Journal of Intelligent Systems and Internet of Things
Vol. 10, No. 02, PP. 90-101, 2023
DOI: https://doi.org/10.54216/JISIoT.100208

3D Hand Pose and Shape Estimation from Single RGB Image for Augmented Reality

Enas Kh. Hassan[1], Jamila Harbi S.[2]
[1] Remote Sensing and GIS Department, College of Science, University of Baghdad, Baghdad, Iraq
[2] Computer Science Department, College of Science, Mustansiriyah University, Baghdad, Iraq
Emails: enas.mkhazal@gmail.com; dr.jameelahharbi@gmail.com
*Corresponding Author: enas.mkhazal@gmail.com

Abstract

In the realm of Human-Computer Interaction (HCI), the importance of hands cannot be overstated. Hands serve as a fundamental means of communication, expression, and interaction in the physical world. In recent years, Augmented Reality (AR) has emerged as a next-generation technology that seamlessly merges the digital and physical worlds, providing transformative experiences across various domains. In this context, accurate hand pose and shape estimation plays a crucial role in enabling natural and intuitive interactions within AR environments. Augmented Reality, with its ability to overlay digital information onto the real world, has the potential to revolutionize how we interact with technology. From gaming and education to healthcare and industrial training, AR has opened up new possibilities for enhancing user experiences. This study proposes an innovative approach for hand pose and shape estimation in AR applications. The methodology commences with the utilization of a pre-trained Single Shot Multi-Box (SSD) model for hand detection and cropping. The cropped hand image is then transformed into the HSV color model, followed by applying histogram equalization on the value band. To precisely isolate the hand, specific bounds are set for each band of the HSV color space, generating a mask. To refine the mask and diminish noise, contouring techniques are applied to the mask, and gap-filling methods are employed. The resultant refined mask is then combined with the original cropped image through logical AND operations to accurately delineate the hand boundaries. This meticulous approach ensures robust hand detection even in complex scenes. To extract pertinent features, the detected hand undergoes two concurrent processes. Firstly, the Scale-Invariant Feature Transform (SIFT) algorithm identifies distinctive keypoints on the hand's outer surface. Simultaneously, a pre-trained lightweight Convolutional Neural Network (CNN), namely MobileNet, is employed to extract 3D hand landmarks, the hand's center (middle finger metacarpophalangeal joint), and handedness information. These extracted features, encompassing hand keypoints, landmarks, center, and handedness, are aggregated and compiled into a CSV file for further analysis. A Gated Recurrent Unit (GRU) is then employed to process the features, capturing intricate dependencies between them. The GRU model successfully predicts the 3D hand pose, achieving high accuracy even in dynamic scenarios. The evaluation results for the proposed model are very promising that the Mean Per Joint Position Error in 3D (MPJPE) is 0.0596 between the predicted pose and the ground truth hand landmarks, while the Percentage of Correct Keypoints (PCK) is 95%. Upon predicting the hand pose, a mesh representation is employed to reconstruct the 3D shape of the hand. This mesh provides a tangible representation of the hand's structure and orientation, enhancing the realism and usability of the AR application. By integrating sophisticated detection, feature extraction, and predictive modeling techniques, this method contributes to creating more immersive and intuitive AR experiences, thereby fostering the seamless fusion of the digital and physical worlds.

Received: April 22, 2023 Revised: July 13, 2023 Accepted: October 11, 2023

Keywords: Pose estimation; Shape Estimation; Hand; 3D space; HSV; SIFT; GRU; MANO

1. Introduction

The human hand is an intricate and versatile tool, playing a pivotal role in how we interact with the world. From mundane tasks to intricate gestures, the hand's dexterity and expressiveness are unparalleled [1]. This significance has led to an increasing emphasis on integrating hand gestures and poses into human-computer interaction (HCI) systems. The advent of augmented reality (AR) further amplifies the importance of accurate hand pose and shape estimation [1].

The ability to estimate accurately the 3D shape and pose of the human hand is a critical frontier. This estimation enables natural and intuitive interactions in AR environments. Tasks such as virtual object manipulation, hand gesture recognition, and immersive experiences all rely on the capability to reconstruct the hand's position, orientation, and articulation. Achieving this level of accuracy and realism in hand pose and shape estimation has become a cornerstone in creating seamless and immersive AR applications [2].

