VOEDHgesture: A Multi-Purpose Visual Odometry/ Simultaneous Localization and Mapping and Egocentric Dynamic Hand Gesture Data-Set for Virtual Object Manipulations in Wearable Mixed Reality

Yemineni Ashok[a], Mukesh Kumar Rohil[b], Kshitij Tandon[c] and Harshil Sethi[d]
Birla Institute of Technology And Science, Pilani (BITS Pilani), Pilani, Rajasthan, India

Keywords: Visual Odometry, Wearable Computing, Augmented Reality, Mixed Reality, Pose Estimation, Simultaneous Localization and Mapping.

Abstract: Visual Odometry/ Simultaneous Localization and Mapping (VO/ SLAM) and Egocentric hand gesture recognition are the two major technologies for wearable computing devices like AR (Augmented Reality)/ MR (Mixed Reality) glasses. However, the AR/MR community lacks a suitable dataset for developing both hand gesture recognition and RGB-D SLAM methods. In this work, we use a ZED mini Camera to develop challenging benchmarks for RGB-D VO/ SLAM tasks and dynamic hand gesture recognition. In our dataset VOEDHgesture, we collected 264 sequences using a ZED mini camera, along with precisely measured and time-synchronized ground truth camera positions, and manually annotated the bounding box values for the hand region of interest. The sequences comprise both RGB and depth images, captured at HD resolution (1920×1080) and recorded at a video frame rate of 30Hz. To resemble the Augmented Reality environment, the sequences are captured using a head-mounted ZED mini camera, with unrestricted 6-DOF (degree of freedom) movements in different varieties of scenes and camera motions, i.e. indoor, outdoor, slow motion, quick motions, long trajectories, loop closures etc. This dataset can help researchers to develop and promote reproducible research in the fields of egocentric hand tracking, visual odometry/SLAM and computer vision algorithms for AR scene reconstruction and scene understanding, etc.

1 INTRODUCTION

Wearable computing moves the computation from desktop computers to body-worn devices and allows the user to interact with computation units whenever and wherever it is needed. Augmented Reality (AR)/ Mixed Reality (MR) Glasses are one category of wearable computers which extend the human-computer interface by presenting the computer's digital information on the surrounding physical world's video. AR/MR Glasses have video cameras to capture the surrounding real-world video and processing power to overlay the graphical content in the surrounding real-world video in physically meaningful locations. Furthermore, within this AR-enhanced visualization context, the graphical content can be interactive and manipulative with proper user interfaces and interaction techniques.

Developing robust AR/MR systems that are able to perform virtual content manipulations in response to human gestures is one of the major challenges in computer vision tasks. The AR/MR systems evaluation mostly relies on the performance of visual odometry and hand gesture recognition modules. Hence, in recent years an increasing number of benchmarks such as KITTI (Geiger et al., 2013), RGB-D SLAM (Sturm et al., 2012), EuRoC MAV dataset (Burri et al., 2016), PennCOSYVIO dataset (Pfrommer et al., 2017), Newer College's Stereo Vision Lidar IMU Dataset (Ramezani et al., 2020), Newer College's Multicam Vision LiDAR IMU dataset (Zhang et al., 2021), Interactive Museum nvGesture(Baraldi et al., 2014), nvGesture(Molchanov et al., 2016), EgocentricGesture(Zhang et al., 2018) etc. have been introduced for benchmarking and to achieve better performance levels in Visual odometry systems and hand gesture recognition systems. However, most of these datasets are specific to the evaluation of either visual odometry tasks or hand gesture recognition tasks.

[a] https://orcid.org/0000-0002-5550-5159
[b] https://orcid.org/0000-0002-2597-5096
[c] https://orcid.org/0009-0007-1105-7853
[d] https://orcid.org/0009-0001-9859-3662

Ashok, Y., Rohil, M., Tandon, K. and Sethi, H.
VOEDHgesture: A Multi-Purpose Visual Odometry/ Simultaneous Localization and Mapping and Egocentric Dynamic Hand Gesture Data-Set for Virtual Object Manipulations in Wearable Mixed Reality.
DOI: 10.5220/0012473900003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 1336-1344
ISBN: 978-989-758-680-4; ISSN: 2184-433X

1.1 Visual Odometer: A "Basic Problem" for Markerless AR

The visual odometry (VO) process measures the agent's (e.g. human, vehicle, robot etc.) egocentric motion by analysing the input of single or multiple cameras fixed to it. VO/SLAM can be applied in various such as wearable computing, robotics, Autonomous vehicles, Micro Aerial Vehicles (MAV), etc. Wearable AR glasses can display the computer's graphical content on real-world video in meaningful full locations, provided that it has the knowledge of where the user is looking. In a localized and relative sense, the determination of tracking and mapping of the camera over a period of time is the major module in the visual odometry framework, and these tracking and mapping calibrations can help in estimating the orientation of the virtual content in every video frame.

