EMOTION SKETCHES: FACIAL EXPRESSION RECOGNITION IN DIVERSITY GROUPS

Alexandru COSTACHE[1], Dan POPESCU[2]

In this paper we present an approach for facial expression recognition in images depicting persons of different ages, genders and ethnicities. We propose "emotion sketches" which are simplified representations of facial expressions, so that classification can be done regardless of the physical characteristics of persons portrayed in the analyzed images. Emotion sketches are based on facial landmarks identified in static images. The paper describes the process of obtaining the emotion sketches and validates the approach by training and testing three neural networks on them. We present and comment our experimental results.

Keywords: facial expression recognition; neural network; facial landmarks

1. Introduction

Facial expression recognition (FER), the practice of using computer vision in determining emotions displayed by persons by analyzing their facial appearance, has been an important research area for the past decades. While human observers still offer the most precise results in determining emotions (e.g., happiness), or states (e.g., pain) in persons, more and more systems have been proposed for automating this process. Important challenges in automated FER come from the quality of analyzed images and from significant differences in the physical appearance of persons. A recurring set-back encountered by researchers is the lack of large, publicly available datasets or databases for FER. Many authors propose either specific feature extraction to be used in analysis, as opposed to images themselves, or expanding the datasets through data augmentation. Paper [1] studies methods of performing accurate FER on datasets with a low number of samples, using a face descriptor to normalize their inputs before using a Convolutional Neural Network (CNN) for emotion detection. Most researchers employ CNNs as they offer good accuracy in visual classification problems.

According to literature, there are two preferred ways to tackle FER: conventional approaches, consisting of face detection, feature extraction and descriptor classification, and deep learning approaches, which reduce the

[1] Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: alexandru.costache0909@gmail.com
[2] Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania

dependence on physical characteristics of faces. Paper [2] suggests such a hybrid method suitable for temporal modelling of expression changes in an individual, while paper [3] gives a comparative study of using various feature extraction methods and various classification algorithms, tested on publicly available databases with annotated Action Units (AU). Paper [4] uses a CNN to detect emotions on the CK+ dataset with AU, reporting an accuracy of over 90%. A recent trend mentioned by paper [5] is the usage of Oriented Fast and Rotated Brief descriptors. The authors of paper [6] use discriminant tensor subspace analysis and extreme learning machines to classify micro-expressions from video clips. In paper [7], authors extract features regarding micro-expressions, using the sparse part of data decomposed by Robust Principal Component Analysis, as opposed to the low rank part usually used. Paper [8] also tackles the problem of micro-expression identification, using AU objective classes and deep learning.

Feature descriptors for facial analysis are often based on facial landmarks (FLs), which can be given as an array of coordinates or can be visually represented as white pixels on a black background. Paper [9] studies fast FL detection, using deep neural networks, fitting algorithms and principal component analysis, obtaining good results on various datasets, including thermal imagery. Some authors rely on cascading CNNs for FL extraction, which can prove efficient with important head pose variations, a situation in which a single CNN has limitations [10][11]. Three categories of FL detection algorithms were described: holistic methods, which build models representing the face appearance, constrained local model methods, which build local appearance models, and regression-based methods, which implicitly capture facial shape information [12].

In recent years, while usage of standard CNNs for image classification remains wide-spread, various alternatives have been proposed to address certain limitations of CNNs, such as its weak understanding of relations between identified parts of an object. Combining FLs into various graphs may give additional information which can aid in classification, such as in paper [13], where authors used Graph CNNs to analyze directed graphs with nodes consisting of FLs. Other network variations were proposed as alternatives for CNNs with varied degrees of success, such as residual networks, addressing the vanishing gradient problem, and capsule networks, which have a better understanding of spatial information in images [14][15].

In this paper, a system performing automated FER is described, with the goal of determining emotions regardless of the age, gender or ethnicity of persons depicted in images. For this, emotion sketches are defined, which are then used to train neural networks to classify the depicted emotions. Chapter II presents the technical details of our approach, chapter III shows experimental results, while chapter IV is dedicated to conclusions and future directions.

