CSE 591 Project Proposal: Stroke Recognition for Gesture Interface

Duo Lu
ASUID: 1205136916
Arizona State University
duolu@asu.edu

Pak Lun (Kevin) Ding
ASUID: 1206493882
Arizona State University
kevinding@asu.edu

***Abstract*—Virtual Reality (VR) headsets and wearable computers equipped with gesture user interface are considered the future of personal computing platforms. However, it is usually not convenient or even impractical to present a keyboard or touchscreen to enter text input. Although voice input is gaining popularity, in many cases it is inappropriate to speak loud. In this project, we propose a system to enable gesture input by recognizing strokes from a piece of in-air-handwriting or hand gesture, obtained by finger motion capturing devices. This system utilize convolutional neural network (CNN) to extract low level motion features from finger movement signals and classify a sequence of motions to a sequence of predefined strokes. Such stroke sequence can be further utilized to recognize handwriting, interpret gesture control command or authenticate a user (*i.e.*, gesture passcode of strokes). The research outcome of our project will be able to build the foundation to enable in-air gesture or handwriting based input for VR and wearable applications.**

I. Proposal

In Virtual Reality (VR) headsets, wearable computers, and other pervasive computing applications, user input is necessary to interact with the computer. In such an environment presenting a keyboard or touchscreen is usually impractical, while gesture based input interface would be efficient and favorable and motion tracking technology is gaining existing systems. For example, popular VR systems such as Oculus Rift, HTC Vive, and Sony PSVR all utilize handheld remote controllers to track users' hands for game playing. Future gesture interface will use hand-wear devices [1] or wearable cameras [2] to capture user's hand motions. However, existing hand motion and in-air-handwriting based input solutions suffer from constrained gesture patterns, limited motion capture resolution, and low recognition rate [3], [4], [5], [6], [7]. In this project, we propose to employ convolutional neural network to classify hand and finger motion to sequence of strokes, and use the strokes as gesture input. We are especially interested in using the stroke sequence as a finger motion passcode to authenticate a user (we define such passcode as FMCode). For example, it would be convenient for a user to write the string "FMCODE" as a password when he or she is required to be authenticated when logging in a virtual website in a VR game.

II. Methodology

Our system contains a finger motion tracking client and an authentication server. The user's finger motion is captured by a wearable device connected to the client machine, and the captured motion signal is sent to the server, where it is processed by the convolutional neural network to obtain a stroke sequence. This sequence is used as the FMCode to authenticate a user through matching with a template generated in the same way at registration stage. The architecture is shown as Fig. 1.

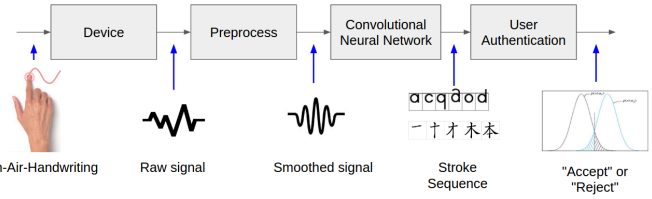

Fig. 1. System architecture.

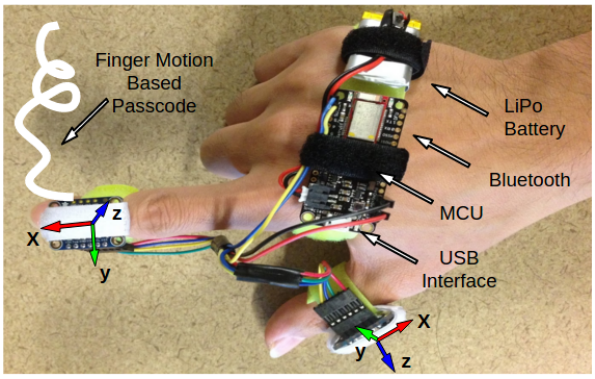

Fig. 2. Finger motion tracking hand band.

We made a prototype of finger motion capture wearable device as a hand band, shown in Fig. 2. (Also there is a left hand version.) It uses an Arduino compatible microcontroller with USB serial port interface, a Bluetooth LE module, a small LiPo battery, and two Inertial Measurement Unit (IMU) chips. Each IMU has a triaxial accelerometer ($\pm 4g$, 14 bit resolution), a gyroscope ($\pm 2000dps$, 16 bit resolution), and a magnetometer ($\pm 1300\mu T$, 13 bit resolution). One IMU is placed on index fingertip and the other is placed on thumb tip. The microcontroller reads out sensor data sampled at 50Hz from both IMUs and sends them either wirelessly or through USB cable (high sampling rate is not necessary because

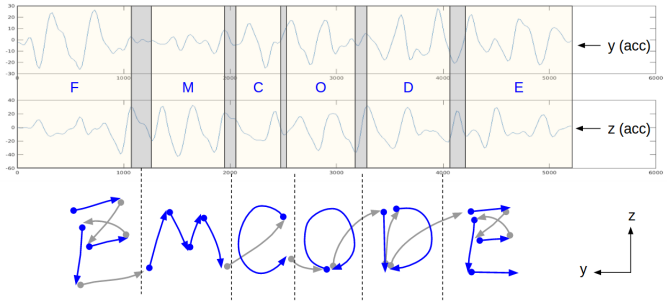

Fig. 3. An example of stroke sequence of in-air-handwriting.

