# Human Pose Estimation through a Novel Multi-view Scheme

Jorge L. Charco[1,2], Angel D. Sappa[1,3] and Boris X. Vintimilla[1]
[1]*ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador*
[2]*Universidad de Guayaquil, Delta and Kennedy Av., P.B. EC090514, Guayaquil, Ecuador*
[3]*Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra, Barcelona, Spain*
{*jlcharco, asappa, boris.vintimilla*}*@espol.edu.ec*

Keywords: Multi-view Scheme, Human Pose Estimation, Relative Camera Pose, Monocular Approach.

Abstract: This paper presents a multi-view scheme to tackle the challenging problem of the self-occlusion in human pose estimation problem. The proposed approach first obtains the human body joints of a set of images, which are captured from different views at the same time. Then, it enhances the obtained joints by using a multi-view scheme. Basically, the joints from a given view are used to enhance poorly estimated joints from another view, especially intended to tackle the self occlusions cases. A network architecture initially proposed for the monocular case is adapted to be used in the proposed multi-view scheme. Experimental results and comparisons with the state-of-the-art approaches on Human3.6m dataset are presented showing improvements in the accuracy of body joints estimations.

## 1 INTRODUCTION

The 2D Human Pose Estimation (HPE) problem is generally tackled by first detecting human body joints (e.g., wrist, shoulder, knee, etc.) and then connecting them to build the human body stick figure. Different solutions have been proposed in the literature (e.g. OpenPose, DeepPose, Stacked Hourglass Networks) and robust solutions obtained when all body joints are detected. However, this problem becomes a challenging one when joints are occluded (e.g., due to self occlusions, which is something common in monocular vision system scenarios). Applications such as human action recognition, augmented reality, healthcare, just to mention a few, have taken advantage of the accuracy of 2D human pose to develop on top of them different solutions. In recent years, convolutional neural networks (CNN) have become a *de facto* tool to tackle most of computer vision tasks; for instance it has been used in image enhancement, object detection, camera pose estimation, just to mention a few, getting better results with respect to classical approaches (e.g.,(Tian et al., 2019), (Wu et al., 2020), (Charco et al., 2018)). Also in the human pose estimation problem we can find different CNN architectures to solve it in a deep learning based framework showing appealing results (e.g., (Wei et al., 2016), (Newell et al., 2016), (Fang et al., 2017), (Cao et al., 2019), (Sun et al., 2019)).

The proposed approaches have used as input a set of images with single or multiple-person to feed the architectures, typically from single-view. Regarding this latter point, multiple-person pose estimation, the number of people in the image increase the computational cost, and hence, also the inference time in real-time. In order to tackle these problems, two approaches have been introduced. The first, known as top-down, localizes the persons in the image and estimate the body joints. The second, referred to as bottom-up, estimates the human body parts in the image and then compute the pose. Despite appealing results obtained on HPE from single-view, the challenge lies in the occlusions of the human body joints in complex poses, causing self-occlusions of certain parts of the human body, in spite of the fact that also the scene could contain multiples moving objects (i.e., bicycles, cars), leading partial occlusion of the human body. In order to overcome this problem, multi-view approaches could be considered; in these cases the human body is captured at the same time from different positions by different cameras. Hence, joints self-occluded in one view can be observed without occlusion by some other camera from other point of view.

The multi-view framework has been already explored to tackle the region occlusion problem in tasks such as 3D-reconstruction, camera pose, autonomous



Charco, J., Sappa, A. and Vintimilla, B.
Human Pose Estimation through a Novel Multi-view Scheme.
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 855-862
ISBN: 978-989-758-555-5; ISSN: 2184-4321

driving, object detection. (e.g., (Xie et al., 2019), (Sarmadi et al., 2019), (Charco et al., 2021), (Hofbauer et al., 2020), (Tang et al., 2018)). For the 2D human pose estimation problem by using a multi-view approach, few works have been proposed. The authors in (Qiu et al., 2019) have proposed a CNN architecture to fuse all features on epipolar line of the images across of all different views. In (He et al., 2020), the authors have proposed to leverage the usage of the intermediate layer to find its corresponding point in a neighboring view, an then combine the features of both views.

