IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13477

An Enhanced Method to Detect Hand Key-points in Single Images using Multiview Bootstrapping

Mohammad Hasan[1], Montasim Al Mamun[2], Abid Hasan[3]

Department of CSE, Bangamata Sheikh Fojilatunnesa Mujib Science and Technology University, Jamalpur[1]

Department of CSE, BAUST, Saidpur[2]

Department of CSE, BAUST, Saidpur[3]

hasan.cse@bsfmstu.ac.bd, mam18@gmail.com, ahasan9@gmail.com

Abstract: Hand key point detection is crucial for facilitating natural human-computer interactions. However, this task is highly challenging due to the intricate variations stemming from complex articulations, diverse viewpoints, self-similar parts, significant self-occlusions, as well as variations in shapes and sizes. To address these challenges, the thesis proposes several innovative contributions. Firstly, it introduces a novel approach employing a multi-camera system to train precise detectors for key points, particularly those susceptible to occlusion, such as the hand joints. This methodology, termed multiview bootstrapping, begins with an initial key point detector generating noisy labels across multiple hand views. Subsequently, these noisy detections undergo triangulation in 3D utilizing Multiview geometry or are identified as outliers. These triangulations, upon re-projection, serve as new labeled training data to refine the detector. This iterative process iterates, yielding additional labeled data with each iteration. The thesis also presents an analytical derivation establishing the minimum number of views necessary to achieve predetermined true and false-positive rates for a given detector. This methodology is further employed to train a hand key point detector tailored for single images. The resultant detector operates in real-time on RGB images and exhibits accuracy on par with methods utilizing depth sensors. Leveraging a single-view detector triangulated over multiple perspectives enables markerless 3D hand motion capture, even amidst complex object interactions.

Keywords: Convolutional Neural Network, Key point detector, Density Network with a Single Gaussian Model, Mixture Density Network, Degree of Freedom.

I. INTRODUCTION

Hand pose estimation stands as a pivotal topic in computer vision, finding myriad applications across human-computer interaction, augmented/virtual reality, and gaming. These applications typically necessitate hand segmentation, articulated hand pose estimation, and tracking. Although recent advancements in body pose estimation[1], [2] can aid in hand detection and segmentation using human body hand joint features, articulated hand pose estimation from monocular RGB images remains a challenging endeavor on several fronts. This complexity arises due to the diverse configurations of human hands, which possess over 20 Degrees of Freedom (DoF). Moreover, hands, being smaller than the body, often occupy only a fraction of the image when the entire body is visible. Additionally, hand keypoints frequently encounter occlusion from other parts of the same hand, the opposing hand, or other body regions.

At present, deep learning techniques demonstrate the highest level of performance in human body pose estimation. This task involves estimating the articulated pose of the body, akin to hand pose estimation. However, body pose estimation generally proves to be less challenging. This is primarily because the body is typically upright, reducing the complexity of the problem.

Additionally, occlusions pose a less frequent and less severe issue in full-body images compared to hand images. Our study delves into deep learning methods designed for hand pose estimation, particularly those employing holistic articulated pose estimation. While pixel-wise pose estimation methods exist, they may be impractical for real-time applications due to their slower processing speed. Moreover, such approaches often fail to leverage crucial holistic hand features due to their focus on individual pixels.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13477

In this study, our focus lies on RGB-based articulated hand pose estimation, a preference rooted in the widespread availability and straightforward deployment of standard color cameras in comparison to depth cameras. Our contribution targets the problem of partial hand pose estimation within individual RGB image frames, with key points of interest including the wrist and fingertips for each digit: thumb, index finger, middle finger, ring finger, and little finger.

We introduce a novel RGB benchmark dataset designed specifically for estimating hand keypoints and conduct evaluations to offer quantitative assessments of current state-of-the-art methods for this task. This dataset encompasses hand gestures alongside keypoint annotations, particularly emphasizing gestures involving rhythmic hand movements. Our motivation stems from the potential utility of tasks involving such movements for cognitive assessments, particularly when integrated with activities involving whole-body motion[3].

There exists a growing need for computational methods aimed at automatically computing various physical performance metrics, thereby enhancing the accuracy and efficiency of human-made assessments. Articulated hand pose recognition assumes a critical role in recognizing and evaluating physical exercises incorporating hand gestures. In Section 4, we delve into the selected hand gestures, elucidating the associated physical exercise tasks and underscoring the significance of articulated hand pose estimation in assessing performance within those tasks. Recognizing rhythmic movements for rapid sequential hand gestures poses additional challenges due to the speed and complexity of the motion. Furthermore, the hand's potential orientation variability and dexterity compound the difficulty in estimating and tracking finger positions.

