Economical Quaternion Extraction from a Human Skeletal Pose Estimate using 2-D Cameras

Sriram Radhakrishna
Department of Computer Science and Enginering
P.E.S University, Banashankari
Bangalore, India
sriram.radhakrishna42@gmail.com

Adithya Balasubramanyam
Department of Computer Science and Engineering
P.E.S University, Banashankari
Bangalore, India
adithyab@pes.edu

***Abstract*—In this paper, we present a novel algorithm to extract a quaternion from a two dimensional camera frame for estimating a contained human skeletal pose. The problem of pose estimation is usually tackled through the usage of stereo cameras and intertial measurement units for obtaining depth and euclidean distance for measurement of points in 3D space. However, the usage of these devices comes with a high signal processing latency as well as a significant monetary cost. By making use of MediaPipe, a framework for building perception pipelines for human pose estimation, the proposed algorithm extracts a quaternion from a 2-D frame capturing an image of a human object at a sub-fifty millisecond latency while also being capable of deployment at edges with a single camera frame and a generally low computational resource availability, especially for use cases involving last-minute detection and reaction by autonomous robots. The algorithm seeks to bypass the funding barrier and improve accessibility for robotics researchers involved in designing control systems.**

***Index Terms*—quaternions, 2-D camera, pose estimation, MediaPipe, low-power, low-latency, embedded computer vision.**

I. Introduction

Essential scene-analysis tasks such as pedestrian detection and localization generally involve the generation of a quaternion to estimate the orientation of a target object. At times, these techniques involve the usage of stereo cameras and inertial measurement units to match feature points [1] or other such methods involving expensive hardware components with a relatively large latency and computational resource utilization.

Intuitively speaking, generalizing the potential use cases, keeping the computation closer to the system and eliminating the expensive hardware involved in the deployed algorithm would be the way forward to tackle the cost and latency issues. In order to do this, the hardware reliant inputs being accepted by existing pose estimation models must be taken note of. For this particular use case of human pose estimation, the depth of the points located within a targeted pose object and the simultaneous threading of two image streams at minimum stands out. In order to address these issues, a system containing a single 2-D camera with a human subject in frame on the hardware side to minimize costs was opted for. Additionally, a specialized deep learning based solution on the software side was implemented to account for the loss of depth perception as well as maintain latency on edge devices with a lower resource availability. The goal is to mimic the quaternion outputs provided by modern inertial measurement units for any given edge between adjacent joints on the human skeletal pose [2]

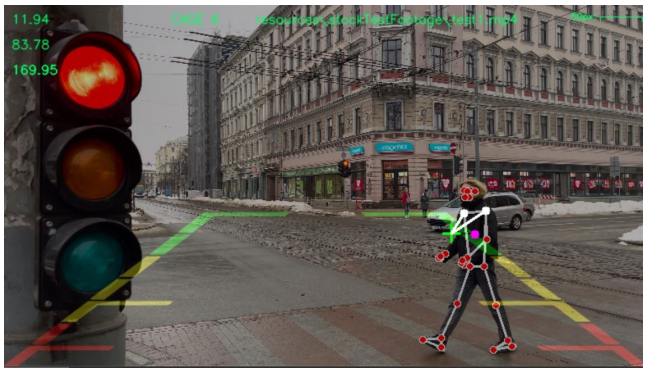

Fig. 1. A demonstration of the novel algorithm for the use case of calculating the angle of orientation of a pedestrian object from a generated quaternion.

II. Related work

Pavvlo et al. in their efforts to develop a quaternion based recurrent model for human motion, made the observation that human motion is a stochastic sequential process with a high level of intrinsic uncertainty [3]. While deep learning based approaches have been very successful in predicting the pose of a human skeleton in both short term [4] [5] and long term [6] applications, they require multi-threaded computations and are generally costlier to implement on hardware. This is especially true for use cases in fields like mobile robotics and autonomous vehicle navigation where the cost factor is usually compromised on in favour of larger core counts on the main edge processing unit.

Although this trade-off has recently found more justifications with the dropping retail costs of Nvidia GPU based

The authors are with the Center for Internet of Things, Department of Computer Science and Engineering, PESU Ring Road Campus (Bangalore, India).
Our code repository can be found at : https://github.com/SR42-dev/human-pose-quaternion-extraction

edge computing units [7], the gap is best bridged by adapting CPU based systems to more efficiently handle the calculations necessary. It was accurately noted by Eberly, D. that rotation matrices and quaternions take over 61 percent fewer calculations to obtain than angle-axis representations of vector rotating operations [8].

