Multiview 3D Pose Estimation

Optional Semester Project Hosted by Computer Vision Lab, EPFL.

Author: **Mohamed Saad Eddine El Moutaouakil**

Supervisors:

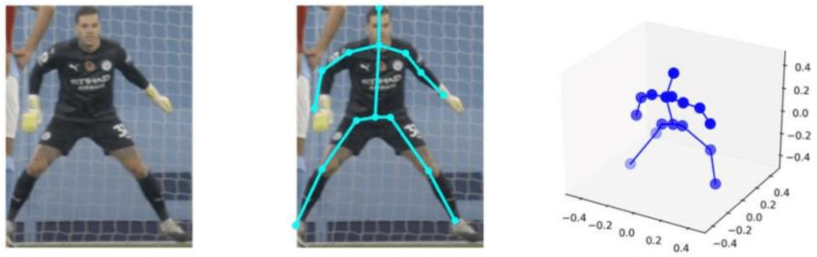
Sena Kiciroglu

Christian Giang

Prof. Pascal Fua

I - Introduction

3D Human Pose estimation is a critical task in computer vision, as it involves determining the position and orientation of a person in an image or video. It has a wide range of applications, such as in autonomous vehicles, analytics, security, and sports. In order to obtain accurate 3D human pose estimates, data acquisition setups typically involve multiple cameras capturing a scene from different angles. This allows for the reconstruction of a 3D model of the human skeleton.



(a) Input Image

(b) 2D Pose Estimation

(c) 3D Pose Estimation

Figure 1 2D and 3D Pose estimation in Sports

Traditional methods for 3D pose estimation from 2D poses rely on algebraic optimization, which is a computationally intensive process that requires proper camera calibration and knowledge of the camera parameters, both intrinsic and extrinsic. Thus, obtaining accurate camera parameters can be labour-intensive and time-consuming.

The goal of this work is to propose and evaluate new architectures capable of learning 3D poses from 2D poses without the prior knowledge of camera parameters. The proposed method aims to overcome the limitations of traditional methods by eliminating the need for labor-intensive camera calibration and the knowledge of camera parameters. This approach aims to make 3D pose estimation more efficient, robust and widely applicable.

II - Previous work

a) Iskakov, Karim et al. "Learnable Triangulation of Human Pose." (2019).

The paper presents two solutions for the task of 3D human pose estimation from 2D images. The first solution, considered as the baseline, is a differentiable algebraic triangulation with an added component of confidence weights estimated from the input images. The second solution is a novel method of volumetric aggregation from 2D feature maps, which is then refined through 3D convolutions. This produces final 3D joint heatmaps and models human pose prior. Both solutions are end-to-end differentiable, which enables direct optimization of the target metric. The authors show the transferability of the solutions across datasets and achieve considerable improvement in multi-view state of the art on the Human3.6M dataset.

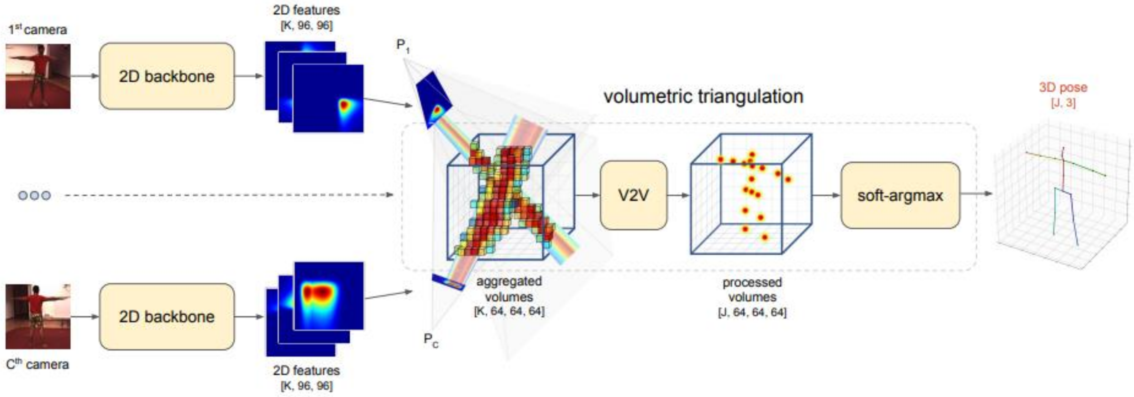


Figure 2 The approach based on volumetric triangulation.

