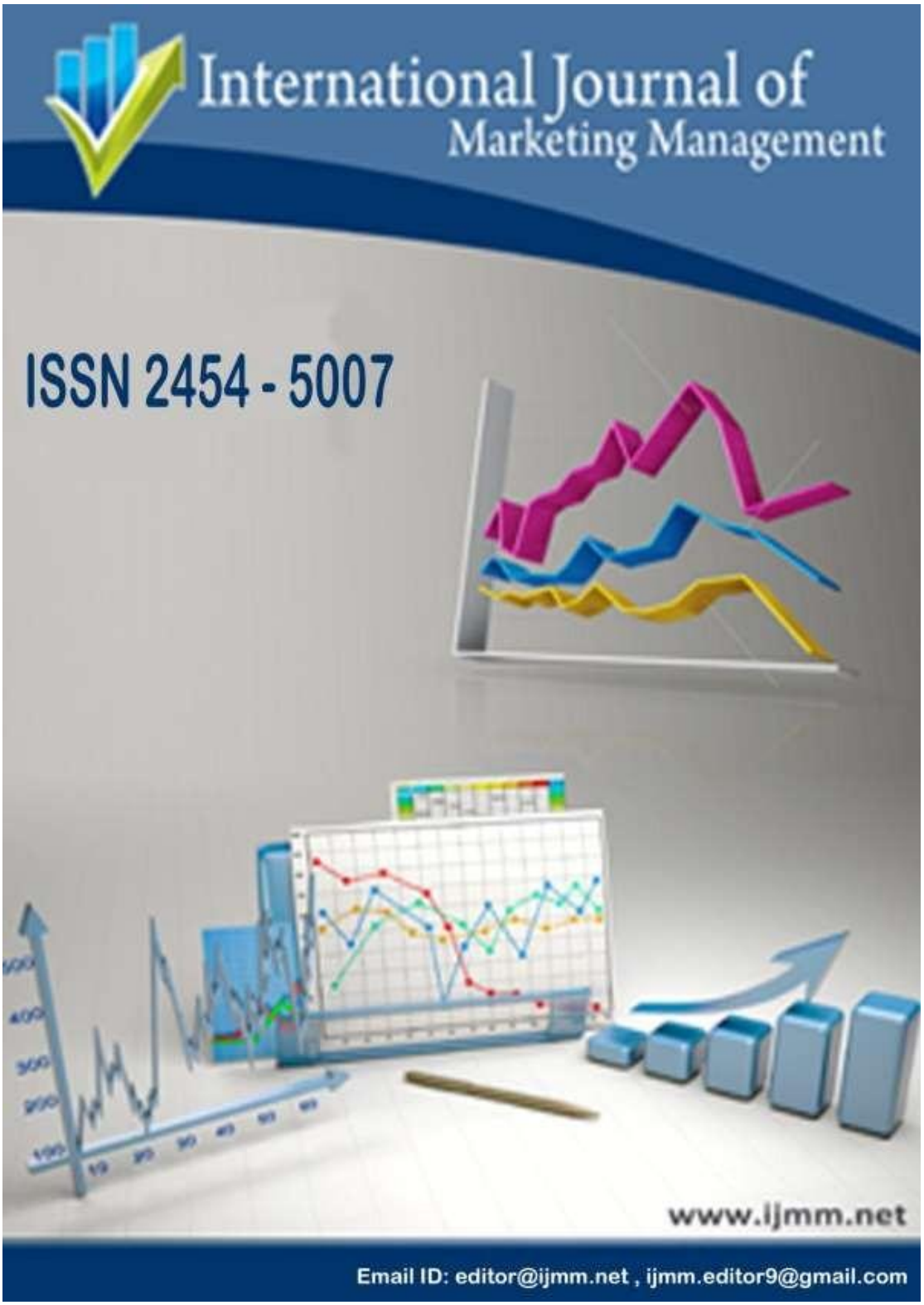

International Journal of
Marketing Management
ISSN 2454 - 5007
www.ijmm.net
Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

Human Face Localization in 3D For Humanoid Robot Vision using ML

[1]Bussa Kumara Swamy,[2]Kali Mahesh,[3]Karne Purnachander,[4]Amgoth Shobhan,[5]M.Sudhakar

[1,2,3,4] Student,Department of ECE,Narsimha Reddy Engineering College,Misammaguda(V),Kompally-500100,Telangana State,India.
[5] Assistant Professor,Department of ECE,Narsimha Reddy Engineering College,Misammaguda(V),Kompally-500100,Telangana State,India.