III. Concept Theory and Implementation Methodology

The following section introduces the thought processes, assumptions and applied quaternion mathematics behind the novel algorithm.

A. Overview

The challenges faced in this approach lie in the implementation of algorithmic solutions to finding the inputs mentioned in the introductory paragraph. First, the problem of calculating the orientation of the human pose object came with the issue of localization of body landmarks with acceptable standards of latency. Following this, was the conception of a mathematical function that extracted the quaternion from the pose data of these points.

To solve the problem of estimating the orientation, we employed MediaPipe, a framework by Alphabet Inc. that provides deployed solutions from deep learning models trained on data for human pose object detection. This package is aimed at edge devices with a low resource utilization. [9]. While the framework achieves admirable detection latencies, an inherent flaw in the system when porting to use cases requiring quaternion generation like robotic navigation [10] and such due to its basis in 3-D cartesian space. An added advantage provided by the novel algorithm is that portability of the model is maintained even when implemented on dynamic frames of reference, e.g.; on a mobile robot in motion.

B. Quaternion Transformations

Before we get to the equations that constitute the novel algorithm, it is important to gain some clarity on why a quaternion is necessary for this use case. We can visualize this by taking the example of a pedestrian obstruction to an autonomous vehicle or robot. For situations requiring a quick response from the robot to a major last-minute change in scene, especially ones which could endanger human lives such as humans crossing the path of a moving vehicle

To provide some context here, the quaternions can be calculated for any adjacent pairs of joints in the human pose, due to which we have narrowed down a use case for the algorithm to the problem of pedestrian orientation detection and modeled our implementation based on that. Hence, the two points chosen to evaluate the pose of our target human are the shoulder points as their relative positions are an accurate descriptor of the orientation of said target [11]. The same equations and thought processes can be used to extract this data for any pair of joints.

To narrow down this choice, all the provided options in the MediaPipe pose solutions documentation were evaluated [12]

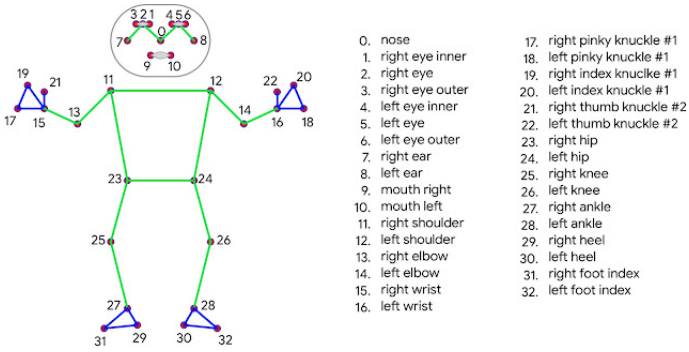

Fig. 2. BlazePose 33 keypoint topology for a human skeleton as COCO (colored with green) superset [12]

for appropriate pairs of points whose projections are sure to change location on a 2-D frame when the target's roll, pitch or yaw changes. Additionally, it was also arbitrarily determined that symmetry must be maintained for these points across the mid-line of the body as this seemed like the intuitive guideline to maintain given the potential use cases of our implementation.

After obtaining the coordinates for these points from the cartesian space estimated by the framework, a rotation matrix for the skeleton was generated by taking the coordinates of the two shoulder points mentioned before. Let us denote the points for the left and right shoulder points on the frame as (xl, yl, zl) and (xr, yr, zr) respectively. Therefore, the Z-axis vector of the rotation matrix can be calculated using the difference between the two points as

$$\vec{z} = \begin{bmatrix} xl & yl & zl \end{bmatrix} - \begin{bmatrix} xr & yr & zr \end{bmatrix} \tag{1}$$

$$\widehat{z} = \begin{cases} \frac{\vec{z}}{|\vec{z}|}; \Delta z \neq 0 \\ \begin{bmatrix} 0 & -1 & 0 \end{bmatrix}; \Delta z = 0 \end{cases} \tag{2}$$

