UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Niyi Solomon Adebayo

Facial Expression Recognition Based on Deep Learning on EMO2018 dataset

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor(s):

Professor, PhD Gholamreza Anbarjafari

Ms, MSc Liina Juuse

Tartu 2021

Abstract:

We conducted data processing of the first systematic database of facial expressions. We proposed two video-based facial expression recognition models capable of discriminating facial expressions in the video into seven basic emotions using deep learning algorithms. The first model is frame attention networks (FAN), which takes video containing facial expressions as its input and produces a stable representation. The system comprises of two modules. The feature embedding module embeds face images into feature vectors, and the frame attention module adaptively cumulate the feature vectors to form a single discriminative video representation. The second model, a deep learning approach based on an attentional convolutional network with a minimal number of convolutional layers, can focus on the vital region of the face. Both methods achieve state-of-the-art performance compared with other high performing models on similar datasets.

Keywords:

Facial expression recognition, frame attention networks, convolutional neural network, spatial transformer network, deep convolutional network.

CERCS:

P176 Artificial intelligence

T111 Imaging, image processing

Näoilme tuvastus masinõppega, EMO2018 andmetel.

Lühikokkuvõte:

Viisime läbi esimese näoilmete süstemaatilise andmebaasi andmetöötluse. Pakkusime välja kaks videopõhist näoilmetuvastuse mudelit, mis on võimelised eristama videos näoilmeid seitsmeks põhiemotsiooniks, kasutades süvendatud õppimise algoritme. Esimene mudel on raami tähelepanu võrgud (FAN), mille sisendiks on näoilmeid sisaldavad videod ja stabiilne esitus. Süsteem koosneb kahest moodulist. Funktsiooni kinnistamise moodul kinnistab näokujutised funktsioonivektoritesse ja kaadritähelepanu moodul koondab funktsioonivektorid adaptiivselt, moodustades ühe diskrimineeriva videoesituse. Teine mudel - süvendatud õppimise lähenemisviis, mis põhineb tähelepanelikul konvolutsioonivõrgustikul, kus on minimaalne arv konvolutsioonilisi kihte - võib keskenduda näo olulisematele piirkondadele. Mõlemad meetodid saavutavad tipptasemel jõudluse võrreldes teiste sarnaste andmekogumite suure jõudlusega mudelitega.

Võtmesõnad:

Näoilme äratundmine, raami tähelepanuvõrgud, konvolutsionaalne närvivõrk, ruumitrafovõrk, sügav konvolutsionaalne võrgustik.

CERCS:

P176 Tehisintellekt

T111 Pilditehnika

Acknowledgements

Foremost, I thank God Almighty for His grace, mercy, and blessings upon my life because, without Him, this research will not be possible.

I am extremely grateful to the University of Tartu and everyone whose help was valuable during my BSc program and in this research.

In particular, I appreciate my Supervisor, Professor Gholamreza Anbarjafari (Shahab) and Miss Liina Juuse for their sincere and valuable guidance throughout this research.

These acknowledgements are not complete, without appreciating my family for their support.

TABLE OF CONTENTS

1. INTRODUCTION

A facial expression can be described as the motions or positions of the muscles beneath the skin of the face. Facial emotion expressions are an inevitable portion of any interpersonal communication. Facial expression recognition (FER) in recent years has been gaining increasing attention in the research community and industry due to its range of applications such as affective computing, intelligent environments, and multimodal human-computer interface (HCI). This research work stems from the need to utilise the multimodal data set code-name EMO2018 data set containing two sub-data types. Namely, video data having a facial expression of the emotion of 108 participants and electroencephalogram (EEG). The recording of brain bioelectric activity (in an EEG method) allows for a more accurate assessment of the extent to which emotions were blocked or expressed (Hosseini et al. 2021). However, on a preliminary investigation, it might be challenging to work with EEG and video data simultaneously because of the complexity of individual data. Hence my proposed line of action will be to work with the somewhat easier data, which is the video data.

1.1 Application

Many potential applications will benefit from automatic facial expressions such as; displays of genuine and unfelt emotional states, improved human-computer interaction, improved human-robot interaction for assistive robotics, and treatment of chronic disorders, assisting investigation conducted by police forces would be just a few.

According to Kulkarni et al. (2019) three significant ways in which emotional facial expressions are intentionally manipulated are: an expression is stimulated when any genuine emotion does not accompany it, masked when the expression corresponding to the felt emotion is replaced by a falsified expression that corresponds to a different emotion or neutralised when the expression of genuine emotion is inhibited while the face remains neutral. It has been suggested that displays of unfelt emotions would be betrayed by the leakage of their genuine emotional states through their nonverbal behaviour, such as subtle facial expressions of short duration (Porter and Brinke, 2008).

Although a variety of studies have focused on the evaluation of how genuine some FEEs might be while relying on the analysis of still images, i.e. static images, not much attention has been paid to dynamic images. In a naturalistic setting, FEEs are always perceived as dynamic facial

displays, and it is easier for humans to recognise facial behaviour in video sequence rather than in still images (Kulkarni et al., 2019).

