Machine Learning: Science and Technology

PAPER • **OPEN ACCESS**

Effectual pre-processing with quantization error elimination in pose detector with the aid of image-guided progressive graph convolution network (IGP-GCN) for multi-person pose estimation

To cite this article: Jhansi Rani Challapalli and Nagaraju Devarakonda 2023 *Mach. Learn.: Sci. Technol.* **4** 025015

View the article online for updates and enhancements.

You may also like

- Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt
M Jain, Balwinder Singh, A A K Srivastava et al.
- Light absorbing organic aerosols (brown carbon) over the tropical Indian Ocean: impact of biomass burning emissions
Bikkina Srinivas and M M Sarin
- Irradiation-induced hardening in fusion relevant tungsten grades with different initial microstructures
Chih-Cheng Chang, Dmitry Terentyev, Aleksandr Zinovev et al.

This content was downloaded from IP address 122.163.64.247 on 11/07/2023 at 12:34

IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 025015 https://doi.org/10.1088/2632-2153/acc9fc

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
17 November 2022

REVISED
15 February 2023

ACCEPTED FOR PUBLICATION
3 April 2023

PUBLISHED
26 April 2023

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Effectual pre-processing with quantization error elimination in pose detector with the aid of image-guided progressive graph convolution network (IGP-GCN) for multi-person pose estimation

Jhansi Rani Challapalli and Nagaraju Devarakonda*

School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India
* Author to whom any correspondence should be addressed.

E-mail: dnagaraj_dnr@yahoo.co.in

Keywords: pose estimation, pre-processing, error lessening, neural network, occlusion, heat maps, graph

Abstract

Multi-person pose estimation (MPE) remains a significant and intricate issue in computer vision. This is considered the human skeleton joint identification issue and resolved by the joint heat map regression network lately. Learning robust and discriminative feature maps is essential for attaining precise pose estimation. Even though the present methodologies established vital progression via feature map's interlayer fusion and intralevel fusion, some studies show consideration for the combination of these two methodologies. This study focuses upon three phases of pre-processing stages like occlusion elimination, suppression strategy, and heat map methodology to lessen noise within the database. Subsequent to pre-processing errors will be eliminated by employing the quantization phase by embracing the pose detector. Lastly, Image-Guided Progressive Graph Convolution Network (IGP-GCN) has been built for MPE. This IGP-GCN consistently learns rich fundamental spatial information by merging features inside the layers. In order to enhance high-level semantic information and reuse low-level spatial information for correct keypoint representation, this also provides hierarchical connections across feature maps of the same resolution for interlayer fusion. Furthermore, a missing connection between the output high level information and low-level information was noticed. For resolving the issue, the effectual shuffled attention mechanism has been proffered. This shuffle intends to support the cross-channel data interchange between pyramid feature maps, whereas attention creates a trade-off between the high level and low-level representations of output features. This proffered methodology can be called Occlusion Removed_Image Guided Progressive Graph Convolution Network (OccRem_IGP-GCN), and, thus, this can be correlated with the other advanced methodologies. The experimental outcomes exhibit that the OccRem_IGP-GCN methodology attains 98% of accuracy, 93% of sensitivity, 92% of specificity, 88% of f1-score, 42% of relative absolute error, and 30% of mean absolute error.

1. Introduction

Human body pose estimation (HBPE) is a process to estimate the human body joint points. The task of HPE to localization of human joints also called as keypoints. It is the study where track the object by identifying the joint points position data analysing the link between the joint points and, later, rebuilding human limbs' methodology to form a skeleton structure. In the last decade, it has been increasingly popular and has been used in a variety of fields, including motion analysis, augmented reality, and virtual reality. Despite the strong performance of recently established deep learning-based systems in estimating human posture, problems still exist due to a lack of training data, crowded background, invisible key points, depth ambiguities, and body occlusion [1].

© 2023 The Author(s). Published by IOP Publishing Ltd

To address these problems [2] proposed a novel pose-based action recognition method was implemented that detect the human poses from videos. In this technique initially obtain K-best pose estimations (PEs) for each frame then select the best poses by applying segmentation and temporal constraints for all the video frames [3] proposed deep expressive model for higher level action in still images to combine data from many noisy sources, such as body part detection and object detection, a deep belief net is developed. Formerly, HPE depended upon hand-labelled features. The PE can be described as a tree structure or graphical paradigm that fails to efficiently address the spatial framework association between the keypoints. The action estimation identification's strength remains bad. Due to the progression of convolutional neural network (CNN) in the HPE discipline, the keypoints identification's execution is highly enhanced.

The pose remains generally conveyed by three angles (pitch, yaw, and roll), which define the head's egocentric orientation. The estimation turns difficult while many persons remain in an image, thereby their faces contain a little support region generally below 100×100 pixels [4]. Although the face position within the image remains notable, one should excerpt the pose angles out of low-definition data. The local features identification, for instance, facial landmarks, remains troublesome in such instances and one could solely employ global visual data, for instance, histogram of oriented gradients (HOG) [5]. A global face describer like this will be employed as input in this study for estimating the three-dimensional pose. Hence, it is necessary to resolve high density to low density resolution mapping issue. This remains renowned that the high-to-low regression will be more difficult since a huge quantity of criteria is required to be estimated. It will be frequently resolved by employing kernel methodologies like Gaussian procedure regression. Nevertheless, it indicates an ad-hoc selection of a kernel function and also the hyper criteria estimation that results in a non-linear/non-convex optimization issue. Lately, a high-to-low regression has been proffered, which learns a low-to-high regression out of what high-to-low expectancy will be next derived centred upon Bayes inversion called the Gaussian mixture of locally-liner mapping paradigm [6]. This technique's benefit above other prevailing linear regression mixture approaches remains that this prevents the estimation of the criteria's enormous quantity related to high-to-low learning.

In traditional methods of HPE, local information is extracted using keypoint detectors to create visual structures [7]. Contextual information is typically required to offer visual representations that can be derived from a broad region around the part [8] or by interaction among detected parts [9] in order to manage challenging circumstances of occlusion or partial vision. PE can generally be viewed from one of two perspectives either as a correlated part detection task or as a regression problem. Finding key points independently is a common goal of detection-based techniques, which are then combined to generate a single pose prediction during post-processing phases. But there exists ambiguity while combining joints it can be eliminated by exploiting the dependency between the joints by various multi stage architectures like [10, 11]. The most effective usage of these techniques is 2D PE. However, because the 3D heat maps require so much memory and processing power, they do not readily generalise to 3D posture estimation. Regression-based techniques, on the other hand, use a function to directly map input images to the positions of body joints. They are generic for both 3D and 2D posture estimation and directly target the problem.

Earlier CNNs are successfully implemented to two-dimensional body PE; specifically, fully convolutional networks [12] could execute pixel-wise joint identification more precisely. It will be compiled as every pixel's pixel-wise classification remaining a joint's position. Hence, PE attempts in generating analogous networks for identifying joints in two-dimension. Joint identification could leverage local patterns very directly than the high resolution via pixel wise classification assisting the network in learning finer feature maps.

The 2D detections and three-dimensional regressions could be later combined with a multi-task setup, either by supplying the 2D detections heat map as an input to a three-dimensional regressor network or by exchanging the heatmaps between identification and high resolution. Nevertheless, there remains no assurance that the regressed three-dimensional joints, when these have to be cast back to two-dimensional, would be in accord with the initial 2D detection heatmap. Additionally, by design, the above said HR's disadvantages will be yet unremoved with this task line.

The rest of the studies in 2DD imply reverse kinematics and employ a paradigm-centred augmentation. Nevertheless, the critical multi-person's self-occlusion generates uncertainty that remain difficult in solving and struggling out of accuracy issues that will be alternatively absent in body PE. This study's inputs include the ensuing,

- The three phases of Pre-processing stages are focused like occlusion elimination (OE), suppression strategy, and heatmap methodology to lessen noise within the database.
- Subsequent to PP, errors will be eliminated by employing the quantization step (QS) by embracing a pose detector (PD).

- Image-Guided Progressive Graph Convolution Network (IGP-GCN) will be built for multi-person pose estimation (MPE) that consistently learns affluent fundamental spatial information by merging features inside the layers.

The rest of this study is organized as follows: Segment 2 highlights a few existing works, Segment 3 discusses the proffered approach and methodologies, Segment 4 exhibits the experimental results and discussion, and, finally, Segment 5 finishes with the conclusion and prospective studies.

2. Related works

Lately, deep learning methodologies evolved as very strong approaches to automatedly learning features out of unprocessed data. Particularly, deep learning methodologies attained appreciable progression in object identification, an issue that attracted the focus of several analyses in the present decade. Video surveillance remains one of the very difficult and basic regions in the security system since this relies completely upon numerous object identification and tracking. This observes humans' behaviour in public for identifying whatever suspecting behaviour.

2.1. Survey on heatmap and key point excerption

The study [13] introduced a novel method for MPE to overcome the scale variant. This work mainly concentrates on scale variation of keypoints within heatmap generation called scale aware heatmap generator. It generates heatmap for each keypoints based on scales variant with modified loss function by weight redistribution that are used to identify the invisible keypoints. This model outperforms 69.5% AP on the COCO dataset.

