ConvPose: Application of CNN for Estimating Human Poses

Vividha[1], Anjali Khandelwal[2], Anubhav Singh[3], Preeti Nagrath[4], Narina Thakur[5]
Bharati Vidyapeeth's College of Engineering[1,2,3,4]
vividha.cse1@bvp.edu.in[1], Anjalikhandelwal2910@gmail.com[2],
Anubhav.singh2359@gmail.com[3], preeti.nagrath@bharatividyapeeth.edu[4], narinat@gmail.com[5]

Abstract*—Human pose estimation has always been a chal- lenging problem, especially in the detection of human joints in the various body poses in order to get a posture frame. This trend leads to several architectures and models that are computationally expensive and require costly pre-training of datasets and hardware equipment, also it leads problem in comparing other techniques with replicate existing outcomes. Therefore, to resolve this issue, this paper, discusses the challenges faced in detecting human joints and give an overview of the substantial research work done in this area. In this paper, an efficient CNN-based regression model has been developed that can easily be trained on mid-run 1inimi GPU towards estimation of body keypoints and identifying the joints that result in low loss value, for the estimation of the MPII dataset and its sub-regions. This technique shows consistent improvement over the dataset where the data was curated and crafted as per the needs. The dataset images were cropped to increase the focus on the humans in the images. This paper modifies the images excluding the not required background details, thus only training the human body image pixels in the proposed network. The paper also shows the results obtained by the proposed model with the modelled dataset.

***Index Terms*—Human pose estimation, MPII human pose dataset, CNN, ConvNets.**

I. Introduction

Human Pose Estimation is the computer vision method to identify the pose of the person in the image through the mapping of body joints in the body. Pose Estimation is being researched for the last 15 years. Human pose estimation finds importance due to its abundance of application and benefits that could be gained from this technology [1].

With a wide range of application in Augmented Real- ity(AR), Security Drones, Sports insights to assisted living and many more pose estimation is high boosting concept in tfhe future. This technology finds its applications ranging from character animation in gaming to clinical medicine

Despite enough research for many years, there are still not satisfactory results that give a perfect solution to all the challenges faced by human pose estimation. Major challenges faced by human pose estimation are: (1) changing human visual appearance in images, (2) changing light conditions (3) variability in human physique, (4) partial occlusions due to self-articulation and layering of objects in the scene, (5) complexity of human skeletal structure and (6) high dimensionality of the pose.

In spite of enough research in this field, there are as yet not agreeable outcomes that give an ideal answer for every one of the difficulties looked by human posture estimation. Significant difficulties looked by human posture estimation like halfway impediments because of self-explanation and layering of articles in the scene or high dimensionality of the posture, multifaceted nature of human skeletal structure, changing human visual appearance in pictures, changing light conditions and inconsistency in human body.

The problem of estimating human poses involves identifying the location of the body key points, which include main parts and joint elements. Identification of human body joints faces challenges of occlusions, small joints and the need to capture context. Over this unrestricted background, uneven lighting, scales, unconstrained human appearance, and some tough poses add on to the troubles in human postures. Figure 1 displays the color coded body joints.

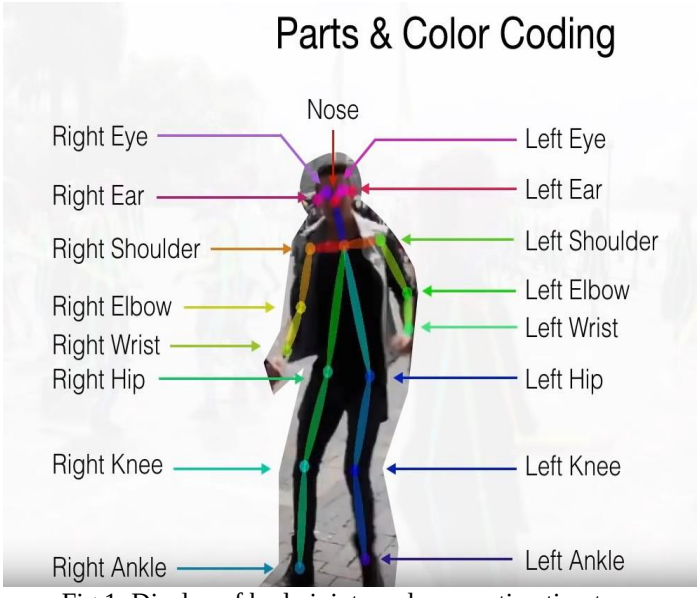

Fig 1: Display of body joints and pose estimation tree

In an approximation of complex and non – linear mapping functions, neural networks have a high capacity. Mapping functions like 2D images of a random person to the joint locations with additional challenges. Fundamentally, convolution neural networks (CNNs) are very similar to ordinary neural networks consisting of neurons with adjusting weights and predispositions [2].

