Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
BM40A0801 Machine Vision and Digital Image Analysis

Sara Heikkinen, 567012
Alex Karonen, 566770

PRACTICAL ASSIGNMENT: IMAGE-LEVEL MICRO-GESTURE CLASSIFICATION

March 30, 2024

Responsibilities:	Sara Heikkinen: Planning of the model architecture, contibuting code, writing the literature review Alex Karonen: Planning of the model architecture, writing the code, writing about methods

CONTENTS

1 INTRODUCTION

Micro-gesture is a unintentional behaviour which is driven by feelings of the individual. These types of gestures are performed without an intention and are usually not noted by the individual. Classifying these type of gestures can reveal the inner feelings behind the gesture. Understanding these type of actions is important since it does play a significant role in human-computer interaction. Humans express their emotions widely non-verbally which makes understanding micro-gestures essential part of understanding a human interaction.

The used dataset is the iMiGUE dataset [1] (more info at: https://github.com/linuxsino/iMiGUE). There are 6 different main classes for the micro-gestures, head, body, hand, head-hand, body-hand and illustrative body language. These are further divided into sub-classes and therefore the dataset have in total of 32 different classes represented. The purpose of this practical assignment is to classify the different micro-gestures represented as images from short video clips.

The distribution of the samples within the classes is extremely unbalanced. This is understandable since some micro-gestures are just more common than some other. Since this dataset is collected from the video clips filmed after some sport tournament, the micro-gestures are not intentionally performed for illustrative purposes, there is no easy way to collect more data of some specific micro-gesture. This will bring challenges to the classification task since training a model with such an unbalanced dataset is not desired situation.

The planning of the implemented model was done together, as so the coding part. The code was done together, Alex writing it and Sara contributing to the implementation. The report is also written together, Sara delved into the related work done in this field, and Alex wrote more about the methods part.

2 RELATED WORK

The used dataset is relatively new and therefore there is not that many articles about this specific dataset but some can be found. In [2] the authors propose a CNN-based network where they incorporate skeletal and semantic embedding loss for gesture classification. It utilizes layers from the ResNet which is residual neural network. The network proposed achieves 64.12% Top-1 accuracy on the iMiGUE dataset.

Many state-of-the-art methods has been used for the micro-gesture recognition [1]. For example Recurrent Neural Networks, Graph Convolutional networks, encoder-decoder architectures and different Convolutional networks have been used to classify the micro-gestures. The highest Top 1 classification accuracy 61.10% is achieved with Temporal Shift model (TSM). TSM method is proposed to specifically for video form data [3]. It utilizes 2 dimensional convolution for cheaper computations but maintains the performance of a 3 dimensional convolution due to temporal shifts.

Most of the models used for classification of this dataset are based on different neural network architecture and convolutions. Convolutions are efficient way of finding underlying patterns in the images and therefore it is justified choice to be used in this type of task. The Pose information (skeleton data) is also widely used method for classifying different gestures [1, 2].

3 METHODS

Since the data is wanted to be classified into different micro-gestures, the problem will become extremely complex since the number of different classes is 32. A multimodal model is proposed here to classify the micro-gestures. As an input for the multimodal model, the original RGB image and a skeleton information created from the image are used. To obtain the skeleton information, Google's MoveNet Single Pose Lightning model is used. It's a pretrained convolutional neural network that predicts human joint locations from RGB images. Example of the skeleton data obtained with the model is seen in the Figure 1.

Due to implemetational difficulties, the MoveNet model could not be attached to the multimodal model itself, thus the skeleton information is created separately and is stored as a dataset and used as a vector form input to the multimodal model. The lightning version of the model was used since it processes smaller images (192 × 192 × 3) and is thus faster than the thunder model. The original 1280 × 720 images have been resized with bilinear interpolation to the desired size.

