https://doi.org/10.31449/inf.v45i4.3470 Informatica **45** (2021) 593–604 **593**

Skeleton-aware Multi-scale Heatmap Regression for 2D Hand Pose Estimation

Ikram Kourbane and Yakup Genc
Department of Computer Engineering, Faculty of Engineering, Gebze Technical University, Kocaeli, Turkey
E-mail: ikourbane@gtu.edu.tr; yakup.genc@gtu.edu.tr

Keywords: hand pose estimation, hand detection, hand dataset, convolutional neural networks, heatmaps

Received: March 14, 2021

Hand pose estimation plays an essential role in sign language understanding and human-computer interaction. Existing RGB-based 2D hand pose estimation methods learn the joint locations from a single resolution, which is not suitable for different hand sizes. To tackle this problem, we propose a new deep learning-based framework that consists of two main modules. The first one presents a segmentation-based approach to detect the hand skeleton and localize the hand bounding box. The second module regresses the 2D joint locations through a multi-scale heatmap regression approach that exploits the predicted hand skeleton as a constraint to guide the model. Moreover, we construct a new dataset that is suitable for both hand detection and pose estimation tasks. It includes the hand bounding boxes, the 2D keypoints, the 3D poses and their corresponding RGB images. We conduct extensive experiments on two datasets to validate our method. Qualitative and quantitative results demonstrate that the proposed method outperforms the state-of-the-art and recovers the pose even in cluttered images and complex poses.

Povzetek: V prispevku je predstavljena učna metoda za nalogo 2D ocenjevanja položaja roke z uporabo monokularne RGB kamere.

1 Introduction

The hands are one of the most important and intuitive interaction tools for humans. Solving the hand pose estimation problem is crucial for many applications, including human-computer interaction, virtual reality, augmented reality and sign language recognition.

The earlier works in hand tracking use special hardware to track the hand, such as gloves and visual markers [1]. But, these types of solutions are expensive and restrict the applications to limited scenarios. Tracking hands without any device or markers is desirable. To this end, several works have been proposed in the literature to tackle this problem [2]. However, markerless hand pose estimation is very challenging due to strong articulations and self-occlusions. Furthermore, the hands have a huge variation in shape, size, skin texture and color.

The rapid development of deep learning techniques revolutionizes complex computer vision problems [3, 4] and outperforms conventional methods in many challenging tasks, including object classification [5], object segmentation [6, 7] and object detection [8, 9]. Hand pose estimation is not an exception and deep convolution neural networks (CNNs) [10] have been applied successfully in [11, 12, 13]. These studies address the scenarios where the hand is tracked via an RGB-D camera. However, depth-enhanced data is not available everywhere, and they need an overhead setup to utilize. Thus, estimating the hand pose from a single RGB image has been an active and challenging area of research, as they are cheaper and easier to use than depth sensors [14, 15, 16, 17].

We can classify RGB-based hand pose estimation methods into two broad categories as regression-based and detection-based. The former approach uses CNNs as an automatic feature extractor to directly estimate the joint locations [14, 18, 19]. Although the regression-based approach is fast at inference time, it remains a difficult optimization problem due to its non-linear nature requiring many iterations and a lot of data for convergence [20].

To overcome these limitations, recent solutions to human and hand pose estimation problems use probability density maps such as the heatmap [16, 21, 22]. They divide the pose estimation problem into two steps. The first one finds a dense pixel-wise prediction for each joint while the second step infers the joint locations by finding the maximum pixel in each heatmap. The heatmap representation helps the neural network to estimate the joint locations robustly and has a fast convergence property.

In this work, we focus on the 2D hand pose estimation from a single RGB image. This task is also challenging due to the many degrees of freedom (DOF) and the self-similarity of the hand. The proposed approach has two principal components; The first one estimates the hand skeleton using the UNet-based architecture [23]. The hand bounding boxes are extracted in a post-processing step from the predicted skeleton. The second part presents a new multi-scale heatmap regression approach to estimate joint locations from multiple resolutions. Specifically, the network output is supervised on different scales to ensure accurate poses for different hand image sizes. This strategy helps the model for better learning of the contextual and the location information. Besides, our method uses the

predicted hand skeleton as additional information to guide the network to predict the 2D hand pose.

We validate the proposed method on a common existing Large-Scale Multi-View hand pose dataset (LSMV) [18]. Furthermore, we create a new dataset suitable for hand detection and 2D pose estimation tasks using leap motion sensors. This dataset includes 60 thousand samples, such that each one contains the hand bounding box, the 2D keypoint, 3D pose and the corresponding RGB image. We extended our experiments to our newly created dataset (GTHD). Results demonstrate that our method generates accurate poses and outperforms three state-of-the-arts [18, 24, 25]. In summary, our contributions are the following:

- We propose a segmentation-based approach for skeleton detection and hand bounding box localization.
- We propose a multi-scale heatmap regression architecture that uses the hand skeleton as additional information to constrain the 2D hand pose estimation task. The reported qualitative and quantitative results demonstrate the competitiveness of the proposed method.
- We introduce a new dataset to validate the hand detection and the 2D pose estimation methods.

We organize the rest of the paper as the following. Section 2 gives the problem definition as well as the related work. Section 3 describes in detail our hand detection and pose estimation approaches and defines the required steps to build our hand pose dataset. Section 4 discusses the conducted experiments and the obtained qualitative and quantitative results. Finally, Section 5 provides the main conclusion of this work and a direction for further research.

2 Related work

2.1 Hand detection

The hand detection task identifies the hand region and distinguishes it from the background. It has many applications including, gesture recognition [26] hand segmentation and hand tracking [27]. Traditional computer vision methods follow a feature extraction and classification scheme for hand detection. They extract skin color features, shape features or combine the two types of features to represent the image [28]. Following, they utilize a classifier to check each pixel, whether it belongs to the hand or not [29].

Deep learning-based methods circumvent such bottlenecks by unifying feature extraction and classification phases. This combined strategy has been outperforming conventional methods for the last five years. For instance, [30] employs two streams Faster RCNN [8] for hand detection. The first stream extracts feature maps from depth video while the second one extracts it from RGB video. After that, they use an alignment stage to connect the two features and they run a region proposal network to classify the pixels. Another method [31] applies multi-scale Faster-RCNN to avoid missing the small hands.

2.2 Hand pose estimation

Estimating the 2D hand pose has been an active and challenging area of research in computer vision. Recently deep learning-based methods achieve competitive performance in this task as well. We can classify these based on the input modality into two broad categories as depth-based and RGB-based. In the former class, several studies achieve accurate 2D pose estimation results for images containing single hand [11, 12, 13, 32]. Also, [33] handles multi-hands using pictorial structure models and Mask-RCNN.