On the contrary to previous approaches, where complex deep learning based architectures feed with images from different cameras acquired at the same time, in the current work a compact architecture, originally proposed for monocular scenarios, is adapted to the multi-view scenario. Actually, the multi-view scenario is considered just during the training stage. Images of the same scene, simultaneously acquired by cameras at different point of views, are acquired and used for the CNN training. The proposed architecture uses a variant of ResNet-152 with learning weights as backbone that was proposed in (Iskakov et al., 2019). The multi-view adapted backbone proposed in the current work is trained considering the set of images acquired as mentioned above. It allows to tackle complex poses and overcome the self-occlusion problem, improving the accuracy of estimated joints, being the basis to solve other related problems, such as 3D human pose estimation.

The remainder of the paper is organized as follows. In Section 2 previous works are summarized; then, in Section 3 the proposed approach is detailed together with a description of the scheme multi-view. Experimental results are summarized in Section 4 together with comparisons with state-of-the-art approachs. Finally, conclusions and future work are given in Section 5.

## 2 RELATED WORK

Vision-based human pose estimation is a challenging problem due to the complexity to extract features from images; this complexity is due to different lighting conditions, complex poses, occlusions, among others. On this basis, CNN models have been used for this purpose due to the capability of analysis of images to extract key features of the human body (e.g., joints) improving state-of-art results. Some works have been proposed for 2D-human pose estimation from a single-view scenario. The authors in (Toshev and Szegedy, 2014) have proposed a Deep Neural Network (DNN) as a regress to get the $(x, y)$ image coordinates of human body joints. Additionally, they propose to use a scheme of a cascade of DNN to increase the precision of estimated coordinates by using higher resolution sub-images for refining the predicted joints. In (Tompson et al., 2015) the authors have proposed a multi-resolution ConvNet architecture to implement a sliding window detector with overlapping contexts to generate heatmaps for each joint. The architecture is fed with images, which are running through multiple resolution banks in parallel, and thus capturing important features at a variety of scales. The proposal is trained by minimizing the Mean Squared-Error (MSE) distance of the predicted heatmap to a targeted heatmap. Similarly to the previous works, the authors in (Carreira et al., 2016) have proposed a convolutional network that takes advantage of hierarchical feature extractor, which introduces a top-down feedback of both input and output spaces. The proposal estimates the current human pose, and the joints with wrong predictions are iteratively improved by feeding back error predictions instead of trying to directly predict the target outputs. The authors in (Newell et al., 2016) have proposed a model that consists of steps of pooling and upsampling layers, which are stacked together. The proposed model extracts features at every scale to capture global and local information of the images. Skip connections are used to preserve spatial information at each resolution.

On the contrary to the previous approaches, in (Xiao et al., 2018) an architecture have been proposed, which consists of a variant of ResNet that includes a few deconvolutional layers at the end. These simple changes preserve better the information for each resolution than one with skip connections. Similarly to the previous approaches, MSE is used as the loss between the predicted heatmaps and the targeted heatmaps. The authors in (Sun et al., 2019) have proposed a novel architecture, which is able to maintain a high-resolution representation through the whole process, i.e., it starts from a high-resolution subnetwork as the first stage, and gradually add high-to-low resolution subnetworks to form more stages that are connected in parallel, instead of recovering the resolution through a low-to-high process.

Just few works have been proposed to solve the human pose estimation problem leveraging any geometry information of the cameras to improve the 2D detector. In (Qiu et al., 2019), the authors have introduced a cross-view fusion scheme into CNN to jointly estimate 2D poses from multiple views. The initial pose heatmaps are generated for each image into a multi-view scheme, the corresponding features