2. LITERATURE REVIEW

Facial expression recognition is the fundamental research direction in many disciplines; computer vision, pattern recognition, artificial intelligence, and human-computer interactions (HCI) in general. Human-computer interaction (HCI), for example, according to Karray et al. (2008), is an essential part of affective computing. The worth of any piece of technology, no matter how sophisticated, is in its functionality and usability to humans. HCI systems such as emotion recognition systems as the world moves towards a state where computers are more ubiquitous, it will become essential that machines recognise and interpret all clues, both implicit and explicit expressed. A natural human-computer interaction cannot be based solely on explicitly stated commands such as traditional computers, through command lines, keyboard etc. Computers will need to detect and decode the numerous behavioural signals based on which to infer one's emotional state at any given moment, e.g. Siri virtual assistant from Apple Inc., Amazon Alexa virtual assistant AI by Amazon. The need to detect and decode the numerous behavioural signals from humans is an essential piece of the puzzle that has to be surmounted to predict accurately people's intentions and future behaviour.

According to Wei and Zhang (2020), before the emergence of deep learning networks, most earlier facial expression recognition algorithms used manual feature extraction and shallow learning. Hitherto to 2013, numerous facial expression recognition tournaments produced some popular data sets such as Fer2013 and EmotiW, in which a large number of authentic expression data was collected. Because of the increased amount of data gathered over these periods, traditional methods have become incapable of meeting the requirements of accuracy, efficiency, performance, and other evaluation metrics. Deep learning network technologies have thus become increasingly used in the field of emotion expression recognition.

The current chapter aims to introduce numerous mainstream facial image pre-processing techniques for both static and dynamic images and discuss the current mainstream feature learning deep networks. Furthermore, it presents popular emotion expression databases and neural networks of static and dynamic images, respectively. Finally, I analyse the deep learning expression recognition systems.

2.1 Pre-processing

One of the first and essential processes in machine learning is to conduct data pre-processing to make data fit for input purposes into a neural network (Al-Jabery et al., 2020). There are numerous variations that are not related to facial expressions themselves but significantly affect the results (e.g., different backgrounds, lighting, head poses). Before training deep neural networks, there is a need to use many pre-processing methods to calibrate and regulate the visual semantic information of faces, and such semantic correction can be done as part of data pre-processing which can be done in 3 stages.

2.1.1 Face extraction

A commonly used technique to carry out face extraction is by adopting the Histogram of Orientation Gradient (HOG). However, according to Domnich and Anbarjafari (2021), this method is not very effective for some dataset because of the rich representation of skin colour in the data. The given facial feature extractor could not extract faces for all subjects. Hence, they proposed a better approach called Multitask Cascade Convolutional Networks (MTCNN).

2.1.2 Decreasing frame size

When datasets contain video data with high frame rates, it is necessary to decrease the frame rate to be able to fit the GPU capacity of the training computer. To achieve this, basic clustering algorithms (such as K-means, KNN) can be used to determine the most distinguishable frames for each video.

2.1.3 Data Augmentation

When videos data in a dataset are limited, as it usually is, and since deep learning methods do better on large datasets. Data augmentation is the go-to technique to manipulate dataset without distortion. Methods employed include horizontal flip, zoom, random rotation and brightness augmentation, among other techniques.

Universally, deep neural networks require a massive amount of training data to ensure good robustness of accuracy in the recognition task. However, most open-source standalone dataset has largely proved to be insufficient, so there is a need to enhance the data in data set (Singh et

al., 2019). Data enhancement or augmentation can be broadly divided into two categories: offline data enhancement and online data enhancement. Offline enhancement refers to main methods through image processing operations to expand the database. These operations include image rotation, horizontal flip, zoom, et cetera. Online data enhancement, on the other hand, refers to methods generally integrated into the deep learning toolboxes. During the training process, the input samples are randomly rotated and horizontally flipped, resulting in a data set 10 times larger than the original database. Because the light intensity of each frame in a dynamic image might be relatively different, many studies of dynamic facial expression recognition utilise frame aggregation. This means a feature vector is utilised to substitute this time series. Frame aggregation is classified into two groups: Decision-level frame aggregation and feature-level frame aggregation. Between them, Kahou et al. (2013) suggested decision-making frame aggregation for:

(a) a sequence of above ten frames, the total number of frames is partitioned into ten independent frames grouped by time, whereas the probability vectors are averaged.

(b) For sequences of less than ten frames, the methodology follows that the frames are expanded to 10 frames by sequence is expanded to 10 frames by evenly duplicating the frames.

2.2 Face alignment

In numerous facial emotion recognition studies, face realignment is a necessary pre-processing procedure (Martinez, 2019). The commonly used face alignment techniques are listed in Table 1 below. Active appearance model (AAM) is an image segmentation algorithm based on the active appearance model. The multi-object tracking (MOT) method and the Dual-Regularised Matrix Factorisation (DRMF) method use a local-based technique to represent the global feature by local appearance information in the neighbourhood of each key point. Besides, some discriminating models directly map the appearance of the image to the position of the marker using a series of regression functions and then display better results, for example, Support distribution machine (SDM), Labeled-Based Forecasting (LBF), and Incremental algorithms. Recently, deep neural networks have become widely used in face localisation. Cascaded CNN algorithm and Multi-Task Cascaded Convolutional Neural Networks (MTCNN) algorithm bootstraps simple network cascades to synthesise a strong classifier to achieve a better face alignment.