RGB-based methods are more challenging and less studied in the literature. Early studies give the cropped hand image as input to the ResNet-based model to directly regress the 2D joint by minimizing the mean-squared error (MSE) between the predicted 2D joint annotations and their ground truth [18]. Recently, [25] employs a graph-based framework to allow features at each node to be represented by 2D spatial confidence maps. Also, [24] propose an adaptive graphical model network that includes two branches of CNNs computing unary and pairwise potential functions and a graphical model to integrate the calculated information. [34] employs a cascaded CNN to predict the silhouette information (mask) and the 2D key-points in an end-to-end manner after localizing the hand region. To perform efficient small hand 2D pose estimation, [35] simultaneously regresses the hand region of interests and hand key-points. Subsequently, it iteratively uses the hand RoIs as feedback information for boosting the hand keypoints estimation performance. [8] proposes the Limb Probabilistic Mask, which uses a Gaussian distribution mask rather than the one-hot mask. To address the self-occlusion issue, it splits the whole hand mask into five fingers and the palm. The 2D pose regression task employs the synthesized hand mask to model the structural relationship between the 2D keypoints. All the aforementioned state-of-the-art methods results are presented in Table 1 that summarized the hand detection and pose estimation techniques. Besides, it shows the used datasets, including LSMV [18], OneHand10K [34], CMU and MPII+NZSL [37]. The results are reported using the mean PCK metric [37], which is widely used to evaluate human and hand pose estimation methods. It considers the predicted joints as correct if the distance to the ground truth joint is within a certain threshold γ. Some approaches use a normalized threshold by dividing all the joints values by the size of the hand bounding box. In this work, we propose a new multi-scale heatmap regression architecture that uses the 2D skeleton as a constraint to accurately estimate the 2D hand pose for small and big hands.

	Hand detection	Model	Estimation method	$meanPCK^{\uparrow}_{a,b}$	Threshold γ
Gomez et al [18]	Faster R-CNN for bounding box detection	ResNet-50	Direct regression	80.74 on LSMV (Self-dataset)	0.01-0.06 (N)
Kong et al [24]	Cropping square image patches of annotated hands	Adaptive graphical model	Heatmaps detection	70.34 on CMU and 85.63 on LSMV	0.01-0.06 (N)
Kong et al [25]	Cropping square image patches of annotated hands	Spatial information aware graph convolutional network	Heatmaps detection	81.72 on MPII+NZSL, 71.65 on CMU and 85.56 on LSMV	0.01-0.06 (N)
Wang et al [34]	Semantic segmentation using CNN	Mask-pose cascaded CNN	Heatmaps detection	90.27 on OneHand10K (Self-dataset) and 74.82 on MPII+NZSL	0.2
Wang et al [35]	Hand region localization using CNN-based bounding box regression	Simultaneously regress the hand region of interests and hand keypoints	Heatmaps detection	0.94 on OneHand10K	0.2
Chen et al [8]	Limb probabilistic mask with splitting the hand into fingers and palm	Nonparametric structure regularization machine	Direct regression	88.46 on OneHand 10k and 80.03 on CMU	0.1-0.3 and 0.04-0.012 (N)

Table 1: Summary of related 2D hand pose estimation approaches and their obtained results. We show the $meanPCK$ metric for defined thresholds on a specific dataset. $\uparrow$: higher is better, a, b: begin and the end of the experimented interval of thresholds γ and N refers to a normalized threshold.

3 Proposed method

Our proposed approach for 2D hand pose estimation uses a skeleton-based approach to detect the hand and extract the bounding boxes. The second part uses the predicted skeleton as a constraint to guide the proposed multi-scale heatmap regression approach to predict the 2D joint locations of the cropped hand.

3.1 Skeleton detection and hand bounding box localization

We represent the detected hand location in an image by a rectangular region with four corners. Faster-RCNN [8] type of deep network models directly regress the four corner coordinates from the given hand image.

Alternatively, we can predict the 2D hand skeleton and extract the bounding box in a post-processing step (Figure 1). Direct regression of the bounding box is useful for hand cropping but cannot be further exploited for other tasks. In contrast, estimating the hand skeleton includes useful information that guides the 2D pose estimation. Also, the segmentation task is less challenging than predicting the bounding box.

Of course, one needs to have the training data with corresponding skeletons. We can obtain this type of data using a 3D hand tracker and an RGB camera to provide the 2D key-points (see Section 3.3). We create the ground truth data for the skeleton by connecting the joints in each finger and attaching the palm to the ends of each finger. Also, we represent each joint location by the standardized Gaussian blob.

We can treat hand skeleton data as a segmentation mask. Thus, we use the well-known UNet architecture [23] since it is one of the best encoder-decoder architectures for semantic segmentation. It has two major properties. The first one is the skip connections between the encoder and the decoder layers that enable the network to learn the location and the contextual information. The second property is its symmetry, leading to better information transfer and performances.

The model outputs single feature maps on which we apply a *sigmoid* activation function to bound the prediction values between 0 (background) and 1 (hand). We localize the bounding box using a post-processing step, in which we identify the foreground pixels, and then we apply a region growing algorithm. In our case, the horizontal and vertical boundaries of the recovered regions are reported as the location of the detected hand. Our model robustly differentiates between the skin of the hand and that of the face. Also, it can detect the hand even in cluttered images or different lightning conditions (see Section 4.2).

Concerning the loss function, we did two experiments. In the first one, we only used the L_1 loss function, which can not robustly localize the skeleton and adversely affecting the bounding box localization results. In contrast, using the combination of L_1 loss (Equation 1) and a *SoftDice*

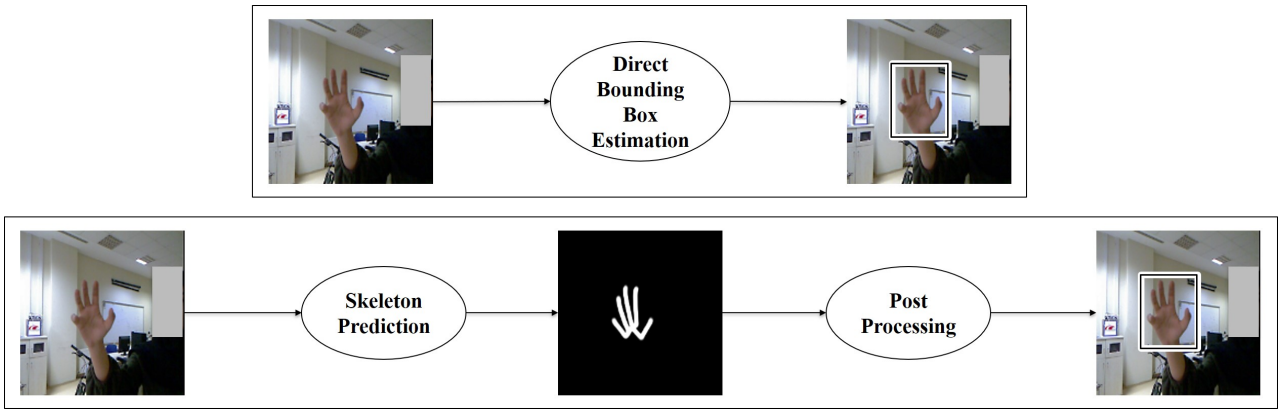

Figure 1: The proposed method for hand bounding box detection. Unlike many deep learning approaches that use Faster R-CNN [8] model to directly estimate the bounding box (top), we predict the skeleton image and infer the bounding box in a post-processing step (bottom).

