

## Pose estimation algorithm for mobile augmented reality based on inertial sensor fusion

Mir Suhail Alam<sup>1</sup>, Malik Arman Morshidi<sup>1</sup>, Teddy Surya Gunawan<sup>1,2</sup>, Rashidah Funke Olanrewaju<sup>1</sup>, Fatchul Arifin<sup>3</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia

<sup>2</sup>School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

<sup>3</sup>Department of Electronic and Informatics Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

### Article Info

#### Article history:

Received May 29, 2021

Revised Mar 16, 2022

Accepted Mar 30, 2022

#### Keywords:

3D virtual object

Augmented reality

Gyroscope sensor

Oriented FAST rotated BRIEF

Pose estimation

### ABSTRACT

Augmented reality (AR) applications have become increasingly ubiquitous as it integrates virtual information such as images, 3D objects, video, and more to the real world, which further enhances the real environment. Many researchers have investigated the augmentation of the 3D object on the digital screen. However, certain loopholes exist in the existing system while estimating the object's pose, making it inaccurate for mobile augmented reality (MAR) applications. Objects augmented in the current system have much jitter due to frame illumination changes, affecting the accuracy of vision-based pose estimation. This paper proposes to estimate the pose of an object by blending both vision-based techniques and micro electrical mechanical system (MEMS) sensor (gyroscope) to minimize the jitter problem in MAR. The algorithm used for feature detection and description is oriented FAST rotated BRIEF (ORB), whereas to evaluate the homography for pose estimation, random sample consensus (RANSAC) is used. Furthermore, gyroscope sensor data is incorporated with the vision-based pose estimation. We evaluated the performance of augmenting the 3D object using the techniques, vision-based, and incorporating the sensor data using the video data. After extensive experiments, the validity of the proposed method was superior to the existing vision-based pose estimation algorithms.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Malik Arman Morshidi

Department of Electrical and Computer Engineering, International Islamic University Malaysia

53100 Jalan Gombak, Kuala Lumpur, Malaysia

Email: mmalik@iium.edu.my

## 1. INTRODUCTION

Augmented reality (AR) is the technology that evolved from virtual reality (VR). VR revolves around the computer-generated environment. However, AR is a technology that combines the real world and computer-generated information. Azuma *et al.* [1] have described AR in a novel way that includes real-time object augmentation. In the actual world, virtual and physical items must be mathematically aligned. The principle of AR is to integrate virtual information such as 3D models, images, text, video, music, and much more to the real environment, which further enhances the real world [2]. In recent years, AR applications have become increasingly ubiquitous. AR technology is applied in a wide range of fields such as tourism, medical, logistics, entertainment, maintenance, and much more [3]–[5]. It plays a huge role in tourism, including transportation, food, accommodation, museum, and more. It was anticipated that business that uses AR applications would get an advantage to make more progress and capture the market [6].

Acquiring 3D pose estimation is the long-standing enigma in computer vision. 3D pose estimation is the process of anticipating an object's orientation or relative position from a defined reference. Many applications take advantage of pose estimation in recognition, localization, object grasping, and mapping. The extraction of features (key points) from a scene/object is the first step in estimating a pose, which is then utilized to determine the relative pose of the object concerning the coordinate frame. However, because the camera projection only includes 2D data from the 3D world, the accuracy of the scene or objects may only be assumed to a limited extent.

AR image registration allows augmenting or overlaying the virtual object into the real environment, allowing augmenting 3D objects using different computer vision techniques. However, 3D registration technology first determines the relationship between the virtual object and the orientation of the displaying device. Besides, the rendered object must be precisely augmented into the real scene to merge the virtual image and the model with the real environment [2], [7].

Vision-based pose estimation in AR applications has been widely explored. However, mobile augmented reality (MAR) applications have certain issues, such as jitter, illumination issue, and more. This paper aims to estimate the pose of an object by combining the vision-based techniques and micro electrical mechanical system (MEMS) sensor (gyroscope) to minimize the jitter issue in the pose estimation. First, the video is composed and then processed using the software. Next, the reference in the video is tracked accordingly. The reference in this scenario is our target in the video, where we intend to augment the virtual object. Although, many studies have been done regarding AR image registration which allows rendering the virtual object into the physical or real environment. However, when there is a change in orientation or illumination, the pose gets affected, disturbing the augmentation. To overcome this, we have recorded video, which comprises of series of movements like rotating, tilting, and different random motions, to check the robustness of our proposed algorithm. Secondly, oriented FAST rotated BRIEF (ORB) and random sample consensus (RANSAC) are used to perform feature extraction and homography, respectively. Finally, the augmentation part is initially done using a vision-based method only, and sensor data (gyroscope) is incorporated only when the extracted features are the threshold. The result significantly improves estimating pose and augmenting the 3D object by incorporating the gyroscopic sensor data.

The arrangement of the rest of the paper is: section 2 presents the related works, and section 3 presents the proposed method to incorporate the sensor data with vision-based pose estimation and augment the 3D object on a target surface. Section 3 further describes the feature detection and matching process, homography estimation, and gyroscope. Finally, section 4 shows the experimental results and discussion, followed by the conclusion in section 5.

## 2. RELATED WORKS

Augmenting virtual objects, whether 2D or 3D, on top of the physical layer is AR. The digital information (2D/3D objects) augmented in the real environment has immensely increased visualization technology. AR applications capture the images of the physical world and represent it with additional layers of data which is then displayed on the different digital screens [8]. Enhancing visual performance using AR technology has become the trend for big brands to capture the market. An approach made by [9] presents an AR framework that uses AR and shader effects. Shaders are scripts that include the mathematical computations and techniques needed to compute the color of each rendered pixel. The animated screen effects and light effects are achieved by shading. Furthermore, the techniques used in the framework make the AR scene more appealing and realistic by overlaying different virtual objects and making it more convenient to experience the 3D augmentation on their mobile devices. However, the systems cannot evaluate the pose estimation of the objects accurately.

Accomplishing the goal of inserting the virtual objects in an image sequence accurately where the 3D objects are rendered and aligned with the real environment is of great importance. Although AR has seamlessly allowed augmenting the 3D objects, the pose estimation or camera localization process issues still exist. In recent years, Marchand *et al.* [10] have presented a brief introduction to the different approaches related to vision-based pose estimation and camera localization issues. Additionally, the gap between practical implementation and theoretical aspects of pose estimation has been reduced. Eyjolfsson and Turk [11] proposed a multisensory method for estimating the transformation of mobile phones between images taken from its camera. The method used inertial sensors to support vision-based pose estimation by warping two images into the same perspective. Adaptive features from accelerated segment test (FAST) feature detectors and image patches are incorporated as key point detections. The results show a considerable improvement in matching the key point between two images. However, the study fails when there is a big transformation due to poor linear movement estimation.