The paper is further organized as follows: In Section 1, we discuss Introduction; in Section 2, we discuss Literature Review; Section 3 describes Methodology; in Section 4 we describe Implementation; in Section 5, we discuss our experimental Result & Analysis.

II. LITERATURE REVIEW

Early research in hand pose estimation initially focused on utilizing RGB data, as demonstrated by Rehg and Kanade [4] who explored applications in vision-based Human-Computer Interaction (HCI). Many of the early methods were fragile, relying on the fitting of intricate 3D models with strong priors, such as principles from physics or dynamics [5], employing multiple hypotheses [6], or utilizing analysis-by-synthesis techniques [7].

These approaches often relied on visual cues like silhouettes, edges, skin color, and shading, which were tested in controlled environments with limited poses and simple movements. Wang and Popovic's method managed to alleviate some of these limitations but necessitated the use of a specialized colored glove. Similarly, multiview RGB methods often rely on fitting complex mesh models (e.g., [1], [3]), achieving impressive accuracy but typically only under highly controlled conditions.

Following the advent of readily available depth sensors, research emphasis shifted towards single-view depth-based hand pose estimation, leading to a proliferation of depth-based methods. These approaches can broadly be categorized into generative methods [8], discriminative methods [9], or hybrid methods [1], [10], [11]. A recent example of a hybrid method by Sharp et al. [10] has showcased practical performance across a wide spectrum, though challenges persist in scenarios involving interactions between hands or hands and objects. Discriminative and hybrid strategies for depth-based hand pose estimation heavily rely on synthetic data. Oberwerger et al. [12], for instance, employ feedback loops to generate synthetic training data for hand pose estimation, driven by similar principles as our work, albeit focusing on generating depth images. Moreover, the semi-automatic data annotation scheme outlined in [13] shares a similar motivation; however, our approach utilizes multi-view geometry and key-point detection to offer automated supervision.

Discriminative methods, particularly those reliant on deep architectures, necessitate extensive annotated training datasets. While synthesizing such datasets for depth maps is comparatively straightforward, generating them for RGB poses significant challenges due to the complexity of rendering, demanding photorealistic appearance and realistic lighting. Multiview bootstrapping presents an approach that facilitates the generation of large annotated datasets using an initially weak detector.

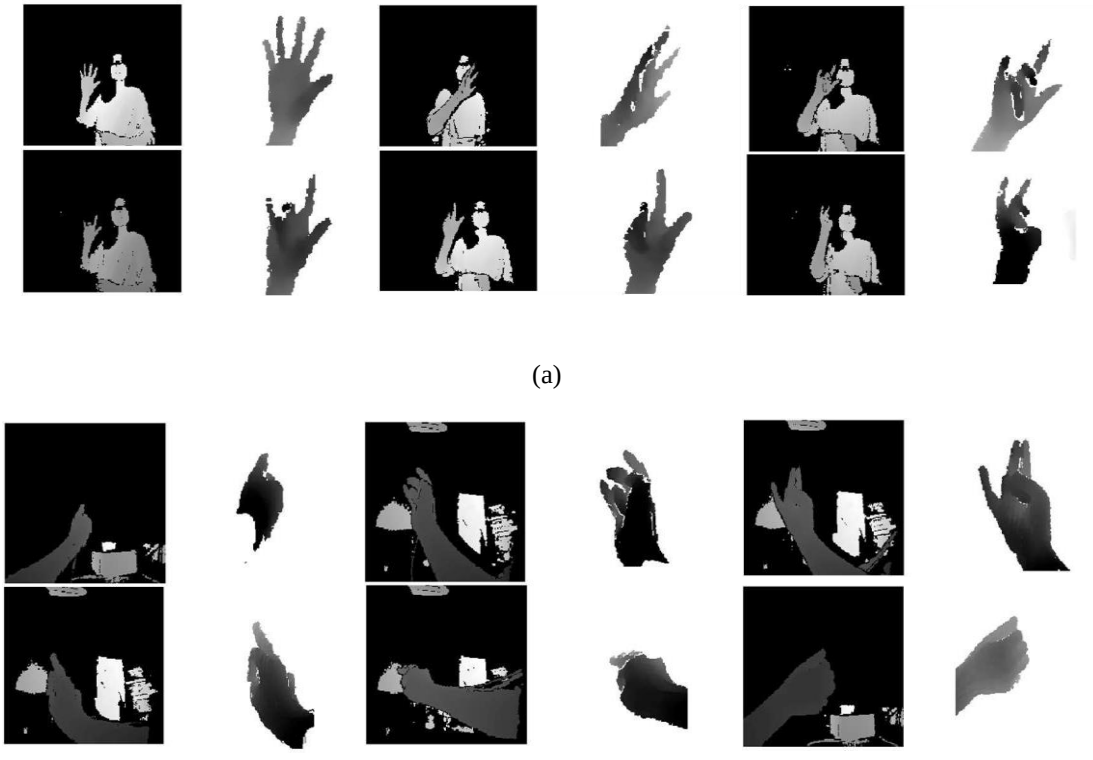
This process, in turn, facilitates the development of the first real-time hand key-point detector for RGB images in real-world settings. When the pose parameters involve joint locations, the pose estimation task can be likened to detecting key points from input images, thereby sharing similarities with other vision problems such as facial landmarking, 6D Object Detection, and human body pose estimation. Of these, hand pose estimation encounters similar challenges to human body estimation, which has witnessed significant advancements in the past decade [14].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13477



