International Journal of Innovation Engineering and Science Research
(ISSN: 2581-4591)

www.ijiesr.com

# A Comparative Study of 2D Human Pose Estimation Methods

**[1]Daniela Hagiescu, [2]Felix Pirvan**
*Advanced Slisys, Bucharest, Romania*

**[3]Lidia Dobrescu**
*[3]Faculty of Electronics, Telecommunication and Information Technology, University POLITEHNICA of Bucharest, Bucharest, Romania*

## ABSTRACT

Human pose estimation is the process of detecting the key points or landmarks of the human body. Face pose estimation and hand pose estimation are the two most common special cases. In this paper, we will focus on the body as a whole. Human pose estimation is used in various subsequent tasks, such as action recognition or motion characterization. We are presenting the main types of approaching the problem and the different techniques and architectures involved. We are also discussing the evaluation datasets available and their features. The field has seen fabulous progress in later years since the advent of deep learning.

***Keywords***—*human pose, pose estimation, keypoint, landmark, human body, deep learning*

## I. INTRODUCTION

Human pose estimation (HPE) involves detecting the positions of the human body parts, given as input some sensor data. Input data often comes from one or more cameras. A single camera can be RGB or RGB-D (depth) camera, while stereo camera systems can also be used. Sometimes video sequences are available, and this helps by imposing additional consistency constraints on the HPE from each single image. Applications of HPE range from movies to healthcare, from virtual reality to autonomous cars and from surveillance to robotics.

Some HPE methods use a body model that makes connections between the skeleton joints, based on prior knowledge about the human body structure. The most commonly used body model is skeleton-based. It describes the connections between different joints of the skeleton. The HPE task comes down to estimating the 2D or 3D coordinates of the skeleton joints.Other methods do not use joint connections, hence they are faster, but can be prone to errors when encoutering an unseen pose. Some HPE methods use a top-down approach, where they first detect all the persons in the image and then estimate the pose of each person. Other methods work bottom-up, first detecting all the joints and then grouping them by persons. The more persons in the image, the more time needed for top-down methods, while the bottom-up methods maintain a constant time. However, when persons overlap, bottom-up methods encounter difficulties matching the joints to the right person. Some methods use other kind of body parts in addition to the joints, like the limbs. Some HPE methods directly regress the coordinates of the joints, while others try to detect the image patches around the joint or use heat maps instead of point locations. Single-stage HPE methods are more compact are easier to train end-to-end, while multi-stage methods can offer more flexibility and make it easier to pinpoint the issues that occur in one stage or another.

In our attempt to summarize the present state of the field, we build upon previous work. In [1], the authors focus on single-camera (monocular) HPE and discuss separately the 2D and 3D methods, present the datasets and the metrics used for evaluation. Even more recently, [2] presents several body models and then many 2D and 3D methods grouped by approach (top-down or bottom-up), including a performance comparison. In this paper, our approach is to select and focus on the most widely used datasets and the most successful methods and their techniques. The field is already vast enough to get lost, so our goal is to bring worward the most promising results.

## II. DATASETS

In order to train and evaluate different HPE methods, several datasets have been made publicly available. They all maintain a competition ladder with the best scoring methods. Those datasets have put together a set of images or video sequences containing the full body of one or more persons. They

defined a set of keypoints and annotated each image with the positions of all defined joints. As you can see in Figure 1, the set of keypoints differs from one dataset to another, making it difficult to use all of