2. Methodology, Algorithms and Implementation

Deep learning is a class of machine learning algorithms suitable for extracting various features from inputs, with varying degrees of complexity. In image processing, most such algorithms are based on artificial neural networks, which can recognize increasingly complex structures depending on their architecture. In this paper, we use neural networks to identify emotions portrayed by facial images of persons. While most FER applications analyze grayscale images of faces, those images often contain information which can be safely discarded (e.g., facial hair). Another characteristic of those application is that they only perform well on faces whose features resemble those used in training (e.g., a network trained using faces of young Asians, will give poor results for faces of old Caucasians). Most image classification applications use CNNs, which are robust to scaling and translation of target objects, but are not robust to rotation, even a slight change in the orientation of an analyzed face making it unclassifiable. Of course, this can be addressed by increasing the training sets to include faces under different angles, but this greatly slows down the training process. Some limitations may be addressed by using FLs, as we have previously approached in paper [16], but only the need for similar facial features is removed, while the rotation problem is aggravated. Also, most authors only use FLs when comparing image sequences, studying FLs' movements rather than positions. In this paper we construct "emotion sketches" with the aid of FLs. Fig. 1 shows the main steps of our approach in performing FER.

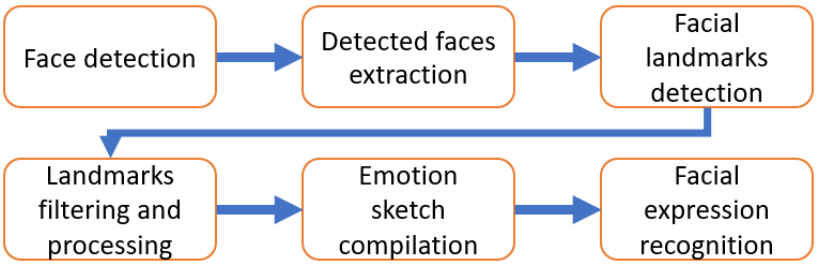

Fig. 1. Input image processing steps

Analysis can be performed on either static images or on video frames. Persons' faces are determined using the Haar Cascade Classifier, which offers results faster than more modern approaches (e.g., MTCNN), while still having an accuracy of over 95%. This is important, as we would like to achieve real-time results. Faces cropped from the source image are rescaled to 250x250 pixels and inputted to the DLIB Facial Landmark Detector, resulting 68 FLs for each face [17]. Image upscaling is an assumed hindrance in our methodology, as it often offers poor results. Also noted is the poor accuracy of the DLIB detector on images with low contrast, as FLs would often be mispositioned, as observed

during experiments. Still, it is the best compromise for relatively fast detection of an appropriate number of FLs with reasonable precision, and it is free to use. The detector outputs a map of sparse points in which correspondence between the FL's index and the anatomic component it represents is constant (e.g., the FL of the tip of the chin always has index 8). A lot of information from the source image is discarded when extracting FLs, leading to incomplete definitions of features (i.e., the system can hardly determine if an FL references an eye or an eyebrow). We reclaim some information by uniting FLs referring the same facial component, constructing a series of disconnected graphs. For this paper, we consider that to determine emotions it is enough to study the eyebrows, the eyes and the mouth of the subject, a total of 5 components, discarding the remaining FLs. Components are scaled and translated towards the center of the image to reduce the differences originating in the natural facial shapes of persons. The result is the "emotion sketch" image (ES) and the steps for obtaining it using FLs are detailed below. Each FL_i has coordinates (x_i, y_i).

1. We extract horizontal – FL_0 and FL_{16} (leftmost and rightmost FLs), and vertical references – FL_{27} and FL_{33} (top and tip of the nose)
2. We define target coordinates – $x_0' = 10$, $x_{16}' = 240$, $y_{27}' = 70$, $y_{33}' = 210$
3. We remove FLs except those of the eyebrows, eyes and mouth
4. We determine new position L_i' (x_i', y_i') using the equations (1) and (2):

$$x_i' = \frac{x_{16}' * (x_i - x_0) + x_0' * (x_{16} - x_i)}{x_{16} - x_0} \quad (1)$$

$$y_i' = \frac{y_{33}' * (y_i - y_{27}) + y_{27}' * (y_{33} - y_i)}{y_{33} - y_{27}} \quad (2)$$

5. We move all FLs for eyes and eyebrows 50 pixels downward and all FLs for the mouth 70 pixels upward
6. We connect FLs for each eyebrow, each eye contour, and mouth contour with 7 pixels thick lines

All numeric values were determined experimentally. Fig. 2 shows a facial image, the result of FLs detection and the corresponding ES. Detected FLs were dilated for better viewing. We chose to keep the resolution for ES at 250x250 pixels.