Table 1

Summary of common methods for face alignment (Wei and Zhang, 2020)

Type		Points	Speed	Performance
Holistic	AAM	68	Fair	Poor generalisation
Part-based	MoT	39/68	Slow	
	DRMF	66	Fast	Good
Cascaded regression	SDM	49		Good/very
	LBF	68		
	Incremental	49	Fast	Good
Deep learning	Cascaded CNN	5		Good/very
	MTCNN	5	Fast	Good

Note. General comparison of holistic, part-based, cascaded regression, and deep learning methods of face alignments models. Points indicate number of facial feature points, speed indicates relative speed of execution of the respective algorithm and performance indicates relative accuracy of the respective models.

2.3 Expression Intensity network

There is a subtle variation in the intensity of expression of videos, and intensity in this case refers to the extent to which frames represent an expression in the dynamic image. Generally, an expression is best expressed roughly in the middle position, which is the intensity peak. Most techniques focus on the neighbourhood of the peak and ignore the trough frames at the beginning and end. Zhao et al. (2016) proposed a Peak-Guided Depth Network (PPDN) for expression recognition with constant intensity. PPDN takes pairs of peak and non-peak facial emotion expressions from the same person as input and then uses the L2 norm loss to minimise the distance between the two images. By using L2 norm loss It modifies the loss function by adding the penalty (shrinkage quantity) equivalent to the square of the magnitude of coefficients. Zhao et al. (2016) used peak gradient suppression (PGS) as a backpropagation mechanism to approximate the characteristics of peak expressions with features of non-peak expressions. At the same time, the gradient information of the peak expression is ignored in the

L2 normalization minimisation to avoid inversion which is a phenomenon where identifying inverted faces compared to upright faces is much more difficult than doing the same for non-facial objects.

2.4 Problem

The major issues are how to achieve the objective of this project which is to design a neural network technique that is capable of discriminating and automatically detecting emotional states (such as expressed or blocked emotion). The outcome of the project is important both from a psychological viewpoint and for understanding human-machine interaction (as discussed in the "Application" section above).

Kulkarni et al. (2019) showed that overall, the problem of recognising whether facial movements are expressions of authentic emotions or not could be successfully addressed by learning spatio-temporal representations of the data. For that purpose, they proposed a method that aggregates features along fiducial trajectories in a deeply learned space.

Finally, in line with the literature, temporal information is essential in improving facial expression recognition algorithms. While the technique used by Kulkarni et al. (2019) was by no means an exhaustive study, they suggested that their approach disregards some of the temporal information for compactness. They then recommended that other, more powerful sequential learning methods, like Recurrent Neural Networks, might be employed with better results. Hence, I will be exploring this technique and different similar ways since the type of data used has dimensionality similarity to the EMO2018 data set.

3 THE AIMS OF THE THESIS

The research project proposes the following main objectives (research questions):

1. To conduct data processing of the first systematic database of facial expressions.
2. To analyse basic facial expressions using deep learning algorithms
    a. Do deep learning algorithms reliably discriminate between basic emotional responses induced by images?

4 MATERIALS AND METHODS

Overview of materials and methods

In this section, we first list the equipment used and later discuss the overview of two models that were used in the implementation of facial emotion recognition architecture. Then, we describe the main phases of the two approaches.

The research of our deep learning model that is capable of discriminating facial emotions expressions consists of 4 phases, which include: (1) Data collection, (2) Data processing, (3) Model design and model training, (4) Model Evaluation.

4.1 Materials

The materials, equipment are more broadly described in the methodology part of this work. However, a brief overview is given as thus:

Hardware:

GoPro HERO-4,

Camera stand,

Halogen lamps,

Computer (3× Nvidia RTX 3060 12GB VRAM)

Software:

There was numerous software used at different stages of the research, and the major ones are listed below:

Python programming language,

NVIDIA® TensorRT™ 8.0.0,

Python OpenCV,

PyTorch

TensorFlow' Keras library

4.2 Methods

4.2.1 Data collection

In our data collection phase, we collected multimodal data (EEG and video) in order to create the first systematic multimodal database of facial emotion expressions.

The recordings of EEG and facial expressions were conducted in an electrically shielded quiet room in the Experimental Psychology Laboratory of the University of Tartu. The room was dimly lit by both sides of the participants' face in order to achieve better measuring quality. The lights were not moved throughout the entire duration of the experiments. The participants were seated on a comfortable office chair approximately 0.9 to 1.2 meters from the computer screen. The camera (GoPro HERO4) recording the facial expressions of participants was attached to the top of the computer screen. The participant was instructed to use a mouse to answer questions provided by the test program. Blocks of emotional pictures or words were displayed for 6 seconds, followed by questions to the participant. The experiment lasted approximately 40 minutes (varied based on the answering speed of the participant). At the time of the experiment, only the participant and experimenter were in the laboratory.

4.2.2 Data processing

In this phase, we focus solely on the video data. To make the long continuous video data useful as input into the model, we trimmed the video into short clips of individual participants using custom made Python algorithm software which uses the trigger beep (the beep in the video was used during the recording to signal to participants to start facial expression) in the video to signal the start of trim of the video and the next beep simultaneously signifies the end of trim of for the current participant and the start of the next. There was a trigger beep in the video clip corresponding to the emotional trigger onset and the start from which the participants start expressing the allocated emotion during the experiment. We further trimmed down the videos into short 6 seconds clips for individual participants using self-made algorithms**.** Each 6-second video contained an expected emotion and the emotion that is being expressed on the video. The next task was to mark the video clip with frame-stamps "the start" and "the end" of the expressed emotion. Video marking or frame marking in this case ensured that each participant video clip contains the expected expressed emotion and no extra facial emotion expression is allowed.