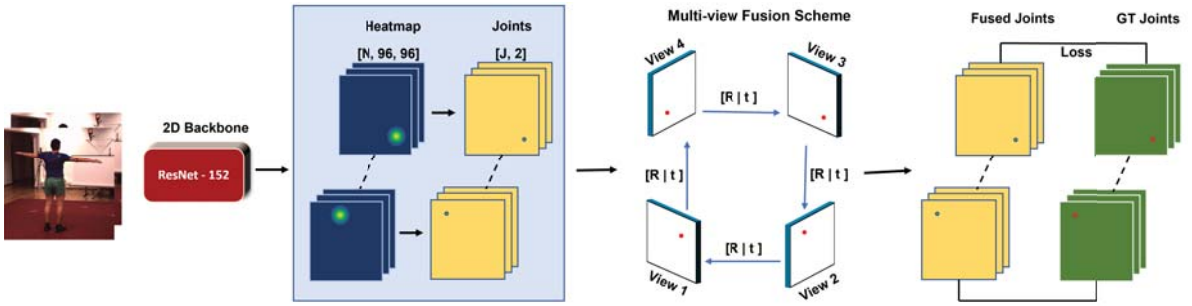


Figure 1: CNN backbone feeds with a set of pairs of images of the same scene simultaneously acquired from different points of view. The multi-view fusion scheme allow to estimate occluded joints with information from other views across of the relative camera pose.

between them are found on the epipolar line, and then they are fused across different views. Similarly to the previous work, the authors in (He et al., 2020) have proposed an architecture to leverage 3D-aware feature in the intermediate layers of the 2D detector, and not only during the final robust triangulation phase. The epipolar transformer is used to augment the intermediate features of a 2D detector for a given view (reference view) with features from neighboring views (source view). The authors in (Remelli et al., 2020) have proposed a novel multi-camera fusion technique, which uses the feature transform layers to map images from multiple-views and exploit 3D geometry information to a common canonical representation by explicitly conditioning them on the camera projection matrix.

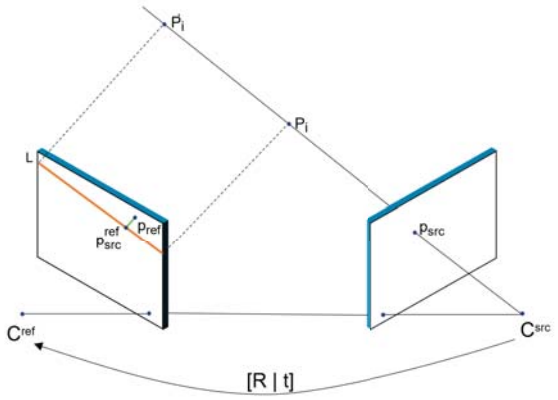


Figure 2: An image point $p_{src}$ back-projects to a ray in 3D defined by the point $p_{src}$ and depth ($Z$). The ray is projected to the image plane of reference view to generate the epipolar line ($L$).

## 3 PROPOSED APPROACH

The proposed approach consists to leverage the multi-view scheme to solve the self-occlusion problems in the 2D human pose estimation. The CNN backbone, proposed by (Iskakov et al., 2019), is used as baseline to be retrained with the proposed multi-view scheme. Basically, this backbone is a variant of Resnet-152 with learnable weights, transposed convolutions, and the number of human body joints as the output channels. For more details see the work mentioned above.

A multi-view system of $C$ calibrated and synchronized cameras with known parameters $R_c$ (i.e., intrinsic and extrinsic parameters), which capture the performance of a single-person in the scene from different views, is used by the proposed model. The images acquired by the multi-view system are denoted as $Im^c$, and organized in pairs of images, which belong to different views, namely, reference view $Im^{ref}$ and source view $Im^{src}$. The output of the backbone is a set of heatmaps for each image. These heatmaps correspond to each human body joint, which are fused across source view considering the confidence of each joint, doing robust the human pose of each view (see Fig. 1). The details of proposed multi-view approach are given below.

### 3.1 Multi-view Scheme

Given a set of pairs of images, the CNN backbone extracts the heatmaps of each joint for each input image separately, which are denoted as $M_\Theta^{ref} = \left\{Im_1^{ref}, ..., Im_i^{ref}\right\}$ and $M_\Theta^{src} = \{Im_1^{src}, ..., Im_i^{src}\}$, where $i$ is the number of joints and, $ref$ and $src$ correspond to the reference and source view respectively. The heatmaps are used to estimate the 2D positions of each joint for each input image. First, it is computed the softmax across the spatial axes; and then, the 2D positions of the joints ($p_{(x,y)}$) are computed as the center of mass of the corresponding heatmaps, which is defined as:

$$p_{(x,y)} = \sum_{u=1}^{W} \sum_{v=1}^{H} h_{i_{(u,v)}} . (\zeta_\Theta(h_{i_{(u,v)}})), \qquad (1)$$

where $\zeta_\Theta$ represents the function softmax; $h$ represents the ROI of the heatmaps of $i$-th joint and W and H correspond to the size of the heatmap ROI.

