

# Deep Learning Based Framework for Hand Keypoints Detection from a Monocular RGB Images



Thesis submitted in partial fulfillment  
for the Award of Degree

*Doctor of Philosophy*

by

*Purnendu Mishra*

*DEPARTMENT OF ELECTRONICS ENGINEERING*  
**INDIAN INSTITUTE OF TECHNOLOGY**  
**(BANARAS HINDU UNIVERSITY)**  
**VARANASI - 221005**

*Roll No. 16091004*

*Year 2022*

## CERTIFICATE

It is certified that the work contained in the thesis titled "*A Deep Learning Based Framework for Hand Keypoints Detection from a Monocular RGB Image*" by *Purnendu Mishra* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all requirements of Comprehensive Examination, Candidacy, and SOTA for the award of Ph.D. Degree.



Supervisor

**Dr. Kishor Sarawadekar**

Assistant Professor,

Department of Electronics Engineering,

Indian Institute of Technology (BHU) Varanasi,

Uttar Pradesh, INDIA 221005.

## DECLARATION BY THE CANDIDATE

I, *Purnendu Mishra*, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of *Dr. Kishor Sarawadekar* from *July 2016* to *Dec 2022* at *Department of Electronics Engineering*, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.* reported in journals, books, magazines, reports, dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my own work.

Date: 22/12/2022

Place: Varanasi

*Purnendu Mishra*  
(Purnendu Mishra)

## CERTIFICATE BY THE SUPERVISOR

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

  
(Dr. Kishor Sarawadekar)

Assistant Professor,  
Dept. of Electronics Engineering,  
Indian Institute of Technology (BHU) Varanasi

  
Signature of Head of Department

विभागाध्यक्ष/Head  
इलेक्ट्रॉनिकी अभियांत्रिकी विभाग  
Department of Electronics Engineering  
भारतीय प्रौद्योगिकी संस्थान  
Indian Institute of Technology  
(बनारस हिन्दू यूनिवर्सिटी)  
(Banaras Hindu University)  
वाराणसी/Varanasi-221005

## COPYRIGHT TRANSFER CERTIFICATE

**Title of the Thesis:** A Deep Learning Based Framework for Hand Keypoints Detection from a Monocular RGB Image


**Name of the Student:** Purnendu Mishra

### Copyright Transfer

The undersigned hereby assigns to the Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy*.

**Date:** 22/12/2022

**Place:** Varanasi

  
(Purnendu Mishra)

**Note:** However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

**Dedicated to**  
**Almighty Mahadev,**  
**Shri Kashi Vishwanath**

# ACKNOWLEDGEMENT

With the completion of the thesis, I wish to express my extreme gratitude to my research supervisor, Dr. Kishor Sarawadekar, for his continuous guidance, encouragement, and critical evaluation throughout this research work. His constant untiring supervision at the vital stages of this work has factored in the smooth completion of the thesis in the most appropriate period of time.

Moving forward, I wish to extend my sincere gratitude to the entire Indian Institute of Technology (BHU), Varanasi, especially the Head of the Department (HoD), Department of Electronics and Engineering, for their encouragement and insightful comments throughout this research work.

This research work would have been intense labor without the constant support of my friends, Dr. Mumtaz Ali Ansari, Shyam Gopal Yadav, Tanushree Meena, and seniors Dr. Gourav Modanwal, Dr. Subiman Chatterjee, who was always ready to guide me throughout my journey.

During this research, my colleagues, Sumit Kumar Yadav, and Bharat Bhushan extended their unconditional support, which contributed to the effortless completion of this work.

Apart from all the people directly contributing to the research, I also want to draw attention to the ones who have been a pillar of strength, even without being directly associated with the investigation. I would like to acknowledge and express my sincere gratitude to the lab Senior Technician, Mr Vinod Kumar Verma, for being constantly around and being just a call away in case of any discrepancy.

