A Spatial Feature Attention Enhanced Facial Action Unit Detection Model for Analyzing the Effects of First Language on Facial Muscle Movements

Mohan Yu

Beijing National Day School

aiden.yu@hotmail.com

Beijing, China

Keywords: Linguistics, Facial Features, Deep Learning, Key Point Detection, Attention Mechanisms

Abstract: Knowledge of one's first language, which is greatly indicative of their cultural background, is a valuable asset in today's diverse society. Different languages have unique phonetic combinations, causing native speakers to use facial muscles in distinct patterns. This paper proposes the spatial feature attention-enhanced facial action unit detection model SFAE-Net to detect facial movements by quantifying Facial Action Units (AUs). SFAE-Net consists of two modules: the face key-point-assisted region learning module (LRL) and the multi-scale region learning module (MSL). LRL uses key points to focus on AU regions, while MSL captures multi-scale features to improve generalizability. Experiments show SFAE-Net achieves an F1-score of 62.7% and accuracy of 79.5%, outperforming state-of-the-art models. The paper also provides an instance of analyzing first language effects on facial muscles using the model.

1. Introduction

In diverse societies, understanding cultural backgrounds is crucial for effective communication. A person's first language strongly reflects their culture, and facial movements while speaking can indicate native language patterns. Different phonetic sounds lead to unique facial muscle usage, making it possible to infer first language from facial features. A deep learning network analyzing the pattern of facial muscles can thus provide us with insight into the subjects' cultural backgrounds by predicting their first language. To realize this design, an antecedent network to accurately recognize facial movement is crucial.

The Facial Action Coding System (FACS) first proposed by Ekman P, Friesen W V, and Hager J C, often used by psychologists to recognize basic emotions, defines a number of Action Units (AUs) that are representative of human facial movements. As Figure 1 shows, AU5, for example describes a lift of the eyelid, while AU7 describes a tightening movement. Many facial movements can be described by a combination of AUs. A surprised look, for example, might feature AU2 (a raising of the outer brow) and AU26 (a jaw drop), while a stressed face might consist of AU4 (lowered brow), AU9 (a wrinkle on the nose), and AU23 (the tightening of the lips). By defining AU, objective quantitative measurements can be made in order to better identify and analyze facial muscle movements.

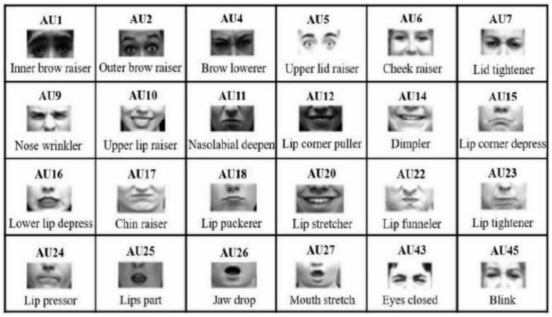

Figure 1: Example Action Units

  DOI: 10.25236/icceme.2025.021

This paper proposes a deep learning network to accurately identify AUs as a stepping stone to constructing the deep learning network able to conclude first language from facial muscle movements.

2. Manuscript

2.1. Model design and overall structure

Upon observing that AUs are distributed in different regions of the face and that separate analysis of facial regions can thus enhance performance in AU detection, Zhao et al. proposed the Deep Region and Multi-label Learning (DRML)[1] network, whose region layer evenly crops the input feature map obtained by a prior convolutional layer into 8x8 blocks. As Figure 2 shows, each of the 64 small blocks are independently learnt by a convolutional layer and then stitched together to be fed into the fully connected layer. This design has fewer parameters and better generalization in AU detection compared to the traditional CNN structure, as it focuses on local features of the input data.