b) **Long Zhao et al. Semantic Graph Convolutional Networks for 3D Human Pose Regression.**

In this paper, the authors examine the challenge of training Graph Convolutional Networks (GCNs) for regression tasks. They identify limitations in current GCN architectures, such as a small receptive field for convolution filters and a shared transformation matrix for each node. To overcome these limitations, the authors propose a novel neural network architecture called Semantic Graph Convolutional Networks (SemGCN) that can handle regression tasks with graph-structured data. SemGCN is capable of learning semantic information, such as local and global relationships between nodes, which is not explicitly represented in the graph. These semantic relationships can be learned through end-to-end training using ground truth data without the need for additional supervision or hand-crafted rules. The authors also investigate the application of SemGCN to 3D human pose regression, as both 2D and 3D human poses can be represented as a structured graph encoding the relationships between joints in the skeleton.

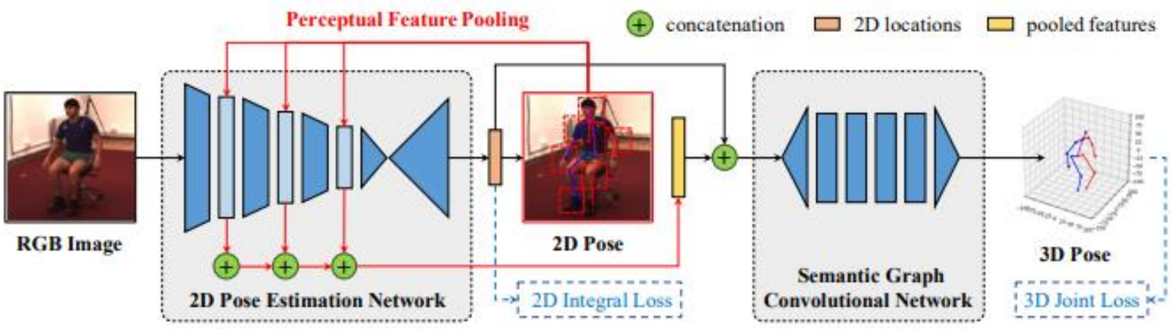


Figure 3 Framework incorporating image features for 3D human pose estimation

III - Dataset

The dataset used was acquired by the computer vision lab at EPFL. It consists of 5 participants each of them performed 4 movements - Squats, Lunges, planks, and pick-ups- in various forms and repetitions. The scene was captured with 4 cameras. The sampling rate was 30 frame per second. The dataset is already processed and contains both 2D and 3D poses ground truths.