(Equation 2) loss with their empirical weights can robustly localize the hand (Equation 3).

$$L_1(x, \hat{x}) = \| x - \hat{x}\|_1^1 \tag{1}$$

$$SoftDice(x, \hat{x}) = 1 - \frac{2\hat{x}^T x}{\|\hat{x}\|_2^2 + \|x\|_2^2} \tag{2}$$

$$Total(x, \hat{x}) = \lambda_1 L_1(x, \hat{x}) + \lambda_2 SoftDice(x, \hat{x}) \tag{3}$$

Where: x, $\hat{x}$, λ_1 and λ_2 represent the ground truth skeleton, the predicted skeleton and the two hyperparameters of the loss function (set to 0.4 and 0.6, respectively). We trained the model for 20 epochs using a batch size of 8.

3.2 Multi-scale heatmaps regression for 2D hand pose estimation

Most of the existing hand pose estimation methods predict the heatmaps at a single-scale. However, the hand in the original image can have several sizes (close/far hands). Hence, when we use a single scale image, the cropped hand image size cannot be suitable for all the dataset samples.

To address this limitation, we propose a multi-scale heatmaps regression architecture that performs the back-propagation process for many resolutions providing better joint learning for both large and small hands. Moreover, the cropped hand image would include some parts of the background. To overcome this problem, we employ the predicted hand skeleton to act as an attention mechanism for the network to focus on hand pixels. This makes the 2D pose regression task less challenging to optimize.

Figure 2 shows our skeleton-aware multi-scale heatmaps network approach for 2D hand pose estimation. We feed the concatenation of the cropped hand image and the predicted skeleton to the first convolution layer. The latter is followed by two downsampling ResNet blocks, two upsampling ResNet blocks, and a final transposed convolution layer that recovers the input resolution. After each downsampling (similarly upsampling), we apply a 1×1 convolution layer followed by a *sigmoid* activation function to output 21 or 20 feature maps representing the heatmaps in GTHD or LSMV datasets, respectively. The heatmaps resolution is divided/multiplied by two after each downsampling/upsampling.

In test time, we calculate a weighted average of the predicted five heatmaps to find the coordinate of the 2D key-points. We formulate the loss function in (Equation 4) as:

$$L(x, \hat{x}) = \sum_{i=1}^{k} \delta_i \| x_i - \hat{x}_i\|_2^2 \tag{4}$$

Where: k is the number of scales including the full resolution output, and δ_i is the weight given for each scale. In our experiments we choose $k = 5$ and δ_i is set be 1, 1/2 and 1/4 for scales 128, 64 and 32 respectively.

3.3 Datasets

Deep learning methods require a large number of labeled data for training. There is a lack of datasets that has RGB hand images with their 2D annotations that we can use to train our proposed approaches. For example, [38] has RGB images with their 2D annotations, but they are both small scale and do not describe the hand by joint annotations.

Our method has been implemented and tested on two different datasets. The first one is LSMV [18], which is one of the large-scale datasets that provide the hand bounding boxes, the 2D key-points as well as the 3D pose. We split the data into 60000, 15000, and 12760 samples for the training set, validation set, and test set, respectively. While LSMV [18] can be used to train and validate the 2D hand pose estimators, it can not be used for hand detection since it does not have images without hands.

To overcome this limitation, and train both the hand detector and the hand pose estimator, we have built our own

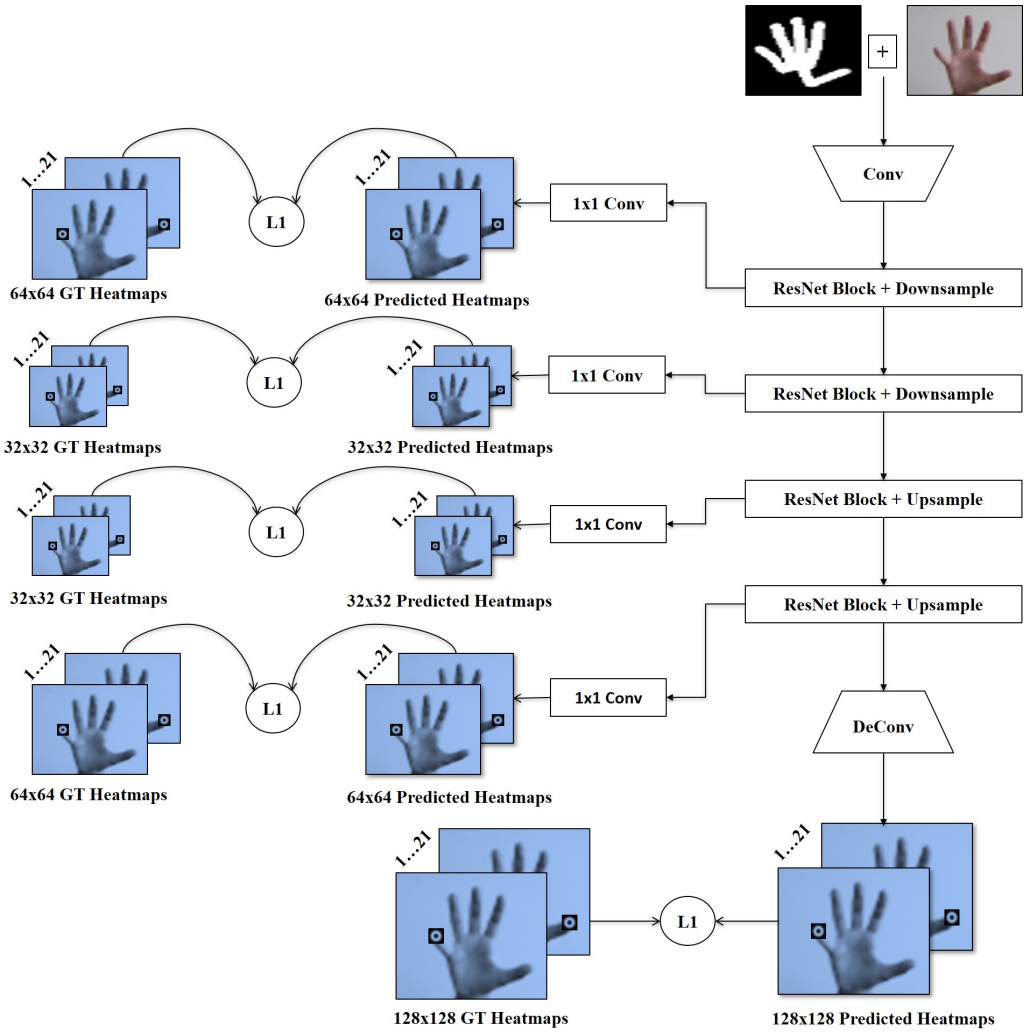

Figure 2: The overall architecture of the proposed 2D hand pose estimation approach uses the hand skeleton as a constraint and estimates the joint heatmaps from multiple scales.