With MEMS sensors becoming more accurate, the camera pose estimation problem turned down. An approach regarding the 3D camera rotation and translation which are the extrinsic parameters of camera

pose estimation, was made by [12], in which the vision data and inertial fusion using simplified structure from motion for pose estimation. A gyroscope sensor was used to estimate the camera rotation parameter, while the translation parameter was estimated separately. Moreover, the camera rotation parameter assists in estimating the translation parameter using image data. Nevertheless, the limitation of the technique lies in the drift problems of gyroscopes that need to be calibrated after a long time. In addition, a pose estimation algorithm using depth information was proposed by [13]. The results shown by the latter study were more accurate than algorithms using both depth information and color information when evaluated against many scenes.

Another research conducted by [14] proposed using integrated position sensors, universal serial bus (USB) position sensors, head-mounted display, and gyroscopic mouse in their City 3D AR project. The research is to be used in the field of architecture and urban planning. However, the project faces challenges in augmenting the 3D objects in the outdoor environment.

### 3. PROPOSED METHOD

This section describes the proposed solution for estimating the object's pose and its reference using vision-based pose estimation along with gyroscope sensor data to precisely augment the 3D virtual object in the real environment. The virtual object to be augmented must remain intact on the target surface without any jitter. So, the position and orientation of the 3D virtual object match the position and orientation of the predefined target surface. Moreover, if the surface on which the augmented 3D object changes its position and orientation, the 3D object also changes accordingly.

We incorporate the sensor data with vision-based pose estimation to reduce the jitter problem while augmenting the 3D object on the target surface to achieve our objective. Figure 1 indicates the procedure that begins with the video data acquisition followed by feature extraction of the target surface using the ORB algorithm and RANSAC to evaluate the homography for pose estimation of the target surface. After estimating the target's pose, the method proceeds towards the augmentation based on vision data alone. However, if the key points are the set threshold because of the significant change in the target's orientation, then we incorporate sensor data that defines the target position in three dimensions with the vision data. Thus, the vision and sensor data are aligned together to enhance the pose accuracy and improve the augmentation.

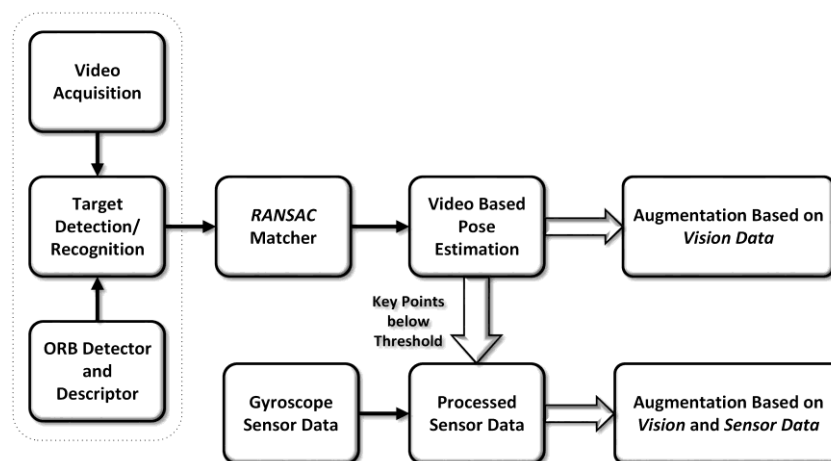


Figure 1. Overview of the stages associated with vision and sensor-based pose estimation

#### 3.1. Video data acquisition

A mobile phone is utilized to capture video data of the reference where the virtual item is to be augmented. Even though the study aims to use real-time video input, a pre-recorded video is used to evaluate several algorithms. It is because the location variables of each key point change from one real-time session to the next when using a real-time video. As a result, the outcomes are not consistent with each execution of the real-time test. However, each key point is precisely the same location with a pre-recorded video for each test iteration, resulting in reliable findings for each test. In addition, we have further elaborated on the resolution and other details in the experimental section.

### 3.2. Target surface recognition

The target detection/recognition begins after the video acquisition. The video consists of many frames and the target surface on which we augment the 3D object will also change. So, to augment the 3D virtual object, we will first detect the surface for virtual object then and only we can augment our virtual object precisely. The target recognition consists of several steps, which are discussed in the following subsections:

#### 3.2.1. Feature extraction

In computer vision and other related image processing applications, two critical tasks are feature detection and matching. Image processing applications are expanding in multiple fields daily. Image matching algorithms are used in a wide range of applications, from simple photogrammetric tasks like feature detection to the development of sophisticated 3D modeling tools and image search engines. The feature of an image is the valuable information that helps solve the computational problem in computer vision applications. Feature extraction is the process of retrieving valuable information (interest points) from the image that includes minimizing the size of information representing a large amount of data. Moreover, image retrieval systems are mostly based on color, shape, texture, and layout [15], [16]. Therefore, the image features should represent some uniqueness, such as corners and edges. Also, it should be scale-invariant or invariant to any transformation.

#### 3.2.2. Feature description

Descriptors provide the representation of the information acquired from the feature and surroundings. Descriptors encapsulate the feature vector of the object to be recognized, and the feature vector contains the descriptors of the interest points in the reference and target image. Many algorithms are present to extract the features from the image and compute its descriptors, such as speeded up robust features (SURF), scale-invariant feature transform (SIFT), ORB, and many more. The algorithm used in our project is ORB, as it provides better performance and minimal computational loads. ORB is an efficient feature detector and descriptor ideal for real-time situations where efficiency and speed are preferred [17], [18]. Also, it is free from any patent protection claims. Both the techniques of detection and description perform better and are low cost.

#### 3.2.3. Feature matching

The task of determining the correlation between the reference and the target image is known as feature matching. The easiest method of doing this is to take the descriptor of each element in the primary set, compute the distance to all the descriptors in the subsequent set and return the nearest one as the best match. However, the matching task depends on the variations within an image and the image type to be matched. Certain parameters need to examine while matching the image: i) scale: the scales of at least two items of the set of images views differ, ii) orientation: the views of the images are rotated concerning one another, iii) affine transformation: whether it be a flat, angular, or textured object, iv) illumination: variation in illumination also arises a typical issue for efficient feature matching, and v) occlusion: two spatially separated objects in the 3D environment might look like one or get interference in the 2D image plane [19]. Moreover, a threshold should be defined on the number of matches found, which further demonstrates the minimum key points matched with reference and gives efficient recognition.

