DEEP LEARNING BASED CLASSIFICATION OF 2D AND 3D IMAGES FOR FACIAL EXPRESSION RECOGNITION: COMPARISON STUDY

Fouzia Adjailia, Diana Olejarova and Peter Sincak

Dep. Cybernetics and Artificial Intelligence, Technical University of Kosice,
Kosice, Slovak Republic.

ABSTRACT

Facial expressions are an important communication channel among human beings. The Classification of facial expressions is a research area which has been proposed in several fields in recent years, it provides insight into how human can express their emotions which can be used to inform and identify a person's emotional state. In this paper, we provide the basic outlines of both two dimensional and three-dimensional facial expression classification with a number of concepts in detail and the extent of their influence on the classification process. We also compare the accuracy of two-dimensional (2D) and three-dimensional (3D) proposed models to analyse the 2D and 3D classification using comprehensive algorithms based on convolution neural network, the model was trained using a commonly used dataset named Bosphorus. Using the same experimental setup, we discussed the results obtained in terms of accuracy and set a new challenge in the classification of facial expression.

KEYWORDS

Convolution neural network, facial expression classification, bosphorus, voxel classification.

1. INTRODUCTION

One of the fundamental human traits is our ability to understand, to some extent, non-verbal signals coming from other people. Mimics, gestures and body language are important part of our daily interactions and are rooted in evolutionary signalling theory [1]. Even through emotion as is doesn't have one specific definition in literature, the term is taken for granted as is. Currently nine emotions are associated with distinct non-verbal expressions and have received cross-cultural support as universal. Research in this topic is supporting the importance of social function of emotion expression [2],[3]. Even through emotions can be derived from many sources (voice, body language, gestures, ...), facial expressions are currently the most popular source. One of the pioneer works by Paul Ekman on this topic identified 6 universal emotions - anger, disgust, fear, happiness, sadness, and surprise [4]. Later, Ekman adopted the FACS from Hjortsjö. The FACS is anatomically based system for describing all visually discernible facial movements. It breaks down facial expressions into individual components of muscle movement and became the standard for facial expression research. [5][6]. Facial expression recognition helps to generate visual representations for person's emotional state. The concept is particularly relevant to learning about people attention, mood and emotions. Facial expression recognition can have broad applications across diverse environments. For example, the concept of facial expression is a basic technique for identification of a person's feedback regarding an experience. it can be used in

David C. Wyld et al. (Eds): CCSEA, BIoT, DKMP, CLOUD, NLCAI, SIPRO - 2020
pp. 225-233, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101017

Health care to recognize and be aware about the emotional state that patients exhibit during their rehabilitation for a better assessment with treatment process by providing more attention to patients who need it [7]. In education in order to have a better understanding of the adaption of learners to the study material, and based on the analysis, an adjustment of the teaching methodology is made. User feedback by monitoring the user's and customers expressions while watching a movie, playing games, or do shopping can be critical for the industry to fundamentally understand the needs of users and customers and get feedback about their services or products for bigger profit and better marketing. In security and safety, applications in surveillance were designed based on facial emotion recognition in order to detect suspicious people.

Emotions can be classified by emotion models. Classification was based on two viewpoints on emotion - either are emotions discrete - the Categorical models, or dimensional and the Dimensional models}. In the categorical models, the core is, that all humans have an innate set of basic, or fundamental emotions, that are universally recognizable. Basic emotions are considered discrete because they can be distinguished by facial expressions and biological processes [8]. Popular example of basic emotions to this day are Ekman's 6 universal emotions [4] that were mentioned earlier. On the other hand, the dimensional models are based on system of emotions, that can be represented in more than one dimension. To achieve this, they incorporate intensity, valence, arousal dimensions according to the needs of the specific model. There have been developed multiple dimensional models, but only few became dominant and widely used like Circumplex model, Plutchik's model see Figure 1.

