Trend Shift in Pose Estimation: Toward Continuous Representation

Sunwoo Bang, Chaeyeong Lee, Junhyeong Ryu, Hyung Tae Lee, and Jeongyeup Paek

Department of Computer Science & Engineering, Chung-Ang University, Seoul, Republic of Korea
{layer97, cxaexeong, rjh6883, hyungtaelee, jpaek}@cau.ac.kr

***Abstract*—Pose estimation is a fundamental task in many applications that utilizing 3D data from sensors like LiDAR and RGB-D cameras. It is particularly crucial in fields where precise position and orientation information are required, such as autonomous driving, cooperative perception, robotics, and augmented reality (AR). To improve the pose estimation, many methods used in other applications are adopted to enhance the network architecture and these methods make significant progress in pose estimation. However, when using deep neural networks (DNNs), the issue of discontinuous rotation representation has emerged, and various studies pointed out that this could be a cause of substantial error. Therefore, we focus on addressing the issue of discontinuities by reviewing the latest research trends in pose estimation published by major academic publishers and provide insights into future directions for pose estimation.**

***Index Terms*—Pose estimation, rotation representation, neural networks**

I. Introduction

3D sensors such as LiDAR, and RGB-D camera provide accurate 3D data representing the real world, which offer the essential cognitive ability to recreate the real world. This cognitive ability enables significant progress in fields like augmented reality [1], autonomous driving [2], inverse kinematics [3], and other applications. At the core of technological advancements based on such 3D data is the pose estimation technique, which estimates the position and orientation of an object or the sensor itself.

Before the widespread adoption of deep neural networks (DNN), pose estimation relied on hand-crafted features or algebraic solutions. However, conventional methods are vulnerable to noisy, cluttered scenes and cannot be generalized for various scenes. Therefore, when the robust feature matching capability of DNN became evident, which improved robustness to complex scenes, using DNNs for pose estimation became dominant.

While deep-learning based solutions have significantly improved the accuracy of pose estimation, errors still persist, and researchers continue to strive to minimize these errors further. One of the reasons for errors is the representation of rotation. Euler angles, unit quaternions, or axis-angle representations are typically used, but the discontinuity in these rotation representations can be a significant vulnerability. We analyze and review research papers to provide insights into the directions and trends of recent studies addressing this issue. Additionally, since pose estimation is an essential task before providing 3D data to recreate the real world, we also explore studies that apply neural networks (*e.g.,* ensemble, attention mechanism) designed for specific field, beyond research on rotation representation.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A4A5034130 & No. RS-2024-00359450), and also by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2022-00156353) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

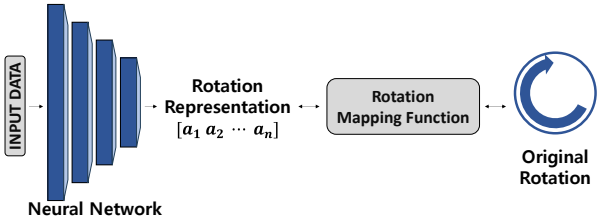

Fig. 1: Overview of rotation estimation

The remainder of this article is organized as follows. We first introduce Euler angles and unit quaternions in Section II and describe the papers on rotation representation in Section III. In Section IV, we discuss non-rotation methods to address discontinuities, and finally, we conclude the article in SectionV.

II. Background

To represent 3D rotation in pose estimation, using Euler angle or unit quaternion is simple and intuitive approach. Euler angles represent the orientation of an object in 3D space using three sequential rotations, as expressed in equation (1), where the rotations performed in XYZ order. Each angle α, β, γ corresponds to a rotation about X, Y, Z axes respectively, where s and c denote the $\sin$ and $\cos$ functions.

$$R_{euler} = \begin{bmatrix} c_\beta c_\gamma & -c_\beta s_\gamma & s_\beta \\ c_\alpha s_\gamma + s_\alpha s_\beta c_\gamma & c_\alpha c_\gamma - s_\alpha s_\beta s_\gamma & -s_\alpha c_\beta \\ s_\alpha s_\gamma - c_\alpha s_\beta c_\gamma & s_\alpha c_\gamma + c_\alpha s_\beta s_\gamma & c_\alpha c_\beta \end{bmatrix} \quad (1)$$

Euler angles are intuitive and easy to understand, but they have the drawback such as gimbal lock [4], where two axes align, causing a loss of one degree of freedom (DoF) and limiting the representation to 2D rotation.