The values of the joints in the world coordinate system $P = (X,Y,Z)$ are obtained using each 2D position of each joint of each image, as show in Eq. (2):

$$x_i = f\frac{X}{Z} \qquad y_i = f\frac{Y}{Z}, \tag{2}$$

where $x, y$ are the 2D position of $i$-th joint obtained in Eq. (1), and $f$ corresponds to the focal length of the camera. Since, the depth $(Z)$ of the joint is unknown, then two values are used to solve the Eq. (2). The first, a value of depth near zero that corresponds to the position close of camera in the world coordinate system; and the second, a value of depth near to the size of the space of the scene. Empirically, this value has been set to $10m$ for the experiments.

For each 2D position of the joint, two points in the world coordinate system using the depth estimation mentioned above are computed with Eq. (2), then they are transformed by using the relative camera pose between both views (reference and source view) and projected to the image plane, as shown below:

$$T_{rel} = Rot_{src} \cdot (T_{ref} - T_{src}), \tag{3}$$

$$Rot_{rel} = Q(Rot_{ref}.T)^{-1} * Q(Rot_{src}.T), \tag{4}$$

$$p^{ref}_{src_{(x,y)}} = \Delta 2D_{ref}(Rot_{rel} \cdot (P_i - T_{rel})), \tag{5}$$

where $Q(.)$ represents the quaternion. $Rot \in \mathbb{R}^{3x3}$ and $T \in \mathbb{R}^{3x1}$ represent the matrix rotation and the vector translation respectively. $P_i$ corresponds to coordinates of the $i$-th joint, obtained in Eq. (2), in the world coordinate system. As the depth $(Z)$ of $i$-th joint is unknown, the linear equation on image plane of reference view is calculated using the points obtained in the Eq. (5). By definition, the depth $(Z)$ of the $i$-th joint in the source view should be any 2D-point on the linear equation of reference view. Given that the linear equation has infinite points, and any of them could be the depth of $i$-th joint in the source view, then, the 2D-point on linear equation used as the depth of the joint in the source view is calculated by using the intersection between the 2D-point of the joint calculated in the reference view $p_{ref_{(x,y)}}$ and the linear equation previously obtained.

Since the $i$-th joint has two different 2D positions in the image plane, the first corresponds to the reference view $p_{ref_{(x,y)}}$, and the second corresponds to the source view, which are projected to the reference view $p^{ref}_{src_{(x,y)}}$ by using Eq. (5), the confidence values of $i$-th are obtained (see Fig. 2). It is calculated as the distance between the ground-truth of 2D position of $i$-th joint and the 2D position of $i$-th joint obtained by Eq. (1). In order to improve 2D position of $i$-th joint in the reference view, the confidence values and 2D positions of $i$-th joints ($p_{ref_{(x,y)}}$ , $p^{ref}_{src_{(x,y)}}$) are used, as shown in Eq. (7).

$$\omega = 1 - \left| \frac{D_\Delta(\hat{\gamma}_i, \gamma_i)}{\sum D_\Delta(\hat{\gamma}_i, \gamma_i)} \right|, \tag{6}$$

$$\delta_{upd_{i_{(x,y)}}} = \omega * p_{i_{(x,y)}}, \tag{7}$$

where $(\hat{\gamma}, \gamma)$ represent the ground truth and prediction of 2D position of $i$-th joint respectively, and $\omega$ corresponds to the confidence of the points of $i$-th joint in the reference view, including the points projected from source view.

Note that $\delta_{upd_{i_{(x,y)}}}$ corresponds to the new 2D positions of $i$-th joint, which has been enhanced with the information and confidence of $i$-th joint obtained from the source view. Finally, the loss function used in the proposed approach is defined as:

$$Loss = \sum_{i=1}^{N} \left\| \delta_{upd_{i_{(x,y)}}} - \hat{p}_{i_{(x,y)}} \right\|_2, \tag{8}$$

where $N$ corresponds to the number of joints, and $\hat{p}_{i_{(x,y)}}$ is the ground-truth of $i$-th joint in image plane.