dataset (GTHD) using an RGB camera and a Leap Motion sensor [39]. It is composed of two subsets; The first one has 60 thousand RGB images with their corresponding hand bounding boxes, 2D keypoints, and 3D pose. The second set has 15 thousand RGB images that present either the background or people who do not show their hands. The new dataset has a large variation in hand poses, backgrounds, skin color and texture

The RGB camera provides an image with a resolution of 640×480 pixels. The leap motion controller is a combination of hardware and software that senses the fingers of the hand to provide the 3D joint locations. Hence, a projection process from 3D space to the 2D image plane is necessary. We achieve this goal in two steps. In the first one, we use OpenCV to estimate specific intrinsic parameters of the camera. In the second step, we estimate the extrinsic parameters between the leap motion controller and the camera. To get the correct pose with its corresponding image, we synchronize the two sensors in time.

Finally, to find the rotation and translation matrices, we manually mark one key-point in a set of hand images and solve the *PnP* problem by computing the 3D-2D correspondences [40]. Figure 3 illustrates the results of the calibration process. We randomly split the GTHD dataset into a training set (75%), a validation set (10%) and a test set (15%).

3.4 Evaluation metrics

We report the performance of the hand skeleton detection module using famous classification metrics, such as *Accuracy*, *Precision*, *Recall* and *F1*. Furthermore, we calculate the Area Under ROC Curve (*AUC*) for GTHD datasets since it measures how well the two classes (Hand and No-Hand) are separable. It calculates the trade-off between the true positive rate and the false positive rate. Also, we report the Intersection over Union (*IOU*) metric to quantify our model performance in the hand bounding box detection task. It evaluates the predicted bounding boxes by comparing them against the ground truth.

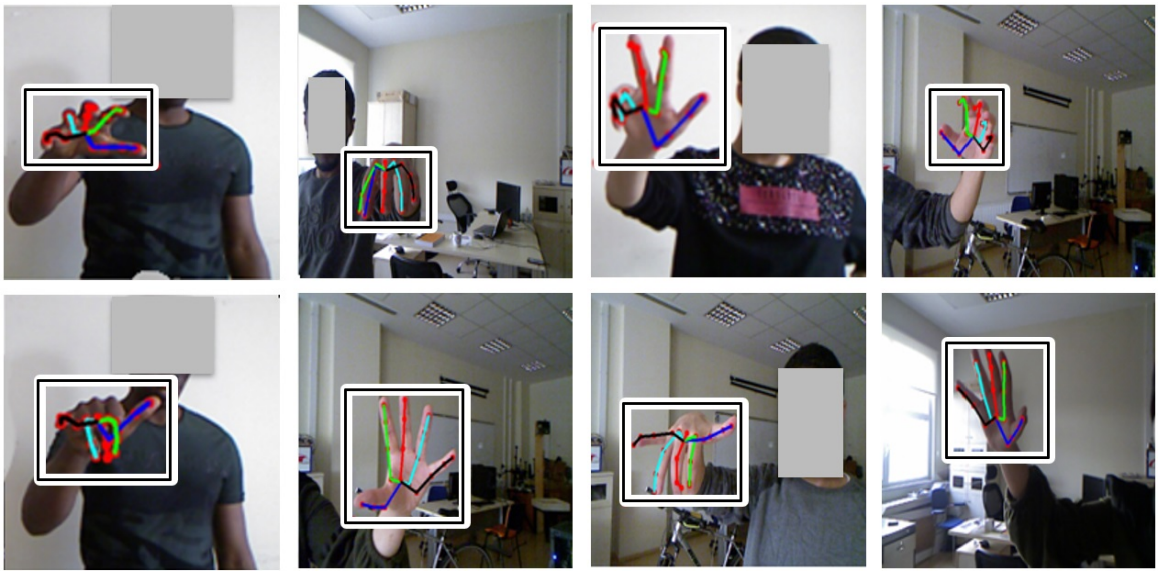

Figure 3: Examples of our hand dataset images having the bounding boxes and 21 joints annotations taken from four subjects and covering many pose and backgrounds.

To quantitatively evaluate the performance of the proposed 2D hand pose regression methods, we use the Probability of Correct Keypoint (PCK) metric [37] as it is used frequently in human and hand pose estimation tasks.We use a normalized threshold by dividing all the joints values by the size of the hand bounding box. Also, for additional quantification of the performance of the proposed method, we report the mean joint pixel error (MJPE) over the input hand image with 128×128 resolution.

4 Experiments

4.1 Implementation details

We train the models for 30 epochs using a batch size of 8 and Adam as an optimizer. We initialize the learning rate to 0.01, and we decrease it after every eight epochs by 10%. We conduct all experiments on NVIDIA GTX 1080 GPU using PyTorch v1.6.0.

Before giving the RGB image as input to the model, we resize and normalize the datasets by subtracting the mean from all the images. The number of heatmaps is the number of joints where we represent each one with a Gaussian blob in a map of the same size of the image. The coordinate of the joint is the location of the highest value in the heatmap. We find them by applying the *argmax* function.

To validate our hand detection approach, we use different degrees of skeleton thickness (1, 3 and 6). In the first case, the skeleton is simply composed of lines. In other cases, it has thicker connections and regions around the joints. Furthermore, we select a threshold that represents the the number of foreground pixels to be the criterion to separate the two classes (Hand and NoHand).

Dataset	Faster-RCNN [8]	Ours
GTHD	0912	0.923
LSMV	0.895	0.917

Table 2: Bounding box evaluation on LSMV and our GTHD dataset with IOU.

4.2 Hand detection and bounding-box estimation results

Our approach for hand bounding box localization can robustly estimate the hand skeleton and localize the hand bounding box for the two datasets. It does not produce any false positives for background images or images with people who do not show their hands (see Figure 4 and Figure 5).