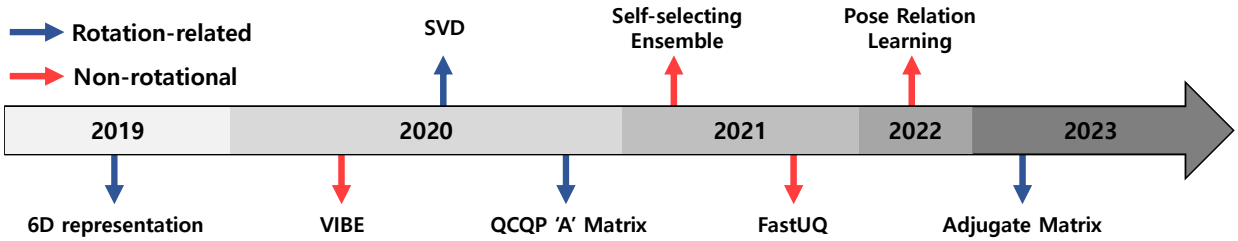

Fig. 2: Timeline of progress in pose estimation. Blue and red arrows refer rotation representation approaches, and non-rotational approach, respectively. Self-selecting ensemble method is non-rotational approach, but it aims to achieve rotation continuity.

In contrast, as shown in equation (2), a quaternion is represented as a four dimensional complex number consisting of one real part and three imaginary components corresponding to i, j, k, where q_w is a scalar part of quaternion, and q_x, q_y, q_z are coefficients of imaginary parts.

$$q = q_w + q_x i + q_y j + q_z k \tag{2}$$

Quaternion can also be represented using trigonometric functions as:

$$q = \cos\left(\frac{\theta}{2}\right) + \sin\left(\frac{\theta}{2}\right)(\mathbf{v}_x i + \mathbf{v}_y j + \mathbf{v}_z k) \tag{3}$$

where θ represents the rotation angle, and $\mathbf{v}_{x,y,z}$ denote the rotation axis, making it more intuitive to understand compared to equation (2).

To perform a rotation using a quaternion, we can use the formula shown in equation (4), where p' represents the rotated quaternion. This is achieved by applying the quaternion q, its conjugate q^*, and the quaternion to be rotated p as follows:

$$p' = qpq^* \tag{4}$$

Furthermore, a quaternion can be converted into a rotation matrix, as shown in equation (5),

$$R_{quat} = \begin{bmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y - q_wq_z) & 2(q_xq_z + q_wq_y) \\ 2(q_yq_x + q_wq_z) & 1-2(q_x^2+q_z^2) & 2(q_yq_z - q_wq_x) \\ 2(q_zq_x - q_wq_y) & 2(q_zq_y + q_wq_x) & 1-2(q_x^2+q_y^2) \end{bmatrix} \tag{5}$$

where q_w is the scalar value of the quaternion, and the others represent the coefficient of imaginary numbers (i, j, k). This rotation matrix R_{quat} allows for the application of 3D rotations in a form that is compatible with linear algebra operations commonly used in graphics and robotics.

Due to robust interpolation method, spherical linear interpolation (SLERP) and absence of gimbal lock issue, unit quaternions are widely used in computer graphics, and computer vision field. However, because of quaternion rotation calculation in Eq. (4), quaternion q and $-q$ represent the same rotation, which create an ambiguity.

Fig. 1 shows the overview of rotation estimation. Rotation representation is simply the output of neural networks, typically expressed as N-dimensional vector, and mapping functions are used to map this vector to a rotation, such as 3×3 rotation matrix and vice versa.

PoseCNN [5] is representitive pose estimation method using unit quaternions, which achieved high accuracy even when multiple objects are occluded in cluttered scene. In this work, it was observed that significant errors within certain rotation angle ranges both in symmetric objects, and non-symmetric objects. While the main causes of this issue are not clearly identified, some researchers assume that it is due to discontinuous rotation representation.