(a)

(b)

Fig. 1: Depth images captured by Intel RealSense SR300 [Intel] and cropped hand areas. (a) images captured in third-person viewpoints; (b) images captured in first-person (ego- centric) viewpoints.

Both models aim to represent articulated objects with numerous degrees of freedom and account for self-occlusions. However, hand pose estimation presents distinct challenges due to intricate variations stemming from high Degrees of Freedom (DoF) articulations, diverse viewpoints, self-similar components, significant self-occlusions, and variations in shapes and sizes.

The CPM (Convolutional Pose Machines) is a convolutional neural network designed for human pose estimation using single 2D human pose estimation datasets like MPII, LSP, and Frames Labelled in Cinema (FLIC). This model utilizes CNNs for human pose estimation, with its primary innovation lying in the utilization of a sequential convolution architecture to capture both spatial and texture information.

This architecture comprises multiple stages within the network, each undergoing supervised training to prevent gradient vanishing in deep networks. Initially, the original image serves as input, while subsequent stages use the feature map from the previous stage.

This approach aims to integrate spatial information, texture information, and central constraints. Additionally, employing multiple scales to process the input feature map and response map within the same convolutional architecture ensures both accuracy and consideration of the distance relationship between key points of each human skeleton.

The overall structure of the CPMs is depicted in Figure 2, where "C" and "MC1, MC2" denote different convolution layers, and "P" represents various pooling layers. The "Centre map" denotes the central point of the human body image, used for aggregating response maps to the image centers. The "Loss" function reflects the minimum output cost, consistent with subsequent figures.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13477

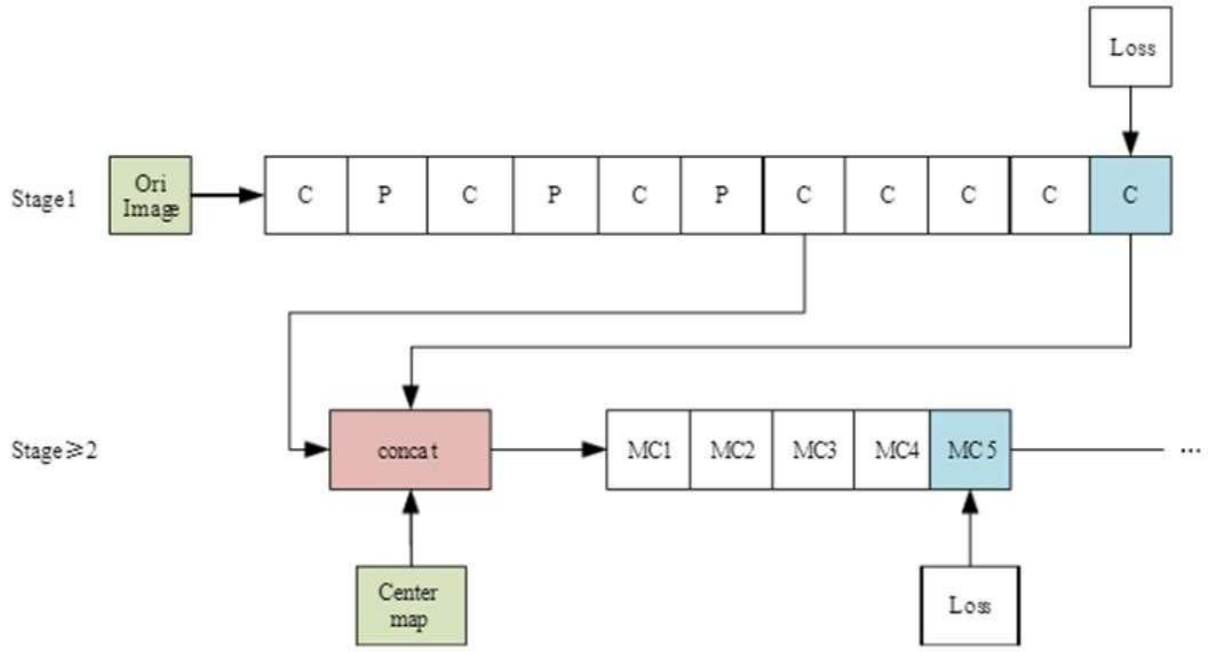


Fig. 2: The overall structure of the convolutional pose machines.