Abstract

Robots, autonomous vehicles, security, and human-computer interaction are just a few of the many computer vision applications that rely on accurate three-dimensional human position data. Improving the Arslan humanoid robot's vision system so it can have a continual connection with humans in its field of view is the major emphasis of this study. Using two identical USB cameras and two pan cameras calibrated with servo motors for precise angle alignment, the system extracts depth information from photographs using human binocular vision principles. Taking pictures, identifying faces, comparing camera counts, and doing 3D localization are all part of the process. The most fundamental objectives are to maintain eye contact with people, to properly determine their exact positions, and to predict their approximate locations. Recent advances in methodology have substantially improved localization accuracy, leading to fewer errors and more reliability. The field of computer vision as a whole benefits from these developments because they pave the way for more natural and quick interactions between humans and computers.

Keywords—3D localization, Depth Perception, Human-Robot Interaction, Stereo Vision

INTRODUCTION

Computer vision is an essential component of AI that enables robots to understand and process visual input. This has led to advancements in robotics, autonomous vehicles, and security systems. Due to inadequate depth awareness, older monovision systems often fail when trying to locate persons in three dimensions, making this a core problem in the field. A multitude of tactics aimed at enhancing the precision of spatial data extraction from images have been introduced by recent developments in camera calibration and posture assessment for 3D reconstruction. One of the most important ways to get the settings of a camera right is via target-based calibration, which makes use of well-established calibration targets like checkerboards. The controlled environment provided by these targets allows for accurate identification of both intrinsic (such as focal length and optical center) and extrinsic (such as rotation and translation vectors) parameters during camera calibration. There has been research into both target-based tactics and self-calibration methods that rely on mathematical modeling. As an example, methods that rely on [1] epipolar geometry use a plethora of photos taken from various angles to calibrate the system. In order to estimate camera settings and reproduce the scene's 3D structure, these methods resolve restrictions that arise from point correspondences between these photographs. Computer vision, artificial intelligence, and 3D printing rely on these methods to accurately estimate the position of the camera in three-dimensional space from two-dimensional images. Methods like Singular Value Decomposition (SVD) and single camera calibration, which employ projection matrices to transform 3D object coordinates into 2D picture coordinates, are crucial for extracting intrinsic and extrinsic factors, as stressed by Siddique et.al. [2].

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

In order to separate intrinsic from extrinsic qualities, SVD is crucial for decomposing the projection matrix into its component parts. Precise 3D reconstructions are made possible by this method's assurance of a constant and precise calibration approach. The purpose of the work "Robust Landmark Selection for 3D Face Pose Estimation" by Cıvır and Topal is to lessen the impact of oscillations in facial expressions on the accuracy of face pose assessment. This research finds the optimal set of landmarks using a variance analysis method, which improves the stability and accuracy of real-time 3D face position estimation. They discovered 29 landmarks for neutral expressions and 30 for expressive emotions, leading to a 36.14% and 14.79% decrease in jitter, respectively. Novel approaches to increasing the accuracy of human position estimation are presented in the paper "Enhancing 2D Point Tracking and 3D Pose Estimation for Human Behavior Analysis" [4] by Ariz et.al. The Lucas-Kanade approach for 2D point tracking and the weighted POSIT methodology for 3D pose estimation, coupled with outlier identification and correction algorithms, result in considerable accuracy improvements. With little extra processing cost, the method increases accuracy by about 30% in noisy conditions and around 15% in normal circumstances. "Staged Cascaded Network for Monocular 3D Human Pose Estimation" by Gao et.al. [5] offers a novel approach to solving the problems of 3D human posture estimation's accuracy and resilience.