Additionally, the peak of the emotion was marked by a third click that is present in-between the two frame-stamps.

The start and end of emotion are defined as muscle tension and relaxation. When lip muscles start to tense up in case of smile expression, it is the start of emotion, and when they relax, then it is the end of emotion. Between the start and end mark of emotion, the third marker, "peak of emotion", was assigned (third frame-point, third press of “S” in the software). The "peak of emotion" mark was placed at the frame in which the person labelling (persons labelling here refers to human labellers who validated each video clip tag correspond to the expressed emotion in them) considered the expression to be the most intense. If there are several similar high-intensity frames, then the middle one was picked. In the case where emotion presented by the subject did not correlate with emotion expected, the following actions were taken: mark "non-correlating expression" was assigned. In case of more than one emotion present in a video clip, the following action was taken: expected emotion mentioned in the GUI (Graphical User Interface) was marked. In case the main emotion starts with some additional muscle movement, for example, lips move to the side, and then the subject expresses surprise, the emotion itself had to be marked instead of any pre-movement that could be linked to personality. In the case of 'poker face' (i.e., no muscle movement in the face), the video clip was marked as is. However, if subjects express non-neutral emotion, the following action was taken: we marked it as non-neutral emotion and placed an additional mark "non-correlating expression".

In the case where emotion intensity changed from heavily expressed to less expressed, the start frame and end frame was marked regardless of the intensity. In the case where the video started with an emotion already being expressed – we marked the first frame and the frame at which the emotion ended. When the subject pointed eyes down or to the side – no special action was taken and we followed the regular labelling protocol. Where the subject started to scratch their body – we placed an additional marker "abnormal limb behaviour" (button “T”). When emotions began to get mixed over time – we marked only the expected emotion. If the subject obstructed his or her face, we placed the marker "heavy or low face occlusion". The marker "heavy occlusion" was chosen when more than 60% of the face was covered. The mark" low-occlusion" was chosen when more than 30% of the face was covered.

If a new edge case is detected and appropriate action can not be taken – we mark such videos as "other" (button “X”). The marks were saved as meta-data in a text file format and named accordingly in the following format with tab separation:

<emotion start frame> <emotion end frame> <emotion peak frame> <non-correlating mark> <abnormal limb behaviour> <low-occlusion> <heavy occlusion> <other>

Example:

3 89 43 1 0 0 0 0

The metadata file is then used in a custom-made Python software to re-trim the video clips and label them as appropriate.

Table 2

GUI interface commands of video trimming algorithm

Command	Button
Forward frame	D
Backward frame	A
Mark frame-stamp	S
Save	E
Quit video	Q
Reset all marks	W
Non-correlating expression	G
Abnormal limb behaviour	T
Low face occlusion	U
Heavy face occlusion	P
Other	X

This labelling software was used to correctly separate the data into 7 label folders containing video data expressing only the correct emotions. The label categories used were anger, disgust,

fear, happiness, sadness, surprise and neutral. A visual sample of the expression category is given in the Figure 1 below.

Figure 1

The facial expressions used to convey happiness, surprise, sadness, fear, disgust, contempt, and anger are similar across developmental periods and cultures. They even occur in blind people who have never witnessed these expressions, leading psychologists to call them universal facial expressions.

Note. This is a copyright-free example of what was used in the experiment.

4.2.3 First method: Implementing Frame Attention Networks

Figure 2 below illustrates the framework for the proposed frame attention network. The model takes as input a video of facial expressions and produces a dimension of fixed features representation. The whole model consists of two modules: the feature embedding module and the frame attention module. The feature embedding module is a deep convolutional neural network which embeds each face image into a feature vector. This deep CNN is initialised on Deep residual networks (RESNET18) pre-trained on ImageNet. ImageNet is an extensive database or dataset of over 14 million images. It was designed by academics intended for computer vision research (Devopedia, 2021). The frame attention module learns two types of attention weights, i.e. self-attention weights and relation-attention weights, which were used to adaptively accumulate the feature vectors to form a unified or single discriminative video representation. A video with n frames is denoted as V, and the frames therein are denoted as I_1, I_2, …I_n, and the facial frame features are denoted as f_1, f_2, …, f_n.

Self-attention weights: As regards individual frame features, the model applies a fully connected layer and a sigmoid function to assign self-attention weights. The self attention weight of the frame number "i" is mathematically defined by:

$$\alpha_i = \sigma(f_i^T \times q^0) \qquad (1)$$

From the above equation, $\mathbf{q^0}$ is the parameter of fully connected layers (FC), $\boldsymbol{\sigma}$ denotes the sigmoidal function. With these self-attention weights, all the input frame features were aggregated into a global representation f'v as follows:

$$f'v = \frac{\sum_{i=1}^{n} \alpha i \times fi}{\sum_{i=1}^{n} \alpha i} \qquad (2)$$

Parameter f'v in equation 2 above represent a video-level global anchor for learning more accurate relation-attention weights.

