AirGestAR: Leveraging Deep Learning for Complex Hand Gestural Interaction with Frugal AR Devices

Varun Jain, Ramakrishna Perla, and Ramya Hebbalaguppe
Smart Machines R&D Group, TCS Research, India

TATA CONSULTANCY SERVICES

ABSTRACT

- Hand gestures provide a natural and an intuitive way of user interaction in AR/VR applications.
- However, most popular devices such as the Google Cardboard and Wearality still employ only primitive modes of interaction such as the magnetic trigger and have limited user-input capability.
- Hololens, Magic Leap, and Meta Glasses capable of instinctual gestures but are expensive and use proprietary hardware.
- We propose a deep learning framework for recognizing **complex 3-dimensional marker-less temporal gestures** (*Bloom, Click, Zoom-In, Zoom-Out*).
- Capable of real-time performance and employs **monocular camera** input from a single smartphone.

DATASET

- Lack of large-scale datasets is a major factor hampering the advent of deep learning in the task of hand gesture recognition.
- We created a dataset of Bloom, Click, Zoom-In, Zoom-Out gestures captured in egocentric view.
- It has 480 videos: 100 videos per gesture for training and 20 videos per gesture in the testing set.
- The videos were recorded using an Android device on a Google Cardboard at a resolution of 320*240, and at 30 FPS.
- 6 subjects with varying skin color were involved in the data collection, ages ranging from 21 to 55.
- The same can be found by scanning the QR code above.

CONCLUSION

- Presented a novel approach for marker-less temporal 3D hand gesture recognition in ego-centric videos.
- Works with just RGB data, without depth information.
- Can enable wider reach of frugal devices for AR applications.
- Also published the gesture dataset used in training and testing of the LSTM network.
- Demonstrated the performance of our proposed approach under realistic conditions and also reported turn-around-time and accuracy of gesture recognition.

PROPOSED FRAMEWORK

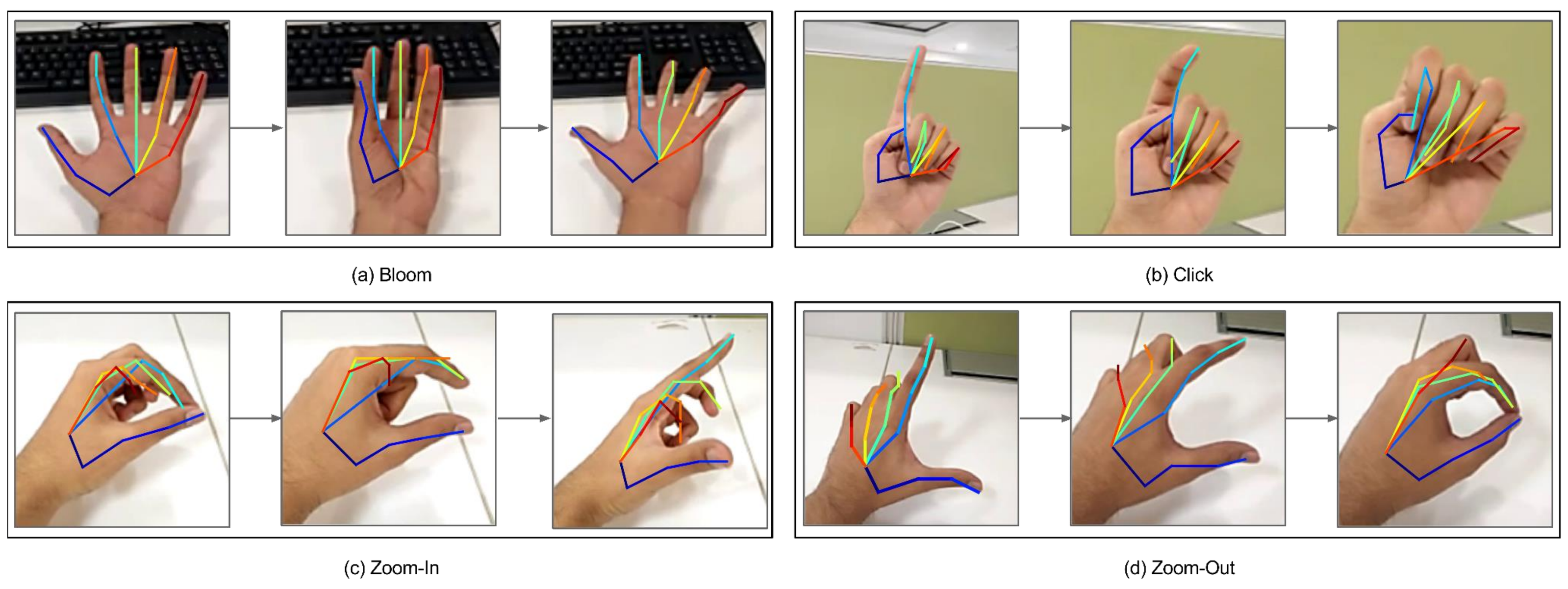
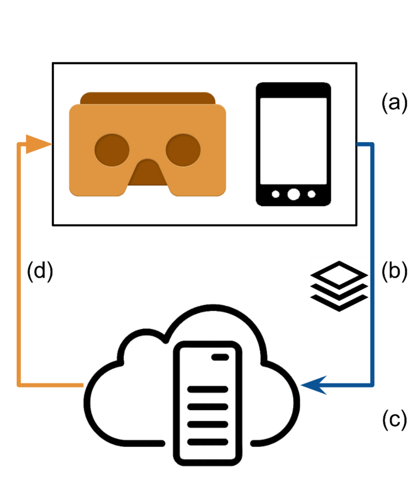


Figure 1. 3D gestures supported. (a) Bloom: for a menu display operation, (b) Click: for a select/hold operation, (c) Zoom-In: for zooming into a scene, and (d) Zoom-Out: for zooming out of a scene.
(a) and (b) have been inspired by the Microsoft Hololens. The 21 key-points as detected by hand pose estimation network are shown as an overlay on input images.