### 3.3. Homography estimation

Once the reference surface is recognized in the current frame and has several legitimate matches, we can continue to appraise the homography between the two images. First, we need to discover the transformation that maps the points from the image plane to the surface plane. Then, the homography matrix equation is used while estimating the pose of the given image. The coordinates P0, P1, P2, and P3 shift an image from the viewer's perspective on an image plane and project it onto a world plane in a three-dimensional environment as shown in Figure 2 [20].

The transformation needs to be updated in every frame we process. Therefore, homography is anticipated to reflect a mapping from one picture plane to the next related to a rigid body transformation. Hence, it is expected that a rigid body keeps its shape during the acquisition of pictures and the transformation occurs only on the projected image surface if a change in camera view [21].

Using an existing algorithm to determine the homography would be easy as the reference and target image matches are found. The RANSAC algorithm can concurrently sort out the outliers based on the assessed model. It is an iterative technique of assessing the parameters of a numerical model from sample data containing both outliers and inliers. This algorithm works well in the presence of a large number of outliers. But determining accurate homography is a bit complicated, as we have to set the model so that there are minimum outliers. With the minimum outliers, the 3D object is augmented more accurately.

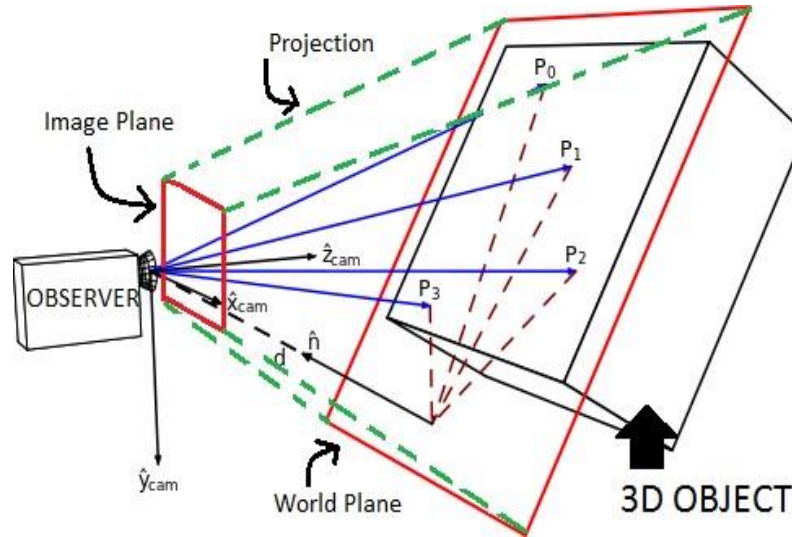


Figure 2. Homography evaluation for pose estimation

**3.4. Pose estimation in augmented reality**

AR has been essentially a multidisciplinary and old field. Although, it is obvious that real-world and virtual world registration problems have piqued people’s interest. But from a wider perspective, this is a motion tracking problem. Different sensors have been considered, such as magnetic, mechanical, inertial, global positioning system (GPS), ultrasonic devices, and more, but unfortunately, there was no silver bullet to mitigate this issue [22]. The method of estimating the camera’s position and orientation from a collection of correspondences between 3D features and their image plane projections is called pose estimation. As a result, any errors in the camera estimation in the global frame would be obvious to the user. As a result, vision-based AR is restricted to a camera pose estimation problem. Three angles of rotation and three angles of translation can be used to reflect a pose. However, at least three points can be used to approximate a 3D pose. Researchers have tried various techniques, including the P-n-P problem and simultaneous localization and mapping (SLAM), to estimate the pose based on the available data, either 3D or 2D. Although the P3P technique mitigates the pose estimation problem, it produces more reliable results by increasing the points.

**3.5. Inertial sensors**

Gyroscopes are mechanical gadgets that are used to measure the angular rate of rotation. MEMS gyroscope and magnetometers innovation currently give this function in various packages that are generally integrated into a sensor module or chip and broadly used in a variety of applications [23]. MEMS gyroscope utilizes a minuscule micromechanical framework on silicon structures, supporting the motion to electrical transducer functions. It is mostly used in the navigation system to deliver the heading estimation.

The gyroscope measures the angular velocity along three axes. As a result, it cannot predict roll, pitch, or yaw. However, as we can see, integrating angular velocity over time produces the angle, which can then be utilized to determine roll, pitch, and yaw changes even though gyroscope readings are sometimes erroneous due to fast motion. In general, we calculate roll and pitch using (1).

$$\phi = \arctan\left(\frac{-a_x}{a_y}\right); \theta = \arctan\left(\frac{a_z}{a_y}\right) \tag{1}$$

The (2)-(4) are used to calculate the orientation and to rotate the acceleration vector [11].

$$\begin{aligned} \theta &= \omega_y \cos \phi - \omega_x \sin \phi \\ \phi &= \omega_z + (\omega_y \sin \phi + \omega_x \cos \phi) \tan \theta \\ \psi &= (\omega_y \sin \phi + \omega_x \cos \phi) / \cos \theta \end{aligned} \tag{2}$$

$$C_L^I = \begin{bmatrix} \cos \psi \cos \theta & \sin \psi \cos \theta & -\sin \theta \\ -\sin \psi \cos \phi + \cos \psi \sin \theta \sin \phi & \cos \psi \cos \phi + \sin \psi \sin \theta \sin \phi & \cos \theta \sin \theta \\ \sin \psi \sin \theta + \cos \psi \sin \theta \cos \phi & -\cos \psi \sin \phi + \sin \psi \sin \theta \sin \phi & \cos \theta \cos \phi \end{bmatrix} \tag{3}$$

$$aI = C_L^I a_L - g \quad (4)$$

where  $C_L^I$  is the rotation matrix used to rotate the local acceleration vector;  $\theta$ ,  $\phi$ ,  $\psi$  depicts *pitch*, *roll*, and *yaw*, respectively. In the MAR application, gyroscopes can assist in recording the changes in the target's orientation to estimate the pose accurately and precisely augment the 3D object on a target surface. So, to enhance the pose estimation accuracy, the gyroscopic sensor can be incorporated with the vision-based method.

### 3. EXPERIMENTAL RESULTS AND DISCUSSION

This section shows the experimentation and the results obtained to estimate the pose and augment the 3D object in a real environment. The steps involved in the proposed algorithm include video data acquisition, keypoint detection and description, homography, pose estimation, and is followed by incorporating the gyroscope sensor data to enhance the efficiency of the augmentation, which can be obtained only by estimating the pose accurately. The tools used during the experimentation are matrix laboratory (MATLAB) and Python using OpenCV (a popular computer vision library). Moreover, the gyroscope sensor reading was recorded using an android physics toolbox sensor suite.

