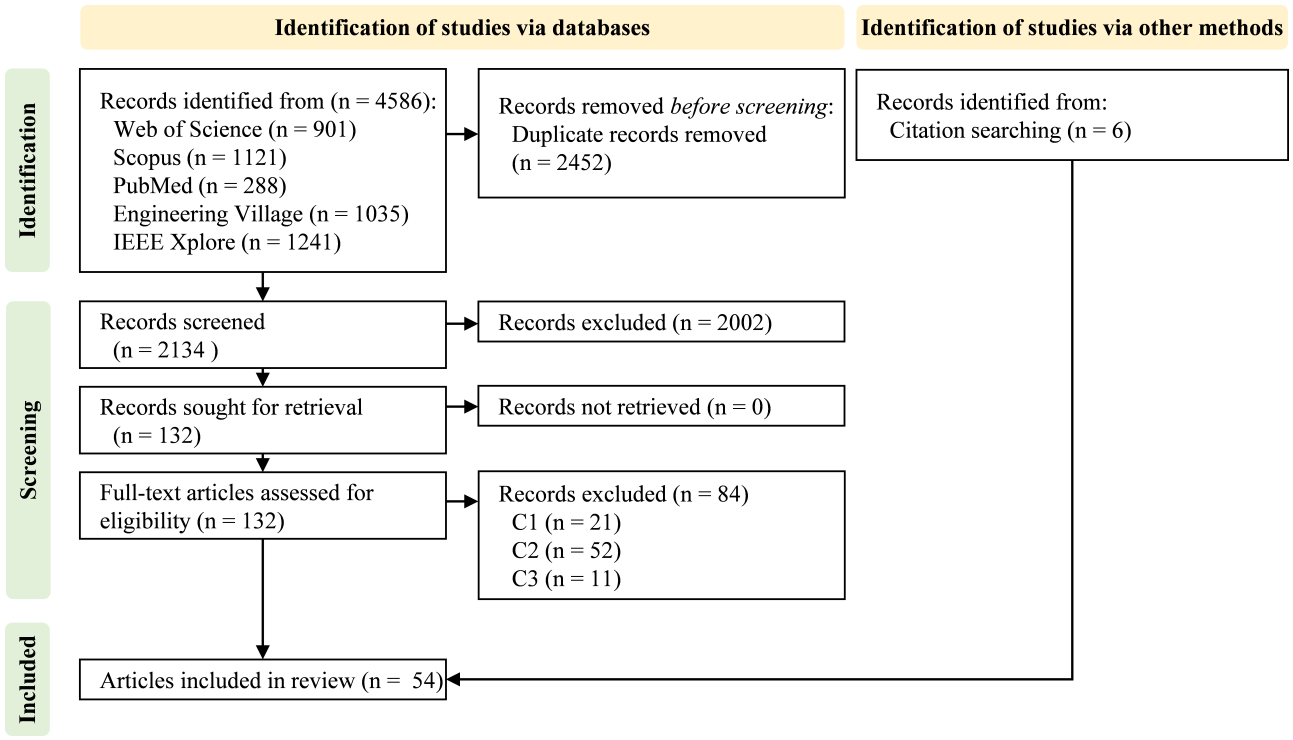

Fig. 1. Review process following guidelines of the PRISMA.

the studies in these aspects in detail and summarized key parameters in Table II where the studies were listed chronologically.

A. Targets for Estimation

For pose estimation, a model of the human body is usually needed to describe the geometry of the body. Since the human body is highly complex with hundreds of joints and degrees of freedom (DoFs), diverse models have been used to simplify the human body to different extents. There are two major types of human models commonly used in these reviewed studies, i.e., skeletal models and volumetric models [104]. The skeletal models only considered the skeletal structure of the human body and the bones are usually described as sticks connected by joints. On the other hand, the volumetric models also include the shape of humans and model the human body surface as meshes often with thousands of vertices. Within the included studies, most of them adopted the skeletal model, while only ten studies utilized the volumetric models [35], [58], [73], [76], [77], [84], [87], [92], [94], [95]. The skinned multiperson linear (SMPL) model has been proposed as an open-source generalized human body model in [105] and it has been applied in seven studies [73], [76], [77], [84], [92], [94], [95]. The default model includes 23 joints and 6890 vertices and some variations of the model have been used, such as in [92] (13 joints) and [73] (48 DoFs, 1000 vertices). During the pose estimation, the shape parameters can be estimated from the sensor fusion [73] while other studies either use parameters from manual tuning [58] or body scanning [76] for pose estimation.

Another major difference between the included studies lies in the targeted body part for capture. The target is an important factor as it is usually predefined based on the application scenarios and it also influences the capture volume and setup of sensor systems. Sensor systems targeting different body parts may not work interchangeably or would achieve unsatisfying performances (e.g., accuracy). The included studies mainly focused on pose estimation of four types of body parts (see Table II), including the full body, upper body, lower body, and hand.

The majority of the studies (24, 44.4%) aimed to estimate full-body pose. Volumetric models have been applied for full-body pose estimation in six studies [76], [77], [84], [92], [94], [95], while other studies used skeletal models with DoFs ranging from 10 to 20. Upper body pose has been estimated in 14 studies (25.9%) and it is usually modeled as four skeletal segments connected by three joints (shoulder, elbow, and wrist). Four of these studies considered both arms [55], [73], [82], [89], while the other studies only considered a single side of the arm. The lower body was estimated in 12 studies and most of them focused on both sides of leg joints [45], [67], [68], [78], [90], [97], [98]. Five studies focused on hand pose estimation [42], [72], [83], [87], [93]. The hand was considered as a single body in [72], [83], and [93] and digital joints were only considered in [42] and [87].

B. Hardware

1) Inertial Sensors: The IMUs can be generally divided into two types depending on whether they have magnetometers (sometimes termed M-IMU or MARG sensors) or not. Without using magnetometers, the systems do not suffer from

TABLE II
PARAMETERS OF THE INCLUDED STUDIES

#	Year	Ref.	Body[1]	Vision[2]	# of IMUs	Location of IMUs	Mag.[3]	Fusion method	Performance/Accuracy
1	2008	[43]	Upper	1 video camera (red patch on the IMU)	1	Left wrist	Yes	A: EKF, B: PF	Wrist joint position A: 0.4–10.56 cm B: 0.68–1.77cm
2	2011	[55]	Upper	1 chest-mounted camera (red/green dots on wrists)	5	Chest, upper arms, forearms	No	EKF	Drift-free wrist joint position
3	2011	[56]	A: Lower B: Upper	1 Kinect	A: 2 B: 1	A: right thigh and shank B: one upper arm	No	KF	Satisfactory results for shoulder, knee, and ankle angle
4	2012	[57]	Full	4 consumer cameras	5	Back, forearms, shanks	Yes	Optimization via interacted simulated annealing	Bone orientation error 10.78°±8.5°
5	2013	[58]	Full	1 depth camera, Kinect	6	Head, trunk, forearms, shanks	Yes	Combination of a generative tracker and a discriminative tracker	Joint position error 75 mm
6	2014	[59]	Full	1 Kinect	9	Chest, upper arms, forearms, thighs, shanks	Yes	Skeleton fusion	Elbow and knee angles error 3.81–14.19°
7	2014	[60]	Upper	1 Kinect	2	Left upper arm and forearm	No	KF	Standard deviation of the joint coordinate measurements 2.20 cm
8	2015	[61]	Upper	1 Kinect	2	Right upper arm and forearm	Yes	UKF	Drift-free position information, elbow angle error <20°
9	2016	[62]	Full	OMC cameras (reflective markers)	5	Waist, thighs, shanks	No	A: EKF B: Periodic-EKF	Right knee angle error A: 2.79°, B: 2.07°
10	2016	[63]	Upper	OptiTrack cameras (spherical markers)	3	Right upper arm, forearm, and hand	Yes	Switches the optical/inertial motion capture according to the detection of markers	Can operate when markers are not detected
11	2016	[64]	Full	A: 8 RGB cameras. B: 7 cameras	5	Waist, forearms, shanks. A: collected B: synthetic	Yes	Optimization-based hybrid tracker	A: mean bone orientation error 15.71°. B: mean joint position error 3.8–5.2 cm.
12	2016	[65]	Lower	1 Kinect	2	Right thigh and shank	Yes	EKF	Knee angle error 3.96°
13	2016	[66]	Upper	1 Kinect	2	Right upper arm and forearm	Yes	UKF	Elbow and wrist axial position error 5.0–10.2 cm, upper arm and forearm axial orientation error 0.03–0.43 rad
14	2017	[67]	Lower	1 depth camera, Kinect	7	Waist, thighs, shanks, feet	Yes	KF	Localization error of 0.214–0.226 m
15	2017	[68]	Lower	1 Kinect	7	Waist, thighs, shanks, feet	No	Spatial calibration based on the Fourier series	Simulations demonstrate the practicability and stability
16	2017	[69]	Full	8 HD video cameras	13	Total Capture's IMU location	Yes	LSTM neural network	Average per joint error 70 mm
17	2017	[70]	Full	1 high-resolution monocular camera	6	Bicycle, trunk, forearms, thighs	No	EKF	Joint angle error within 3°.
18	2017	[71]	Upper	1 Kinect	2	Right upper arm and forearm	No	Weighted sum using Euler angles	Elbow angle error 4.1°, elbow and wrist joint position error 2.6 and 2.9 cm
19	2018	[45]	Lower	1 Google Glass camera	3	Google Glass and feet	No	Batch least-square algorithm	Head and feet position error 0.05–0.14 m
20	2018	[72]	Hand	1 camera (1 ArUco marker on the hand)	1	Hand	No	KF	Motion of the hand and fingers can be obtained correctly
21	2018	[73]	Upper	1 Kinect	3	Kinect, wrists	Yes	Optimizing energy function	Joint position error 25–45 mm
22	2018	[74]	Full	1 Kinect	2	Right upper arm and forearm	Yes	Replace arm pose from Kinect with IMU data	Wrist joint position error 5.49–8.04 cm
23	2018	[75]	Lower	1 camera (circular dot patterns on IMUs)	2	Right shank and foot	Yes	Correct IMU reference frame using visual data	Ankle joint angle error 2.8°
24	2018	[76]	Full	A: 1 camera B: A hand-held smartphone camera	A: 13 B: 17 or 9–10	A: Total Capture's IMU location. B: all major bone segments	Yes	Minimize reprojection error with the Levenberg-Marquardt algorithm	A: MPJPE of 26 mm and MPJAE of 12.1°. B: assignment precision of 99.3%
25	2018	[77]	Full	1 Kinect	8	Upper arms, forearms, thighs, shanks	Yes	Joint optimization scheme	A: max error 65.5 mm. B: Average joint tracking error 15.5–43.5 mm
26	2019	[78]	Lower	1 depth-sensing camera	—	—	—	Bayesian regularization artificial neural network	Knee roll/pitch mean square error 38.78/16.08°

