Mixed Recognition of Hand and Body Motions Based on Concurrent CNN-BiLSTM

--Final Inspection Report

Supervisor:Song Xuan

Team members:Jiang Yuchen,Chen Jiyuan

1.Abstract

This is a report for the final inspection. Conclusions of the second inspection and work of final stage are concluded in the passage. Furthermore, summary of this term and future work are written at the end of the passage.

2.Conclusions of 2nd Inspection

It's reminded that we are behind schedule and we need to focus on our aim which is to build up a hand motion detection system. We need to build up our deep learning model as soon as possible.

3.Literature Review

1. Sensor in phones data collecting

Data collecting mainly processes by acceleration sensors.There are three kinds of acceleration data which are X,Y and Z axis data.We will collect acceleration data together with timeline.

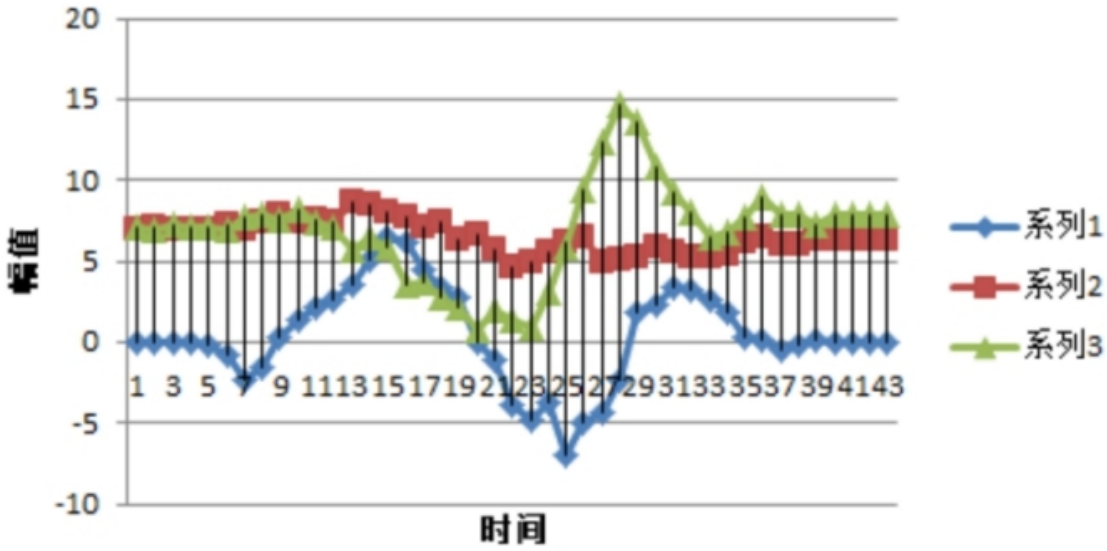

Figure 1. acceleration signal from 3 axises[2]

Also,signals will be different as sensors are bound on different positions like wrist,arm an d ankle.The scientists find that age and gender influence the acceleration data.