But, CNN is computationally lot better than traditional neural network ensuring translational invariance of the objects in the image. Neural Network generates a lot of parameters since every layer is fully connected to the previous layer. But on CNN, a neuron in a layer is connected to neurons of other layers who are spatially close to them thus trimming a large number of connections between layers. Also, CNN has two non-linearity functions-Pooling and Relu [3]. Pooling acts on the input image and passes the maximum or minimum values only reducing the size of the output and reducing extra parameter to learn. Helping CNN to limit computational time. It is not required to design features and detectors for parts instead it learns them from the data. Involving full transformation of the input image instead just focusing on predefined local detectors thus limiting the output.

The author presents a simple and efficient solution for pose estimation through MPII Dataset giving promising results. The paper proposes to use human parsing information all the more adequately and effectively to adapt better posture estimation models and enhancing their execution. [4]

This paper discusses considering the problem of human pose estimation as a regression problem of estimation of 16 human body joints which determines the pose.

In a nutshell, the model is generating pose skeleton of the human from the input image. This paper proposes a CNN model which is straightforward and efficient and could be trained on a mid run GPU. The proposed model uses simple functions like relu and connectivities that makes it an efficient and easily usable. This paper also explains the results obtained which were compared by a loss function. This loss is mapped giving a 2inimize2d2on of accuracy and capability of the model. The decrease in loss function add up to the accuracy of the model.

II. Problem Statement

The objective is to estimate, the pose of humans in the input images. For this estimation the network is fed with the raw images and the expected output is a vector of coordinates of the body key points. The objective is to find x-y pixel coordinates of 16 human body joints by training a regression CNN that compensates the loss.

III. Related Work

Progressive activity extractors, for example [5] Convolutional Networks(ConvNets) [6] that have accomplished noteworthy execution on an assortment of order errands utilizing absolutely feedforward preparing. Feedforward preparing and can learn rich portrayals of the info space, however, don't unequivocally display conditions in the yield spaces that are very organized. Therefore the author proposes a framework in their research work that continuously grows the expressive intensity of progressive element extractors to incorporate both info and yield spaces, by presenting top-down criticism.

The researcher thinks [7] about the assignment of the 2inimize2d human pose by identifying the multi-person in realistic images. They propose a methodology that determines the number of people present in the related image and also recognized the blocked body parts, they have created a framework with CNN-based part finders.

The author proposes a work [8] and design of convolutional systems that work on conviction maps from past stages, that provide assessments for part areas, without any need for unequivocal graphical model-style deduction. They exhibit extraordinary approach on standard benchmarks including the MPII, LSP, and FLIC datasets.

The researcher present [9] a novel design that incorporates a proficient position refinement display that is prepared to assess the joint counterbalance areas inside a little area of the picture. This refinement demonstrate is together prepared to assess the joint counterbalance area inside a little area of the picture.

Contributing through the state-of-the-art of articulated pose estimation also in the multi-person pictures proposing improved body parts detectors and an incremental streamlining methodology producing multi-individual posture estimation results. [10]

CNNs have always been primarily used for classification problems. A classification CNN model can be converted to a regression model for localization by replacing the last classification layer with a regression layer for which the activation is real-valued predictions. Chen and Yuille proposed a graphical method to estimate the pose estimation using a graphical model. The graphical model took advantage of local image measurements to detect both parts (or joints) and spatial relations between them. The authors represented the spatial relationship between joints by a combined model over a variety of spatial relationships exploiting the extreme power of DCNNs [11]. The presented method outperformed the state of the art methods on the datasets like LSP [12] and FLIC [12]. It also showed good results on the Buffy dataset.

Sermenet and others proposed an embedded strategy with a single CNN for object detection, recognition and localization [13]. Bearman and Cond have also developed their own CNN model for human pose estimation and activity classification [14].

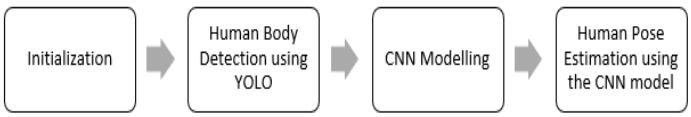

Fig 2: Flowchart of the Proposed System

At the abnormal state, human activities can regularly be precisely identified due to a combination of body pose, movement, and interaction with scene objects [15]. In any case, because of the challenging nature of this problem, most recent models for activity recognition generally depends on holistic representations that extricate appearance and motion features from the video from which the images are taken. Recently, Toshev and others [16] demonstrated the advantage of reasoning about poses by using the regression model rather than the classification in a simple but comprehensive fashion. The pose estimation was defined as a body joint regression problem, a CNN model of 7 layers was used to predict each joint location using a full image as an input. A state-of-the-art PCP score (0.61) was achieved using this method and it proved to be much simpler than past techniques based on explicitly designed feature representations and graphical models.