2. Related Work

Recent years have witnessed remarkable progress in 3D hand shape and pose estimation, largely propelled by the surge in deep learning techniques and the accessibility of affordable depth sensors. These advancements have opened avenues to capture intricate hand poses and shapes from real-world high-definition images. Here are some leading researches in the field for the last few years:
Liang et.al, [1] presented a novel approach to optimize the leaf weights in a Hough forest to aid global hand pose estimation with a single depth camera. Unlike traditional Hough forest, they propose to learn the vote weights stockpiled at the leaf nodes of a forest in an upright way to minimalize average shape prediction error, so that hazy votes are largely inhibited during prediction fusion, according to the results using optimized leaf weights improved the pose estimation on both real and synthetic image datasets. Ref33 et. al, proposed a tracking method that combines a CNN with a kinematic 3D hand model, which led to better generalization of unseen data, their approach is robust to varying camera viewpoints and occlusions, and leads to anatomically acceptable as well as temporally smooth hand motions. They proposed a novel approach for the generation of synthetic training data based on a geometrically coherent image-to-image translation network for training CNN. Ref34 et.al, introduced an approach for real-time, accurate and robust hand pose estimation from moving egocentric RGB-D cameras in chaotic real environments. This approach employees two successively applied CNNs to localize the hand and regress 3D joint locations. Hand localization is accomplished by using a CNN to estimate the 2D position of the hand center in the input image, even in the existence of occlusions and chaos. The combination of localized hand position and the corresponding input depth value is utilized to generate a normalized cropped image to be fed into a second CNN to regress relative 3D hand joint locations in real time. For added robustness, accuracy, and temporal stability, they refine the hand pose estimation using a kinematic shape tracking energy. This approach proved the ability of achieving low errors even under scene clutter and difficult occlusions. Remelli et. al, introduced a robust methodology for the personalization of sphere-mesh tracking model of user using a collection of depth measurements. Building and performance of shape-space is comparable to shape-spaces composed from datasets of carefully standardized models by reparametrizing the geometry of the tracking template as a first step, and introducing a multi-stage calibration optimization. Their parameterization decouples the DoF for pose and shape, consequentially improving the convergence properties. Analytically differentiable multi-stage standardization pipeline optimizes for the model in the natural low-dimensional space of local anisotropic scaling, leading to an efficient solution easily embedded in other tracking/ standardization algorithms [2]. Spurr et.al, proposed an approach to learn a statistical hand model characterized by a cross-modal trained latent space via a generative deep neural network. Using an objective function from the variation lower bound of the variational auto-encoder (VAE) framework and conjointly optimize the resulting cross-modal Kullback-Leibler (KL) the posterior reconstruction objective and divergence, naturally conceding a training regimen that leads to a comprehensible latent space across several modalities such as 2D key point detections, RGB images, or 3D hand configurations. Furthermore, it concedes a straightforward manner of using semi supervision. This latent space can be used immediately to estimate 3D hand poses from RGB images [3]. Zimmermann et.al, propose a deep network that learns a network-implicit 3D articulation prior. Along with detected key points in the images, this network produces good estimates of the 3D shape. Introducing a large-scale 3D hand pose dataset based on synthetic hand models for training the implicated networks. Since the performance of their system based on image annotation, the system yielded unpromising results with the lack of annotated large-scale real world image dataset and diverse pose statistics [4]. Ge et.al, suggested 3D hand shape estimation system that takes a depth image of a hand as the input and outputs a set of 3D hand joint. The hand depth image is mapped to a set of 3D points. The 3D point set is down sampled and normalized in an oriented bounding box (OBB). A hierarchical PointNet appropriates N points as the input to obtain hierarchical hand features and regress the

3D hand shape. Moreover, for improving the estimation accuracy of the locations of the fingertip, a fingertip refinement network is deliberated [5]. Ge et.al, suggested a multi-view CNN-based method for 3D hand shape estimation. To exploit better 3D information in the depth image, projection of the point cloud produced from the query depth image onto multiple views of two projection settings and incorporate them for robust shape estimation. Training multi-view CNNs to learn the mapping from projected images to heat maps, which replicate probability distributions of joints on each view. Multi-view heat-maps are then compound to estimate optimal 3D hand shape with learned shape priors, while the erratic information in multi-view heat-maps is concealed using a view selection method [6]. Ge et.al, suggested a Point-to-Point Regression PointNet that takes directly the 3D point cloud as an input and produces point-wise estimations, such as, unit vector fields on the point cloud, heat-maps, representing the imminence and direction from every point in the point cloud to the hand joint. The estimations of the point-wise are used to conclude 3D joint locations with weighted fusion. Stacked network architecture for PointNet with intermediate supervision is applied to improve the capturing of 3D spatial information in the point cloud [7]. Malik et.al, introduced a fully supervised deep neural network, which learns to estimate jointly a full 3D hand mesh interpretation and shape from a single depth image. Which improves the CNN architecture used to estimate parametric depictions such as bone scales, hand shape, and complex shape parameters. Afterwards, hand shape layer, included inside the deep framework, results hand mesh, and 3D joint positions [8]. Panteleris et.al, utilize the latest innovations of deep learning, fusing them with the power of generative hand shape estimation techniques to achieve real-time monocular 3D hand shape estimation in unobstructed scenarios. Pre-trained network of OpenPose is used for hand cropping in the image, estimating the 2D joint locations of hand. Afterwords, non-linear least-squares minimization fits a 3D model of the hand to the estimated 2D joint positions, producing the 3D hand pose [9]. Ge et.al, proposed an algorithm to extract Image-based features by 2D CNNs that are not directly suitable for 3D hand shape estimation due to the absence of 3D spatial information. The proposed 3D CNN-based method, taking a 3D volumetric depiction of the hand depth image as input and obtaining 3D features from the volumetric input, consequentially capturing the 3D spatial formation of the hand and precisely regress full 3D hand shape in a single pass. 3D data augmentation is performed in order to make the 3D CNN robust to variations in global orientations, and hand sizes on the training data. Applying the 3D deep network structure and leveraging the comprehensive hand surface as transitional supervision for learning 3D hand shape from depth images [10]. Taylor et.al, propose a method for 3D hand shape estimation from a monocular image through 2.5D shape depiction. The depiction estimates shape up to a scaling factor that can be estimated as well if the hand size is given in priory. CNN architecture is used to learn depth maps and heat map distributions. Ref53 et.al, suggest to leverage the depth images that can be simply attained from commodity RGB-D cameras through training, while during testing phase, only RGB inputs for 3D joint predictions is considered. This help lessen the burden of the costly 3D observations in real-world dataset. The weakly supervised method, adapting from fully- glossed synthetic dataset to weakly labeled real-world dataset with the regularization of depth that generates depth maps from projected 3D shape and works as weak supervision for 3D shape regression [11]. Boukhayma et.al, present an end-to-end deep learning methodology that foresees 3D hand shape from RGB images in real world. The network comprises of a fixed model-based decoder, and a deep convolutional encoder. The encoder foresees a set of view and hand parameters using an input image and 2D hand joint locations obtained from a separate CNN. The decoder encompasses two components: A pre-computed articulated mesh distortion hand model that produces a 3D mesh from the hand parameters, and a re-projection module managed by the view parameters that projects the produced hand into the image domain [12]. Ge et.al, proposed generating a full 3D mesh and 3D hand joint locations of the hand directly from a single monocular RGB image. Explicitly, the input is a single RGB image focused on a hand, which is passed throughout two-stacked hourglass networks to deduce 2D heat-maps. The estimated 2D heat-maps, along with the image feature maps, are coded as a dormant feature vector using a residual network that encloses eight residual layers and four max pooling layers. The encoded dormant feature vector is then the input to the Graph CNN to deduce the 3D coordinates of hand mesh. Ref61 et.al, suggested an approach known as pose guided structured region ensemble network (Pose-REN) to increase the performance of hand shape estimation. Underneath the supervision of an primarily estimated shape, this approach extricates boroughs from the feature maps of CNN and produces more ideal and delegate features for hand shape estimation. The feature regions are then combined hierarchically according to the topology of hand joints by tree-structured fully connected to revert the cultivated hand shape. The latent hand shape is attained by an iterative cascaded algorithm [13]. Guo et.al, proposed an end-to-end network for forecasting 3D hand shape from a single RGB image. By extracting, several feature maps from different resolutions and generate parallel feature fusion, and produce prototype for graph based convolutional neural network section to infer the initial 3D hand key points. Subsequently, used3D geometric knowledge and 2D spatial relationships to build a self-supervised module to diminish domain gaps between 2D and 3D space. Finally, the final 3D hand pose is computed by averaging the 3D hand shapes from the graph convolutional neural network output and the self-supervised module output [14]. Cai et.al, suggested leveraging the depth images that are effortlessly attained from RGB-D cameras through training, while through testing RGB inputs are taken only for 3D joint predictions. Alleviating the encumbrance of the expensive 3D remarks in real-world dataset. This proposes a weakly supervised approach, adapted from fully marked synthetic dataset to weakly labeled real-world single RGB dataset with the support of a depth regularization that operates as weak supervision for 3D shape estimation. To utilize more the physical scheme of 3D hand shape, novel CVAE-based statistical structure is proposed to include the pose-specific subspace from RGB images that can then be employed to predict the 3D hand joint locations [15].