Similarly, the X and Y axis vectors are generated as such -

$$\widehat{x} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \times \widehat{z}; iff \, \widehat{x} \neq 0 \, else \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \tag{3}$$

$$\widehat{y} = \widehat{z} \times \widehat{x} \tag{4}$$

... which allows us to derive our rotation matrix [13] using the camera look-at method [14] as

$$R_{3,3} = \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} \widehat{x} & \widehat{y} & \widehat{z} \end{bmatrix} \tag{5}$$

Subsequently, the quaternion $\mathbf{Q}$ was obtained in a standard form from this rotation matrix [13] as

$$\mathbf{Q} = a + bi + cj + dk \tag{6}$$

... where the coefficient values can be mapped according to equations 7 through 10 [15].

$$a = \frac{1}{2}\sqrt{|1 + r_{00} + r_{11} + r_{22}|} \tag{7}$$

$$b = \frac{r_{21} - r_{12}}{4a} \tag{8}$$

$$c = \frac{r_{02} - r_{20}}{4a} \tag{9}$$

$$d = \frac{r_{10} - r_{01}}{4a} \tag{10}$$

In order to extract a usable real world statistic from these equations and to test the practicality of the novel algorithm, the direction being faced by the human pose object in the camera frame was extracted by applying certain transformations to the obtained quaternions (hereby referred to as the angle of orientation of the human pose object). Let it be noted at this point that the quaternion values returned do in fact contain some noise due to implicit measurement uncertainities in xl, yl, zl, xr, yr and zr (refer to equation 1) as illustrated in the section IV, sub-section D 'Kalman Filter to eliminate quaternion noise'. These uncertainties were compensated for with the implementation of a 1-D Kalman filter for the angle of orientation of the human pose object as it was emperically the most accurate approach found post-testing.

This was done by extracting the angle of rotation from the angle-axis form of the quaternion, as the axis vector itself was implicitly made to be oriented upwards from the head of the human object in the frame when the rotation matrix was extracted from the model.

Assuming our quaternion to be of the form ...

$$\mathbf{Q} = cos\theta + sin\theta(xi + yj + zk) \tag{11}$$

Theta was extracted as ...

$$\theta = \arccos(\frac{a}{\sqrt{a^2 + b^2 + c^2 + d^2}}) \tag{12}$$

Theta was then transformed to obtain the angle of orientation in our frame of reference such that the axis of reference was the horizontal across the camera feed window.

$$\theta = \frac{\theta \cdot \frac{180^2}{\pi}}{45} - 180 \tag{13}$$

The transformation applied here takes care of the conversion from radians to degrees as well as the fact that an x degree rotation in the quaternion translates to a rotation of 2x in a real life scenario. The equation was arrived at after collecting the values of theta for a 180 degree rotation of the human object and mapping them to the real world angle being faced by the same. The angle being calculated was essentially the one included by a perpendicular from the line connecting the shoulder points to the horizontal of the camera frame. This scene can be visualized as shown in figure 3, not taking into account the fact that the camera and the shoulder points don't exist at the same height.

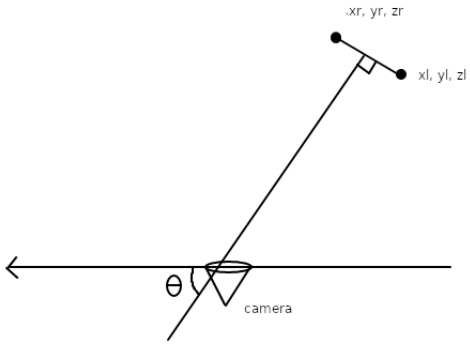

Fig. 3. A simplistic visualization of the use case to detect the angle of orientation of a pedestrian object taking just the shoulder points.

C. Algorithm

Prior to covering the main algorithm, do note that the *getRotationMatrix* function accepts 6 floating point values denoting xl, yl, zl, xr, yr and zr respectively (Refer to equation 1) and returns the rotation matrix formulated in equation 5 by iterating through the calculations in equations 2 through 4. The *calculateQuaternion* function accepts a rotation matrix R as given in equation 5, calculates **Q** as in equation 6 by calculating its components according to equations 7 through 10 and returns the required quaternion.