TABLE II
(Continued.) PARAMETERS OF THE INCLUDED STUDIES

27	2019	[79]	Full	A: 8 cameras B: 6 video cameras	13	Total Capture's IMU location	Yes	LSTM	A: average per joint per frame error 42.6 mm. B: resolved poses can accurately reflect the image
28	2019	[80]	Upper	1 Microsoft Lifecam camera (colored markers on IMUs and the elbow joint)	2	Left upper arm and forearm	Yes	Kalman filter	Elbow angle error of 2.4°
29	2019	[81]	Full	1 video camera	12	Head, back, waist, chest, upper arms, forearms, thighs, shanks	Yes	Optimization-based IK	Allows for reducing the number of IMUs while keeping accuracy
30	2019	[82]	Upper	2 or 3 Kinect V2	6 or 7	chest, upper arms, forearms, left hand or both hands	Yes	Average angle from two systems	None
31	2019	[83]	Hand	1 Leap Motion camera (6 retro-reflective markers)	2	Camera and hand	Yes	Multi-layer perceptron model	Hand pose error 2–5°
32	2020	[84]	Full	2 Logitech C920 webcams	2	Right upper arm and forearm	Yes	Optimization	Able to track arm joint angles
33	2020	[85]	Full	8 cameras	13	Total Capture's IMU location	Yes	Two-stage fully 3D network	Mean joint position error 28.9 mm
34	2020	[35]	Full	1 camera	13	Total Capture's IMU location	Yes	Full-body pose optimization	Mean position error 13.5 cm, mean orientation error 8.83°
35	2020	[44]	Full	A: 8 cameras B: 7 cameras	13	Total Capture's IMU location	Yes	Pose optimization	A: mean position error 26.1 mm, mean orientation error 7.5°
36	2020	[86]	Lower	4 Kinects	4	Chest, sacrum, right thigh and shank	Yes	Trigonometry	Replicating knee and hip sagittal angles and keeping the output within the prediction bands
37	2020	[87]	Hand	1 Kinect	—	Hand and fingers (synthetic)	—	Gradient descent adaptive particle swarm optimization	Mean distance error 9.34 mm
38	2020	[88]	Full	4 cameras	8	Upper arms, forearms, thighs, shanks. A: collected, B: synthetic	Yes	Orientation regularized network	A: MPJPE 24.6 mm, B: average error over all joints 21.7 mm
39	2020	[89]	Upper	1 Kinect	5	Wrists, elbows and the hip	Yes	Optimization	Accurately capture the pose of the operator
40	2021	[90]	Lower	1 Kinect	4	Thighs and shanks	Yes	UKF	Knee and ankle joint position error 4.87–6.38 cm, step length error 0.03 cm
41	2021	[91]	Full	Front and rear cameras of an iPhone	1	Phone	Yes	Optimization-based IK solver	Joint position error 20.9 cm
42	2021	[92]	Full	3 head-mounted cameras	4	Forearms, shanks,	Yes	Temporal visual-inertial orientation network	Joint position error of 3.17 cm, bone orientation error 8.76°
43	2021	[93]	Hand	1 head-mounted RGB-D camera	3	Camera, right upper arm, forearm	Yes	EKF	Hand position error <1.0 or 5.77 cm (occluded)
44	2021	[94]	Full	1 head-mounted camera	17	—	Yes	Joint optimization	Stable and physically correct pose estimation
45	2021	[95]	Full	10 RGB cameras	15	Head, upper back, sacrum, upper arms, forearms, hands, thighs, shanks, feet	Yes	Optimization	Knee and hip angle error of 3.7° and 2.7°
46	2021	[96]	Full	1 Kinect	15	—	Yes	Use depth information to obtain body parameters and initial pose	Obtained motion capture data accurately reflected the motion acquired by the optical equipment and the motion of the actual subject
47	2021	[42]	Hand	1 head-mounted stereo camera, ZED Mini	7	Dorsum of the hand, metacarpal/proximal phalanges of the thumb, proximal/intermediate phalanges of the index/middle fingers	No	EKF	Keypoint position error 12.68 mm
48	2021	[97]	Lower	2 USBFHD01 cameras (ArUco markers on IMUs)	3	Waist, feet	No	EKF	Lower body joint angle error 3.5°
49	2021	[98]	Lower	1 RGB-D RealSense d435 camera	4	Thighs, shanks	No	Factor graph with a sliding window filter formulation	Smoother 3D joint trajectories

TABLE II
(Continued.) PARAMETERS OF THE INCLUDED STUDIES

50	2021	[99]	Full	A: 1 fixed camera B: 1 moving camera (ArUco markers on IMUs)	A: 3 B: 7	A: right upper arm, forearm, and hand B: waist, thighs, shanks, feet	Yes	Interpolate during gaps	Position and orientation could be measured robustly even if the marker was lost.
51	2022	[100]	Full	1 head-mounted camera	6	Head, pelvis, hands, legs	Yes	BiRNN network with a ConvLSTM	Joint position error 50.51 mm, angle error 11.31°
52	2022	[101]	Upper	1 chest-mounted camera (ArUco markers on the wrist)	3	Chest, right upper arm and forearm	No	EKF	Joint angle error 3.4–10.1°
53	2022	[102]	Upper	1 camera (ArUco markers on IMUs)	3	Left upper arm, forearm, and hand	No	EKF	Joint angle error 2.7°
54	2022	[103]	Lower	1 RGB camera	1	Right foot	No	Replace foot pose from camera with IMU data	Error of peak ankle angle 3.2–11.57°

[1] Targeted body is classified as Full = full body, Upper = upper body or arm, Lower = lower body or leg, and Hand.
[2] The number of cameras indicates those used in this study, not the total number in the corresponding dataset. This also applies to the number of IMUs. The location of the cameras is omitted if they are set in the environment. Contents in the parentheses describe the markers used for vision.
[3] This column indicates whether the IMUs include magnetometers.
—: not available from the paper. A and B: indicate different settings examined in the paper. KF: Kalman filter. PF: particle filter. EKF: extended Kalman filter. UKF: unscented Kalman filter. LSTM: long short-term memory. IK: inverse kinematics. MPJPE: mean per joint position error. MPJAE: mean per joint angular error. Total Capture's IMU location: head, upper back, lower back, upper arms, forearms, thighs, shanks, feet.

magnetic distortion or ferromagnetic interference and thus are suitable for challenging environments such as industrial or indoor spaces [55]. However, the absolute heading angle (relative to the north of the Earth) cannot be estimated and the drifting issue due to the integration process also requires visual information to correct. The majority of the studies (37, 68.5%) used IMUs with magnetometers, while only 15 studies (27.8%) used IMUs without magnetometers and one study used only gyroscopes [70].

The major metrological characteristics of the IMUs include the measurement noise and bias values of individual sensors and they usually need to be calibrated either by the manufacturer or the user [106]. The measurement noise is commonly used to set the covariance matrix in filters [43], [45], [55], [61], [66], [67], [83], and the bias values need to be removed for correction [45], [62], [70], [71], [80], [83], [90]. For studies directly using the orientation outputs from IMUs, the orientation accuracy is sometimes reported [57], [64], [65], [82]. The measurement range of the individual sensor is also reported in [101]. In most studies, the metrological characteristics are not reported, while the brand or chip model of the IMU is directly stated. The most widely used IMUs were from Xsens Technology, including MT9 [43], MTw [44], MTx [58], and MTi series [67]. Some other studies used products from X-IO Technologies [59], InterSense [61], ZMP Inc. [63], Shimmer [65], [86], Noitom [77], Notch Wearable [82], InvenSense [56], Delsys Inc. [103], Wit Inc. [101], Trivisio [55], and Bosch Inc. [93].

2) Cameras: Different types of cameras, including IR cameras, RGB cameras, and depth cameras, have been applied in the reviewed studies. Products, such as Microsoft Kinect and RealSense, include both RGB cameras and depth cameras for easy integration. IR cameras were commonly used in OMC systems where the trajectories of markers are captured via reflected IR lights as used in [62] and [63]. RGB cameras are widely used in real life as webcams, surveillance cameras, and phone cameras. The commercial cameras adopted in the studies include Manta G-145 Allied Vision Technologies [70], PointGrey Grasshopper 3 machine vision cameras [44], LogitechC920 webcams [84], and iPhone cameras [91].

Depth cameras provide range images that are useful for 3-D reconstruction. The Intel RealSense camera includes two IR cameras to calculate depth information based on stereo triangulation and it was applied in [98]. The Leap Motion IR stereo camera is based on a similar theory but with a smaller size and is more suitable for wearable devices [83]. Kinect cameras calculate the depth information based on structured light (Kinect V1) or time of flight (Kinect V2) and have been applied in 19 of the reviewed studies (35.2%). Of these studies, four of them utilized the point cloud depth image [58], [73], [77], [87] and the camera can be replaced by other depth cameras, while others directly used the skeletal information that can be retrieved from Kinect cameras via application programming interfaces (APIs).