In light of these challenges, this work aims to contribute to the advancements in 3D hand pose and shape estimation for AR environments. The following is the outline of the article. Section 2 defines the 3D hand pose and shape estimation from single RGB image. Section 3 discusses the steps followed to extract features, train the model and construct the hand shap. Section 4 contains the experimental findings. The proposed 3D hand pose and shape estimation, which was implemented in Python, is concluded in Section 5.

3. Methodology

To address the challenge of 3D hand shape estimation for augmented reality applications, a comprehensive approach combining computer vision techniques and deep learning methodologies is proposed. The methodology comprises several stages, each contributing to accurate hand pose and shape estimation from a single RGB image Figure (1) shows the proposed system structure:

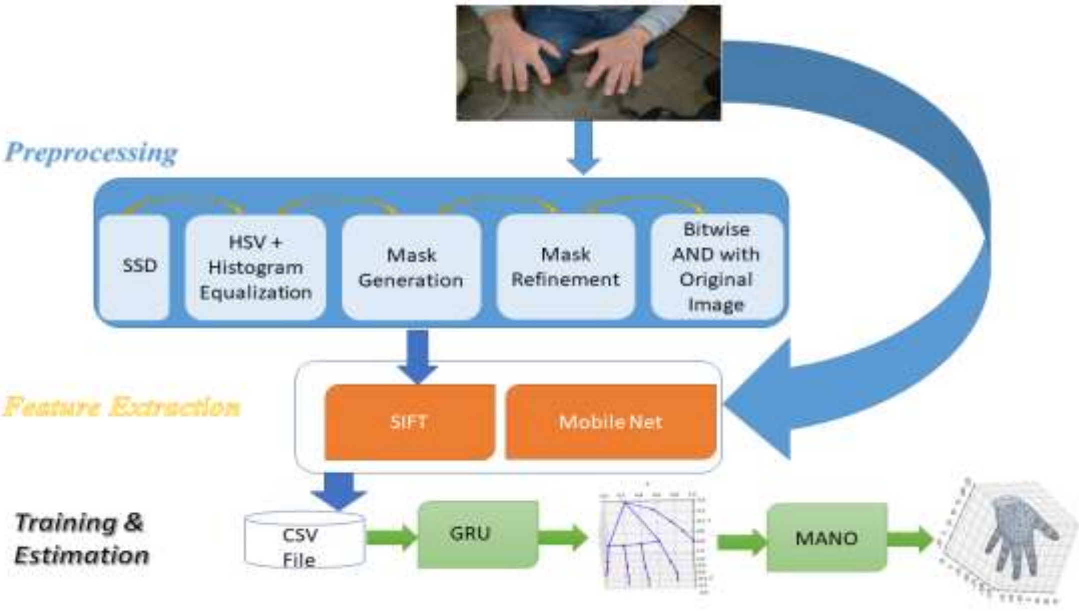

Figure 1: the proposed system structure

1. Data Preprocessing:

Hand Detection and Cropping: A pre-trained Single Shot Multi-Box Detector (SSD) is employed to detect and localize the hand object within the RGB image [16]. The detected hand region is then cropped for further processing, as showing in figure 1.