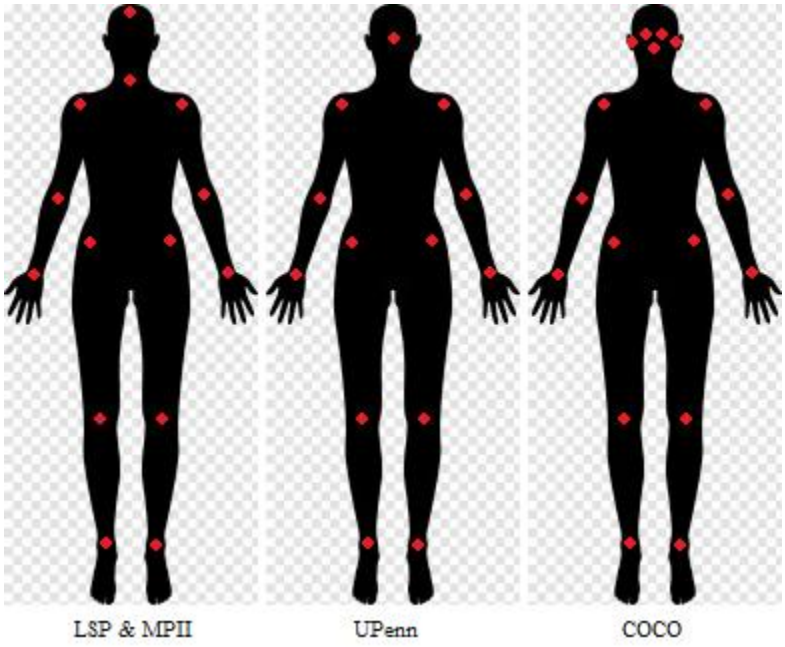


**Fig. 1 Keypoints annotated by some of the most used datasets**

them together.

Leeds Sports Pose (LSP) has two versions: original and extended. Together, they contain 12.000 images of persons during sport activities, crawled from Flickr. Each person is annotated with a 14 keypoints: top of the head, neck, shoulders, elbows, wrists, hips, knees and ankles. A visibility flag for each key pointis also stored. The images are scaled as to make the annotated person 150 pixels in height.

MPII Human Pose is a dataset put together by the Max Planck Institute for Informatics. It contains around 25.000 images extracted from Youtube videos. The images may contain more than one persons, for a total of 40.000 annotated persons. It uses the same set of keypoints as the LSP dataset. It also annotates the type of activity the person is doing, for a total of 410 activity types. The activities are grouped by category, e.g. bicycling, dancing, home repair, music playing, self care, transportation and others.

Penn Action is dataset built by the University of Pennsylvania. It contains over 2300 video sequences of maximum resolution 640x480 pixels, with all frames annotated, making up for 330.000 frames with 330.000 instances of 2300 persons. However, these numbers should be regarded with caution, as the frames from the same video are not independent. There are 13 annotated keypoints (head, shoulders, elbows, wrists, hips, knees and ankles), as well as 15 actions (e.g. baseball swing, bowling, squats, tennis serve). Labels also include the visibility of each keypoint, the viewpoint (front, back, left, right), the bounding box of the person.

COCO-WholeBody is an extension of the COCO dataset with around 200.000 images and 250.000 annotated persons. In addition to the 17 generic body key points (nose, eyes, ears, shoulders, elbows, wrists, hips, knees and ankles), it also defines and annotates 68 key points for face, 42 for hands and 6 for feet. The bounding box for the body is also provided.

TABLE I. DATASETS

| Dataset | Images | Videos | Persons | Key points | Actions |
|---|---|---|---|---|---|
| LSP | 12.000 | - | 12.000 | 14 | 8 |
| MPII | 25.000 | - | 40.000 | 14 | 410 |
| UPenn | - | 2300 | 2300/330.000 | 13 | 15 |
| COCO | 200.000 | - | 250.000 | 17 | - |

## III. Best Ranking Methods

**Zoom Net** [3] is a deep neural network introduced by the same team that built COCO-Whole Body dataset. It takes into account the hierarchical structure of the human body to solve scale variations of certain body parts of the same person. The Faster RCNN architecture is used to extract the person bounding boxes from the image. So this paper is using a top-down approach. Then for each person, ZoomNet first extracts some features from the image, then detects the body key points , then, based on the position of the hands and head, zooms in to detect the hand and the face key points . Hence, it has 4 parts (subnets): FeatureNet, BodyNet, FaceHead, HandHead. BodyNet uses HRNet-W32 as backbone, while the heads for face and hand use HRNetV2p-W18 as backbone. The authors point out several factors that influence the accuracy of the system. Using ground truth (GT) bounding boxes for the persons, the accuracy improved by 23.6%. Medium scale persons yield better results than large scale persons, because the accuracy is measured relative to the person size.

**Key point Communities** [4] detects key points not only on persons, but also on objects (e.g. cars). The method is based on the concept of community. It builds a bottom-up graph of key points and uses a connectivity measure to group them by person or object. The key points are assigned weights according to their importance relative to an ego graph. Groups of key points can be important, even if the individual key points within the group are not so important. The ego graphs are based on the euclidean distance between pairs of key points . The method uses the concept of graph centrality that gives importance to the more central nodes of the graph, like in the case of a social network. The backbone is ShuffleNetV2. Ablations are performed with respect to the key pointweighting method.