The study [14] proffers a bottom-up technique for posing analysis and movement detection. The authors propose a Strong Pose system, which handles association among object-part by employing part-based modelling. The convolution network in this paradigm identifies powerful keypoint heatmaps and estimates their correlative displacements permitting keypoints to be grouped into human instances. Additionally, this employs the keypoints for creating body heatmaps, which could decide the human body's location within the image. This model was trained on COCO dataset with Resnet-101 and Resnet-152 architectures which outperforms average precision of 0.70 and 0.725.

The study [15] models a lightweight bottleneck block having a re-parameterized framework. This creates and enhances the feature maps diverseness. Next, the authors present a multi-branch framework and a single-branch framework within the bottleneck block. In the training stage, a multi-branch framework will be comprised for enhancing the estimated precision. In the deploying stage, single-branch framework will be employed for enhancing the paradigm reference speed. Which almost reduced the computational cost. This model outperforms 74.1% on COCO dataset and the network architecture is same as HRNet with resolution 128×128.

The study [16] proffers a network named GroupPoseNet (GPN) employing a categorizing scheme for dealing with this issue. GPN excerpts the left-hand and right-hand features accordingly and, hence, prevent the collaborative attachment betwixt the communicating hands. Authorized by a new up-sampling block named multi branch framework Block, this anticipates two-dimensional heatmaps in an advancing manner by merging image, hand pose, and multi-scale features. GPN remains efficient and strong to crucial occlusions. For attaining an effectual three-dimensional hand rebuilding, the authors model a transformer operation-related reverse kinematics unit (called TikNet) for mapping three-dimensional joint positions to the MANO hand paradigm's hand shape and pose criteria.

The study [17] suggests a multi-person PE algorithm aims upon the double anchor embedding (DAE) that exhibits that bottom-up algorithms will remain yet challenging in accuracy. Initially, to lessen the identification job's designing complexity, the authors split the human joints into top and bottom half categories that remain inwardly consistent and greatly compared. Subsequently, a new joint affinity cue, named DAE, will be modelled that could assist the network efficiently in excerpting the data of local as well as global contexts thereby could better handle occluded scenes and intricate postures.

The study [18] introduces a multi-hop attention graph (MAGC) convolution network for excerpting strong person joint feature (JF) information by residual attention technique while reducing the effect of environmental noise. The transfer of higher order graph features' inside MAGC facilitates the network for learning the hidden association betwixt features. The authors as well present the self-attention semantic perception layer that could adaptatively choose additional discriminant features for additionally reinforcing the transfer of beneficial data.

The study [19] suggests a solution for resolving issues with 3D human PE by taking depth information into consideration. In order to do this, a cross-modality CNN training strategy was used, along with the concept of a batch normalization layer within the RGB-pretrained 2D CNN model to reduce the distribution divergence between the RGB and depth data during training. The normal vector map is combined with the raw depth data in order to incorporate additional 3D descriptive information. Even yet, performance can be improved by using local refinement with coarse-to-fine human posture estimation. While the best method for determining the local observation scale is not fully discussed. In line with this, a multi-scale local refinement network is suggested, with the tiny local region concentrating on capturing the fine information. On the other hand, the vast local region has more comprehensive semantic contextual data.

2.2. Survey upon CN for PE

The study [20] proffers a methodology to address the issues such as keypoint representation quantization error (QE). In this methodology, the observed keypoint coordination representation distribution probability will be extracted by a CNN, and the cross-entropy will be created with the estimated probability distribution as loss function. To minimize the Kullback–Leibler distance between the estimation and the ground truth, the CNN will be augmented, and the HM's coordinates will be finally positioned.

The study [21] highlights employing a densely connected convolutional module (DCCM) as the NN's fundamental unit. For every DCCM layer, feature maps, which are entirely generated by the former layers, will be connected as the input, and the output feature maps will be provided to every layer. The experimental outcomes upon the MPII human pose dataset and LSP dataset exhibit that this methodology could obtain corresponding execution when this needs fewer criteria so that greater criteria efficacy could be attained.

The study [22] recommends visual control system comprising a visual perception module (VPM) and a robot manipulator administrator. The VPM merges deep CNNs (DCNNs) and a totally linked conditional random field layer for identifying an image semantic segmentation function that could give steady and precise object classification outcomes in a disordered environment. The object PE unit applies a paradigm-related PE methodology for analysing the 3D pose of the target for picking control. Furthermore, the proffered data augmentation model automatedly creates new training data for training the DCNNs.

The study [23] suggests employing global relation reasoning (GRR) graph CNNs (GRR-GCNN) for effectually catching the global associations amidst disparate body joints. GRR-GCN projects all the features in the original coordinate space to a graph space. Within the graph space, such features will be depicted by an array of nodes for creating a fully connected (FC) graph whereupon GRR will be executed by graph convolution. Node features will be projected back to the Coordinate space after GRR to undergo further processing.

Even though HPE remains the most complex challenging task and vital execution enhancement were accomplished in the last some years, a few listed precisions in the techniques have been achieved via multiple post-processing phases or a few schemes employed in the dataset competency. For an instance, executing multi-scale feature analysis, enhancing outcomes by a different methodology, or the assessment of accuracy at one image scale when speed will be registered at one more scale. Such post processing phases intrude on the decision in detecting an algorithm's robustness and efficiency Hence, assessing a methodology devoid of whatsoever post processing phases and schemes remains additionally objective and extra invaluable for the study and practical implementation.

2.3. Background of error identification prior to PE

MPE remains very difficult than a single person since the position and persons within an image remain unfamiliar. Generally, the problem could be resolved by employing one of the two techniques.

- The simple technique remains integrating a person detector initially, ensued by estimating the portions and, later, computing the pose of every man. This methodology is called the top–down technique.
- One more technique remains in identifying entire portions in the image ensued by relating portions appertaining to unique people. This methodology is called the bottom-up technique.
- Generally, the top–down technique remains simpler for applying than the bottom-up techniques since pose detection remains very simplest than including relating algorithms. This remains difficult for deciding whichever technique possesses comprehensive finer execution as this actually since in lands amidst the pose detection and relating algorithms that remain finer.

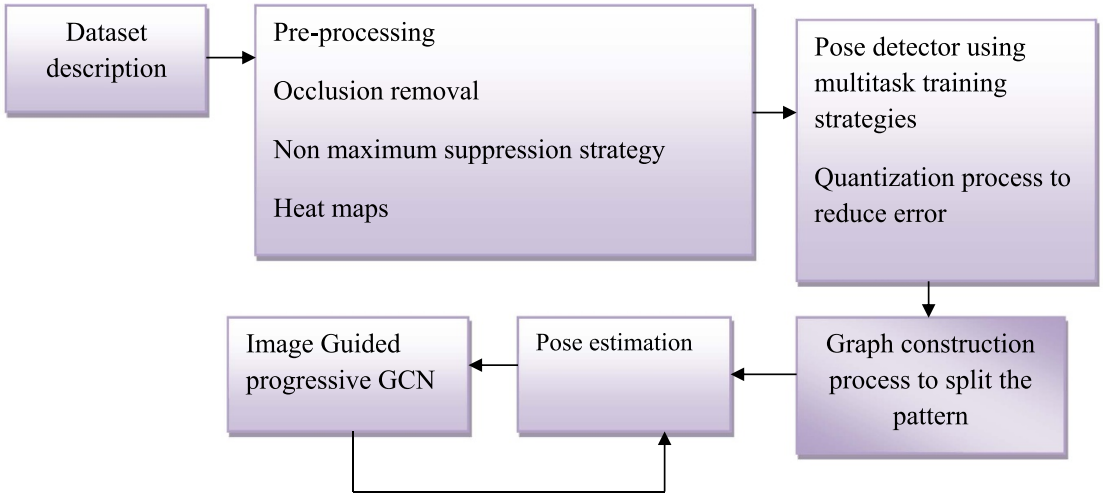

Figure 1. Overall architecture for pose estimation.

3. System paradigm

Figure 1 depicts the work flow of the proposed work. Initially, the input dataset is trained; during this procedure, the image will be pre-processed by embracing transition-related OE, non-maximum suppression strategy (NMSS), and Gaussian heat map methodology. In pre-processing, the image will be recalled and the noise will be eliminated at first; later, the image will be cleared. The noise-eliminated image will be sent to dance PD mode employing multitask training strategies in which the quantization procedure aids in lessening the error. Then, the IGP-GCN framework would be trained to employ the method of training dataset for classification.