The initial phase of the CPMs comprises a fundamental convolutional neural network (indicated by white convolutions), tasked with directly generating response maps for each key point of the human skeleton from input images. The entire model encompasses response maps for 14 key points of the human skeleton and an additional background response map, resulting in a total of 15 layers of response maps.

The network architecture remains consistent across stages ≥ 2. In subsequent stages, a feature image with a depth of 128, derived from stage 1, serves as input. This input undergoes fusion through a concept layer, integrating three types of data: texture features, spatial features, and center constraints (wherein the center point of the human body image aggregates the response maps to the image centers).

III. METHODOLOGY

This technique essentially involves a learning process conducted through a Multiview system. In our everyday experiences, we observe objects from various angles, each presenting a different shape. Developing computer vision algorithms for camera networks necessitates an understanding of the relationships between images of the same scene captured from different viewpoints. A strategy employed in this regard entails utilizing a multi-camera system to train detailed detectors for keypoints vulnerable to occlusion, such as hand joints. This approach, termed Multiview Bootstrapping, unfolds as follows: initially, an initial keypoint detector generates noisy labels across multiple hand views. Subsequently, these noisy detections undergo either 3D triangulation using Multiview geometry or are flagged as outliers. The resulting reprojected triangulations serve as new labeled training data to refine the detector. This iterative process repeats, yielding additional labeled data with each iteration.

a. Dataset

We present the Hand Keypoint Dataset (HKD), containing annotated RGB images captured while participants engage in rhythmic finger movements. Our dataset comprises 782 color image frames collected from four distinct participants, constituting a novel benchmark dataset for hand Keypoint detection and/or tracking from RGB images. The dataset includes original frames annotated with key points, as well as annotated cropped frames loosely centered around the hand's centroid in the frame. Annotations cover six hand keypoints: W (wrist), TT (tip of the thumb), IT (tip of the index finger), TM (tip of the middle finger), TR (tip of the ring finger), and TL (tip of the little finger). Additionally, the dataset includes annotations for the hand centroid location in the original RGB frames. During data collection, participants executed rapid sequential finger gestures outlined in Section 3, performing these movements thrice with varying hand orientations relative to the camera. The dataset encompasses hand movements from four participants (two male, two female), with annotations manually conducted by two annotators utilizing a standardized annotation toolkit developed by our team.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13477

Additional details of our dataset,

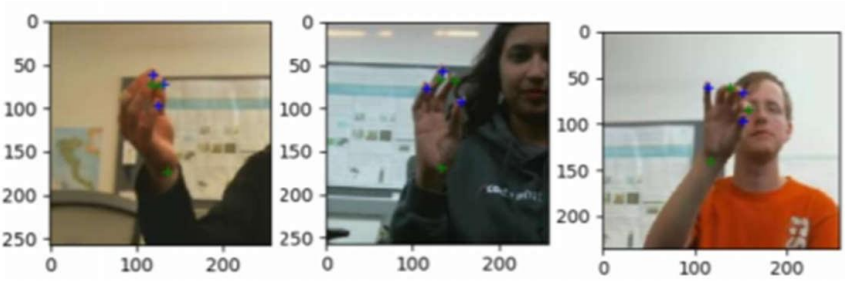


Fig. 3: Example annotations of cropped images from HKD dataset.