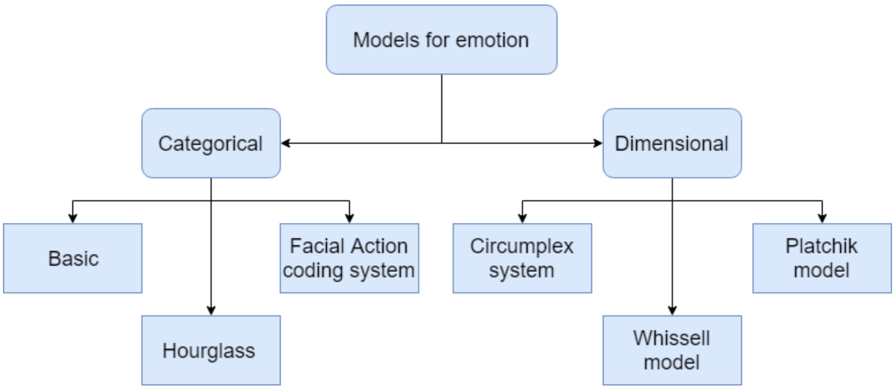


Figure 1. Emotion models.

The main contributions of this paper are as follows:

- Help new researchers understand basic notions of facial emotion recognition and its key components.
- Provide a new assimilation after reviewing the literature and related work.
- Introduce the standard 3D dataset called Bosphorus for facial emotion recognition databases along with their characteristics.
- Propose an experimental setup for comparison between 2D and 3D classification.

2. Literature Overview

In the current world, technology is advancing in a rapid speed. Certain devices and systems are getting more available, and affordable for the mass usage. One of these devices is camera. Its main function is capturing 2D images of 3D surroundings. This data can be used in multiple ways, but I would point out the one that is on the constant rise for a year now - emotion recognition. There is no doubt, that humans are naturally able to detect emotions of others based on their facial expressions. But why should be technology able to do the same? The number of fields that can benefit from emotion recognition is huge. Few of great examples are: detecting emotions during interviews, increasing safety levels for cars, video-game testing, improving emotional well-being of employees, marketing, evaluation of the engagement of students during classes (face-to-face or online), even lie detection. We can gradate tasks connected to emotion recognition to two groups: emotion recognition from 2D images and a level above is personalized emotion recognition.

[9] came up with DeXpression (Deep Convolutional Neural Network for Expression Recognition). It was the convolutional NN architecture for facial expression recognition and was independent of any hand-crafted feature extraction and performed better than the earlier proposed convolutional NN based approaches in the time of that research. Datasets used for this research were MMI (ongoing project, that aims to deliver large volumes of visual data of facial expressions) and CK+ (Extended Cohn-Kanade - labelled facial videos captured in a controlled environment). The architecture of this network can be broken into four parts, significant components are two FeatEx blocks (Parallel Feature Extraction), which were inspired by GoogleNet. These blocks consist of Convolutional, Pooling and ReLU (Rectified Linear Unit) layers and are split into two paths for a diverse input representation. They present the use of filters of different size as the reflection of the various scales at which faces can appear. The network was also built with cutting down the computational efforts. In conclusion, their network performed similarly to the state-of-the-art at the time - in average, a recognition accuracy for the CK+ dataset was 99.6\% and 98.36\% for the MMI dataset. Some misclassifications happened in the first few images of the sequence when the labelled emotion was not yet displayed. It was also pointed out, that with a closer look at some misclassified images within the datasets that the problem might be in the way people show emotions rather than in the NN. For example, image which depicts a person with a wide open mouth and open eyes and is labelled as Fear in dataset is classified as Surprise in the NN because other images depicting Surprise usually show people with wide open mouths and eyes.

Another research focused on NN architecture was [10]. They developed the architecture based on the knowledge that facial expressions are the results of specific facial muscles. They also wanted to solve the challenge of training deep model with small dataset and avoiding transfer learning. The idea is to drive the model to the relevant features (eyes, eyebrows, wrinkles in specific places, mouth, ...). They split their network to three parts: facial-parts component (mapping input to a relevance map, representing the probability that pixel is relevant for expression recognition), representation component (hidden learning representation is filtered by previously learned relevance map to respond strongly only on the most relevant features) and classification component (classification of highly discriminated representation of facial expression from previous part). They proposed three types of regularization based on the level of the available data annotations. Fully supervised (FS) regularization relies on class labels as well as coordinates of recognition key-points that create target relevance maps. Weakly supervised (WS) regularization does not require annotations because the loss function is defined to compensate this information. Hybrid fully and weakly supervised (HFWS) regularization is based on an idea to combine the strengths of previous methods and suppress their weaknesses. Predicted relevance maps of the model with fully supervised regularization can be seen on Figure 2. This model was tested on well-known FER datasets - CK+, JAFFE (Japanese Female Facial Expressions - acquired in controlled environment), SFEW (Static Facial Expressions in the Wild - selected frames from movies resembling the real world environment) and FER2013 (Facial Expression Recognition 2013 - images with spontaneous expressions collected in non-controlled scenarios) the results shown that it is comparable and even outperforms some of the state-of-the-art methods. The combination of fully and weakly supervised regularization was also proven to be better than the former methods, on CK+ in average the accuracy was 92.54% for FS, 93.37% for WS and 93.64% for HFWS.