Determination of 3D-human pose [17] from monocular RGB image data is a very challenging issue. As the 2D data results in extensive ambiguities that make depth analysis difficult. The researchers have exhibited an adaptable statistical framework determining the ability of deep neural network to make such surmisings with sensible exactness. Many substan- tial models use various coordinates that produce methods that are memory-concentrated and not differentiable. The authors propose enhancements to arrange 3D data and forecast the pre- defined unnecessary attributes by predicting two-dimensional heatmaps. The researcher Margi33pose model [17], determine reasonable heatmaps after differentiability.

The authors addressed [18], the more troublesome instance of utilizing a solitary camera as well as not utilizing markers going legitimately from 2D appearance to 3D geometry. Deep learning approaches have demonstrated noteworthy capacities to selectively learn 2D appearance highlights. The missing thing is the means by which to coordinate 2D, 3D, and worldly data to recoup 3D geometry and record for the vulnerabilities emerging from the discriminative model. Notably, the strategy suggested does not involve synchronized 2D-3D training data and applies toin-the-wild images shown with the MPII dataset. The researcher [19] proposes a start to finish design for identification for body joints of 2D and 3D human pose in normal images. The key to our methodology is the age and scoring of various posture proposition per picture, which en-ables us to foresee 2D and 3D stances of numerous individuals at the same time. The method recovers full-body 2D and 3D attributes, analyzing conceivable body parts when the people are somewhat blocked or truncated by the boundary of the picture. Additionally, it demonstrates promising outcomes on genuine pictures for both single and multi-individual subsets

of the MPII 2D present benchmark.

Using specific vector machines[20], hierarchical SVMs[21], or random forest classifiers[22], The estimation of 3D pose is conceived as a regression problem with the functions retrieved. Recently, CNNs also received a lot of recognition for determining the 3D human pose. However, it is often conceived as a regression problem rather than a classification problem as the search space in 3D is much larger than the 2D image space. Sijin Li and others used CNNs to explicitly learn 3D human pose from the input images [23]. CNNs learn about the relative 3D location to the parent joint through regression. They have used sliding window design 2D component detectors for each joint. They consider that the loss function, combining 2D joint classification with 3D joint regression, helps improve the 3D pose estimation performance. Weichen Zhang and others improved the 3D pose estimation performance by incorporating a formal system for learning CNNs [24]. Tekin and others recently proposed a structured prediction framework, which uses an auto-encoder to learn 3D pose representations [25]. It regresses to a high-dimensional pose representation learned by an autoencoding to explicitly encode dependencies between joints. The authors also stated that the temporary video sequence information also helps in predicting the result of a more accurate pose estimate. Zhou and others used the 2D pose estimate result for the reconstruction of a 3D pose [26]. As a weighted sum of shape bases comparable to the traditional non-rigid structure from motion, they formed a 3D pose and developed an EM algorithm that formulates the 3D pose as a latent variable when 2D pose estimation results are visible. It addresses the 3D pose estimation problem for a monocular image sequence which incorporates 2D, 3D and temporal information to compensate for inconsistencies in model and calculations. The approach does not require synchronized 2D-3D training data, i.e. only 2D pose annotations are required to train the CNN joint regressor and a separate 3D dataset for learning the sparse 3D base.

Pons-Moll and others proved in their research that although CNNs dominate in the estimation of the 3D pose, regression forests were also used to efficiently extract 3D pose bit descriptors [27]. The representations of input and output are important too. Ionescu and others found that to locate the person, the input image is generally cropped to the subject's bounding box before estimating the 3D pose [28]. Video input provides temporal indications that translate into increased precision [29]. The downside of movement conditioning is the increased dimensionality of the input, which requires movement databases with ample movement variance, which are much harder to capture than pose data set. The fixed camera positioning offers additional height indications in controlled conditions [30]. Since monocular reconstruction is ambiguous in size, 3D joint positions relative to the pelvis are commonly used as the production, with normalized subject height. Zhou and others used a kinematic model to calculate the joint angles of a skeleton from the single images [31].

The problem of inferring 3d joints out of their 2d projections are traceable back to the classic work of Lee and Chen [32]. They showed that, given the lengths of the bone, the problem lies in a binary decision tree, where each split corresponds to two potential joint states with respect to its parent. This binary tree can be pruned on the basis of joint constraints, although it has seldom led to a single solution. Jiang used a broad pose database to address ambiguities based on neighboring queries closest to him [33]. Interestingly, Gupta and others, the one who integrated temporal constraints during the search, recently revisited the concept of using nearest neighbors for refining the outcome of