Various devices and equipment like RGB-D cameras, such as Microsoft Kinect, Intel RealSense, Stereolab's ZED, etc., can significantly help to push the evaluation of benchmark's state-of-the-art forward. RGB-D cameras provide the 3D structure of the environment in addition to texture information, and these depth maps help simplify the complexities involved in SLAM's initialization process. In this paper, we utilize one of the recent RGB-D cameras, ZED mini, to capture real-world scenes.

1.2 Hand Gesture Recognition: A Basic Problem in Human Wearable Computer (AR/MR) Interaction

Hand gestures are the natural and intuitive way to interact with wearable computers like AR/MR glasses. AR/MR glasses can capture hand gestures with a head-mounted camera. This captured video introduces a distinct human-centric perspective of the surrounding visual world, thereby exhibiting new hand gesture recognition system characteristics 1) Egocentric motion: As the camera is attached to the user's head as in figure 1, the camera motion can be significantly affected by quick and sudden head motions, especially when the user performs gesture while walking. 2) Hands in short-range: Hands could partly or fully be out of video frames because of the close proximity between the camera and hand.

Currently, it is difficult to find a benchmark dataset that can be useful in developing both camera pose estimation and ego-centric dynamic hand gesture recognition methods. Most of the hand gesture recognition-related datasets provide bounding boxes and labels

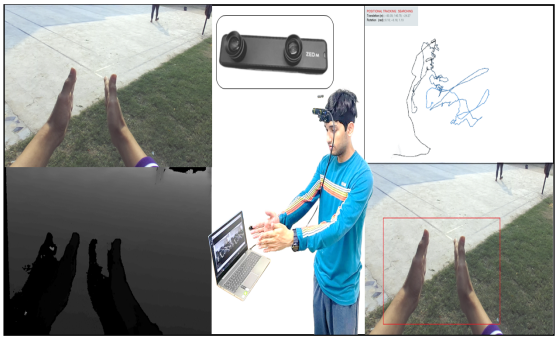

Figure 1: (Middle) Recording platform (A subject with a head-mounted camera for capturing the hand gestures), RGB (left-top) and depth (left-bottom) images generated by the ZED camera, trajectory (Right top) and bounding box labels (Right bottom).

for action recognition, sign language understanding, and hand detection tasks only. They do not explicitly provide any ground truth trajectory calibrations for camera pose estimation. The datasets like Interactive Museum (Baraldi et al., 2014) and EgocentricGesture (Zhang et al., 2018) are the two public datasets available for the evaluation of egocentric gesture recognition tasks. However, these datasets provide only the spatial and temporal-related information between the frames for human-computer interaction tasks. In summary, we take advantage of a Visual inertial stereo ZED mini camera to develop novel challenging benchmarks for visual odometry/SLAM and egocentric dynamic hand gesture recognition. The major contributions of this work include:

1. In this paper, we provide 264 real RGB-D egocentric hand gesture sequences that are captured in various illumination conditions like in outdoor sunlight, indoor sunlight, nighttime in artificial light, and time in dim lighting conditions, along with the trajectory of ground truth to fully quantify the accuracy of a given VO/SLAM system.
2. This dataset sequence consists of 40 basic dynamic or static hand gesture classes that are useful for hand gesture recognition tasks for controlling virtual content manipulations. These are collected from 25 human subjects of varied age groups and genders and in various illumination conditions, e.g., daytime, nighttime with artificial light, and evening with light illuminations.
3. We show the performance analysis on traditional as well as recent RGB-D SLAM models and hand gesture recognition models.

With all these novelties, our VOEDHgesture dataset is the first of its kind to be useful for developing two basic modules e.g. tracking and interaction of wearable AR/MR glasses.

Table 1: Comparison of Related Data sets and Benchmarks.