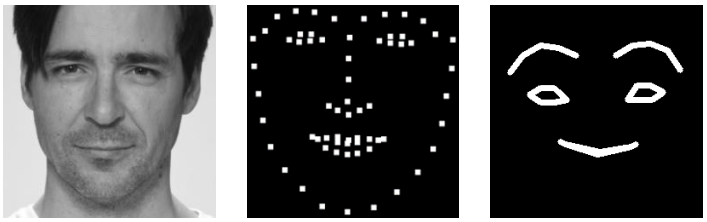

Fig. 2. Original facial image (left), facial landmarks image (center), emotion sketch (right)

A recurring problem of FER is the limited availability of free, extensive datasets, as CNNs, often used in image classification problems, require large amounts of training samples for good results. Additional challenges come from the different appearances of persons, from age and ethnicity to the amount of facial hair. Networks are usually trained and tested on datasets depicting subjects with similar features (e.g., JaffeDBase, often used by researchers, contains only images of Japanese women) and each network offers varying usage of resources and precision of results. Using ES, we reduce the differences in facial appearance for persons of different ages and ethnicities, allowing us to combine multiple FER datasets into one. Still, the number of images was insufficient for determining a large number of emotions, so we settled for 4 classes: positive (i.e., happiness), negative (i.e., anger, disgust and sadness), awe (i.e., fear and surprise), plus neutral (i.e., the natural, relaxed facial appearance).

We constructed our dataset using images from the FEI Face Database [18], the JaffeDBase [19], the California Facial Expressions (CAFE) dataset [20], all available online, plus images we captured during our previous work. About half of the images depict Caucasian persons, the other half depicting East-Asian persons. Some images needed manual cropping and alignment to be suitable for our method. All images are captured from a frontal perspective and depict adults, with ages varying between 20 and 80 years old, about 75% being females. These images were processed to obtain corresponding ES images. We want to analyze how efficient the usage of emotion sketches is for FER using neural networks. As the combined dataset we constructed is still relatively small, with a little over 600 samples, we thought of a way to increase the accuracy of the networks using transfer learning. Considering the simple appearance of our ES images, as disconnected graphs, we considered the problem of classifying to be similar to that of Optical Character Recognition (OCR). So, we chose to refine the weights of the networks that were first trained on large datasets, such as MNIST or ImageNet [21][22]. Fig. 3 shows an example of ES for each of the 4 classes we want to distinguish.

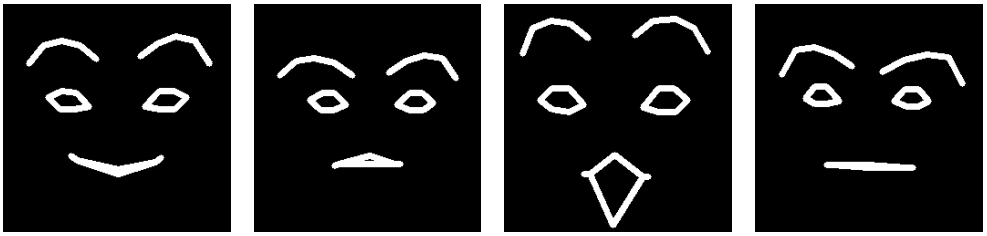

Fig. 3. Emotion sketches – positive, negative, awe and neutral (from left to right)

To test our approach, we chose three neural networks, VGG11, ResNet50 and CapsNet, that researchers previously successfully proposed for FER. VGG11 is a type of CNN, proposed in 2015, that obtained good results in classification