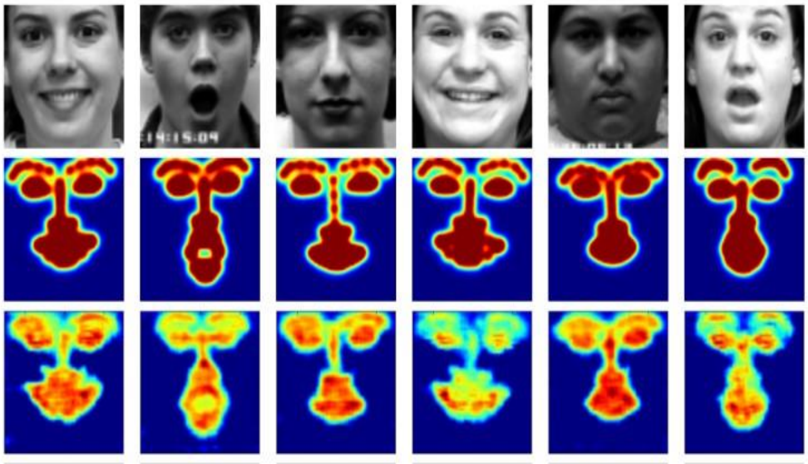


Figure 2. Examples of relevance maps with fully supervised regularization top row: input images middle row: target relevance maps bottom row: predicted relevance maps.

3. DATASET DESCRIPTION

One of the most known and used databases for emotion recognition from 3D object is Bosphorus database [11]. The Bosphorus Database is intended for research on 3Dand 2D human face processing tasks including expression recognition, facial action unit detection, facial action unit intensity estimation, face recognition under adverse conditions, deformable face modelling, and 3D face reconstruction. There are 105subjects and 4666 faces in the database. There are 24 landmarks with 2D and 3Dcoordinates. 2D landmarks were manually put on every image and 3D landmarks were calculated using the 3D-2D correspondences. Facial data are acquired using structured-light based 3D system. Acquisitions are single view, and subjects were made to sit at about 1.5 meters away from the 3D digitizer.

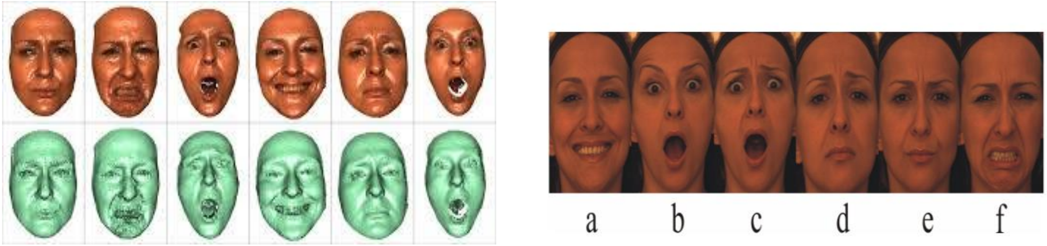


Figure 3. 3D scans from Bosphorus dataset (left) 2D images (right)

4. EXPERIMENTAL SETUP

In order to allow the classification process to be carried out in a more sophisticated way, we avoided applying any major pre-processing techniques for both 2D and 3D images as well as avoid applying data augmentation. We used 453 two dimensional as well as 453 three dimensional images. We split the provided images into 80% training and 20% for validation.

Figure 4. shows the analyse of the training and validation set to understand the labels present in the data and the overall distribution of labels.