The video is recorded using an Android smartphone Xiaomi Mi A1 that has a 20-megapixel camera. The resolution of the video is 1080×1920 pixels while recording the video, and the duration of the video is 34 seconds with a rate of 30 frames per second which results in 1020 frames. However, we have used plenty of videos with different video lengths during experimentation, but eventually, we selected the video of 34 seconds for the results. In addition, the video comprises a series of camera movements like tilting, rotating, and random motions to show the system's robustness.

#### 4.1. Quantitative analysis

##### 4.1.1. Ground truth analysis

The key points that appear in all the video frames are the ground truth of our method. To process the video and extract all the frames, we used MATLAB. After extracting the frames, we used the ORB algorithm to obtain all the frames key points. The ORB algorithm is a hybrid of a modified FAST detector and a modified binary robust independent elementary features (BRIEF) descriptor. FAST detects key points by scanning the pixel along with its neighboring pixel  $p$  within the radius  $r$ . The new pixel  $p$  detected as a key point is determined by the surrounding pixels within the radius  $r$ . If their intensity differs significantly from that of the candidate pixel  $p$ , only the new key-point is detected. Figures 3(a) and 3(b) depicts the key points (green dots) of our target surface and strong 5 key points in each frame of the recorded video respectively. Since each frame is only 2D after extraction of the frames from the video, the position of the key points is only denoted in  $x$  and  $y$  coordinates. Figure 4 shows the ground truth which encapsulates all the keypoints in the frames.

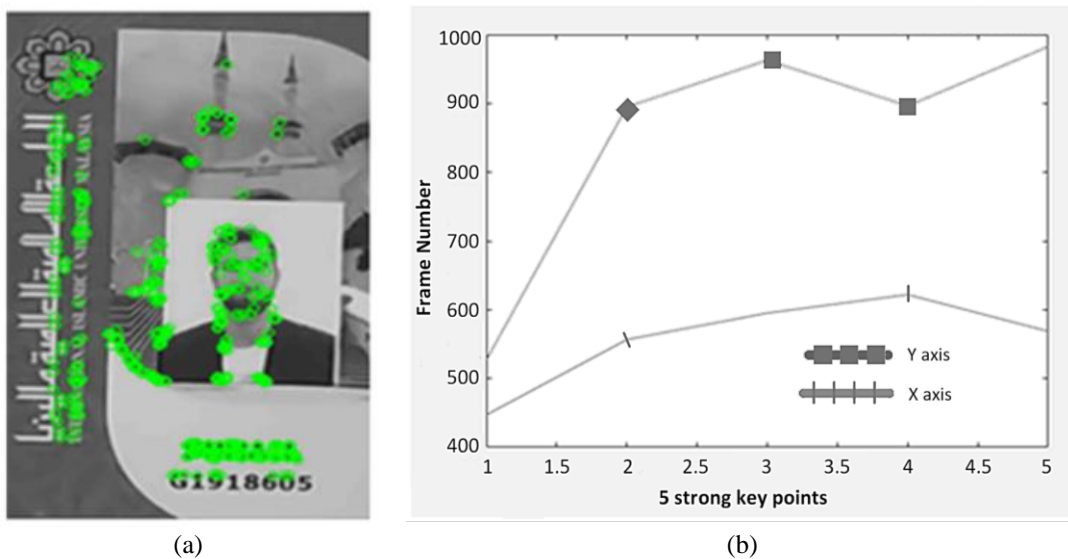


Figure 3. The key points (a) key points of the target and (b) 5 strong key points in each frame

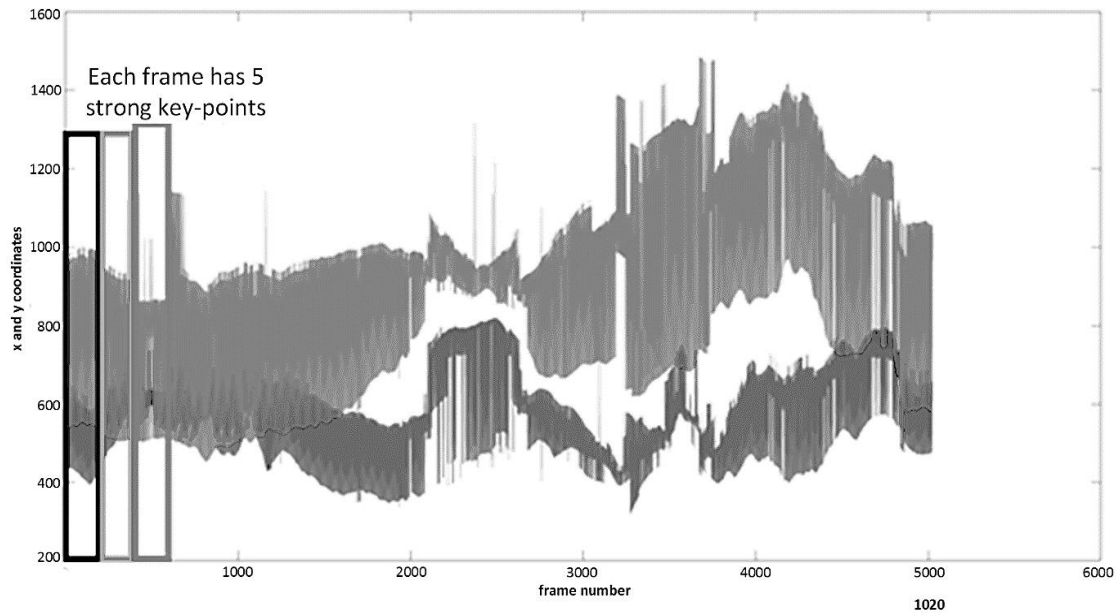


Figure 4. Ground truth of all video frames

#### 4.1.2. Vision-based analysis

The vision-based approach for determining the mobile device or static object pose estimation is widely used. The vision can be taken as a video or an image to interpret the environment. It represents a spatial connection between the captured 2D image and the 3D points on the scene. The use of AR markers makes the alignment of the virtual object with real environments very efficient way. Moreover, it improves stability and reduces requirements for computation. Furthermore, while analyzing the other vision-based approaches such as [24], [25], they have used enhanced techniques. However, the efficiency and robustness of a vision-based method depend on the performance of the feature extractor, i.e., inadequate feature extraction in images results in failure of the pose estimation. So, robust pose estimation based on vision is yet to be achieved. However, the incorporation of sensors such as gyroscopes, accelerometers with vision-based techniques has minimized this issue.