problems, including on the MNIST dataset of handwritten digits. It contains 11 weigh layers and has a reported accuracy of 70.4% (i.e., a top-1 error rate of 29.6%) in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [23]. It is a simple, feed-forward network that we have implemented in Python using Keras and Tensorflow, starting from online article [24]. Weights for transfer learning were obtained by training the CNN on the MNIST dataset. ResNet50, also published in 2015, is a deep neural network which uses residual functions to address the problem of vanishing gradients. It contains 50 layers, with 3-layered bottleneck blocks, and it obtained a top-5 error rate of 3.57% in ILSVRC [14]. We used the model available in Keras for Pyton, with weights already initialized by training on ImageNet dataset, so for transfer learning we only changed the input shape and the number of outputs [25]. CapsNet is a new type of network proposed in 2018, that uses capsules, which implement groups of neurons that encode spatial information and the probability of an object being present in the image. The main advantage is the understanding of objects as sums of parts: while CNNs classify an object based on the existence of parts (e.g., a face must have two eyes and a mouth), CapsNet takes notice of the relations between the parts (e.g., the eyes must be colinear and the mouth must form an isosceles triangle with them) [15]. For this network, we adapted an implementation available online [26], using Python, Keras and Tensorflow. We used a 9x9 kernel for the convolutional layers with 16x16 strides for down-sampling and increasing computation speed, while setting a number of 128 filters for the first convolutional layer. We use 3 iterations for the routing algorithm when training on MNIST to obtain the weights for transfer learning, and 5 iterations when training with ES images. No GPU-acceleration was used during development. All training, validation and querying of the networks was done on an Intel Core i7 processor. Code for ES computation was written in Java, using the OpenCV library.

3. Experimental Results and Discussions

The ES dataset we constructed contains 668 images. We use 556 images for training, 90 for validation, and 22 for querying the networks at a later time. We pretrained VGG11 and CapsNet using the MNIST dataset, while for ResNet50, we downloaded weights obtained for the ImageNet dataset, which were available online. Table I shows processing times for various stages in ES computation, for 7 images. We consider face detection and extraction from source images as preliminary steps; therefore, they are not specified in the table. Resulting images were saved as .png to avoid compression which could lead to the appearance of artifacts. All times are measured in seconds. Approximately 3 seconds is necessary to obtain an ES for a single face, mainly due to the long time needed by the DLIB detector to extract FLs. Unfortunately, this makes them

impractical to use in real-time applications without employing hardware acceleration or finding means to optimize the process.

Table I

Emotion sketches computation times

Image	FLs detection	FLs processing	Sketch drawing	Total
1	2.0871	0.8791	0.3460	3.3122
2	2.1022	0.8645	0.2398	3.2065
3	1.9802	0.7842	0.2895	3.0539
4	1.9299	0.9011	0.3201	3.1511
5	1.8932	0.8940	0.2930	3.0802
6	1.9345	0.8862	0.3287	3.1494
7	1.8763	0.7953	0.2514	2.9230

Fig. 4 shows the original face images and the resulting ES images for the 7 images in Table I. The images were chosen to depict 7 emotions, 6 usually researched in FER, namely anger, disgust, happiness, fear, sadness and surprise, with a neutral expression added.

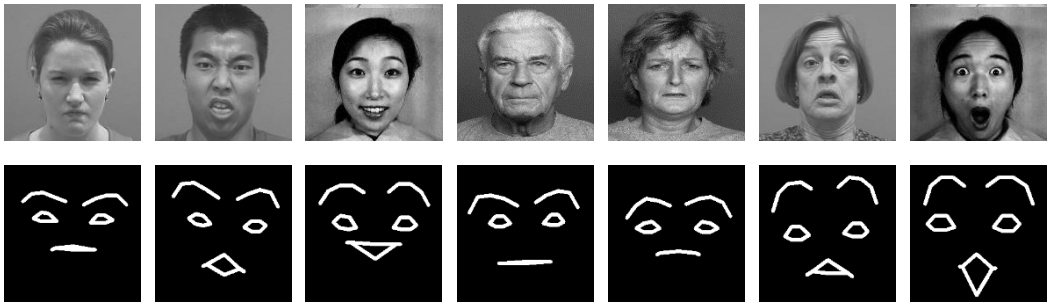

Fig. 4. Images depicting the expressions, from left to right:
'Angry', 'Disgusted', 'Happy', 'Neutral', 'Sad', 'Scared' and 'Surprised'
(top row – source image; bottom row – emotion sketches)

Ethnicity, gender and age of subjects greatly varies. These ES images correctly depict each expression, making it easy for a human observer to distinguish them. As we mentioned earlier, experimentation has shown us our dataset is too small to allow detection of all 7 expressions, so we grouped them into 4 new expression classes which we attempted to classify using the neural networks: 'Negative' (i.e., anger, disgust and sadness), 'Positive' (i.e., happiness), 'Awe' (i.e., fear and surprise), and 'Neutral'. For the first set of experiments, we trained the networks using only the ES dataset, while for the second set we performed transfer learning using the MNIST and ImageNet datasets. Table II notes the number of epochs and batch sizes, plus the loss, accuracy (Acc) and training time for each configuration, except for training done on the MNIST dataset. The epochs and batch sizes were determined following multiple trial-and-error attempts, considering a balance between training time and obtained

accuracy. The loss, accuracy and training time are given as shown by Keras at the end of the training process.