**HR Pose** [5] focuses on learning high-resolution representations even in the deepest levels of the convolutional neural network. To this end, the network keeps the big feature maps until the end, but in parallel also adds 3 more series of smaller feature maps, which repeatedly receive information from the biggest feature maps through fusions. Each series of feature maps has the same width throughout, which is half the width of the previous series maps. High-to-low fusions are strided convolutions, while low-to-high fusions are realized through up-sampling. The network regresses a heat map for each keypoint. The authors experimented with two architecture sizes: HRNet-W32 with feature map widths of 256, 128, 64 and 32, and HRNet-W48 with feature map widths of 384, 192, 96 and 48. The method was trained and evaluated separately on MPII and COCO datasets.

**HPR Net** [6] is a bottom-up method that detects all the 133 key points in the COCO-WholeBody dataset in a single shot. It addresses the scale issue of the different body parts by building a point representation of body parts and then regressing them all at once. The face key points are regressed relative to the center of the face and similarly, the hand key points relative to the hand center. Tipically for bottom-up methods, the time consumed is constant: it does not depend on the number of persons in the image. After an Hourglass-104 backbone that extracts the shared features, the network has 7 heads, one for each of the person center heatmap, person center correction, person box, body key pointheatmaps, body key pointoffsets, hand key pointoffsets, and face key pointoffsets.

**Open Pose**[7] is a bottom-up method that learns to associate body parts to persons in a greedy manner, thus being able to perform the computations in real time. The key pointlocations and their associations are learned simultaneously by the two branches of the network. Key pointlocations are predicted as heat maps, while the association between key points is represented as vector fields, which help decide which key points are connected and thus construct a skeleton for each person in the image. The network has several stages, each stage taking as input the output of the previous one, concatenated to the image features. The field of affinities is defined on a limb (the segment between two connected key points ) as unit vectors with direction along the limb. The field takes zero values elsewhere. In terms of speed, this method achieved 8.8 fps (frames per second) on a video with 19 persons.

**Soft-Gated Skip Connections** [8] (SGSC) proposes gated skip connections with learnable parameters for each channel, instead of the usual plain skip connections, a mechanism meant to control the data flow through each channel. In addition, the authors use a hybrid network constructed by combining a heavier Hourglass and a lighter U-Net architecture, achieving the same performance as the heavy Hourglass but with 3 times fewer parameters. Regarding the skip connections, this method proposes two novelties. First, the gate per se is represented by a special module composed of three convolutions. Second, the encoder features are merged into the decoder by concatenations instead of summation.

**Cascade Feature Aggregation** [9] (CFA) proposes a deep neural network composed of multiple Hourglass stages with abundant feature aggregations and fusions between stages, meant to improve the key point localization, but also to make the network robust to unusual poses, occlusions and low resolution images. Final prediction is averaged across the predictions (heat maps) of each stage except the first. The typical hourglass architecture is improved by using ResNet blocks in both encoder

and decoder. As it is hard to achieve convergence from scratch with random initialization for a network like this with more than 3 stages, the authors trained first a 3-stage network and than used the learned parameters to initialize the 4th stage, then train the final 4-stage network.

**TransPose** [10] uses the Transformer architecture, which is able to reveal the local dependencies that the network constructs and uses to predict a certain keypoint. For example, the attention maps of the Transformer reveal that predicting the position of an occluded left ankle depends on the left knee and left hip, but also on the right ankle and right knee positions. These dependencies are image-specific. The network starts with some convolutional blocks (the backbone), followed by three attention stages. For the preliminary convolutional blocks, it uses only the first part of HRNet, accounting for a small fraction of their total number of parameters. The subsequent attention stages are implemented by standard Transformer encoders. A head is appended to predict the key pointheat maps.

**UniPose** [11] achieves human pose estimation in a single stage, incorporating contextual segmentation and joint localization. It uses a waterfall architecture for progressive filtering and in the same time keeps the multi-scale receptive field typical for pyramidal configurations.The network has a ResNet-101 backbone, followed by a WASP (waterfall atrous spatial pooling) module, followed by a decoder that outputs heat maps for the key points . Dilated (atrous) convolutions increase the receptive field while avoiding downsampling.This network also has an extension, UniPose-LSTM, dedicated to pose estimation from video sequences. The LSTM module is placed after the UniPose network and predicts another set of heat maps for the key pointlocations. The LSTM also receives as input its own predictions from the previous frame, achieving temporal consistency.