pose inference [34]. Another approach to assemble information about 3d human pose from datasets is by developing overcomplete bases appropriate for depicting human poses as sparse combinations, trying to lift the pose to a scientifically valid Hilbert space kernel (RHKS)[28], or by generating novel priors from specialized datasets of severe human poses [35]. Pavlakos and others implemented a deep, convolutional neural network built on the stacked hourglass architecture which maps probability distributions in 3d space instead of regressing 2d joint probability heatmaps [36, 37]. Moreno-Noguer trains to predict from 2-to-3dimensional space a matrix (DM) of pair size [38]. Distance matrices are probabilistic to rotation, translation, and reflection, thus multidimensional mapping is complemented with prerequisite human poses [35] to rule out uncertain predictions. A primary motivation underneath Moreno-Noguer's DM regression approach, as well as Pavlakos's volumetric approach, is the idea that 3d key point prediction from 2d detections is exceptionally difficult. Pavlakos, for instance, describe a baseline where a direct 3d joint representation is used alternatively, with results being far less accurate than using volumetric regression.

Bogo and others used a full-body generative volumetric model [39]. Ionescu, Tekin and others considered latent joint representations, obtained through Kernel Dependency Estimation and autoencoders, to implement joint dependence during the 2D-to-3D inference [40]. EDMs have originally been used in similar fields, for example in modal analysis to estimate the shape basis [41], to represent protein structures [42], to locate sensor networks [43] and to resolve kinematic constraints [44]. It is worth pointing out that Geodesic Distance Matrices (GDMs) are preferred to EDMs for 3D shape recognition tasks, as they are invariant to isometric deformations [45]. But GDMs are not ideal for our problem for the same cause, because multiple deformations in yield the same GDM. The form that generates a particular EDM, on the other hand, is special (up to translation, rotation, and reflection), and can be estimated by multidimensional scaling [46].

Zuffi and Black used a 3D mesh model to study and match 3D scans of synthetic specimens [47]. Chen and others recently trained a human 3D pose regressor on images of virtual training made from such a 3D mesh-model [48]. Similarly, Huang and others used synthetic data generated by a game engine to train a human detector for unusual pedestrians [49]. In both cases, it took a stage of domain adaptation to generalize to real pictures. A scene-specific pedestrian detector was trained without real data by Hattori and others [50], while Enzweiler and Gavrila synthesized virtual samples with a generative model to improve the efficiency of a discriminative model in classification[51]. In [52], 2D character images were animated with the fitting and deformation of a 3D mesh model. Later, Pishchulin and others increased the labeling of training images with similarly small disturbances [53]. Those methods involve complete human segmentation in the photos. Through applying geometric transformations to the first frame of a video series [54], conceptual poses for tracing were synthesized.

Activity acknowledgment and human posture estimation are related to each other but still treated in a different respect. Therefore, to resolve this issue, the author [55] developed a multi-task framework and design a CNN-regression approach that estimates the poses by identifying the key joints as well as do activity classification.

The author proposes an existing framework [56] on video-based pose estimation and identifying real videos having numerous person involved. The paper presents a system that gathers, comment on and discharges another dataset that mainly focuses on recordings with numerous individuals named with an individual for tracking and verbalized human pose. The paper presents huge scale benchmark or video-based human posture estimation.

IV. METHODOLOGY

Identification and distinguishing the position and direction of an article mean and mainly focusing on key point areas that exactly depicts the item is the task for estimation of posture. This research work contains the strategy adopted for distinguishing the significant human body joints restricted in the body using the agile technologies and another scale benchmark dataset. Figure 2 displays the flowchart depicting the proposed system.

A. Dataset

MPII human posture dataset [57] has been 4inimize containing 25k pictures and 40k around individuals showing 410 exercises additionally characterized in classifications. YouTube video pictures were extricated and furnished with going before and following unexplained outlines. Additionally, the dataset also contains 20 comment record that depicts the body joints. The design contains a picture of w*h size creating a yield including 2d limitation of body joints incorporated into the comment record. A review of the dataset is shown in figure 3 The annotation file takes the name of the picture, directions of the head rectangle, individual scale with respect to 200px tallness, unpleasant human position in the picture, individual driven body joint explanation: x, y directions of joint, joint ids, joint perceivability, video list, preparing/testing pictures, picture position in video, rectangle shape id of isolated people, movement id and video id, activity and category name. The annotation file contains 16 joints 1 wrist, 1 elbow, 1 shoulder, r shoulder, r elbow, r wrist, head top, upper neck, thorax, pelvis, l ankle,1 knee, 1 hip, 1 pelvic, r ankle, r knee, r hip as shown in table 1.