This research work has been an incredible journey where mentors like Dr. Kishor Sarawadekar have installed his belief and have motivated me through the various research paper submissions. He relentlessly devoted himself to my betterment and the entire department.

Apart from the academic guidance, he has been a GUIDE in a true sense who dilated every situation and has always advised the most appropriate solutions to every

obstacle.

I also wish to acknowledge the incredible guidance and support of the entire institute, especially during the pandemic years. I, in particular, am grateful for all the opportunities that came my way.

Along with all the academic guidance, I wish to express my sincere regards to my parents, Nirbhaya Kumar Mishra, Ranjana Mishra, Bijay Kumar Mishra and Anju Mishra. My siblings Navendu Mishra and Kalindi Mishra, along with my Brother-in-Laws Dhananjay Jha and Alok Mishra. They all have been instrumental in keeping my mental strength intact and always ensure positivity and zeal throughout this journey.

I want to give special thanks to my wife, Minakshi Mishra, who has stood by me through all my travails, my absences, my fits of pique and impatience. She gave her support and help, discussed ideas, and prevented several wrong turns. She also supported the family during much of my research phase.

Last but not least, I would like to thank God Almighty for bestowing his power and positivity which was key to this entire research process.

**(Purnendu Mishra)**

# Contents

List of Figures	xii
List of Tables	xv
List of Symbols	xvii
List of Abbreviations	xix
Preface	xxi
<b>1 Introduction</b>	<b>1</b>
1.1 Hand Pose Estimation . . . . .	5
1.1.1 Importance of Hand Pose Estimation . . . . .	5
1.1.2 Types and Approaches . . . . .	6
1.1.3 Challenges . . . . .	8
1.1.4 Datasets . . . . .	11
1.2 Artificial Intelligence and Deep Learning . . . . .	11
1.2.1 Deep Learning . . . . .	13
1.2.2 Convolution Neural Network . . . . .	16
1.2.3 Applications of CNN . . . . .	18
1.2.4 Advantages and Disadvantages . . . . .	20
1.2.5 Hyper-parameter Tuning . . . . .	20
1.3 Motivation . . . . .	28
1.4 Contribution of the Dissertation . . . . .	28
1.5 Organization of the Dissertation . . . . .	29
<b>2 Partial Hand Keypoints Detection</b>	<b>31</b>
2.1 Overview . . . . .	31
2.2 Related Works . . . . .	32

2.3	Fingertips Detection . . . . .	37
2.4	Partial Keypoints Detection with Multi-label Classification . . . . .	38
2.4.1	Fingertips detection with multi-label classification . . . . .	41
2.4.2	The CNN architecture . . . . .	42
2.4.3	Model optimization . . . . .	43
2.4.4	Experimental Analysis . . . . .	44
2.4.5	Discussion . . . . .	47
2.5	Anchors-based Fingertips Detection . . . . .	48
2.5.1	The CNN architecture . . . . .	51
2.5.2	Inference . . . . .	51
2.5.3	Model optimization . . . . .	53
2.5.4	Experimental Analysis . . . . .	53
2.5.5	Discussion . . . . .	56
2.6	Nearest Neighbor Fingertips Detection . . . . .	57
2.6.1	Working principle . . . . .	59
2.6.2	The CNN architecture . . . . .	62
2.6.3	Decoding . . . . .	64
2.6.4	Model optimization . . . . .	65
2.6.5	Experimental Analysis . . . . .	65
2.6.6	Discussion . . . . .	73
<b>3</b>	<b>Full Hand Keypoints Detection</b>	<b>77</b>
3.1	Overview . . . . .	77
3.2	Related Works . . . . .	79
3.3	Single Hand Keypoints Detection . . . . .	81
3.3.1	Working principle . . . . .	82
3.3.2	Hand localization . . . . .	84
3.3.3	Feedback Inference . . . . .	85
3.3.4	Network architecture . . . . .	85
3.3.5	Decoder . . . . .	87
3.3.6	Model optimization . . . . .	87
3.3.7	Experimental Analysis . . . . .	88
3.3.8	Discussion . . . . .	96
3.4	Double Hand Keypoints Detection . . . . .	97
3.4.1	Working principle . . . . .	97
3.4.2	Hand ROIs Detection . . . . .	100

