People as Points:
Robust Human-Robot Interactions - Real-Time Tracking and Detection for Face and Hand

Than Le[1,2], Long H. Nguyen[1], Tai T. L. Nguyen[1] and Trung A. N. Hoang[1]

***Abstract*— The pandemic such as SARS or Covid-19 have impact to safety human life while it still loss many people deaded and currently continue losing many human-life all around the world. Hence, we need solutions to reduce the impact of these pandemic. In this paper, we propose the people of points for robust human-robot interaction methodologies in order to increasing the interaction. Namely, we use the deep learning method to extract the learning features.**

I. INTRODUCTION

Era of resolution of Robust Human-Machine Interaction is currently the challenge where there are many pandemics such as Severe acute respiratory syndrome (SARS) or Coronavirus disease 2019 (COVID-19). The impact of Covid-19 We are currently development the robust human-robot interaction. There are many applications potentially in human-robot interaction, such as detect action real-time (or video, image, etc) activities negative impact into children.

There are currently limited performance of speed and accuracy [1]. Recently, its approach can carry out the flexible or mobility based on human-robot interaction. Specifically, we use the transfer learning to make the robust deep learning and explainable [9], [1], [2], [4], [6] based on object detection. We also try to implement the hand data set for tracking system.

II. APPROACH: PEOPLE AS POINTS

Consider input image $I(x) \in \mathcal{R}^{W \times H \times 3}$, where W, H as size of width and height accordingly, and x as number of input image. Robot will be a key potential replacing human-human interaction in near future. Where Recent research [1] The bounding box of human k according the category c_k can be formalized: $[x^1_{(k)}, y^1_{(k)}, x^2_{(k)}, y^2_{(k)}]$, then its center point is lie at p_k to be determined:

$$p_k = \left[\frac{x^1_{(k)}, x^2_{(k)}}{2} \quad \frac{y^1_{(k)}, y^2_{(k)}}{2} \right] \tag{1}$$

After that it's used the keypoint estimation $\widehat{Y}$ to predict the people center points. Moreover, we need to recursive the people size s_k:

*This work was supported by Thu Dau Mot University and Ton Duc Thang University organizations

[1] is with Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh city, Vietnam. Email: than.ld@ieee.org

[2] is with Faculty of Engineering and Technology, Thu Dau Mot University, Binh Duong Privance, Vietnam. Email: ledinhthan@tdtm.edu.vn

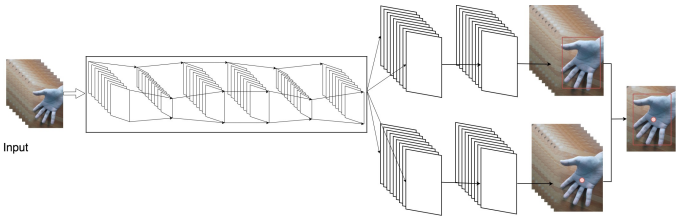

Fig. 1: People as Points: Architecture Frame

$$s_k = \left[x^2_{(k)} - x^1_{(k)} \quad y^2_{(k)} - y^1_{(k)} \right] \tag{2}$$

We use a single size prediction $\widehat{S} \in \mathcal{R}^{\frac{w}{R} \times \frac{H}{R}}$

For feature representation, we use the Deep Layer Aggregation Network [8] to extract the feature selections. It's used to add the deconvolutional and deformable convolutional neural network.

Loss function is one of most important factors, where we can defined an efficiency way to learn the parameters. In this case, we use $\mathcal{L}1$ loss function at the center point

$$\mathcal{L}_{size} = \frac{1}{N} \sum_{k=1}^{N} |\widehat{S}_{pk} - s_k| \tag{3}$$

The training objective formalized (illustrated Figure 4):

$$\mathcal{L}_{TotalLoss} = \mathcal{L}_k + \lambda_{size}\mathcal{L}_{size} + \lambda_{off}\mathcal{L}_{off} \tag{4}$$

Where $\mathcal{L}_{size}$ as $\mathcal{L}1$ Norm Offset Loss show by:

$$\mathcal{L}_{off} = \frac{1}{N} \sum_{p} |\widehat{O}_{\tilde{p}} - (\frac{p}{R} - \tilde{p})| \tag{5}$$

and $\mathcal{L}_k$ formalized:

$$\mathcal{L}_k = \frac{-1}{N} \sum_{xyz} \begin{cases} (1 - \widehat{Y}_{xyz})^{\alpha} log(\widehat{Y}_{xyz}) & \text{if } Y_{xyz} = 1 \\ (1 - Y^{\beta}_{xyz})(\widehat{Y}_{xyz})^{\alpha} log(1 - \widehat{Y}_{xyz}) & \text{otherwise} \end{cases} \tag{6}$$

For pose estimation we use key-point skeleton recognition by using transfer learning from objects of points [3], [4]. It's used Deep Layer Aggregation [7] to train. We use PC high performance to training it according to configuration: 32GB RAM, GTX NVIDIA 2070, i7 to do experiments.