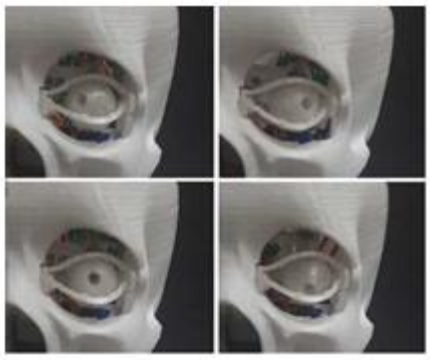
The proposed method employs an upgraded fusion module and composite residual module cascaded network architecture to boost network performance and data flow. An initial 2D landmark detector is used to identify key human features in images; then, a cascaded 3D coordinate regression model is used to generate a 3D posture using the 2D joints and geometric data. This problem is addressed in the research by developing a stereo vision system to enhance the ability of the humanoid robot Arslan [6] to make eye contact. Improving human localization accuracy and reliability using a stereo vision technology is the primary objective of this work. Two identical USB cameras and two meticulously calibrated pan cameras with servo motors capture images of their environment in Arslan's approach. The reason why pan motion is the major emphasis is because the application requires Arslan to have his eyes inside the viewfinder at all times. This enables for more precise control and alignment of the camera.

METHODS

A System for Humanoid Robot Vision in Arslan The humanoid robot Arslan's [6] skull, jaw, and cervical vertebrae are designed to mimic those of a human head. A satisfactory range of motion is achieved via Arslan's nick mechanism, which permits axial rotation of 41° on one side and flexion/extension of 45° (Figure 1)

Figure 1. Arslan's Physical Structure and Nick Rotations [6]

Arslan's eyes can only move in certain limited ways as they revolve around their axis of motion: With an adduction angle of 19°, an abduction angle of 34°, a supraduction angle of 0°, and an infraduction angle of -34°, 20 degrees. By adjusting these angles, Arslan is able to keep the Vestibulo-Ocular Reflex (VOR) within a reasonable range, which allows for stable vision even when the head moves (Figure 2).



International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

Figure 2 Arslan's Eyes Rotation [6].

In stereo vision, two cameras are positioned side by side to capture images of a subject from slightly different angles (Figure 3). Through the analysis of differences between the two images, this technique allows for depth awareness, mimicking human binocular vision.

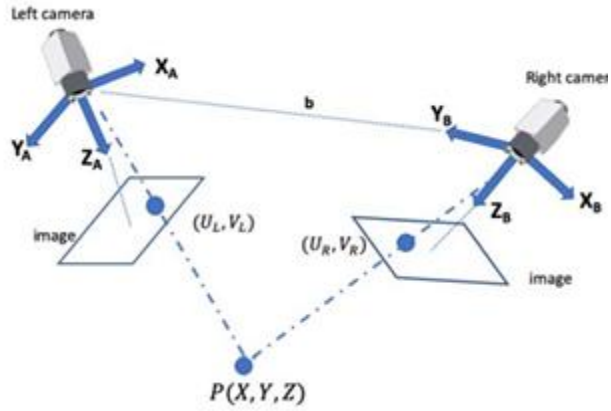


Figure 3. Stereo Vision System Layout [7]

One method for automatically classifying faces in images is face detection. Finding a face and its position is the goal of this essential subset of object detection [8]. An alternate perspective sees this task as a point detection problem, where the object is described by a single point—typically, the center of the face bounding box—that is of interest for face identification. You Only Look Once (YOLO), Fast R-CNN, Haar Cascade, and Regions with Convolutional Neural Networks (R-CNN) are just a few of the several face recognition technologies available. D. Recognition of Faces Data representation and feature extraction must be robust for face recognition to be accurate [9]. Key facial features must be encoded while noise and extraneous information are minimized for face data to be described successfully. To extract distinct face features from unprocessed image data, feature extraction methods like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and algorithms based on deep learning are essential. These methods simplify very high-dimensional data by drawing attention to the most crucial features for recognition. Local Binary Pattern Histogram (LBPH), Eigenface, and Fisherface are just a few of the several face recognition systems available. Methodology for Optimizing Face Count (E) To effectively identify and locate people, the conversion stereo vision localization approach uses two cameras to compute correct angles, allowing for successful tracking. Figure 4 shows a situation where two cameras are positioned with one facing forward and the other facing inward. In order to achieve pinpoint human localization, the system can identify and match users, as well as calculate pan angles (1 and 2) to align cameras. By computing tilt angles (α 1 and α 2), the system modifies vertical alignment. With these angles in hand, the stereo vision system can precisely position the cameras to follow objects in the viewing area with pinpoint accuracy. This method improves the stereo vision system's overall accuracy and reliability in many applications, such as computer vision tasks and human-robot interaction.