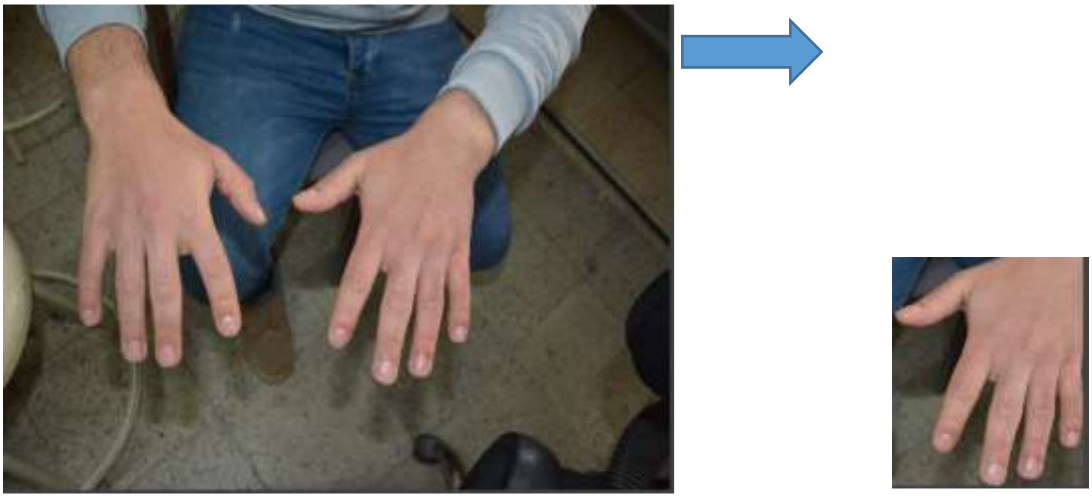

Figure 1: Hand Object Cropping

Color Space Transformation: The cropped hand region is converted to the HSV color space. This transformation facilitates better handling of lighting variations and enhances the visibility of the hand's features [17].

Histogram Equalization: Histogram equalization is applied to the value channel of the HSV image. This step enhances the contrast of the hand object, improving the distinctiveness of its features, as shown in figure (2).

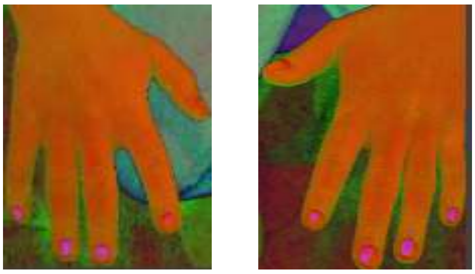

Figure 2: HSV and Histogram Equalization for Value channel

Mask Generation: Based on specific bounds set on the hue, saturation, and value bands of the HSV image, a mask is generated to isolate the hand object from the background, as shown in Figure (3). These specific bounds are set after comprehensive experiments to choose the best combination [18]:

$$0 \geq H \geq 20 \tag{1}$$

$$20 \geq S \geq 255 \tag{2}$$

$$70 \geq V \geq 255 \tag{3}$$

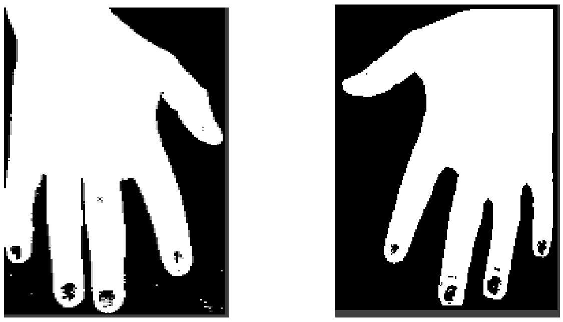

Figure 3: Mask Generation

To reduce noise and highlight the hand's boundaries image Contouring and Flood Filling is used, the generated mask undergoes image contouring to refine the mask boundaries. Subsequently, a gap-filling process is employed to eliminate any remaining noise and inconsistencies in the mask [19] [20]; the result of the refined mask is shown in Figure (4).

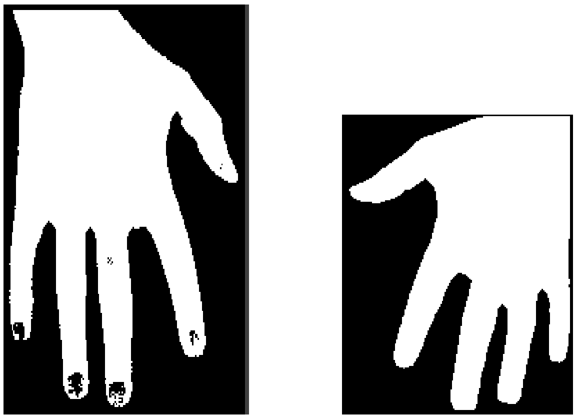

Figure 4: Refined Hand Mask

Mask and Image Fusion: The refined mask is combined with the original cropped image using a bitwise AND operation. This step effectively extracts the hand object from the background, as shown in Figure (5).