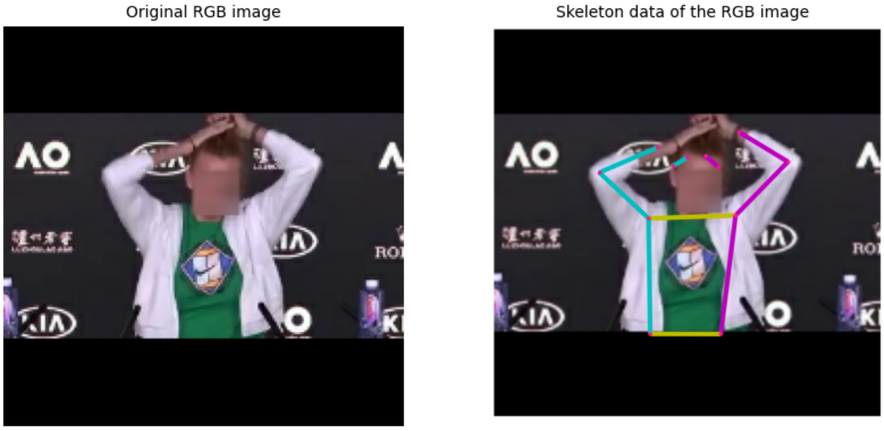

Figure 1. Visualization of the original image (left) and skeleton data on the original example image (right)

The model is implemented with Tensorflow and its Keras interface [4]. The multimodal model is divided to the skeleton side (MoveNet+Dense) and image side, where a convolutional neural network is used. For the convolutional side of the multimodal network, a pretrained EfficientNetV2B0 [5] without the top classification layers was used as its base with two dense layers (sizes 256 and 128 respectively) for transfer learning. EfficientNets are desinged to be smaller but have the efficiency of a bigger neural network to get better results.

The outputs from the skeleton data, and the CNN are concatenated with a concatenate layer and the result is fed to dense classification layers. Therefore obtained information from both, RGB image and skeleton data, is utilized for the classification task. A visualization of the multimodal model is seen in the Figure 2. First two image inputs are processed on their own and then their information is combined and classified.

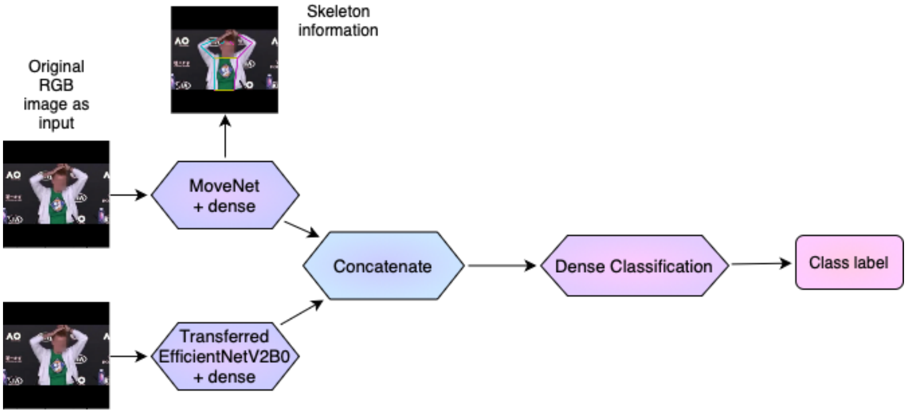

Figure 2. Visualization of the classification procedure

The whole model has been trained for 20 epochs over the training data with validation done also each epoch. Cross-entropy loss (from a softmax output) is used as the loss function in the model training as it is commonly used in a multi-class classification case. The Adaptive Moment Estimation with a Strongly Non-Convex Decaying Learning Rate (AMSGrad) version of the Adam-optimizer is used with a starting learning rate of 0.001. A learning rate reducer is used if the validation loss plateaus and a early stopping callback is used if the validation loss worsens within 4 epochs from the minima.

4 EXPERIMENTS & RESULTS

The data was divided into training and validation data with random split of 80/20 % respectively. The division was done so that from every separate class 20% of the samples were assigned as the validation data in a way that images from one video clip could not be in both train and validation data. Dividing the data in this manner it is ensured that the measured accuracy actually represents the ability to classify different micro-gestures rather than individuals. No separate test dataset was created here and all the results are considering the validation dataset. All the separate frames from the video clips were used as an input for the convolutional neural network. The progression of the training can be seen in Figure 3. Although the loss goes down steadily, the prediction accuracy seems to rise very slowly if at all.