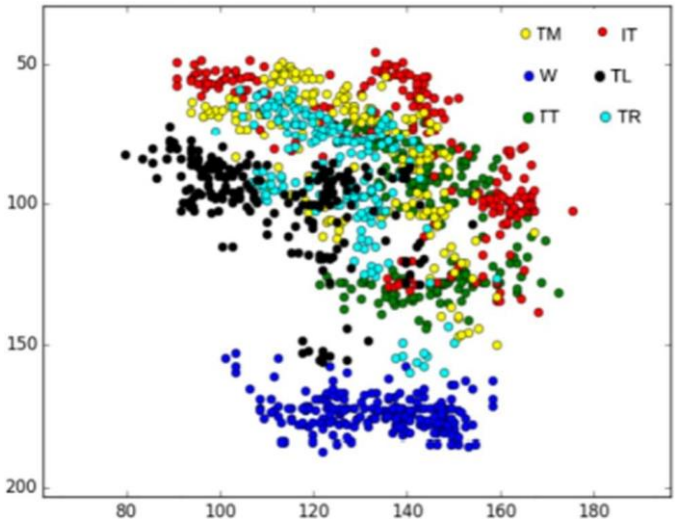


Fig. 4: Variance of hand key points of Subject1 in HKD.

b. Proposed Method

We configured our network module and dataset using OpenCV, along with prototext and caffemodel files. Subsequently, we processed an RGB image containing a hand. The process of detecting keypoints resembles identifying notable points on the hand, akin to human hand joints. We established 22 points to pinpoint hand keypoints and established pairs of keypoints to connect specific points, akin to a human skeleton. Once this method localizes all the keypoints, we depict the skeleton by drawing lines between the pairs and represent the keypoints as circles. Ultimately, we present the output, showcasing the detected keypoints and skeleton.

IV. IMPLEMENTATION

We implement the procedure outlined in the preceding section to ascertain the hand's pose and individuate the five fingers individually. Employing the Multiview Bootstrapping method, we detect all 22 keypoints, proceeding to estimate the hand's pose based on these detected keypoints. Ultimately, we categorize the five fingers independently, marking the culmination of our implementation.

a. Image Preprocessing

Initial image processing is undertaken to ready it for primary processing or subsequent analysis. In this stage, we begin with an image and reduce its dimensions through resizing. Initially, we determine the original height and width of the provided image, calculating the aspect ratio accordingly. Following this, we establish a default height of 368 pixels, while the width is determined based on the aspect ratio and height. The width is computed as the product of the aspect ratio and height. For processing with the Multiview Bootstrapping method, four image parameters are required: batch size, channel, height, and width. Utilizing OpenCV and its "blobFromImage" function, we amalgamate all necessary information required for the network.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13477

b. Localize Keypoints

The result consists of 22 matrices, each representing the Probability Map of a keypoint. To pinpoint the precise keypoints, we initially resize the probability map to match the dimensions of the original image. Subsequently, we identify the keypoint's location by detecting the maximum value within the probability map, achieved through the minmaxLoc function in OpenCV.

We then illustrate the detected points on the image, labeling them with corresponding numbers. Specifically, the wrist is denoted by point 0, while points 4, 8, 12, 16, and 20 signify the tip points of each finger.

c. Estimate Hand pose

We define pairs of key points as like joint on human hand. We define 20 pose pairs to connect the detected keypoints. The pose pairs are like,

[[0, 1], [1, 2], [2, 3], [3, 4],

[0, 5], [5, 6], [6, 7], [7, 8],

[0, 9], [9, 10], [10, 11], [11, 12],

[0, 13], [13, 14], [14, 15], [15, 16],

[0, 17], [17, 18], [18, 19], [19, 20]]

We will use the detected points to get the skeleton formed by the key points and draw them on the image.