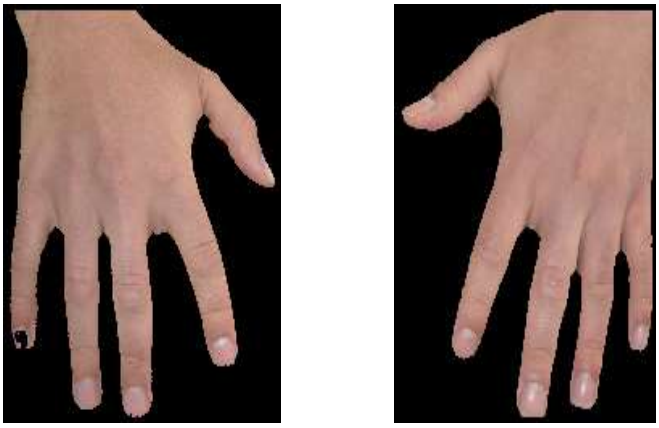

Figure 5: Hand object Detection

2. Feature Extraction:

Keypoint Detection with SIFT: Scale-Invariant Feature Transform (SIFT) is applied to the hand image to detect distinctive keypoints on the hand's outer surface. These keypoints capture salient regions and aid in subsequent analysis [21], the detected keypoints are shown in Figure (6).

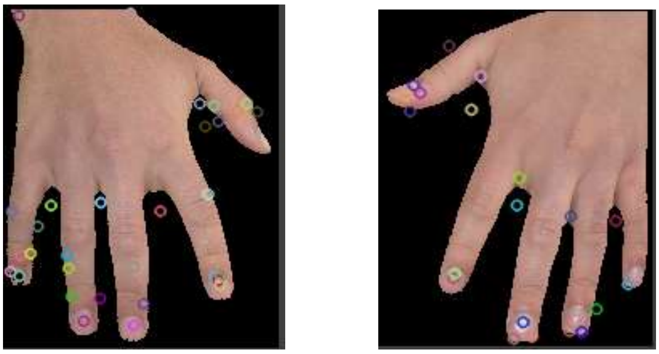

Figure 6: SIFT Detected Keypoints

MobileNet-based Feature Extraction: A pre-trained, lightweight convolutional neural network MobileNet, is employed to extract 3D hand landmarks, the hand's center (middle finger MCP), and handedness [1]. MobileNet's efficiency makes it suitable for real-time applications [22], as shown in figure (7) below:

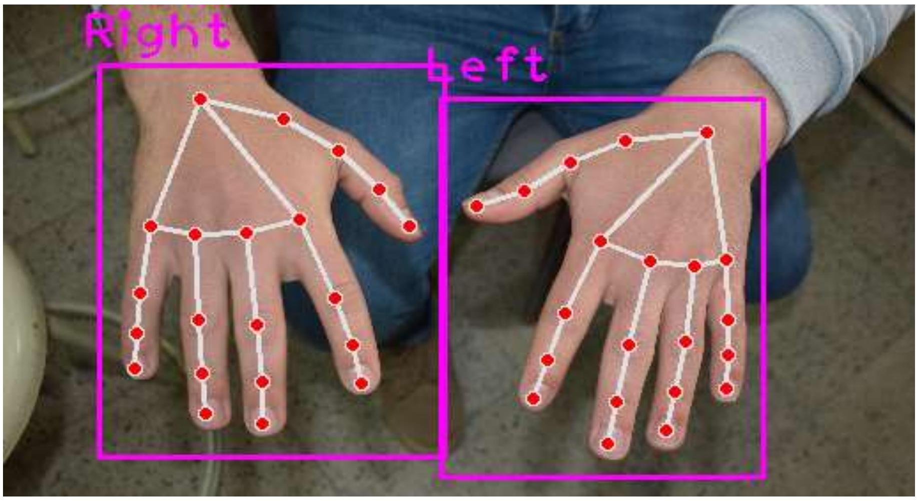

Figure 7: Hand landmarks and handedness information

3. Feature Fusion and Sequence Preparation:

CSV Feature Storage: The detected SIFT keypoints, hand center (MCP), 3D hand landmarks, and handedness information are all stored in a structured CSV file. This file serves as input for the subsequent stages. The stored features are organized into sequences, capturing dependencies that represent the evolution of the hand pose.

4. Deep Learning Model:

GRU-based Model: A Gated Recurrent Unit (GRU) is employed as the core of the deep learning model. GRUs are adept at capturing sequential dependencies in data.

Feature Sequence Input: The prepared sequences from the CSV file are fed into the GRU model. The model learns to capture the intricate relationships between keypoints, landmarks, hand center, and handedness. To train the model a group of hyper parameters are chosed (epochs:50, Batch Size: 512, GRU input layer: 128, Dense Unit: number of joints*3), for the update gate Tanh activation is used while for forget gate sigmoid activation is used [23].

The trained GRU model outputs predicted 3D hand poses. These predictions encapsulate the hand's orientation, articulation, and pose, as shown in figure (8).