3.1. Dataset description

3.1.1. Indian classical dance dataset

The dataset employed in this study contains 626 video recordings gathered out of YouTube, which appertains to the ensuing seven dance formats: Bharatnatyam, Kathak, Kuchipudi, Manipuri, Mohiniyattam, Odissi, and Sattriya. This has been assured that these videos remain clear and efficient having minimal background activity. Every class comprises 30 video clips having a maximal resolution of 400 A-U350 and of 25 s maximal running time. Optimization has been performed with the videos for classes Manipuri, Kuchipudi, and Mohiniattam on YouTube. In the course of data processing, the video segments have been additionally clipped into five to six seconds chunks of frames at 25 fps for creating a maximal of 150 frames. The training to test proportion for analysis was chosen as 7:3. The consequent dataset put forth multiple complications encompassing varied illumination modifications, dancers' shadow effects on the dais, same dance postures, and so on. The less accuracy of the SJ coordinates and the portion of body parts (BPs) missing in a few concatenations turns this dataset so hard. A sum of 420 videos has been taken into account for training intentions. The rest of the videos in every class have been regarded for testing.

3.1.2. UCF-101 dataset

This is a benchmark dataset consists of 101 action categories grouped into 25 groups. Each group consists of 4–7 videos of action. For example, sports category includes videos such as baseball pitch, basketball shooting, bowling, boxing, cricketing etc. For our experimentation we train and test the proffered model with sports videos. Which detects multi PE in videos.

3.2. OE

Occlusion remains a complex issue for tracing humans bound by various situations. Because of inconstant demonstration and comprehensive poses sequence that humans could adopt, identification of persons remains a difficult task either in an image or a video. In real time environment, a notable quantity of partial

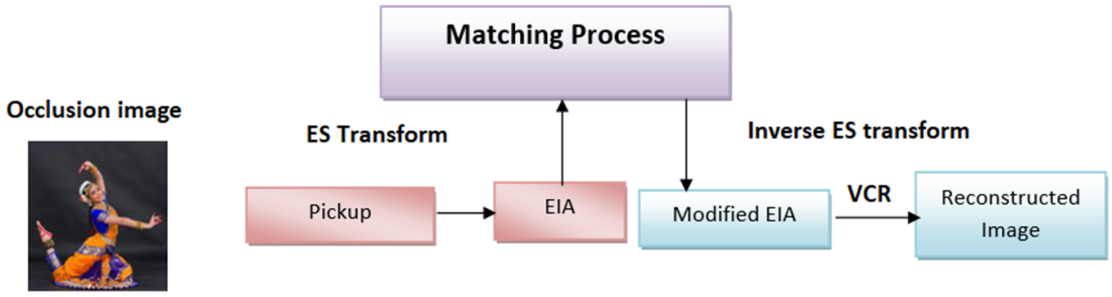

Figure 2. OE procedure.

occlusion (PO) happens as wayfarers move in the proximity of the rest of the objects. Published literature upon tracing system advancement remains chiefly concentrated upon tracing an object moving in indoor settings. The notable progression in person identification and extended tracing have permitted in identifying and tracing of several persons simultaneously in complex scenes. However, systems remain heretofore constantly confronted by PO and complete occlusion that happen usually in practical complexity scenes. This study provides an OE methodology by optimizing the Output image via the unfamiliar occlusion removal. Figure 2 illustrates this system's conception employing the OE methodology. The initial procedure remains to employ a calculative transition between the elementary image array (IA) and sub-IA (SIA). The next procedure is eliminating the unknown occlusion within the SIA by employing dissimilarity information by the sub-image block corresponding algorithm that remains famous in the stereo vision.

The recorded elemental IA (EIA) would be converted into SIA for the proffered OE methodology. This conversion is called ES transform. In other words, we excerpted the similar location for entire EIAs and pixels gathering of similar location has been acquired as SIA. This ES transform can be applied on single pixel excerption or multi pixel excerption. Assume that s_x and s_y indicate the pixels quantity for every elemental image (EI), and l_x and l_x indicate the EI in the x and y axes accordingly. Next, the whole EIs that are indicated as E, turn into $(n_x = s_x\, l_x) \times (n_x = s_y l_y)$ pixels. When $(m \times m)$ pixels have been gathered, the m pixel-related SIA could be computed as,

$$S(i,j) = E\left(t_x s_x + q_x m + r_x, t_y s_y + q_y m + r_y\right) \tag{1}$$

$p_x = \left[\frac{i}{ml_x}\right]$, $q_y = \left[\frac{j}{ml_y}\right]$, $p_x = i\%\left(ml_x\right)$, $p_y = j\%\left(ml_y\right)$, $t_x = \left[\frac{p_x}{m}\right]$, $t_y = \left[\frac{p_y}{m}\right]$, $r_x = i\%m$ and $r_y = j\%m$. $[x]$ remains the gauss function that indicates the biggest integer below or equivalent to the number x, and $a\%b$ indicates the remainder on the division of a by b using the equation (1). We can produce a SIA based on arbitrary pixels from EIA. When the number of pixels increases the resolution of sub images would be increased but distorted at some degree.

A big sampling interval $(z > z_{\max})$ leads to sampling crossing from the correlating lenslet. It might lead to image distortion in SIA. For preventing this scenario, the maximal pixel number (MPN) $m_{\max}$ could be computed for the provided distance z. In which there remains no distortion in the SIA. The MPN could be acquired at the maximal distance $z_{\max}$ in which $n_x = Mm_{\max}$. This could be provided as,

$$m_{\max} = (m_{\max,x} = \frac{g}{z} n_x = \frac{n_x}{M} m_{\max,y} = \frac{g}{z} n_y = \frac{n_y}{M}). \tag{2}$$

In which $M = z/g$. It has been noticed that the SIA has been created having a high resolution by employing the MPN.

3.3. Non maxima suppression strategy in pre-processing

When the occlusion has been eliminated, the NMSS would occur in which the self-similarity, or preference to be chosen as an example, would be naturally selected as a function of the object detector's score: the powerful the output, the most probable a data point must be chosen. The similarity between two windows depends upon their intersection over union (IoU) as $s(i,j) = \frac{i \cap j}{i \cup j} - 1$. In this, the indices indicate the windows' area. It conveys the common area degree that covers within the image correlated with the entire area covered that remains a fine indication of how probable they define the similar object. False positives (FPs) are object hypotheses that belong in fact to the background. Hence, these must not be selected to any cluster or selected

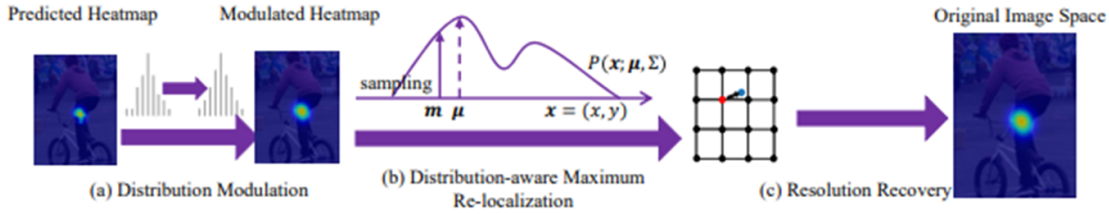

Figure 3. Heatmaps dispensation for pose detector.

as an example. For preventing getting invalid clusters, the relaxation should be rewarded by a penalty for not designating datapoint to any other cluster.

$$I_i\left(c_{i1},\ldots..c_{iN}\right)=\begin{cases}-\infty \text{ if } \sum\limits_j c_{ij}>1\\ \mu \text{ if } \sum\limits_j c_{ij}=0\\ 0 \text{ otherwise}\end{cases} \tag{3}$$

where I_i would be weighted; thus, we could fix $\mu=-1$ without loss of generality. Nevertheless, by default, the affinity-propagation-clustering (APC) does not directly penalize selecting examples, which remain so near to one another as long as theysignify their particular clusters. When this terminology would prefer not to choose windows in a similar neighbourhood, this would not prevent this rigorously too. It would yet permit APC for choosing several objects in immediate proximity. It is indicated by $R=\sum_{i\neq j}R_{ij}\left(C_{ii},C_{jj}\right)$, the new array of repellent local functions, in which $i\neq j$

$$R_{ij}\left(C_{ii},C_{jj}\right)=\begin{cases}r(i,j) \text{ if } c_{ii}=c_{jj}=1\\ 0 \text{ otherwise}\end{cases}. \tag{4}$$

That is, a novel terminology is added for each data point pair that remains active only when the two points are instances. This pair would be penalized by the amount of $r(i,j)$ repellence cost, repeatedly, the repellence cost would be placed between two windows upon their IoU as $r(i,j)=\frac{i\cap^j}{i\cup^j}$. Notice that R_{ij} and R_{ji} denote the similar local function. Nevertheless, the two notations would be sustained for simpleness.

3.4. Heatmaps methodology in pre-processing

The head pose (HP) assessment's output normally possesses two classes: direct regression and transforming into a classification issue that could be known as a 'soft label issue'. Permitting the network for outputting the angle value straight for augmentation learning remains an exceedingly non-linear procedure; the loss function weight limit would be fairly weak, and the feature map's spatial information would be missed. If the HP assessment's output remains transformed into a classification issue the image would be regarded as an entirety, and, hence, it remains frequently requisite for PP the image initially and clip out the head region; or else, the paradigm remains arduous to train. Nevertheless, heatmaps would shortly be a generally employed methodology in HBPE. The fundamental technique remains that a single joint point correlates to a single heatmap. This technique's benefit remains that the output encompasses the two—classification as well as the regression. The classification has been split into two levels—classifying the disparate heatmaps, which differentiate the disparate joint points, and classifying the foreground and background in a single heat map as illustrated in figure 3.