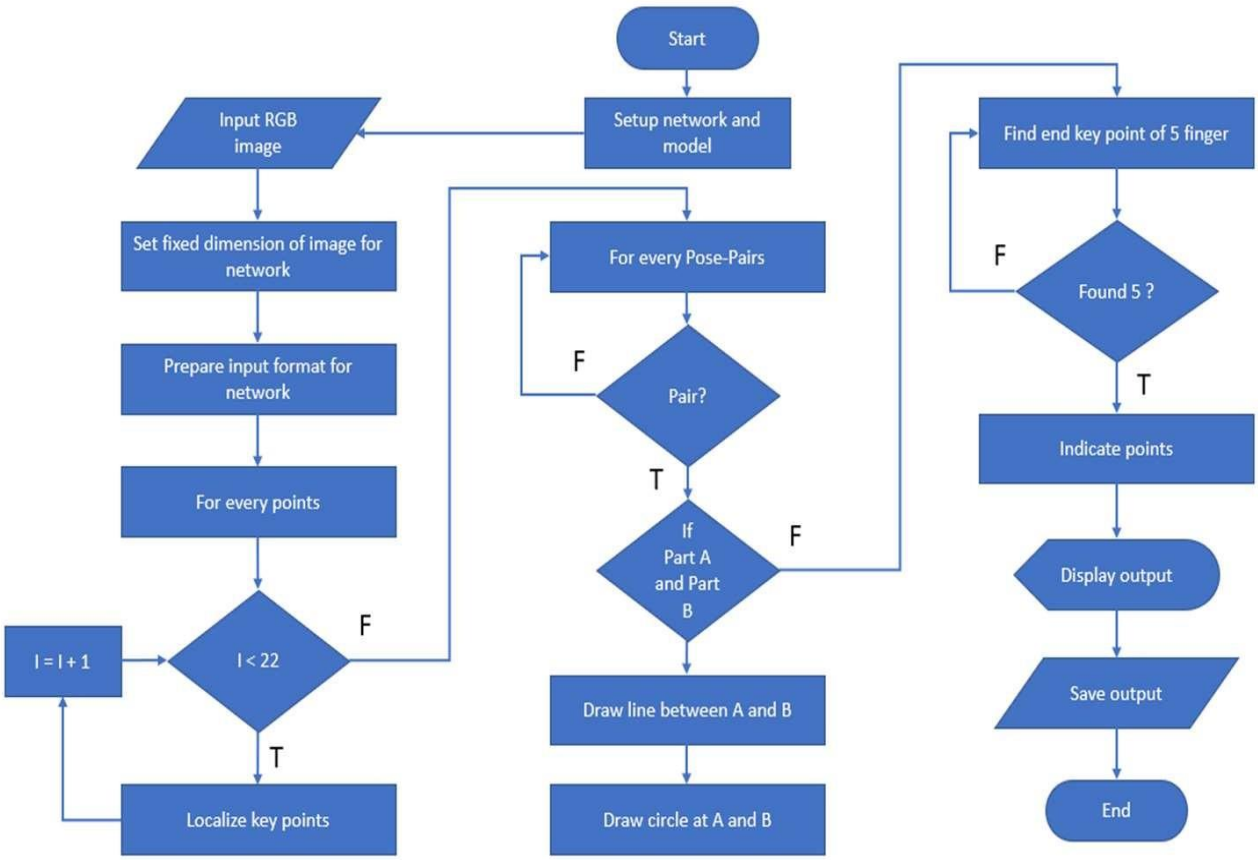


Fig. 5: Flowchart of proposed methodology

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13477

d. Finger Classification

We noticed that all the fingers start from 0 point and end with 4, 8, 12, 16, 20 respectively.When the detector d_i {if i ϵ [0…4]} It is indicating as number 1 finger, When, d_i {if i ϵ [0, 5…8]} It is indicating as number 2 finger, When, d_i {if i ϵ [0, 9…12]} It is indicating as number 3 finger,When d_i {if i ϵ [0, 13…16]} It is indicating as number 4 finger, When, d_i {if i ϵ [0, 17…20]} It is indicating as number 5 finger. So, we need to identify the pixel value of the needed points and indicate with the indexing number.

e. Visualization

The visualization of the image is facilitated by a function within the OpenCV module. Utilizing OpenCV, we display our outcome. Specifically, we employ distinct colors for individual fingers and denote each finger with a straight line originating from the top of the image, accompanied by an index.

V. RESULT AND ANALYSIS

None of the existing hand pose estimation datasets we evaluated were suitable for our intended application: encompassing general, real-world images depicting everyday hand gestures and activities. Consequently, we manually annotated two publicly available image collections: (1) The MPII Human Pose dataset [18], sourced from YouTube videos curated specifically to portray ordinary human activities, and (2) Images from the New Zealand Sign Language (NZSL) Exercises conducted by Victoria University of Wellington [2], featuring individuals using NZSL to narrate stories. We opted for the latter dataset due to its diverse range of hand poses, resembling those encountered in conversational contexts, which are less prevalent in the MPII dataset.

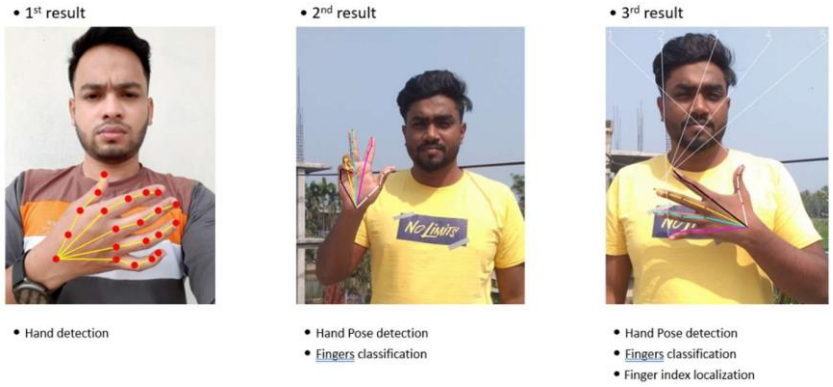


Fig. 6: Expected result after successful implementation.