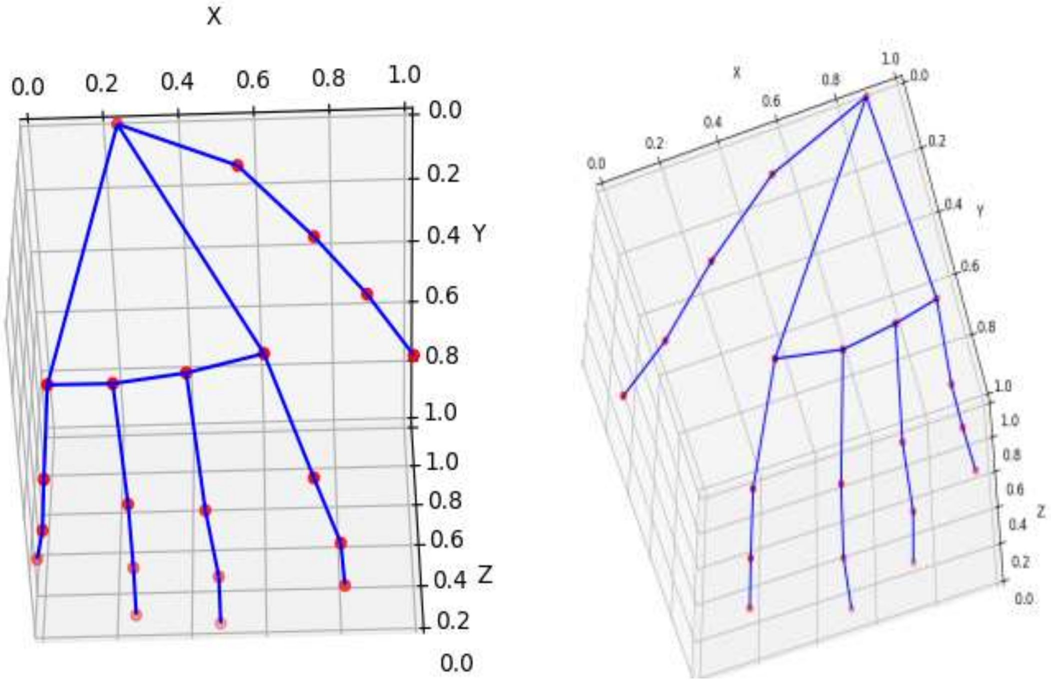

Figure 8: Estimated 3D hand Pose

By utilizing pre-trained MANO (hand Model with Articulated and Non-rigid deformations) [24]model after calculating the hand shape parameters (betas) and the pose parameters calculated from predicted joints locations and the topology connecting them based on hand anatomy a hand mesh was created. This mesh is composed of vertices and triangles that collectively depict the hand's shape [1]. The generated mesh is then can be visualized in augmented reality environments, providing users with a representation of their hand's pose and shape, as shown in Figure (9).

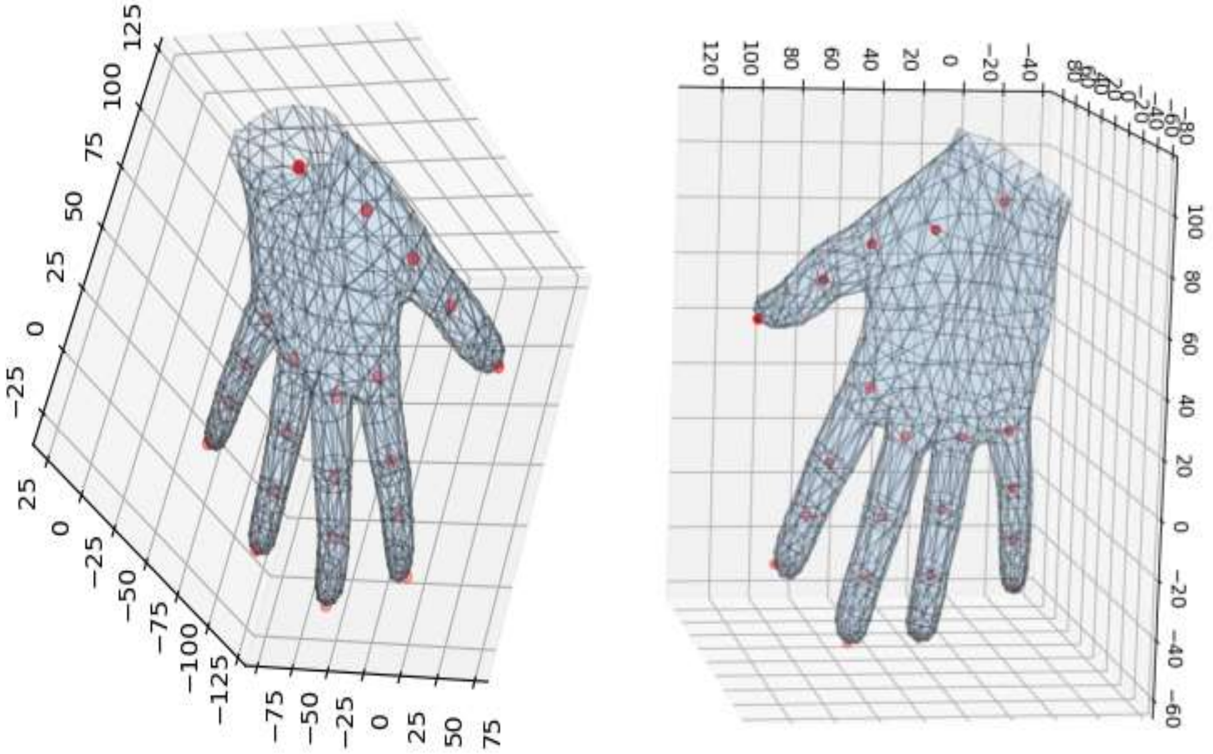

Figure 9: Estimated 3D hand Shape

In conclusion, the proposed methodology encompasses several key stages, from data preprocessing to deep learning-based feature extraction and prediction. By leveraging a combination of computer vision techniques and deep learning models, the approach aims to accurately estimate the 3D pose and shape of the human hand from a single RGB image. This estimation is vital for creating immersive and natural interactions in augmented reality applications.