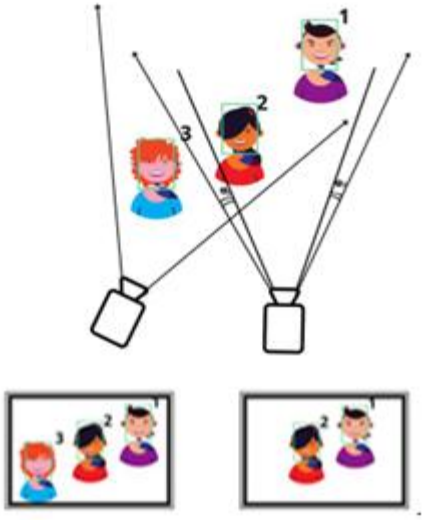


Figure 4. Maximizing Faces Count.

The algorithm that was used is this: 1. Make use of the left and right cameras to capture photographs. 2. Find faces using the front and rear cameras. 3. Check how many faces there are. If there are more faces picked up by the left camera than by the right camera, then: A. First, we'll use the footage from the left camera for our training. We will make an effort to move the appropriate camera. B-In the event that the total number of faces picked up by both cameras is more than zero: 2. We will be using the correct camera for our training. B.2-We will make an effort to move the left camera. C-In the event that the quantity of faces discovered by the two cameras is

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

complementary: First, we'll choose between the left and right cameras for our training. C.2-Copy the matching process to the second camera. The fourth criterion for 3D localization is whether or not all of the faces seen in the two cameras are identical. 4.1- Since he didn't fit, we'll compare the disproportionately huge faces of ancient people and pan to them. 4.2- Verify the camera's face detection capabilities; face size in one camera won't matter if the angle is more than 90 degrees, and vice versa for negative 90 degrees.

RESULTS

Evaluating Distance Profile Our model's performance in the right camera throughout many epochs is shown in the table below. Since both the face's breadth and its distance from the camera are rational, the former may serve as a crucial metric for the latter. In Tables I and II, for the left and right cameras, each row represents a separate training run, and in Figures 5 and 6, the corresponding columns show performance metrics at different points in the training process, denoted by epochs (60, 70, 80, 90, 100, 110, 120, 130, 140, 150, and 170).

Table I. Right Camera Distance Estimation

	60	70	80	90	100	110	120	130	140	150	170
	130	129	120	107	93	88	78	71	65	62	55
	132	125	119	106	94	85	79	72	67	61	56
	131	127	118	105	95	86	80	71	66	63	57
	131	128	122	104	94	87	81	70	65	64	58
	130	129	121	106	96	85	82	71	66	61	56
AVG	130.8	127.6	120	105.6	94.4	86.2	80	71	65.8	62.2	56.4
STD	0.83	1.67	1.58	1.14	1.14	1.30	1.58	0.70	0.83	1.30	1.40

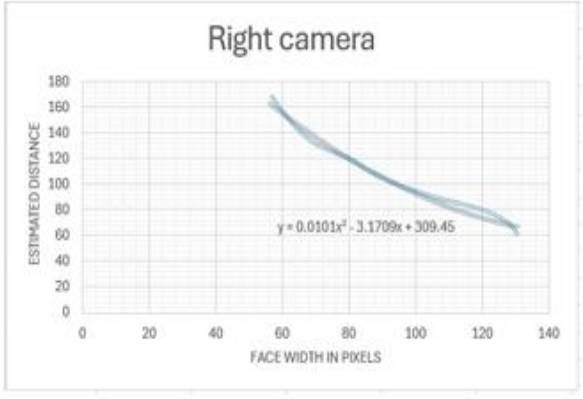


Figure 5. Right Camera Profile

Table II. Left Camera Distance Estimation

	60	70	80	90	100	110	120	130	140	150	170
	147	127	121	102	91	83	77	69	65	61	55
	146	129	120	103	93	82	76	70	67	60	56
	144	128	117	98	94	81	78	71	64	61	57
	146	130	118	99	95	85	79	70	63	59	55
	147	131	120	101	93	84	77	69	65	60	54
AVG	146	129	119.2	100.6	93.2	83	77.4	69.8	64.8	60.2	55.4
STD	1.22	1.58	1.64	1.58	1.48	1.58	1.14	0.83	1.48	0.83	1.14