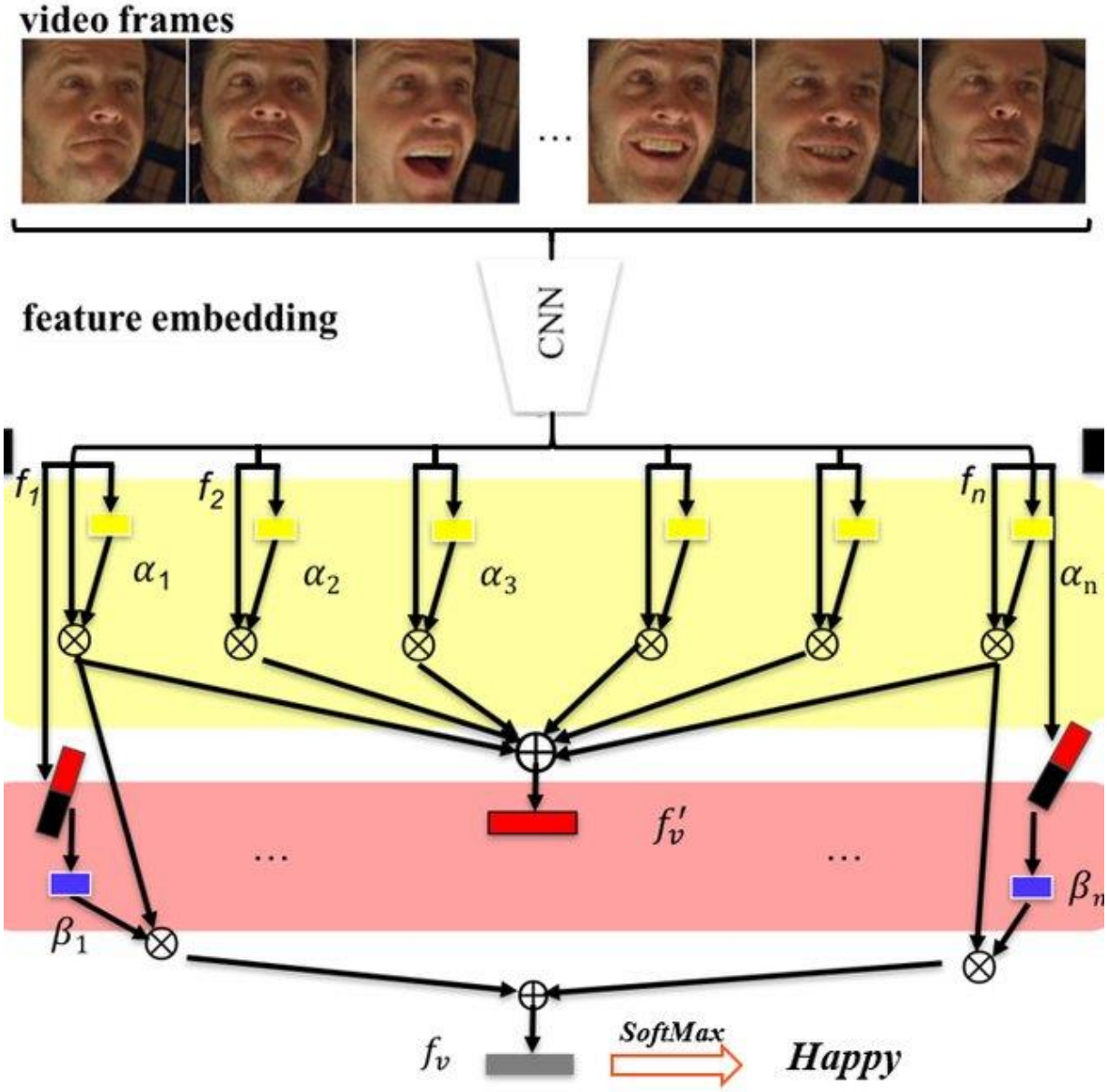

Figure 2

Pictorial representation of the proposed frame attention network architecture (FAN)

Relation-attention weights: Meng et al. (2019) suggested that learning weights from both a global features and local features is more reliable.

The self-attention weights learns with individual frame features and non-linear mapping. Since f'v innately contains the whole video, the attention weights can furthermore be refined by modelling the relationship between frame features and this global representation f'v.

The relation-attention weight of frame number "i" is formulated in equation 3 as:

$$\beta i = \sigma([f_i : f'v]^T \times q^1) \quad (3)$$

From equation3 above, q^1 is the parameter of fully connected layers (FC), and σ denotes the sigmoid function.

Conclusively, with self-attention weights and relation-attention weights, The model aggregates all the frame features into a new compressed feature parameter denoted as:

$$f_v = \frac{\sum_{i=0}^{n} \alpha i \times \beta i \, ([fi: f'v]}{\sum_{i=0}^{n} \alpha i \times \beta i} \quad (4)$$

The video frame was processed for face detection and face alignment in the python Dlib toolbox (King, 2009). The ratio of the bounding boxes were extended by 25% and then resize and then cropped faces to scale of 224×224 pixels. We implemented our method in the Pytorch toolbox (Paszke et al., 2019) which an open source machine learning library based on the Torch library developed by Facebook's AI Research lab (FAIR). The feature embedding was initialised by using ResNet18 (He et al., 2016) which has been pre-trained on MSCeleb-1M face recognition dataset and FER Plus expression dataset. For training, The batch size was 48 instances with K numbers of frames in every instance. For frame sampling in video, the videos were first split into K parts and then select randomly a frame from each segment. By default, K was set to 3 which divides the video into 3 segments: beginning of expression, peak of expression and end of expression.