Dataset	Sensors	Scenario	Type	Platform	Size
New College Data set (Smith et al., 2009)	Stereo camera	Outdoor	Real	Robot	1 sequence
KITTI (Geiger et al., 2012)	Stereo Camera	Outdoor	Real	car	22 sequences
RGB-D SLAM (Sturm et al., 2012)	RGB-D camera	Indoor	Real	Robot/ Handheld	39 sequences
ICL-NUIM (Handa et al., 2014)	RGB-D Camera	Indoor	Synthetic	Handheld	8 sequences
KITTI (Geiger et al., 2013)	Stereo Camera	Dynamic outdoor	Real	car	400 Scenes
TUM-mono-VO dataset (Engel et al., 2016)	Monocular camera	Indoor& Outdoor	Real	Handheld	50 Sequences
EuRoC MAV dataset (Burri et al., 2016)	stereo Camera, IMU	Machine hall& Room	Real	MAV	11 Sequences
UMich NCLT dataset (Carlevaris-Bianco et al., 2015)	LiDAR	Indoor& Outdoor	Real	Robot	27Sequences
PennCOSYVIO dataset (Pfrommer et al., 2017)	stereo Camera, IMU	Indoor& Outdoor	Real	Handheld	27Sequences
TUM VI Dataset (Schubert et al., 2018)	stereo Camera, IMU	Indoor& Outdoor	Real	Handheld	28 sequences
RIDI Dataset (Yan et al., 2018)	Smartphone with IMU	Indoor& Outdoor	Real	Human body worn	60 sequences
ADVIO (Cortés et al., 2018)	Smartphone with IMU	Indoor& Outdoor	Real	Handheld	23 sequences
Newer College's Stereo Vision Lidar IMU Dataset (Ramezani et al., 2020)	LiDAR,IMU	outdoor	Real	Handheld	9 Sequnces
TartanAir (Wang et al., 2020)	-	Indoor& Outdoor	Synthetic	-	1037 Sequnces
Newer College's Multicam Vision LiDAR IMU dataset (Zhang et al., 2021)	LiDAR,IMU	outdoor	Real	Handheld	6 Sequences
VOEDHgesture (ours)	Stereo camera with RGB-D Data Output	Indoor & outdoor	Real	Head Mounted	264 Sequences

Notes: "-" means Data not Avilable

2 RELATED WORK

Most of the available public benchmark datasets are specific to either VO/SLAM or hand gesture recognition tasks. Hence we are analysing both of them separately in their subsections as below.

2.1 VO/ SLAM Datasets

Based on the sensor carrying platform, VO/ SLAM datasets can be divided into two categories: 1) a vehicle (e.g. robot, car, MAV) attached and 2) a human carried. Vehicle-based datasets are often used to evaluate pose estimate variants in robotics and autonomous driving applications, while human-carrying datasets are used to evaluate fields where sudden and quick camera motions are presented, e.g. Augmented Reality, Mixed Reality, etc. Focusing on the sensors and environment modules, Table1 provides information about the datasets that are useful for the evaluation of VO/ SLAM methods and the same datasets are summarized below.

2.1.1 Vechicle Based Datasets

RGB-D dataset (Sturm et al., 2012) stands as one of the initial RGB-D datasets. These dataset sequences are captured with a Kinect sensor and are collected in two sensor-carrying platform scenarios: 1) a robot and 2) a human handheld. All the videos are recorded in indoor environments along with ground truth trajectory calibrations. The ground truth is calculated with the integration of a high-precision motion capture system. Similarly, The ICL-NUIM (Handa et al., 2014) dataset was gathered with a handheld RGB-D camera, in which artificial sensor noise is added to get a realistic feeling in the indoor room sequences visuals.

New College Data set (Smith et al., 2009) was one of the first stereo datasets. Its sequences are captured in an outdoor environment of the New College Grounds in Oxford with a robot's attached BumbleBee stereo camera. Similarly, The KITTI (Geiger et al., 2012) and KITTI (Geiger et al., 2013) datasets are the stereo datasets released for research on autonomous vehicle navigation. All the kitti sequences are captured in the outdoor environment with a stereo camera attached to the car platform.

EuRoC MAV dataset (Burri et al., 2016) gathers stereo-inertial sequences with an onboard Micro Aerial Vehicle (MAV) in two environments: 1) Machine Hall and 2) Vicon Room. This dataset's ground truth was captured using a motion capture system and a laser tracker.

2.1.2 Human Based Datasets

TUM-mono-VO (Engel et al., 2016) dataset was captured using a hand-held monocular camera in different environment conditions that range from narrow indoor corridors to wide outdoor scenes. PennCOSYVIO (Pfrommer et al., 2017) and TUM VI (Schubert et al., 2018) datasets are focused on visual-inertial SLAM evaluation techniques. Both PennCOSYIO and TUM VI datasets employ stereo cameras to generate indoor and outdoor sequences and these hold with a human hand. RIDI Dataset (Yan et al., 2018) and ADVIO (Cortés et al., 2018) are specifically focused on mobile-based visual-inertial odometry. ADVIO employs a Google Pixel smartphone and an Apple iPhone for their test equipment. Similarly, the subjects wore four smartphones (e.g. google Tango phone, LenovoPhad2 Pro. etc.) on

Table 2: Comparison of different Benchmark Data sets for Hand Gesture Estimation.