An important metrological characteristic related to vision processing is the image resolution and the typical values (in pixels) reported in the reviewed studies include 656 × 368 [44], 640 × 480 [83], [84], [95], 720 × 520 [103], 800 × 600 [64], 1280 × 720 [42], 1392 × 1040 [70], and 1920 × 1080 [95]. The field of view measured in angles is also diverse depending on the cameras [42], [55], [71], [92], [96] and affects the capture volume. In studies using the data extracted from the images, such as the location of a feature in the image plane or the spatial location of the skeletal joint, the measurement error is a source of error for the final body pose estimation [43], [45], [62], [65], [66], [75].

3) Markers: Markers are usually used in company with cameras for easy detection and low computation cost. They are also robust in different backgrounds and lighting conditions. Reflective markers [see Fig. 2(a)] are commonly used in OMC systems to reflect IR light for capture [62], [63]. Markers of solid color are also popularly used for capture with RGB cameras. Tao and Hu [43] used a red patch on an IMU [see

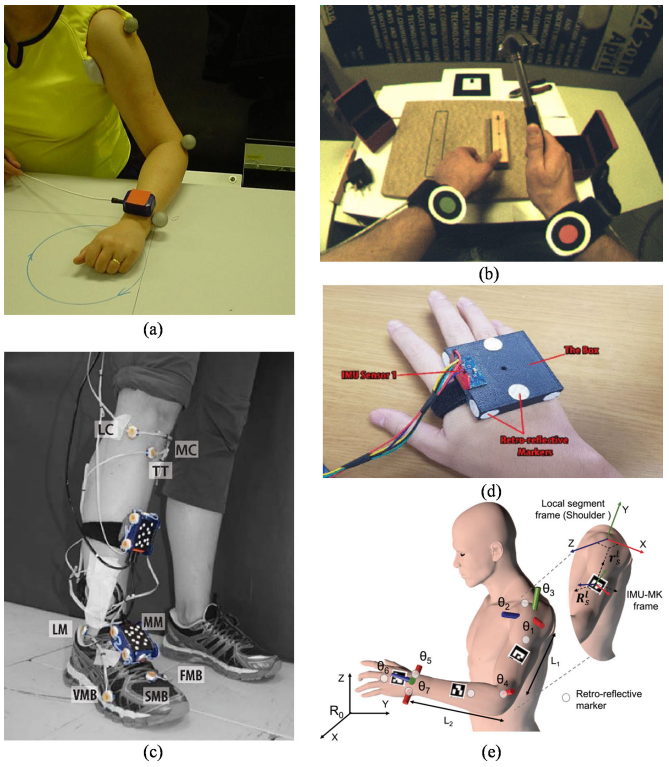

Fig. 2. Marker-based motion capture setups in the reviewed studies. (a) Red color patch on an IMU and reflective markers on the skin [43], (b) color dots on wrists [55], (c) dotted patterns on IMUs [75], (d) IR-light reflective markers on a box [83], and (e) ArUco markers on IMUs [102].

Fig. 2(a)] to detect the motion of the IMU and the wrist joint center. Bleser et al. [55] attached a red and green dot [see Fig. 2(b)] on the wrists to capture the wrist joint motion. Black dot markers were attached to gyroscopes to detect the motion of sensor modules [70]. Circular dot patterns [see Fig. 2(c)] were used in [75] to detect the pose of the sensor module with Vogiatzis and Hernández's algorithm [107]. Multiple circular markers were attached to a box [see Fig. 2(d)] to reflect IR lights, which can be captured by the Leap Motion camera in [83]. Multiple color dots were attached to a glove and they can be detected with a head-mounted stereo camera to compute the position of the hand [42]. The 2-D square ArUco markers [see Fig. 2(e)] were applied in five studies for direct plate pose estimation [72], [97], [99], [101], [102].

C. System Integration

Regarding the sensor location, IMUs are generally attached to certain body segments of interest with straps or tapes. It is usually needed to attach an IMU on each segment to reduce the possibility of indeterminacy when fewer IMUs are used. Cameras are typically set in the environment, while it is also possible to apply cameras worn on the human body for mobile body pose estimation. The sensor systems in the reviewed studies are very diverse since different types, numbers, and locations of IMUs and cameras can be integrated, particularly when targeting different body parts as detailed in the following.

1) Full Body: Full-body pose estimation based on OMC systems usually requires tens of markers attached to the body. Missing, unlabeled, or mislabeled markers commonly occur and require an offline process. To deal with these issues, Joukov et al. [62] proposed to use IMUs (waist, thighs, and ankles) and kinematic models in compensation for marker trajectories to achieve smooth joint angles.

The Kinect camera was combined with 6–15 IMUs over the body for full-body pose estimation using either skeletal pose outputs or purely depth data. Destelle et al. [59] attached nine IMUs on the body (chest, upper arms, forearms, thighs, and shanks) to achieve better accuracy than purely using the Kinect camera. Hwang et al. [96] proposed a system to accurately capture body motion during a golf swing by combining the orientation information from 15 IMUs over the body and joint position from a Kinect camera facing the subject for automatic segment length estimation. Helten et al. [58] only used the depth information from the Kinect camera for fusion with orientation data from six IMUs on the head, waist, forearms, and shanks. A similar setup with different IMU locations (upper arms, forearms, thighs, and shanks) was used by Zheng et al. [77].

For a single RGB camera-based system, the IMUs help to solve the ill-posed problem of reconstructing 3-D poses from 2-D images. Setups with a single RGB camera and multiple IMUs are used in four studies for full-body pose estimation. Twelve IMUs (head, back, waist, chest, upper arms, forearms, thighs, and shanks) were used in [81], and similarly, 13 IMUs (head, sternum, pelvis, upper and lower limbs, and feet) were used in [35] with ground contact constraint taken into consideration. von Marcard et al. [76] proposed methods using a moving (handheld) camera to estimate the pose of a single person equipped with 13/17 IMUs or two persons with 9–10 IMUs each in the outdoor environment. Dot markers were used together with a bike-mounted camera and gyroscopes on both the human body (trunk, forearms, and thighs) and bikes for rider pose estimation [70].

The combination of multiview vision and multiple IMUs on the body has been a popular solution explored for full-body pose estimation [see Fig. 3(a)]. Different number combinations have been used, sucas four cameras + five IMUs (back, forearms, and shanks) in [57], four cameras + eight IMUs (upper arms, forearms, thighs, and shanks) in [88], eight or seven cameras + five IMUs (waist, forearms, and shanks) in [64], four or eight cameras + 13 IMUs (head, upper back, lower back, upper arms, forearms, thighs, shanks, and feet) in [44], [69], and [79], and ten cameras + 15 IMUs (head, upper back, sacrum, upper arms, forearms, hands, thighs, shanks, and feet) in [95].

Body pose estimation using body-worn egocentric cameras was examined in only four studies. Ahuja et al. [91] utilized the front and rear cameras on an iPhone and the built-in IMU to generate the full-body animation of the user. Cha et al. [92] used the three cameras installed on a glass [Fig. 3(b), two cameras facing downward and one camera facing forward] to reconstruct the full-body pose based on an incomplete view of the body. Guzov et al. [94] used only a single head-mounted camera facing forward together with 17 Xsens IMUs on the body to achieve body pose estimation and positioning in a known environment simultaneously. Kim and Lee [100] further

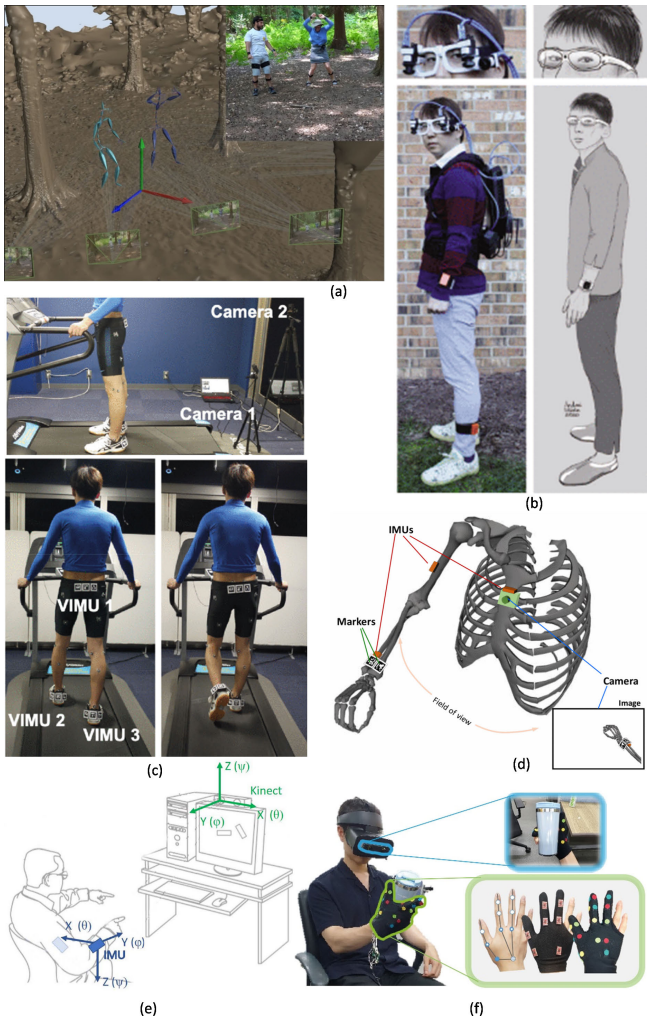

Fig. 3. Visual–inertial sensor systems proposed in the literature. (a) Multiview vision and IMUs on the body [44], (b) head-mounted device with three cameras and IMUs on the wrists and ankles [92], (c) two RGB cameras and three IMU-marker modules on the lower body [97], (d) chest-worn camera and three IMUs on the upper body [101], (e) Kinect camera and two IMUs on the arm [71], and (f) head-mounted stereo camera and markers on the hand [42].

reduced the number of IMUs to six on the head, pelvis, hands, and legs.

