Gesture Controlled UI/UX Navigation using Neural Networks

Rajat Bisht
Khoury College, Northeastern University
bisht.r@northeastern.edu

1 Overview

With emergence of new human computer interaction technologies in form of Virtual Reality(VR) and Augmented Reality(AR), hand gesture recognition using computer vision has gained significant importance. Hand gestures play a crucial role in navigation within these meta-realities, necessitating faster, seamless, and highly accurate mapping and classification of gestures. This project is (very loosely) based on sixth sense technology[1] developed by Pranav Mistry in association with MIT Media labs. Building upon the idea, this project proposes an effort in generating optimal hand-gesture recognition software using webcam for controlling mouse pointers using 'Hand-Landmarks First' approach for gesture recognition as a proof of concept expandable to AR/VR environments.

This project will take some inspiration from Google's Mediapipe project for hand recognition[2] and add a hand gesture detection network for categorization of gestures for mapping with mouse pointer operations. The implementation process for hand detection from images will be divided into two stages: first, the localization and identification of the hand/palm region of interest (RoI), and second, the extraction and embedding of hand landmark information within the detected hand object. Based on the knowledge accumulated form survey paper by Kaur et.al [7] on object detection strategies and bottlenecks, the following strategy is proposed. Implementation of hand/palm detection will be done by using SSD[3] for extracting anchors, followed by a feature pyramid networks[4] for feature extraction and considering hard example dataset's impact for a small objects like hand/palm as compared to soft example like wall/background [6], the project will lastly employ focal loss[5] for better generalization of learning over the trained deep learning network. Second step will be to pipeline extracted hand RoI to the landmarks detection neural net which will use convolutional neural network for determining landmarks and hand presence. Since, the scope of this project is to mimic mouse pointer actions, left or right handedness categorization of hand will be omitted as design choice, but can be incorporated as future work. Finally, a feed-forward neural net will be implemented in application layer for categorizing hand gestures and synchronizing with mouse pointer actions.

2 Datasets

For training, evaluation and testing of this project, datasets used can be classified as still images and video frames. Since learning curve and generality of learned knowledge in a deep neural net is highly contingent on incorporation of both hard and soft examples while leaning heavily upon heavy examples[6], the dataset used are based on this knowledge rationale. Therefore, still image dataset can be further understood as soft examples using HaGRID dataset from Kaggle[8] and hard examples using hand database from CMU[9]. Video frame datasets are incorporated for training the application to learn movement tracking and swiping actions, for which Jester dataset[10] and gesture recognition dataset[11] will be used. Datasets will be cleaned and tailored to fit the specific scope of this application, aiming to optimize training time and improve performance.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

References

[1] Mistry, P., & Maes, P. (2009). SixthSense: a wearable gestural interface. In *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation* (pp. 85-85).

[2] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.

[3] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.

[4] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[5] Ross, T. Y., & Dollár, G. K. H. P. (2017, July). Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2980-2988).

[6] Kishida, I., & Nakayama, H. (2019). Empirical study of easy and hard examples in cnn training. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part IV 26* (pp. 179-188). Springer International Publishing.

[7] Kaur, R., & Singh, S. (2023). A comprehensive review of object detection with deep learning. *Digital Signal Processing, 132,* 103812.

[8] Alexander Kapitanov, Andrey Makhliarchuk, Karina Kvanchiani, & Aleksandr Nagaev. *HaGRID - HAnd Gesture Recognition Image Dataset.* (n.d.). Www.kaggle.com.

`https://www.kaggle.com/datasets/kapitanov/hagrid/data`

[9] *Hand Database - CMU Panoptic Dataset*. (2017). Cmu.edu.

`http://domedb.perception.cs.cmu.edu/handdb.html`

[10] *Papers with Code - Jester (Gesture Recognition) Dataset*. (2022). Paperswithcode.com.

`https://paperswithcode.com/dataset/jester-gesture-recognition`

[11] Venkata, A. (2024). *Gesture Recognition Dataset*. Kaggle.com.

`https://www.kaggle.com/datasets/abhishek14398/gesture-recognition-dataset/data`