Dataset	Modality	Sensors	Resolution	Number Subjects	Gesture Classes	Environme	View	Dynamic/ Static
American Sign Language	RGB	-	320 × 243	1	40	1	Fisr,Second	-
Cambridge Gestures	RGB	-	320 × 240	2	9	5	Second	Static
ChAirGesture	RGB, Depth,IMU	Kinect	640 × 480	10	10	2	second	Static, Dynamic
SKIG	RGB, Depth	Kinect	640 × 480	6	10	3	Second	Dynamic
Chalearn Dataset	RGB, Depth	kinect	640× 480	27	20	1	Second	Static
Interactive Museum	RGB	-	800 × 450	5	7	1	Fisrt	Static, Dynamic
nvGesture	RGB, Depth, Stereo IR	SoftKinect	320 × 240	20	25	1	Second	Static, Dynamic
EgocentricGesture	RGB, Depth	Re	640 × 480	50	83	6	First	Static, Dynamic
ArASL	RGB	-	64 × 40	40	32	-	Second	Static
IPN Hand	RGB	-	640 × 480	50	14	28	Second	Static,Dynamic
HANDS	RGB,Depth	Kinect	960 × 540	5	29	5	Second	Static
VOEDHgesture (ours)	RGB, Depth	ZED mini	1920 × 1080	25	40	10	First	Static,Dynamic

Notes: "-" means Data not Avilable

their body during the recording process of the RIDI Dataset (Yan et al., 2018). Newer College's Stereo Vision Lidar IMU Dataset (Ramezani et al., 2020) and Newer College's Multicam Vision LiDAR IMU dataset (Zhang et al., 2021) employ powerful LiDAR scans for determining the ground truth and the data is recorded with a hand-held camera.

2.2 Hand Gesture Recognition Dataset

In the field of human hand gesture recognition, most of the established datasets are captured for the application of sign language prediction. So, these datasets do not include the VO/SLAM's trajectory ground truth and were captured in second-person view. In the second-person view, the subject performs hand gesture activity in front of a camera. The camera faces the subject at a relatively near distance and acts like a receiver in the scene. Whereas in the first-person view (Ego-Centric vision), the camera is mounted on a subject itself and acts as a performer in the scene. Based on the view, sensor modalities, Table 2 provides information about some of the hand gesture datasets and the same are briefly discussed below.

For the RGB-related datasets, the American Sign Language (Starner et al., 1998) dataset (ASL dataset) collected the data for the sign language purpose with both first and second-person approaches. The ASL dataset contains 2,500 images with 40 classes of gestures in a single subject in only one indoor environment. Similar to ASL dataset, Interactive Museum (Baraldi et al., 2014) also adopts the egocentric approach and provides RGB data for gesture recognition and segmentation tasks. It contains 7 gesture classes from 5 subjects and also includes a sample of dynamic hand gestures. Cambridge hand gesture dataset (Kim et al., 2007) is also a similar kind of dataset that provides RGB sequences for action/gesture classification. However, all the 900 sequences with nine gesture classes presented in this dataset are collected in a second-view approach. Although these kinds of datasets, without any ego-centric vision videos, are not helpful for the evaluation of wearable computer interaction techniques, these are useful for the results comparison of gesture recognition tasks. Recent RGB datasets ArASL (Latif et al., 2019) and IPN Hand (Benitez-Garcia et al., 2021) are with a second-person approach. The AirASL dataset has static 54000 images, which are useful for Arabic Sign language understanding. IPN Hand provides more than four thousand images to understand both static and dynamic hand gestures.

The SKIG (Liu and Shao, 2013) and Chalearn (Escalera et al., 2013) datasets provide RGB and depth sequences, which are collected through the second-view approach. SKIG dataset's 1080 sequences are collected from six subjects with ten hand gesture classes. Chalearn dataset contains 15,000 images, which are captured with a Kinect camera for solving pose estimation problems. In addition to RGB-D information, the ChAirGesture dataset provides inertial measurements for their hand gesture sequences. In the recording setup of ChAirGesture data collection, an accelerometer is attached to the human hand to gather the inertial calibrations as per the hand movement. The nvGesture dataset (Molchanov et al., 2016) includes stereo IR input and RGB-D information in the dataset. This dataset's sequences are collected over 20 subjects with 25 gesture classes in a simulated driving environment. The HANDS (Nuzzi et al., 2021) dataset also provides RGB and depth frames for human-robot interaction. It contains 29 unique gesture classes that can be formed using single or both hands.

In this paper, we are introducing a novel RGB-D dataset to develop fundamental VO/SLAM and gesture recognition modules of AR/ MR system, which can be useful for the evaluation of methods that are useful for projecting the virtual object's visuals as per the camera orientation and to manipulate the virtual content as per the human's hand gestures.

3 DATA ACQUISITION AND DATA ANALYSIS

Developing a comprehensive, realistic and large-scale benchmark for the evaluation of the above-mentioned tasks presents several challenges, such as capturing a large amount of data in real time, creating the ground truth while minimizing the need for extensive supervision, and selecting appropriate video frames related to each gesture class and hand region of interest in each frame. In this section, we discuss our approaches to addressing these challenges.

