Team: Coxozaurus

Real-Time Gesture Recognition System for Smart Device Control and Sign Language Translation

Project scope

Gesture recognition can be used to facilitate human-computer interaction by offering an alternative way of communication. In this scope, it can be used in environments that must follow strict hygiene standards or present several constraints for manual operation. An alternative would be voice recognition, but this is not applicable in noisy conditions and cannot be used by people with speech disabilities or the deaf and hard-of-hearing community. Another possibility is specialized hardware like Kinect or Leap Motion, which limit widespread adoption. Gesture recognition can also be used for sign language translation, making everyday conversations easier for the struggling population. Edge AI is meant to reduce latency in both applications and address the privacy concerns that come with cloud-based processing, which is what most of the existing solutions use. There is also the need for stable internet connectivity to provide seamless interactions and communication. This is solved by offline solutions that operate on-device.

Literature review

Recurrent 3D convolutional neural networks have been used to enhance hand gesture recognition by Molchanov et al. [1]. They presented how temporal convolutions effectively capture motion dynamics. To address the real-time constraints for such applications, Köpüklü et al. [2] proposed a lightweight network architecture that can achieve competitive accuracy on the Jester dataset, proving that edge devices can be used for smart device control scenarios.

Huang et al. [3] proposed a 3D CNN approach to address the sign interpretation problem and achieved promising results on isolated Chinese Sign Language (CSL) gestures. To address the continuous nature of sign language, Koller et al [4] introduced CNN-HMM models and tested them on the RWTH-PHOENIX-Weather dataset achieving a lower word error rate at the expense of higher computational costs.

To further optimize the existing architectures for deployment on resource-constrained devices, Han et al. [5] proposed deep compression techniques that combine pruning, quantization and Huffman coding to reduce the size of neural networks by 35-49x without a significant loss in accuracy. Another approach was proposed by Howard et al. [6] and consisted of using depthwise separable convolutions to reduce the number of parameters by 75% which made mobile and embedded vision applications feasible.

Furthermore, real-time hand landmark detection was addressed by Bazarevsky et al. [7] by implementing MediaPipe hands, a lightweight pipeline that achieves 21-keypoint hand tracking at 30+ FPS on mobile devices using a two-stage detector-tracker architecture. Their work was extended by Zhang et al. (2020), which introduced MediaPipe Holistic that is capable of tracking hands, pose and face, enabling capturing the full expressiveness of sign language.

References

[1] Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., & Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4207-4215.

[2] Köpüklü, O., Gunduz, A., Kose, N., & Rigoll, G. (2019). Real-time hand gesture detection and classification using convolutional neural networks. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1-8.

[3] Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2257-2264.

[4] Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding, 141, 108-125.

[5] Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. International Conference on Learning Representations (ICLR).

[6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[7] Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M. (2020). BlazePose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204.

[8] 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*.