The recognition of objects and feature matching in AR is vital under uncontrolled, real-world conditions. It is important to create object-based environmental representations and to manipulate objects. Object recognition means recognizing a particular object or reference (e.g., ID card) in our work. Numerous approaches are presently being used to detect, recognize, and classify objects with scale-invariant descriptors and detectors. Among the algorithms available include ORB, SIFT [26], SURF [27], Gaussian of difference (DOG), and many more. Object detection and acknowledgment can be done by computer vision, which detects an object in an image or video. The object recognized is used to identify an object position or a scene [28], [29]. Object recognition is based on some criteria that include appearance, feature, and color based. However, each algorithm has its advantages and disadvantages.

The algorithm used in our experiment is ORB, being the better feature detector and descriptor and suitable for real-time situations that favor efficiency and speed. In the vision-based system, the object's points of interest in the image match it with the reference in a similar scene or image. After extracting the features, we have to appraise the homography between the frames. RANSAC [30] is used here because it can calculate parameters with high precision even if it includes a significant number of outliers. Figures 5(a) and 5(b) showed the feature matching of the frame and homography estimation using RANSAC, respectively.

Figure 6 shows the frames concerning the matching key points. The vision-based method has good matching points until there is illumination and tilt in our reference. We have set a threshold for key points. Until the key points are above the threshold, our augmentation based on vision works well. When the matching key points get reduced due to the tilt or illumination in the target, it cannot estimate the pose well, affecting the augmentation. Frames from 400 to 700 have interest points the set threshold because of the target's random motions and orientation changes. Furthermore, the decrease in key points affects pose estimation and augmentation of a 3D object. To overcome this issue, we have used inertial sensors to incorporate the gyroscope data that can help in mitigating the problem. Figure 7 shows that vision-based pose estimation works well when key points are above the threshold, which helps in 3D augmentation. However,



we incorporate the sensor data with vision data to enhance pose estimation accuracy when the key points are the threshold.

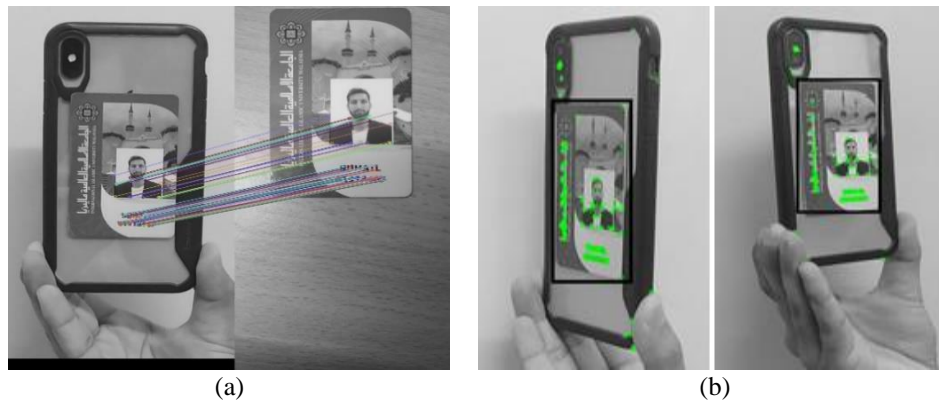


Figure 5. Feature matching and homography (a) frame feature matching and (b) homography using RANSAC

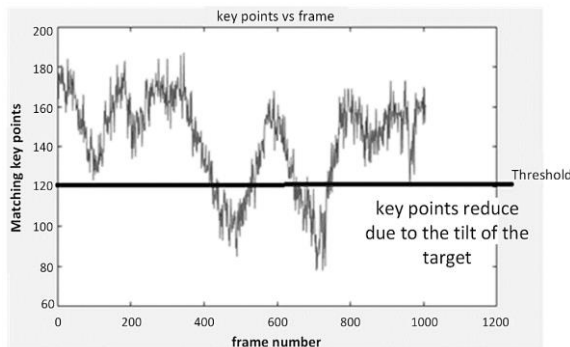


Figure 6. Matching key points versus frame number

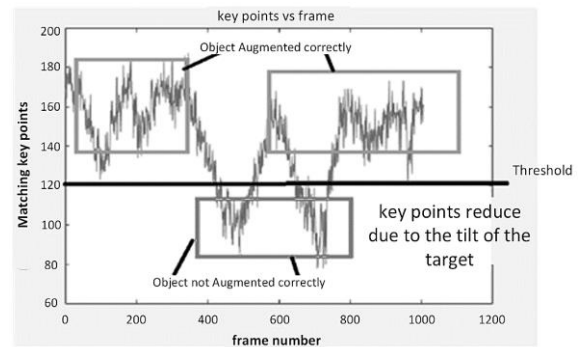


Figure 7. Matching key points above and below the threshold

#### 4.2. Sensor fusion

Sensor fusion aims to improve pose estimation performance by combining and integrating vision data with the sensor data. It is not possible to provide exact details using an inertial measurement unit (IMU) sensor alone. Hence vision is often used. However, vision data alone cannot accommodate occlusion, rapid movement due to the camera's scope. Thus, the fusion of both IMU (gyroscope data) and vision data would better estimate poses with the deficiency in the vision-based method [31].

Different approaches based on sensor fusion have been made by [32], [33]. Assa and Janabi-Sharifi [32] has proposed an extended Kalman filter (EKF) based sensor fusion approach for pose estimation where multiple camera measurements were fused using a combination two-camera configuration. Assa and Janabi-Sharifi [33] proposed the new techniques of multi-camera sensor fusion based on virtual visual serving for efficient and reliable pose estimation. However, the methods primarily concentrate on the fusion of pose estimation, which has yet to be implemented in visual guidance applications such as grasping. Furthermore, the sensor fusion techniques assume that the target is in the intersection field of view (FOV) of all the cameras [34], [35]. Another main issue in using EKF based multi-camera method is the increase in computational cost.

In our experimental method, we mounted the target on the mobile phone and recorded the inertial sensor data (gyroscope) using that phone, and we have placed the camera in front of our target. A gyroscope senses the changes in the target's position and measures these angular rates of rotation. Built-in gyroscopes in our handled devices help us measure these readings. Our application is then recording it for further analysis. It assists in enhancing the accuracy of the estimated pose. Figure 8 shows a graphical representation of gyroscope data in the x, y, and z-axis separately.