## 3.2 Dataset and Metrics

The experiments are conducted on one large-scale pose estimation public dataset with multi-view synchronized images and evaluated using the JDR(%) metric. This section will breafly describe both of them, dataset and used metrics.

### 3.2.1 Human 3.6m

The *Human*3.6*m* dataset was proposed by (Ionescu et al., 2014), and it is currently one of the largest publicly available human pose estimation benchmark. It can be used with monocular or multi-view setups. Four synchronized and calibrated digital cameras were used to capture 3.6 million frames with a single-person. The motions are performed by 11 professional actors (6 males, 5 female) in different activities such as taking photo, discussion, smoking among other. In the current work, subjects 1, 5, 7, and 8 are used for trained the proposed approach; while subjects 9 and 11 are used just for testing. Images from all the cameras are used during the training and testing process.

Table 1: Comparison of 2D pose estimation accuracy on Human3.6m dataset using JDR(%) as metric. "$-$": these entries were absent. $*$: approach presented in (Qiu et al., 2019). $\Upsilon$ trained again by (He et al., 2020). $\Psi$ approach presented in (He et al., 2020). R50 and R152 are ResNet-50 and ResNet-152 respectively. Scale is the input resolution of the network.

| | Net | scale | shlder | elb | wri | hip | knee | ankle | root | neck | head | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sum epipolar line $^{*}$ | R152 | 320 | 91.36 | 91.23 | 89.63 | 96.19 | 94.14 | 90.38 | - | - | - | - |
| Max epipolar line $^{*}$ | R152 | 320 | 92.67 | 92.45 | 91.57 | 97.69 | 95.01 | 91.88 | - | - | - | - |
| Cross-View fusion $^{*\Upsilon}$ | R50 | 320 | 95.6 | 95.0 | **93.7** | 96.6 | 95.5 | 92.8 | 96.7 | 96.5 | 96.2 | 95.9 |
| Cross-View fusion $^{*\Upsilon}$ | R50 | 256 | 86.1 | 86.5 | 82.4 | 96.7 | 91.5 | 79.0 | **100** | 93.7 | 95.5 | 95.1 |
| Epipolar transformer $^{\Psi}$ | R50 | 256 | 96.44 | 94.16 | 92.16 | 98.95 | **97.26** | 96.62 | 99.89 | 99.68 | 99.63 | 97.01 |
| **Mview-Joints (ours)** | R152 | 384 | **99.65** | **97.31** | **93.70** | **99.22** | 97.24 | **97.45** | 99.83 | **99.82** | **99.75** | **98.22** |

Table 2: Comparison of average median Euclidean distance error between Mview-Joints and Learning triangulation backbone proposed by (Iskakov et al., 2019) on Human3.6m. *Backbone*: Resnet 152 with pretrained weight (Iskakov et al., 2019).

| | Net | shlder | elb | wri | hip | knee | ankle | root | neck | nose | belly | head | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Learning triangulation | Backbone | **7.84** | 8.00 | 7.40 | **7.55** | 7.45 | 9.70 | 5.75 | **5.86** | 6.46 | 6.47 | 6.57 | 7.18 |
| **Mview-Joints (ours)** | Backbone + Multi-view | 7.88 | **6.73** | **7.08** | 7.62 | **6.82** | **9.19** | **5.24** | 6.05 | **5.29** | **6.15** | **3.25** | **6.48** |

#### 3.2.2 Metrics

The metric to be used to evaluate the performance of the obtained results is an important factor. In the human pose estimation problem, the Joint Detection Rate (JDR) is generally used. The JDR measures the percentage of *successfully* detected joints, assuming as a successful detection those joints where the distance between the estimated and the ground truth joint is smaller than a given threshold; in the current work this threshold has been defined as half of the head size, as proposed in [2]. In the current work, in addition to the JDR, the Euclidean distance error for every estimated joint with respect to the corresponding ground truth has been also computed. These Euclidean distance error values help to determine the accuracy of each joint of the estimated human pose.