In the implementation, as recommended by Meng et al. (2019), the stochastic gradient descent (SGD) method was used for optimisation with a momentum of 0.9 and a weight decay of 10^{-4}.

Using the momentum, Instead of depending only on the current gradient to update the weight, gradient descent with momentum substitute the current gradient with momentum. This is an accumulation of gradients. This accumulation is the exponential moving average of present and previous gradients. The learning rate (lr) was initialised to 0.1 and modified to 0.02 at 250 epochs, as recommended by Meng et al. (2019) and training stopped after 500 epochs.

4.2.4 Second method: Implementing deep-emotion using attention convolutional network

The deep-emotion facial expression recognition using attention convolutional network framework was proposed by Minaee et al. (2019). It uses an end-to-end deep learning framework based on an attentional convolutional network to classify the underlying emotion in the face images. Often times, improving a deep convolutional neural network relies on adding more neurons or layers or by better regularisations (e.g., spectral normalisation), especially for classification problems with a large number of classes. However, for facial expression recognition, due to the small number of classes, it was shown that using a convolutional network with less than 10 layers and attention was able to achieve good results, comparable to state-of-the-art models in several databases. In facial expression recognition, we only need to attend to the specific regions of the face to get a sense of the underlying emotion. Based on this observation, we added an attention specific algorithm, through spatial transformation network into our framework to focus on important region of the face (Minaee et al., 2019). Figure 3 below illustrates the model architecture.

The architecture consists of two main parts namely: feature extraction and the spatial transformation part (the localisation network). The feature extraction part consists of four convolution layers as depicted in Figure 3, each two convolution layers are followed by max-pooling layer and rectified linear unit (ReLU) activation function. They are then followed by a dropout layer and two fully-connected (FC) layers. The spatial transformation part consists of two convolution layers and each of the two convolution layers are followed by max-pooling and rectified linear unit ReLU and two fully-connected layers. After regression of the transformation parameters, the input is transformed to the sampling grid, depicted as "$T_\theta(G)$" producing the reshape data. The spatial transformer module fundamentally tries to focus on the most relevant part of the slide of images, by estimating a sample over the attended region. Here an affine transformation was used.

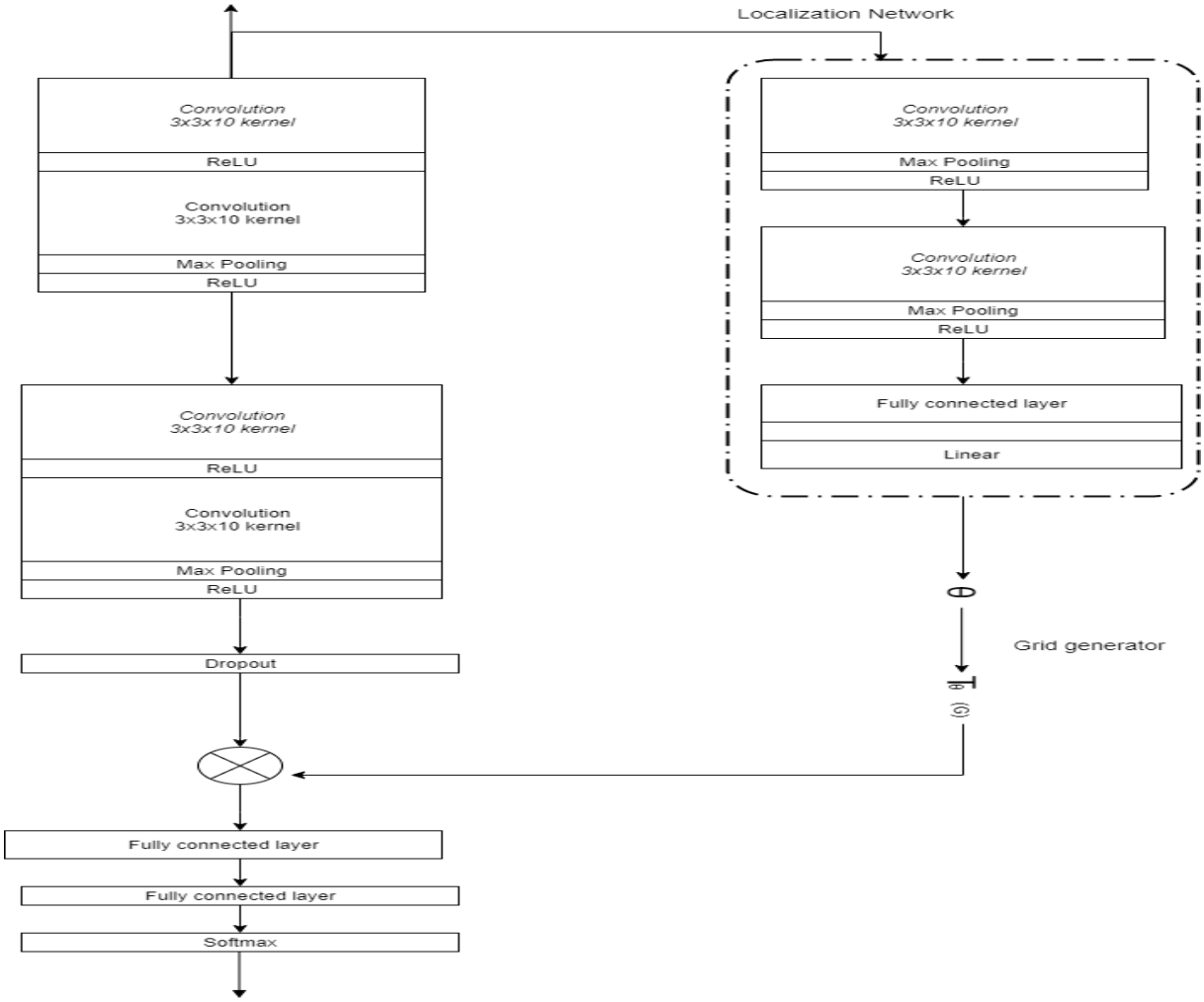

Figure 3

The proposed network architecture for deep-emotion using attention convolutional network (Minaee et al., 2019)