Figures 8 and 9 shows the gyroscope data (x, y, and z axis) independently and combined respectively. In addition, Figure 9 indicates that when there is a big orientation change of our target, the key points of our frames also get reduced. Thus, affecting the accuracy of the estimated pose and augmentation as well. Our proposed algorithm uses this sensor data and the vision data to enhance the accuracy of the estimated pose. When the key points get reduced the set threshold due to illumination or orientation changes, we fuse the sensor data with vision data to improve pose estimation and augmentation accuracy.

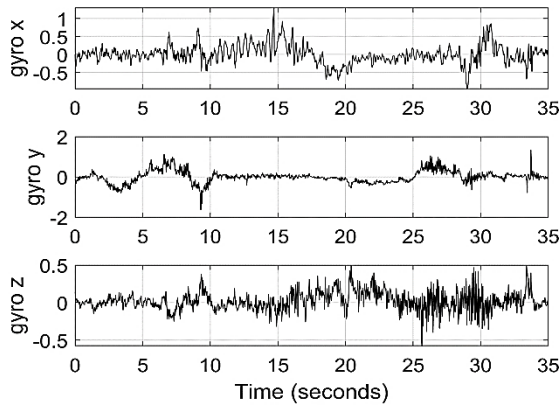


Figure 8. Sensor data of x-axis, y-axis, z-axis

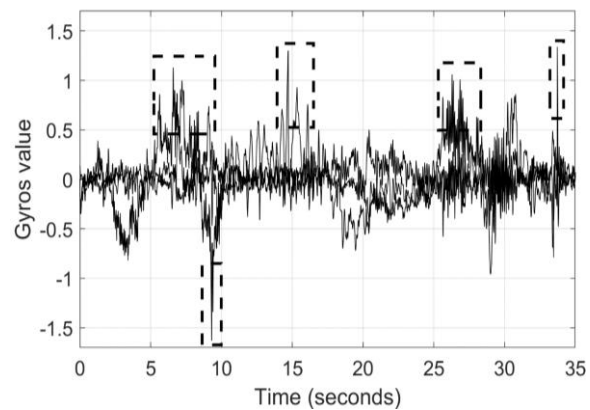


Figure 9. Combined gyroscopic data

#### 4.3. Qualitative analysis

We found significant improvement in pose estimation performance after applying the proposed algorithm based on sensor fusion (gyroscope data) and vision data. However, there were some good results while using vision data when the target's orientation did not tilt that much. Nevertheless, it could not estimate the pose accurately when there is a significant change in the target's orientation. Furthermore, due to a significant change in the target's position, many key points were not extracted, eventually affecting the augmentation.

Figure 10 shows frames from 431-433, which has a significant change in the target, causing jitter in the estimated pose and reducing the key points. Therefore, the performance in these frames degraded even though the target is visible and in the frame. Nonetheless, we used the gyroscopic data, which measures the angular rate of motion as mentioned earlier. Since the target could not estimate the pose properly due to a major change in the orientation, we fused the sensor data with vision data, we observed a significant improvement in the performance of the results, as shown in Figure 11. When the key points get reduced the defined threshold, the sensor data aids the vision data in figuring out the pose and simultaneously helps in augmentation. Figure 10 shows the same frame as in Figure 11, but the latter is based on vision data, and the other is after incorporating the sensor data with it.

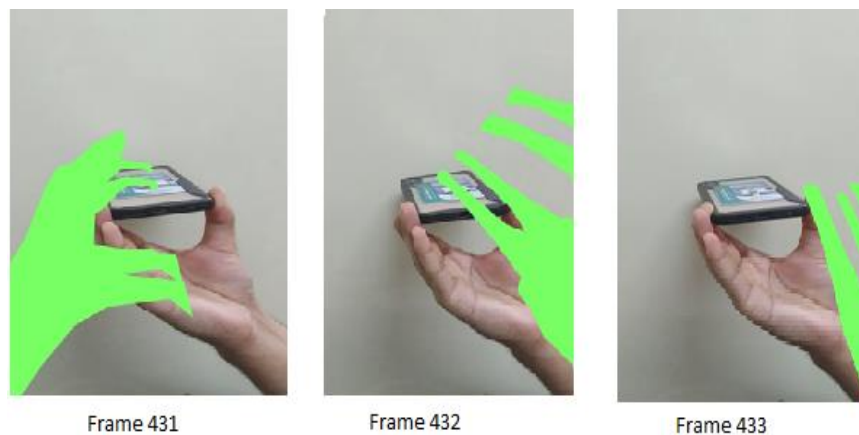


Figure 10. 3D augmentation frames based on vision data

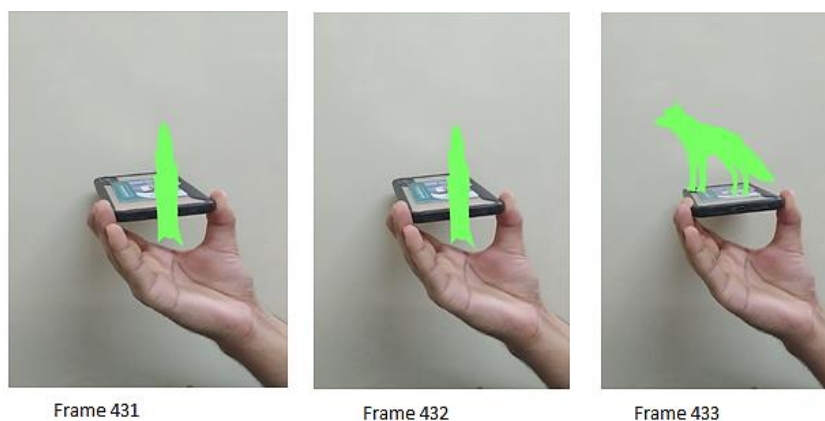


Figure 11. 3D augmentation frames based on vision data and sensor fusion

#### 4. CONCLUSION AND FUTURE WORKS

A unique method of pose estimation based on vision data incorporating sensor data (gyroscope) and 3D augmentation was introduced with low computational cost. Augmentation of the virtual 3D object in the real environment is greatly significant due to the increased demand for AR in every field, such as medical, tourism, education, and more. This paper presented an algorithm that incorporates sensor data with vision-based pose estimation to reduce the jitter issue due to illumination or change in orientation in the frame sequence while estimating the pose. We used the ORB algorithm for the feature detection and description process because of its minimal computational loads and is ideal for real-time situations. The RANSAC algorithm determined the homography of the reference and target image, which is robust to sorting out the inliers and outliers. The performance of vision-based pose estimation was compared with and without gyroscope sensor data. The experimental results show better performance using vision-based pose estimation with sensor fusion. Finally, the virtual 3D object was augmented on the predefined surface whose rotation changes continuously, the issue of change in the orientation and jitter was minimized. In future work, the limitations of the sensors will be addressed, such as the effect of drift on gyroscope and poor accuracy in magnetometer and accelerometers under fast motion.