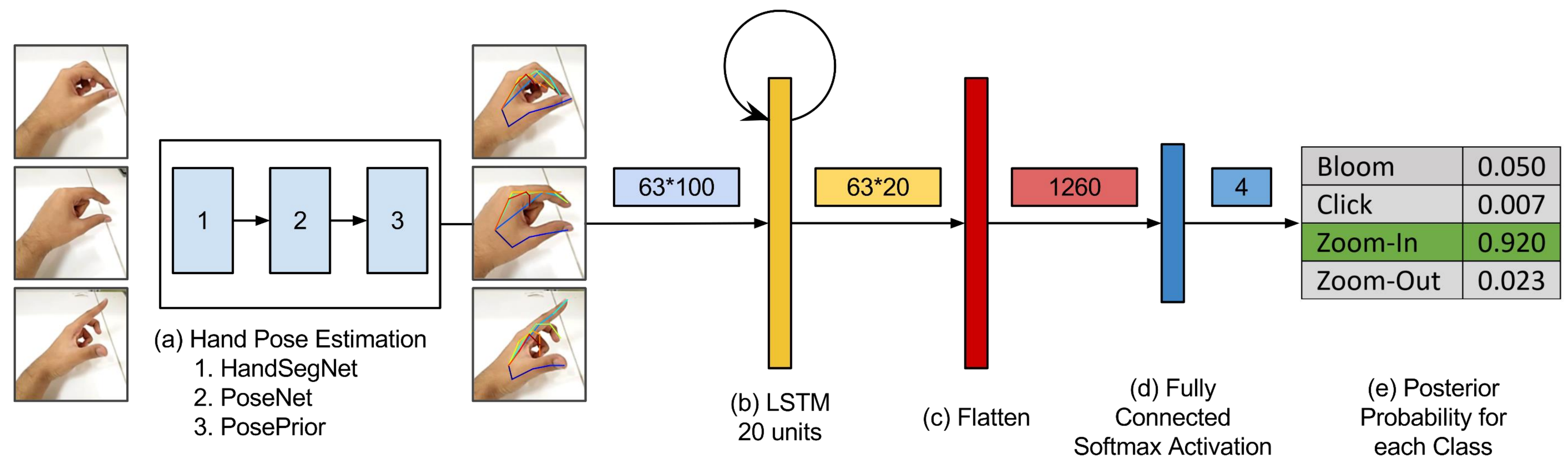


Figure 2 (L). System architecture: each smartphone sends down-scaled video frames to a server which runs the AirGestAR gesture recognition framework. The result is then communicated back to the device.

Figure 2 (R). The proposed AirGestAR neural network framework for gesture recognition (Note output size after each layer).
(a) The input video frames are first fed to hand pose estimation model proposed by Zimmermann and Brox [1].
(b) to (e) shows our proposed LSTM architecture. An LSTM layer consisting of 20 LSTM cells takes input 3 coordinates for each of the 21 key-points and 100 frames at a time.
At the end, a fully connected layer outputs posterior probabilities corresponding to each of the 4 gestures.

EVALUATION AND FUTURE WORK

- Table 1 shows a confusion matrix for the experiments. We achieved an accuracy of **93.75%** with 2 cases of misclassification out of 80. The presence of a gesture is detected when the probability of a gesture is more than 70%. Otherwise, no gesture-detection is reported.
- The average response time of the proposed framework is found to be **0.8s** on GPU configuration.

Predicted / True	Bloom	Click	Zoom-In	Zoom-Out	Unclassified
Bloom	**20**	0	0	0	0
Click	0	**16**	0	2	2
Zoom-In	0	0	**20**	0	0
Zoom-Out	0	0	0	**19**	1

Table 1. Confusion matrix for the proposed framework yielding an accuracy of 93.75% with 2 cases of misclassification out of 80.

- Our LSTM-only architecture is capable of delivering frame rates of up to 107 on GPU implementation. However, the hand pose estimation network works at 9 FPS.
- On deeper analysis, we find that the click and the zoom-out gestures are highly correlated since both involve movement of index finger towards the palm. The bloom and the zoom-out gestures fare well due to their unique nature and the fact that the bloom gesture exploits most of the 21 key points being detected.
- We would like to explore the possibility of tuning the model with more synthetic images that are a better representation of gestures commonly used in FPV applications.
- We would also like to look at the possibility of exploiting powerful System on Chip (SoC) architectures to port the Tensorflow models to an embedded board such as the ODROID 6 board. This has been successfully explored by Baraldi et al. [2]

Contact

Varun Jain
Email: varun14170@iiitd.ac.in
Website: varunj.github.io

Ramakrishna Perla
Email: r.perla@tcs.com

Ramya Hebbalaguppe
Email: ramya.hebbalaguppe@tcs.com

References

[1] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single RGB images. *CoRR*, abs/1705.01389, 2017.

[2] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 688–693, 2014.