3.1 Methodology, Sensors and File Formats Used

We use a ZED mini visual inertial stereo camera, one of the few that can accurately measure the 3D information of the environment with depth accuracy such that the maximum error is only of $< 1.5\%$ when capturing scene from 3M baseline and error is up to 7% when capture scene distance is above 3m. The camera captures different images of the scene with slightly different perspectives from its two lenses and generates depth information with stereo depth sensing technology. A baseline of 63 mm separates these lenses to match the average distance between human pupils. All data was captured at a high resolution of 1920×1080 .

ZED Mini Calibration. We followed the (Geiger et al., 2012) process to calibrate the ZED mini intrinsic and extrinsic parameters. We placed a checkboard pattern in front of the zed camera and detected corners in our calibration images. The process of matching corners between the checkboard and the camera and optimizing them with reprojection errors will give us the calibration parameters.

Groundtruth. The ground truth for the trajectory of the VO/ SLAM system is directly obtained from the RGB-D/ IMU positional tracking module's output of the ZED mini camera. Despite using online crowdsourcing to annotate tools to generate dynamic hand gesture ground truth labels, we manually selected the appropriate video frames related to each gesture and assigned tracklets to hand regions of interest in each frame in the form of a bounding box. So that it is potentially able to create bounding boxes even in motion blur, truncated, occluded and semi-occluded situations.

File Formats. Each sequence data is provided in TGZ file format and contains the following files/ directories:

- "rgb/". A directory provides RGB frames of the sequence in .png format

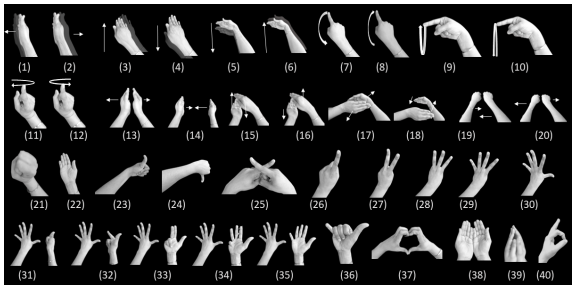

Figure 2: The illustration of 40 gestures classes present in the proposed VOEDHgesture dataset.

Table 3: Possible Hand Gestures for Virtual Content Manipulations.

S.N	Category	Tasks	Manipulation	Action
1	Manipulative	Transform	Along X-Axis Towards Right	Move Towards Right
2			Along X-Axis Towards Left	Move Towards Left
3			Along Y-Axis Towards Downwards	Move Towards Downwards
4			Along Y-Axis Towards Upwards	Move Towards Upwards
5			Along Z-Axis Towards Right	Move Towards Subject
6			Along Z-Axis Towards Left	Moving Away Form Subject
7		Rotation	Along X-Axis(Roll) Towards Clockwise	Move Towards Right
8			Along X-Axis(Roll) towards Anti Clockwise	Move Towards Left
9			Along Y-Axis(Pinch) towards Clockwise	Move Apart From Subject
10			Along Y-Axis(Pinch) towards Anti Cockwise	Move Towards Subject
11			Along Z-Axis(Yaw) towards Clockwise	Move Upwards
12			Along Z-Axis(Yaw) Towards Anticlockwise	Move Downwards
13		Scale	Along X-Axis Enlarge	Two Hands Move Apart
14			Along X-Axis Shrink	Two Hands Move Together
15			Along Y-Axis	Two Hands Move Apart
16			Along Y-Axis	Two Hands Move Together
17			Along Z-Axis	Two Hands Move Apart
18			Along Z-Axis	Two Hands Move Together
19			Scale Uniform (enlarge)	Two Hands Move Apart
20			Scale Uniform (shrink)	Two Hands Move Together
21		Bounce	Bouncing	Constant Position
22	Control		Insert New Virtual Object	Constant Position
23			Lock The Virtual Object Position	Constant Position
24			Unlock The Virtual Object Poistion	Constant Position
25			Erase All AR Visuals	Constant Position
26	Speed Control		1×	Constant Position
27			2×	Constant Position
28			3×	Constant Position
29			4×	Constant Position
30			5×	Constant Position
31			6×	Constant Position
32			7×	Constant Position
33			8×	Constant Position
34			9×	Constant Position
35			10×	Constant Position
36	Other Gestures			Constant Position
37				Constant Position
38				Constant Position
39				Constant Position
40				Constant Position

Figure 3: Some hand gesture samples to illustrate the complexity of our dataset (A) Gesture classes with a single hand in library environment(left) and outdoor environment(right) (B) Gesture classes with both hands in a reading room (left) and on an outside road (right) (C) Gestures with motion blur in indoor environment (D) Hand gesture in dynamic background in outside environment(left) and in inside environment (right) (E) Hand gesture in artificial light(low illumination) in a mechanical workshop (left) and in outside playground (right) (F) Gestures with hand out of the frame in outside park environment (left) and in a corridor (right) (G) Gestures collected in mechanical workshop environment (H) Gestures in occluded environment.

