Machine learning based 2D pose estimation model for human action recognition using geometrical maps

M.V.D. Prasad, K. Durga Bhavani, S. M. Tharun Kumar, P.V.V. Kishore, M. Teja Kiran Kumar, E. Kiran Kumar, D. Anil Kumar

Abstract*: Human action recognition has been a most emerging topic in computer vision because of its several applications like surveillance cameras, human machine interaction and video retrieval. Later human action recognition got better results using machine learning algorithms when compared to computer vision algorithms. In next level pose estimation technique has been introduced and it has drawn more attention for its ability to segment a human body and for detection of joints. Now, in this paper we are developing a human action recognition framework using pose estimation to extract geometrical features and these features are inputted to a sequential convolution neural network for recognizing action performed by the subjects.

***Index Terms*: Human action recognition, machine learning algorithms, pose estimation, sequential convolution neural network.**

I. INTRODUCTION

Human pose estimation has gained exceptional progress from the fast development of varied deep CNN models. This is regularly because of deep neural systems are solid at approximating convoluted and non-straight mapping capacities from impulsive individual pictures to the joint areas even at the nearness of at freedom material body look, viewing conditions and background noises. In this work, we concentrated on the issue of human poses captured in complex backgrounds, which involves following and the assessing the pose of every human in time instance. The Challenges here are involved, including present changes, impediments and the nearness of various covering occurrences. Latest video present estimation strategies use hand planned graphical models or whole number program advancements over casing based key point expectations to compute the final predictions over time. While such methodologies have demonstrated great execution, they require hand-coding of improvement imperatives and may not be adaptable past short video cuts because of their computational complexity. In particular, the following enhancement is in charge of connecting outline level forecasts, and the system has no instrument to improve the estimation of key-points by utilizing temporary data. This infers that if a key-point is poorly detected in a given frame, e.g., due to partial identification or movement obscure, the forecast can't be improved regardless of corresponded, perhaps less equivocal, data being close by in neighboring casings. To address this restriction, we propose a basic and compelling methodology which use the present cutting edge strategy in posture forecast [2] and broadens it by incorporating fleeting data from neighboring video outlines by means of a novel 3D CNN architecture. Comparable the same number of vision errands, the advancement on human posture estimation issue is essentially best in class by profound learning. Since the pioneer work in [5], the execution on the MPII benchmark [4] has turned out to be soaked in three years, beginning from about 80% PCKH@0.5 [6] to over 90% [3]. The advancement on the later and testing COCO human posture benchmark [4] is much quicker. The mAP metric is expanded from 60.5 (COCO 2016 Challenge victor [5] to 72.1(COCO 2017 Challenge champ [3] in one year. With the brisk development of posture estimation, an additionally testing undertaking of "synchronous posture discovery and following in the wild" has been presented as of late [5].

II. RELATED WORK

Human Pose Estimation The past five years have collected a huge progress of human pose estimation in the deep learning region [5]. Despite the clear performance increases, these prior works focus only on improving the pose estimation accuracy by using complex and computationally expensive models whilst largely ignoring the model inference cost issue. This significantly restricts their scalability and deploy ability in real world applications particularly with very limited computing budgets available. In the literature, there are a few recent works designed to improve model efficiency. For example, Bulat and Tzimiropoulos built parameter binaries CNN models to accommodate resource-limited platforms [3]. But this method leads to dramatic performance drop therefore not satisfied for reliable utilization. In most cases, high accuracy rates are required. Rafi et al. exploited good general-purpose practices to improve model efficiency without presenting a novel algorithm [6]. Further, this method does not provide quantitative evaluation on the trade-off between model efficiency and effectiveness. Rather than these past techniques, we deliberately examine the posture estimation proficiency issue under the state of protecting the model execution rate so the came about model is progressively usable and dependable in genuine application situations.

Revised Manuscript Received on December 22, 2018.

First Author name, His Department Name, University/ College/ Organization Name, City Name, Country Name.
Second Author name, His Department Name, University/ College/ Organization Name, City Name, Country Name.
Third Author name, His Department Name, University/ College/ Organization Name, City Name, Country Name.

Learning Distillation The target of information refining is worried about data exchange between various neural systems with particular limits [6]. For example, Hinton et al. effectively utilized a well-prepared expansive system to help train a little system [4]. The method of reasoning is a misuse of additional supervision from an instructor model, spoke to in type of class probabilities [2], highlight portrayals [3], or a between layer stream [3]. This guideline has likewise been as of late connected to quicken the model preparing procedure of substantial scale disseminated neural systems [4], to exchange learning between different layers [2] or between numerous preparation states [6]. Past the traditional two phase preparing based disconnected refining, one phase online information refining has been endeavored with included benefits of progressively proficient improvement [3] and increasingly successful learning [6]. In addition, information refining has been misused to distil simple to-prepare extensive systems into harder-to-prepare little systems [2].