Heatmaps employment assists in avoiding the lesser input resolution's usage for quicker paradigm deduction. It is presumed that the anticipated Heatmap ensues a two-dimension Gaussian dispensation, similar to the actual Heatmap. Hence, the anticipated Heatmap can be portrayed by,

$$G(X,\mu,\partial)=\frac{1}{(2\pi)\,\delta^{1/2}}\exp\left(-\frac{1}{2}(x-\mu).T\delta^{-1}(x-\delta)\right). \tag{5}$$

In which X represents a pixel position within the estimated Heatmap, μ represents the Gaussian mean correlating to the intended assessed joint point. The covariance δ represents a cross-wise matrix, similar to that employed in coordinate encoding. To decrease the approximation difficulty, we use logarithm to convert

the actual exponential shape G to a quadratic shape P to simplify inference through keeping the actual maximal activation region defined as

$$(X, \mu, \partial) = \ln(G) = -\ln(2\pi) - \frac{1}{2}\ln(|\partial|) - \frac{1}{2}(x-\mu)^T \delta^{-1}(x-\mu). \tag{6}$$

Particularly, to match the requirement of our method we proffer Gaussian kernel K with a similar variation as the Training data for smoothening out the impacts of multi-peaks within the Heatmap h by,

$$h' = K * h. \tag{7}$$

In which $*$ indicates the convolution operation (CO). For sustaining the initial Heatmap's dimension, h' would be lastly measured, thereby its maximal activation remains equivalent to that of h through the ensuing transition:

$$h' = \frac{h' - \min(h')}{\max(h') - \min(h')} * \max(h) \tag{8}$$

In which max() and min() given the input matrix's maximal and minimal values accordingly. In this experimental assessment, it is authenticated that the distributed modification additionally enhances the coordinate decoding methodology's execution.

Significantly without any algorithm modification the earlier HPE methodologies effortlessly profited from distribution aware coordinate representation of keypoints.

3.4.1. Pose detection employing multi-tasking

The heterogeneous multi-task architecture comprises two kinds of jobs: (i) a pose regression job in which the goal remains in predicting the human body joint positions within an image, and (ii) an array of body part identification jobs in which the aim remains in classifying, in any case, a window within the image comprises the particular body part. In the ensuing, we assume that the bounding box surrounding the human has previously given, for instance, employing an upper body identifier.

3.4.2. Joint points regression

The regression task remains in estimating the joint points position for every human body part. Every joint points coordinate will be considered as the target values. Entire coordinates would be normalized with the bounding box's dimension, thereby their values would remain in the range of [0, 1]. The squared error would be employed as the cost function for this regression job as,

$$E_r(J'_i, J_i) = \|J_i - J'_i\|. \tag{9}$$

In which J_i and J'_i represents the real and estimated locations for the ith joint accordingly.

3.4.3. Body part identification

For the body part identification jobs, the aim remains in deciding in any case a provided window within the image comprises a particular body part. Consider P remain the sum of BPs' quantity, and L remain the overlapping windows' quantity within the bounding box. For the pth body part, the L classifiers are trained, particularly $C_{p,1}, \ldots . C_{p,L}$, for deciding, in any case, the lth window comprises body part p. Notice that a particular classifier is trained for every position L that permits the body part identifier for learning a position-specified appearance of the body part and also position-specified contextual data with the rest of the body part s. For instance, a lower arm in the bounding box's upper corner would very probably remain vertical or cross-wise. In this training set, the annotated body part s would be portrayed as sticks. This, for training the body part identifiers, it is necessary to initially detect the windows within the training set, thereby it comprises every body part. A window would be regarded to possess a body part when the body part's part within the window remains not less than a specific length compared with the body part's sum length. Particularly, the ensuing formulation is employed for transforming the body part's stick annotation p into a binary label detecting its existence/non-existence within the lth window.

$$y_{p,l} = \begin{cases} 1 \text{ if len}\left(\text{window}_1 \cap \text{stick}_p\right) > \beta.\text{len}\left(\text{stick}_p\right) \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

In which stick_p denotes the pth body part's section, and $\text{window}_1 \cap \text{stick}_p$ depicts the part of the stick_p within the window 1. β denotes a fixed threshold that is experientially fixed as $\beta = 0.3$ in the entire experimentations. Lastly, computing the binary indicator $y_{p,l}$ for every window 1 leads to a binary indicator map for part p. Notice that several body part s are permitted to emerge in a similar window and as well permit a single body part to emerge in multiple windows. For every identification job for part p and window 1, the cross-entropy error function as,

$$E_d\left(y'_{p,l}, y_{p,l}\right) = -y_{p,l}\log\left(y'_{p,l}\right) - 1\left(1 - y_{p,l}\right)\log\left(1 - y'_{p,l}\right). \tag{11}$$

In which $y_{p,l}$ indicates the actual label, and $y'_{p,l}$ indicates the correlating identification probability out of the classifier.

3.5. QE elimination employing optimization architecture

Generally, low-pass filters could be implemented to remove the stair-like figure. Entire high-frequency (HF) elements, encompassing the QE and the edge data, would be eliminated simultaneously. The three-dimensional conception could be extremely affected by the unfinished edge data. Conversely, in another multi-person, although the bilateral filters eliminate HF when sustaining edges, deciding the filter kernel's appropriate dimension remains vital. Filtering with a low-dimension kernel remains ineffective, however, artifacts like halo effects emerge while filtering with a high-dimension kernel. Hence, this study proffers optimization architecture for capturing the QE. Provided the initial signal I and the QS q, it is presumed to employ round-off in quantization, and the quantized signal lX could be designed by,

$$\frac{\boldsymbol{q}}{2} \leqslant \boldsymbol{I} - \boldsymbol{X} \leqslant -\frac{\boldsymbol{q}}{2}. \tag{12}$$

As aforesaid, normally, the QS (sampling interval) within the spatial domain remains very little when compared with the intensity domain for depth pixels; in other words, the pixels must remain smooth in spatial domain. The favoured predicted signal I' could be estimated by cross-wise line segments and horizontal line segments. Concisely, the resultant energy function would be employed for recovering pels out of the quantized pels:

$$I' = \operatorname{argmin}\left(X - I'\right) + \log\left(x\right). \tag{13}$$

Even though limb orientation errors are not collected, joint point errors yet can generate alongside the skeleton tree and likely gather into big errors for joints at the leaf node. For instance, a position shifting in the left shoulder will result in a similar position shifting quantity to the left elbow and also the left wrist. For resolving the issue, extended intentions must be regarded thereby the three-dimensional orientations would be collectively optimized. In this methodology, as the entire phases have been modelled to be distinguishable, the 3D pose loss could be straight employed as an extended intention and train the paradigm end-to-end. In this, the loss for 3D pose is employed:

$$L_{\text{pose}} = \sum_i \sum_k \left| y^i_{k-} y^{i'}_k \right|. \tag{14}$$

In which y^i_k and $y^{i'}_k$ portray the estimated and ground truth three-dimensional positions for joint k in training instance i. In the experimentation, it is observed that the end to end training could accelerate the convergence and also enhance the prediction's accuracy. Altogether, for a T phase paradigm, the comprehensive loss function remains:

$$L = \sum_{T=1}^{t} \partial_1 L^t_1 + \partial_2 L^t_2 + L^t_{\text{pose}}. \tag{15}$$

In which ∂_1, ∂_2 and ∂_3 manage every intention's relative significance. In this experimentation, it is fixed that $\partial_1 = 0.1, \partial_2 = 1$ and $\partial_3 = 1$.

3.6. Graph building procedure for splitting the pattern

Generally, the human KPs build a prominent graph structure centred upon the human body shape, and they possess clear neighbouring associations with one another. Thus, it is regarded that the keypoints localization could be derived finer with the aid of the data suggested by this association. For example, in this architecture, when it is familiar that the guided point remains left elbow, the left wrist's guided point must incline to possess a greater response upon left wrist postulated, since left wrist remains neighbour to left elbow. Thus, additional guidance could be imposed on those keypoints features rather than treating them independently. For taking benefit of this data implied in the graph architecture aforesaid, a graph pose refinement module has been proffered for designing it and, later, refining those keypoints' features. A graph has been constructed and Gaussian convolution for every keypoint has been performed. The output embedding feature could be calculated as,

$$g_k = \frac{1}{z_k} \sum_{s_{k'} \in N(k)} \omega'_k, T_{k'}(f_{k'}) \tag{16}$$

$$\omega'_k = \begin{cases} h_{k'} \mathbb{I}(\boldsymbol{R_{k'}}) \text{ where } k' \neq k \\ 1 \text{ where } k' = k \end{cases} \tag{17}$$

In which $N(k)$ portrays a point set comprising the guided point s_k and its neighbours $T_{k'}$ for the linear transition out of guided point s_k to $\boldsymbol{S_{k'}}$ and $\mathbb{I}$ for the indicator function. $z_k = \sum_{s_{k'} \in N(k)} \omega'_k$, will be employed for normalization. $\boldsymbol{R_{k'}}$ remains a Boolean kind criterion encoding the guided point dependability that functions for filtering out inferior quality points.