2) Lower Body: Lower body pose estimation is essential for gait analysis and localization-related studies. Kinect cameras have been considered as a promising alternative for the low-cost home-based system. Tannous et al. [65] applied two IMUs on the thigh and shank and a Kinect camera for knee angle measurement during home-based rehabilitation. Abbasi et al. [90] combined four IMUs on the thighs and shanks of both sides with a Kinect camera to estimate lower body joint position and step length. For complete lower body measurement, Han et al. [68] integrated seven IMUs on the waist, thighs, shanks, and feet with a Kinect camera facing perpendicularly to the walking direction. Cho et al. [67] used a similar sensor setup to achieve simultaneous motion tracking and localization. To enable a 360° view of measurement, Rodrigues et al. [86] integrated four Kinect cameras together with four IMUs on the chest, sacrum, thigh, and shank for accurate hip and knee angle measurement. Bersamira et al. [78] applied a depth camera (possibly Kinect) and IMUs for estimating gait parameters and knee angles, while details of the setup were not presented. Mitjans et al. [98] used an RGB-D camera (Intel RealSense d435) facing along the walking direction and four IMUs on thighs and shanks to obtain smooth joint trajectories.

A single RGB camera was used to identify special circular dot patterns attached to IMUs for pose estimation [75]. The IMU-marker modules were attached to the shank and foot to compensate for the magnetic field distortion close to the floor. Mallat et al. [97] attached ArUco markers on three IMUs worn on the waist and both heels and estimated the lower body kinematics by fusing marker pose captured by two precalibrated RGB cameras and IMUs [see Fig. 3(c)]. Yamamoto et al. [103] set only one IMU on the foot to estimate foot orientation, which was used to compute ankle joint angle with shank orientation from an RGB camera.

Head-mounted cameras can hardly capture lower body motion and thus were applied to estimate body motion from environmental information. Ahmed and Roumeliotis [45] used a head-mounted camera-IMU module together with two IMUs on the feet for the estimation of lower body posture.

3) Upper Body: Human upper body movements are usually performed in a relatively limited space, and thus, a smaller capture volume is required than full-body or lower body motion capture. However, due to the diversified and dexterous arm motion, markers on the human body in traditional marker-based OMC systems may be easily occluded. Kobayashi et al. [63] proposed a system integrating IMUs on the upper arm, forearm, and hand with an OptiTrack OMC system to ensure continuous arm-pose detection and real-time robot teleoperation even with losing track of markers.

The Kinect camera has been widely used in upper body pose estimation [see Fig. 3(e)]. Kalkbrenner et al. [60] set two IMUs on the upper arm and forearm and a Kinect camera to detect the position of the shoulder, elbow, and wrist joint center for fusion. This setup was also used in [61], [66], [71], and [74], while different calibration and fusion methods were presented. Jatesiktat et al. [73] set two IMUs on the left and right wrists and an additional IMU on the Kinect camera to capture the pose of both arms by fusing the depth and inertial data. To avoid occlusions, Tarabini et al. [82] used two or three Kinect cameras together with six or seven IMUs on the upper body (chest, upper arms, forearms, and hands) to capture the motion of both arms of workers close to machines.

The system with a single RGB camera and several IMUs and markers on the human arm has been proposed in three studies. Tao and Hu [43] used a red patch on the IMU attached to the wrist to estimate the position of the wrist and compensate for the sensor noise and model inaccuracy. Orozco-Soto et al. [80] attached two IMUs on the upper arm and forearm with colored markers on the IMUs. An additional color belt was set at the elbow joint and these colored markers can be captured by the camera to calculate the centroid for further fusion. Mallat et al. [102] used three IMU-ArUco modules attached to the upper arm, forearm, and hand together with a monocular camera to fuse the marker pose with IMU raw

data for driftless arm-pose estimation. Markerless arm motion capture based on RGB cameras was only used in [84], where two precalibrated webcams and two IMUs on the upper arm and forearm were combined.

Chest-mounted cameras have also been proposed to estimate upper body pose as the hand may be captured for certain tasks. Bleser et al. [55] attached five IMUs on the upper body (chest, upper arms, and forearms) and two colored dot markers on the wrist for fusion. Li et al. [101] used ArUco markers on the forearm-mounted IMU instead of the dot marker to obtain the marker orientation and position from a single monocular camera [see Fig. 3(d)]. A calibration technique without using magnetometers was also proposed to remove the influence of magnetic interference.

4) Hand: Cameras are usually set in the environment to capture the hand pose in a predefined area. Chan et al. [72] attached an IMU module with an ArUco marker on it to the back of the hand. A table-mounted camera was used to estimate the marker pose, which was fused with the IMU orientation. Zhang et al. [87] proposed to use of a depth camera (Kinect) to capture the depth range and an IMU-equipped data glove to form a multimodal initializer for rapid hand pose estimation.

Head-mounted cameras have been proposed for hand capture as humans tend to gaze at the hand area during manipulations. Vu and You [83] combined multiple markers and an IMU in a box attached to the hand, whose pose can be estimated with a head-mounted Leap Motion camera. Gosala et al. [93] attached IMUs on the head, upper arm, and forearm and estimated the hand position by fusing IMU data together with the wrist joint position from a head-mounted RGB-D camera. Lee et al. [42] attached seven IMUs on the hand and three fingers together with 37 visual markers of four colors to estimate hand pose [see Fig. 3(f)].

D. Data Extraction

Estimating human pose from visual and inertial data usually takes multiple steps in the data process. Some intermediate data types will be generated by processing raw data and act as inputs for the fusion process. Table III shows the combination of data used for the fusion process.

1) Data From IMUs: Data extracted from the IMUs system can be divided into three types, i.e., raw data, IMU orientation, and body pose. The IMUs typically consist of sensor chips, including accelerometers, gyroscopes, and magnetometers, which can output raw data, including accelerations, angular velocities, and magnetic strength, respectively. These raw data may need to be corrected with some calibration process [102] to reduce or eliminate the error due to drift, noise, and nondiagonal axis issues. Some calibration processes may have been conducted by the vendors, while others, such as drift due to temperature and local magnetic field, need to be conducted onsite by the users or with online algorithms [28].

From the raw data or corrected raw data, the orientation of the IMU in the Earth-fixed global frame (based on the gravity and magnetic field) can be estimated with diverse algorithms [25]. Many products can also directly provide the orientation data as outputs.

TABLE III
DATA EXTRACTION AND FUSION METHODS IN THE REVIEWED STUDIES

Inertial + visual	Deterministic	Filter	Optimization	Learning
Raw + marker 2D position	0	3 [43], [55], [70]	0	0
Raw + marker 3D position	1 [99]	5 [42], [62], [97], [101], [102]	0	0
Raw + skeletal 2D pose	0	0	0	1 [98]
Raw + skeletal 3D pose	0	2 [56], [61]	1 [95]	0
Orientation + marker 3D pose	2 [63], [75]	2 [72], [93]	0	1 [83]
Orientation + silhouettes	0	0	2 [57], [64]	0
Orientation + skeletal 2D pose	1 [86]	0	4 [35], [44], [84], [88]	2 [79], [85]
Orientation + skeletal 3D pose	4 [59], [68], [74], [103]	5 [60], [65-67], [90]	3 [81], [89], [91]	2 [69], [92]
Orientation + point cloud	0	0	3 [58], [73], [77]	0
Orientation + body location	0	0	2 [45], [94]	1 [100]
Orientation + joint angle	1 [71]	0	0	0
Body pose + skeletal 2D pose	0	0	1 [76]	0
Body pose + skeletal 3D pose	2 [82], [96]	1 [80]	0	0
Body pose + point cloud	0	0	1 [87]	0
Unclear	0	0	0	1 [78]

Body pose can be obtained from the IMU orientations based on the skeletal structure [76], [82]. To obtain the anatomical joint angles, the calibration of the transform from the sensor frame to the segment frame (i.e., sensor-to-segment calibration) is required. Since the human body is often modeled as a tree structure, the position of body segments or joints may also be obtained from IMU orientations with forward kinematics. However, the position of the root segment cannot be obtained unless it is assumed to be static. The position data in dynamic motion cannot be obtained purely from the IMUs due to the severe drifting issue as sensor noises are inevitable and will accumulate along with time.

2) Data From Cameras: Depending on the type of cameras and markers adopted, various data types can be obtained from image processing. For marker-based systems, the position of markers, such as colored dots in the 2-D image plane, is direct information that can be obtained by identifying the markers from the images [43], [55], [70]. Marker position in 3-D space can be obtained with stereovision when the same marker can be identified from multiple cameras [42], [62], [63]. The pose

of 2-D marker plates can be computed from either multiple dotted or spherical markers or a special pattern such as the ArUco marker [72].

From images with the human body captured, the silhouettes of the body in the images can be obtained by identification and segmentation [57], [64]. Skeletal pose in 2-D space may also be obtained [35], [98] with methods such as OpenPose [108]. The skeletal pose in 3-D space can be obtained by lifting the 2-D poses to 3-D with multiviews [95]. The Kinect cameras can also directly output the skeletal pose in 3-D space [61], [82]. The silhouettes with depth information may be obtained from depth cameras by processing the point cloud information [58], [77]. As for cameras looking outside of the body, the camera's 3-D position can be calculated with localization techniques in a known environment [45], [94].

E. Fusion Algorithms

Various methods can be applied to fuse the inertial and visual data. On the one hand, different levels of fusion can be achieved according to the types of data obtained from the initial process. On the other hand, different algorithms can be applied to the same type of data [52], [109]. The data and algorithms applied in the fusion process of the reviewed studies are summarized in Table III. The fusion methods in the reviewed studies can be classified into four types as follows.