To test the hand tracking we use the our

Fig. 2: Hand Pose Tracking and Detection

Fig. 3: Real-time Face Detection: LEFT - Image Face Detection; CENTER - Image Multi-Face Detection; RIGHT - Real-Time Camera Detection

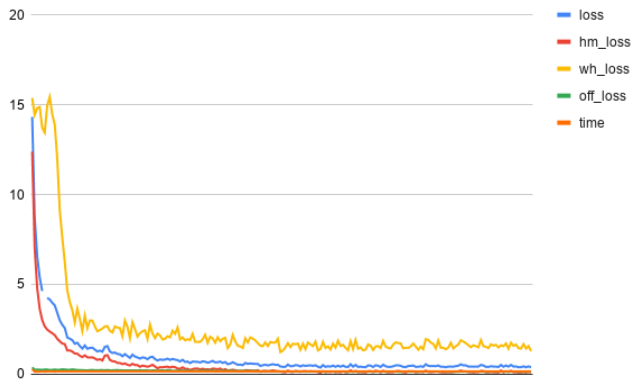

Fig. 4: Training Parameters

A. Real-Time Face Detection

(illustrated Figure 3)Face Detection for Human-Robot Interaction is setups the same with experience in [1], [2]. It's tested as one class for only face detection.

We use crashed images by ourselves for face detection. We use the Intersection of Union (IOU) to do two regular tasks in object detection, segmentation, and tracking.

B. Hand Pose Estimation

We use CMU Panoptic Dataset [8] for Hand Pose Estimation, the network which we use is Deep Layer Aggregation(DLA) [7] instead of ResNet [1], [6], [9] for image classification since DLA achieves better performance with fewer parameters. The average accuracy after 30 epochs of training is 30(percent), only when the hand is at far distance, if the hand is too close, the model can't predict since there's no data relate to close hand pose detection. The input shape for DLA network is 128 x 128 and with the output shape of 1 after the network. The result shows at Figure 2

III. CONCLUSIONS

In conclusion, we propose the robust human-robot interaction based on transforming the multiple objects to multiple points. It illustrates enough for robust do simple tasks in learning and adaptive in localization and prediction in object detection and segmentation based on bounding boxes and pose of human hand estimation.

ACKNOWLEDGMENT

We would like to thanks Thu Dau Mot University and Ton Duc Thang University supported grants for our research activities.

REFERENCES

[1] T. D. Le, D. T. Huynh and H. V. Pham, "Efficient Human-Robot Interaction using Deep Learning with Mask R-CNN: Detection, Recognition, Tracking and Segmentation," 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, 2018, pp. 162-167, doi: 10.1109/ICARCV.2018.8581081.

[2] Q. H. Nguyen, T. N. P. Tran, D. D. Huynh, A. T. Le and T. D. Le, "Real-Time Localization and Tracking System with Multiple-Angle Views for Human Robot Interaction," 2017 First IEEE International Conference on Robotic Computing (IRC), Taichung, 2017, pp. 316-319, doi: 10.1109/IRC.2017.53.

[3] Zhou, Xingyi and Wang, Dequan and Krähenbühl, Philipp, "Objects as Points". arXiv preprint arXiv:1904.07850, 2019.

[4] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision, pages 734–750, 2018.

[5] Xingyi Zhou, Vladlen Koltun, Philipp Krähenbühl, *Tracking Objects as Points*, arXiv technical report (arXiv 2004.01177)

[6] Le, Than (2020): Efficient Post-Contour Correctness in Object Detection and Segmentation. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.11603487.v1

[7] Yu, Fisher et al. "Deep Layer Aggregation." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 2403-2412.

[8] Simon, Tomas et al. "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 4645-4653.

[9] Le, T.D. and Pham, H.V. (2020). Intelligent Data Analysis. In Intelligent Data Analysis (eds D. Gupta, S. Bhattacharyya, A. Khanna and K. Sagar). doi:10.1002/9781119544487.ch5