The aggregated algorithmic flow for quaternion generation from the 2-D image of the skeletal pose was framed to proceed as follows -

Input: Video feed API v
Output: **Q**

Initialisation :
v = VideoCapture Class [16]
LOOP Process :
while True **do**
  img = v.frame // assigning an image requested from the API to the variable img at the time of request
  pose = poseDetector(img) // detects the existence of a skeletal pose in the frame
  // getting the requested landmarks from the MediaPipe framework
  if pose != None **then**
    xl, yl, zl = pose.getLandmarks('Left Shoulder')
    xr, yr, zr = pose.getLandmarks('Right Shoulder')
    R = getRotationMatrix(xl, yl, zl, xr, yr, zr)
    Q = calculateQuaternion(R)
    return **Q**

12: **end if**
13: **end while**

IV. Results and Supporting Statistics

The system was tested using a standard wide-angle USB 2.0 camera (specifications elaborate on in sub-section C) and yielded a frame rate of 24 per second on average. Taking the reciprocal of the frame rate gives us the testing latency of the algorithm as a whole. This value came out to be 41.67 milliseconds. A comparison of this method with even a rudimentary artificial neural network with three hidden layers of twenty neurons with respect to the number of mathematical operations required to generate a prediction [17] illustrates the advantage of such a set up, given that neural networks are the current industry standard for such tasks [18]

A snapshot of the execution sequence can be seen in figure 4 where **Q** was obtained as 0.63 - 0.12i + 0.31j + 0.62k, as visualized in figure 6. The accuracy of the same is verified in sub-section B 'Model Accuracy Verification'. All results obtained in this section were taken from the 25th test iteration of the novel algorithm after they were deemed satisfactory.

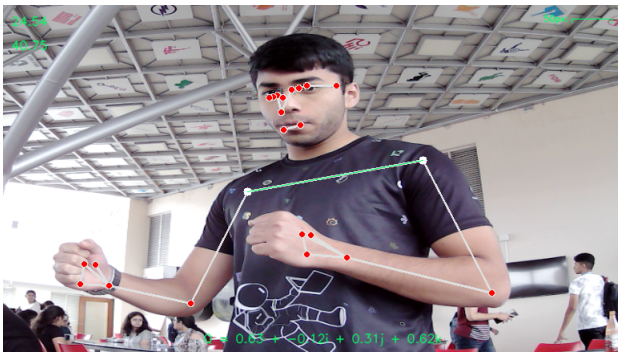

Fig. 4. A demonstration of the quaternion extraction from the pose using the novel algorithm.

A. Testing Conditions

The scenes employed in the testing of the novel algorithm were intentionally chosen to reflect the systems intended use case according to the authors, i.e.; estimating the angle of orientation of a pedestrian object from the point of view of an autonomous robot. As such, they were made to contain multiple human objects to demonstrate the arbitrary selection of the skeletal frame with the highest confidence value due to the singularly threaded nature of the novel algorithm.

The domain taken for all angles of testing given the use case were all human pose objects with an orientation between 10 and 180 degrees with respect to the axis determined by the quaternion. As such, the test cases covered involved pedestrian objects approaching the robots path from 8 different orientations -

- Obliquely to the right, in the direction of the robots motion.
- Obliquely to the left, in the direction of the robots motion.
- Obliquely to the right, against the direction of the robots motion.
- Obliquely to the left, against the direction of the robots motion.
- Perpendicularly to the right, cutting across the path of the robot.
- Perpendicularly to the left, cutting across the path of the robot.
- Directly towards the robot.
- Facing away from the robot but still obstructing its path.

The angle domain of 10 to 180 degrees was maintained through cases where the pedestrain was facing both towards and away from the camera by extrapolating the direction of motion of the pedestrian object on the frame and classifying them into fuzzy states for the same.

B. Model Accuracy Verification

Due to current resource limitations, the angle of orientation of the human pose object was extracted to verify the accuracy of the quaternion model by cross-checking the angle being faced by the user of the program as illustrated by equations 11 through 13.