#### ACKNOWLEDGEMENTS

This research was supported by the Ministry of Education Malaysia (MOE) through Fundamental Research Grant Scheme (FRGS) (Ministry Project ID: FRGS/1/2018/ICT01/UIAM/02/1) under Grant (FRGS19-068-0676). The authors would also like to express their gratitude to the International Islamic University Malaysia, University of New South Wales, and Universitas Negeri Yogyakarta.




#### REFERENCES

- [1] R. Azuma, Y. Baillet, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, "Recent advances in augmented reality," *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34–47, 2001, doi: 10.1109/38.963459.
- [2] Y. Chen, Q. Wang, H. Chen, X. Song, H. Tang, and M. Tian, "An overview of augmented reality technology," *Journal of Physics: Conference Series*, vol. 1237, no. 2, p. 22082, Jun. 2019, doi: 10.1088/1742-6596/1237/2/022082.
- [3] M. E. C. Santos, J. Polvi, T. Taketomi, G. Yamamoto, C. Sandor, and H. Kato, "Toward standard usability questionnaires for handheld augmented reality," *IEEE Computer Graphics and Applications*, vol. 35, no. 5, pp. 66–75, 2015, doi: 10.1109/MCG.2015.94.
- [4] G. Westerfield, A. Mitrovic, and M. Billingham, "Intelligent augmented reality training for motherboard assembly," *Int. Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 157–172, 2015, doi: 10.1007/s40593-014-0032-x.
- [5] K. C. Brata and D. Liang, "Comparative study of user experience on mobile pedestrian navigation between digital map interface and location-based augmented reality," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 2, pp. 2037–2044, Apr. 2020, doi: 10.11591/ijece.v10i2.pp2037-2044.
- [6] E. Özkul and S. T. Kuşlu, "Augmented reality applications in tourism," *International Journal of Contemporary Tourism Research*, pp. 107–122, Dec. 2019, doi: 10.30625/ijctr.625192.
- [7] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic, "Augmented reality technologies, systems and applications," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 341–377, Jan. 2011, doi: 10.1007/s11042-010-0660-6.
- [8] P. Dalsgaard and K. Halskov, "3d projection on physical objects," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 2011, pp. 1041–1050, doi: 10.1145/1978942.1979097.
- [9] H. Pai, "An imitation of 3D projection mapping using augmented reality and shader effects," in *2016 International Conference on Applied System Innovation (ICASI)*, May 2016, pp. 1–4, doi: 10.1109/ICASI.2016.7539879.
- [10] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: a hands-on survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 12, pp. 2633–2651, Dec. 2016, doi: 10.1109/TVCG.2015.2513408.




- [11] E. Eyjolfssdottir and M. Turk, "Multisensory embedded pose estimation," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, Jan. 2011, pp. 23–30, doi: 10.1109/WACV.2011.5711479.
- [12] S. Wei, Z. He, and W. Xie, "Relative pose estimation algorithm with gyroscope sensor," *Journal of Sensors*, vol. 2016, pp. 1–8, 2016, doi: 10.1155/2016/8923587.
- [13] M. Li and K. Hashimoto, "Accurate object pose estimation using depth only," *Sensors*, vol. 18, no. 4, Mar. 2018, doi: 10.3390/s18041045.
- [14] A. Cirulis and K. B. Brigmanis, "3D outdoor augmented reality for architecture and urban planning," *Procedia Computer Science*, vol. 25, pp. 71–79, 2013, doi: 10.1016/j.procs.2013.11.009.
- [15] S. M. Lajevardi and Z. M. Hussain, "Automatic facial expression recognition: feature extraction and selection," *Signal, Image and Video Processing*, vol. 6, no. 1, pp. 159–169, Mar. 2012, doi: 10.1007/s11760-010-0177-5.
- [16] M. Narottambhai and P. Tandel, "A survey on feature extraction techniques for shape based object recognition," *International Journal of Computer Applications*, vol. 137, no. 6, pp. 16–20, Mar. 2016, doi: 10.5120/ijca2016908782.
- [17] A.-K. Al-Tamimi, A. Qasaimeh, and K. Qaddoum, "Offline signature recognition system using oriented FAST and rotated BRIEF," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, pp. 4095–4103, Oct. 2021, doi: 10.11591/ijece.v11i5.pp4095-4103.
- [18] P. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proceedings of the British Machine Vision Conference 2013*, 2013, pp. 13.1–13.11, doi: 10.5244/C.27.13.
- [19] M. M. El-gayar, H. Soliman, and N. Meky, "A comparative study of image low level feature extraction algorithms," *Egyptian Informatics Journal*, vol. 14, no. 2, pp. 175–181, Jul. 2013, doi: 10.1016/j.eij.2013.06.003.
- [20] O. Chum, T. Pajdla, and P. Sturm, "The geometric error for homographies," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 86–102, Jan. 2005, doi: 10.1016/j.cviu.2004.03.004.
- [21] D. Monnin, E. Bieber, G. Schmitt, and A. Schneider, "An effective rigidity constraint for improving RANSAC in homography estimation," in *Advanced Concepts for Intelligent Vision Systems*, Springer Berlin Heidelberg, 2010, pp. 203–214.
- [22] G. Welch and E. Foxlin, "Motion tracking: no silver bullet, but a respectable arsenal," *IEEE Computer Graphics and Applications*, vol. 22, no. 6, pp. 24–38, Nov. 2002, doi: 10.1109/MCG.2002.1046626.
- [23] P. Kumari, K. Singh, and A. Singal, "Reducing the hygroscopic swelling in MEMS sensor using different mold materials," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 494–499, Feb. 2020, doi: 10.11591/ijece.v10i1.pp494-499.
- [24] H. Yu, Q. Fu, Z. Yang, L. Tan, W. Sun, and M. Sun, "Robust robot pose estimation for challenging scenes with an RGB-D camera," *IEEE Sensors Journal*, vol. 19, no. 6, pp. 2217–2229, Mar. 2019, doi: 10.1109/JSEN.2018.2884321.
- [25] X. Zhang, Z. Jiang, H. Zhang, and Q. Wei, "Vision-based pose estimation for textureless space objects by contour points matching," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 5, pp. 2342–2355, Oct. 2018, doi: 10.1109/TAES.2018.2815879.
- [26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
- [27] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008, doi: 10.1016/j.cviu.2007.09.014.
- [28] P. Loncomilla, J. Ruiz-del-Solar, and L. Martínez, "Object recognition using local invariant features for robotic applications: A survey," *Pattern Recognition*, vol. 60, pp. 499–514, Dec. 2016, doi: 10.1016/j.patcog.2016.05.021.
- [29] T. S. Gunawan, I. Z. Yaacob, M. Kartiwi, N. Ismail, N. F. Za'bah, and H. Mansor, "Artificial neural network based fast edge detection algorithm for MRI medical images," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 7, no. 1, pp. 123–130, Jul. 2017, doi: 10.11591/ijeecs.v7.i1.pp123-130.
- [30] M. A. Fischler and R. C. Bolles, "Random sample consensus," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, doi: 10.1145/358669.358692.
- [31] M. Alatise and G. Hancke, "Pose estimation of a mobile robot based on fusion of IMU data and vision data using an extended kalman filter," *Sensors*, vol. 17, no. 10, Sep. 2017, doi: 10.3390/s17102164.
- [32] A. Assa and F. Janabi-Sharifi, "A robust vision-based sensor fusion approach for real-time pose estimation," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 217–227, Feb. 2014, doi: 10.1109/TCYB.2013.2252339.
- [33] A. Assa and F. Janabi-Sharifi, "Virtual visual servoing for multicamera pose estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 20, no. 2, pp. 789–798, Apr. 2015, doi: 10.1109/TMECH.2014.2305916.
- [34] W. Liu, J. Hu, and W. Wang, "A novel camera fusion method based on switching scheme and occlusion-aware object detection for real-time robotic grasping," *Journal of Intelligent and Robotic Systems*, vol. 100, no. 3–4, pp. 791–808, Dec. 2020, doi: 10.1007/s10846-020-01236-7.
- [35] G. A. Cardona, J. Ramirez-Rugeles, E. Mojica-Nava, and J. M. Calderon, "Visual victim detection and quadrotor-swarm coordination control in search and rescue environment," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2079–2089, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2079-2089.