B. Preprocessing

The annotations had a ton of subtleties not required for this undertaking, so just the required fields were extracted from annotations and were converted into a CSV file. The annotation CSV file had total 36 columns containing 32 fields as X and Y coordinates of 16 body joints, and other columns include image name, activity name, category name and scale. Since the dataset contains absolute image coordinates of images varying with sizes like 1080px and 720p (4inimiz.), it becomes

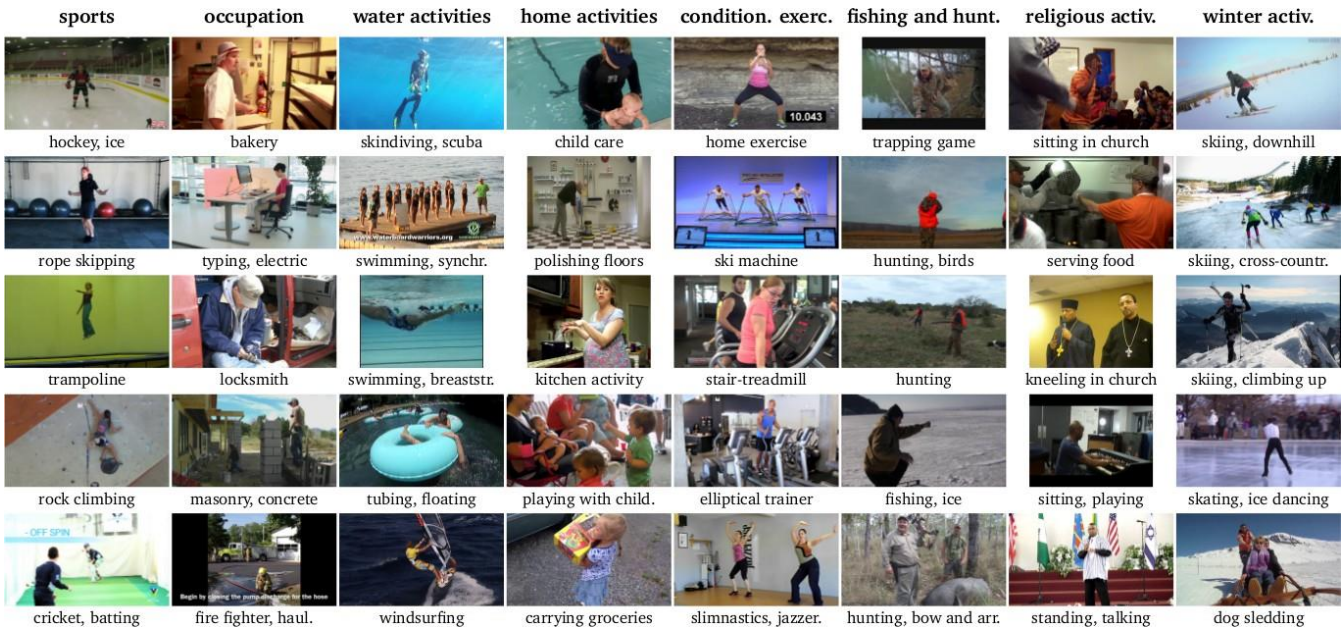

Fig. 3: An overview of MPII human pose dataset images

Joint	:	r	r	l	l	l	pelvis	thorax	upper	head	r	r	r shoul-	l shoul-	l el-	l
	ankle	knee	hip	hip	knee	ankle			neck	top	wrist	elbow	der	der	bow	wrist
id	)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Table 1: Human body joints and their Ids in annotation file

essential to normalize the images and its annotations points for better training and computation. The images were 5inimize5d w.r.t. bounding boxes around a human body and cropping images with respect to the box. The box is defined by three parameters centre coordinates, width w, height h. Bounding box generation can be done manually but for large datasets like MPII authors decided to use pre - trained regression models for bounding box prediction for humans. The YOLO v3 [23] model was used to get bounding boxes. The coordinates of the top left corner of the box were defined by

$$x = int(centerX - (width/2)) \quad (1)$$

$$y = int(centerY - (height/2)) \quad (2)$$

Width and height were denoted as width, height. Thus, for cropping the image w.r.t. bounding box generated by YOLO v3 [23] authors normalized the image as

$$crop\_img = image[y : y + h, x : x + w \quad (3)$$

Cropped image was saved with the help of OpenCV. The images were scaled and normalized to a smaller size image which only contains the useful part of the image but the problem was that the points given in training dataset were according to the full resolution images which contain useful and unnecessary information and background view. So for this task, the coordinates of the annotation were rescaled and normalized w.r.t. newly generated images after cropping. The annotations had some null entries and those rows were deleted in CSV file. The refined dataset then contained 17K images with complete annotations well-structured in the CSV file.

From the images the part containing humans was cropped out, reducing the size of images hence increasing the processing speed using YOLO [58]. The annotations were also accordingly scaled. The cropped images were converted into the RGB image of a common size of 144*144. The final dataset comprised around 15K cropped RGB images of size 144*144 and output is 32 points denoting the 16 body joints. An example of cropped images with joints located on it is shown in figure 4.

C. The Network Architecture

The developed model takes an image of size (144*144 pixels) as input shape and outputs the pixel coordinates of each body key point. The approach used Keras(an open-source neural network library capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML) for developing the model and Tensorflow was used as backend.