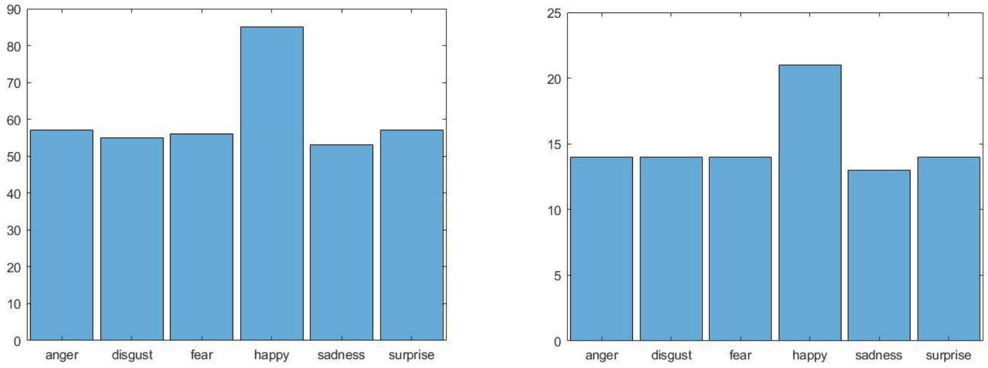


Figure 4. Distribution of the training (left) and validation (right) sets.

The following hyperparameters were used for both experiments A1 and A2:

- **Mini Batch Size:** a mini batch is a subset of the training set that is used to evaluate the gradient of the loss function and update the weights.
- **Learn Rate Schedule:** Option for dropping the learning rate during training
- **Learn Rate Drop Period:** Number of epochs for dropping the learning rate
- **Max Epochs:** Maximum number of epochs to use for training
- **Initial Learn Rate:** Initial learning rate used for training.
- **Solver name:** Solver for training network.

Table 1. Hyperparameters used in our experiment.

Mini Batch Size	20
Learn Rate Schedule	piecewise
Learn Rate Drop Period	40
Max Epochs	40
Initial Learn Rate	0.2
Solver name	Stochastic Gradient Descent with Momentum

The first experiment in described in section 4.1 where 2D images from Bosphorus dataset are classified using 2D convolution neural network. However, experiment carried out using 3D images are presented in section 4.2.

4.1. Experiment A1

4.1.1. Data pre-processing

Bosphorus dataset provides 453 2D images figure 5. The pre-processing step includes a set of steps as follows:

Table 2. Pre-processing steps.

1	Grey scale
2	Resize the images to 30x30

4.1.2. Architecture

The chosen network follows the architecture:

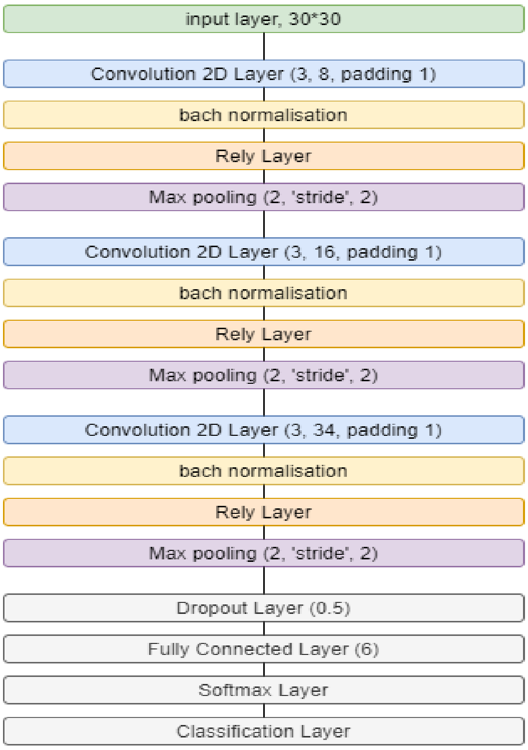

input layer, 30*30
Convolution 2D Layer (3, 8, padding 1)
bach normalisation
Rely Layer
Max pooling (2, 'stride', 2)
Convolution 2D Layer (3, 16, padding 1)
bach normalisation
Rely Layer
Max pooling (2, 'stride', 2)
Convolution 2D Layer (3, 34, padding 1)
bach normalisation
Rely Layer
Max pooling (2, 'stride', 2)
Dropout Layer (0.5)
Fully Connected Layer (6)
Softmax Layer
Classification Layer

Figure 5. Neural network architecture.

4.2. Experiment A2

4.2.1. Data pre-processing

Bosphorus dataset provides binary files on the form of (.bnt), these files contain Nx5 matrix where columns are 3D coordinates and 2D normalized image coordinates respectively. 2D coordinates are normalized to the range [0,1]. First, we created point cloud images figure 6. We applied uniform sampling to the point as a pre-processing step.