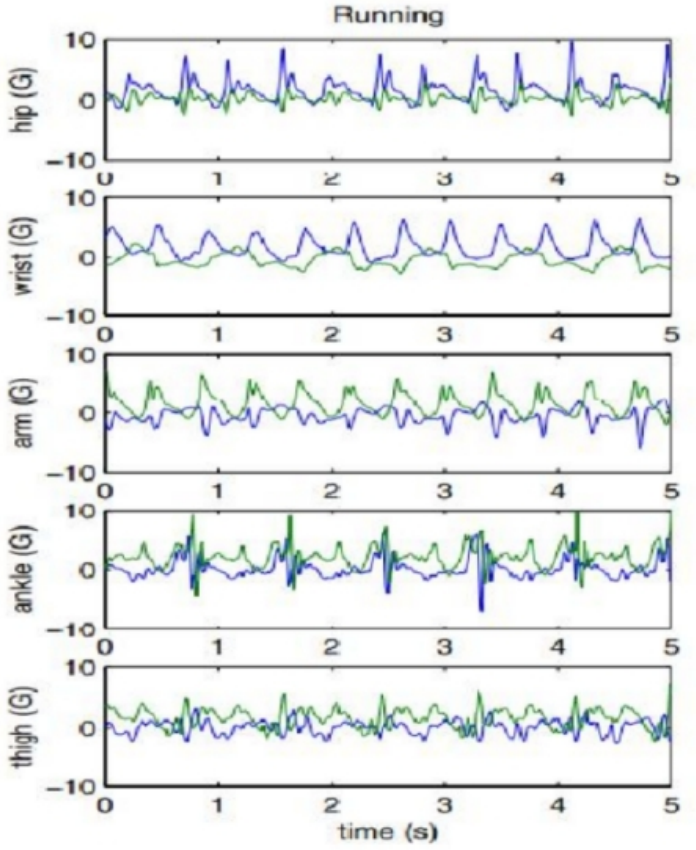

Figure 2 different acceleration signals from different part of body[1]

Besides,we can apply gyroscope in devices to record other data,for example, angular momentum,wh ich will be able to enrich variables so that we can analyze more details about actions and get more accura te result.

2. Data from sensors preprocess.

According to collected data,we will extract some important features for analysis.There are three mai n methods for extracting:time field ,frequency field and time-frequency-mixed field analyze.

The table below shows some important variables.For example,Amplitude area will show the degree of moving amplitude.

特征类型	特征名称
时域特征	平均值
	均方根
	信号幅值面积
	信号幅值向量
	相关系数
	能量
	平均绝对偏差
	时域积分
	姿势方向
频域特征	FFT 系数
	频域熵
	能量密度
	谱系数
时频特征	小波系数

Figure 3 common features[1]

After selecting the features , a data cleaning is needed to denoise the data. In many occasions we jus t use the following method to smooth the data.

$$a_{now} = \frac{a_{i-domain} + a_{i-domain+1} + \ldots + a_i + \ldots + a_{i+domain}}{window - size} \quad (1)$$

Figure4 smooth algorithm[2]

The data before and after the smooth is shown as follows:

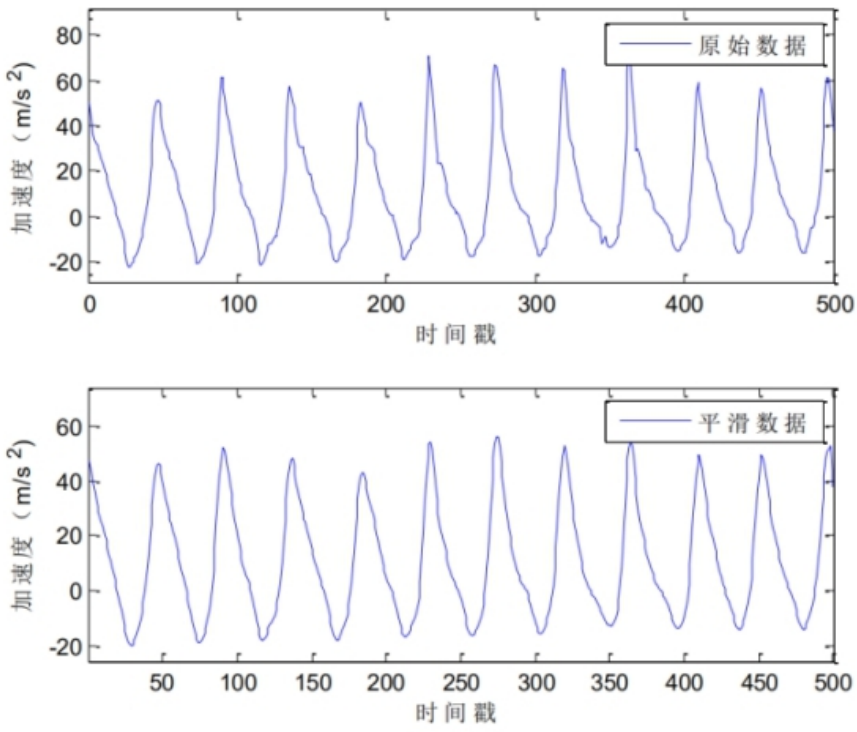

Figure5 data comparison of smooth[2]

Also for the reason that the moving data produced by sensors are related with time, we must know h ow to calculate the distance between two sequences of data which may differ in time slice length. Fo

rtunately, DTW comes to rescue.