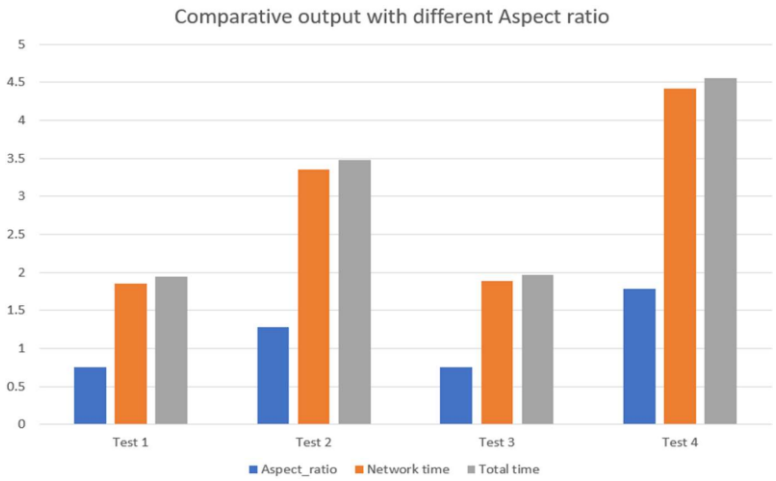


Fig. 7: Comparative output with different Aspect ratio

DOI: 10.17148/IJARCCE.2024.13477

a. Robustness to View Angle

We assess the enhancement in the detector's resilience to varying viewing angles by gauging the proportion of outliers during 3D reconstruction. To establish ground truth, we meticulously scrutinize our most accurate 3D reconstruction outcome, selecting only frames that are correctly reconstructed. We define the view angle in terms of azimuth φ and elevation θ relative to a stationary hand positioned at the origin.

From an intuitive perspective, angles where $\varphi = \{-180, 0, 180\}$ (providing a direct view of the palm or backhand) are deemed easier due to reduced self-occlusion. Conversely, angles at $\varphi = \{-90, 90\}$ offer a side view of the hand, from the thumb to the little finger or vice versa, resulting in increased occlusion. Similarly, angles at $\theta = \{90, -90\}$ present a view from the fingertips to the wrist, and vice versa, representing the most challenging perspectives. We contrast the initial iterations of the "Mix" detector, which swiftly adapt to diverse viewing angles. This comparison is visualized as a heatmap, wherein hand detections are binned based on the azimuth and elevation of each example. The percentage of outliers is determined by considering all examples falling within each bin.

b. Comparison to Depth-based Methods

We evaluate the effectiveness of our method using a publicly available dataset curated by Tzionas et al.. While numerous datasets are commonly employed for assessing depth-based methods, many lack corresponding RGB images or have annotations that are solely applicable to depth images. Datasets containing RGB images with precise manual annotations are scarce; therefore, the dataset provided by [18] aligns best with our method's evaluation requirements. Employing the 2D Keypoint detector "Mix 3" on the RGB images from the dataset, we analyze sequences featuring single-hand motion, hand-hand interaction, and hand-object interaction.

For direct comparison, we utilize the average pixel errors in keypoint locations as outlined in Table 1. It's noteworthy that the referenced method relies on a sophisticated 3D hand template, utilizing depth data and tracking, resulting in several seconds of processing time per frame. In contrast, our approach achieves comparable performance in single-hand and hand-object scenarios using only per-frame RGB detection, capable of real-time operation with GPU acceleration. Performance diminishes in hand interaction scenarios, where our detector may erroneously identify occluding hands. Simultaneous detection of joints on both hands would offer advantages in such cases, rather than treating each hand independently as our current approach does.

c. Summery

We assessed Multiview bootstrapping by implementing Algorithm 1 using three initial detectors. All three detectors adhere to the architecture outlined in Sect. 4, but are trained on three distinct sets of initial training data T0: (1) "Render": a preliminary collection of synthetically generated images of hands, totaling approximately 11,000 examples, (2) "Manual": manual annotations extracted from the MPII and NZSL training sets discussed earlier, and (3) "Mix": a fusion of rendered data and manual annotations. For Multiview bootstrapping, we utilized images from the Panoptic Studio dataset [8]. Specifically, we employed 31 HD camera views and four sequences featuring hand motions, leveraging the provided 3D body pose [8] to estimate occlusions and bounding boxes for hand detection.