## 4 EXPERIMENTAL RESULTS

As mentioned above, the multi-view approach is proposed to tackle challenging scenarios where self-occlusions of joints happen resulting in difficult 2D human pose estimation. This section presents details on the experimental results by training the proposed multi-view scheme with Human 3.6m dataset (Ionescu et al., 2014). The proposed approach was implemented with Pytorch and trained with NVIDIA Titan XP GPU and Intel Core I9 3.3GHz CPU. Adam optimizer is used to train the network with a learning rate of $10^{-5}$ and batch size of 32 (i.e., eight human poses simultaneously captured from four different points of view).

### 4.1 Training of Multi-view Scheme

The CNN backbone used in the proposed architecture was initialized with the weights of Resnet-152 pretrained by (Iskakov et al., 2019). The network architecture was trained on Human 3.6m dataset. As pre-processing dataset, the images were cropped according to the bounding box of the person and resized to 384x384 pixels; then, the mean value of intensity of pixels was computed and subtracted from the images. For the training process, a set of 60k images were used to feed to the network, which was trained until 20 epochs; it takes about 120 hours. The preprocessing mentioned above has been also used during the evaluation phase. In the evaluation a set of 8k images have been considered.

### 4.2 Results and Comparisons

Experimental results obtained with the proposed architecture are presented in Table 1, which shows some joints and compare them with state-of-the-art CNN-based methods by using the JDR metric. The proposed approach referred to as Mview-Joints outperforms the previous works on most of body joints. The improvement is most significant for the shoulder, elbow, and ankle joints, which increment from 96.44% to 99.65%, from 95.00% to 97.31% and from 96.62% to 97.45%, respectively. The average JDR of body joints obtained by Mview-Joints improves the results of Epipolar transformer (He et al., 2020) about 1%, and with respect to Cross-View fusion (Qiu et al., 2019) about 3% approximately.

Additionally, median Euclidean distance error is used to evaluate the accuracy of prediction of the pro-


Figure 3: Challenging poses, the multi-view scheme takes advantage from the additional view with respect to the backbone—single view—proposed by (Iskakov et al., 2019).

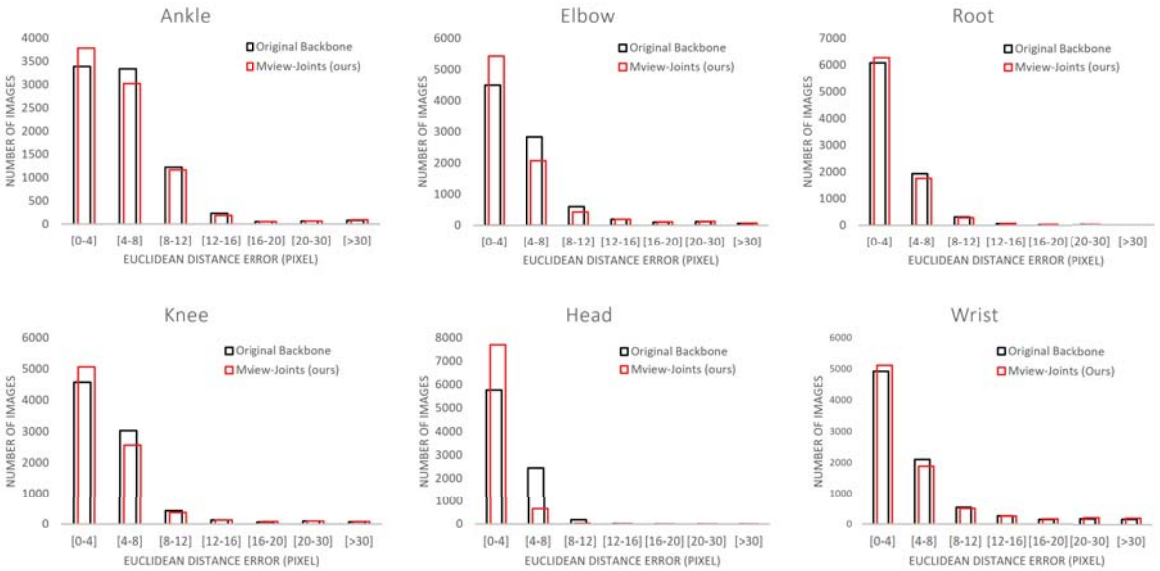


Figure 4: Comparison of Euclidean distance errors between the backbone approach (Iskakov et al., 2019) and the proposed Mview-Joints for six different body joints.