---

3.4.3	Network architecture . . . . .	103
3.4.4	Model optimization . . . . .	104
3.4.5	Experimental Analysis . . . . .	105
3.4.6	Comparison . . . . .	105
3.4.7	Discussion . . . . .	110
<b>4</b>	<b>Multiple Hands Keypoints Detection</b>	<b>113</b>
4.1	Overview . . . . .	113
4.2	Related Works . . . . .	115
4.3	Methodology . . . . .	116
4.3.1	Training Pipeline . . . . .	119
4.3.2	Inference Pipeline . . . . .	123
4.4	Results . . . . .	123
<b>5</b>	<b>Conclusion and Future Scope</b>	<b>127</b>
	<b>References</b>	<b>133</b>
	<b>List of Publications</b>	<b>153</b>



# List of Figures

1.1	The images left to right illustrate the development in the interface through which we interact with the telephone device. . . . .	2
1.2	Illustration of the evolution of technology that enables the way we take pictures using our smartphone. . . . .	3
1.3	Some examples for hand pose applications area. . . . .	7
1.4	A human hand and its 21 keypoint (or joints). . . . .	9
1.5	Sub-fields of Artificial Intelligence [1]. . . . .	12
1.6	Making an analogy between human and machine ways of gaining knowledge and performing intended tasks. . . . .	16
1.7	The basic working principle of an artificial neuron. . . . .	17
1.8	A simple neural network architecture. . . . .	17
1.9	Representation of digital image in two different forms. . . . .	18
1.10	A classic CNN architecture called LeNet [2] . . . . .	19
1.11	A few examples for CNN application areas. . . . .	20
1.12	A representation of loss function in three dimensions. . . . .	22
1.13	An example of a saddle point. . . . .	23
1.14	The effect of the value of the power on the form of the polynomial learning rate scheduling curve. . . . .	24
1.15	Different warm restart strategies. . . . .	25
1.16	Classification accuracy for different learning rate schedules on CIFAR-10 dataset. . . . .	26
1.17	Classification accuracy for different learning rate schedules on CIFAR-100 dataset. . . . .	26
2.1	A type of data glove used for capturing hand joint motion. . . . .	33
2.2	An example of a keypoint represent using the Gaussian heatmap. . . . .	36
2.3	Encoding multiple keypoints in different channels with Gaussian heatmap (Source [3]). . . . .	37

2.4	The positions of important hand keypoints . . . . .	38
2.5	Flow diagram illustrating the process of fingertips detection. . . . .	39
2.6	Issue of ambiguous points being present with direct regression-based fingertips detection . . . . .	40
2.7	Example of multi-label classification for finger identification. . . . .	41
2.8	The DNN model architecture used for fingertip(s) position estimation. .	42
2.9	The hand localization results on the samples from the SCUT-Ego-Gesture dataset. . . . .	44
2.10	The fingertips detection results for different hand gestures on the samples from the SCUT-Ego-Gesture dataset. . . . .	46
2.11	The distribution of fingertips for different gestures along the hand bounding box. . . . .	48
2.12	Local regression using prior known points called anchors for estimation of fingertips position. . . . .	49
2.13	The framework used for fingertips detection in the anchor-based detection method. . . . .	50
2.14	The model architecture used for fingertips detection with the anchor-based method. . . . .	52
2.15	The effect of accuracy on fingertips estimation when the number of markers is varied. . . . .	54
2.16	Performance comparison of anchor-based fingertips detection method in terms of PCK. . . . .	55
2.17	Sample fingertips detection results on the SCUT-Ego-Gesture dataset with the anchor-based method. . . . .	57
2.18	Sample fingertips detection results on HGR dataset with the anchor-based method. . . . .	58
2.19	Demonstrating the effectiveness of rotation invariance of anchor-based fingertips' detection algorithm. . . . .	58
2.20	The framework for nearest-neighbor fingertips detection process. . . . .	59
2.21	The grids formed by the pose particles and the 4-nearest pose particles to all visible fingertips. . . . .	60
2.22	The 4-nearest pose particle to the thumb's tip and the position vectors from the nearest pose particles . . . . .	60
2.23	The CNN model architecture used in nearest-neighbor based fingertips detection process. . . . .	63