For the multi-view pose graph (MPG), the vertices portray the two-dimensional keypoints within a particular camera view. The features have been connected and initiated entire nodes within the graph: (i) visual features R acquired out of the two-dimensional backbone networks' feature maps at the projected two-dimensional position, (ii) one-hot portrayal of the joint kind R, and (iii) normalized original three-dimensional coordinates R. The MPG comprises two edge kinds: (i) single-view edges, which link dissimilar kinds of two KPs within the canonical skeleton form in a particular camera view, and (ii) crow-view edges, which link similar kinds of 2 KPs in different views. One hot feature vector (FV) $\mathbb{R}^2$ is employed for differentiating the two kinds of edges.

3.7. IGP-GCN building

The main method proffered in this study remains IGP-GCN for correction. In this network, the image context and pose structure clues of invisible joints inference are fused. The particulars of every layer and ResGCN Attention Blocks (ResGCNAB) will be explained in additional materials.

- The predictable position of imperceptible joints from the base module is occasionally far from their exact locations and this makes it a complex challenging task to directly revert their displacements. Hence, we design an intuitive coarse-to-fine learning procedure has been designed in the coordinate-based module, which constructs a progressive.

  GCN architecture and influences the performance steadily by enforcing multi-scale image features in a progressive way.
- There exists a lack of local context information due to coordinate-based module. As a result, the concerned IFs for every joint points have been excerpted and merged into the module. That is, the PE outcomes have been enhanced by integrating image featuremap (FM)s $F'_1, F'_2,$ and F'_3. Particularly, cascaded ResGCNAB have been designed for grasping the beneficial data, which has been saved in the FMs yet missed in the original pose $\widehat{p}_i$. The three FMs have been arranged out of coarse for fining as per the receptive fields' dimension. Next, a grid sample methodology has been utilized, which attains the jth JF by excerpting the feature positioned in $\left\langle x_i^j, y_i^j \right\rangle$ upon the concerned coordinate weight FM. Each pose results in three node FVs $J'_1, J'_2,$ and J'_3 are excerpted ensuing this procedure.

3.7.1. Self attention module

Shuffled attention mechanism (SAM) [24] would be employed in the multilevel network's final module for shuffling and weighting the output functions. As illustrated in figure 4, SAM's initial unit remains the residual connections' channel shuffling. Subsequent to shuffling, a $1 * 1$ CO and a sigmoid activation function (SAF) would be implemented for attaining the space attention α. SAM's last portion remains the

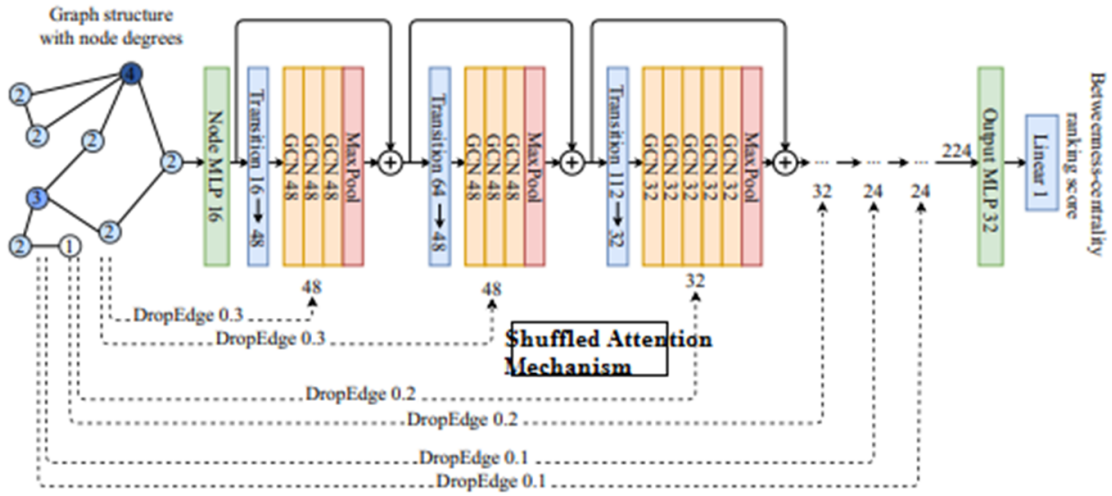

Figure 4. IGP-GCN framework.

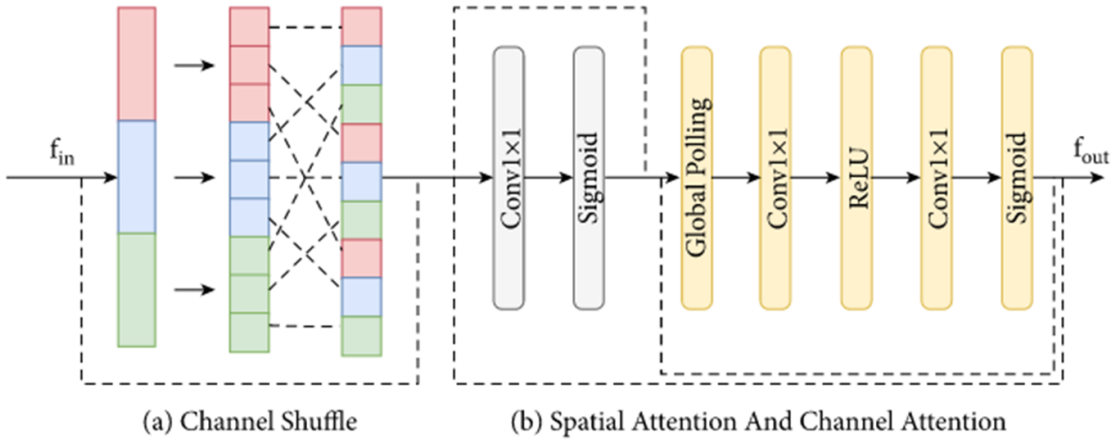

Figure 5. Shuffle attention mechanism architecture.

channel attention (CA) that comprises a global pooling, two $1 * 1$ COs, a ReLU activation function (ReLUAF), and a SAF for acquiring the CA vector β.

3.7.2. Channel shuffle (CS) operation

In this we apply CS operation instead of dense convolution to attain the feature communication. The CS operation could be designed as a process shown in figure 5(a), it is composed of 'reshape-transpose-reshape' procedures. Consider that the IP layer would be split into arrays, the IP feature would be reshaped into $G \times N$ sizes wherein N represents the channels' quantity within every array. Next, the features would be translated into (N, G) sizes to guarantee that separate groups are used as the input for the subsequent group convolution process. Lastly, this is reshaped into dimensions, thereby the data could move between different arrays. The shuffled feature would be fused with the initial by component-wise sum for establishing the CS module's output.

Assume that the SAM's input remains f_{in}; it as well remains the final multi-stage polymerization module's output. The CS could be derived as,

$$f_{\text{CS}}^{\text{out}} = \text{CS}\,(f_{\text{in}}) + f_{\text{in}}. \tag{18}$$

In which CS (.) portrays the CS operation, and $f_{\text{CS}}^{\text{out}}$ portrays the CS module's output.

3.7.3. Attention Mechanism

Spatial attention (SA): the feature map results in the keypoints position's unwanted outcomes because of the regions' presence in the spatial data, which remains unrelated to keypoints. The purpose of SA mechanism shown figure 5(b) is to decrease the inference of the unrelated areas and highlight the areas related to locating task by weight the feature map. The spatial-wise attention weight $\propto$ would be created by a CO ensued by a sigmoid function upon the IP. The SA could be derived by,

$$\propto = \text{sigmoid}\left(\text{Conv}\left(W, f_{\text{CS}}^{\text{out}}\right)\right). \tag{19}$$

In which $Conv(.)$ indicates the CO, W denotes the CO's learning weight, and sigmoid $(.)$ denotes the activation function. Lastly, the self-attention weight $\propto$ would be resized, and the output would be described in the following expression. $f_{\text{out}}^{\text{at}}$ denotes the SpAM's output.

$$f_{\text{out}}^{\text{at}} = f_{\text{CS}}^{\text{out}} * (\propto +1). \tag{20}$$

Channel Attention(CA): FM's every channel remains the correlating convolutional layer's feature activation. As a convolution just performs in a local space, it remains difficult for acquiring adequate data for excerpting the association betwixt the channels. Motivated by the Squeeze-and-Excitation Network [25] that employs an excitation unit for learning the FM's weight of every convolution layer, CA is considered the procedure to adaptably choose the convolution layer. In this squeeze phase, the SpAM's output feature $f_{\text{out}}^{\text{at}}$ would be employed as CA input. The whole spatial feature upon the channel would be encoded as a global feature and employ global mean pooling upon $f_{\text{out}}^{\text{at}}$ for producing channel statistics $z \in R^c$ that is described by,

$$z_t = \frac{1}{H * L} \sum_{i=1}^{H} \sum_{j=1}^{L} U_t(i,j). \tag{21}$$

In which z_t indicates the tth component in Z , and U_t indicates the output of the tth convolution kernel within the CA network. The squeeze procedure acquires the global definition attributes, yet we require one more procedure for capturing the association between channels. This should be capable of learning the nonlinear association between every channel. Furthermore, the learnt association remains compatible since the multichannel feature has been permitted rather than the onehot form. Hence, a Sigmoid gating procedure would be employed for channel statistics (z) described as,

$$\beta = \text{signoid}\left(\text{Conv}\left(W_2, \text{ReLU}\left(\text{Conv}\left(W_1, Z\right)\right)\right)\right). \tag{22}$$

In which $W_1 \in R^{c \times c}$ and $W_2 \in R^{c \times c}$ portray the learning criteria within the two FC layers, and ReLU $(.)$ represents the ReLUAF.