The correct threshold for selecting *Hand* from *NoHand* depends on the data. A robust threshold should eliminate the noise and be in an interval that does not miss samples from the dataset distribution. In other words, the selected threshold should decrease both the false-negative rate (adding samples from the *NoHand* class) and false positive rate (missing samples from the *Hand* class) to achieve high performance and robustly detect the hand. Figure 7. shows that selecting a threshold from the interval $[200, 400]$ is the best choice for our dataset. Also, the thickest skeleton representation seems to be more robust to the noise. It outperforms the other representations and achieves a higher performance ($AUC = 0.99$). Finally, our approach records high scores of $Accuracy = 0.99$, $Precision = 0.97$, $Recall = 0.99$ and $F1 = 0.98$.

We do not report the AUC for the LSMV dataset because it does not have images without hands. Nevertheless, we predict the skeleton representation to extract the

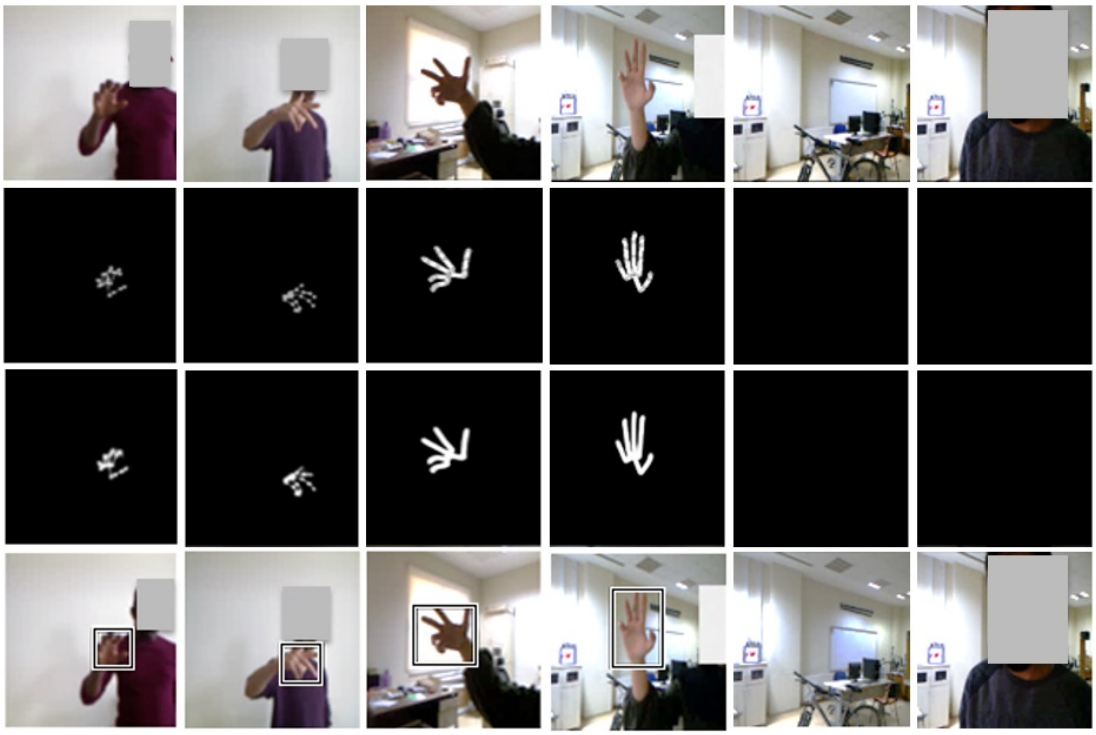

Figure 4: The results of the skeleton estimation and the bounding box localization on the GTHD dataset using thick and thin skeleton representations. The rows from top to down show: the input image, the ground truth skeleton, the predicted skeleton, and the obtained bounding boxes.

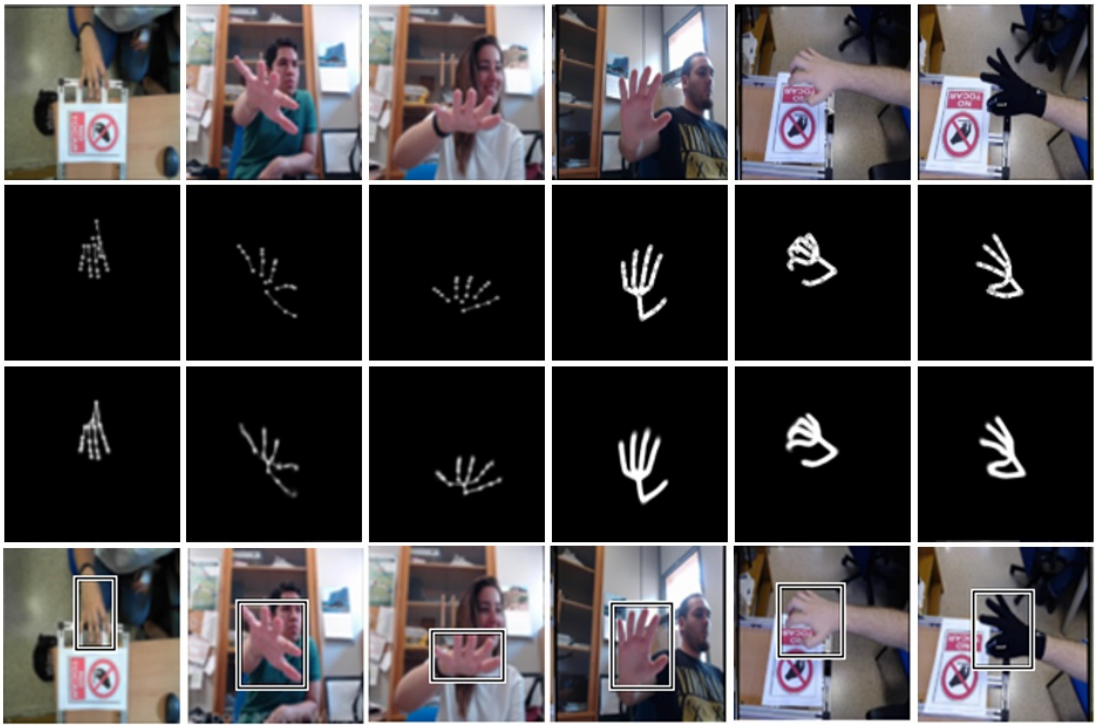

Figure 5: The results of the skeleton estimation and the bounding box localization on the LSMV dataset [18] using thin and thick skeleton representations. The rows from top to down show: the input image, the ground truth skeleton, the predicted skeleton, and the obtained bounding boxes.