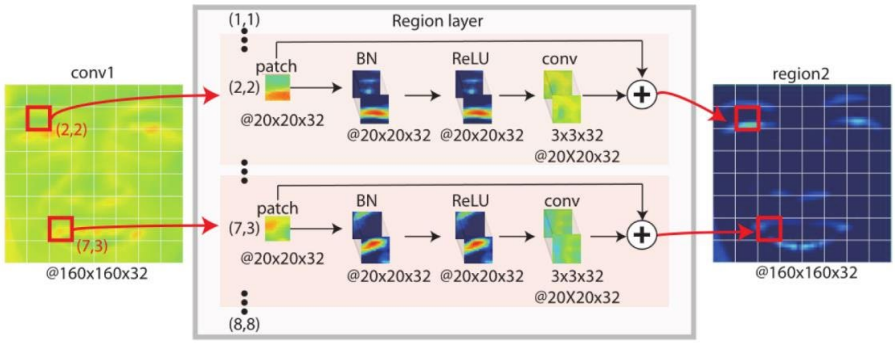

Figure 2: The key Region Layer proposed in DRML[1]

AUs are defined to describe facial muscle movements. Based on this theory, the proposed model features a multi-scale region learning module to allow the model to learn facial regions of different sizes. Another module that first extracts facial key points to analyze local features of their surrounding regions is also used, as these key points often indicate the center of the AU and can ensure specific, focused learning of relevant facial regions. The two modules are combined in order to obtain a comprehensive analysis and thus a detection of maximum accuracy. With AUs often located at facial key points such as the eyes, nose, mouth, etc., key points are highly indicative of AU locations, and more and more studies are exploring the learning of AU features at facial key points[2]. The EAC-Net proposed by Li et al. [3], for example, features an augmentation network that uses an attention mechanism to focus on regions that facial layers are local to. The coordinates of the centroids of 20 AUs are first found, and their surrounding features are weighted according to Manhattan distance from the centroids. Logically, the greater the distance, the smaller the weight, and vise versa. As demonstrated in Figure 3, the Dlib[4] tool is first used to detect face key points, locating the centroid of each AU, and each facial region is then weighted according to distance from the closest centroid. Therefore, the network is allowed to focus on learning the AU location based on this attention graph.

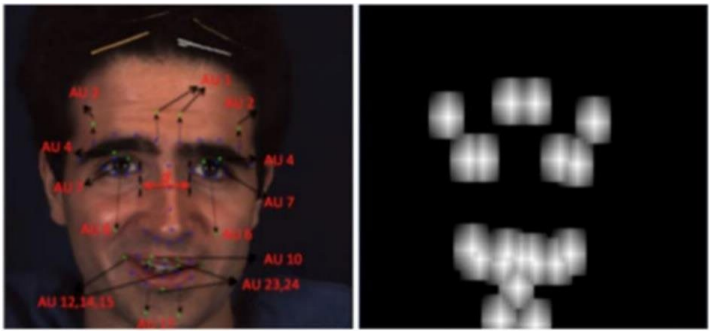

Figure 3: EAC-Net attention map generation process [3]

By applying weight according to distance from these key points, such as eyes, cheeks and eyebrows etc., allow concentrated attention on these regions of interest, and therefore not only reduce channels and increase efficiency, but also eliminates distraction from data on irrelevant areas. In addition, considering the individual difference between faces and images, using face key points as landmarks to navigate regions for learning can make specification on each individual image and adjust to specific individual facial distributions, therefore greatly eliminating the negative affect of individual differences in AU location.

The *SFAE-Net* consists of two sub-networks: the Landmark-assisted Region Learning (LRL) module for detection guided by facial key point landmarks, and the Multi-Scale Region Learning (MSL) module. The network feeds images into the two modules separately, each getting detail-enhanced facial AU features. LRL uses the detected facial key points to generate attention maps with which AU regions are weighted and learned, obtaining locally strengthened AU feature f1 and face key point feature f2; at the same time, MSL obtains the multi-region-aware feature f3 without relying on facial key point detection. The feature maps f1, f2 and f3 are then fed into the fully connected layer for the final AU prediction outcome.

2.2. LRL and MSL modules

Facial key points have an extremely close relationship with AUs: on the one hand, they indicate the center of AUs, which is of great help in focused local learning; on the other hand, AU activation triggers regular movement of facial key points, e.g., AU1 (eyebrow inner lifting) and AU2 (eyebrow outer lifting) results in upward motion of key points surrounding the eyebrow, while AU25 (lips slightly open) and AU26 (chin drop) leads to movement of key points around the corners of the mouth. Therefore, the correlation between the two can be exploited to improve AU detection performances. A feature extraction module extracts the features used for face key points and AU detections. Figure 4 shows its detailed structure, consisting of one ordinary convolutional layer followed three partitioned ones of different scales. The global facial feature is obtained first through the first convolutional layer, and is fed into the next layer where the global feather is divided into 8x8 blocks that are then learned separately, and then stitched together to be passed into the next layer. The same process is repeated for the following two convolutional layers, where the input is divided into blocks of 4x4 and 2x2, respectively. The output of the three partitions convolutional layers are connected to yield a multi-scale feature map which is next fed into subsequent modules for facial key point and AU detections.