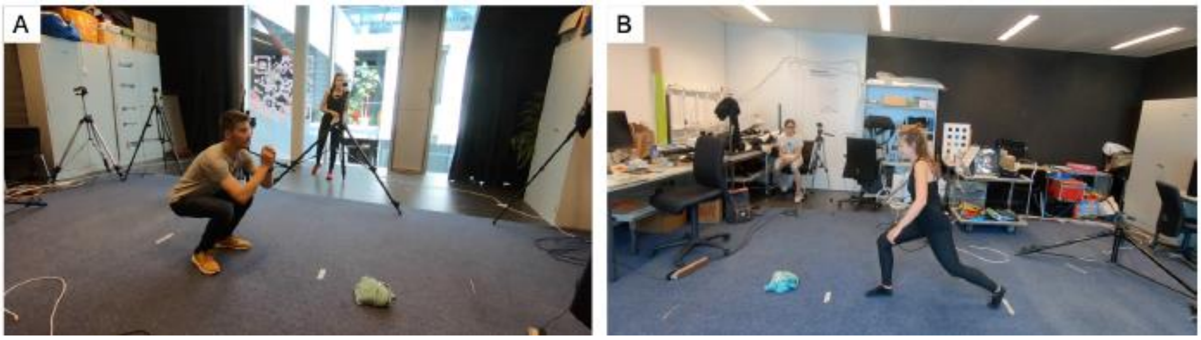


Figure 4 Participants performing exercices

IV - Methods

a) Baseline

As a baseline, the architecture and framework proposed by Zhao shown in figure 5 was adapted to our dataset. New custom dataloaders were introduced to replace the original loaders made for the 3.6M human dataset. The framework was also tweaked to match our dataset.

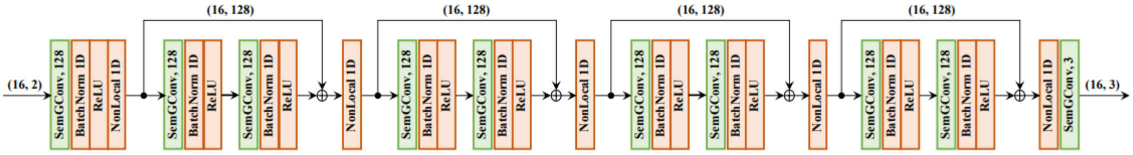


Figure 5 Semantic GCN proposed in Zhao's work

b) Custom method: Aggregations sum and concatenation.

The original architecture taken from Zhao’s work consider each view independently, which misses the opportunity to leverage more data contained in different views of the same frame.

To enable this feature, we proposed an end-to-end trainable multi-input architecture, taking as input the four 2D views of a frame, passing through half of the original architecture. The intermediate outputs of each GCN are aggregated and fed to the remaining half of the architecture producing a single 3D coordinates.

The aggregation used are summation (figure 6), and concatenation (figure 7)

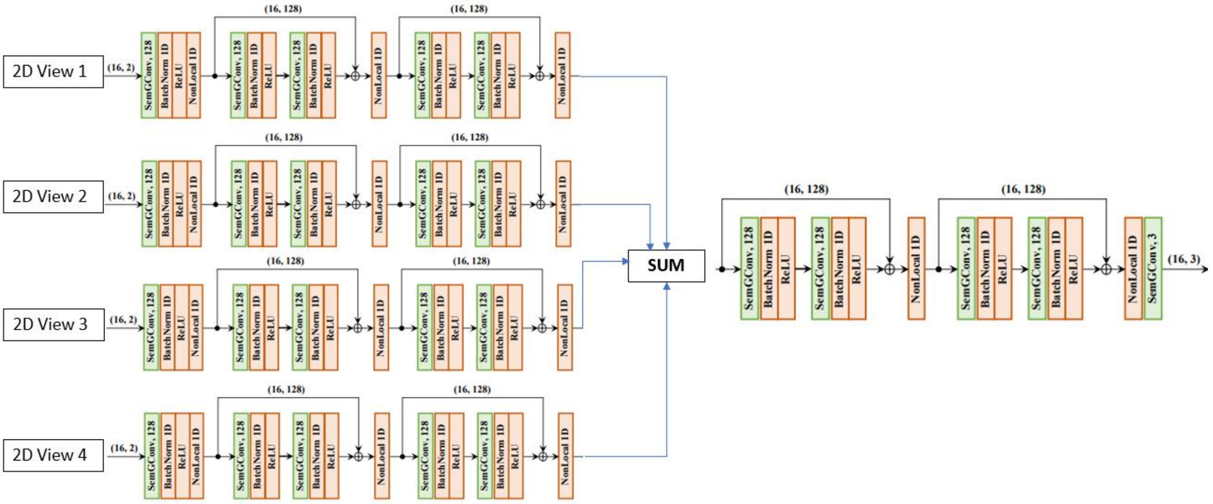


Figure 6 Our custom architecture using sum as a function to aggregate intermediate outputs