1) Deterministic Method: A direct fusion method is to compute the body pose from the poses obtained from vision and IMUs independently with simple deterministic methods (11 studies, 20.4%) such as logical or trigonometric methods and weighted sum. Kobayashi et al. [63] calculated the upper body pose from both the IMU system and the OMC system and the final pose was selected based on the error between them. Destelle et al. [59] proposed a hierarchical approach for full-body pose estimation with a Kinect camera and IMUs. The obtained orientation from IMUs was used to correct the body orientation from Kinect systems while maintaining the proximal joint position. Glonek and Wojciechowski [71] estimated the elbow angle from both systems and then computed the final pose as the weighted sum of two angles. Kempfle and Laerhoven [74] replaced the upper arm and forearm orientation from the Kinect camera system with results from IMU systems for the calculation of the wrist joint center. Similarly, Yamamoto et al. [103] used the shank angle from an RGB camera and the foot angle from the IMU system to calculate the ankle angle since the foot angle from the camera is less accurate for capturing at a long distance. Tarabini et al. [82] took the mean angle (a special weighted-sum method) as the final pose from two systems. Trigonometry was used by Rodrigues et al. [86] to compute the joint angles from four Kinect cameras and four IMUs. The segment length was calibrated with the skeletal pose from the Kinect camera to compute the positional information such as foot distance with orientation from IMUs in [96]. Since the visual information may have gaps due to occasional occlusions, the IMU data were used for integrational interpolation during the fusion process in [99]. Lebel et al. [75] used the orientation from a camera to correct the IMU-to-segment transformation.

2) Filter-Based Methods: They are popular for state estimation from multimodal noisy sensor data in real time (18 studies, 33.3%). The popular methods include the Kalman filter [56], [60], [67], [72], [80] and its variants such as the extended Kalman filter [42], [55], [62], [65], [70], [93], [97], [101], [102] and unscented Kalman filter [61], [66], [90] for nonlinear systems. The particle filter is also suitable for nonlinear systems and was used in [43].

3) Optimization-Based Methods: They have been widely used in the studies (17 studies, 31.5%). A cost function with multiple terms is commonly used and many algorithms can be used to solve the optimal body states. Popularly used methods include dynamic programming [88], batch least-squares (BLS) algorithm [45], interacted simulated annealing [57], and Levenberg–Marquardt algorithm [76].

4) Learning-Based Methods: They have become popular for sensor fusion in recent years (eight studies, 14.8%). Trumble et al. [69] applied a long short-term memory (LSTM) network for pose estimation from multiview vision and an additional layer was used to fuse the 3-D pose from both systems. Huang et al. [85] applied a network to fuse orientation from IMUs and the voxel heatmap of joints from eight views. A temporal visual–inertial orientation network was used by Cha et al. [92] to fuse IMUs orientations and 3-D joint positions. Mitjans et al. [98] applied a deep neural network to fuse the raw IMU signals and 2-D skeletal poses from an RGB-D camera. Gilbert et al. [79] fused the poses from both systems computed independently and an additional fully connected layer of neural network was used to fuse the two modalities.

F. Evaluation and Accuracy

Public datasets are useful to evaluate the performance of the proposed sensor systems and fusion algorithms and it would be easier to compare the accuracy between different studies. There are several datasets published for full-body pose estimation, as listed in Table IV. RGB cameras are the most popularly adopted in these datasets for data collection. Three of these datasets are based on multiple RGB cameras fixed in the indoor environment. The sensor combination includes eight cameras + ten IMUs [64], eight cameras + 13 IMUs [69], and 140/31 cameras + 15 IMUs [95]. Four datasets were collected in outdoor environments with one handheld camera + 17 IMUs [76], six cameras + 13 IMUs [110], seven cameras + 13 IMUs [44], or one head-mounted camera + 17 IMUs [94]. However, no ground truth from the OMC system was provided in the outdoor collection. Two datasets used a single depth camera together with two IMUs on the wrists [73] or six IMUs on the body [58]. Both datasets were collected indoors as the depth camera Kinect cannot work well in outdoor environments. Of these datasets, the Total Capture dataset [69] has been reused in five later studies, while other datasets were rarely tested in other studies.

The metrics used for pose estimation in the reviewed studies included both quantitative metrics and qualitative metrics. For qualitative evaluation, the studies usually gave descriptions that the estimated body motion (e.g., joint angles) was satisfactory [56], drift-free [55], smooth [98], or with no

TABLE IV
DATASETS PROPOSED AND USED IN THE REVIEWED STUDIES

#	Dataset Name	Year	Ref.	Camera	IMU	Other Data	Actions	Frames	Applications
1	Helten et al.	2013	[58]	1 Kinect camera	6. Head, trunk, forearms, shanks	Ground truth pose parameters and joint positions from PhaseSpace OMC system	Punching, kicking, rotating on the spot, sidewards and circular walking	6K, 30 Hz	[77]
2	TNT15	2015	[64]	8 RGB cameras	10. chest, waist, shanks, thighs, upper arms, lower arms	3D laser scans and registered meshes of each actor	Walking, running on the spot, rotating arms, jumping and punching.	13K, 50 Hz	—
3	TotalCapture	2017	[69]	8 video cameras	13. Total Capture's IMU location	3D joint positions and angles from Vicon system	ROM, Walking, Acting, Running and Freestyle	1.9M, 60 Hz	[35], [44], [76], [79], [100]
4	TotalCapture Outdoor	2017	[110]	6 video cameras	13. Total Capture's IMU location	—	—	—	[44]
5	NTU Mocap dataset	2018	[73]	1 Kinect camera	3. wrists and Kinect camera	Joint center positions from Vicon system and Plug-in-Gait model	Arm raising, shoulder lifting, torso tilting, air punching, forearm rolling, hair combing, jumping, etc.	39311, 30 Hz	—
6	3D Poses in the Wild	2018	[76]	1 hand-held smartphone camera	17 on a subject or 9–10 for each subject (2 subjects), one on phone	Subjects' SMPL from scanning	Shopping, doing sports, hugging, discussing, capturing selfies, riding bus, playing guitar, relaxing	51000, 30Hz	—
7	Outdoor Duo	2020	[44]	7 RGB cameras	13. Total Capture's IMU location	—	Fast motion, close interaction, occlusion, and props	—	—
8	HPS dataset	2021	[94]	1 head-mounted camera	17. Xsens mo-cap system	—	Exercising, reading, eating, lecturing, using a computer, making coffee, dancing.	300K, —	—
9	CMU Panoptic Dataset 2.0	2021	[95]	140 VGA cameras, 31 HD cameras, 10 Kinect cameras	15. Head, upper back, sacrum, upper arms, forearms, hands, thighs, shanks, feet	—	Hopping, range of motion, stair ascend and descend, ramp ascend and descend, sit-to-stand, walking, jogging, walking in a figure 8, and high-jump activities	86 subjects 6.5 min, 29.97Hz	—

Total Capture's IMU location: head, upper back, lower back, upper arms, forearms, thighs, shanks, feet.

gaps [63], [99]. Some studies presented demos or animations of the constructed body pose [45], [72] and claimed that the methods were effective [84], practical [68], and stable [94]. For quantitative evaluations, the common metrics for evaluations include joint positions (24 studies, 44.4%), body segment orientations (six studies, 11.1%), and joint angles (17 studies, 31.5%). There are also other metrics that are rarely used, such as the silhouette overlap ratio [64], person assignment precision [76], and percentage of correct key points (PCK) [42], [88].

The accuracy reported in the reviewed studies with the major metrics is summarized in Table V based on the body parts and metrics. However, it should be noted that the definition of these metrics may also be different in the review studies. For example, joint position error was mostly computed as the Euclid distance in 3-D space, while a few studies reported the axial position error [43], [60], [66]. Similarly, segment orientation error that was usually considered as the geodesic distance between orientation quaternions was sometimes also reported as the angle error decomposed in Euler angles [66], [91]. It should also be noted that the error could be significantly affected by the number of sensor units and space for capture. Therefore, accuracy may not be a deterministic indicator for performance comparison between different studies regarding the algorithm or hardware.

IV. DISCUSSION

Human pose estimation based on visual–inertial fusion has been advancing rapidly in the recent two decades. Diverse hardware systems and fusion algorithms have been presented in the included studies. Different types of cameras and IMUs have been applied in the sensor systems. The number of cameras varies from 1 to 10 and the number of IMUs ranges from 1 to 17. It is also identified that the systems are designed for capturing the motion of different body parts, including the full body, upper body, lower body, and hand. Different application scenarios, such as a walkway, indoor space, and outdoor environment, are also targeted in these studies. Practitioners applying these systems to future applications need to note these conditions to achieve the expected performance. By reviewing the state-of-the-art research, several challenges that need to be dealt with in future studies are also discovered and discussed in the following.

A. Mobility

Using wearable sensors to achieve long-term monitoring is an important trend in human pose estimation. However, the cameras in most of the reviewed studies (44 studies, 81.5%) are fixed in the environments surrounding the targeted subjects, particularly when estimating human full-body pose and lower body pose. Such setups are also similar to traditional OMC

TABLE V
ACCURACY REPORTED IN THE REVIEWED STUDIES

	Full body	Upper body	Lower body	Hand
Joint position (cm)	7.5 [58] 3.8-5.2 [64] 7.0 [69] 5.49-8.04 [74] 2.6 [76] <6.55/1.55-4.35 [77] 4.26 [79] 2.89 [85] 13.5 [35] 2.61 [44] 2.46/2.17 [88] 20.9 [91] 3.17 [92] 5.051[100]	0.68-1.77 [43] 2.2 [60] 5.0-10.2 [66] 2.6/2.9 [71] 2.5-4.5 [73]	5-14 [45] 4.87-6.38 [90]	0.934 [87] <1.0/5.77 [93] 1.268 [42]
Segment orientation (°)	10.78 [57] 15.71 [64] 8.83 [35] 7.5 [44] 8.76 [92]	1.72-24.64 [66]	—	2-5 [83]
Joint angle (°)	3.81-14.19 [59] 2.07/2.79 [62] <3 [70] 12.1 [76] 3.7/2.7 [95] 11.31[100]	<20 [61] 4.1 [71] 2.4 [80] 3.4-10.1 [101] 2.7 [102]	3.96 [65] 2.8 [75] 8.78/16.08 [78] 3.5 [97] 3.2-11.57 [103]	—

systems where multiple cameras are set in a predefined area. The capture area is fixed and predefined, which restricts the application scenarios of such systems.

For mobile pose estimation, three types of setups have been proposed. The first method is carrying the cameras by a mobile platform, such as a robot or a person [76]. This may be extended to future work with cameras mounted on social robots [111], [112]. Only a single camera has been used and this may limit the capture accuracy as occlusion and blur may easily occur. Multiple mobile vision has not been tested and a major challenge might be the online calibration of the variable camera extrinsic parameters.