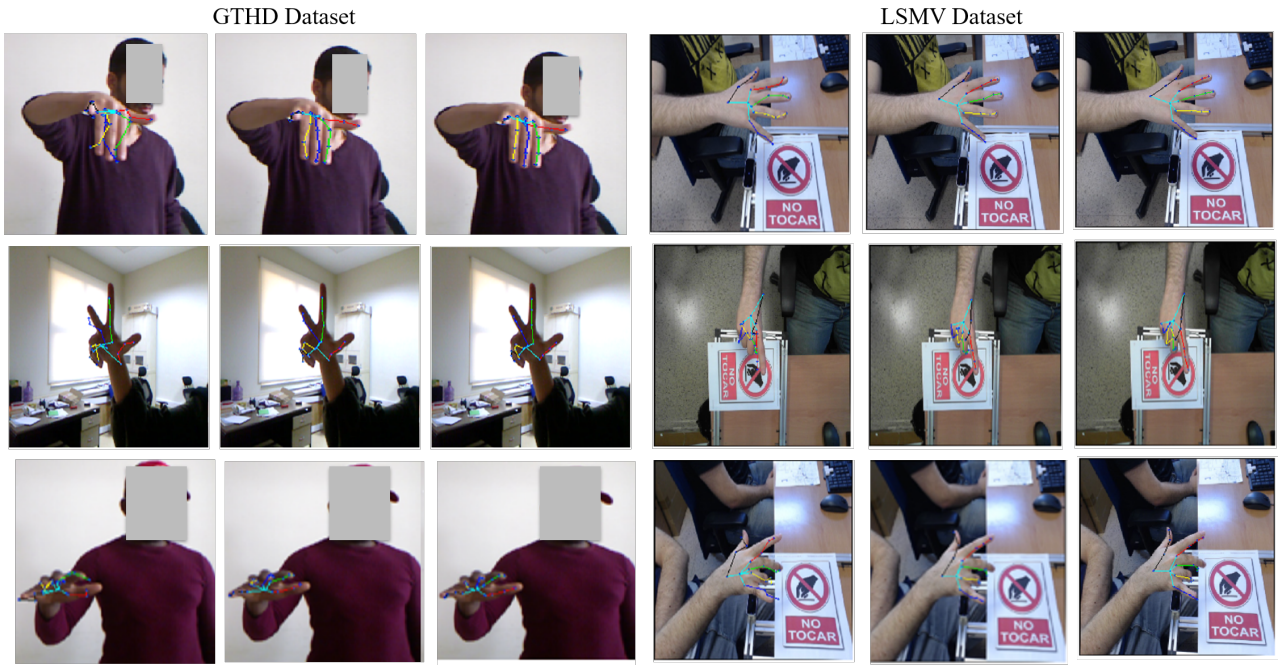

Figure 6: Qualitative results for 2D hand pose estimation on GTHD and LSMV datasets. The columns from left to right in each image show: the direct regression proposed in [18], our proposed skeleton aware multi-scale heatmaps regression and the ground truth joints.

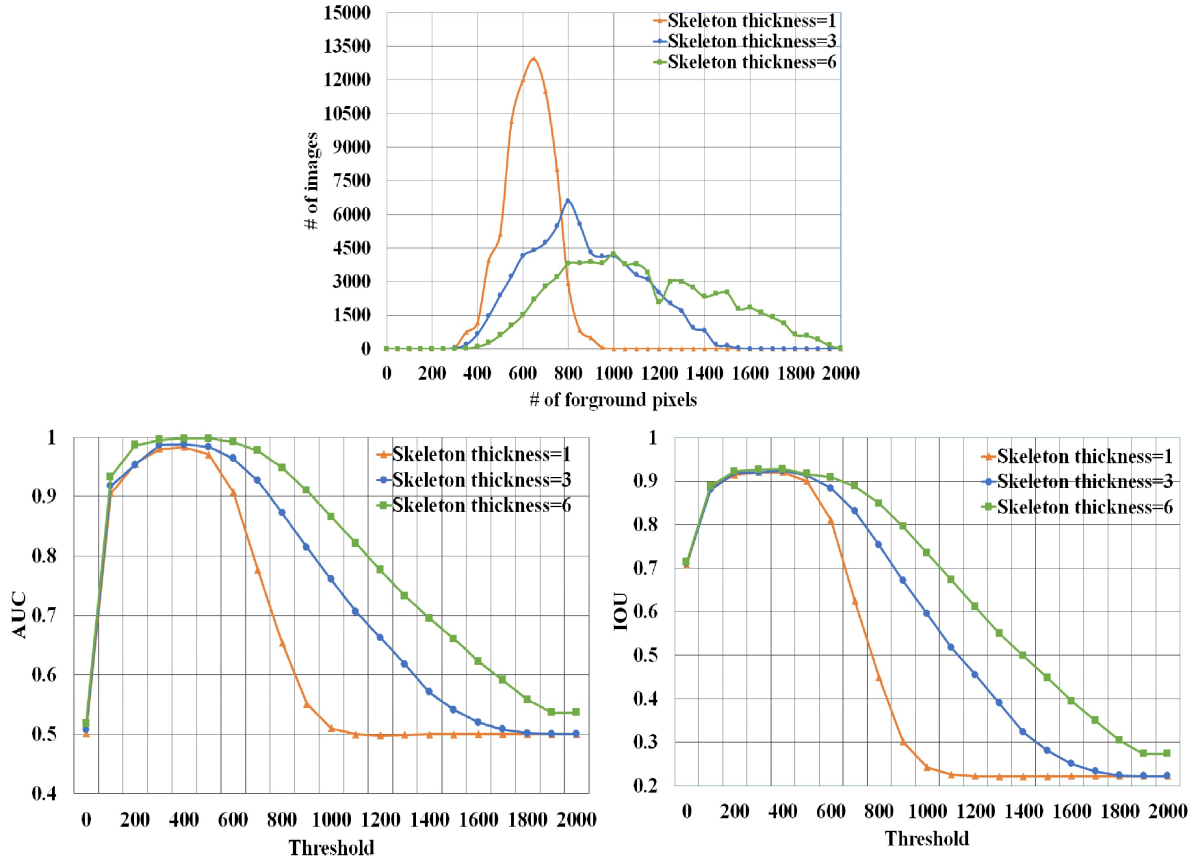

Figure 7: The impact of the threshold selection on the performances.