This particular model was then trained by optimising a loss function using a stochastic gradient descent approach (Adam optimiser). The loss function is essentially the summation of two losses, the classification loss (cross-entropy) and the regularisation term, which is the L2 norm normalisation of the weights in the last two fully connected layers.

$$L_{overall} = L_{classifier} + \lambda \, || \, \omega_{(fc)} \qquad (5)$$

The regularisation weight represented by Lambda (λ) in equation 1 above is tuned on the validation set (validation set is 10 percent of the total dataset). Adding both the dropout and L2 regularisation enables to train this model from scratch, even on smaller datasets.

The model was trained for 500 epochs from scratch on a crypto-mining computer graphically accelerated with three parallel synchronised NVidia RTX 3060 GPU, each with 12 GB of VRAM. The network weights were initialised with random Gaussian variables with zero mean and 0.05 standard deviation. For optimisation, Adam optimiser was used with a learning rate of 0.005 with weight decay. Minaee et al. (2019) reported that the Adam optimiser seemed to be performing slightly better than any other optimisers.

5 RESULTS

In this section, we report the test accuracies of our models on EMO2018 datasets on test data.

EMO2018 dataset contains 24,192 samples. The training data contains 16,934 samples (70 percent), validation contains 2,419 samples (10 percent of the dataset), and test data had 4,839 samples (20 percent of the dataset). By test accuracy, we mean classification accuracy of our models, which is the ratio of the number of correct predictions to the total number of input sample video.

5.1 Evaluation of Frame attention network

I evaluated the frame attention network model on our dataset (EMO2018 dataset) with comparisons to other several state-of-the-art models on a similar dataset in Table 3 below. We achieved a test accuracy of 92.30% on the test dataset. For a relatively fair comparison, I only detail the results obtained from the best single models in previous published works in literature. Unlike in our case, most of the comparison methods conduct data selection manually. The convolution neural network model of Zhang et al. (2017) only uses the last peak frames. Jung et al. (2015) selected a sequence of fixed-length for each input video.

5.2 Evaluation of attention convolutional network (Deep-Emotion)

We used the entire 16,934 sample videos in the training set to train the model, validated on 2,419 validation video clips, and report the model accuracy on 4,839 videos in the test set. We achieved a test accuracy of 89.69% on the test dataset. The results with a comparison of previous works on a similar dataset is shown in Table 3.

Table 3

Evaluation of our FAN and Deep-Emotion with a comparison to state-of-the-art methods reported in literature on CK+ database

S/N	Method	Training data	Test data	Accuracy(%)
1	FAN on EMO2018 dataset	All frames	All frames	92.30
2	Deep emotion on EMO2018	All frames	All frames	89.69
3	FAN on CK+ dataset	All frames	All frames	94.80
	Deep-emotion on CK+ dataset	All frames	All frames	98.00
4	PHRNN on CK+ dataset	All frames	All frames	98.50
5	DTAGN on CK+ dataset	Fixed length	Fixed length	97.25
6	CNN + Island loss on CK+ dataset	The last three frames and the first frame	The last three frames and the first frame	94.35
7	LOMO by on CK+ dataset	All frames	All frames	92.00
8	Fisher face on CK+ dataset	All frames	All frames	89.20
9	Salient Facial Patch on CK+ dataset	All frames	All frames	91.80
10	Convolution neural network and support vector machine on CK+ dataset	All frames	All frames	95.31
11	Incremental Boosting Constitutional Neural Networks on CK+ dataset	All frames	All frames	95.10

Note. Only those methods evaluated with 7 classes are included. Except for rows 1 and 2, other values were sourced from Meng et al. (2019) and Minaee et al. (2019).

6 DISCUSSION

In this section, we discuss the results obtained from from our evaluation and answer if our research goals were met.

Research goal one: To conduct data processing of the first systematic database of facial expressions.

From the results obtained from the "Data pre-processing" sub-section of "Methods", we were able to split the video stream into short videos containing individual participants. Furthermore, we annotated and trimmed the video data into clips that contained only the expected emotions. We carried out image processing on the dataset to make it fit as input into our model.

Research goal two: To analyse basic facial expressions using deep learning algorithms

From the result obtained, it can be seen that we successfully implemented a framework for facial expression recognition using an attentional convolutional network (Deep emotion) and frame attention network (FAN) and both deep learning algorithms reliably discriminate between basic facial emotional expressions.

SUMMARY

We propose two video-based facial expression recognition models using frame attention networks (FAN) and deep attentional convolutional network (deep emotion). Both models were capable of discriminating the 7 basic human facial expressions with accuracies similar to the state-of-the-art models on similar dataset. The frame attention network contains a self-attention module and a relation-attention module. The experiments on our dataset gave test accuracy of 92.30 % on the test dataset while deep attentional convolutional network (deep emotion) gave accuracy of 89.69 % on the test dataset.