The second method that has been used in several studies is setting a camera worn on the body facing outside [45], [94], [100]. The camera motion and localization can be achieved with feature tracking methods [113], hierarchical structure-based localization algorithms [114], or commercial VR products [100]. The IMUs on the body compensate for the position/velocity information for estimating body pose. However, a major drawback is that it can only obtain the pose of a single segment. A previous study [21] applied 16 cameras to estimate body pose by capturing the outside world, but this may not be practical since image processing would be too computationally costly.

The last method is wearing cameras that can capture part of the body such as the hand or wrist [42], [55], [83], [91], [92], [101]. Purely vision-based pose estimation with head-mounted cameras was also designed in [22] and [115]. Combining IMUs on segments can compensate for the partial views of the body and would be a promising solution for the mobile egocentric motion capture method.

Increasing mobility using wearable cameras or cameras on mobile robots would be useful for future applications in real life and deserves more effort. However, wearable cameras also introduce ethical and privacy issues since consent could be difficult to obtain from third parties for vision in real-life usage [116], [117], [118].

B. Calibration

Calibration is a critical procedure for human motion capture, which has a direct and vast impact on the accuracy of the sensor systems. A combination of visual and inertial sensors leads to more sensor units for calibration but also provides the possibility of simplifying or automatizing the calibration process with additional information. The calibration process involves multiple aspects, including sensor intrinsic calibration, sensor-to-sensor calibration, and sensor-to-segment calibration [101]. Calibrating intrinsic parameters of IMUs (scaling factors and misalignment matrices) and cameras (focal length, optical center, and distortion coefficients) have been rather mature [106], [119], [120], and many constant parameters have been provided with the commercial products such as Xsens IMUs and ZED cameras. However, some parameters, such as sensor drifts, need to be estimated and compensated based on static states as in [55] and [66]. Manual calibration is still sometimes needed before use, particularly for magnetometers [121], [122].

Sensor-to-sensor calibration relies more on manual operation and calibration would be needed in each usage since the camera and IMUs are independent, unlike rigidly assembled visual–inertial systems on mobile robots [123], [124]. For example, camera-to-camera transformation needs to be calibrated once set up with marker wands [97] or cubes [64], and repetitions may be needed once bad calibration results or incidental movements occur. Cameras with multiple built-in cameras (e.g., Kinect cameras and ZED cameras) have constant extrinsic parameters, but their field of view is limited. The transformation between cameras and IMUs is required during the fusion process. This could be achieved with rigidly attached IMU–camera modules and the alignment may be calibrated with the magnetometer measurement [93]. A calibration cube with IMU inside can be used for the calibration between the IMUs system and the multicamera system [57]. It should be noted that both methods rely on the accurate calibration of the magnetometers.

The sensor-to-segment alignment is a common and challenging issue in wearable sensors for body pose estimation. The alignment is usually achieved via manual placement [67], [69], [98] or predefined pose [35], [71] such as T-Pose in full-body Xsens IMU systems. Such methods rely on the accuracy of the operation and usually result in low performance. Wand-based manual calibration methods can achieve better accuracy but also requires assistance from the operator [102]. Vision-based 3-D skeletal identification has been extensively studied and the joint position information has been applied for a simpler calibration process [42], [93].

However, this also heavily relies on the accuracy of visual identification. A simple and stable calibration process requires more research efforts to enable the wide application of visual–inertial systems.

C. Real-Time Pose Estimation

Real-time pose estimation is highly desirable in many application scenarios, such as gaming, telerehabilitation, and robot control. Users prefer systems that are plug-and-play, rather than those that require complicated setup processes and professional data processing. In the reviewed studies, only a small portion of them was run in real time [35], [55], [58], [65], [93]. While many studies also indicated the capability of real-time operation [77], [85], [86], [92], several factors beyond the calibration process, such as the sensor synchronization and computation cost, still hinder the real-time application of these systems.

Time synchronization between different sensors is also required for the fusion process since the visual and inertial signals are usually retrieved with different beginning time and at different sample frequencies due to hardware differences. When a hardware trigger is not available, the start time may be determined with predefined cues, such as clapping motion [57], vertical jumps [59], [86], [95], foot stamp [35], [64], and LED flashing [77]. However, these methods could be difficult for online processing and may lead to large synchronization errors. To deal with the difference in sampling rate, offline interpolation or resampling has been used in many studies [44], [66], [68], [71], [90], [100], but it is not suitable for online processing and the sensor noise may also introduce large error for interpolation intervals when using high-order methods such as splines. Unlike the IMU–camera system on mobile robots, hardware synchronization for the wearable system could be challenging when a large number of sensor units are used. Wireless stable synchronization methods might be more suitable and desirable.

The computation cost is usually from the vision process and fusion process. Marker-based methods usually cost less time than markerless methods, but they could be less generalizable. Thus, markerless capture is still the future trend. For data fusion, the optimization and learning-based methods usually cost a large amount of time and result in low frequency, such as 0.067 fps [73], 6.7 fps [88], 5–19.9 fps [91], and 1.9–4.5 fps [98]. Such methods also usually require higher performance hardware such as graphics processing units (GPUs) [44], [77], [81], [85], [88], [92], [100]. Affordable high-performance processors and low computation demanding algorithms are highly desirable, particularly for mobile or outdoor applications.

D. Evaluation and Accuracy

Quantitative evaluation has been made in the majority of the reviewed studies. However, most of them were evaluated with self-collected data and the public datasets have been rarely used in the reviewed studies (see Table IV). Public datasets with both visual and inertial data are still lacking, particularly for capturing part of the body. This may be due to the fact that partial-body capture is less generalizable and the sensor system setup is not unified. Markerless human motion capture is getting more prevalent, while datasets with wearable devices, such as prosthetics and exoskeletons, are lacking. Explorations are desirable on whether these markerless methods can be extended to these diverse scenarios.

Directly selecting and applying the methods for practitioners could be challenging as the datasets and metrics used for evaluation are diverse across studies. The readers should be cautious when comparing the accuracy of studies even using the same metrics. The expression of joint angles is not consistent due to the differences in the models. This could be particularly challenging when applying these results to the medical analysis of human movement based on anatomical structure. The transform is needed when extending the results for analyzing musculoskeletal models [81]. The tasks and movements used for evaluation also need to be standardized, partially in some specific application scenarios, such as rehabilitation, daily living, and gaming.

V. Conclusion

A systematic review of advances in visual–inertial fusion-based human pose estimation has been conducted in this work. Significant progress has been made in both hardware design and computation algorithms mostly in the past two decades. The sensor systems are designed for different targeted body parts and thus include diverse designs in aspects such as the type, number, and location of cameras and IMUs. Sensor fusion at different levels has been proposed ranging from raw data fusion to fusion of estimated poses from individual systems with different algorithms. We further identified the challenges in the current systems for pose estimation and discussed the possible directions for future research. Developments in the following aspects are highly desirable.

1) Increase the mobility of the sensor system. This can reduce the restriction on the capture volume and may be achieved with wearable cameras/IMUs or cameras on mobile robots.
2) Simplify the calibration procedure. Future algorithms may take advantage of the abundant and complementary sensor data to achieve automatic calibration of sensor-to-sensor and sensor-to-segment parameters.
3) Develop real-time processing methods. Attention should be paid to increasing the practicality of the system and enabling plug-and-play capability.
4) Standardize the benchmark process and metrics. Studies should report the accuracy in widely accepted metrics and factors affecting the performance should also be discussed to guide future applications. With the contributions from future research at the hardware and software level, wide and mature applications of these visual–inertial sensors will be achieved.

References

[1] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia-Mag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[2] Y. Zhang, K. Chen, J. Yi, T. Liu, and Q. Pan, "Whole-body pose estimation in human bicycle riding using a small set of wearable sensors," *IEEE/ASME Trans. Mechatronics*, vol. 21, no. 1, pp. 163–174, Feb. 2016.

[3] V. Hoang and K. Jo, "3-D human pose estimation using cascade of multiple neural networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2064–2072, Apr. 2019.

[4] W. Kim, J. Sung, D. Saakes, C. Huang, and S. Xiong, "Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose)," *Int. J. Ind. Ergonom.*, vol. 84, Jul. 2021, Art. no. 103164.

[5] W. Xu, D. Xiang, G. Wang, R. Liao, M. Shao, and K. Li, "Multiview video-based 3-D pose estimation of patients in computer-assisted rehabilitation environment (CAREN)," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 2, pp. 196–206, Apr. 2022.

[6] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1691–1703, Sep. 2012.

[7] J. Shotton et al., "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[8] B. Artacho and A. Savakis, "UniPose: Unified human pose estimation in single images and videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7033–7042.

[9] J. Li et al., "Real-time human motion capture based on wearable inertial sensor networks," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8953–8966, Jun. 2022.

[10] T. Li, L. Wang, J. Yi, Q. Li, and T. Liu, "Reconstructing walking dynamics from two shank-mounted inertial measurement units," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 6, pp. 3040–3050, Dec. 2021.

[11] K. D. Nguyen, I.-M. Chen, Z. Luo, S. H. Yeo, and H. B.-L. Duh, "A wearable sensing system for tracking and monitoring of functional arm movement," *IEEE/ASME Trans. Mechatronics*, vol. 16, no. 2, pp. 213–220, Apr. 2011.

[12] S. Hu, M. Dai, T. Dong, and T. Liu, "A textile sensor for long durations of human motion capture," *Sensors*, vol. 19, no. 10, p. 2369, May 2019.

[13] D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. Whittlesey, *Research Methods in Biomechanics*. Champaign, IL, USA: Human Kinetics, 2013.

[14] P. Eichelberger et al., "Analysis of accuracy in optical motion capture—A protocol for laboratory setup evaluation," *J. Biomech.*, vol. 49, no. 10, pp. 2085–2088, Jul. 2016.

[15] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D human pose estimation: A review of the literature and analysis of covariates," *Comput. Vis. Image Understand.*, vol. 152, pp. 1–20, Nov. 2016.