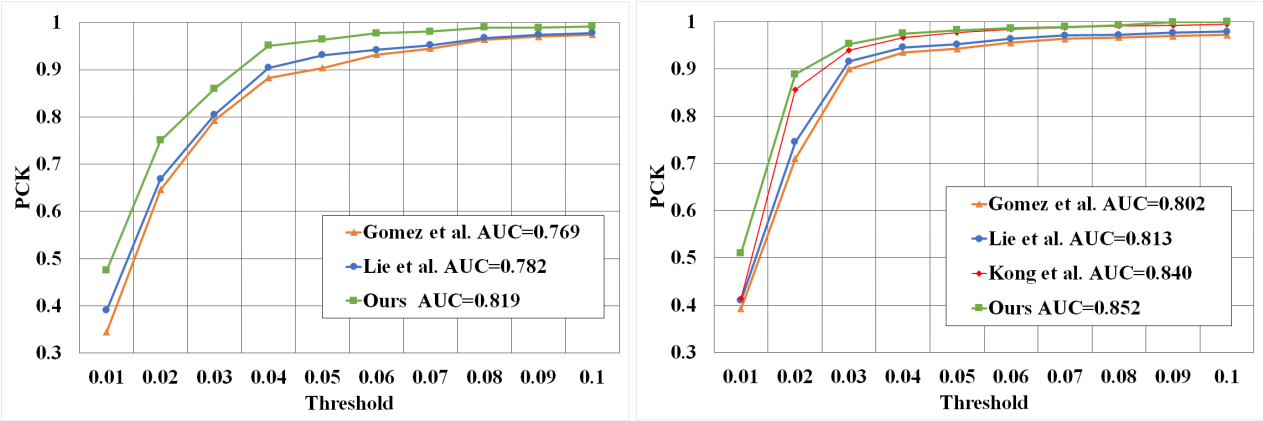

Figure 8: Quantitative comparison of the proposed 2D hand pose estimation with the other methods [18, 41, 24] using PCK metric. Left for GTHD and right for LSMV.

hand bounding boxes and perform our proposed 2D hand pose estimation method (Figure 5). Also, we report *IOU* in Table 2 showing that the proposed method outperforms Faster-RCNN [8].

To show the influence of the hand detection step on the bounding box localization performance, we record AUC and IOU metrics for different thresholds on the GTHD dataset. Figure 7 shows that the performance of the bounding box localization is strongly related to skeleton detection. Also, the thickest skeleton representation seems to be more robust to the noise. It outperforms the other representations and achieves a higher performance (0.99 in AUC and 0.92 in IOU).

4.3 Pose estimation results

The proposed method can robustly estimate the 2D hand pose even in the cases of complex poses and cluttered images. Figure 6 shows some randomly selected test images on LSMV and GTHD datasets.

We compare the proposed pose estimation approach against three-deep learning-based methods [18, 24, 25] on the LSMV dataset. Our baseline is [18] that uses ResNet-50 architecture [5] to directly regress the 2D joints from RGB images. The other deep-based methods [24, 25] are two of the existing state-of-the-art in 2D hand pose estimation. The proposed skeleton-aware multi-scale heatmaps regression method outperforms [18, 24, 25] since it learns the joint location from many resolutions. It reports the highest PCK across all the thresholds (Table 3).

To further demonstrate the effectiveness of the proposed approach, we conduct additional experiments on the GTHD dataset. In the first one, we perform two state-of-the-art methods [18, 41]. The second experiment applies single-scale heatmap regression using UNet architecture [23] on 128×128 resolution images. The third experiment performs our multi-scale heatmaps regression without the skeleton information. In the last experiment, we perform our skeleton aware multi-stage heatmap regression architecture shown in Figure 2. We can see from Figure 8 that our method achieves a high PCK score (0.98) with a small threshold in LSMV and GTHD datasets. Furthermore, the hand skeleton representation improves the proposed multi-scale heatmaps regression method since it constrains the 2D pose estimation task (Table 4).

Estimating the 2D hand pose using the single-scale heatmaps regression outperforms the direct regression since the detected heatmaps help CNNs to learn better the joint locations and converge faster (Figure 8). Finally, our proposed method for 2D hand pose estimation provides more improvement for our dataset since it has more complex poses, face occlusion cases, and lighting condition variations (Figure 8 and Table 4).

5 Conclusion

In this work, we propose a new learning-based method for 2D hand pose estimation. It performs multi-scale heatmaps regression and uses the hand skeleton as additional information to constrain the regression problem. It provides better results compared with the direct regression and single-scale heatmaps regression. Also, we present a new method for hand bounding box localization that first estimates the hand skeleton and then extracts the bounding box. This approach provides accurate results since it learns more information from the skeleton. Furthermore, we introduce a new RGB hand pose dataset that can use both for hand detection and 2D pose estimation tasks.

For future work, we plan to exploit our 2D hand pose estimation method to improve the 3D hand pose estimation from an RGB image. Also, we plan to incorporate other constraints that can restrict the hand pose estimation problem.

Threshold of PCK	0.01	0.02	0.03	0.04	0.05	0.06	meanPCK
Gomez et al [18]	39.27	71.12	90.43	93.56	94.38	95.69	80.74
Kong et al [24]	41.38	85.67	93.96	96.61	97.77	98.42	85.63
Kong et al [25]	41.27	85.89	93.82	96.43	97.61	98.29	85.56
Ours	**51.02**	**88.91**	**95.30**	**97.54**	**98.27**	**98.63**	**88.27**

Table 3: Comparison with the state-of-the-art methods on the LSMV datasets with the PCK metric.

Methods	GTHD	LSMV
Gomez et al [18]	13.20	10.00
Lie et al. [41]	6.25	8.05
Single-scale [23]	7.33	5.87
Ours w/o skeleton	5.89	4.95
Ours	**5.51**	**4.67**

Table 4: Comparison with the state-of-the-art methods on GTHD and LSMV datasets with Mean pixel errors.

References

[1] El-Sawah, A., Georganas, N. D., & Petriu, E. M. (2008). A prototype for 3-D hand tracking and posture estimation. IEEE Transactions on Instrumentation and Measurement, 57(8), 1627-1636. https://doi.org/10.1109/TIM.2008.925725

[2] Chen, T. Y., Wu, M. Y., Hsieh, Y. H., & Fu, L. C. (2016, December). Deep learning for integrated hand detection and pose estimation. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 615-620). IEEE. https://doi.org/10.1109/icpr.2016.7899702

[3] Samed, S., Ferhat, C., Kevser, S. (2021, Vol 45, No 1). A Generative Model based Adversarial Security of Deep Learning and Linear Classifier Models. Informatica (pp.33-64). https://doi.org/10.31449/inf.v45i1.3234

[4] Biserka, P., Tatjana, P., Natasa, S., Aleksandra, S., Mirjana, K. (2021, Vol 45, No 3). Machine Learning with Remote Sensing Image Data Sets. Informatica (pp.347–344). https://doi.org/10.31449/inf.v45i3.3296

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.1109/cvpr.2016.90

[6] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440). https://doi.org/10.1109/cvpr.2015.7298965

[7] Sina, S., Sara, K. (2020, Vol 44, No 4). Teeth Segmentation of Bitewing X-Ray Images Using Wavelet Transform. Informatica (pp.421–426). https://doi.org/10.31449/inf.v44i4.2774