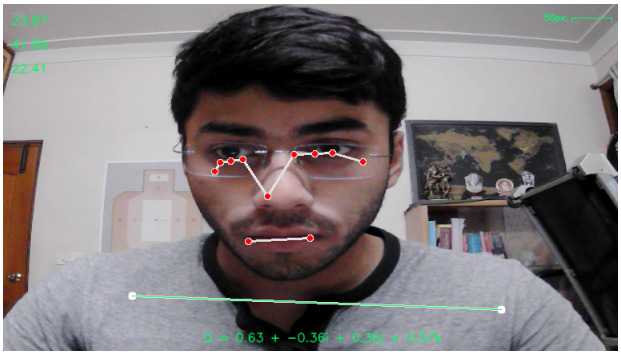

Fig. 5. A representation of a stationary human pose object taken for initial quaternion accuracy evaluation. Note that only the pose object itself and the shoulder landmarks need to be detected for all calculations following the same in the algorithm.

In a preliminary test of a stationary pose object as illustrated in figure 5, it was noted that the components of the quaternion were generated with the following variances and standard deviations over the time-span of the execution sequence of the stationary scene -

- Real component variance and standard deviation : 3.08e-4, 1.76e-2
- i component variance and standard deviation : 9.83e-4, 3.14e-2
- j component variance and standard deviation : 3.83e-4, 1.96e-2
- k component variance and standard deviation : 2.39e-5, 4.88e-3

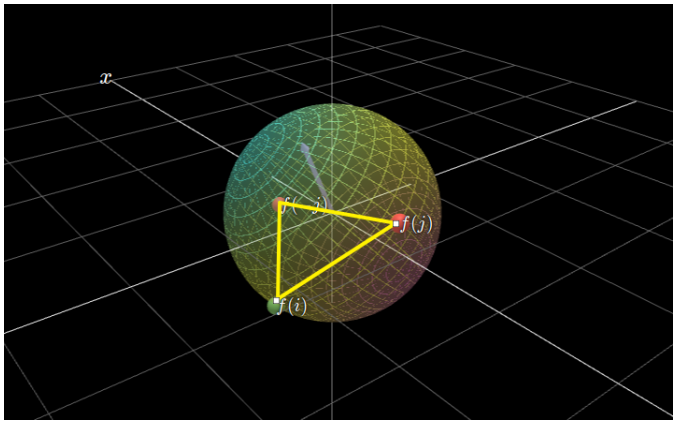

Fig. 6. A representation of the quaternion generated from the pose estimate in figure 4 with the three vertices of the triangle denoting the positions of the shoulder points and the direction of orientation respectively. [19].

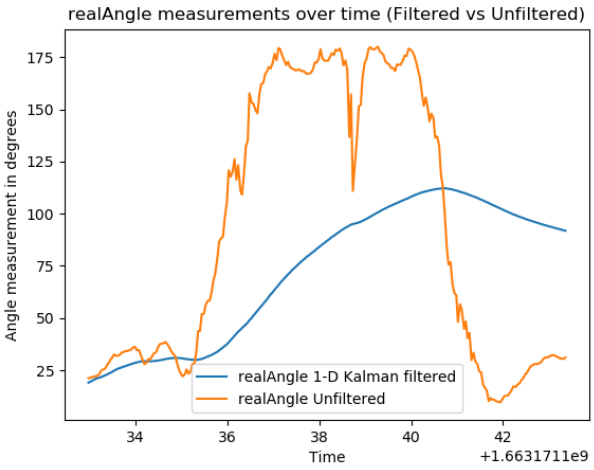

Fig. 7. A line graph representation of the filtered values from the theta measurement, referred to here as realAngle.

C. System specifications

This system was tested on a platform with the following specifications -

- Intel Core i5 7200U processor
- 8GB RAM
- 512 GiB Solid State Drive
- External USB 2.0 30 FPS 2MP 'Passport' camera with a resolution of 1920x1080 and view angle of 110 degrees.

D. Kalman filter to eliminate quaternion noise

In order to obtain the angles of orientation quoted in sub-section B, a Kalman filter in one dimension was applied to the theta readings in order to get a stable feed [20].

To re-iterate, the angle measurements were transformed as given in equation 13 to account for the fact that all values for theta were taken with respect to an axis horizontally bisecting the frame, which was effectively achieved by taking test cases of human pose objects with erect postures. Refer to figure 7 for a visualization of the same.

The filter designed was made to be dynamic for a system assumed to be perfect, i.e.; the predicted covariance equation assumed the model uncertainty to be zero. This was because of the random nature of the system and no viable physical equations present to accurately define the motion of the pose object. Hence, the filter was made to accept the average of the last 10 readings in the window as the prediction when the real measurement was deemed to be not viable. For scale, 24 measurements were calculated from the generated quaternion every second.