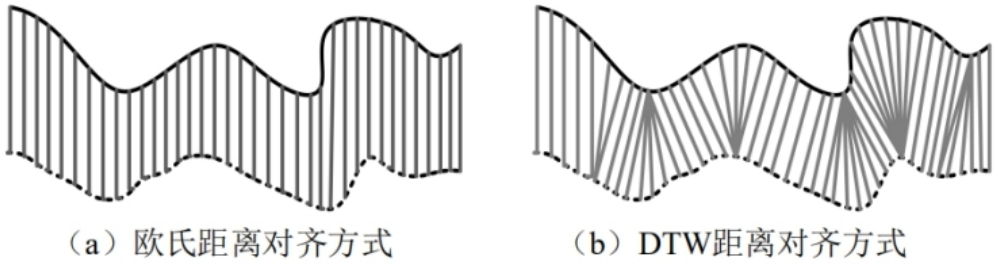

Figure6 DTW algorithm matching[1]

The figure shows that in DTW, time slot are twisted so that the feature points can be matched with the corresponding ones clearly. In this way, we can calculate the distance between any tw o time-
sequence data easily without worrying if the two data are of different length.3. Model analysis Most articles about gesture caption is now based on a famous machine learning model calledS upport Vector Machine (SVM).SVM is originally designed to find a hyperplane by the maxi mum interval learning method, so that the distance between the vectors belonging to different s ets and the hyperplane is maximized.

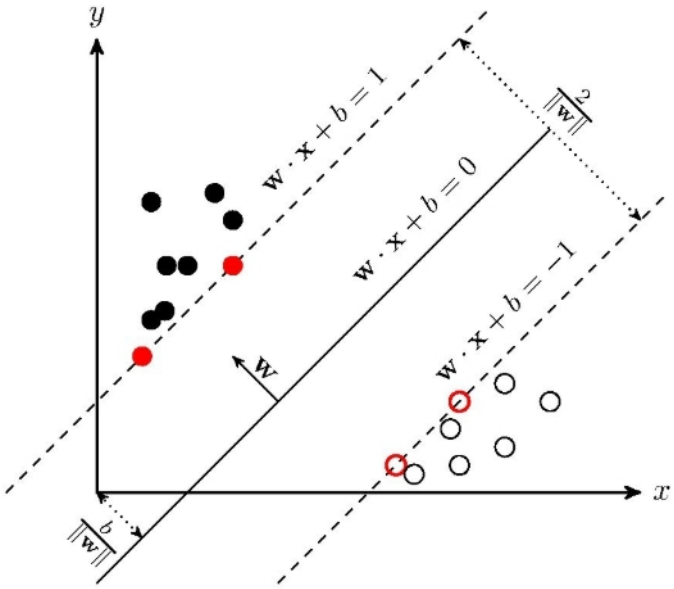

Figure7 SVM finding hyperplane

Even though single SVM can only separate two different data sets. But after applying k different SV M to k data sets, we can build a multivariable classifier with the nameone vs rest.So now the way is clear to recognize different gestures when we calculate the vectors distance by DTW distance.

3.Model Design

A. Data set

Our data is collected from smart phones which provide many categories of data, such as acceleration and angular speed. Based on APP: phyphox which can collect data of smart phones in certain time sequence. Here is an example of one data set which are acceleration data without g in three axes.