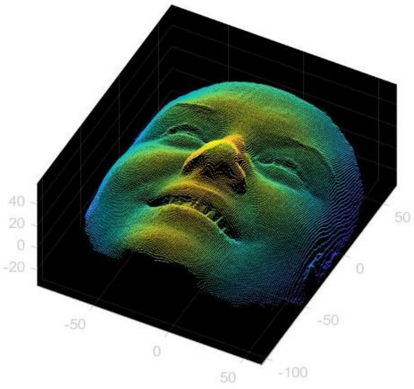


Figure 6. Point cloud file from Bosphorus data.

4.2.2. Voxelization

We generate the dataset for our work based on the pipeline show in Fig. 7.

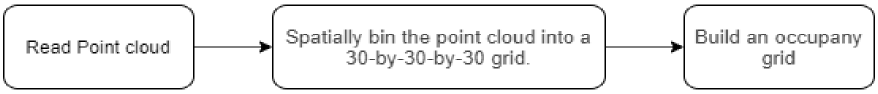


Figure 7. Pipeline for voxelization of the point cloud data.

Data generated is presented in figure.8

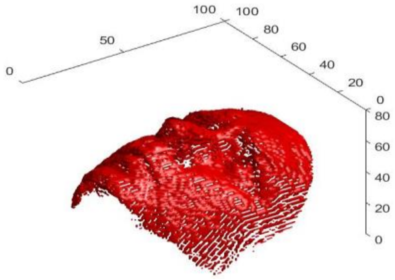


Figure 8. Voxel representation for Bosphorus data.

4.2.3. Architecture

The chosen network follows the architecture:

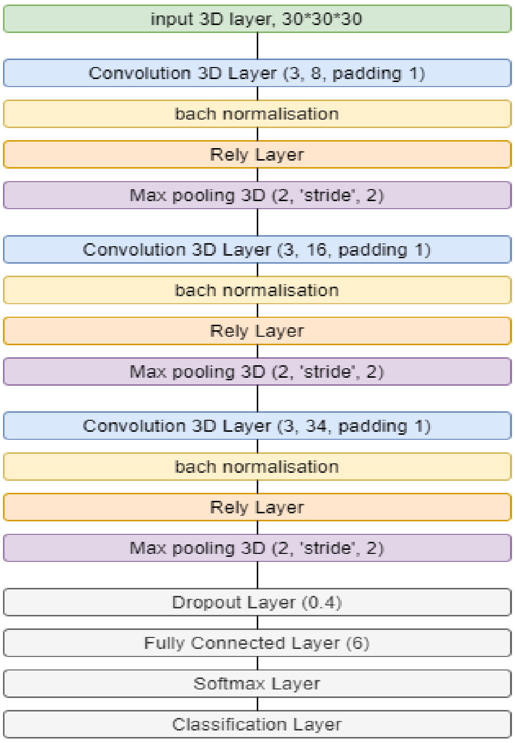

input 3D layer, 30*30*30
Convolution 3D Layer (3, 8, padding 1)
bach normalisation
Rely Layer
Max pooling 3D (2, 'stride', 2)
Convolution 3D Layer (3, 16, padding 1)
bach normalisation
Rely Layer
Max pooling 3D (2, 'stride', 2)
Convolution 3D Layer (3, 34, padding 1)
bach normalisation
Rely Layer
Max pooling 3D (2, 'stride', 2)
Dropout Layer (0.4)
Fully Connected Layer (6)
Softmax Layer
Classification Layer

Figure 9. Network architecture.

5. Results and Discussion

In order to evaluate the performance of the two experiments, we must compare results obtained from training and validation of the classification. The accuracy of 2D classification achieved 85.00 % for the training and 75.56% in the validation. As for the training loss it reached 0.37 and 0.70 for the validation loss. However, in the 3D classification, we obtained 65.0% of training accuracy and 48.89% of validation accuracy. As well as 0.4 in the training loss and 0.6 in the validation loss.

The confusion matrix is shown in figure 11. Based on the confusion matrix, it is shown that the network predicts well accurately on. However, for surprise and fear, the model gives wrong predictions.