**Multi-Stage Pose Network**[12] (MSPN) proposes a multi-stage design, cross-stage aggregation of the features, and coarse-to-fine supervision. The network is composed of 8 feature pyramids grouped in 4stages. In each stage, the first pyramid is series of 4 downscaling layers, while the second is a sequence of 4 upscaling layers, informed by the corresponding layers from the first pyramid. The output of the top-most layer of the second pyramid in the first stage is the input for the first pyramid in the second stage. Feature aggregation between stages is preceded by 1x1 convolutions. Each layer contributes to the prediction.

**Spatial Context Network** [13] (SCN) uses contextual information in two different ways. Cascade Prediction Fusion (CPF) is a technique that accumulates prediction from the previous stage and guides the prediction of the following stage. Pose Graph Neural Network (PGNN) captures the relations between human joints as a graph, with messages passing through the graph edges, between connected joints. Ambiguities in key pointlocations from earlier stages are gradually resolved in later stages. CPF comes first, and its predictions are refined by PGNN. The body model includes not only skeleton-based connections, but also some long-distance connections (e.g. between ankle and hip). The backbone is an 8-stack Hourglass. Ablation studies show 0.4 metric points improvement when adding CPF over the backbone baseline and a further 0.8 points when also adding PGNN.

**OmniPose** [14] is an improvement on UniPose [11], by the same authors. It uses multi-scale feature representations which incorporates contextual information through the innovative Waterfall module, which uses a large receptive field while keeping the high resolution of the feature maps. Also, the WASP module from [11] now acts also as a decoder, reducing the network complexity. The backbone is a modified 3-stage HRNet, where deconvolutions with Gaussian heat maps modulations replace the standard upsampling. Separable convolutions are used to reduce the number of parameters.

**Adversarial Data Augmentation** [15] (ADA) is a technique meant to unify data augmentation and training in the same process, using a generative adversarial network (GAN). The generator produces progressively harder augmentations trying to fool the discriminator, which in turn makes progressively accurate predictions. The two components are jointly trained. The training network (the discriminator) is U-Net shaped and decides between standard and generated augmentations, according to a typical HPE loss function. The augmentation network (the generator) is informed by the encoding part of the discriminator's U-Net and outputs distributions of mixed Gaussians, from which scaling and rotations are sampled. Occlusions are generated in the U-Net at the smallest scale.

**Pyramid Residual Modules** [16] (PRMs) learn convolutional filters on various scales of the input features, aiming to better detect the key points in unusual poses or in foreshortened body parts, where the relative scale of certain parts to the others is rather uncommon.The network is composed of stack Hourglass modules, preceded by PRMs. Each PRM has a branch for each scale and the features on each branch are downscaled and then upscaled. In addition, the authors note that Xavier initialization leads to increasing variance in the multi-branch networks, so they lay out a theoretical basis for a new type of initialization for such networks.

**LSTM Pose Machines** [17] (LSTM PM) is designed to estimate human pose in video sequences. Specific challenges are the consistency from frame to frame that can cause flickering, as well as the

low quality of some frames, due to motion blur. The authors note that a multi-stage convolutional neural network (CNN) with sharing weights can be rewritten as a recurrent neural network (RNN), which is well suited for the task. Long Short-Term Memory (LSTM) units are inserted between the frames to ensure temporal geometric consistency. The optimal number of LSTM iterations is found to be 5, corresponding to the number of past frames that are still useful for the current frame prediction.

**Thin-Slicing Network** [18] (TSN) targets video sequences and uses a body model as prior knowledge, in order to ensure temporal consistency. The network can represent both the appearance and the spatio-temporal relationship between the key points . It takes several consecutive frames as input and predicts initial key pointlocations, while also computing the dense optical flow between the frames. Then a spatio-temporal layer passes messages iteratively through the edges of a loopy graph representing a spatio-temporal view of the body model, yielding the final predictions. For example, the right ankle from the current frame is linked with the right knee from the current frame, but also with the right ankle from the previous frame.

**DarkPose** [19] relies on extracting the right key pointlocation from the predicted heat map, as well as the encoding of the GTs as heat maps. Usually the heat maps are predicted at a lower resolution than the original and they have to be scaled back. It is also supposed that the heatmaps would have near-Gaussian shape, but the authors found that this is often not the case. So they first modulate the low resolution heat maps through a convolution with a Gaussian kernel. On the other hand, the GT encoding suffers from quantization error, when the heat map is generated after the image is scaled down. So the authors renounced the quantization altogether. Best results are achieved with an HRNet backbone. This method still works fairly well for low resolution inputs, for with a much reduced network complexity is needed, so it can be very fast.