- "depth/". A directory provides grayscale depth images of the sequence in .png format
- "associations.txt". A text file contains a matrix of size $N \times 4$ (One row by frame). Each row represents a consecutive list of RGB and depth frames of the sequence (time_stamp RGB_filename time stamp Depth_filename).
- "gestureinforamtion.txt". A textfile contains a matrix of size $N \times 7$ (One row by frame). Each row represents a consecutive list of hand regions of interest and gesture class number (Format: sequence_name, timestamp, x,y, width, height, gesture class)
- "groundtruth.txt". A text file contains a matrix of size $N \times 7$. each row represents the positional and rotation vectors (format: time_stamp tx ty tz qx qy qz qw).
- "calibration.yaml". A text containing the calibration parameters of the camera.

In addition to the above files, we provide train_handgesture.txt and test_handgesture.txt that provide the information related to train and test sequences respectively.

The subjects wear the ZED mini camera to their head using a strap-mounted belt, as illustrated in figure 1 and are directed to perform all the hand gestures classes illustrated in figure 2 in nine different environments, these environments include six indoor scenes (e.g. Hostel room, reading room, library, lobby, mechanical lab) and three outdoor scenes (Outdoor lobby, road or pathway, park). To simulate all possible scenarios of wearable counting, we defined four different scenarios: 1) Both the subject and background are stationary, 2) The subject is stationary against a dynamic background, 3) The subject is dynamic (Walking) against a static background, and 4) Both subject and background are dynamic. During the data collection process, in the beginning, we guide the subjects in the execution of each gesture along with brief descriptions and provide them with the corresponding gesture names as per the list of gesture names in the table 3. The subjects are informed of the gesture name and instructed to execute it accordingly. Each session involves the continuous performance of all 40 gestures in any random order. These are then captured and recorded as a video.

3.2 Characteristics of the Dataset

The sequences in the dataset could have been characterized along gestures classes, subjects, ego-centric motion, illumination conditions and cluttered background, as illustrated in figure 3 and the same is explained below:

Gesture Classes. As we are designing the gesture classes for wearable-computer interaction, the gesture classes should be meaningful, and easily memorable to users. Following these principles, we have categorized the gestures into four main categories: 1) Manipulative, 2) Control, 3) Speed control, and 4) Other gestures as in table 3. The manipulative class gestures are used to instruct the computer for position change, rotational change and scale change of virtual content.

Except for the bouncing operation, all remaining operations are designed as dynamic gestures and can be performed with both hands of humans. The speed gestures are designed as a static operation and are useful to control the speed of the manipulative operation. The control gestures are also static and are useful for controlling the AR system. The others are not related to our virtual object manipulative system but can used for the learning process.

Subjects. The limited intra-class variation due to a small number of subjects problem can potentially be addressed by increasing the number of subjects. So, we invited 25 subjects to expand our data collection efforts. Twenty males and five females are within the group of 25 subjects, and these subjects' average age is 25, the range is [20,40].

Egocentric Motion. When individuals utilize wearable computers, they are frequently in motion, typically walking. This can lead to significant egocentric motion, resulting in a change in orientation and motion blur in sequences. Hence, we provided the ground truth trajectory for evaluating the pose estimation process and incorporated these motion blur sequences in our dataset for effective training of the gesture recognition model.

Illumination Conditions. As baseline models are vision-based, these are very sensitive to illumination change. Hence, to evaluate the robustness of the model, we have collected sequences of three different illumination settings: 1) In the daylight, 2) at night time with artificial light, and 3) In little illumination conditions.

Clutter Background. To capture more realistic videos, we recorded the scenes with static backgrounds adorned with everyday items and dynamic backgrounds with pedestrians walking into the camera.

4 BENCHMARK EVALUATION AND ANALYSIS

4.1 Visual Odometry/ Simultaneous Localization and Mapping

We executed three different VO/SLAM approaches 1) ORB-SLAM2 (Mur-Artal and Tardós, 2017), 2) ElasticFusion (Whelan et al., 2015), and 3) Maskfusion (Runz et al., 2018) on seven different sequences of our "VOEDHgesture" dataset 1) en1: indoor library 2) en2: Inside a room while subject is walking 3) en3: inside a mechanical workshop in low light illumination 4) en4: inside a mechanical workshop in the day time with normal sunlight 5) en5: On a road while dynamic objects are moving in background 6) en6: in a motion blur due to student hand or head movements and 7) en7: hand captured partially for the performance evaluation. From the evaluation of the above SLAM models on our dataset found that these algorithms couldn't deliver better results in complex, realistic situations like en3.