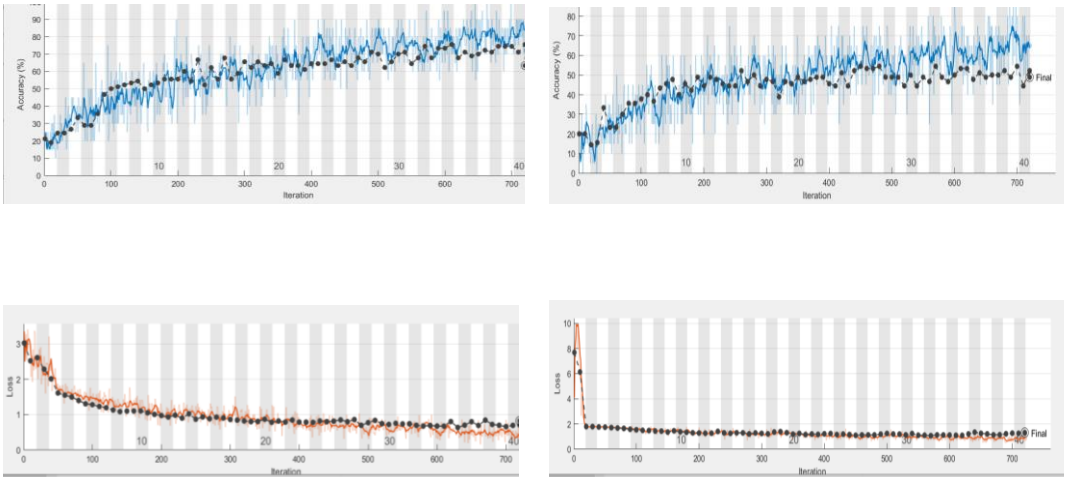
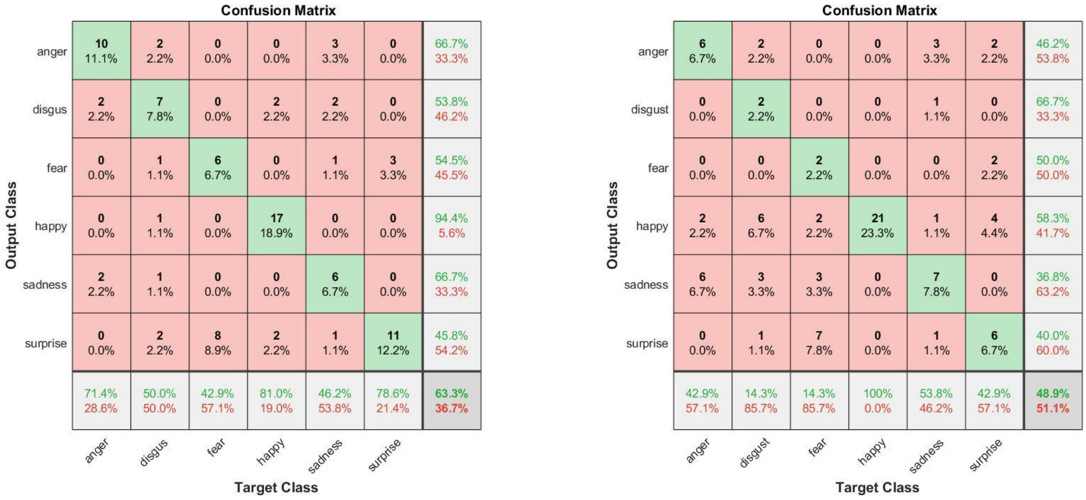


Figure 10. Training progress for experiment A1: Accuracy (top left), Loss (bottom left). Training progress for experiment A2: Accuracy (top right), Loss (bottom right)



Confusion Matrix

Output Class	anger	disgus	fear	happy	sadness	surprise	
anger	10 11.1%	2 2.2%	0 0.0%	0 0.0%	3 3.3%	0 0.0%	66.7% 33.3%
disgus	2 2.2%	7 7.8%	0 0.0%	2 2.2%	2 2.2%	0 0.0%	53.8% 46.2%
fear	0 0.0%	1 1.1%	6 6.7%	0 0.0%	1 1.1%	3 3.3%	54.5% 45.5%
happy	0 0.0%	1 1.1%	0 0.0%	17 18.9%	0 0.0%	0 0.0%	94.4% 5.6%
sadness	2 2.2%	1 1.1%	0 0.0%	0 0.0%	6 6.7%	0 0.0%	66.7% 33.3%
surprise	0 0.0%	2 2.2%	8 8.9%	2 2.2%	1 1.1%	11 12.2%	45.8% 54.2%
	71.4% 28.6%	50.0% 50.0%	42.9% 57.1%	81.0% 19.0%	46.2% 53.8%	78.6% 21.4%	63.3% 36.7%