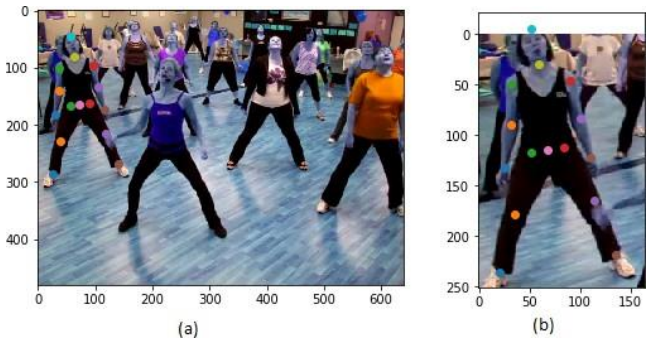

Fig. 4: (a) depicting body joints in the original image (b) depicting joints in the cropped image

The model is sequential containing 13 layers with 3 convolution2d layers, 3 maximum pooling 2d layers, 3 dropout layers, 1 flattening layer and 3 dense layers. The architecture of the model is shown in figure 5.

The input shape of the model is (144,144,3). the activation function used for Conv2D layers is ”relu”. The maxPooling2D layer has stride shape of (2,2). First Dropout layer gad the rate of 0.1, second had it of 0.25 and third of 0.5. The first two dense layers had 500 units with activation function as ”soft- max” and the third layer had 32 units with a linear activation function. A log file was also maintained for easy analysis which was done using ModelCheckpoint also detecting the early stopping.

ModelCheckpoint also detecting the early stopping. The model was compiled with loss type as the mean absolute error using Adam optimizer.

Mean absolute error (MAE) [59] is a measure of the sum of absolute errors. MAE measures, without considering their direction, calculating the linear score. This means that all the individual differences have the same weight as in the test sample, the absolute difference between prediction and true observation.

Considering a scatter plot of n points, where point i has coordinates (xi, yi), the Mean Absolute Error is given by:

$$MAE = 1/n(\sum_{n=1}^{n} |yi|) \quad (4)$$

The data was fitted in this model and with a ratio of 0.2 for validation. The training x data contain the images and training y data contain the CSV file.

V. Results

After training for 30 epochs with the loss function, we ended with losses as 0.0864 in training and 0.0899 in validation. And with Mean Squared Error Loss function, we ended with 0.0864 as training MSE and 0.0899 as validation MSE. These epochs were trained upon 14959 images training on 11967 images and validating upon 2992 images. The dataset is providing a rough scale for training as well as testing images. These results are much better as compared to produced by the images used

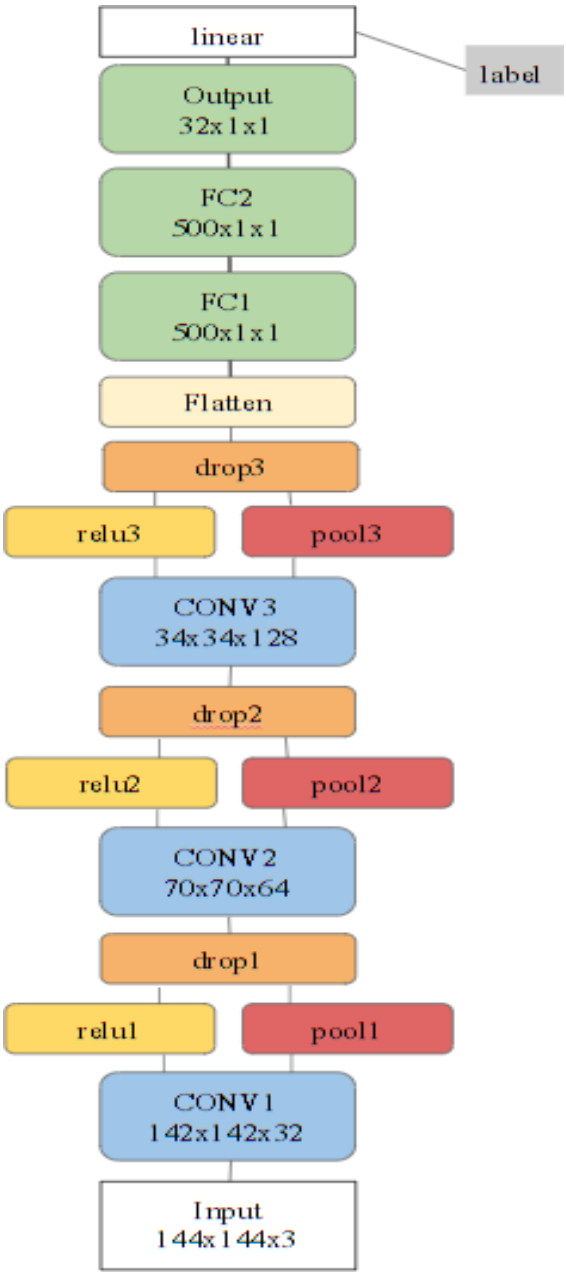

Fig. 5: CNN architecture of Regression for identifying key-points. The final layer outputs a 32-dimensional vector representing the x and y coordinates of each of the 16 key points. Before applying YOLO to crop the images focusing only the human body since the final output is based only upon the same. The graph in figure 6 shows Training and Validation loss as well as mean absolute error showing a near decrease in the graph and then becoming nearly constant after a certain epoch which is after 11 epochs. Both graphs suffers similar behaviour. Figure 7 displays the human joints generated by the model.