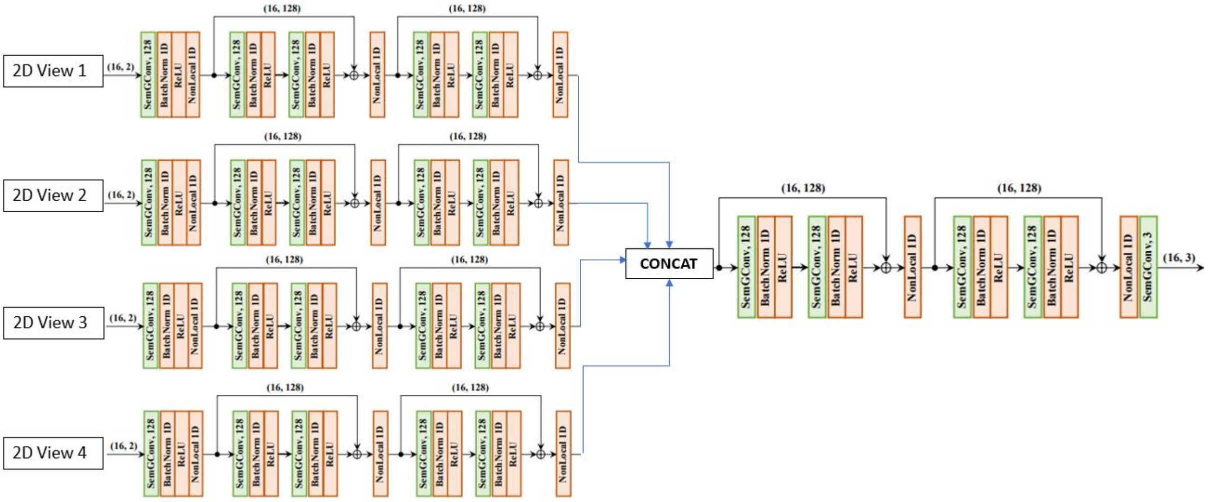


Figure 7 Our custom architecture using concatenation as a function to aggregate intermediate outputs

c) Ablation study: Multilayer Perceptron.

GCNs provide by design a low bias, given that joints are represented as graph that the model learn to represent as closely as possible to a human skeleton. Such a model implies a big number of parameters and thus a longer time to train.

An ablation study was performed to verify whether simple models like multilayer perceptron (MLP) known for having small number of parameters could efficiently do the task of 3D pose estimation.

Keeping the same framework, and the same dataset, the four 2D views of each frame were concatenated, flattened, and fed to a multi-layer perceptron that predicts the 3D view. The number of hidden layers affecting the number of parameters is a hyperparameter that will be optimised during the training and evaluation phase.