III. Rotation Representation Approach

Many prior works are striving to reduce the error of pose estimation. We can categorize their efforts in rotation representation (rotation-related) approaches, which try to achieve rotational continuity through rotation representation, and non-rotational approaches. In this section, we categorize prior works focusing on 'continuity'. Fig. 2 shows the chronological overview of prior research improving pose estimation.

A. Orthogonalization

Zhou et al. [6] pointed out that high error rates in certain rotation angle ranges in PoseCNN [5], then argued these errors are due to discontinuity of conventional rotation representation. Then, they proposed the 6D rotation representation. 6D means two 3D orthogonal vectors gain from first and second column vectors of 3×3 rotation matrix. The third column vector can be restored from the two vectors, therefore, the rotation matrix can be represented with six elements. The mapping function to representation space is defined as:

$$g\left(\begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}\right) = \begin{bmatrix} c_1 & c_2 \end{bmatrix} \tag{6}$$

where c_n represents the column vector of the rotation matrix respectively. 6D representation [6] was a significant milestone in the progress of pose estimation. This work not only proposed a solution to a critical issue in pose estimation but also sparked the widespread interest in the field of rotation representation.

Levinson et al. [7] proposed rotation representation using singular value decomposition (SVD) orthogonalization, which needs 9D network output. This work is similar to Zhou et al. [6] in that both representations involve orthogonalization process. 6D representation uses the Gram-Schmidt orthogonalization process, whereas the method proposed in this paper uses SVD orthogonalization. While $SVD(M) = U\Sigma V^T$, they train the network with $SVDO(M) = UV^T$ and $SVDO^+(M) = U\Sigma' V^T$

B. Symmetric matrix

Peretroukhin et al. [8] proposed another rotation representation adopting a quadratically-constrained quadratic program (QCQP). This paper pointed out that prior rotation representation including 6D representation [6] is a representation for certain 'point', which cannot represent the uncertainty of neural networks. Then they proposed 'A' matrix which is a symmetric 4×4 matrix using a 10D vector and can cover Bingham distribution over unit quaternions. Using dispersion of distribution, this method can effectively detect and reject out-of-distribution (OOD) data, which is called dispersion thresholding (DT) without any additional stochastic processes. 'A' matrix is represented as:

$$A(\theta) = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ \theta_2 & \theta_5 & \theta_6 & \theta_7 \\ \theta_3 & \theta_6 & \theta_8 & \theta_9 \\ \theta_4 & \theta_7 & \theta_9 & \theta_{10} \end{bmatrix} \tag{7}$$

where $A(\theta)$ is a quadratic cost function of QCQP, parameterized by θ. Parameter θ_n are learned by neural network through calculating sum and difference between each point within the point cloud. As a result, 'A' matrix is equivalent to $\sum_{k=1}^{N} B_k$, and B_k is defined as:

$$B_k = \begin{bmatrix} a_1^2 + a_2^2 + a_3^2 & a_3 s_2 - a_2 s_3 & a_1 s_3 - a_3 s_1 & a_2 s_1 - a_1 s_2 \\ a_3 s_2 - a_2 s_3 & a_1^2 + s_2^2 + s_3^2 & a_1 a_2 - s_1 s_2 & a_3 s_1 - a_1 s_3 \\ a_1 s_3 - a_3 s_1 & a_1 a_2 - s_1 s_2 & a_2^2 + s_1^2 + s_3^2 & a_2 a_3 - s_2 s_3 \\ a_2 s_1 - a_1 s_2 & a_3 s_1 - a_1 s_3 & a_2 a_3 - s_2 s_3 & a_3^2 + s_1^2 + s_2^2 \end{bmatrix} \tag{8}$$

where a_n, s_n represent $x_n - y_n$ and $x_n + y_n$ with two point x and y, respectively.

Lin et al. [9] pointed out that the approach proposed by Xiang [10] is an *ad hoc* method (it will be introduced later), and that it conceals the simple relationship between quaternions with a complicated approach. Then they proposed an adjugate matrix that also represents rotation with 4×4 symmetric matrix using a 10D vector. The adjugate matrix is based on root mean squared deviation (RMSD), and the expanded Frobenius loss function proposed by Bar-Itzhack [11]. The adjugate matrix is represented as:

$$M_{adj} = \begin{bmatrix} q_w^2 & q_{wx} & q_{wy} & q_{wz} \\ q_{wx} & q_x^2 & q_{xy} & q_{xz} \\ q_{wy} & q_{xy} & q_y^2 & q_{yz} \\ q_{wz} & q_{xz} & q_{yz} & q_z^2 \end{bmatrix} \tag{9}$$

where q_{mn} represent $q_m q_n$ of quaternions.