[16] Z. Liu, J. Huang, J. Han, S. Bu, and J. Lv, "Human motion tracking by multiple RGBD cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 2014–2027, Sep. 2017.

[17] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 529–545.

[18] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.

[19] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 561–578.

[20] S. Dubey and M. Dixit, "A comprehensive survey on human pose estimation approaches," *Multimedia Syst.*, vol. 29, no. 1, pp. 167–195, Feb. 2023.

[21] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion capture from body-mounted cameras," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011.

[22] H. Rhodin et al., "EgoCap: Egocentric marker-less motion capture with two fisheye cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, Nov. 2016.

[23] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors," Xsens Motion Technol. BV, Enschede, The Netherlands, Tech. Rep. 1, 2009, pp. 1–7, vol. 1.

[24] J. M. Lambrecht and R. F. Kirsch, "Miniature low-power inertial sensors: Promising technology for implantable motion capture systems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 6, pp. 1138–1147, Nov. 2014.

[25] M. Nazarahari and H. Rouhani, "40 years of sensor fusion for orientation tracking via magnetic and inertial measurement units: Methods, lessons learned, and future challenges," *Inf. Fusion*, vol. 68, pp. 67–84, Apr. 2021.

[26] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, "Estimation of IMU and MARG orientation using a gradient descent algorithm," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, Jun. 2011, pp. 1–7.

[27] D. Laidig, I. Weygers, S. Bachhuber, and T. Seel, "VQF: A milestone in accuracy and versatility of 6D and 9D inertial orientation estimation," in *Proc. 25th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2022, pp. 1–6.

[28] B. Fan, Q. Li, and T. Liu, "How magnetic disturbance influences the attitude and heading in magnetic and inertial sensor-based orientation estimation," *Sensors*, vol. 18, no. 2, p. 76, Dec. 2017.

[29] M. Mundt et al., "Assessment of the measurement accuracy of inertial sensors during different tasks of daily living," *J. Biomech.*, vol. 84, pp. 81–86, Feb. 2019.

[30] A. G. Cutti, G. Paolini, M. Troncossi, A. Cappello, and A. Davalli, "Soft tissue artefact assessment in humeral axial rotation," *Gait Posture*, vol. 21, no. 3, pp. 341–349, Apr. 2005.

[31] R. V. Vitali and N. C. Perkins, "Determining anatomical frames via inertial motion capture: A survey of methods," *J. Biomech.*, vol. 106, Jun. 2020, Art. no. 109832.

[32] B. Fan, Q. Li, T. Tan, P. Kang, and P. B. Shull, "Effects of IMU sensor-to-segment misalignment and orientation error on 3-D knee joint angle estimation," *IEEE Sensors J.*, vol. 22, no. 3, pp. 2543–2552, Feb. 2022.

[33] Y. Tao, H. Hu, and H. Zhou, "Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation," *Int. J. Robot. Res.*, vol. 26, no. 6, pp. 607–624, Jun. 2007.

[34] F. Santoso, M. A. Garratt, and S. G. Anavatti, "Visual–inertial navigation systems for aerial robotics: Sensor fusion and technology," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 260–275, Jan. 2017.

[35] T. Kaichi, T. Maruyama, M. Tada, and H. Saito, "Resolving position ambiguity of IMU-based human pose with a single RGB camera," *Sensors*, vol. 20, no. 19, pp. 1–12, 2020.

[36] Z. Huai and G. Huang, "Robocentric visual-inertial odometry," *Int. J. Robot. Res.*, vol. 41, no. 7, pp. 667–689, 2022.

[37] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual–inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[38] L. Jinyu, Y. Bangbang, C. Danpeng, W. Nan, Z. Guofeng, and B. Hujun, "Survey and evaluation of monocular visual–inertial SLAM algorithms for augmented reality," *Virtual Reality Intell. Hardw.*, vol. 1, no. 4, pp. 386–410, Aug. 2019.

[39] W. Fang, L. Zheng, H. Deng, and H. Zhang, "Real-time motion tracking for mobile augmented/virtual reality using adaptive visual–inertial fusion," *Sensors*, vol. 17, no. 5, p. 1037, May 2017.

[40] J. Male and U. Martinez-Hernandez, "Recognition of human activity and the state of an assembly task using vision and inertial sensor fusion methods," in *Proc. 22nd IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2021, pp. 919–924.

[41] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 1, pp. 51–61, Feb. 2015.

[42] Y. Lee, W. Do, H. Yoon, J. Heo, W. Lee, and D. Lee, "visual–inertial hand motion tracking with robustness against occlusion, interference, and contact," *Sci. Robot.*, vol. 6, no. 58, Sep. 2021, Art. no. eabe1315.

[43] Y. Tao and H. Hu, "A novel sensing and data fusion system for 3-D arm motion tracking in telerehabilitation," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 5, pp. 1029–1040, May 2008.

[44] C. Malleson, J. Collomosse, and A. Hilton, "Real-time multi-person motion capture from multi-view video and IMUs," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1594–1611, Jun. 2020.

[45] A. Ahmed and S. Roumeliotis, "A visual–inertial approach to human gait estimation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4614–4621.

[46] T. Li and H. Yu, "Upper body pose estimation using a visual–inertial sensor system with automatic sensor-to-segment calibration," *IEEE Sensors J.*, vol. 23, no. 6, pp. 6292–6302, Mar. 2023.

[47] N. U. Khan and W. Wan, "A review of human pose estimation from single image," in *Proc. Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Jul. 2018, pp. 230–236.

[48] J. Zhao, "A review of wearable IMU (inertial-measurement-unit)-based pose estimation and drift reduction technologies," *J. Phys., Conf. Ser.*, vol. 1087, no. 4, 2018, Art. no. 042003.

[49] Y. Niu, J. She, and C. Xu, "A survey on IMU-and-vision-based human pose estimation for rehabilitation," in *Proc. 41st Chin. Control Conf. (CCC)*, Jul. 2022, pp. 6410–6415.

[50] G. Huang, "visual–inertial navigation: A concise review," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9572–9582.

[51] M. Servières, V. Renaudin, A. Dupuis, and N. Antigny, "Visual and visual–inertial SLAM: State of the art, classification, and experimental benchmarking," *J. Sensors*, vol. 2021, pp. 1–26, Feb. 2021.

[52] S. Majumder and N. Kehtarnavaz, "Vision and inertial sensing fusion for human action recognition: A review," *IEEE Sensors J.*, vol. 21, no. 3, pp. 2454–2467, Feb. 2021.

[53] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 4405–4425, Feb. 2017.

[54] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Systematic Rev.*, vol. 10, no. 1, p. n71, Mar. 2021.

[55] G. Bleser, G. Hendeby, and M. Miezal, "Using egocentric vision to achieve robust inertial body tracking under magnetic disturbances," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 103–109.

[56] A. P. L. Bó, M. Hayashibe, and P. Poignet, "Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 3479–3483.

[57] G. Pons-Moll, L. Leal-Taixe, J. Gall, and B. Rosenhahn, "Data-driven manifolds for outdoor motion capture," in *Proc. 15th Int. Workshop Theor. Found. Comput. Vis.*, 2012, pp. 305–328.

[58] T. Helten, M. Müller, H. Seidel, and C. Theobalt, "Real-time body tracking with one depth camera and inertial sensors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1105–1112.

[59] F. Destelle et al., "Low-cost accurate skeleton tracking based on fusion of Kinect and wearable inertial sensors," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 371–375.

[60] C. Kalkbrenner, S. Hacker, A. E. Algorri, and R. Blechschmidt-Trapp, "Motion capturing with inertial measurement units and Kinect," in *Proc. Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2014, pp. 120–126.

[61] Y. Tian, X. Meng, D. Tao, D. Liu, and C. Feng, "Upper limb motion tracking with the integration of IMU and Kinect," *Neurocomputing*, vol. 159, pp. 207–218, Jul. 2015.

[62] V. Joukov, R. D'Souza, and D. Kulic, "Human pose estimation from imperfect sensor data via the extended Kalman filter," in *Proc. 15th Int. Symp. Express Robot.*, 2016, pp. 789–798.

[63] F. Kobayashi, K. Kitabayashi, K. Shimizu, H. Nakamoto, and F. Kojima, "Human motion caption with vision and inertial sensors for hand/arm robot teleoperation," *Int. J. Appl. Electromagn. Mech.*, vol. 52, nos. 3–4, pp. 1629–1636, Dec. 2016.

[64] T. V. Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and IMUs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1533–1547, Aug. 2016.

[65] H. Tannous et al., "A new multi-sensor fusion scheme to improve the accuracy of knee flexion kinematics for functional rehabilitation movements," *Sensors*, vol. 16, no. 11, p. 1914, Nov. 2016.

[66] A. Atrsaei, H. Salarieh, and A. Alasty, "Human arm motion tracking by orientation-based fusion of inertial sensors and Kinect using unscented Kalman filter," *J. Biomech. Eng.*, vol. 138, no. 9, Sep. 2016, Art. no. 091005.

[67] S. Y. Cho, S. Y. Lee, J. H. Lim, and S. J. Park, "Simultaneous motion tracking and localisation of a person based on the integration of multiple IMUs and depth camera," *IET Radar, Sonar Navigat.*, vol. 11, no. 11, pp. 1679–1688, Nov. 2017.

[68] S. Han, H. Zhang, X. Wang, L. Xu, and N. Zheng, "A rehabilitation gait training system for half lower limb disorder," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017, pp. 3841–3847.

[69] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.

[70] X. Lu, K. Yu, Y. Zhang, J. Yi, J. Liu, and Q. Zhao, "Whole-body pose estimation in physical rider-bicycle interactions with a monocular camera and wearable gyroscopes," *J. Dyn. Syst. Meas. Control-Trans. ASME*, vol. 139, no. 7, 2017, Art. no. 071005.

[71] G. Glonek and A. Wojciechowski, "Hybrid orientation based human limbs motion tracking method," *Sensors*, vol. 17, no. 12, p. 2857, Dec. 2017.