	A	B	C	D	E	F	G
1	Time (s)	Linear Ac	Linear Ac	Linear Ac	Absolute acceleration (m/s^2)		
2	0	-0.0738	-0.0839	0.0549	0.124498		
3	0.01	-0.0738	-0.0839	0.0549	0.124498		
4	0.02	1.1033	1.1099	1.2462	2.000541		
5	0.03	0.8588	1.1295	1.4711	2.043879		
6	0.04	0.8986	1.1483	1.1084	1.831564		
7	0.05	1.0461	1.1381	1.2261	1.973048		
8	0.06	0.105	1.2129	1.4852	1.920409		
9	0.07	0.2824	1.2024	1.9864	2.339081		
10	0.08	0.1791	1.2414	2.2682	2.591888		
11	0.09	-0.2777	1.6347	2.0199	2.613304		
12	0.1	0.1815	1.2467	1.6072	2.04213		
13	0.11	0.2676	1.0675	1.5749	1.921321		
14	0.12	2.0325	1.8268	0.9609	2.896823		
15	0.13	1.3075	1.6443	1.3285	2.485597		
16	0.14	1.2738	1.3543	0.9939	2.108206		
17	0.15	0.9614	1.1433	0.9346	1.762073		
6763	67.61	0.5449	0.3524	1.9757	2.079541		
6764	67.62	0.5269	0.4545	2.4772	2.573075		
6765	67.63	0.4997	0.5739	1.3856	1.580806		
6766	67.64	0.3813	0.668	-0.1922	0.792814		
6767	67.65	0.0553	0.7315	-1.622	1.780178		
6768	67.66	-0.3571	0.6088	-2.5385	2.634794		
6769	67.67	-0.5378	0.2004	-3.0461	3.099696		
6770	67.68	-0.5235	-0.0596	-2.537	2.591134		
6771	67.69	-0.5486	-0.0905	-0.8901	1.04949		
6772	67.7	-0.3893	-0.0437	-0.1698	0.426962		
6773	67.71	-0.4047	-0.0502	0.1216	0.425545		
6774	67.72	-0.1167	-0.1215	0.5432	0.568724		

Fig.8 Current data set: acceleration data without g

We have designed 6 different hand gestures and 3 body motions which we use to collect and clean data.

Here are also some constraints on data collecting. We define one motion time slot is 3 s. Since when experimenting we set collecting rate as 100 HZ, so we will get a piece of data every 0.01 s.

Besides, when collecting data, the interval between two data should be above 2 seconds so that our program can partition it successfully and two motion interval will not interrupt.

B. Pre-process of data

We apply smoothing and normalizing to preprocess the data before it is sent to model for training,