While these past works above exchange classification level discriminative learning, our strategy exchanges more extravagant organized data of thick joint certainty maps. An increasingly comparable work is the most recent radio signs based posture model that likewise receives the possibility of information refining [3]. In any case, this technique focuses at utilizing remote sensors to handle the impediment issue, as opposed to the model proficiency issue as we consider here.

Along these lines, to take care of previously mentioned issues we are actualizing another structure utilizing Geometric component based human activity acknowledgment (HAR) utilizing pre-trained present estimation models. Pre-trained present estimation models will give the joint area in a human body by utilizing these areas we determined the joint separations between every single imaginable joint. These determined joints are inputted to a convolution neural system (CNN) for preparing to perceive the activities.

III. METHODOLOGY

A. Pose estimation models:

Pose estimation models are trained using human action videos with the corresponding joint locations (keypoints) as the labels. After successful training of the human action videos using deep learning algorithms pose estimation models gains the ability to locate the joint position in 2D plane $l_j = \left(x_j, y_j\right) \in R^{2\times J} \ \forall\, j = 1\ to\ \mathrm{J}$, where J is the number of joints.

Fig 1. Joint location obtained for subject 1 and subject 2

B. Joint locations to geometrical features:

Joint locations are now used to create a geometrical maps(GM's) by calculating the joint features between obtained joints from pose estimation models. The joint distances (JD) of a human action video consisting of 'T' frames and 'j' joints is calculated by using Euler's distance formula $JD_j^t = \left\| l_j^t - l_{j+1}^t \right\| \ \forall\, j = 1\ to\ J,\ t = 1\ to\ T$ and these JD's with 'J' joints of size $\left(\frac{J\times(J-1)}{2}\right)\times T$ are obtained i.e., MPI have 15 joint locations obtained from pose estimation models and these having a size of $\left(\frac{15\times(15-1)}{2}\right)\times T = (105\times T)$ for T frames. The obtained matrix is having a size 105 rows and T columns. 105 is the number of all possible distances are calculated using 15 joints.

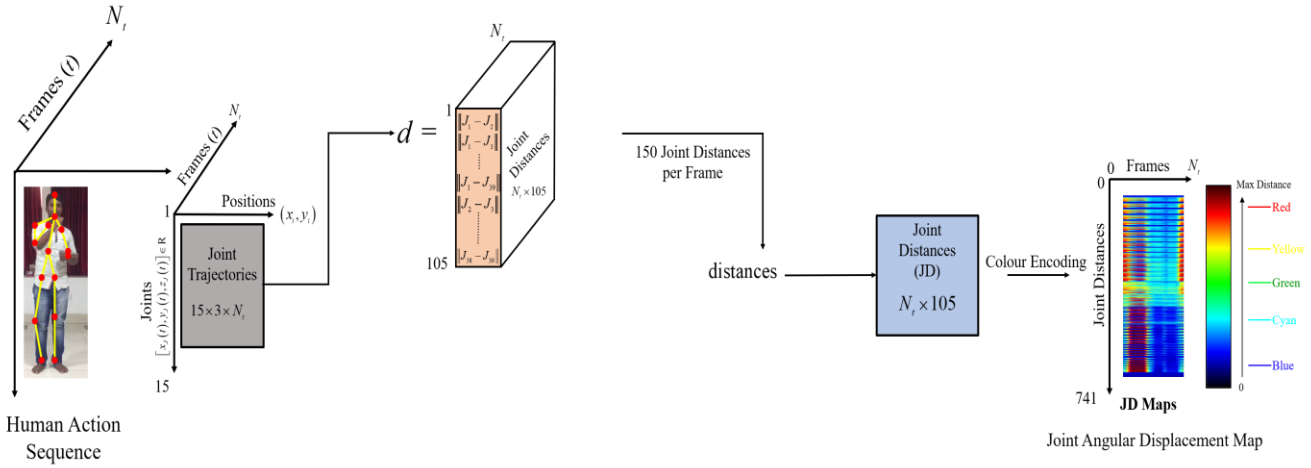

Fig 2. Geometrical feature namely joint distances creation and procedure

C. Geometrical features to CNN:

Fig 3. Shows the CNN architecture here geometrical feature maps are inputted for recognizing human actions and CNN architecture requires very few layers and it requires very less time. Weights are initialized randomly using

gaussian distribution function.