5. Evaluation Metrics

The evaluation of the 3D hand pose and shape estimation model involves using appropriate metrics that measure the accuracy and performance of the predicted results compared to ground truth data. Several metrics commonly used in this context include:

1. Mean Per Joint Position Error(MPJPE)

Mean Per Joint Position Error (MPJPE) is a commonly used evaluation metric in computer vision and pose estimation. It serves to measure the average Euclidean distance between the ground truth joint positions and the predicted joint positions. The metric operates on a set of N joints, where each joint is represented by a 3D coordinate (x, y, z) in the world coordinate system [25].

To calculate MPJPE, we start by obtaining the predicted joint positions, denoted as P = {P1, P2, ..., PN}, from a pose estimation algorithm using a given input data. Similarly, we have the ground truth joint positions, denoted as G = {G1, G2, ..., GN}, for the same input data. Pi represents the predicted position of joint i, and Gi represents the ground truth position of joint i.

The MPJPE is then computed as the average Euclidean distance between the predicted and ground truth joint positions across all joints. It can be expressed by the equation:

$$\text{MPJPE} = (1/N) * \Sigma(||Pi - Gi||) \qquad (4)$$

, where ||.|| denotes the Euclidean distance between two 3D points. The summation is performed over all N joints. The qualitative results of using this metric on single image is 0.01648 between predicted pose and the ground truth landmarks.

2. Percentage of Correct Keypoints (PCK)

Percentage of Correct Keypoints (PCK) is a broadly used evaluation metric in computer vision, specifically designed to assess the accuracy of pose estimation. It measures the percentage of keypoints that are correctly localized based on a predefined threshold distance. PCK provides a quantitative measure of an algorithm's ability to accurately estimate the positions of keypoints within an image.

PCK is typically calculated by counting the number of predicted keypoints that fall within a certain distance threshold from their corresponding ground truth keypoints. The threshold is usually defined relative to a reference length, such as the torso length or the diagonal of the bounding box. The percentage is then computed by dividing the count of correctly localized keypoints by the total number of keypoints and multiplying by 100%.

$$\text{PCK} = (\text{Number of correctly localized keypoints}) / (\text{Total number of keypoints}) * 100\%. \quad (5)$$

The resulting value represents the percentage of keypoints correctly localized by the algorithm within the specified threshold. After setting the number of points in training phase for 21, this makes the PCK metric value is 100% that the detected joints locations are perfectly detected [26].

The evaluation metrics used on selected images of both datasets RHD_published_v2 and the self-collected Two hand dataset are shown in table (1) below:

Table 1: MPJPE and PCK on sample tested images

Image ID	Type	Dataset	PCK	MPJPE
DSC_6067	**Right**	**Ours**	**95%**	**0.0596**
DSC_6067	**Left**	**Ours**	**66%**	**0.0922**
DSC_6423	**Right**	**Ours**	**90%**	**0.0682**

DSC_6423	**Left**	**Ours**	**85%**	**0.0667**
DSC_6646	**Right**	**Ours**	**76%**	**0.0712**
DSC_6646	**Left**	**Ours**	**95%**	**0.0608**
DSC_6772	**Right**	**Ours**	**71%**	**0.0792**
DSC_6772	**Left**	**Ours**	**---**	**---**
03728	**Right**	**RHD_published_v2**	**80%**	**0.0711**
03728	**Left**	**RHD_published_v2**	**61%**	**0.0889**
01749	**Right**	**RHD_published_v2**	**85%**	**0.0637**
01749	**Left**	**RHD_published_v2**	**76%**	**0.0798**

6. Comparison to Previous Results

The efficacy of any novel approach in this field is often evaluated through rigorous comparisons with existing methods and benchmarks. Such comparisons not only establish the state-of-the-art but also shed light on the advancements achieved, limitations encountered, and the potential avenues for further refinement.

In this study, we delve into the realm of hand pose estimation, building upon the foundations laid by previous works in the domain. We present an in-depth analysis of our methodology and results, juxtaposed against the backdrop of pioneering studies that have paved the way for advancements in hand pose estimation. Our aim is to provide a comprehensive assessment of the current state of the art by considering the strengths and weaknesses of both our approach and those that precede it, Table (2) show the results conducted compared to previous studies.

Table 2: MPJPE and PCK of our method compared to previous results

Method	**Dataset**	**MPJPE**	**PCK**
Ours	**Two-Hands Dataset**	**0.646**	**95.238**
Ours	**RHD-published-v2**	**0.1344**	**28.571**
[12]	**MPII+NZSL**	**---**	**51.87**
[29]	**Stereo**	**---**	**55**
[4]	**S-val**	**---**	**30**

7. Conclusion

In conclusion, this study focused on the critical task of 3D hand pose and shape estimation for augmented reality applications. The importance of accurate hand shape estimation in human-computer interaction, augmented reality, and various other domains was highlighted. The rapid advancements in deep learning and the availability of depth sensors have propelled the development of accurate hand shape estimation models. However, challenges persist due to variations in hand shapes, occlusions, viewpoint changes, and limited labeled data. The proposed approach, which combines pretrained models for hand detection, color space transformation, and neural networks for 3D landmark extraction, demonstrated promising results. The use of machine learning techniques, such as CNNs and GRUs, allowed us to predict 3D hand poses with reasonable accuracy. The integration of these predictions with 3D mesh generation techniques produced visually plausible hand shapes. Although this approach is well defining an accurate hand pose but still give less accurate results when it come to the shape of hand with difficult hand pose.