Table 4: Visual Odometry Evaluation.

Sequence	ORB-SLAM2	ElasticFusion	MaskFusion
en1sub1	0.010	0.020	0.030
en2sub2	0.009	0.013	0.018
en3sub3	0.0680	0.070	0.072
en4sub2	0.005	0.009	0.018
en5sub1	0.025	0.018	0.019
en1sub2	0.022	0.038	0.041

Comparison of ATE-RMSE(m)

ORB-SLAM2 (Mur-Artal and Tardós, 2017) is a real-time feature-based, RGB-D visual odometry and SLAM library that employs bundle adjustment to build globally consistent sparse reconstruction. ORB-SLAM2 utilizes the RGB-D sensors' depth information to generate the current frame's feature coordinates. ORB_SLAM2 prefers sparse construction of globally consistent trajectories instead of dense scene reconstruction and trajectories with full details. However, ORB_SLAM2 precise keyframe positions could efficiently generate accurate reconstruction calibrations by fusing the position values into depth maps.

ElasticFusion (Whelan et al., 2015) is also an RGB-D SLAM that captures RGB and depth information for dense reconstruction of the surrounding environment. However, instead of pose graph optimisation, ElasticFusion utilizes a surfel-based map-centric approach to achieve non-rigid deformation and loop-closing properties in the map.

Maskfusion (Runz et al., 2018) is one of the RGB-D SLAMs that can employ semantic information of objects to efficiently deal the non-rigid and dynamic scene situations. This method utilizes Mask-RCNN to detect, recognize, track and reconstruct the multiple moving objects and also a geometry-based segmentation method to increase the object boundaries in the object mask.

We utilized Absolute Trajectory Error (ATE) (Sturm et al., 2012) for evaluating the VO/SLAM models. ATE is the Root Mean Square Error (RMSE) distance between estimated trajectory values and ground truth trajectory values, and it can defined as below:

$$\mathrm{ATE}_{rmse}\left(F_{1:n}\right) := \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\mathrm{trans}\left(Q_i^{-1}SP_i\right)\right\|^2\right)^{1/2} \quad (1)$$

Here, $F_{1:n}$ represents the frames presented in the sequence from frame numbers 1 *to* n and $P_{1:n}$ and $Q_{1:n}$

represents the calibrated and ground truth trajectory values. S is the rigid body transformation to properly align the estimated and ground truth trajectory's coordinate systems.

As listed in table4. Our evaluation of the dataset on VO/SLAM algorithms delivers better results in occluded environments with static background scenes like a library, room, machine lab, and playground without human movements. However, it faces challenges during the evaluation of complex sequences like outdoor environments with dynamic backgrounds in low illumination conditions. ElasticFusion (Whelan et al., 2015) delivers a little lower performance than ORB-SLAM2 in dealing with occluded static environments; however, it delivers better results in situations like dynamic background scenes. However, delivers low accuracy in low illumination with minimal texture conditions. Maskfusion (Runz et al., 2018) also delivers similar results in static, occluded backgrounds, but in dynamic background scenes, it delivers better performance than ORBSLAM2 and Elastic_fusion. However, it also could not deliver better results in low illumination conditions.

4.2 Hand Gesture Classification on Continuous Data

The hand region occupies a partial part of the images, not a full image. So, training the model with segmented hand regions of interest will deliver better performance. In our dataset files, gestureinforamtion.txt contains the hand region of interest bounding box information and consecutive frame's time stamp along with gesture class. These two data columns enable the spatiotemporal relation of the content of the gestures. With our dataset, The gesture detection and classification tasks initially investigated the effect of the count of input frames. Table 5 shows the VOEDHgesture accuracy on state-of-the-art classification models. Secondly, we investigated the effect of RGB and RGB+DEPTH features on the performance of the model. We observed that RGB with depth maps delivers better performance than a single RGB model. We set the learning rate and the batch size as per the specification of Yifan's (Zhang et al., 2018) and used Jaccard index (Zhang et al., 2018) for evaluating the continuous hand gestures, and it is defined as follows.

$$J_s = \frac{1}{l_s} \sum_{i=1}^{L} \frac{G_{s,i} \cap P_{s,i}}{G_{s,i} \cup P_{s,i}} \tag{2}$$

Here, $G_{s,i}$ and $P_{s,i}$ represents the ground and estimated classes of i^{th} gestures label for the sequence s with l_s classes.

Table 5: Gesture Classification Evaluation on Continuous Data.