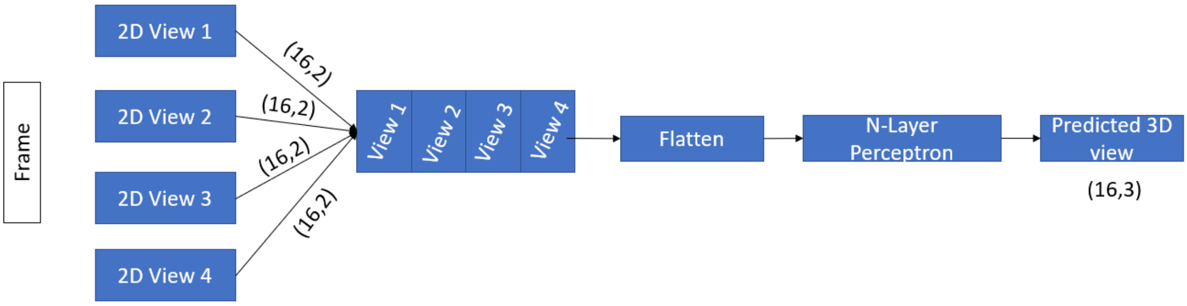


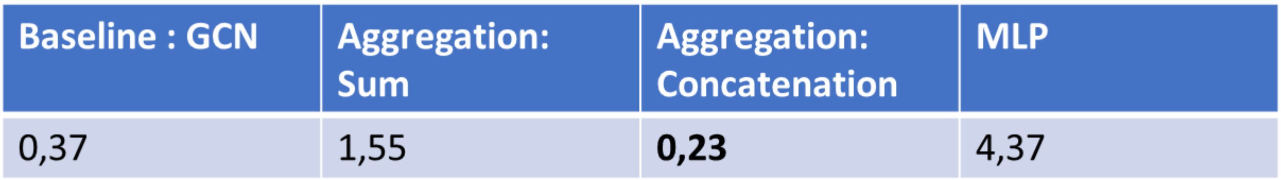
Figure 8 Framework used for multi-layer perceptron

V- Evaluation protocol

The evaluation protocol used in this work is the one that is mostly used in literature. A sample of individuals is used for training while the others serve for validation and testing. In our case, 4 participants performing all the exercises were assigned to training, while one participant was assigned to validation.

VI - Results and discussion

All the models were trained and evaluated with the Mean Squared Error (MSE) loss function. The coordinates used as input are the 2D camera coordinates, and as output the 3D world coordinates.



Baseline : GCN	Aggregation: Sum	Aggregation: Concatenation	MLP
0,37	1,55	**0,23**	4,37

Figure 9 Global MSE Loss on test set (Multiplied by 1000)

Overall, GCNs perform consistently better than MLP. The concatenation of the 4 views of the intermediate outputs presented in our method is the best performing model. It was expected that the use of concatenation would perform better than the sum because the latter don't enable the model to learn the contribution of each view. The baseline and the concatenation method perform very well visually (Figure 10) which indicates the model's ability to learn both the 3D structure of a human body and the camera parameters required to situate it.

MLPs on the other hand remain an interesting choice. Even though they perform many times worse on average than GCN, the visual evaluation shows that they can still learn and represent acceptable 3D human estimations. Figure 10 shows the accuracy of this simple model on chosen representations of Squats, Lunges, planks, and pick-ups. The high error on MLPs on one side, and the good visual results on the other side indicate the MLPs' high variance.

Concerning the time required to train each architecture, we observed that MLPs are 4 times faster to train than GCNs.