Lastly, the CA weight β would be learnt by SAM. SAM's output could be produced by,

$$f_{\text{out}}^{\text{SAM}} = f_{\text{out}}^{\text{at}} * (\beta + 1). \tag{23}$$

SAM module's loss could be described as,

$$L_{\text{SAM}} = \frac{1}{K} \sum_{j=1}^{K} (Y_j^{\text{SAM}} - Y_j')\text{^}2. \tag{24}$$

In which Y_j^{SAM} remains the jth KP's heatmap anticipated by SAM's feature.

4. Performance analysis

The experimental outcome has been performed by assessing criteria employed for assessment including accuracy, sensitivity, specificity, f1-score, kappa score, relative absolute error (RAE), and mean absolute error (MAE). Such criteria have been correlated with two advanced methodologies like DCNN and GRR-GCNN with the proffered Occlusion Removed_Image-Guided progressive Graph CN (OccRem_IGP-GCN).

Accuracy provides the capability of the comprehensive anticipation generated by the paradigm. True positive and true negative give the ability to anticipate the data's existence and non-existence. FP and false negative (FN) provide the wrong anticipations done by the employed paradigm.

Table 1. Accuracy correlation.

No. of epochs	DCNN	GRR-GCNN	OccRem_IGP-GCN
100	89	91	92
200	90.1	92	93
300	90.5	93	94
400	91	94	96
500	92.3	95.8	97.5

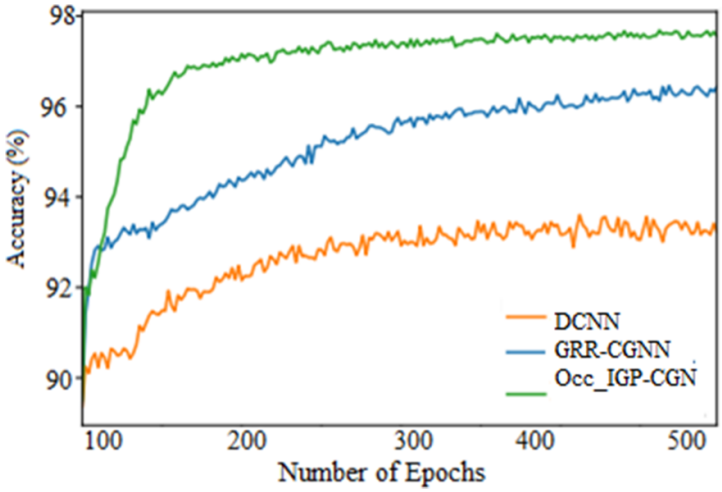

Figure 6. Accuracy correlation.

Table 2. Sensitivity correlation.

No. of epochs	DCNN	GRR-GCNN	OccRem_IGP-GCN
100	86	88	89
200	88	89	90
300	89	90	91
400	90	91	92
500	91	92.5	93

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{25}$$

Table 1 exhibits the accuracy correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Figure 6 exhibits the accuracy correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered Occlusion Removed_Image Guided Progressive Graph Convolution Network (OccRem_IGP-GCN) methodologies in which the X-axis portrays the epochs' quantity employed for the assessment, and the Y-axis portrays the accuracy values acquired in percentage. While correlated, the prevailing DCNN and GRR-GCNN methodologies attained 94% and 95% of accuracy accordingly, whereas the proffered OccRem_IGP-GCN methodology attained 98% of accuracy that remains 4% finer than DCNN and 3% finer than GRR-GCNN methodologies.

Sensitivity predicts the classification paradigm's efficacy. This remains the probability of data's positive anticipation that is detected as well named TP Rate and described by,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{26}$$

Table 2 exhibits the sensitivity correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Figure 7 exhibits the sensitivity correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies in which the X-axis portrays the epochs' quantity employed for

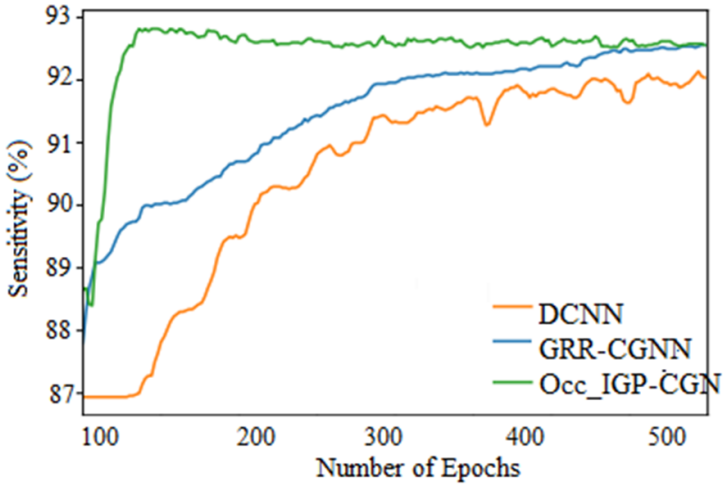

Figure 7. Sensitivity correlation.

Table 3. Specificity correlation.

No. of epochs	DCNN	GRR-GCNN	OccRem_IGP-GCN
100	84	85	87
200	85	87	89
300	85	88	90
400	86	89	91
500	87	90	92

the assessment, and the *Y*-axis portrays the sensitivity values acquired in percentage. While correlated, the prevailing DCNN and GRR-GCNN methodologies attained 90% and 91% of sensitivity accordingly, whereas the proffered OccRem_IGP-GNN methodology attained 93% of sensitivity that remains 3% finer than DCNN and 3% finer than GRR-GCNN methodologies.

Specificity remains the TN's probability that is properly detected and as well named TN Rate. This is computed by,

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{27}$$

Table 3 exhibits the specificity correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Figure 8 exhibits the specificity correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies in which the *X*-axis portrays the epochs' quantity employed for the assessment, and the *Y*-axis portrays the specificity values acquired in percentage. While correlated, the prevailing DCNN and GRR-GCNN methodologies attained 87% and 90% of specificity accordingly, whereas the proffered OccRem_IGP-GCN methodology attained 92% of specificity that remains 5% finer than DCNN and 3% finer than GRR-GCNN methodologies.

F1-score will be employed for deciding the anticipation execution. This remains the weighted mean of precision and recall. The value of one remains the finest whereas zero remains the poorest. F1-score in no way regards TNs and can be computed by,

$$\text{f1} - \text{Score} = \frac{2 * P * R}{P + R}. \tag{28}$$

Table 4 exhibits the f1-score correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Figure 9 exhibits the f1-score correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies in which the *X*-axis portrays the epochs' quantity employed for

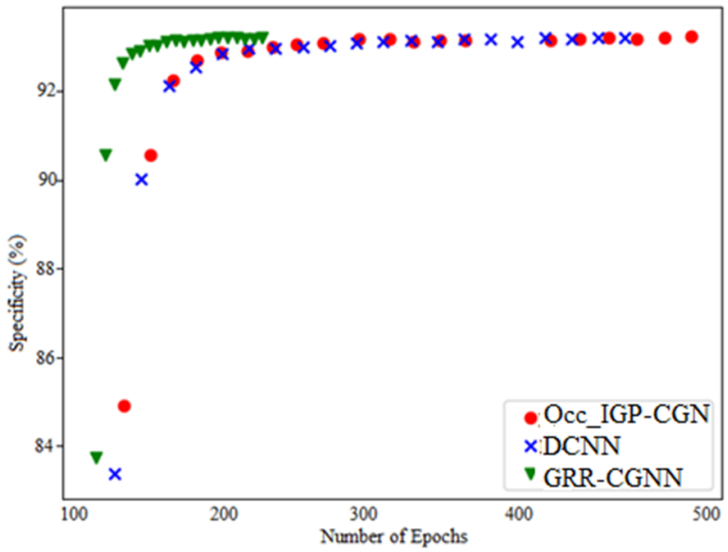

Figure 8. Specificity correlation.