VI. Conclusion

CNN, fundamentally utilized for characterization could likewise be utilized for relapse, it has the advantage that it encom-

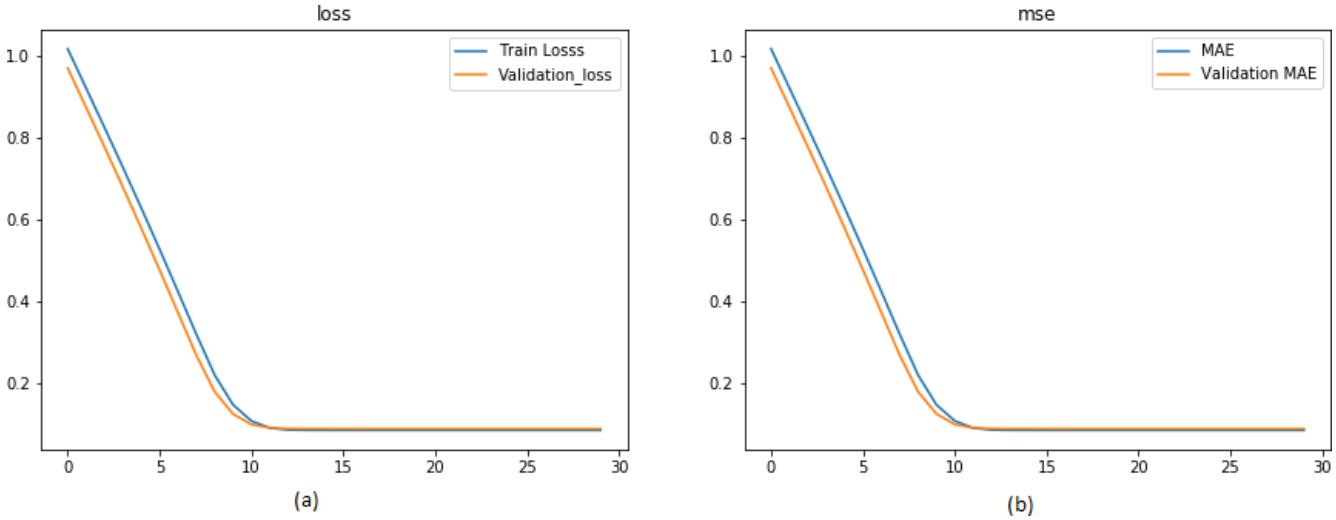

Fig 6: Graphs for training and testing loss on normalized data (a): Training loss and Validation Loss (b): Mean Absolute Error in Training and Validation

Fig. 7: Human Joints generated by the model

pass and accept the whole image as contribution for each key joints rather working on nearby indicators. The proposed work accomplish focused outcomes on testing scholarly dataset. This research work exhibited a less complex CNN model for the relapse issue of human posture estimation. There is as yet an ability to achieve amazingly better results with more processor power and space (the profundity of their backslide CNN was constrained by the RAM of the GPU it was set up on). By modifying the base learning rate and learning rate approach and attempt to unmistakable sorts of energy update and tuning of regularization quality could be engaged upon. To minimise the gap between preparing and approval leads to further adjust the hyperparameters of the creator's illustrate. Further, this result could be utilized to order a picture in different classes or anticipate the action. Various things with a mix of joint estimation and activity portrayal could be attempted to check in the case of knowing the territories of joints in an image improves the climate control system action course of action execution of a Convolutional Neural Network.

REFERENCES

[1] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," *arXiv preprint arXiv:1901.01760*, 2019.

[2] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou, "Good practices on building effective cnn baseline model for person re-identification," in *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, vol. 11069. International Society for Optics and Photonics, 2019, p. 110690I.

[3] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 41–50.

[4] [Online]. Available: https://www.cs.ubc.ca/ lsi-gal/Publications/SigalEncyclopediaCVdraft.pdf

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4733–4742.

[6] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep convnet for multi-label classification with partial labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 647–657.