## BIOGRAPHIES OF AUTHORS






**Mir Suhail Alam**    received his Bachelor's degree in Information Technology from The University of Agriculture Peshawar, Pakistan, and a M.Sc. degree in Computer and Information Engineering from International Islamic University Malaysia (IIUM), Malaysia. He worked as a Graduate Research Assistant with the Department of Electrical and Computer Engineering, IIUM. His research interests include machine learning, artificial intelligence, augmented reality, and computer vision. He can be contacted at email: suhialam@gmail.com.






**Malik Arman Morshidi**    graduated from Western Michigan University, the USA, in 1999 with BSc in Computer Science. Upon graduation, he started his career as System Engineer at MacroHard AUM Sdn Bhd. Later, he joined Office Equipment and Communication Sdn Bhd (OEC) in 2000 as a System Analyst where he worked directly under the supervision of Software Team, Unit Bisnes Fasiliti (UBF), Tenaga Nasional Berhad (TNB), Petaling Jaya. In 2001 he joined Irhamna IT Sdn Bhd (IIT) as Chief Software Engineer. During his tenure in both OEC and IIT (where both are TNB vendors), he was responsible for developing and managing many software development projects for TNB and was the recipient of the Certificate of Excellence for this achievement. He joined the academic profession in 2003 as Assistant Lecturer at International Islamic University Malaysia (IIUM) in the Faculty of Engineering. He completed his MSc in Computer Systems Engineering from Universiti Putra Malaysia in 2007, whose research was on the vision and navigation system for an autonomous mobile robot for plant watering. In 2013, he completed his Ph.D. in Engineering specializing in Image Processing and Augmented reality from the University of Warwick, UK. He can be contacted at email: mmalik@iium.edu.my.






**Teddy Surya Gunawan**    received his BEng degree in Electrical Engineering with cum laude award from Institut Teknologi Bandung (ITB), Indonesia, in 1998. He obtained his M.Eng degree in 2001 from the School of Computer Engineering at Nanyang Technological University, Singapore, and a Ph.D. degree in 2007 from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia. His research interests are speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award in 2018 from IIUM. He is currently an IEEE Senior Member (since 2012), was chairman of IEEE Instrumentation and Measurement Society-Malaysia Section (2013 and 2014), Professor (since 2019), Head of Department (2015-2016) at Department of Electrical and Computer Engineering, and Head of Programme Accreditation and Quality Assurance for Faculty of Engineering (2017-2018), International Islamic University Malaysia. He has been a Chartered Engineer (IET, UK) since 2016 and Insinyur Profesional Utama (PII, Indonesia) since 2021, a registered ASEAN engineer since 2018, and ASEAN Chartered Professional Engineer since 2020. He can be contacted at email: tsgunawan@iium.edu.my.



**Rashidah Funke Olanrewaju**    (Senior Member, IEEE) was born in Kaduna, Nigeria. She received the B.Sc. degree (Hons.) in software engineering from University Putra Malaysia in 2002, and the M.Sc. and Ph.D. degrees in computer and information engineering from the International Islamic University Malaysia (IIUM) Kuala Lumpur, in 2007 and 2011, respectively. She is currently an Associate Professor with the Department of Electrical and Computer Engineering, IIUM, leading the Software Engineering Research Group (SERG). She is an Executive Committee Member of technical associations, like IEEE Women in Engineering and Arab Research Institute of Science and Engineers. In addition, she represents her university, IIUM, at the Malaysian Society for Cryptology Research. Her current research interests include MapReduce optimization techniques, compromising secure authentication and authorization mechanisms, secure routing for ad-hoc networks, and formulating bio-inspired optimization techniques. She can be contacted at email: frashidah@iium.edu.my.



**Fatchul Arifin**    received a B.Sc. in Electric Engineering at Universitas Diponegoro in 1996. Afterward, he received a Master's degree from the Department of Electrical Engineering, Institut Teknologi Bandung (ITB) in 2003 and a Doctoral degree in Electric Engineering from Institut Teknologi Surabaya in 2014. Currently, he is the lecturer at both the undergraduate and postgraduate electronic and informatic programs of the Engineering Faculty at Universitas Negeri Yogyakarta. His research interests include but are not limited to artificial intelligent systems, machine learning, fuzzy logic, and biomedical engineering. He can be contacted at email: fatchul@uny.ac.id.