**Cascaded Pyramid Network** [20] (CPN)is a top-down approach that, after person bounding box detection, has two stages, called GlobalNet and RefineNet. GlobalNet is a feature pyramid that detects the easier key points , but struggles with the harder one, affected by occlusion or difficult background. RefineNet integrates the feature representation from the first stage and uses hard key pointmining to give more weight in the loss to the harder key points . The authors also investigate different non-maximum suppression (NMS) thresholds for the bounding box detection, and find that Soft-NMS is the best choice. Also, as backbone, Resnet-50 is found to be superior to 8-stage Hourglass. In addition, ensemble models give the best results.

**OpenPifPaf** [21] proposes a bottom-up, single-stage network for real-time key pointdetection and tracking, applicable not only to persons, but also to cars and animals, with direct applicability to self-driving cars and delivery robots. It defines a spatio-temporal pose as a graph spanning multiple frames. Composite Intensity Fields (CIF) are confidence maps that reach maximal values in the vicinity of key pointlocations. Composite Association Fields (CAF) regress the locations of source and target joints, for each limb and for each point in the image, and also the size of each joint. The network can be trained jointly on multiple datasets. At training time, the input is several consecutive frames, while at inference time an additional layer is inserted that merge the features cached from the previous frame into the current features. The body model has redundant connections, which helps with the occluded joints or sparse pose, where the visible body parts of a person are not direct neighbors.

**Pose Residual Network** [22] (PRN) is a bottom-up method that simultaneously handles key pointdetection, person detection and semantic segmentation. After the ResNet-101 backbone, the network splits in two branches, the key pointsubnet and the person subnet, the latter predicting bounding boxes, as well as segmentation. The two branches are then rejoinedby the final module, the actual PRN, which groups the key points by person. The key pointsubnet is composed of three successive feature pyramids.

## IV. Summary of the Techniques

Most methods use an established backbone, sometimes in a slightly modified version, or taking only the relevant portion of it. Some methods use entirely custom built networks. The most common backbones are Hourglass, HRNet, ResNet and ShuffleNet. **Hourglass** is used by 6 methods ([6], [8], [9], [13], [15], [16]) and was first introduced as U-Net [23]. It shrinks progressively the input while encoding it, then the feature maps are enlarged back to original size in a symmetric manner.For the HPE task, there are usually many (8) Hourglass modules stacked back to back. **HRNet** [24] is used by 5 methods ([3], [5], [10], [14], [19]). It has 4 stages, from the first stage that contains feature maps only at the input resolution, to the last stage that contains feature maps at 4 different resolutions.**ResNet** [25] is used by 4 methods ([11], [12], [20], [22]) and it was the first to use residual (skip) connections to address the issue of vanishing gradients in deep networks.**ShuffleNet** [26] is used by 2 methods ([4], [21]). It uses group convolutions and channel shuffle to reduce the computational cost of the network.

The techniques used by the most successful methods can be summarized as follows:

- Multi-branch networks or subnets:[7], [16], [22]
- Averaging the predictions from several stages: [9], [12], [13]
- Graph neural network to emulate relationships between key points : [13], [18], [21]
- Body model: [3], [18]
- Vector fields for key pointaffinity: [7], [21]
- Feature aggregation between stages: [9], [12]
- WASP module for progressive filtering: [11], [14]
- LSTM for temporal processing: [11], [17]
- Gaussian heat map modulations for upscaling features [14] and predictions [19]
- Groups of key points : [4]
- Training progressively the multi-stage network: [9]
- Transformer as attention module: [10]
- GAN to generate hard augmentations during training: [15]
- Weight initialization specific to multi-branch networks: [16]
- Optical flow between frames: [18]
- Hard key pointmining: [20]
- Soft-NMS [27] to choose the person bounding box: [20]
- Semantic segmentation head: [22]

## V. METRICS AND EVALUATION

The metric for the COCO dataset [28] is mean average precision (mAP), i.e. the mean over all classes (key pointtypes) of the average over all recall thresholds of the precisions of each keypoint. Object Key pointSimilarity (OKS) is used to measure the similarity between the prediction and the GT and it depends on the Euclidean distance between prediction and GT, the scale of the human body and the standard deviation of the human GT annotations.