[72] T. K. Chan, Y. K. Yu, H. C. Kam, and K. H. Wong, "Robust hand gesture input using computer vision, inertial measurement unit (IMU) and flex sensors," in *Proc. IEEE Int. Conf. Mechatronics, Robot. Autom. (ICMRA)*, May 2018, pp. 95–99.

[73] P. Jatesiktat, D. Anopas, and W. T. Ang, "Personalized markerless upper-body tracking with a depth camera and wrist-worn inertial measurement units," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1–6.

[74] J. Kempfle and K. van Laerhoven, "PresentPostures: A wrist and body capture approach for augmenting presentations," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 318–323.

[75] K. Lebel, M. Hamel, C. Duval, H. Nguyen, and P. Boissy, "Camera pose estimation to improve accuracy and reliability of joint angles assessed with attitude and heading reference systems," *Gait Posture*, vol. 59, pp. 199–205, Jan. 2018.

[76] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 614–631.

[77] Z. Zheng et al., "HybridFusion: Real-time performance capture using a single depth sensor and sparse IMUs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 389–406.

[78] J. N. Bersamira et al., "Human gait kinematic estimation based on joint data acquisition and analysis from IMU and depth-sensing camera," in *Proc. IEEE 11th Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ., Manage. (HNICEM)*, Nov. 2019, pp. 1–6.

[79] A. Gilbert, M. Trumble, C. Malleson, A. Hilton, and J. Collomosse, "Fusing visual and inertial sensors with semantics for 3D human pose estimation," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 381–397, Apr. 2019.

[80] S. M. Orozco-Soto, A. I. Pérez-Sanpablo, P. Vera-Bustamante, and J. M. Ibarra-Zannatha, "Development of a visual–inertial motion tracking system for muscular-effort/angular joint-position relation to obtain a quantifiable variable of spasticity," in *Proc. 4th Int. Symp. Wearable Robot.* Cham, Switzerland: Springer, 2019, pp. 210–215.

[81] V. Samy, K. Ayusawa, Y. Yoshiyasu, R. Sagawa, and E. Yoshida, "Musculoskeletal estimation using inertial measurement units and single video image," in *Proc. IEEE Int. Conf. Adv. Robot. Social Impacts (ARSO)*, Oct. 2019, pp. 39–44.

[82] M. Tarabini et al., "Real-time monitoring of the posture at the workplace using low cost sensors," in *Proc. 20th Congr. Int. Ergon. Assoc.*, 2019, pp. 678–688.

[83] L. C. Vu and B. You, "Hand pose detection in HMD environments by sensor fusion using multi-layer perceptron," in *Proc. 1st Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2019, pp. 218–223.

[84] R. J. Cotton, "Kinematic tracking of rehabilitation patients with markerless pose estimation fused with wearable inertial sensors," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 508–514.

[85] F. Huang, A. Zeng, M. Liu, Q. Lai, and Q. Xu, "DeepFuse: An IMU-aware network for real-time 3D human pose estimation from multi-view image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 418–427.

[86] T. B. Rodrigues, D. P. Salgado, C. Ó. Catháin, N. O'Connor, and N. Murray, "Human gait assessment using a 3D marker-less multimodal motion capture system," *Multimedia Tools Appl.*, vol. 79, nos. 3–4, pp. 2629–2651, Jan. 2020.

[87] T. Zhang, H. Xia, C. Zhang, and Z. Zeng, "MultiModal, robust and accurate hand tracking," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 1886–1890.

[88] Z. Zhang, C. Wang, W. Qin, and W. Zeng, "Fusing wearable IMUs with multi-view images for human pose estimation: A geometric approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2197–2206.

[89] J. Zhang, P. Li, T. Zhu, W. Zhang, and S. Liu, "Human motion capture based on Kinect and IMUs and its application to human-robot collaboration," in *Proc. 5th Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Dec. 2020, pp. 392–397.

[90] J. Abbasi, H. Salarieh, and A. Alasty, "A motion capture algorithm based on inertia-Kinect sensors for lower body elements and step length estimation," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102290.

[91] K. Ahuja, S. Mayer, M. Goel, and C. Harrison, "Pose-on-the-go: Approximating user pose with smartphone sensor fusion and inverse kinematics," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–12.

[92] Y. Cha et al., "Mobile. Egocentric human body motion reconstruction using only eyeglasses-mounted cameras and a few body-worn inertial sensors," in *Proc. IEEE Virtual Reality 3D User Interfaces (VR)*, Mar. 2021, pp. 616–625.

[93] N. Gosala et al., "Self-calibrated multi-sensor wearable for hand tracking and modeling," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 3, pp. 1769–1784, Mar. 2023.

[94] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4316–4327.

[95] E. Halilaj, S. Shin, E. Rapp, and D. Xiang, "American society of biomechanics early career achievement award 2020: Toward portable and modular biomechanics labs: How video and IMU fusion will change gait analysis," *J. Biomech.*, vol. 129, Dec. 2021, Art. no. 110650.

[96] S. Hwang, K. Ko, and S. B. Pan, "Motion data acquisition method for motion analysis in golf," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 2, p. e5215, Jan. 2021.

[97] R. Mallat et al., "Sparse visual–inertial measurement units placement for gait kinematics assessment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1300–1311, 2021.

[98] M. Mitjans et al., "visual–inertial filtering for human walking quantification," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13510–13516.

[99] K. Ogata, H. Tanaka, and Y. Matsumoto, "Simple three-dimensional motion measurement system using marker-IMU system," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 7073–7076.

[100] M. Kim and S. Lee, "Fusion poser: 3D human pose estimation using sparse IMUs and head trackers in real time," *Sensors*, vol. 22, no. 13, p. 4846, Jun. 2022.

[101] T. Li, X. Wu, H. Dong, and H. Yu, "Estimation of upper limb kinematics with a magnetometer-free egocentric visual–inertial system," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 1668–1674.

[102] R. Mallat, V. Bonnet, M. A. Khalil, and S. Mohammed, "Upper limbs kinematics estimation using affordable visual–inertial sensors," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 1, pp. 207–217, Jan. 2022.

[103] M. Yamamoto, K. Shimatani, Y. Ishige, and H. Takemura, "Verification of gait analysis method fusing camera-based pose estimation and an IMU sensor in various gait conditions," *Sci. Rep.*, vol. 12, no. 1, p. 17719, Oct. 2022.

[104] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 4–18, Oct. 2007.

[105] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[106] D. Tedaldi, A. Pretto, and E. Menegatti, "A robust and easy to implement method for IMU calibration without external equipments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 3042–3049.

[107] G. Vogiatzis and C. Hernández. (Dec. 2022). *Automatic Camera Pose Estimation From Dot Pattern* [Online]. Available: http://george-vogiatzis.org/calib/

[108] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[109] J. Llinas and D. L. Hall, "An introduction to multi-sensor data fusion," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jan. 1998, pp. 537–540.

[110] C. Malleson, A. Gilbert, M. Trumble, J. Collomosse, A. Hilton, and M. Volino, "Real-time full-body motion capture from video and IMUs," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 449–457.

[111] E. Chong, A. Southerland, A. Kundu, R. M. Jones, A. Rozga, and J. M. Rehg, "Visual 3D tracking of child–adult social interactions," in *Proc. Joint IEEE Int. Conf. Develop. Learn. Epigenetic Robot. (ICDL-EpiRob)*, Sep. 2017, pp. 399–406.

[112] M. Garcia-Salguero, J. Gonzalez-Jimenez, and F.-A. Moreno, "Human 3D pose estimation with a tilting camera for social mobile robot interaction," *Sensors*, vol. 19, no. 22, p. 4943, Nov. 2019.

[113] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, vol. 2, 1981, pp. 674–679.

[114] P. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12708–12717.

[115] D. Tome et al., "SelfPose: 3D egocentric pose estimation from a headset mounted camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6794–6806, Jun. 2023.

[116] A. R. Doherty et al., "Wearable cameras in health: The state of the art and future possibilities," *Amer. J. Preventive Med.*, vol. 44, no. 3, pp. 320–323, 2013.

[117] L. E. Meyer et al., "Using wearable cameras to investigate health-related daily life experiences: A literature review of precautions and risks in empirical studies," *Res. Ethics*, vol. 18, no. 1, pp. 64–83, Jan. 2022.

[118] V. Shipp, A. Skatova, J. Blum, and M. Brown, "The ethics of wearable cameras in the wild," in *Proc. IEEE Int. Symp. Ethics Sci., Technol. Eng.*, May 2014, pp. 1–5.

[119] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[120] U. Qureshi and F. Golnaraghi, "An algorithm for the in-field calibration of a MEMS IMU," *IEEE Sensors J.*, vol. 17, no. 22, pp. 7479–7486, Nov. 2017.

[121] T. Beravs, S. Beguš, J. Podobnik, and M. Munih, "Magnetometer calibration using Kalman filter covariance matrix for online estimation of magnetic field orientation," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 8, pp. 2013–2020, Aug. 2014.

[122] Y. Wu, D. Zou, P. Liu, and W. Yu, "Dynamic magnetometer calibration and alignment to inertial sensors by Kalman filtering," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 2, pp. 716–723, Mar. 2018.

[123] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 1280–1286.

[124] J. Rehder and R. Siegwart, "Camera/IMU calibration revisited," *IEEE Sensors J.*, vol. 17, no. 11, pp. 3257–3268, Jun. 2017.

Tong Li (Member, IEEE) received the B.S. and Ph.D. degrees in mechatronic engineering from Zhejiang University, Hangzhou, China, in 2015 and 2020, respectively.

He is currently a Research Fellow with the Department of Biomedical Engineering, National University of Singapore, Singapore. His current research interests include wearable sensors, assistive devices, and human modeling.

Haoyong Yu (Senior Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1988 and 1991, respectively, and the Ph.D. degree in mechanical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2002.

He is currently an Associate Professor with the Department of Biomedical Engineering, National University of Singapore, Singapore. His areas of research include medical robotics, rehabilitation engineering and assistive technologies, and system dynamics and control.