2.24	Performance comparison at different numbers of nearest-neighbor pose particles . . . . .	68
2.25	Performance comparison for different grid sizes formed by the pose particles	69
2.26	Comparative analysis in terms of PCK on two different datasets. . . . .	70
2.27	Comparative analysis in terms of PCK on RHD dataset . . . . .	70
2.28	Qualitative comparison of fingertips' detection on OneHand10K dataset samples with various methods . . . . .	71
2.29	Performance comparison for single and double hand. . . . .	73
2.30	Sample results on double hand fingertips detection. . . . .	74
3.1	Framework for single hand keypoints detection. . . . .	82
3.2	Illustration of the process of single hand keypoints detection. . . . .	83
3.3	Illustration of the process of single hand bounding box detection. . . . .	84
3.4	The CNN architecture used for single hand keypoints detection. . . . .	86
3.5	The CNN architecture used for single hand keypoints detection. . . . .	86
3.6	The figure demonstrates heatmaps obtained to detect various hand features. The top left image shows the hand ROI and the model's ROI detection probability is shown using the heatmap in the top right. The bottom two images show the two hand keypoints detection probability in terms of the heatmap. . . . .	89
3.7	The normalized histograms obtained from probability map generated by the CNN model. . . . .	90
3.8	Example showing improvement in keypoints detection accuracy with feedback inference . . . . .	91
3.9	The PCK curve for different grid sizes. . . . .	92
3.10	The effect of grid size of hand keypoints detection process. . . . .	93
3.11	Sample results with grid-based hand keypoints detection algorithm on three different datasets. . . . .	94
3.12	Comparison of proposed grid-based keypoints detection method of various datasets . . . . .	95
3.13	An illustration of the method used for the estimation of 2D hand keypoints position from an RGB image . . . . .	99
3.14	In order to localize the hand in a color image, the input image is divided into $M \times M$ grid. Then grid cells covering the hand region in the image are identified. Separate channels are used for the detection of each hand. Here, (a) the original image and white cells in the grid represent the hand region for (b) the right hand, and (c) the left hand. . . . .	100

3.15	The segregated view of the output array of a prediction layer. Separate channels are used for hand localization and keypoints detection of left and right hands. The first two channels are used for Hand ROIs detection and the rest of the others are used for hand keypoints detection. . . . .	101
3.16	The proposed CNN model for simultaneous hand localization and keypoints detection from an RGB image. The details of the network's configuration are provided in Section 3.4.3. . . . .	104
3.17	Quantitative analysis in terms of PCK values at different values of threshold distance. Separate analysis has been conducted for single and double-hand cases. (a) Single hand, and (b) double hands image samples from the RHD dataset. (c) Single hand, and (d) double hands samples for InterHand2.6M dataset. . . . .	107
3.18	Sample results of hand keypoints detection from different datasets. For a single hand, the skeleton of each finger is shown with different colors. In double hand, the right-hand skeleton is shown in red color while the left-hand skeleton is shown in cyan color. The fingertips and finger joints are marked with white dots on the skeleton. . . . .	109
4.1	Increasing the number of output channels in the prediction layer to perform double hand keypoints detection. . . . .	114
4.2	The architecture used for hand ROIs detection. Source [4] . . . . .	117
4.3	The flow diagram for training pipeline. . . . .	120
4.4	The flow diagram for inference pipeline. . . . .	122
4.5	The sample results of hand ROI and keypoints detection . . . . .	124