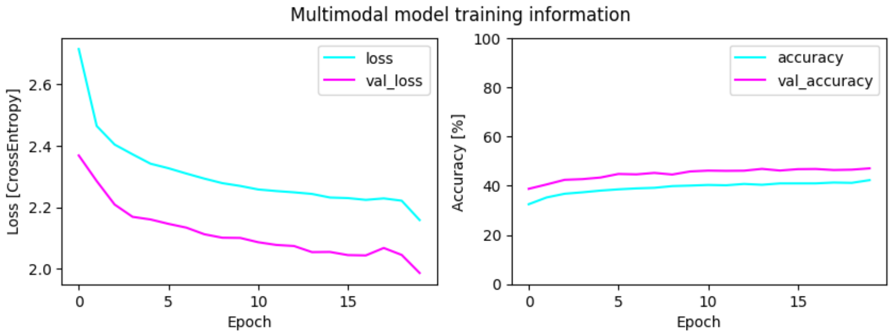

Figure 3. Visualization of the training loss value for train and validation datasets (left) and the accuracy value for training and validation datasets (right)

The model accuracy was tested on the validation split of the dataset. From the 12088 images in the dataset, the model correctly predicted 47.08 % of them (Top 1 accuracy). When testing the top 5 accuracy i.e. how often is the correct label is in the top 5 most probable labels in the model predictions, the model achieved a 89.26 % accuracy. From the testing predictions we determined that the model seems to overfit towards the class 32 ("illustrative body language"), as it is the largest individual class in the training data. This is expected behaviour as with unbalanced data the models usually overfit to the largest individual classes.

Without taking into account the time that it takes the Movenet to create the skeleton dataset, the model processes the whole validation set in approximately 92.6 seconds which means that the model is able to process 130.6 frames per second. However when the data processing into the skeleton datapoints is taken into account the runtime over doubles as the data has to be processed twice. The total runtime goes up to approximately 218.4 seconds on the validation data which means that the rate of data processing is 55.3 frames per second.

5 DISCUSSION

Some of the gestures are nearly imperceptible and almost similar as some other gesture inside the same main class. This feature brings a great challenge for the classification problem. Since the faces of the individuals have been blurred, some of the gestures become almost impossible to recognize since no clear information is seen from the head area. Therefore many similar head-related gestures are nearly impossible to classify correctly even manually.

The performance of the model could be improved with suitable hyperparameter optimization and ablation study. With hyperparameter optimization one will try different values for example for the initial learning rate or batch size and see how the changes will affect the model performance. Based on the results, the most optimized hyperparameter values could be chosen to be used in the final model. The ablation study aims to study which of the components in the model do affect the model performance significantly and which other do not [6]. This will increase the understanding of how the model behaves. If some component does not affect the performance considerably, it could be even removed or simplified from the final model. This type of optimization might decrease the complexity of the model and therefore the training and testing times would decrease.

In this task the data between different classes is extremely unbalanced which brings its own challenges to the training of the model. The problem could be reduced with data augmentation, where one creates more data samples for small classes and modifies the new samples in some minor ways so that they are not just copies of the existing samples [7]. The modifications could be mirroring, rotating, adding noise or cropping. This method will bring more training data and will balance the class distribution and thus it may improve the classification accuracy since the model does have more samples to learn the smaller class's features. Now based on the obtained results, most of the predictions are on class 32 (illustrative body language) since this class has the most number of samples. By augmenting the data on other classes this type of behaviour could be reduced.

Since the MoveNet was not able to be implemented inside the multimodal model used, the time taken to test the model's accuracy was quite long. For further development of this application, the MoveNet could be implemented inside the model and thus the time taken to test the performance could decrease significantly. Then it would be enough to feed only one image to the model as an input and it would also decrease the complexity of the model codes.

REFERENCES

[1] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10631–10642, June 2021.

[2] Kun Li, Dan Guo, Guoliang Chen, Xinge Peng, and Meng Wang. Joint skeletal and semantic embedding loss for micro-gesture classification, 2023.

[3] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding, 2019.

[4] Martín Abadi *et al.* . TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. [Online]. Available at: `https://tensorflow.org/`.

[5] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[6] Xuan Li, Congao Wang, Ding Ma, and Xiangqian Wu. Feature refinement from multiple perspectives for high performance salient object detection. In Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji, editors, *Pattern Recognition and Computer Vision*, pages 56–67, Singapore, 2024. Springer Nature Singapore.

[7] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.