Target Class

Confusion Matrix

Output Class	anger	disgust	fear	happy	sadness	surprise	
anger	6 6.7%	2 2.2%	0 0.0%	0 0.0%	3 3.3%	2 2.2%	46.2% 53.8%
disgust	0 0.0%	2 2.2%	0 0.0%	0 0.0%	1 1.1%	0 0.0%	66.7% 33.3%
fear	0 0.0%	0 0.0%	2 2.2%	0 0.0%	0 0.0%	2 2.2%	50.0% 50.0%
happy	2 2.2%	6 6.7%	2 2.2%	21 23.3%	1 1.1%	4 4.4%	58.3% 41.7%
sadness	6 6.7%	3 3.3%	3 3.3%	0 0.0%	7 7.8%	0 0.0%	36.8% 63.2%
surprise	0 0.0%	1 1.1%	7 7.8%	0 0.0%	1 1.1%	6 6.7%	40.0% 60.0%
	42.9% 57.1%	14.3% 85.7%	14.3% 85.7%	100% 0.0%	53.8% 46.2%	42.9% 57.1%	48.9% 51.1%

Target Class

Figure 11. confusion matrix for Experiment A1(left)and A2(right)

6. CONCLUSION

This paper has demonstrated a comparison study for facial expression recognition, for this purpose convolution neural network was used to classify 2D and 3D images from the Bosphorus dataset, for a sophisticated training process we applied the same experimental setup and hyperparameters. Among the observer studies, it appears that 2D classification overall produces slightly more accurate results than 3D classification. Further studies should be performed to create more 3D data and to develop methods for 3D data augmentation as well as proposing a better voxel classification model.

ACKNOWLEGMENTS

This research work was supported by APVV project 015-0730 "Cloud Based Human Robot Interaction" and FEI grant "Increasing Server Performance for Deep Learning" for the year 2020.

REFERENCES

[1] M. E. McCullough and L. I. Reed, "What the face communicates: Clearing the conceptual ground," Current Opinion in Psychology, vol. 7. Elsevier B.V., pp. 110–114, 01-Feb-2016, doi: 10.1016/j.copsyc.2015.08.023.

[2] M. Cabanac, "What is emotion?," Behav. Processes, vol. 60, no. 2, pp. 69–83, Nov. 2002, doi: 10.1016/S0376-6357(02)00078-5.

[3] J. L. Tracy, D. Randles, and C. M. Steckler, "The nonverbal communication of emotions," Current Opinion in Behavioral Sciences, vol. 3. Elsevier Ltd, pp. 25–30, 01-Jun-2015, doi: 10.1016/j.cobeha.2015.01.001.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124–129, Feb. 1971, doi: 10.1037/h0030377.

[5] "Handbook of Emotion Elicitation and Assessment - Google Books." https://books.google.sk/books?hl=en&lr=&id=9xhnDAAAQBAJ&oi=fnd&pg=PA203&dq=P.+Ekman+-+W.+V.+Friesen.+Facial+Action+Coding+System:+ A+Technique+for+the+Measurement+of+Facial+Movement.+1978.&ots=nLIwd-hBz0&sig=q74leRchuBapEuo2qY PbC6Ex3Ts&redir_esc=y#v=onepage&q=P. Ekman - W. V. Friesen. Facial Action Coding System%3A A Technique for the Measurement of Facial Movement. 1978.&f=false (accessed Jun. 21, 2020).

[6] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283, doi: 10.1145/2993148.2993165.

[7] M. Alhussein, "Automatic facial emotion recognition using weber local descriptor for e-Healthcare system," Cluster Comput., vol. 19, pp. 99–108, 2016, doi: 10.1007/s10586-016-0535-3.

[8] G. Colombetti, "From affect programs to dynamical discrete emotions," Philos. Psychol., vol. 22, no. 4, pp. 407–425, Aug. 2009, doi: 10.1080/09515080903153600.

[9] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep Convolutional Neural Network for Expression Recognition," Sep. 2015, Accessed: 21-Jun-2020. [Online]. Available: http://arxiv.org/abs/1509.05371.

[10] P. M. Ferreira, F. Marques, J. S. Cardoso, and A. Rebelo, "Physiological inspired deep neural networks for emotion recognition," IEEE Access, vol. 6, pp. 53930–53942, 2018, doi: 10.1109/ACCESS.2018.2870063.

[11] A. Savran et al., "Bosphorus database for 3D face analysis," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5372 LNCS, pp. 47–56, doi: 10.1007/978-3-540-89991-4_6.