Equations 14 through 16 cover the updation of the Kalman gain, state and covariance respectively, where k is the Kalman gain, p is the covariance, r was the angle measurement uncertainty (emperically determined as 0.5 after testing), x was previous measurement and z was the current measurement itself -

$$k = \frac{p}{p+r}\,;\; kalman\; gain\; calculation \tag{14}$$

$$x = x + k(z - x)\,;\; state\; updation \tag{15}$$

$$p = (1-k)p\,;\; covariance\; updation \tag{16}$$

In order to demonstrate the translation of this system into use cases in the real world, we chose to build a preliminary prototype of a pedestrian intent classification system [21], where the co-ordinates of a pedestrian object on the frame as well as take the theta output value mentioned earlier to were recorded in a window of values to project the future path of motion on to the same frame. The classification was done using a fuzzy state approach taking these inputs into account. A demonstration of this is illustrated in figure 8, where one should take note of the third value in the top-right of the overlay reading 172.04. This was the angle of orientation of the pedestrian object returned by the novel algorithm in an on-ground test. This can be visually verified by the fact that the referenced pedestrian object was walking almost perpendicularly across the projected path of the robot. Also note that the image here has been flipped horizontally, and hence does not seem consistent with the previously mentioned frame bisecting axis definition at first glance.

The intention classifier was able to accurately classify approximately 75 percent of the co-ordinates of the human pose object on the frame 2 seconds into the future with a radius of accuracy of 50 pixels.

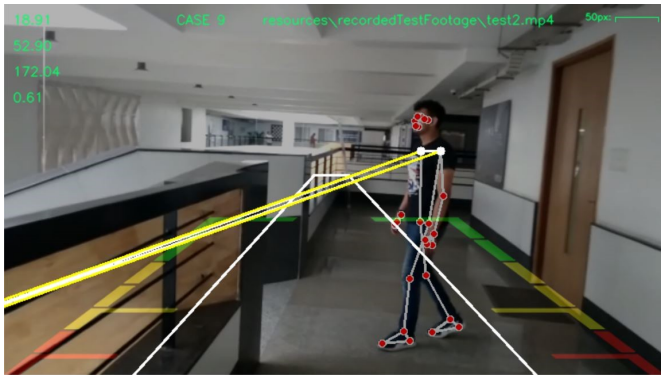

Fig. 8. A demonstration of the fuzzy pedestrian intent classification algorithm prototype implemented from a ground vehicle perspective to demonstrate the practicality of quaternion extraction from a 2-D camera frame containing a human skeletal pose object.

V. Conclusions and End Notes

Therefore, the conclusion was drawn that the novel algorithm was suitable for extracting the pose of a human object using 2-D cameras and satisfied the requirements of preserving costs due to minimal hardware usage as well as those of latency and performance as the algorithm can be implemented on hardware [22] and performs within reasonable limits of accuracy. Notwithstanding the latter, its single threaded nature as well as low computational resource usage make it suitable for edge deployment applications, although it should be noted that the number of threads required increases linearly with the number of pose objects detected.

A. Optimizations and Future Scope

The novel algorithm can be optimized in a variety of ways, with the primary approach being multi-threading of the video feed as well as the implementation of the algorithm for detecting multiple human objects and processing their pose estimations in parallel [23]. Expanding on the use case mentioned in section IV, sub-section A 'Testing Conditions', the model can be used to implement a more elaborate version of the pedestrian intent classifier demonstrated in sub-section D of the same section with more reliable prediction of whether or not a pedestrian object will cross the path in front of an autonomous robot such that its motion may endanger the lives of the same [24].

Additionally, optimizing the algorithm to make use of the large core count of a GPU also has promising prospects [25] [26], although a deeper literature survey into the topic would have to be conducted to say so for sure.

References

[1] K. Fathian, J. P. Ramirez-Paredes, E. A. Doucette, J. W. Curtis, and N. R. Gans, "Quaternion based camera pose estimation from matched feature points," *arXiv preprint arXiv:1704.02672*, 2017.

[2] A. Jouybari, A. Ardalan, and M. Rezvani, "Experimental comparison between mahoney and complementary sensor fusion algorithm for attitude determination by raw sensor data of xsens imu on buoy," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 497–502, 2017.