Table 4. F1-score correlation.

No. of epochs	DCNN	GRR-GCNN	OccRem_IGP-GCN
100	78	79	80
200	80	81	82
300	82	83	84
400	84	85	86
500	86	87	88

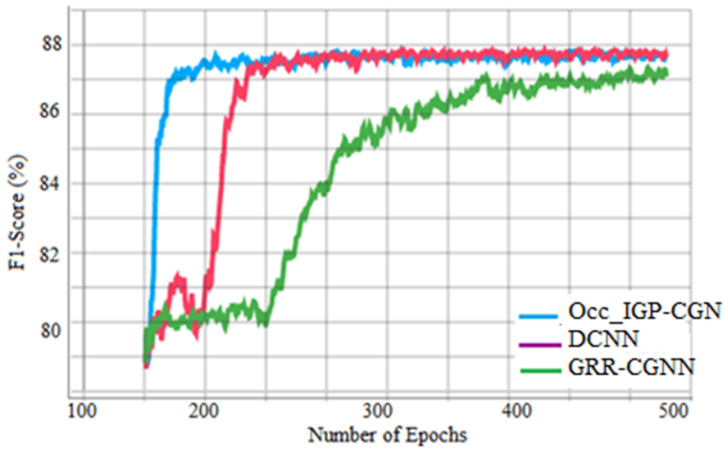

Figure 9. F1-score correlation.

the assessment, and the *Y*-axis portrays the f1-score values acquired in percentage. While correlated, the prevailing DCNN and GRR-GCNN methodologies attained 86% and 87% of the f1-score accordingly, whereas the proffered OccRem_IGP-GCN methodology attained 88% of f1-score that remains 2% finer than DCNN and 1% finer than GRR-GCNN methodologies.

RAE indicates the proportion that correlates a mean error (residual) with the errors generated by a trivial or naive paradigm. It is computed by,

Table 5. RAE correlation.

No. of epochs	DCNN	GRR-GCNN	OccRem_IGP-GCN
100	72	59	50
200	69	57	49
300	60	56	47
400	55	54	44
500	50	49	42

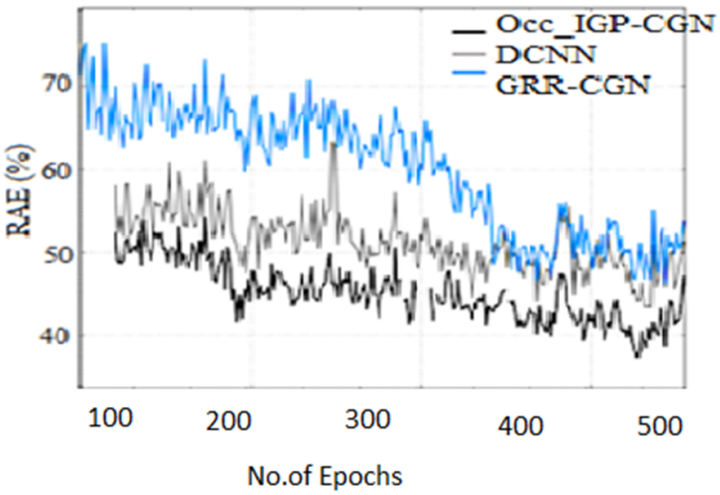

Figure 10. RAE correlation.

$$\text{RAE} = \frac{\sum_{i=1}^{n}(pi - Ai)^2}{\sum_{i=1}^{n} Ai}. \tag{29}$$

Table 5 exhibits the RAE correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Figure 10 exhibits the RAE correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies in which the *X*-axis portrays the epochs' quantity employed for the assessment, and the *Y*-axis portrays the RAE values acquired in percentage. While correlated, the prevailing DCNN and GRR-GCNN methodologies attained 50% and 49% of RAE accordingly, whereas the proffered OccRem_IGP-GCN methodology attained 42% of RAE that remains 8% finer than DCNN and 7% finer than GRR-GCNN methodologies.

MAE remains an errors measurement betwixt coupled observances exhibiting a similar phenomenon. Instances of *Y* vs. *X* encompass correlations of anticipated vs. noticed, next time vs. original time, and a single computation approach vs. alternate computation approach. It is calculated by,

$$\text{MAE} = \sum_{i=1}^{n}(yi - xi). \tag{30}$$

Table 6 exhibits the MAE correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Figure 11 exhibits the MAE correlation betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies in which the *X*-axis portrays the epochs' quantity employed for the assessment, and the *Y*-axis portrays the MAE values acquired in percentage. While correlated, the prevailing DCNN and GRR-GCNN methodologies attained 44% and 40%of MAE accordingly, whereas the proffered OccRem_IGP-GCN methodology attained 30% of MAE that remains 14% finer than DCNN and 10% finer than GRR-GCNN methodologies.

Table 7 exhibits the comprehensive correlation for diverse criteria betwixt the prevailing DCNN and GRR-GCNN with the proffered OccRem_IGP-GCN methodologies.

Table 6. MAE correlation.

No. of epochs	DCNN	GRR-GCNN	OccRem_IGP-GCN
100	55	50	40
200	50	48	46
300	48	46	45
400	46	44	32
500	44	40	30

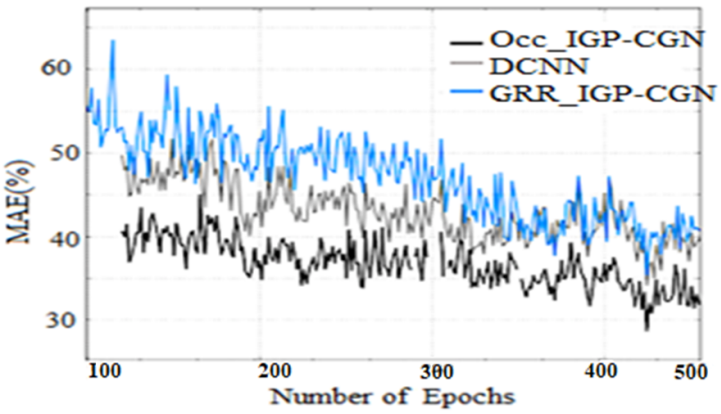

Figure 11. MAE correlation.

Table 7. Comprehensive correlation of the prevailing and proffered methodologies.

Criteria	DCNN	GRR-GCNN	OccRem_IGP-GCN
Accuracy (%)	94	95	98
Sensitivity (%)	90	91	93
Specificity (%)	87	90	92
F1-score (%)	86	87	88
RAE (%)	50	49	42
MAE (%)	44	40	30

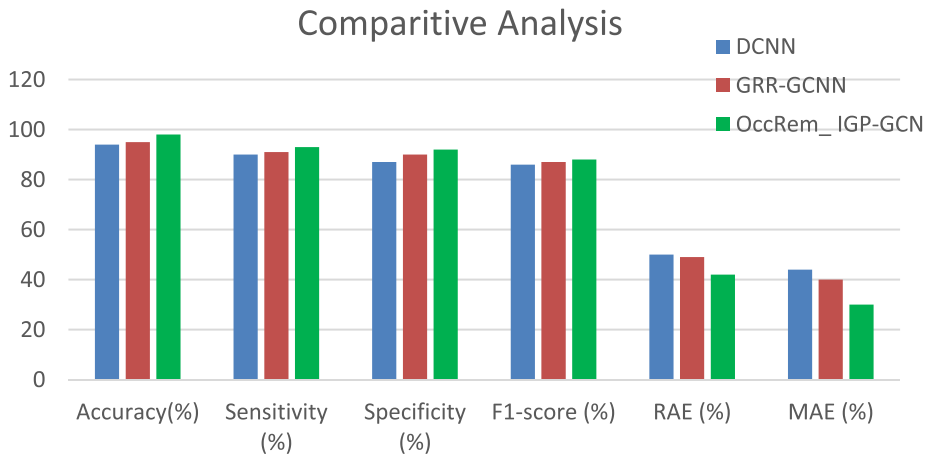

Figure 12. Performance graph correlation with other methodologies.

Figure 12 depicts the performance graph of proffered method OccRem_IGP-GCN by comparing with the existing methodologies DCNN, GRR-GCNN. Compared to other methodologies the proffered methods show best performance.

Figure 13. Validation results on ICD video sequences.

The proffered model was tested with classical dance video for estimating poses of multiple persons. Figure 13 shows some validation results on Indian classical dance videos. Figure 14 depicts some validation results on UCF-101 dataset. This PD model automatically corrects the wrong poses.