# List of Tables

1.1	A list of publicly available datasets for hand keypoints detection	12
1.2	Advantages and disadvantages of a CNN model . . . . .	21
1.3	Test accuracy of different learning rate policies on CIFAR-10, CIFAR-100, and tiny ImageNet datasets with moderate data augmentation . . . . .	27
2.1	Performance on hand detection . . . . .	45
2.2	Comparative results for various performance metrics for fingertips detection with the multi-label classification process. . . . .	46
2.3	Performance comparison of the anchor-based fingertips detection method . . . . .	56
2.4	Effect of different parameters viz. backbone model, N-nearest neighbor, and grid size on the performance of the proposed model tested on three different datasets. . . . .	67
2.5	Comparison Results of the proposed method for fingertips detection with different methods on the various datasets . . . . .	72
3.1	Performance comparison of the grid-based hand keypoints algorithm with different methods. The PCK value is calculated for the threshold value of $\sigma = 0.2$ . . . . .	96
3.2	The performance of the proposed model in terms of PCK value at $\sigma = 0.2$ . . . . .	106
3.3	Hand Keypoints detection performance comparison on several existing datasets in terms of PCK . . . . .	108
4.1	Performance comparison of the proposed multi-hand keypoints detection with three different keypoints detection algorithms and other deep learning-based methods . . . . .	125



# List of Symbols

<b>Symbol</b>	<b>Description</b>
$\sigma$	Distance threshold in PCK calculation
$\chi$	Chi function
$\mathbb{N}$	Number of samples
$\mathbb{R}$	Real number
$w$	Width of the bounding box
$h$	Height of the bounding box
$S$	RGB image dimension
$C$	Probability of a grid cell
$M$	Number of grids
$U$	Numbers of grid rows
$V$	Number of grid columns
$I$	RGB image
$\delta_p$	Grid cell probability threshold
$\delta_m$	Multi-label classification probability threshold



# Abbreviations

<b>Abbreviation</b>	<b>Description</b>
ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
AR	Augmented Reality
ASL	American Sign Language
ANN	Artificial Neural Network
CAGR	Compound Annual Growth Rate
CDF	Cumulative Distribution Function
CIFAR	Canadian Institute of Advanced Research
CLR	Cyclic Learning Rate
CMU	Carnegie Mellon University
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
DOF	Degree of Freedom
F1Score	Harmonics Precision-Recall Mean
FCN	Fully Connected Network
FFNN	Feed-Forward Neural Network
FN	False Negative
FP	False Positive
FPHA	First Person Hand Action
FPN	Feature Pyramid Network
GAP	Global Average Pooling
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HCI	Human Computer Interaction
HGR	Hand Gesture Recognition
IOU	Intersection Over Union

<b>Abbreviation</b>	<b>Description</b>
LR	Learning Rate
MHP	Multi-view Hand Pose
ML	Machine Learning
MLP	Multi-Layered Perceptron
MR	Mixed Reality
NN	Neural Network
NUI	Natural User Interface
PC	Personal Computer
PCK	Probability of Correct Keypoint
POV	Point of View
POLY	Polynomial Learning Rate
ReLU	Rectified Linear Unit
RHD	Rendered Hand Pose
RNN	Residual Neural Network
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SGD	Stochastic Gradient Descent
SGDR	Stochastic Gradient Descent with warm restart
SLR	Sign Language Recognition
SSD	Single Shot Detector
STB	Stereo Hand Pose Tracking Benchmark
TN	True Negative
ToF	Time of Flight
TP	True Positive
VGG	Visual Geometry Group
VR	Virtual Reality
WHO	World Health Organization
YOLO	You Only Look Once
YOLSE	You Only Look what You Should See