[3] D. Pavllo, D. Grangier, and M. Auli, "Quaternet: A quaternion-based recurrent model for human motion," *arXiv preprint arXiv:1805.06485*, 2018.

[4] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4346–4354.

[5] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2891–2900.

[6] D. Holden, T. Komura, and J. Saito, "Phase-functioned neural networks for character control," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.

[7] D. Schneider, "Deeper and cheaper machine learning [top tech 2017]," *IEEE Spectrum*, vol. 54, no. 1, pp. 42–43, 2017.

[8] D. Eberly, "Rotation representations and performance issues," *Magic Software: Chapel Hill, NC, USA*, 2002.

[9] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[10] S. Sarabandi and F. Thomas, "A survey on the computation of quaternions from rotation matrices," *Journal of Mechanisms and Robotics*, vol. 11, no. 2, 2019.

[11] S. Rungruangbaiyok, R. Duangsoithong, and K. Chetpattananondh, "Shoulder angle measurement (sam) system for home-based rehabilitation using computer vision with a web camera." *Songklanakarin Journal of Science & Technology*, vol. 43, no. 5, 2021.

[12] V. Bazarevsky and I. Grishchenko, "On-device, real-time body pose tracking with mediapipe blazepose," Aug 2020. [Online]. Available: https://ai.googleblog.com/2020/08/on-device-real-time-body-pose-tracking.html

[13] W. R. Hamilton, *Elements of quaternions*. London: Longmans, Green, & Company, 1866.

[14] J. C. Prunier, "Camera look-at method," Dec 2016. [Online]. Available: https://www.scratchapixel.com/lessons/mathematics-physics-for-computer-graphics/lookat-function

[15] M. D. Shuster *et al.*, "A survey of attitude representations," *Navigation*, vol. 8, no. 9, pp. 439–517, 1993.

[16] D. L. Baggio, *Mastering OpenCV with practical computer vision projects*. Packt Publishing Ltd, 2012.

[17] S.-C. Wang, "Artificial neural network," in *Interdisciplinary computing in java programming*. Springer, 2003, pp. 81–100.

[18] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.

[19] B. Eater and G. Sanderson, "Visualizing quaternions, an explorable video series," Sep 2018. [Online]. Available: https://eater.net/quaternions

[20] Z. Huang, P. Du, D. Kosterev, and B. Yang, "Application of extended kalman filter techniques for dynamic model parameter calibration," in *2009 IEEE Power & Energy Society General Meeting*. IEEE, 2009, pp. 1–8.

[21] J.-Y. Kwak, B. C. Ko, and J.-Y. Nam, "Pedestrian intention prediction based on dynamic fuzzy automata for vehicle driving at nighttime," *Infrared Physics & Technology*, vol. 81, pp. 41–51, 2017.

[22] J. Hegarty, J. Brunhaver, Z. DeVito, J. Ragan-Kelley, N. Cohen, S. Bell, A. Vasilyev, M. Horowitz, and P. Hanrahan, "Darkroom: compiling high-level image processing code into hardware pipelines." *ACM Trans. Graph.*, vol. 33, no. 4, pp. 144–1, 2014.

[23] K. Roszyk, M. R. Nowicki, and P. Skrzypczyński, "Adopting the yolov4 architecture for low-latency multispectral pedestrian detection in autonomous driving," *Sensors*, vol. 22, no. 3, p. 1082, 2022.

[24] B. Völz, K. Behrendt, H. Mielenz, I. Gilitschenski, R. Siegwart, and J. Nieto, "A data-driven approach for pedestrian intention estimation," in *2016 ieee 19th international conference on intelligent transportation systems (itsc)*. IEEE, 2016, pp. 2607–2612.

[25] G. A. Laguna-Sánchez, M. Olguín-Carbajal, N. Cruz-Cortés, R. Barrón-Fernández, and J. A. Álvarez-Cedillo, "Comparative study of parallel variants for a particle swarm optimization algorithm implemented on a multithreading gpu," *Journal of applied research and technology*, vol. 7, no. 3, pp. 292–307, 2009.

[26] K. Fatahalian and M. Houston, "A closer look at gpus," *Communications of the ACM*, vol. 51, no. 10, pp. 50–57, 2008.