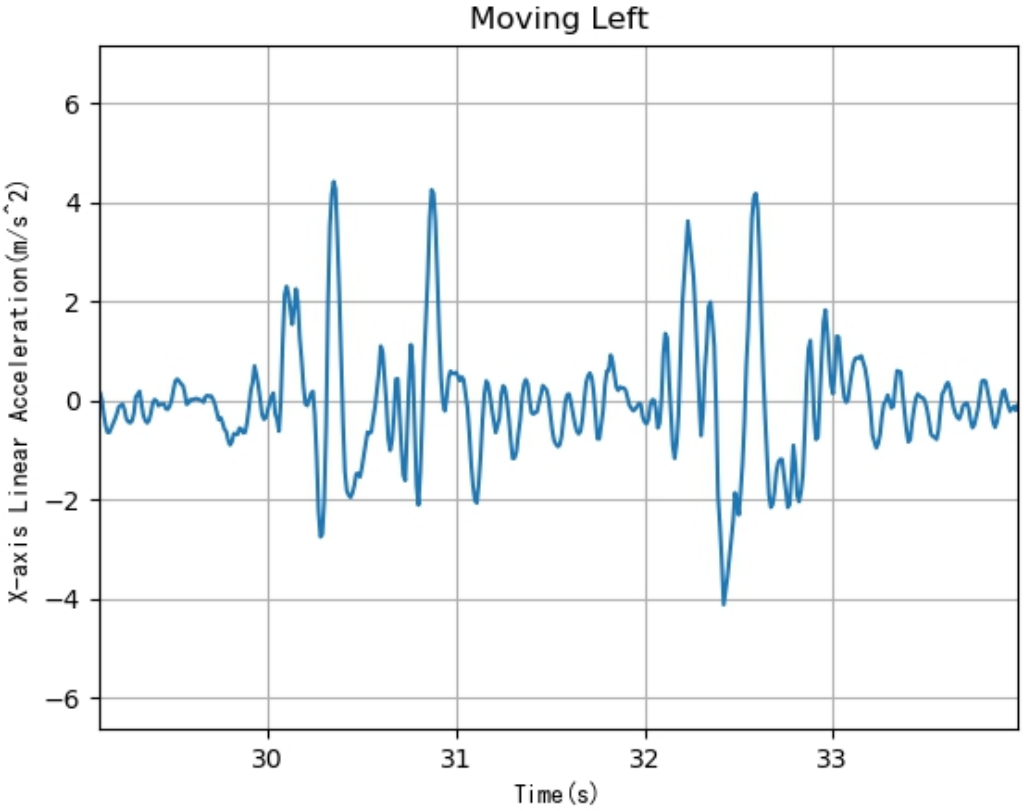

Fig.9 Raw data

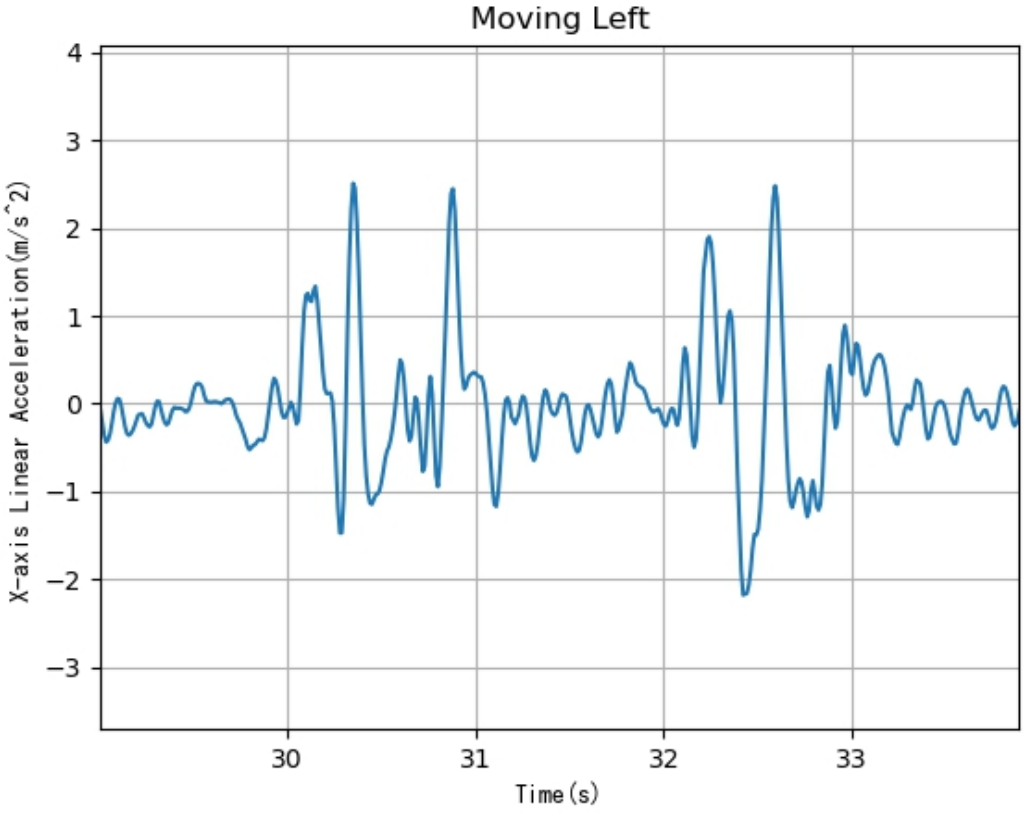

Fig.10 Smoothed data

Figures shown above are raw data and smoothed data. Figures below shows the equations which are used to smooth.