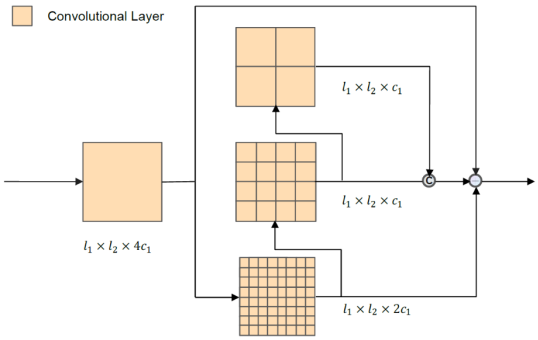

Figure 4: The feature extraction module

While key points are often used in previous works to localize AU regions, some AUs are associated with multiple regions, making it insufficient to obtain regional features based on key points alone. Other works seeking to locate AUs without relying on facial key points, such as the DRML network, on the other hand, often deploy an overgeneralized way of dividing facial regions, overlooking variations such as the location or size of the muscle that exist between each subject, and thus making the model vulnerable to these inevitable individual differences. To cope with this, the MSL in *SFAE-Net* learns the global feature map at various scales, effectively reducing the impact of individual differences.

The inverted residual module of MobileNetV2 [5], a network proposed by M. Sandler et al. is

integrated into this part of the model. MobileNetV2 is a lightweight yet accurate detection network. While MobileNetV1 uses traditional convolutions that require large amount of computation, the use of depth-wise convolution in MNetV2 reduces computation, enabling the processing of more parameters with greater speed. This modification allows inverted residual architecture to first use a 1*1 expansion convolution to produce parameters, enhancing the accuracy of later classification. The obtained highdimension features are then fed into the 3*3 depth-wise convolutional layer to be processed with reduced computation and thus increased efficiency, while high accuracy is maintained by the increased parameters. Another 1*1 projection convolution is then executed to compress the data, and prediction is made by the fully connected layer. The production of parameters done by the expansion layer and use of the highly efficient depth-wise convolution resulted in MobileNetV2 showing landslide low number of parameters and at the same time maintaining high accuracy in the image classification experiments featured in the paper in which it was first proposed. For *SFAE-Net*, a modified version of the inverted residual structure, which is the backbone of MobileNetV2, constitutes the MSL module in order to do lightweight and accurate multi-scale learning. As Figure 5 shows, the 7x7-sized global feature map is interpolated and average-pooled to obtain feature maps of different sizes 9x9 and 5x5, respectively. Each of the three feature maps is expanded by factor 6 using the 1*1 expansion convolution in order to produce parameters and increase accuracy of the prediction later made in the fully connected layer. A depth-wise 3*3 is then used to efficiently extract features with reduced computation. Another projection 1*1 convolutional layer is finally used to again reduce the parameters the fully connected layer has to process in order to enhance efficiency. The resulting feature maps are normalized and ReLU-activated after each step. They are then averagepooled into 5x5 sizes and stitched together to obtain multi-scale region feature f3. The general process is demonstrated as follows.

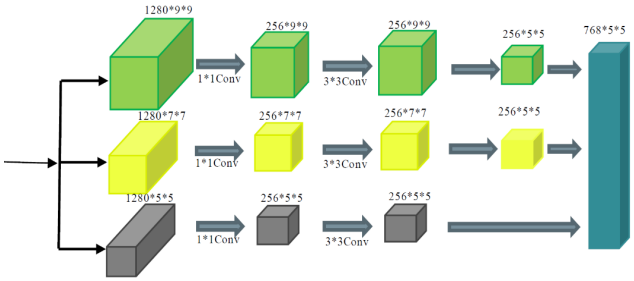

Figure 5: Multi-scale regional learning module

Finally, global feature f is obtained by uniting f1, f2, and f3, and is fed into the fully connected layer to obtain the final prediction of AU probability.