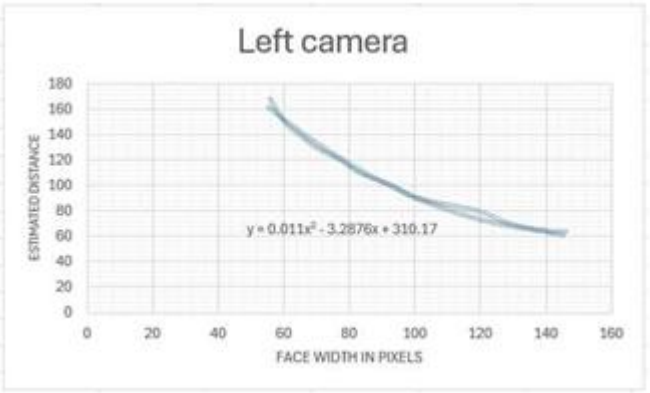


Figure 6. Left Camera Profile

Finding the Distance The depth distance Z is determined by averaging the two estimations using the aforementioned polynomials (Table III). The average depth is determined using a number of test lengths (65, 75, 115, 125, 140, and 170 cm), and its error values are quite low when compared to each estimate (Figure 7). The fact that the inaccuracy owing to discrete error grows as the face distance rises should also be noted.

Table III Depth Calucation.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

	Expected distance 65cm			Error	Expected distance 75cm			Error
	Right Est.	Left Est.	Z Cal.		Right Est.	Left Est.	Z Cal.	
	66.09	63.34	64.71	0.44	70.91	73.81	72.36	3.52
	62.58	63.9	65.24	0.81	70.26	76.49	73.38	1.41
	64.51	61.44	64.48	1.16	70.91	74.46	72.69	0.94
	65.02	61.57	63.44	1.61	71.58	73.15	72.31	0.52
	64.19	6404	64.06	0.97	72.26	75.88	74.05	2.41
AVG	64.47	62.85	64.38	1.00	71.18	74.76	72.96	1.76
STD	1.28	1.26	0.67	0.43	0.76	1.40	0.74	1.21

	Expected distance 115cm			Error	Expected distance 125cm			Error
	Right Est.	Left Est.	Z Cal.		Right Est.	Left Est.	Z Cal.	
	114.51	111.59	113.05	1.70	121.08	116.8	119	4.84
	111.98	112.42	112.30	0.66	121.08	120.05	120.6	1.35
	108.29	113.06	110.97	1.18	116.41	116.82	117.6	2.45
	111.98	112.62	110.30	0.60	119.41	117.88	119.8	1.87
	117.1	116.85	116.96	6.04	119.78	116.82	118.3	1.28
AVG	112.77	113.31	112.72	2.04	119.55	117.67	119.04	2.36
STD	3.28	2.05	2.61	2.28	1.91	1.41	1.18	1.46

	Expected distance 140cm			Error	Expected distance 170cm			Error
	Right Est.	Left Est.	Z Cal.		Right Est.	Left Est.	Z Cal.	
	130.86	125.60	128.23	8.41	150.81	140.99	145.99	14.12
	129.42	125.60	127.10	0.88	152.44	146.01	146.01	0.01
	130.86	123.35	130.10	2.36	159.10	149.86	149.86	2.64
	133.77	127.88	130.83	0.56	150.81	140.59	140.59	6.19
	130.86	123.35	127.10	2.85	154.08	142.23	142.23	1.17
AVG	131.15	125.16	128.67	3.01	153.45	143.94	144.94	4.83
STD	1.59	1.89	1.72	3.17	3.44	3.94	3.63	5.69

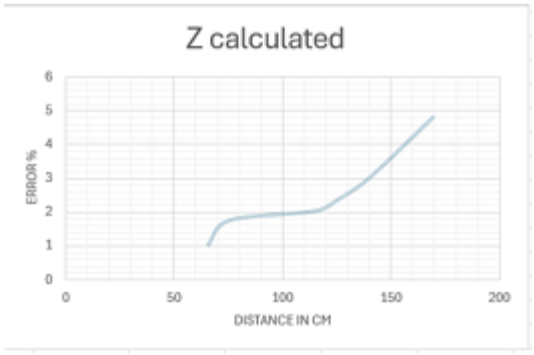


Figure 7. Depth Calculation Error.