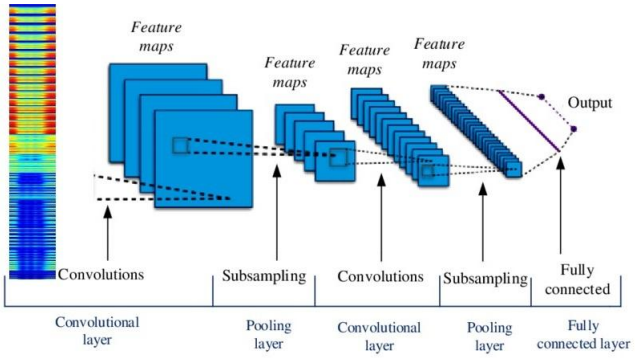

Fig 3. CNN architecture trained with geometrical features

The whole program for GM+CNN has done in Python 3.7 by using libraries Keras and Tensorflow on a HPC (High Performance Computer) available in our university. HPC consists of two NVIDIA Tesla K20 graphics card with 6 nodes and Training takes almost 28 hours. We used stochastic gradient descendant (SGD) as optimizer and its expression is described in eqn.1 and categorical cross entropy (CCE) is a loss function in SGD the expression is shown in eqn.2.

$$\alpha_{sgd} = \alpha_{sgd} - \eta.\nabla_{\theta} J(\alpha_{sgd}, x, y) \quad (1)$$

Where α_{sgd} is the parameter of model vector, η is the learning rate and $\nabla_{\theta} J(\alpha_{sgd}, x_i, y_i)$ is the gradient parameter of x input data to y output class.

$$CCE = -\frac{1}{N}\sum_{j=0}^{M}(y_j.\log(y_j) + (1-y_j).\log(1-y_j)) \quad (2)$$

Where y_j is the predicted label and y_j is the Actual label and CCE finds the loss based on these two weight functions.

IV. RESULTS AND ANALYSIS

In this section, we described about the proposed framework when comparing with the already existing models. Table 1 and table 2 gives the recognition rates of the proposed framework and by comparing with other models our proposed method got better result.

In table 1, we tabulated a recognition rates in different cross subject, cross view and cross scale variations. These variations are inputted to a state-of-the-art method to claim our novelty and we found that our proposed method got better results.

In table-2 we tabulated the overall performance of the proposed method by comparing with the state-of-the-art and our method got better results increasing in recognition rates to 84.10 % when MPI data is inputted to the proposed framework and got 82.07 % when COCO data is inputted to a proposed network.

Table-1. Recognition rates of different methods on publicly available datasets

Datasets		Recognition rates of different methods on publicly available datasets in %				
		Son [1]	Hu Yu [2]	Xue [3]	Hao Chen [4]	GFM+CNN (proposed method)
Cross Subject	MPI [3]	85.81	88.32	88.14	82.19	94.58
	COCO [4]	82.34	89.64	87.25	83.14	88.24
Cross View	MPI [3]	82.24	82.49	84.94	85.24	90.54
	COCO [4]	78.48	80.15	85.19	77.24	87.91
Cross Scale	MPI [3]	76.42	77.51	76.31	77.39	88.35
	COCO [4]	73.24	72.29	73.21	72.28	76.14

Table-2. Overall Recognition rates of different methods on publicly available datasets

Methods	MPI	COCO
Son [1]	78.02	74.93
Yu Hu [2]	80.69	77.25
Xue [3]	81.88	76.04
Hao Chen [4]	77.55	73.87
GFM+CNN (proposed method)	84.1	82.07

Fig 4 and 5 shows the confusion matrix for cross subject and cross view variations. When observing the below confusion matrix, we can see some actions are mismatched i.e., bed and bedsheet actions are almost equal with very slight variations and unable to predict by the proposed method in both the cases.