MODEL	INPUT	JACCARD	
		RGB	RGB+DEPTH
VGG-16+LSTM (Zhang et al., 2018)	1216	0.680	0.73
VGG-16+LSTM (Zhang et al., 2018)	1617	0.60	0.71
C3D+STTM (Zhang et al., 2018)	1216	0.820	0.910
C3D+STTM (Zhang et al., 2018)	1617	0.801	0.89
RESNeXT-101	1216	0.650	0.691
RESNeXT-101	1617	0.67	0.68
C3D+LSTM+RSTTM(Zhang et al., 2018)	1216	0.85	0.91
C3D+LSTM+RSTTM(Zhang et al., 2018)	1617	0.83	0.88

5 CONCLUSIONS

We introduced a novel benchmark dataset for evaluating both hand gesture recognition and RGB-D VO/ SLAM systems. The benchmark dataset provides RGB images, depth images, ground truth of the pose calibrations trajectory and spatiotemporal information of the hand region of interest. The novel dataset provides more diversified background scenes than the existing datasets, as it is collected in ten environments in three different illumination conditions. After evaluating our dataset on different conventions RGB-D SLAM ORB-SLAM2, ElasticFusion, and Maskfusion, we find that ORB_SLAM2 delivers better performance for RGB-D SLAM in static occluded results. ElasticFusion and Maskfusion have delivered better performance even in dynamic environments. However, these SLAM methods delivered low accuracy in low-illumination environments. We also investigated our dataset on hand gesture classification tasks. We obsessed that RGB images along with depth images, delivered better performance than RGB images alone. our dataset VOEDHgesture can help for further exploration of research works: 1) Investigation into spatiotemporal modeling is possible 2) hand gesture detection, recognization, and tracking is possible.

REFERENCES

Baraldi, L., Paci, F., Serra, G., Benini, L., and Cucchiara, R. (2014). Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 688–693.

Benitez-Garcia, G., Olivares-Mercado, J., Sanchez-Perez, G., and Yanai, K. (2021). Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4340–4347. IEEE.

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M. W., and Siegwart, R. (2016). The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163.

Carlevaris-Bianco, N., Ushani, A. K., and Eustice, R. M. (2015). University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 35(9):1023–1035.

Cortés, S., Solin, A., Rahtu, E., and Kannala, J. (2018). Advio: An authentic dataset for visual-inertial odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 419–434.

Engel, J., Usenko, V., and Cremers, D. (2016). A photometrically calibrated benchmark for monocular visual odometry. *arXiv preprint arXiv:1607.02555*.

Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H., Sigal, L., Argyros, A., Sminchisescu, C., et al. (2013). Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 365–368.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE.

Handa, A., Whelan, T., McDonald, J., and Davison, A. J. (2014). A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE.

Kim, T.-K., Wong, S.-F., and Cipolla, R. (2007). Tensor canonical correlation analysis for action classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Latif, G., Mohammad, N., Alghazo, J., AlKhalaf, R., and AlKhalaf, R. (2019). Arasl: Arabic alphabets sign language dataset. *Data in brief*, 23:103777.

Liu, L. and Shao, L. (2013). Learning discriminative representations from rgb-d video data. In *Twenty-third international joint conference on artificial intelligence*.

Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., and Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215.

Mur-Artal, R. and Tardós, J. D. (2017). Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262.

Nuzzi, C., Pasinetti, S., Pagani, R., Coffetti, G., and Sansoni, G. (2021). Hands: an rgb-d dataset of static hand-gestures for human-robot interaction. *Data in Brief*, 35:106791.

Pfrommer, B., Sanket, N., Daniilidis, K., and Cleveland, J. (2017). Penncosyvio: A challenging visual inertial odometry benchmark. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3847–3854.

Ramezani, M., Wang, Y., Camurri, M., Wisth, D., Mattamala, M., and Fallon, M. (2020). The newer college dataset: Handheld lidar, inertial and vision with ground truth. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4353–4360. IEEE.

Runz, M., Buffier, M., and Agapito, L. (2018). Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE.

Schubert, D., Goll, T., Demmel, N., Usenko, V., Stückler, J., and Cremers, D. (2018). The tum vi benchmark for evaluating visual-inertial odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1680–1687.

Smith, M., Baldwin, I., Churchill, W., Paul, R., and Newman, P. (2009). The new college vision and laser data set. *The International Journal of Robotics Research*, 28(5):595–599.

Starner, T., Weaver, J., and Pentland, A. (1998). Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1371–1375.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. (2012). A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE.

Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., and Scherer, S. (2020). Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE.

Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., and Davison, A. (2015). Elasticfusion: Dense slam without a pose graph. Robotics: Science and Systems.

Yan, H., Shan, Q., and Furukawa, Y. (2018). Ridi: Robust imu double integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 621–636.

Zhang, L., Camurri, M., and Fallon, M. (2021). Multi-camera lidar inertial extension to the newer college dataset. *arXiv preprint arXiv:2112.08854*.

Zhang, Y., Cao, C., Cheng, J., and Lu, H. (2018). Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050.