Examples for Evaluation A example test of an actual application run is shown in the accompanying figure. Figure 8 shows that although the right camera records all three faces, the left camera only records two.

Figure 8. Testing Sample of multiple people in the view.

After the Haar Cascade detection phase, the three faces in the right camera are trained using Eigenface, and the left camera recognizes matched faces. After that, you can figure out the depth of the faces, and to bring in the third person, you need turn the left camera clockwise.

CONCLUSION

The ability to precisely pinpoint human positions in three dimensions is crucial in several computer vision applications. This study's overarching objective is to improve the Arslan humanoid robot's visual system so that it can maintain constant visual contact with its surroundings. Based on the same principles as human binocular vision, the system uses two identical USB cameras and two pan cameras calibrated with servo motors to precisely match their angles in order to extract depth information from the pictures. It consists of taking photos, detecting faces, comparing the amount of faces from different cameras, and executing 3D localization. The primary goals are to make and keep eye contact with individuals, to precisely determine their precise positions, and to estimate their approximate places. Recent methodological developments have greatly enhanced localization accuracy, leading to a significant decrease in mistakes and an increase in dependability. These advancements improve computer vision technology as a whole by allowing for more intuitive and responsive interactions between computers and humans.

REFERENCES

[1]. Pan, N. H. W., Li, N. Y., Gao, N. C. M., & Lei, N. Y. (2013). A method for Multi-Cameras Human Tracking Based on Reverse Epipolar Line Geometry.

ISSN 2454-5007, www.ijmm.net
Vol. 14 Issue. 2, April 2026

IEEE Conference Anthology. https://doi.org/10.1109/anthology.2013.6784877

[2]. Siddique, T. H. M., Rehman, Y., Rafiq, T., Nisar, M. Z., Ibrahim, M. S., & Usman, M. (2021). 3D Object Localization Using 2D Estimates for Computer Vision Applications.

[3]. Cıvır, C., & Topal, C. (2019). Robust Landmark Selection for 3D Face Pose Estimation.

[4]. Ariz, M., Villanueva, A., & Cabeza, R. (2019). Robust and accurate 2D-tracking-based 3D positioning method: Application to head pose estimation. Computer Vision and Image Understanding, 180, 13–22.

[5]. Gao, B. K., Zhang, Z. X., Wu, C. N., Wu, C. L., & Bi, H. B. (2022b). Staged cascaded network for monocular 3D human pose estimation. Applied Intelligence, 53(1), 1021–1029. https://doi.org/10.1007/s10489-022-03516-1

[6]. I. Elaff, "Robotic Information System (RIS): Design of Humanoid Robot's Head Based on Human Biomechanics", El-Cezeri Journal of Science and Engineering, vol. 10, no. 2, pp. 420–432, 2023, doi: 10.31202/ecjse.1249294. https://doi.org/10.31202/ecjse.1249294

[7]. Perez, H.; Tah, J.H.M."Towards Automated Measurement of As-Built Components Using Computer Vision." Sensors 2023, 23:7110. https://doi.org/10.3390/s23167110

[8]. Wei Chen, Yan Li, Zijian Tian, Fan Zhang, "2D and 3D object detection algorithms from images: A Survey" Array, 2023,19:100305, https://doi.org/10.1016/j.array.2023.100305.

[9]. Ahsan, M.M.; Li, Y.; Zhang, J.; Ahad, M.T.; Gupta, K.D. Evaluating the Performance of Eigenface, Fisherface, and Local Binary Pattern Histogram-Based Facial Recognition Methods under Various Weather Conditions. Technologies 2021, https://doi.org/10.3390/technologies9020031