Table II

Neural network training configurations and results

Network	Training case	Epochs	Batch size	Loss	Acc (%)	Training time (s)
VGG11	ES and randomly initialized weights	20	64	0.0668	47.66	634
VGG11	MNIST and randomly initialized weights	24	256	-	-	-
VGG11	ES and transfer learning with weights initialized by training on MNIST	32	128	3.0568	54.44	1002
ResNet50	ES and randomly initialized weights	100	16	1.7417	55.56	1304
ResNet50	ES and transfer learning with weights initialized by training on ImageNet	20	32	0.5998	78.57	3960
CapsNet	ES randomly initialized weights	8	24	0.2575	61.22	2726
CapsNet	MNIST and randomly initialized weights	16	300	-	-	-
CapsNet	ES and transfer learning with weights initialized by training on MNIST	100	32	0.5394	64.15	12553

As expected, the more complex the network, the longer the time needed for training. It is interesting that CapsNet gave the best accuracy when training on a small dataset, consisting only of the ES images, but the accuracy obtained after transfer learning was lower than that of ResNet50. After each network was trained and validated, we predicted 22 query images, all depicting persons different from those in the training and validation sets. The output layers of VGG and ResNet50 use Softmax activation function, while the output layer of CapsNet uses Sigmoid function. We proportionally distributed the Sigmoid probabilities in the [0, 1] interval for easier comparison with Softmax probabilities. Table III shows a subset of prediction results on 4 images, for VGG11, ResNet50 and CapsNet, respectively. The subset was chosen to represent all 4 expression classes. Images are identified by an index assigned when constructing the dataset.

Table III

Prediction results

Image index	Network	Negative emotions	Positive emotions	Neutral emotions	Awe emotions	Predicted emotion
3	VGG11	0.6573	0.0146	0.0548	0.2733	Negative
	ResNet50	0.9981	0.0003	0.0000	0.0016	
	CapsNet	0.5362	0.0241	0.0538	0.3859	
9	VGG11	0.0987	0.0117	0.4349	0.4547	Awe

	ResNet50	0.0105	0.0006	0.0051	0.9838	
	CapsNet	0.3048	0.0616	0.2918	0.3418	
11	VGG11	0.0028	0.9962	0.0007	0.0002	Positive
	ResNet50	0.0000	1.0000	0.0000	0.0000	
	CapsNet	0.1010	0.4834	0.3071	0.1084	
16	VGG11	0.4022	0.0119	0.5097	0.0762	Neutral
	ResNet50	0.3710	0.0945	0.4938	0.0407	
	CapsNet	0.4079	0.0712	0.4380	0.0830	

Table IV compares the predicted emotions given by each network for the 22 query images. The ground truth was established by human observation of the source image. The confidence column shows the greatest probability given for a correct prediction by any network. Incorrect results are grayed-out, while the correct results with the greatest confidence score are given in bold-italic.

Table IV

Prediction validation

Image	VGG11	ResNet50	CapsNet	Ground truth	Confidence (%)
1	Negative	***Negative***	Neutral	Negative	99.49
2	Positive	Positive	Positive	Negative	-
3	Negative	***Negative***	Negative	Negative	99.81
4	Negative	***Negative***	Negative	Negative	99.97
5	Neutral	Positive	***Negative***	Negative	52.64
6	Negative	***Negative***	Negative	Negative	99.93
7	Negative	Negative	Negative	Awe	-
8	Neutral	***Awe***	Neutral	Awe	99.96
9	Awe	***Awe***	Awe	Awe	98.38
10	Neutral	***Awe***	Awe	Awe	99.95
11	Positive	***Positive***	Positive	Positive	100
12	Positive	***Positive***	Positive	Positive	100
13	Positive	***Positive***	Positive	Positive	100
14	Neutral	Negative	***Neutral***	Neutral	41.38
15	Negative	Negative	***Neutral***	Neutral	46.82
16	***Neutral***	Neutral	Neutral	Neutral	50.97
17	***Negative***	Negative	Negative	Negative	99.04
18	***Negative***	Negative	Negative	Negative	96.77
19	Negative	***Negative***	Negative	Negative	99.92
20	Awe	***Awe***	Awe	Awe	99.99
21	Awe	***Awe***	Awe	Awe	86.73
22	Awe	***Awe***	Awe	Awe	99.99