The metric for LSP and UPenn datasets [29] is the probability of correct key point(PCK), which depends on the Euclidean distance between prediction and GT and the size of the person bounding box. The metric for MPII dataset (PCKh) [30] is a variant of PCK which depends on the size of the person's head, instead of its whole body size.

Table II shows the evaluation scores of the methods presented above. All times are obtained by the authors on an Nvidia GTX 1080 Ti GPU. Scores marked with asterisk (*) are obtained by training on additional data. Methods marked with a plus sign (+) were using an ensemble of models. Method complexity is expressed in GFlops (floating-point operations) and it depends on the network input size, hence on the dataset, therefore an interval is given for methods evaluated on multiple datasets. Time and complexity should be proportional.

TABLE II. EVALUATION OF HPE METHODS

| Method | Backbone | Complexity [GFlops] | Time [ms] | mAP on COCO | PCKh@0.5 on MPII | PCK on LSP | PCK@0.2 on UPenn |
|---|---|---|---|---|---|---|---|
| DarkPose | HRNet-W48 | 32.9 | N/A | **77.4** | - | - | - |
| OmniPose | custom HRNet | 22.6-37.9 | N/A | 76.4 | - | **99.5** | **99.4** |
| ZoomNet | HRNet-W32 | 27.36 | 175 | 74.3 | - | - | - |
| CPN+ | ResNet-50 | 13.9 | N/A | 73.0 | - | - | - |
| OpenPifPaf | ShuffleNetV2K30 | N/A | 152 | 70.9 | - | - | - |
| PRN | ResNet-101 | N/A | N/A | 69.7 | - | - | - |
| KC | ShuffleNetV2 | N/A | 93 | 69.6 | - | - | - |
| SGSC | custom Hourglass | 9.9 | N/A | - | **94.1*** | 94.8 | - |

| Method | Backbone | Complexity [GFlops] | Time [ms] | mAP on COCO | PCKh@0.5 on MPII | PCK on LSP | PCK@0.2 on UPenn |
|---|---|---|---|---|---|---|---|
| CFA | custom Hourglass | 73 | N/A | - | 93.9* | - | - |
| TransPose | reduced HRNet | 21.8 | 27 | - | 93.5* | - | - |
| UniPose | ResNet-101 | N/A | N/A | - | 92.7 | 94.5 | 99.3 |
| MSPN | 4x ResNet-50 | 19.9 | N/A | - | 92.6 | - | - |
| SCN | 8x Hourglass | N/A | N/A | - | 92.5 | 94.0 | - |
| HRPose | HRNet | 9.5-32.9 | N/A | 65.9 | 92.3 | - | - |
| PRMs | 8x Hourglass | 14.7 | N/A | - | 92.0 | 93.9 | - |
| ADA | GAN with U-Net | N/A | N/A | - | 91.5 | 94.5 | - |
| HPRNet | Hourglass-104 | N/A | 101 | 59.4 | - | - | - |
| OpenPose | custom | N/A | 100 | 56.3 | 88.8 | - | - |
| LSTM PM | custom | N/A | N/A | - | - | - | 97.7 |
| TSN | custom | N/A | N/A | - | - | - | 96.5 |

## VI. Conclusions

State-of-the-art HPE methods all use deep neural networks, but a great variety of techniques. Most networks are pretty heavy and cannot be expected to run in real time, but some of them have lighter versions designed with speed in mind, while not compromising the accuracy too much. Top-down methods tend to perform better but are slower then bottom-up methods. Increasing the efficiency to make the methods more practical should be an important objective.

While very good results were obtained in most of the cases, there still remain challenges like unusual poses, occlusions, crowded people and low resolution images, or motion blur in the case of video sequences. In spite of this, some of the datasets have almost been saturated (see the high scores on LSP and UPenn), so maybe there is a need for a new, more difficult dataset. Synthetic data has not been used much and there is a virtually unlimited amount that can be generated, although it will bring the problem of domain adaptation.

## Acknowledgment

The results presented in this work concerns the research carried out for the "SENTIR" research project, co-financed through the European Regional Development Fund, POC–A.1–A1.2.1–D–2015 grant, research, development and innovation supporting economic competitiveness and the development of businesses.