[7] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[8] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[9] J. Tompson, R. Goroshin, A. Jain, Y.LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.
[10] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.
[11] L. Hansen, M. Siebert, J. Diesel, and M. P. Heinrich, "Fusing informa- tion from multiple 2d depth cameras for 3d human pose estimation inthe operating room," *International journal of computer assisted radiology and surgery*, pp. 1–9, 2019.
[12] S. Liu, Y. Yin, and S. Ostadabbas, "In-bed pose estimation: Deep learning with shallow dataset," *IEEE journal of translational engineering in health and medicine*, vol. 7, pp. 1–12, 2019.
[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
[14] A. Bearman and C. Dong, "Human pose estimation and activity classi- fication using convolutional neural networks," *CS231n Course Project Reports*, 2015.
[15] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet condi- tioned pictorial structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 588–595.
[16] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.
[17] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "3d human pose estimation with 2d marginal heatmaps," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1477– 1485.
[18] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Monocap: Monocular human motion capture using a cnn coupled with a geometric prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 901–914, 2018.
[19] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net++: Multi-person 2d and 3d pose detection in natural images," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
[20] Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. IEEE transactions on pattern analysis and machine intelligence 28(1) (2006) 44–58
[21] Ionescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: 2011 International Conference on Computer Vision, IEEE (2011) 2220–222
[22] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM 56(1) (2013) 116–124
[23] Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision, Springer (2014) 332–347
[24] Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2848–2856
[25] Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3d human pose with deep neural networks. arXiv preprint arXiv:1605.05180 (2016)
[26] Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
[27] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2337–2344, 2014.
[28] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1661–1668, 2014.
[29] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[30] Y. Yu, F. Yonghao, Z. Yilin, and W. Mohan. Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps. In European Conference on Computer Vision (ECCV), 2016.
[31] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In ECCV Workshop on Geometry Meets Deep Learning, 2016.
[32] H. J. Lee and Z. Chen. Determination of 3D human body postures from a single view. Computer Vision, Graphics and Image Processing, 30:148–168, 1985.
[33] H. Jiang. 3d human pose reconstruction using millions of exemplars. In ICPR, pages 1674–1677, Aug 2010.
[34] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham. 3D Pose from Motion for Cross-view Action Recognition via Non-linear Circulant Temporal Encoding. In CVPR, 2014.
[35] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In CVPR, 2015.
[36] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In CVPR, 2017.
[37] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.
[38] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In CVPR, 2017.
[39] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In European Conference on Computer Vision, 2016.
[40] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7):1325–1339, 2014.
[41] A. Agudo, J. M. M. Montiel, B. Calvo, and F. MorenoNoguer. Mode-Shape Interpretation: Re-Thinking Modal Space for Recovering Deformable Shapes. In Winter Conference on Applications of Computer Vision, 2016.
[42] A. Kloczkowski, R. L. Jernigan, Z. Wu, G. Song, L. Yang, A. Kolinski, and P. Pokarowski. Distance Matrix-based Approach to Protein Structure Prediction. Journal of Structural and Functional Genomics, 10(1):67–81, 2009.
[43] P. Biswas, T. Liang, K. Toh, T. Wang, and Y. Ye. Semidefinite Programming Approaches for Sensor Network Localization With Noisy Distance Measurements. IEEE Transactions on Automation Science and Engineering, 3:360–371, 2006.
[44] J. Porta, L. Ros, F. Thomas, and C. Torras. A Branch-andPrune Solver for Distance Constraints. IEEE Transactions on Robotics, 21:176–187, 2005.
[45] D. Smeets, J. Hermans, D. Vandermeulen, and P. Suetens. Isometric Deformation Invariant 3D Shape Recognition. Pattern Recognition, 45(7):2817–2831, 2012.
[46] I. Borg and P. Groenen. Modern Multidimensional Scaling: Theory and Applications. Springer, 2005.
[47] S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In CVPR, 2015.
[48] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3D pose estimation. In 3DV, 2016.
[49] S. Huang and D. Ramanan. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In CVPR, 2017.
[50] H. Hattori, V. N. Boddeti, K. M. Kitani, and T. Kanade. Learning scene-specific pedestrian detectors without real data. In CVPR, 2015.
[51] M. Enzweiler and D. M. Gavrila. A mixed generativediscriminative framework for pedestrian classification. In CVPR, 2008.
[52] A. Hornung, E. Dekkers, and L. Kobbelt. Character animation from 2D pictures and 3D motion data. ACM Trans. Graph., 26(1), 2007.
[53] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In CVPR, 2012.
[54] D. Park and D. Ramanan. Articulated pose estimation with tiny synthetic videos. In CVPR ChaLearn Looking at People Workshop, 2015.
[55] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,

2018, pp. 5137–5146.
[56] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5167–5176.
[57] [Online]. Available: http://human-pose.mpi-inf.mpg.de/
[58] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779– 788.
[59] A. Camero, J. Toutouh, and E. Alba, "A specialized evolutionary strategy using mean absolute error random sampling to design recurrent neural networks," *arXiv preprint arXiv:1909.02425*, 2019.