Activity and Emotion Detection for “Smart Homes”

Inimfon Akpabio

March 2, 2021

Abstract

Advances in IoT (Internet of Things) and artificial intelligence have facilitated the deployment of smart home devices that enable hospitals and health workers monitor the state and well-being of patients, particularly elderly people, who often live alone. Such technologies can be crucial in tracking the health of a patient and ultimately, providing emergency care to such individuals if needed. In this project we propose a model to detect “brushing teeth” in a smart home setting. Our model starts with VGG16 network used for feature extraction and then feeds into GRU + Fully-connected layers to perform the final prediction. The feature extractor is trained separately to improve convergence. For each video, we sample only 5 frames. The rationale is that brushing is a perpetual motion, hence, this should be sufficient. Random still-images from the internet were used to train the feature-extractor while a mixture of videos from both the Kinetics700 and HMDB51 datasets were used to train the overall classifier. The model currently has an accuracy of 70% which leaves plenty room for improvement.

1. Topic

The selected action for this project is “brushing teeth”.

2. Motivation

Oral hygiene is an extremely important and yet often overlooked facet of human life. Failure to maintain proper oral health can result in heavy buildup of plaque which is created by bacteria accumulation on teeth. Such unchecked buildup of plaque can ultimately lead to gingival inflammation as well as tooth deterioration/decay [1]. Hence, monitoring activities related to oral care such brushing teeth, can prove to be extremely useful for Oral hygienists and dentists alike. There have been multiple attempts to monitor brushing teeth using signals from wearable devices [2]. While these methods can be useful for monitoring, they are inevitably limited due to the one-dimensional nature of this approach. Hence, it makes sense to utilize visual data as a means for detection. The benefit of the visual approach is that it affords more features for better detection and can be easily collected using video capture devices. A lot of research work has been done in the field Action/emotion detection in videos in recent years. Advances in computer-vision AI and the advent of competitions worldwide [3] have significantly increased interest in video based approaches for action detection. Similarly to [4], we will attempt to build a deep network that can detect brushing teeth, however, we will not use any sensor data and rely solely on visual data from two second clips.

3. Related Works

In [4], image data is combined with wearable sensor data to detect 16 different techniques of brushing teeth. In this approach, a single stream convent trained to recognize activities in images is used to extract features. The second to last dense layer of this convnet is then merged with the output of an LSTM network which processes the signals from the wearable device. Three different models based on CIFAR-10, Inception-V3 and VGG19 were developed. The VGG-19 was able to achieve an impressive classification accuracy of 85.4%. It must be noted however, that these models were trained to perform intra-class prediction according to the bass brushing techniques. Our model, on the other hand will be performing inter-class prediction between brushing teeth and not brushing teeth.

4. Proposed model

While this model design is not directly based on any particular paper, there are a couple of research endeavors that have adopted similar designs to tackle video and speech related problems. For instance, [5] uses VGG-16 layers to extract features from a video clip and then feeds this data to Bi-directional LSTM network to generate captions for the video.

Fig 1 shows the complete model design. The model starts with a VGG16 network which has been separately on still images for feature extraction. At the end of the VGG16 network, we perform a global maxpool which basically takes the maximum value from each filter. This helps to reduce the dimensionality of features and promotes spatial hierarchy. Next using time-distribution, the features are fed into a GRU network with 64 units. GRU have been shown to be comparable to LSTM in performance and yet more efficient since they lack a memory cell [7]. Finally the output the GRU is fed into a fully connected network that performs prediction. The fully connected network consists of two layers with 1024 and two units respectively. The second layer applies Softmax activation on the 2 units to perform binary-classification.

As a pre-processing step, video clip frames must be resized to (224, 224) and scaled to [0, 1]. The input to the model is an array of 5 image frames sub-sampled from a 2 second clip and has a dimension of (5, 224, 224, 3). It outputs an array of size 2 representing the prediction probabilities for “brushing teeth” and “not brushing teeth”.

5. Dataset

Feature Extractor:

Since it is relatively difficult to find datasets for brushing teeth, a variety of sources were combined. To train the feature classifier, still-images of people brushing teeth were downloaded from the GettyImages website. Images from other classes were obtained from the Stanford40 action dataset [8]. Initially, the images were simply divided into “brushing teeth” and “not brushing teeth” and fed directly to the model for retraining. However, the model exhibited very poor performance, yielding only about 52% accuracy on train and validation sets. To alleviate this, the classes were expanded to eight. The eight classes were comprised of “brushing teeth” and 7 “not brushing teeth” classes. This simple, yet effective adjustment

increased both train and validation performance by at least +30%. Data Augmentation was also applied to prevent overfitting.

Full model:

Finding a video dataset to train the full model proved even more difficult. The Kinetics700 dataset is a large datasets that contains a myriad of human-performed actions and has been widely explored for video recognition tasks in recent years [9]. Due to its massive size, it could take days to download the entire dataset even with a stable internet connection. Hence, it made sense to download only the brushing teeth class and combine this with the HMDB51 dataset which similarly contains videos of human-performed actions [10]. For training, the videos were separated into two classes: “brushing teeth” and “miscellaneous” which represents the collection of all “not brushing teeth” videos. Both classes currently have a size of 255 and 363. As a preprocessing step, the brushing teeth videos were down-sampled to 2s to localize the action and thereby improve convergence during training. Data Augmentation is applied to increase generality and curb overfitting.