REFERENCES

1. Al-Jabery, K. K., Obafemi-Ajayi, T., Olbricht, G. R., & Wunsch D. C. II. (2020). *Data pre-processing.* In K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, D. C. Wunsch II (Eds.), *Computational learning approaches to data analytics in biomedical applications.* (pp. 7-27). Academic Press. https://doi.org/10.1016/B978-0-12-814482-4.00002-4
2. Devopedia (2021). "ImageNet." Version 16, April 7. Accessed 2021-04-07. https://devopedia.org/imagenet
3. Domnich, A., & Anbarjafari, G. (2021). Responsible AI: Gender bias assessment in emotion recognition. *arXiv preprint.* https://arxiv.org/abs/2103.11436
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90
5. Hosseini, M. P., Hosseini, A., & Ahi, K. (2020). A Review on Machine Learning for EEG Signal Processing in Bioengineering. *IEEE reviews in biomedical engineering*. https://doi.org/10.1109/RBME.2020.2969915
6. Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2983-2991). https://doi.org/10.1109/ICCV.2015.341
7. Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., ... & Wu, Z. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 543-550). https://doi.org/10.1145/2522848.2531745
8. Karray, F., Alemzadeh, M., Abou Saleh, J., & Arab, M. N. (2008). Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, *1*, 137-159. https://10.21307/ijssis-2017-283
9. King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, *10*, 1755-1758. https://10.1145/1577069.1755843
10. Kulkarni, K., Corneanu, C., Ofodile, I., Escalera, S., Baro, X., Hyniewska, S., ... & Anbarjafari, G. (2018). Automatic recognition of facial displays of unfelt emotions. *IEEE transactions on affective computing*, *1*. https://doi.org/10.1109/taffc.2018.2874996

11. Martinez, B., Valstar, M. F., Jiang, B., & Pantic, M. (2017). Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, *10*(3), 325-347. https://doi.org/10.1109/TAFFC.2017.2731763
12. Meng, D., Peng, X., Wang, K., & Qiao, Y. (2019, September). Frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 3866-3870). IEEE. https://arxiv.org/abs/1907.00193
13. Minaee, S., Minaei, M., & Abdolrashidi, A. (20). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, *21*(9), 3046. http://doi.org/10.3390/s21093046
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, *32*, 8024–8035. https://arxiv.org/abs/1912.01703
15. Porter, S., & Ten Brinke, L. (2008). Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological science*, *19*(5), 508-514. http://doi.org/10.1111/j.1467-9280.2008.02116.x
16. Singh, P., Kar, A., Singh, Y., Kolekar, M., & Tanwar, S., (2019). Proceedings of International Conference on Robotics and Intelligent Control (ICRIC 2019)
17. Wei, H., & Zhang, Z. (2020). A survey of facial expression recognition based on deep learning. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)* (pp. 90-94). IEEE. https://doi.org/10.1109/iciea48937.2020.9248180.
18. Zhang, K., Huang, Y., Du, Y., & Wang, L. (2017). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, *26*(9), 4193-4203. https://doi.org/10.1109/TIP.2017.2689999
19. Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., & Yan, S. (2016). Peak-piloted deep network for facial expression recognition. In *European conference on computer vision* (pp. 425-442). https://doi.org/10.1007/978-3-319-46475-6_27.

Appendix

I. Training configuration for residual neural network (resnet18)

General train configuration available on PyTorch RESNET documentration ResNet | PyTorch

- lr - initial learning rate.
- epochs - the count of training epochs.
- batch_size - batch sizes for training (train) stage.
- input_size - input images dimension width and height in pixels.
- gpu_devices - list of selected GPU devices indexes.
- data_workers - how many subprocesses to use for data loading.
- dataset_tags - mapping for split data to train (train) and validation (val) parts by images tags. Images must be tagged by train or val tags.
- weights_init_type - can be in one of 2 modes. In transfer_learning mode all possible weights will be transfered except last classification layers. In continue_training mode all weights will be transfered and validation for classes number and classes names order will be performed.
- val_every - how often perform validation. Measured in number(float) of epochs.
- allow_corrupted_samples - number of corrupted samples ерфе can be skipped during train(train) or validation(val)
- lr_descreasing - determines the learning rate policy. patience - the number of epochs after which learning rate will be decreased, lr_divisor - the number learning rate will be divided by.

Full training configuration example:

```
{
 "dataset tags": {
   "train": "train",
   "vale": "val"
 },
 "batch_size": {
   "train": 64,
   "val": 64
 },
 "data_workers": {
   "train": 8,
   "val": 8
```

```
  },
  "input_size": {
    "width": 224,
    "height": 224
  },
  "allow_corrupted_samples": {
    "train": 16,
    "val": 16
  },
  "epochs": 500,
  "val_every": 1,
  "lr": 0.001,
  "lr_decreasing": {
    "patience": 30,
    "lr_divisor": 10
  },
  "weights_init_type": "transfer_learning"
}
```

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, Niyi Solomon Adebayo,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

Facial Expression Recognition Based on Deep Learning on EMO2018 dataset,

supervised by Professor Gholamreza Anbarjafari (Shahab) and Ms Liina Juuse

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Niyi Solomon Adebayo

20/05/2021