The combination of the MSL and LRL units integrates the advantage of the two novel analytic perspectives, namely, the independent accuracy of region-cropping learning and the concentrated attention of landmark-assisted learning. It also compensates for the flaws of using either unit separately, where the case-specificity of LRL makes up for the generalization of MSL, and the more global view of MSL makes up for the oversight of low-wight areas posed by the LRL's attention mechanism.

The AU detection problem can be regarded as a multi-label classification problem whose results are in probability of 0 to 1. Naturally, the closer the probability is to 1, the more likely the AU is present, and the closer it is to 0, the less likely.

2.3. Experimental Results and Analysis

2.3.1. Dataset and experimental setup

Just like other computer vision tasks, AU detection requires large-scale datasets with label annotations for training in order to have better detection performance. However, labeling AUs in face images is much more challenging and complicated compared to labeling simple face or other target objects. This time-consuming and laborious task requiring professionally trained labelers to complete, and current datasets based on the FACS system is very limited. In this paper, the BP4D-

Spontaneous dataset[6, 7] established by Binghamton and the University of Pittsburgh is used. Figure 6 shows some BP4D sample examples. The dataset includes video samples of 41 subjects of different races and genders, annotated of 12 commonly used AUs (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24).

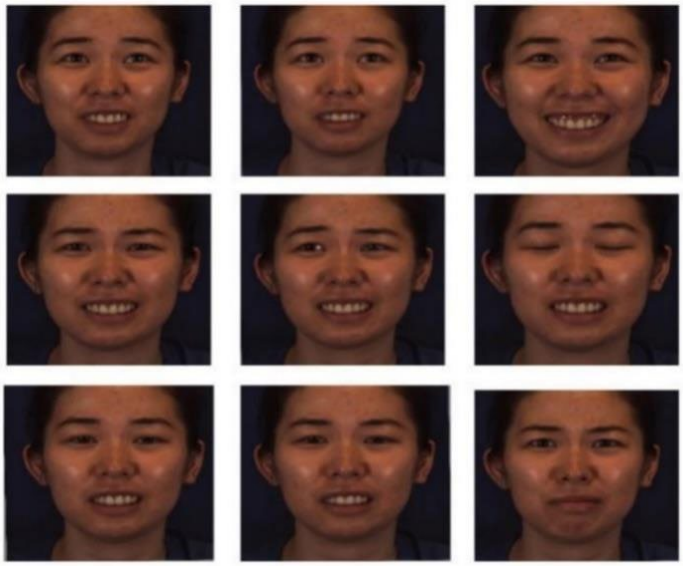

Figure 6: BP4D sample example

The effect of deep learning training depends heavily on the quality and diversity of the training set, and therefore a series of prepossessing operations need to be done on the dataset. Each face image is repositioned and reoriented with planar rotation, scaling and translation to obtain 200x200 sizes. To increase diversity of the data, the images were further randomly mirror-flipped, rotated, and cropped.

The experiment was conducted on Ubuntu 16.04 system using Pytorch, and accelerated using NVIDA3080Ti graphics card. The network is trained for 12 cycles with initial learning rate of 0.01, where learning rate is multiplied by 0.3 every two cycles to obtain optimal result.

2.3.2. Evaluation and analysis

The most commonly used metrics for evaluation AU detection models are F1-score and accuracy. The two are used together in this paper to assess the proposed network. Accuracy, among the most commonly used methods of assessing models, shows the proportion of made predictions that are proved true. It indicates how "reliable" the model is, and is the intuitive way to evaluate *SFAE-Net*. As detection tasks often feature imbalanced datasets, where one category has significantly fewer instances than the other; after all, the activation of an AU is only "true" for one small region over the entire face. For such cases, the F1-score, given by the harmonic mean of recall rate and precision, considers both qualities with equal weight and thus gives a more comprehensive evaluation.

The accuracy and F1-score of *SFAE-Net* are compared with other mainstream methods, including DRML[1], EAC-Net[3], DSIN [8], and CMS [9]. The table 1 below shows experimental results.

Table 1 Experimental results.