CapsNet correctly classified 18 out of 22 query ES images and is the only network to correctly classify all the ‘Neutral’ samples. Table V show the confusion matrices for the three networks over the query dataset [27]. P denotes positive emotions, Nl – neutral, Nv – negative and A denotes awe emotions. True positive results (TP), true negative results (TN), false positive results (FP), false

negative results (FN), accuracy, precision, recall and F1-score is given for each emotion, then macro values are given for each network.

Table V

Confusion matrices

Model	VGG11				ResNet50				CapsNet			
Predicted Emotion	P	Nl	Nv	A	P	Nl	Nv	A	P	Nl	Nv	A
TP	3	2	7	4	3	1	7	6	3	3	7	5
TN	18	16	11	15	17	19	10	15	18	17	12	15
FP	1	3	2	0	2	0	3	0	1	2	1	0
FN	0	1	2	3	0	2	2	1	0	0	2	2
Accuracy	0.95	0.81	0.81	0.86	0.90	0.90	0.77	0.95	0.95	0.90	0.86	0.90
Precision	0.75	0.40	0.77	1.00	0.60	1.00	0.70	1.00	0.75	0.60	0.87	1.00
Recall	1.00	0.66	0.77	0.57	1.00	0.33	0.77	0.85	1.00	1.00	0.77	0.71
F1-Score	0.85	0.50	0.77	0.72	0.75	0.50	0.73	0.92	0.85	0.75	0.82	0.83
Model Accuracy	0.8636				0.8864				0.9091			
Model Precision	0.7319				0.8250				0.8063			
Model Recall	0.7540				0.7421				0.8730			
Macro F1-Score	0.7155				0.7275				0.8160			

All ‘Positive’ samples were correctly classified by all three networks, as expected considering how easy it is to distinguish them visually (i.e., smiling supposes the corners of the mouth to be upturned). Two images were not correctly classified by any network. Viewing them revealed inaccuracies in the ES images due to the DLIB detector mispositioning FLs. This is visible in Fig. 5, with a ‘Negative’ emotion being misclassified as ‘Positive’.

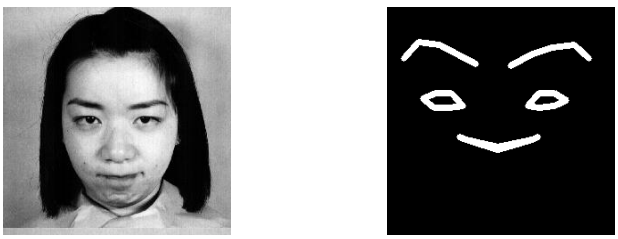

Fig. 5. Mislabeling due to emotion sketch inaccuracy for the image with index 2

Prediction for each query image is very fast, at around 100 milliseconds regardless of the network, so reducing the time in which we can obtain an ES for a facial image can make our method suitable for real-time applications. Accuracy may be improved by altering the ES to utilize more FLs as information we discarded as redundant may contain subtle aspects of visual representation of

emotions (e.g., differences between ‘Anger’ and ‘Sadness’). ES images proved efficient as a lightweight, visual feature descriptor in our quest to determine four emotion classes regardless of the age, gender or ethnicity of the subjects.

4. Conclusions and Future Work

In this paper, we have described a novel approach on automated FER, defining and using emotion sketches for simplified expression representations starting from facial landmarks. We have tested the efficiency of emotion sketches by using them to train three neural networks of different types in order to classify facial expressions in ‘Positive’, ‘Negative’, ‘Awe’ and ‘Neutral’. The prediction results were encouraging, with over 70% accuracy given by each network on a query dataset. Emotion sketches proved efficient as the original images from which the dataset was created depicted persons of various ages, genders and ethnicities, yet the physical characteristics had no evident influence in the results of FER.