6. Model Training and Performance

Hyper-parameters like batch-size, epochs and learning rate were selected by testing random values in fixed ranges and choosing the values that yielded the best convergence. About 80 epochs were used to train the model. Initially, the model began to exhibit heavy overfitting. This unwanted phenomenon was curbed by adding L2 regularization and dropout of 20% to the fully connected layers. On the training set, the model achieves a good accuracy of about 87% and about 80% on the validation set. More data will be provided in the next report.

7. Youtube

On the curated youtube test set, the performance drops to 70%. This is probably due to the limited size of the training set compared to the test set. The FPR and FNR are both 40% and 19% respectively. Hence there is still plenty room for improvement.

8. Improve Accuracy

Next we will explore advanced techniques such as optical flow.

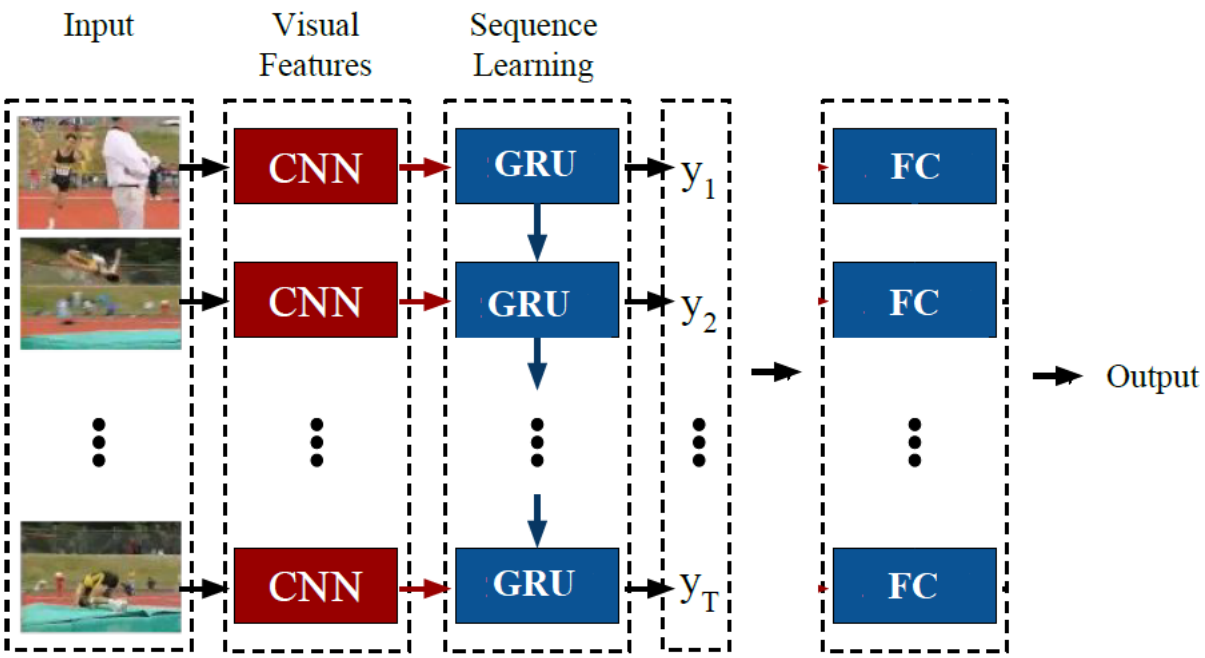

Fig 1: Model Diagram

Source: Adapted from [6]

References

[1] A. D. Association. Tackling tooth decay. Journal of American Dental Association, 2013.

[2] Huang, Hua. (2016). Toothbrushing Monitoring using Wrist Watch. 202-215. 10.1145/2994551.2994563.

[3] Dhall, Abhinav & Goecke, Roland & Gedeon, Tom & Sebe, Nicu. (2016). Emotion recognition in the wild. Journal on Multimodal User Interfaces. 10. 10.1007/s12193-016-0213-z.

[4] M. Jiang et al., "Teeth-Brushing Recognition Based on Deep Learning," 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taichung, Taiwan, 2018, pp. 1-2, doi: 10.1109/ICCE-China.2018.8448684.

[5] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen and X. Li, "Describing Video With Attention-Based Bidirectional LSTM," in IEEE Transactions on Cybernetics, vol. 49, no. 7, pp. 2631-2641, July 2019, doi: 10.1109/TCYB.2018.2831447.

[6] Donahue, J.; Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 677–691

[7] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, Dec. 2014, [online] Available: http://arxiv.org/abs/1412.3555.

[8] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.

[9] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987, 2019.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In 2011 International Conference on Computer Vision, pages 2556–2563, 2011.