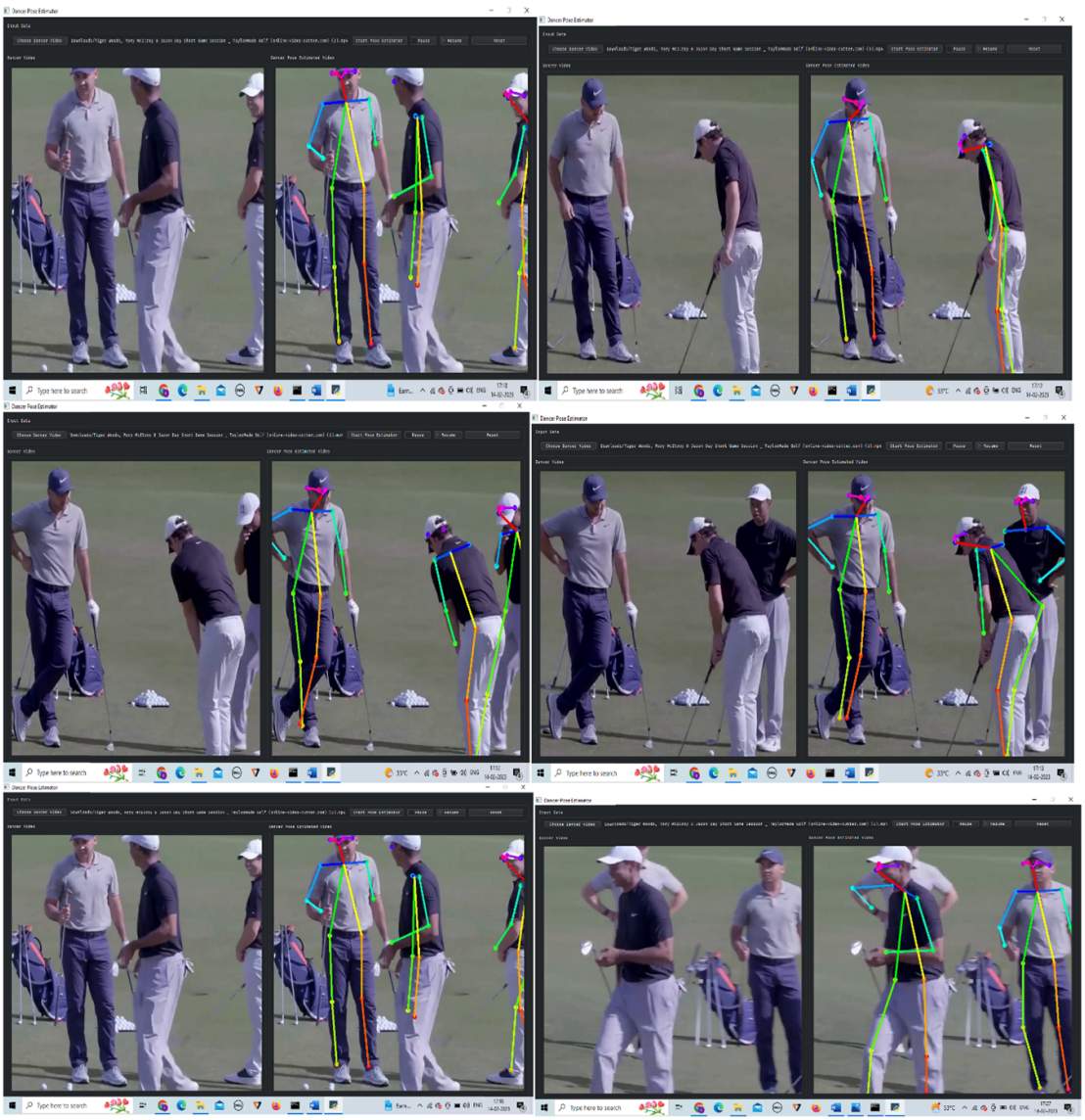

Figure 14. Validation results on UCF-101 dataset.

5. Conclusion

This study introduces a new design model called OccRem_IGP-GCN for attaining the effectual network for HPE and a new learning framework (LF) for efficiently training this effectual network. From what we have known, this remains the foremost trial in analysing OccRem_IGP-GCN designed with the feature model that greatly lessens the calculative price. Additionally, we learned the convergence conduct and devised a new LF to speed up its convergence and enhance its accuracy. This methodology enables the low-latency and low-energy cost implementation as needed in the non-GPU settings. Comprehensive experimentation has been performed, and it has been found that the proffered OccRem_IGP-GCN attained 98% of accuracy, 93% of sensitivity, 92% of specificity, 88% of f1-score, 42% of RAE, and 30% of MAE.

Data availability statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Conflict of interest

On behalf of all authors the corresponding author states that there is no conflict of Interest.

Consent to participate

All Authors consented for participating in this study.

Funding

No funding was received for conducting this study.

Web links

www.kaggle.com/datasets/pevogam/ucf101?resource=download

ORCID iDs

Jhansi Rani Challapalli https://orcid.org/0000-0002-9113-3823
Nagaraju Devarakonda https://orcid.org/0000-0003-4864-8482

References

[1] Lu J, Yang T F and Zhao B 2021 *A Review of Deep Learning-Based Human Pose Estimation* (Beijing: Laser & Optoelectronics Progress)
[2] Wang C Y, Wang Y Z and Yuille A L 2013 An approach to pose-based action recognition *IEEE Conf. on Computer Vision and Pattern Recognition* (*Portland, OR, USA*) pp 915–22
[3] Liang Z J, Wang X L, Huang R and Lin L 2014 An expressive deep model for human action parsing from a single image *IEEE Int. Conf. on Multimedia and Expo (ICME))* (*Chengdu*) pp 1–6
[4] Murphy-Chutorian E and Trivedi M M 2019 Head pose estimation in computer vision: a survey *IEEE Trans. Pattern Anal. Mach. Intell.* **31** 607–26
[5] Dalal N and Triggs B 2015 Histograms of oriented gradients for human detection *IEEE Conf. on Computer Vision and Pattern Recognition* vol 1 pp 886–93
[6] Deleforge A, Forbes F and Horaud R 2014 High-dimensional regression with Gaussian mixtures and partially-latent response variables *Stat. Comput.* **25** 893–911
[7] Felzenszwalb P F and Huttenlocher D P 2005 Pictorial structures for object recognition *Int. J. Comput. Vis.* **61** 55–79
[8] Fan X, Zheng K, Lin Y and Wang S 2015 Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (*June 2015*)
[9] Yang Y, Baker S, Kannan A and Ramanan D 2012 Recognizing proxemics in personal photos *2012 IEEE Conf. on Computer Vision and Pattern Recognition* (*June 2012*) pp 3522–9
[10] Chu X, Yang W, Ouyang W, Ma C, Yuille A L and Wang X 2017 Multi-context attention for human pose estimation (arXiv:1702.07432)
[11] Carreira J, Agrawal P, Fragkiadaki K and Malik J 2016 Human pose estimation with iterative error feedback *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 4733–42
[12] Long J, Shelhamer E and Darrell T 2015 Fully convolutional networks for semantic segmentation *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 3431–40
[13] Yu H, Du C and Yu L 2022 Scale-aware heatmap representation for human pose estimation *Pattern Recognit. Lett.* **154** 1–6
[14] Ahmad N and Yoon J 2021 StrongPose: bottom-up and strong keypoint heat map based pose estimation *2020 25th Int. Conf. on Pattern Recognition (ICPR)* (*January*) (IEEE) pp 8608–15
[15] Wang W, Zhang K, Ren H, Wei D, Gao Y and Liu J 2022 UULPN: an ultra-lightweight network for human pose estimation based on unbiased data processing *Neurocomputing* **480** 220–33
[16] Gao C, Yang Y and Li W 2022 3D interacting hand pose and shape estimation from a single RGB image *Neurocomputing* **474** 25–36
[17] Zhang Z, Luo Y and Gou J 2021 Double anchor embedding for accurate multi-person 2D pose estimation *Image Vis. Comput.* **111** 104198
[18] Gao W, Liu L, Zhu L and Zhang H 2022 Visible–infrared person re-identification based on key-point feature extraction and optimization *J. Vis. Commun. Image Represent.* **85** 103511
[19] Zhang B, Xiao Y, Xiong F, Wu C, Cao Z, Liu P and Zhou J T 2022 3D human pose estimation with cross-modality training and multi-scale local refinement *Appl. Soft Comput.* **122** 108950
[20] Wang X, Hu X, Li Y and Jiang C 2022 Multi-modal human pose estimation based on probability distribution perception on a depth convolution neural network *Pattern Recognit. Lett.* **153** 36–43
[21] Wang Z, Liu G and Tian G 2018 A parameter efficient human pose estimation method based on densely connected convolutional module *IEEE Access* **6** 58056–63
[22] Lin C M, Tsai C Y, Lai Y C, Li S A and Wong C C 2018 Visual object recognition and pose estimation based on a deep semantic segmentation network *IEEE Sens. J.* **18** 9370–81
[23] Wang R, Huang C and Wang X 2020 Global relation reasoning graph convolutional networks for human pose estimation *IEEE Access* **8** 38472–80
[24] Bai Y F, Zhang H B, Lei Q and Du J X 2021 Multistage polymerization network for multiperson pose estimation *J. Sens.* **2021** 1–10
[25] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proc. IEEE Conf. on Computer Visionand Pattern Recognition* (*Salt Lake City, Utah*) pp 7132–41