References

[1] Liang, H., Yuan, J., Lee, J., Ge, L. and Thalmann, D., 2017. Hough forest with optimized leaves for global hand pose estimation with arbitrary postures. *IEEE Transactions on Cybernetics*, *49*(2), pp.527-541.

[2] Obeid, N. (2023). On The Product and Ratio of Pareto and Erlang Random Variables. International Journal of Mathematics, Statistics, and Computer Science, 1, 33–47. https://doi.org/10.59543/ijmscs.v1i.7737

[3] Spurr, A., Song, J., Park, S. and Hilliges, O., 2018. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 89-98).

[4] Zimmermann, C. and Brox, T., 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision* (pp. 4903-4911).

[5] Ge, L., Cai, Y., Weng, J. and Yuan, J., 2018. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8417-8426).

[6] Ge, L., Liang, H., Yuan, J. and Thalmann, D., 2018. Robust 3D hand pose estimation from single depth images using multi-view CNNs. *IEEE Transactions on Image Processing*, *27*(9), pp.4422-4436.

[7] Ge, L., Ren, Z. and Yuan, J., 2018. Point-to-point regression pointnet for 3d hand pose estimation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 475-491).

[8] Malik, J., Elhayek, A., Nunnari, F., Varanasi, K., Tamaddon, K., Heloir, A. and Stricker, D., 2018, September. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In *2018 International Conference on 3D Vision (3DV)* (pp. 110-119). IEEE.

[9] Panteleris, P., Oikonomidis, I. and Argyros, A., 2018, March. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 436-445). IEEE.

[10] Ge, L., Liang, H., Yuan, J. and Thalmann, D., 2018. Real-time 3D hand pose estimation with 3D convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, *41*(4), pp.956-970.

[11] Taylor, J., Stebbing, R., Ramakrishna, V., Keskin, C., Shotton, J., Izadi, S., Hertzmann, A. and Fitzgibbon, A., 2014. User-specific hand modeling from monocular depth sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 644-651).

[12] Boukhayma, A., Bem, R.D. and Torr, P.H., 2019. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10843-10852).

[13] Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J. and Yuan, J., 2019. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10833-10842).

[14] Guo, S., Rigall, E., Qi, L., Dong, X., Li, H. and Dong, J., 2020. Graph-based CNNs with self-supervised module for 3D hand pose estimation from monocular RGB. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(4), pp.1514-1525.

[15] Cai, Y., Ge, L., Cai, J., Thalmann, N.M. and Yuan, J., 2020. 3D hand pose estimation using synthetic data and weakly labeled RGB images. *IEEE transactions on pattern analysis and machine intelligence*, *43*(11), pp.3739-3753.

[16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.

[17] Hema, D. and Kannan, D.S., 2019. Interactive color image segmentation using HSV color space. *Sci. Technol. J*, *7*(1), pp.37-41.

[18] Hassan, E.K. and Saud, J.H., 2023, February. HSV color model and logical filter for human skin detection. In *AIP Conference Proceedings* (Vol. 2457, No. 1). AIP Publishing.

[19] Arbelaez, P., Maire, M., Fowlkes, C. and Malik, J., 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *33*(5), pp.898-916.

[20] C. Bond, "An Efficient and Versatile Flood Fill Algorithm for Raster Scan Displays," 2011.

[21] Wu, J., Cui, Z., Sheng, V.S., Zhao, P., Su, D. and Gong, S., 2013. A Comparative Study of SIFT and its Variants. *Measurement science review*, *13*(3), pp.122-131.

[22] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[23] Abbas, S.K. and George, L.E., 2020. The Performance Differences between Using Recurrent Neural Networks and Feedforward Neural Network in Sentiment Analysis Problem. *Iraqi Journal of Science*, *61*(6).

[24] Romero, J., Tzionas, D. and Black, M.J., 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

[25] Habibie, I., Xu, W., Mehta, D., Pons-Moll, G. and Theobalt, C., 2019. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10905-10914).

[26] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C. and Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4903-4911).

[27] Sharma, R.P. and Verma, G.K., 2015. Human computer interaction using hand gesture. *Procedia Computer Science*, *54*, pp.721-727.

[28] Aliprantis, J., Konstantakis, M., Nikopoulou, R., Mylonas, P. and Caridakis, G., 2019, January. Natural Interaction in Augmented Reality Context. In *VIPERC@ IRCDL* (pp. 50-61).

[29] Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D. and Theobalt, C., 2018. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 49-59).

[30] Abbas, A.H., Arab, A. and Harbi, J., 2018. Image compression using principal component analysis. *Mustansiriyah Journal of Science*, *29*(2), p.01854.

[31] Abbas, A.H., 2011. Mathematical Morphology Operations on Grayscale Image. *Journal of the College of Basic Education*, *17*(67), pp.105-115.

[32] Abdullah, R.M., Alazawi, S.A.H. and Ehkan, P., 2023. SAS-HRM: Secure Authentication System for Human Resource Management. *Al-Mustansiriyah Journal of Science*, *34*(3), pp.64-71.