IV. Non-rotational approach

Research to improve pose estimation using methods other than rotation representation has been performed. Since the widespread adoption of DNNs, many fields have observed improvement driven by the application of neural networks designed for specific fields. Among the various ideas for improvement, two of the most representative notable methods are network ensemble and attention mechanism. Table I shows the various methods to improve the robustness of pose estimation. Even if they use the same method, the purposes could be different.

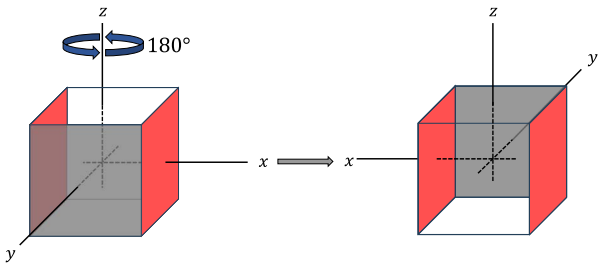

Fig. 3: Rotational symmetry about $z - axis$. Without grey plane, rotation cannot be perceived.

A. Ensemble

Network ensemble is a method to improve approximation accuracy by using multiple networks. Multiple outputs could be used to quantify and indicate uncertainty [12], or combine multiple models with different characteristics [15].

Fig. 3 shows the issue from symmetric objects. While the object appears the same, the pose is changed, and this phenomenon causes ambiguity in rotation estimation. However, many studies do not handle the symmetric objects separated from other objects. Xiang [10] pointed out that methods proposed in prior works [6]–[8] successfully reduced the average error, but still performed large maximum error of $90°$ to $180°$ especially on symmetric objects. They argued that the large maximum error is due to topological error in neural networks, which is a cause of rotation discontinuity. Then, Xiang [10] proposed a self-selecting ensemble method using four training functions that are selected based on conditions. This method solved the discontinuity problem without proposing a new rotation representation. In the experiments, a significant reduction in maximum error is observed.

Shi et al. [12] pointed out the unreliability of deep learning-based pose estimation. Then, they proposed the fast uncertainty quantification (UQ) method, which ensembles multiple heterogeneous models. They first estimate pose respectively, then calculate average disagreement, which refers to errors between different estimated poses.

B. Attention mechanism

The attention mechanism is a widely used method in language models, which can applied to improve pose estimation robustness. By using attention weight, the system can identify the importance of different inputs, allowing the network to focus more on elements that have a greater impact on the estimation Prior pose estimation research, for example, Kocabas et al. [14] adopted the self-attention mechanism on pose estimation to improve human pose estimation in sequential input data such as video. They used the self-attention mechanism in motion discriminator, which enabled the network to learn relative importance between each frame, and assign higher attention weight to them.

Hoang et al. [13] utilized the self-attention mechanism to capture the correlation between objects. They divide correlation into inter-part correlation and inter-instance correlation,

Category	Method	Description	Reference
Rotation-related	Orthogonalization	Incorporate orthogonalization within neural network output	[6], [7]
	Symmetric Matrix	Symmetric 4×4 matrix for rotation representation	[8], [9]
Non-rotation	Ensemble	Use multiple network/learning function	[10], [12]
	Attention Mechanism	Weight to highlight important characteristics	[13], [14]

TABLE I: Different methods of pose estimation improvement

then calculate them separately. Through correlation information, the system could perform relatively accurate pose estimation even when objects are heavily occluded by others.

C. Progression through multi-approach

We notice that the both rotation representation approach and non-rotational approach could applied simultaneously. For example, the method proposed by Xiang [10] shows that even though this method aimed to achieve rotational continuity through network topology, novel rotation representation approaches can improve the accuracy further. Not only in the self-selecting ensemble, Hoang et al. [13] also adopted 6D representation to represent 3D rotation. Furthermore, state-of-the-art pose estimation methods including human pose estimation and object pose estimation are adopting the rotation representation method in their algorithms [16]–[21]. Therefore, adopting continuous rotation representation could be an easy, plug-and-play method for improving pose estimation.