[8] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 91-99. https://doi.org/10.1109/tpami.2016.2577031

[9] Stefan, K., Martin, G., Hristijan, G., Matjaz, G. (2021, Vol 45, No 2). Analysis of Deep Transfer Learning Using DeepConvLSTM for Human Activity Recognition from Wearable Sensors. Informatica (pp.289–296). https://doi.org/10.31449/inf.v45i2.3648

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105. https://doi.org/10.1145/3065386

[11] Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics (ToG), 33(5), 1-10. https://doi.org/10.1145/2629500

[12] Spurr, A., Song, J., Park, S., & Hilliges, O. (2018). Cross-modal deep variational hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 89-98). https://doi.org/10.1109/cvpr.2018.00017

[13] Wan, C., Probst, T., Van Gool, L., & Yao, A. (2018). Dense 3d regression for hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5147-5156). https://doi.org/10.1109/cvpr.2018.00540

[14] Zimmermann, C., & Brox, T. (2017). Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE international conference on computer vision (pp. 4903-4911). https://doi.org/10.1109/iccv.2017.525

[15] Spurr, A., Song, J., Park, S., & Hilliges, O. (2018). Cross-modal deep variational hand pose estimation.

In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 89-98). https://doi.org/10.1109/cvpr.2018.00017

[16] Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., & Theobalt, C. (2018). Ganerated hands for real-time 3d hand tracking from monocular rgb. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 49-59). https://doi.org/10.1109/cvpr.2018.00013

[17] Santavas, N., Kansizoglou, I., Bampis, L., Karakasis, E., & Gasteratos, A. (2020). Attention! a lightweight 2d hand pose estimation approach. IEEE Sensors Journal, 21(10), 11488-11496. https://doi.org/10.1109/jsen.2020.3018172

[18] Gomez-Donoso, F., Orts-Escolano, S., & Cazorla, M. (2019). Large-scale multiview 3d hand pose dataset. Image and Vision Computing, 81, 25-33. https://doi.org/10.1016/j.imavis.2018.12.001

[19] Carreira, J., Agrawal, P., Fragkiadaki, K., & Malik, J. (2016). Human pose estimation with iterative error feedback. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4733-4742). https://doi.org/10.1109/cvpr.2016.512

[20] Bulat, A., & Tzimiropoulos, G. (2016, October). Human pose estimation via convolutional part heatmap regression. In European Conference on Computer Vision (pp. 717-732). Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_44

[21] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., & Murphy, K. (2017). Towards accurate multi-person pose estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4903-4911). https://doi.org/10.1109/cvpr.2017.395

[22] Iqbal, U., Molchanov, P., Gall, T. B. J., & Kautz, J. (2018). Hand pose estimation via latent 2.5 d heatmap regression. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 118-134). https://doi.org/10.1007/978-3-030-01252-6_8

[23] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[24] Kong, D., Chen, Y., Ma, H., Yan, X., & Xie, X. (2019). Adaptive graphical model network for 2d handpose estimation. arXiv preprint arXiv:1909.08205.

[25] Kong, D., Ma, H., & Xie, X. (2020). Sia-gcn: A spatial information aware graph neural network with 2d convolutions for hand pose estimation. arXiv preprint arXiv:2009.12473.

[26] Ren, Z., Meng, J., Yuan, J., & Zhang, Z. (2011, November). Robust hand gesture recognition with kinect sensor. In Proceedings of the 19th ACM international conference on Multimedia (pp. 759-760). https://doi.org/10.1145/2072298.2072443

[27] Hammer, J. H., Voit, M., & Beyerer, J. (2016, July). Motion segmentation and appearance change detection based 2D hand tracking. In 2016 19th International Conference on Information Fusion (FUSION) (pp. 1743-1750). IEEE.

[28] Kumar, A., & Zhang, D. (2006). Personal recognition using hand shape and texture. IEEE Transactions on image processing, 15(8), 2454-2461. https://doi.org/10.1109/tip.2006.875214

[29] Ong, E. J., & Bowden, R. (2004, May). A boosted classifier tree for hand shape detection. In Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings. (pp. 889-894). IEEE. https://doi.org/10.1109/afgr.2004.1301646

[30] Liu, Z., Chai, X., Liu, Z., & Chen, X. (2017). Continuous gesture recognition with hand-oriented spatiotemporal feature. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 3056-3064). https://doi.org/10.1109/iccvw.2017.361

[31] Hoang Ngan Le, T., Zheng, Y., Zhu, C., Luu, K., & Savvides, M. (2016). Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 46-53). https://doi.org/10.1109/cvprw.2016.13

[32] Garcia-Hernando, G., Yuan, S., Baek, S., & Kim, T. K. (2018). First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 409-419). https://doi.org/10.1109/cvpr.2018.00050

[33] Duan, L., Shen, M., Cui, S., Guo, Z., & Deussen, O. (2018). Estimating 2d multi-hand poses from

single depth images. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops (pp. 0-0). https://doi.org/10.1007/978-3-030-11024-6_17

[34] Wang, Y., Peng, C., & Liu, Y. (2018). Mask-pose cascaded cnn for 2d hand pose estimation from single color image. IEEE Transactions on Circuits and Systems for Video Technology, 29(11), 3258-3268. https://doi.org/10.1109/tcsvt.2018.2879980

[35] Wang, Y., Zhang, B., & Peng, C. (2019). Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization. IEEE transactions on image processing, 29, 2977-2986. https://doi.org/10.1109/tip.2019.2955280

[36] Chen, Y., Ma, H., Kong, D., Yan, X., Wu, J., Fan, W., & Xie, X. (2020). Nonparametric structure regularization machine for 2d hand pose estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 381-390). https://doi.org/10.1109/wacv45572.2020.9093271

[37] Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1145-1153). https://doi.org/10.1109/cvpr.2017.494

[38] Pisharady, P. K., Vadakkepat, P., & Poh, L. A. (2014). Hand posture and face recognition using fuzzy-rough approach. In Computational Intelligence in Multi-Feature Visual Pattern Recognition (pp. 63-80). Springer, Singapore. https://doi.org/10.1007/978-981-287-056-8_5

[39] Potter, L. E., Araullo, J., & Carter, L. (2013, November). The leap motion controller: a view on sign language. In Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration (pp. 175-178). https://doi.org/10.1145/2541016.2541072

[40] Beardsley, P., Murray, D., & Zisserman, A. (1992, May). Camera calibration using multiple images. In European Conference on Computer Vision (pp. 312-320). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-55426-2_36

[41] Li, S., & Chan, A. B. (2014, November). 3d human pose estimation from monocular images with deep convolutional neural network. In Asian Conference on Computer Vision (pp. 332-347). Springer, Cham. https://doi.org/10.1007/978-3-319-16808-1_23