# Teeth-Brushing Recognition Based on Deep Learning

*Ming-Xiu Jiang, Yan-Ming Chen, Wei-Hsiang Huang, Po-Hao Huang, Yu-Hsiang Tsai,*
*Yu-Hsuan Huang, Chen-Kuo Chiang,*
*National Chung Cheng University, Taiwan. ckchiang@cs.ccu.edu.tw*

***Abstract***

In this paper, we propose a multi-stream deep learning framework to tackle the activity recognition problem of sixteen kinds of Bass brushing methods by brushing photos and sensor data. This is a challenge task because the model needs to infer the relevance between images and sensor data. In order to solve this problem, CNN model is exploited to learn the spatial features from images and the LSTM model is used to learn the temporal features from sensor data. Then, a fusion scheme is proposed for prediction. Experimental results show that our model achieves high accuracy by using both images and sensor data under the constraint that the dataset is still quite limited.

## I. INTRODUCTION

Recent activity recognition methods use image sequences or sensor data. However, sensor data cannot be used alone for tooth-brushing recognition because the way of left and right brushing are similar, i.e. brushing the right row outside and the left row inside. We can know which side we brush when combining sensor data with images. When use only images, we cannot tell the upper teeth or the lower teeth on the same side because the distance of the upper and lower teeth is close and may be covered by lips. Using sensor data could improve the result because brushing the teeth in upper and lower row are different. Compared with existing method, we integrate images captured by mobile lens with sensor data from the 12-axis sensor data of smart bracelet to achieve precise recognition of sixteen kinds of Bass teeth-brushing.

## II. RELATE WORK

In [1], the dataset is composed of single three-dimension space data from six-axis sensors. It can recognize the angles of hand swinging and compute brush positions based on acceleration and magnetic sensors. To estimate the relative positions using model trained by sensor data, we use multi-stream deep learning architecture to combine images and sensor information for recognition.

## III. PROPOSED METHOD

In activity recognition method, it usually captures statistical information through the mean, variance or entropy to extract features. However, these methods are not effective enough for the recognition of multiple actions. Since it can only capture the linear structure of feature space. Deep learning is regarded as one possible method with the greatest potentials to solve this problem with its non-linear structures.

Motivated by [3], we first extend the single-stream ConvNets for recognizing activities using brush photos. To prevent from overfitting, we pre-trained CNN layers using the weights from existing CNN models. Secondly, we propose to combine Long Short-Term Memory (LSTM) for classifying wearable sensor data within one unified framework. We merge the last dense layer of sensor model and second last dense layer of image model to obtain the final predictions.

To improve convergence speed and avoid halting at local optimal in our multi-stream model, we pre-train image and sensor model separately before training the unified model**.**

In Table I, Inception V3, Cifar-10 and VGG-19 [4] have been used for training model using image data. It shows that best recognition accuracy is from VGG model. So, VGG-19 model is exploited for image data in our multi-stream model. In practice, we used VGG-19 pre-trained model to decrease spending on training model. In the last layer of VGG-19 model, we added two fully-connected layer and one softmax layer, mapping results to sixteen kinds of brushing actions.

Table I、Accuracy by different CNN models.

| **Model Name** | **Testing Accuracy** |
|---|---|
| Cifar-10 **[5]** | 67.1% |
| Inception-V3 **[6]** | 80.6% |
| VGG-19 | 85.4% |

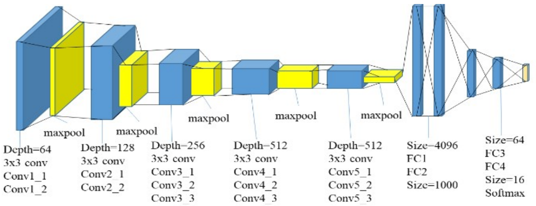

Figure 1. VGG-19 structure [4].