V. Conclusion

Tremendous efforts are being dedicated to developing accuracy and robustness in pose estimation within academia. We reviewed prior efforts to improve accuracy and robustness of pose estimation, and classified prior approaches into two major categories. Based on our observation, integrating continuous rotation representation with other approaches can lead to significant progress in pose estimation. By integrating multiple approaches across different categories, it is possible to significantly enhance the accuracy and reliability of pose estimation, the core of other 3D applications.

Future pose estimation research should focus not only on neural network architectures, but also on continuity of rotation representations. We aim our comprehensive survey to serve as a reference and directional guide to many fellow researchers who wish to study pose estimation in the future.

References

[1] W. Pang, C. Xia, B. Leong, F. Ahmad, J. Paek, and R. Govindan, "UbiPose: Towards Ubiquitous Outdoor AR Pose Tracking using Aerial Meshes," in *Proceedings of The 29th Annual International Conference On Mobile Computing And Networking (Mobicom'23)*. ACM, Oct. 2023.

[2] L. Liu, H. Li, Y. Dai, and Q. Pan, "Robust and Efficient Relative Pose With a Multi-Camera System for Autonomous Driving in Highly Dynamic Environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 8, pp. 2432–2444, 2017.

[3] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3383–3393.

[4] S. Kim and M. Kim, "Rotation Representations and Their Conversions," *IEEE Access*, vol. 11, pp. 6682–6699, 2023.

[5] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[6] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the Continuity of Rotation Representations in Neural Networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5745–5753.

[7] J. Levinson, C. Esteves, K. Chen, N. Snavely, A. Kanazawa, A. Rostamizadeh, and A. Makadia, "An Analysis of SVD for Deep Rotation Estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 554–22 565, 2020.

[8] V. Peretroukhin, M. Giamou, D. M. Rosen, W. N. Greene, N. Roy, and J. Kelly, "A Smooth Representation of Belief over SO(3) for Deep Rotation Learning with Uncertaintys," *arXiv preprint arXiv:2006.01031*, 2020.

[9] C. Lin, A. J. Hanson, and S. M. Hanson, "Algebraically Rigorous Quaternion Framework for the Neural Network Pose Estimation Problem," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 097–14 106.

[10] S. Xiang, "Eliminating Topological Errors in Neural Network Rotation Estimation Using Self-selecting Ensembles," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–21, 2021.

[11] I. Y. Bar-Itzhack, "New Method for Extracting the Quaternion from a Rotation Matrix," *Journal of guidance, control, and dynamics*, vol. 23, no. 6, pp. 1085–1087, 2000.

[12] G. Shi, Y. Zhu, J. Tremblay, S. Birchfield, F. Ramos, A. Anandkumar, and Y. Zhu, "Fast Uncertainty Quantification for Deep Object Pose Estimation," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5200–5207.

[13] D.-C. Hoang, J. A. Stork, and T. Stoyanov, "Voting and Attention-Based Pose Relation Learning for Object Pose Estimation From 3D Point Clouds," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8980–8987, 2022.

[14] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video Inference for Human Body Pose and Shape Estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.

[15] P. F. Proença and Y. Gao, "Deep Learning for Spacecraft Pose Estimation from Photorealistic Rendering," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6007–6013.

[16] Y. Li, Y. Mao, R. Bala, and S. Hadap, "MRC-Net: 6-DoF Pose Estimation with MultiScale Residual Correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 476–10 486.

[17] D. Rondao, N. Aouf, and M. A. Richardson, "ChiNet: Deep Recurrent Convolutional Learning for Multimodal Spacecraft Pose Estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 2, pp. 937–949, 2022.

[18] V. Mollyn, R. Arakawa, M. Goel, C. Harrison, and K. Ahuja, "IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–12.

[19] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas, "3D Human Pose Estimation via Intuitive Physics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 4713–4725.

[20] J. Li, K. Liu, and J. Wu, "Ego-Body Pose Estimation via Ego-Head Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 142–17 151.

[21] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6D Rotation Representation For Unconstrained Head Pose Estimation," in *IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2496–2500.