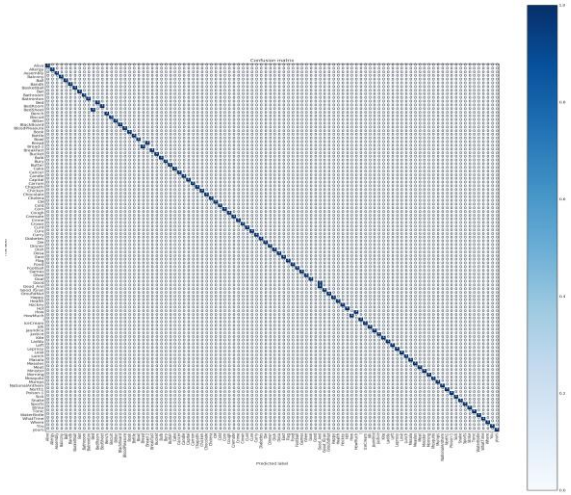

Fig 4. Recognition rates for Cross Subjects

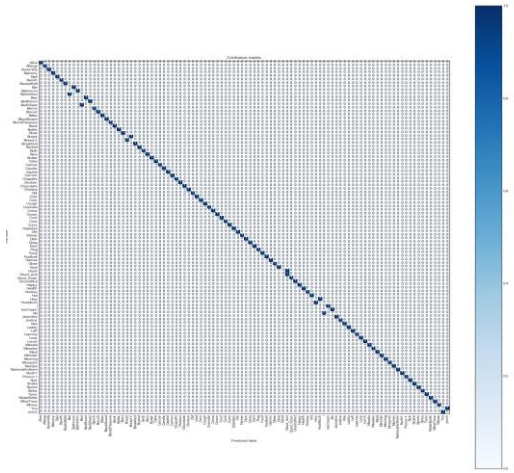

Fig 5. Recognition rates for Cross View

Fig 6 shows the training accuracies and validation accuracies plot with respect to epochs when increase in epochs the accuracy rate of the model is increasing gradually. Fig 7 shows the training loss and validation loss plot with respect to epochs when increase in epochs the loss rate of the model is being decreased, such that the model design and performance parameters are investigated and found that our model is getting better results when compared to state of the art models.

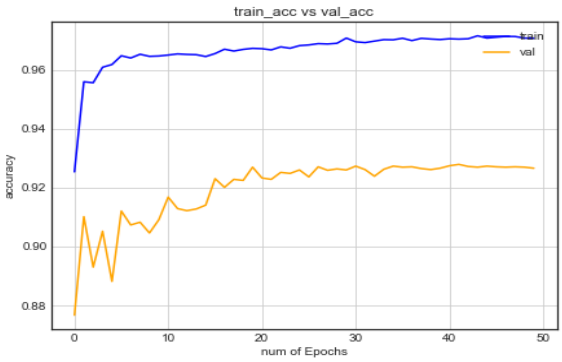

Fig 6. Training and validation accuracies

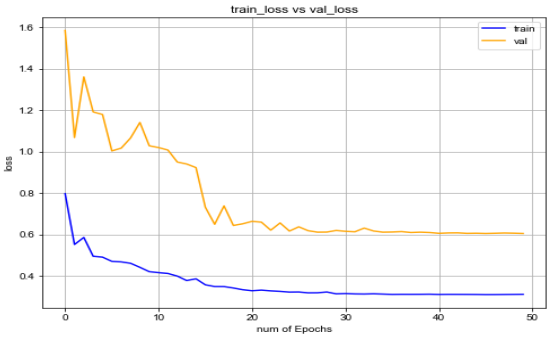

Fig 7. Fig 6. Training and validation losses

V. CONCLUSION

We conclude that our proposed method is performing well as per the experimentation we done using CNN architecture using geometrical maps obtained from the pose estimation models. We found that action are recognized with increased in percentages and time taken for the recognition is also very less. When coming to real time it is taking time for capturing human action and calculating joint location which, we will try to decrease in our future work.

REFERENCES

1. Son, Young-Jun, and Ouk Choi. "Image-based hand pose classification using faster R-CNN." In *2017 17th International Conference on Control, Automation and Systems (ICCAS)*, pp. 1569-1573. IEEE, 2017.
2. Hu, Yu, Yongkang Wong, Wentao Wei, Yu Du, Mohan Kankanhalli, and Weidong Geng. "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition." *PloS one* 13, no. 10 (2018).
3. Xue, Li-wei, Li-guo Chen, Ji-zhu Liu, Yang-jun Wang, Qi Shen, and Hai-bo Huang. "Object recognition and pose estimation base on deep learning." In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1288-1293. IEEE, 2017.
4. Haochen, Li, Zheng Bin, Sun Xiaoyong, and Zhao Yongting. "CNN-Based Model for Pose Detection of Industrial PCB." In *2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pp. 390-393. IEEE, 2017.
5. Guo, Fei, Yifeng He, and Ling Guan. "RGB-D camera pose estimation using deep neural network." In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 408-412. IEEE, 2017
6. Cui, Jiyun, Hao Zhang, Hu Han, Shiguang Shan, and Xilin Chen. "Improving 2D face recognition via discriminative face depth estimation." In *2018 International Conference on Biometrics (ICB)*, pp. 140-147. IEEE, 2018.
7. Chang, Feng-Ju, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. "FacePoseNet: Making a case for landmark-free face alignment." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1599-1608. 2017
8. Mehta, Dushyant, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. "Monocular 3d human pose estimation in the wild using improved cnn supervision." In *2017 International Conference on 3D Vision (3DV)*, pp. 506-516. IEEE, 2017.
9. Ge, Liuhao, Hui Liang, Junsong Yuan, and Daniel Thalmann. "Robust 3D hand pose estimation from single depth images using multi-view CNNs." *IEEE Transactions on Image Processing* 27, no. 9 (2018): 4422-4436.
10. Ghezelghieh, Mona Fathollahi, Rangachar Kasturi, and Sudeep Sarkar. "Learning camera viewpoint using CNN to improve 3D body pose estimation." In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 685-693. IEEE, 2016.