## References

[1] Y. Chen, Y. Tian, M. He, "Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods", Computer Vision and Image Understanding (CVIU), arXiv:2006.01423, 2020

[2] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, M. Shah, "Deep Learning-Based Human Pose Estimation: A Survey", unpublished, arXiv:2012.13392, 2020

[3] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, P. Luo, "Whole-Body Human Pose Estimation in the Wild", European Conference in Computer Vision (ECCV), arXiv:2007.11858, 2020

[4] D. Zauss, S. Kreiss, A. Alahi, "Key pointCommunities", International Conference on Computer Vision (ICCV), arXiv:2110.00988, 2021

[5] K. Sun, B. Xiao, D. Liu, J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1902.09212, 2019

[6] N. Samet, E. Akbas, "HPRNet: Hierarchical Point Regression for Whole-Body Human Pose Estimation", unpublished, arXiv:2106.04269, 2021

[7] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1611.08050, 2017

[8] A. Bulat, J. Kossaifi, G. Tzimiropoulos, M. Pantic, "Toward Fast and Accurate Human Pose Estimation via Soft-gated Skip Connections", unpublished, arXiv:2002.11098, 2020

[9] Z. Su, M. Ye, G. Zhang, L. Dai, J. Sheng, "Cascade Feature Aggregation for Human Pose Estimation", unpublished, arXiv:1902.07837, 2019

[10] S. Yang, Z. Quan, M. Nie, W. Yang, "TransPose: Key pointLocalization via Transformer", International Conference on Computer Vision (ICCV), arXiv:2012.14214, 2021

[11] B. Artacho, A. Savakis, "UniPose: Unified Human Pose Estimation in Single Images and Videos", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:2001.08095, 2020

[12] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, J. Sun, "Rethinking on Multi-Stage Networks for Human Pose Estimation", unpublished, arXiv:1901.00148, 2019

[13] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, J. Jia, "Human Pose Estimation with Spatial Contextual Information", unpublished, arXiv:1901.01760, 2019

[14] B. Artacho, A. Savakis, "OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation", unpublished, arXiv:2103.10180, 2021

[15] X. Peng, Z. Tang, F. Yang, R. Feris, D. Metaxas, "Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1805.09707, 2018

[16] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, "Learning Feature Pyramids for Human Pose Estimation", International Conference on Computer Vision (ICCV), arXiv:1708.01101, 2017

[17] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, L. Lin "LSTM Pose Machines", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1712.06316, 2018

[18] J. Song, L. Wang, L. Van Gool, O. Hilliges, "Thin-Slicing Network: A Deep Structured Model for Pose Estimation in Videos", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1703.10898, 2017

[19] F. Zhang, X. Zhu, H. Dai, M. Ye, C. Zhu, "Distribution-Aware Coordinate Representation for Human Pose Estimation", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1910.06278, 2020

[20] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, "Cascaded Pyramid Network for Multi-Person Pose Estimation", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1711.07319, 2018

[21] S. Kreiss, L. Bertoni, A. Alahi, "OpenPifPaf: Composite Fields for Semantic Key pointDetection and Spatio-Temporal Association", unpublished, arXiv:2103.02440, 2021

[22] M. Kocabas, M. S. Karagoz, E. Akbas, "MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network", European Conference in Computer Vision (ECCV), arXiv:1807.04067, 2018

[23] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", Inrernational Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), arXiv:1505.04597, 2015

[24] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, "Deep High-Resolution Representation Learning for Visual Recognition", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3349-3364, 1 Oct. 2021

[25] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", unpublished, arXiv:1512.03385, 2015

[26] X. Zhang, X. Zhou, M. Lin, J. Sun, "ShuffleNet: An Extremely Efficient Convolution Neural Network for Mobile Devices", Conference on Computer Vision and Pattern Recognition (CVPR), arXiv:1707.01083, 2018

[27] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, "Improving Object Detection with One Line of Code", International Conference on Computer Vision (ICCV), arXiv:1704.04503, 2017

[28] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, Y. Wang, "AI Challenger: A Large-Scale Dataset for Going Deeper in Image Understanding", IEEE International Conference on Multimedia and Expo (ICME), arXiv:1711.06475, 2019

[29] Y. Yang, D. Ramanan, "Articulated Human Detection with Flexible Mixture-of-Parts", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2878-2890, Dec. 2013

[30] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, "2D Human Pose Estimation: New Benchmark and State-of-the-Art Analysis", 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686-3693, 2014