During bootstrapping iterations, frames were discarded if they exhibited an average number of inliers < 5 or an average reprojection error > 5, with a detection confidence threshold of λ=0.2. Throughout the process, we manually discarded no more than 15 incorrectly labeled frames. It's important to note that the detector requires a bounding box surrounding the hand to predict the keypoints. Therefore, for optimal results, the hand should be positioned close to the camera or cropped using a hand detector before being inputted to the network. Additionally, the provided code is designed to detect only one hand at a time; however, it can be easily adapted to detect multiple hands by utilizing the probability maps and implementing certain heuristics.

VI. CONCLUSION AND FUTURE WORKS

This paper introduces two advancements: (1) the inaugural real-time hand Keypoint detector demonstrating practical utility in uncontrolled RGB video settings; and (2) the pioneering markerless 3D hand motion capture system, capable of reconstructing intricate hand-object interactions and musical performances autonomously. We ascertain that extensive training sets can be constructed through Multiview bootstrapping, enhancing both the quality and quantity of annotations.

Our approach can be applied to generate annotations for any Keypoint detector susceptible to occlusions (e.g., body and face). The creation of a large annotated dataset often poses a significant bottleneck for numerous machine learning and computer vision tasks, and our method offers a means to refine weakly supervised learning by leveraging Multiview geometry as an external source of supervision. As a prospective avenue of exploration, enhancing the method's robustness

DOI: 10.17148/IJARCCE.2024.13477

to discern between right and left hands, optimizing the algorithm, and implementing it in real-world problem-solving scenarios would facilitate the development of even more comprehensive datasets that closely mirror real-world capture conditions.

REFERENCES

[1] M. A. Fischler and R. C. Bolles, "Random sample consensus," *Commun ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, doi: 10.1145/358669.358692.

[2] S. Pivac Alexander, M. Vale, and R. McKee, "E-learning of New Zealand Sign Language: Evaluating learners' perceptions and practical achievements," *New Zealand Studies in Applied Linguistics*, vol. 23, pp. 60–79, Apr. 2017.

[3] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys, "Motion Capture of Hands in Action Using Discriminative Salient Points," 2012, pp. 640–653. doi: 10.1007/978-3-642-33783-3_46.

[4] J. M. Rehg and T. Kanade, "DigitEyes: vision-based hand tracking for human-computer interaction," in *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, IEEE Comput. Soc. Press, pp. 16–22. doi: 10.1109/MNRAO.1994.346260.

[5] Shan Lu, D. Metaxas, D. Samaras, and J. Oliensis, "Using multiple cues for hand tracking and model refinement," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, IEEE Comput. Soc, pp. II-443–50. doi: 10.1109/CVPR.2003.1211501.

[6] S. Recker, M. Hess-Flores, and K. I. Joy, "Statistical angular error-based triangulation for efficient and accurate multi-view scene reconstruction," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, IEEE, Jan. 2013, pp. 68–75. doi: 10.1109/WACV.2013.6475001.

[7] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-Based 3D Hand Pose Estimation from Monocular Video," *IEEE Trans Pattern Anal Mach Intell*, vol. 33, no. 9, pp. 1793–1805, Sep. 2011, doi: 10.1109/TPAMI.2011.33.

[8] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Tracking the articulated motion of two strongly interacting hands," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2012, pp. 1862–1869. doi: 10.1109/CVPR.2012.6247885.

[9] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Hand Pose Estimation and Hand Shape Classification Using Multi-layered Randomized Decision Forests," 2012, pp. 852–863. doi: 10.1007/978-3-642-33783-3_61.

[10] T. Sharp *et al.*, "Accurate, Robust, and Flexible Real-time Hand Tracking," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2015, pp. 3633–3642. doi: 10.1145/2702123.2702179.

[11] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt, "Fast and robust hand tracking using detection-guided optimization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 3213–3221. doi: 10.1109/CVPR.2015.7298941.

[12] M. Oberweger, P. Wohlhart, and V. Lepetit, "Training a Feedback Loop for Hand Pose Estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015, pp. 3316–3324. doi: 10.1109/ICCV.2015.379.

[13] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit, "Efficiently Creating 3D Training Data for Fine Hand Pose Estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 4957–4965. doi: 10.1109/CVPR.2016.536.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *." [Online]. Available: https://youtu.be/pW6nZXeWlGM

[15] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-Time Joint Tracking of a Hand Manipulating an Object from RGB-D Input," *ArXiv*, vol. abs/1610.04889, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:16440200

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv 1409.1556*, Apr. 2014.

[17] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 4645–4653. doi: 10.1109/CVPR.2017.494.

[18] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2014, pp. 3686–3693. doi: 10.1109/CVPR.2014.471.