high frequency components over 10Hz is filtered away in the preprocessing stage detailed in the next section). It should be noted that this hand-wear device is not the only form of such device, and we treat the device as a part of the gesture user interface, not a dedicated device only for user authentication. As long as the gesture user interface is able to track user's fingertip movement without constraints on the motion, and provide us acceleration and angular speed information with enough resolution, our authentication method can be applied. For example, contactless motion tracking systems wearing on the VR headset like Microsoft HoloLens can also be used to obtain the FMCode signal if they provide enough resolution, sample rate and field of view (though we do not have such wearable camera devices). We choose inexpensive wearable sensor because they put virtually no restrictions on the finger motions, which is very important to let the user move in his or her own convention, so that the finger motion can be generated in a stable and repeatable fashion.

We use a convolutional neural network to classify each piece of motion signal of in-air-handwriting to a sequence of predefined strokes. An example of the first author writing "FMCODE" is shown in Fig. 3. Currently our predefined strokes have 5 categories: horizontal stroke, vertical stroke, diagnal stroke, half circle, full circle. According to the direction, each categories have several subcategories. In total there are 20 different types of strokes.

Our neural network design resembles AlexNet [8]. The first few convolutional layers work as feature extractors, and the last few fully connected layers work as classifiers. The input of the network is the signal, which contains multiple channels (such as acceleration in x, y, z direction). Such signals can be regarded as an 1D image with multiple color channels. We will start with a "bags-of-strokes" model, *i.e.,* a piece of in-air-handwriting is just a bags of strokes regardless of their sequence inside a bag, while each bag corresponds to a segment of the signal. In this way, the 20 types of strokes will serve as 20 labels, and a piece of in-air-handwriting can have multiple labels as long as it contains the corresponding stroke. For example, for the string "FMCODE", we can define 6 bags, and each bags should output strokes contained by only one letter. Although the writing speed varies, based on our data, for each second there are on average 2.4 to 4 strokes. The final output of the fully connected layers will be processed with a softmax of 20 outputs for each bag. This network is mainly designed for pretraining of the feature extration layers (*i.e.,* the convolutional layers). Because collecting and labeling long pieces of in-air-handwriting is difficult, for pretraining, we will collect many samples of short pieces of in-air-handwriting, where each of them will contain only 3 to 5 strokes. Later we will extends this simple model to sequence of strokes (*i.e.,* multiple bags in sequence where each bag will roughly contains only one stroke or no stroke) to accommodate long pieces of in-air-handwriting. This is done by modifying the fully connected layers of the network to small bags of locally connected layer. For example, consider the signal is sampled at 50 Hz, if one "bag" spans over 0.2 second, the nodes of the first bag will be connected with all those convolutional layers that take input from the first 10 samples of the signal, and so on. If this 10 samples correspond to a stroke, then the output of this bag layer should be the stroke label. These locally connected bag layers share parameters, just like the convolutional layers. In this way, by replicate the convolutional layers and bag layers, our network is able to process signal of arbitrary length. Clearly the variation of writing speed will influence the recognition capability, but based on our current dataset (collected by ourselves), such variation may be still manageable. We will also consider using Recurrent Neural Network (RNN), especially Long-Short-Term-Memory (LSTM).

The authentication will be based matching the stroke sequence of a login signal with that of a template signal saved at registration stage. Given enough length for the stroke sequences, and given the assumption that each user will write a unique passcode, our system should be able to identify a user efficiently.

REFERENCES

[1] "CyberGrasp," http://www.cyberglovesystems.com/cybergrasp/, accessed: 2017-2-1.

[2] "Microsoft HoloLens," https://www.microsoft.com/microsoft-hololens/en-us, accessed: 2017-2-1.

[3] C. Amma, M. Georgi, and T. Schultz, "Airwriting: a wearable handwriting recognition system," *Personal and ubiquitous computing*, vol. 18, no. 1, pp. 191–203, 2014.

[4] C. Qu, D. Zhang, and J. Tian, "Online kinect handwritten digit recognition based on dynamic time warping and support vector machine," *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE*, vol. 12, no. 1, pp. 413–422, 2015.

[5] D. Moazen, S. A. Sajjadi, and A. Nahapetian, "Airdraw: Leveraging smart watch motion sensors for mobile human computer interactions," in *Consumer Communications & Networking Conference (CCNC), 2016 13th IEEE Annual*. IEEE, 2016, pp. 442–446.

[6] H. Wen, J. Ramos Rojas, and A. K. Dey, "Serendipity: Finger gesture recognition using an off-the-shelf smartwatch," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 3847–3851.

[7] Y. Zhang and C. Harrison, "Tomo: Wearable, low-cost electrical impedance tomography for hand gesture recognition," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 2015, pp. 167–173.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.