TIME	F1-score	Accuracy
DRML	48.3%	76.4%
EAC-Net	55.9%	75.2%
DSIN	58.8%	72.9%
CMS	60.6%	78.3%
Ours	62.7%	79.5%

Seeing the resulting data as shown in the table, it can be found that the *SFAE-Net* proposed in

this paper shows both better F1-scores and accuracy than any other methods in the experiment done on the BP4D dataset. The proposed model outperforms both DRML [1], which uses only regional learning, and EAC-Net [3], which uses only landmark-assisted learning, in both F1-score and accuracy assessments, meaning that the integration of these two advantageous perspectives can push a more comprehensive analysis of the data, further improving detection performance over either used separately.

Experimental results also show superior performance of *SFAE-Net* over DSIN and CMS, two other established, accurate networks, which further proves the integration effective.

3. Conclusion

The *SFAE-Net* proposed in this paper, a Spatial Feature Attention Enhanced facial action unit detection model designed for analyzing the relationship between mother tongue and facial muscle features, works by combining two main modules, namely, Landmark-assisted Regional Learning module which learns facial features according to an attention map obtained through weighting facial regions based on distance from key points, which also effectively make specification to each individual case in response to individual differences, and the Multi-Scale Region Learning module, which learns facial features at different scales in order to exclude individual variations and improve generalizability. The features separately obtained by the two modules are integrated and fed into the fully connected layer for prediction of the presence of AU. Experimental results show that integration of the two perspectives in *SFAE-Net* significantly improves AU detection performance.

Although the model outperforms all others that are tested in the experiment, displaying integrated advantage of the two featured analytical perspectives, an F1-score of only 62.7% and accuracy of 79.5% is far from ideal. A considerable part of the reason is that a dataset of merely 41 elements is not nearly enough to sufficiently train a deep learning model on. To further improve the performance of the model, datasets of such kind with larger sizes are required in order for more sufficient training to be done. In addition, other analytical perspectives, such as the relationship between the presence of multiple AUs, since the activation of one AU is very likely accompanied by that of other(s), can be considered and be integrated into the current model in order to make up for flaws such as focusing only on identifying each AU independently while overlooking such inter-AU relationships – flaws that are not yet resolved by the current model design.

As shown in the example featured above, the proposed model can be used to analyze facial muscle featured made to be by the speaking of first language. An extended deep learning model can be constructed in the future such that an AU detection network, like the *SFAE-net* proposed in this paper, acts as a specific feature extraction module obtaining AU distribution features from the input video sample and feeds them into the deeper layers of the model, where the feature data are processed and weighted to be fed into the fully connected layer in order to make predictions at the subject's mother tongue. However, the absence of a sizable and quality dataset, one that is essential to making objective analysis, drawing justified and generalizable conclusions, and not to mention constructing and training deep learning models for such analytic task, still poses a grave challenge.

For future studies of the subject matter, such as the deep learning model for determining first language based on facial features as conceived above, the presence of a suitable dataset is crucial. The dataset needs to contain video data of subjects speaking or reading out a text labeled of their mother tongue. A considerable size of the dataset is also required, as the hypothetical model, just as al deep learning models, needs to train on sufficient data in order to obtain generalizability and accurate performance. Being the very foundation of the model, the AU detection network must also be improved to yield satisfactory accuracy, through the means mentioned above such as training on more data and 12 integrating more analytical perspectives, in order for the hypothetical model to obtain descent performance.

References

[1] Zhao, Kaili, Wen-Sheng Chu, and Honggang Zhang. "Deep region and multi-label learning for

facial action unit detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] Li, Wei, Farnaz Abtahi, and Zhigang Zhu. "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[3] Li, W., Abtahi, F., Zhu, Z., and Yin, L.: 'EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40, (11), pp. 2583-2596

[4] Kazemi, Vahid, and Josephine Sullivan. "One millisecond face alignment with an ensemble of regression trees." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[5] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Zhang, Xing, et al. "A high-resolution spontaneous 3d dynamic facial expression database." 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, 2013.

[7] Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J.M.: 'BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database', Image and Vision Computing, 2014, 32, (10), pp. 692-706

[8] Corneanu, Ciprian, Meysam Madadi, and Sergio Escalera. "Deep structure inference network for facial action unit recognition." Proceedings of the European conference on computer vision (ECCV). 2018.

[9] Sankaran, Nishant, et al. "Representation learning through cross-modality supervision." 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019.