posed multi-view scheme with respect to the CNN backbone proposed by (Iskakov et al., 2019), whose results are shown in Table 2. The body joints that improvement significant the accuracy are the elbow, wrist, knee, nose, head. The median Euclidean distance errors for these joints improve by 15.88%, 4.32%, 8.46%, 18.11% and 50.53% respectively the results obtained with CNN backbone proposed by (Iskakov et al., 2019). Some challenging poses are shown in Fig 3 where the multi-view scheme takes advantage of the different views. The histograms of accuracy of obtained body joints are shown in Fig. 4. Most of the predicted 2D positions of body joints are in the ranges [0-4] pixels and [4-8] pixels by using Euclidean distance error for the proposed approach, compared with the approach presented in (Iskakov et al., 2019).

## 5 CONCLUSIONS

This paper addresses the challenging problem of the human pose estimation when the joints are occluded. A monocular network architecture that takes advantage of multi-view scheme is proposed to accurately estimate the human pose. This scheme is motivated by the reduced information to predict more precisely occluded joints when only one view is used. Experimental results and comparisons are provided showing improvements on the obtained results. The manuscript shows how estimated joints of other views can help to estimate occluded joints more accurately. The obtained precision of body joints is the base to solve others related problems as 3D human pose estimation, action recognition among others. Future work will be focused on extending the usage of multi-view environments to leverage the geometry of the scene, and thus, improve the 3D human pose.

## ACKNOWLEDGEMENTS

This work has been partially supported by the ESPOL projects EPASI (CIDIS-01-2018), TICs4CI (FIEC-16-2018) and PhysicalDistancing (CIDIS-56-2020); and the “CERCA Programme/Generalitat de Catalunya”. The authors acknowledge the support of CYTED Network: “Ibero-American Thematic Network on ICT Applications for Smart Cities” (REF-518RT0559) and the NVIDIA Corporation for the donation of the Titan Xp GPU. The first author has been supported by Ecuador government under a SENESCYT scholarship contract CZ05-000040-2018.

## REFERENCES

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186.

Carreira, J., Agrawal, P., Fragkiadaki, K., and Malik, J. (2016). Human pose estimation with iterative error

feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742.

Charco, J. L., Sappa, A. D., Vintimilla, B. X., and Velesaca, H. O. (2021). Camera pose estimation in multi-view environments: From virtual scenarios to the real world. *Image and Vision Computing*, 110:104182.

Charco, J. L., Vintimilla, B. X., and Sappa, A. D. (2018). Deep learning based camera pose estimation in multi-view environment. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 224–228. IEEE.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343.

He, Y., Yan, R., Fragkiadaki, K., and Yu, S.-I. (2020). Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788.

Hofbauer, M., Kuhn, C. B., Meng, J., Petrovic, G., and Steinbach, E. (2020). Multi-view region of interest prediction for autonomous driving using semi-supervised labeling. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). 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.

Iskakov, K., Burkov, E., Lempitsky, V., and Malkov, Y. (2019). Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727.

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer.

Qiu, H., Wang, C., Wang, J., Wang, N., and Zeng, W. (2019). Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4342–4351.

Remelli, E., Han, S., Honari, S., Fua, P., and Wang, R. (2020). Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6040–6049.

Sarmadi, H., Muñoz-Salinas, R., Berbís, M., and Medina-Carnicer, R. (2019). Simultaneous multi-view camera pose estimation and object tracking with squared planar markers. *IEEE Access*, 7:22927–22940.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703.

Tang, C., Ling, Y., Yang, X., Jin, W., and Zheng, C. (2018). Multi-view object detection based on deep learning. *Applied Sciences*, 8(9):1423.

Tian, C., Xu, Y., Fei, L., Wang, J., Wen, J., and Luo, N. (2019). Enhanced cnn for image denoising. *CAAI Transactions on Intelligence Technology*, 4(1):17–23.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656.

Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732.

Wu, M., Yue, H., Wang, J., Huang, Y., Liu, M., Jiang, Y., Ke, C., and Zeng, C. (2020). Object detection based on rgc mask r-cnn. *IET Image Processing*, 14(8):1502–1508.

Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481.

Xie, H., Yao, H., Sun, X., Zhou, S., and Zhang, S. (2019). Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2690–2698.