For sensor data, Long Short-Term Memory (LSTM) is used as training model. After the pre-training, we train multi-modal with the weight of the pretrain model (the yellow part of the picture below). The prediction is made by concatenating the last dense layer of sensor modal and second last dense layer of image model and merging the softmax score of lower channel (VGG-19) and the upper channel (LSTM). SGD is used as optimizer, Learning rate is set to 0.001, and use cross entropy as loss function.

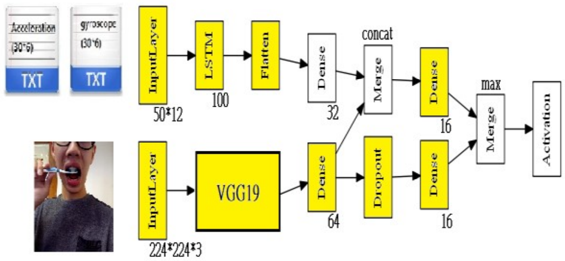

Figure 2. Multi-model structure.

## IV. EXPERIMENTAL RESULTS

We collected data by smart bracelets (collection rate is 25 fps) and frontal camera of smart phones from 74 people throughout a month. The number of training samples is 12,186 and the test sample is 4,107. The image size is 224*224*3. Then, the data samples are manually labeled according to Bass brushing method.

The accuracy of classifying 16 different movements of Bass teeth-brushing is depicted by the confusion matrix in Figure 3. We can see that the accuracies of movements No.5 (palatal surface) and No.16 (right upper buccal surface) are lower. The reason is that the palatal surface and the lip surface actually looks similar in the image, as well as the right upper buccal surface, right upper lingual surface and the right upper occlusal surface. This is why the prediction errors of image model is higher than the multi-model. After combining with sensor data (the right picture), we can see significant improvement on both of the movements.

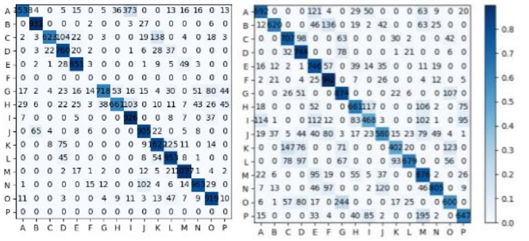

Figure 3. Confusion matrix (left) of image model and (right) multi-model.

For multi-model, we concatenate the second dense layer of the upper pipeline and the first dense layer of the low pipeline. In addition, an additional dense layer is included in the low pipeline. The results are shown in Table II. We can see that the accuracy of multi-model achieves higher accuracy of 89.3% compared to VGG19 and LSTM.

TABLE II、The comparison of test accuracy.

| | Testing Accuracy |
|---|---|
| VGG19 | 85.4% |
| LSTM | 72.5% |
| Multi-model | **89.3%** |

## V. CONCLUSION

Due to the coverage of lips, images of teeth-brushing are not enough for brushing recognition. By using sensor data only, there are many similar brushing movements at different teeth positions. This leads to the success of integrating image model and sensor model fusion within one deep learning model and introduce significant improvement over the accuracy of single model. In the future, it is possible to extend the image model to RNN/LSTM for further recognition improvement.

## REFERENCE

[1] K.-H. Lee, J.-W. Lee, K.-S. Kim, D.-J. Kim, K. Kim, H.-K. Yang, K. Jeong, and B. Lee, Tooth brushing Pattern Classification using Three-Axis Accelerometer and Magnetic Sensor for Smart Toothbrush, International Conference of EMBS Cité Internationale, 2007.

[2] S. Song, V. Chandrasekhar, B. Mandal, L. Li, J.-H. Lim, G. S. Babu, P.-P. San, and N.-M. Cheung, Multimodal Multi-stream Deep Learning for Egocentric Activity Recognition, CVPRW 2016.

[3] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition-Visual Geometry Group, arXiv:1409.1556v6.

[4] A. Krizhevsky, Convolutional Deep Belief Networks on CIFAR-10(2010).

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-ResNet and the Impact of Residual Connections on Learning-Google Inc. 1600 Amphitheatre Parkway Mountain View, arXiv:1602.07261v2.