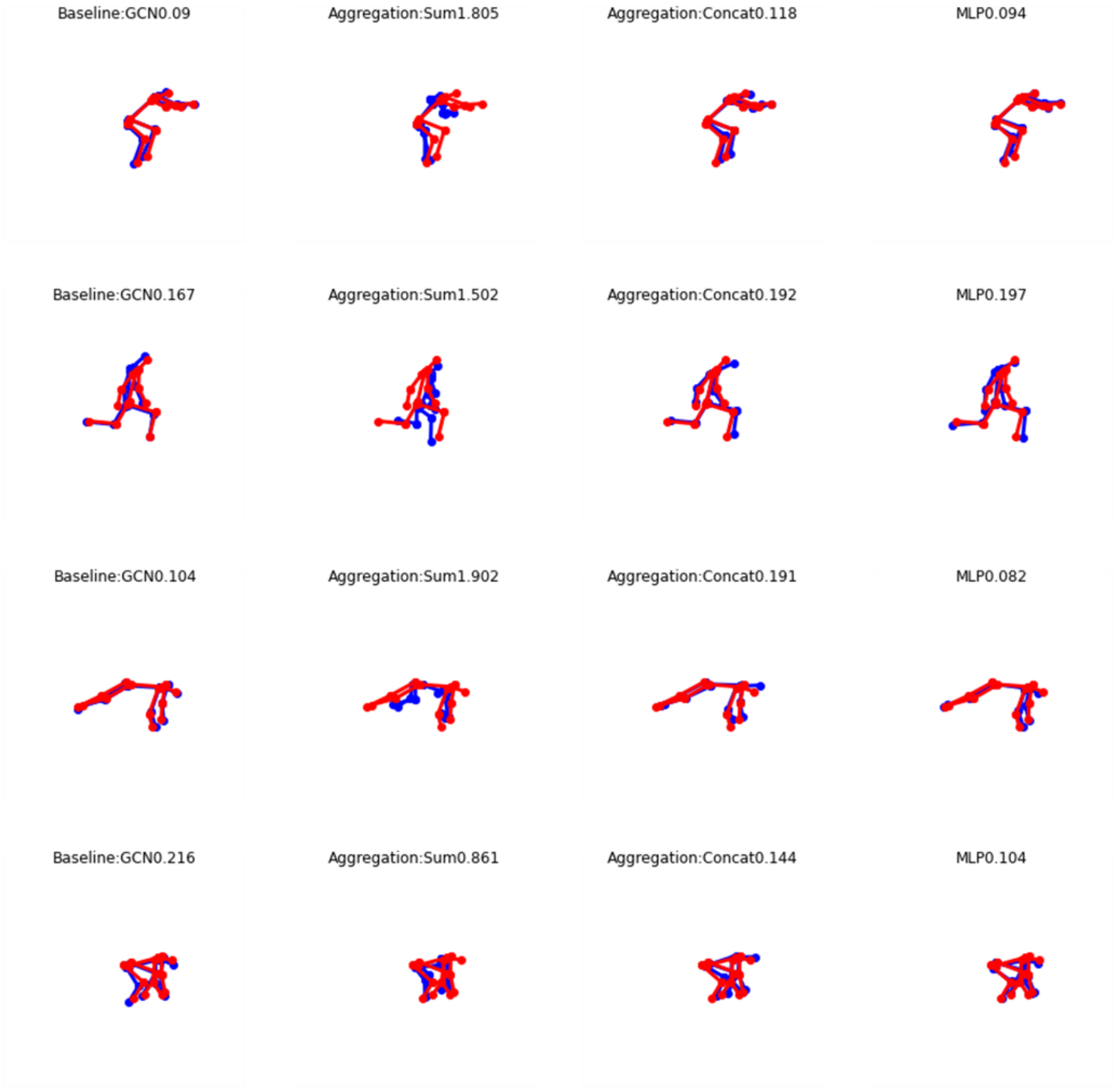


Figure 10 Visual representations of the models' performance with the losses on individual poses.

VII - Conclusion

In this work we tackled the (un)calibration problem in computer vision. It is a crucial step in all computer vision data acquisition setups. Many works addressed this problem through the use of monocular methods, which does not leverage the presence of different views of the same scene to get accurate 3D estimations. This justifies the need of more tailored multiview methods. In this context, we explored both GCNs and MLPs. GCNs perform consistently better than MLPs but require longer time to train given their high number of parameters compared to MLPs.

More effort is to be done to get more accurate and robust methods to perform Multiview 3D pose estimation. By design, our methods are prone to overfit the camera setup which affects the model transferability. A future work could explore the same methods with multiple datasets like the Human 3.6M dataset or try other augmentations like hallucination. Another direction could be to address the problem of high variance of simple models like MLPs.

VIII - References

Iskakov, Karim et al. "Learnable Triangulation of Human Pose." (2019).

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

Hugues Vinzant. “3D Pose Based Motion Correction for Physical Exercises” [Master Project in Life Science engineering, Computer Vision Lab, EPFL].