$$a_{now} = \frac{a_{i-domain} + a_{i-domain+1} + \ldots + a_i + \ldots + a_{i+domain}}{window - size} \quad (1)$$

Fig.11 Equations[1]

C. Deep Learning model

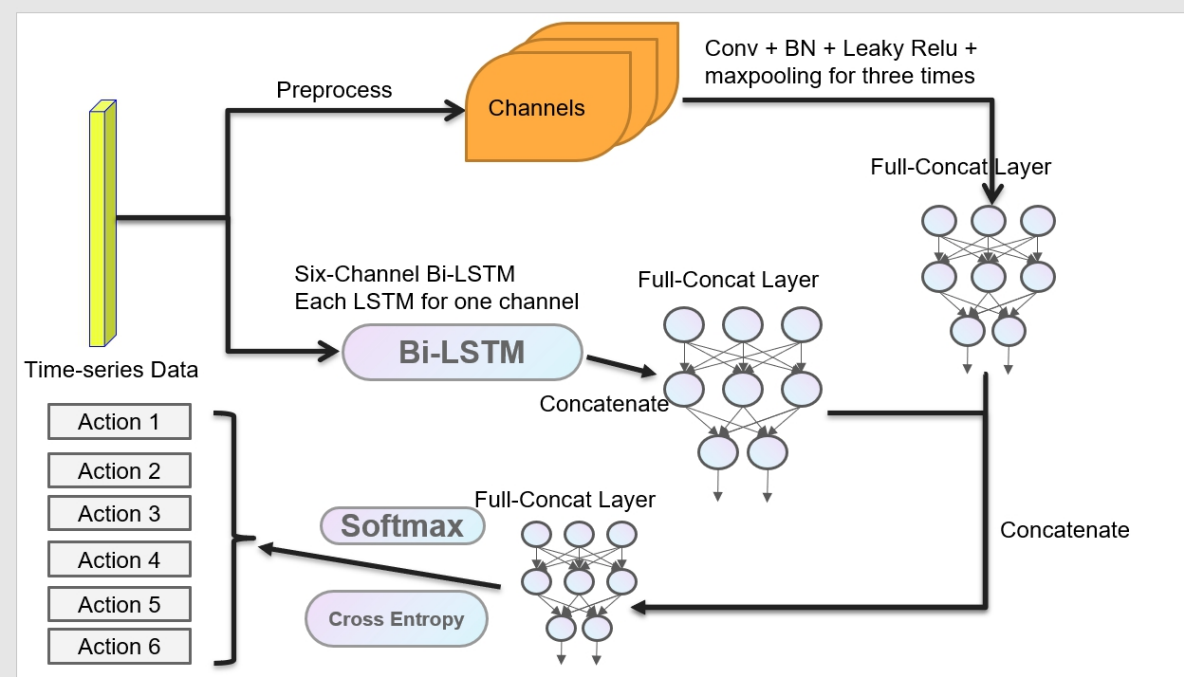

Fig.12 Neuronal network of CNN and Bi-LSTM

Our neuronal network is designed as above. Training data will be run both in LSTM layer and CNN layer. Features which are output from CNN and LSTM will be combined to final layer.

5.Future Work

After the project ends, we will finish our experiment part in our essay so that it becomes complete.

If possible, we will get more data to enrich our model.

6.Reference

[1]龙秋玲.基于改进 CNN-LSTM 的人体行为识别研究[D].成都: 电子科技大学软件工程, 2020:35.

[2]Kim, C., Li, F., & Rehg, J. M. (2018). Multi-object tracking with neural gating using bilinear lstm. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 200-215).

At last, thanks to all professors and seniors who give us great help in this term! Thanks for your help!