Unfortunately, the system can’t be used efficiently for real time videos, as the process of obtaining emotion sketches is time consuming. However, this may be addressed by algorithm optimization and using hardware acceleration. An important limitation of the facial landmark detector we used is its imprecise analysis of images with poor contrast, so resulting emotion sketches may not correctly represent the intended expression.

In the future, we want to optimize the processing to achieve real time facial expression recognition. Also, considering we only used part of the facial landmarks offered by the detector, and that those facial landmarks are not always correctly placed, we will study the opportunity to create our own facial landmarks extraction system, which can both enhance result precision and decrees resource usage.

R E F E R E N C E S

1. *A. De Souza, A. Lopes, E. Aguiar, T. Oliveira-Santos* Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order, in Pattern recognition, vol. 61, 2016.
2. *B. C. Ko* A Brief Review of Facial Emotion Recognition Based on Visual Information, in Sensors, vol. 18, 2018.
3. *D. Mehta, M. F. H. Siddiqui, A. Y. Javaid* Recognition of Emotion Intensities Using Machine Learning Algorithms: A Comparative Study, in Sensors, vol. 19, 2019.
4. *D. Y. Liliana* Emotion recognition from facial expression using deep convolutional neural network, in Journal of Physics: Conference Series, vol. 1193, 2019.
5. *T. Kundu, C. Saravanan* Advancements and recent trends in emotion recognition using facial image analysis and machine learning models, ICEECCOT, Mysuru, India, December 2017.
6. *S.-J. Wang, H.-L. Chen, W.-J. Yan, Y.-H. Chen, X. Fu* Face Recognition and Micro-expression Recognition Based on Discriminant Tensor Subspace Analysis Plus Extreme Learning Machine, in Neural Process Letters, 2013.

7. *S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, C.-G. Zhou* Micro-Expression Recognition Using Robust Principal Component Analysis and Local Spatiotemporal Directional Features, ECCV, Zurich, Switzerland, September 2014.
8. *W. Merghani, A. K. Davidson, M. H. Yap* A Review on Facial Micro-Expressions Analysis: Datasets, Features and Metrics, in CoRR, 2018.
9. *M. Kopaczka, J. Schock, D. Merhof* Super-realtime facial landmark detection and shape fitting by deep regression of shape model parameters, in CoRR, 2019.
10. *S. Mahpod, R. Das, E. Majorana, Y. Keller, P. Campisi* Facial Landmark Point Localization using Coarse-to-fine Deep Recurrent Neural Network, in CoRR, 2018.
11. *X. Chen, L. Huang, Y. Chen* Facial Landmark Detection based on Cascade Neural Network, in Journal of Physics: Conference Series, vol. 1193, 2019.
12. *Y. Wu, Q. Li* Facial Landmark Detection: a Literature Survey, in CoRR, 2018.
13. *D. Liu, H. Zhang, P. Zhou* Video-based Facial Expression Recognition using Graph Convolutional Networks, ICPR2020, Milan, Italy, January 2021.
14. *K. He, X. Zhang, S. Ren, J. Sun* Deep residual learning for image recognition, in Microsoft Research, 2015.
15. *S. Sabour, N. Frosst, G. Hinton* Dynamic routing between capsules, in CoRR, 2017.
16. *A. Costache, D. Popescu, L. Ichim* Facial Expression Detection by Combining Deep Learning Neural Networks, ATEE, Bucharest, Romania, 25 March 2021.
17. Facial landmarks with dlib, OpenCV, and Python. Available online: pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python (accessed on 23 March 2021)
18. FEI Face Database. Available online: https://fei.edu.br/~cet/facedatabase.html (accessed on 21 March 2021)
19. The Japanese Female Facial Expression (JAFFE) Database. Available online: zenodo.org/record/3451524#.Xmt3LnIzaUk (accessed on 21 April 2021)
20. CAFE Dataset, M. Daley, G. Cottrell and J. Reilly, 2001, Available online: http://www.cs.ucsd.edu/users/gary/CAFE/ (accessed on 16 February 2021)
21. The MNIST database of handwritten digits. Available online: yann.lecun.com/exdb/mnist/ (accessed on 15 December 2020)
22. ImageNet. Available online: https://www.image-net.org/index.php (accessed 4 April 2021)
23. *K. Simonyan, A. Zisserman* Very deep convolutional networks for large-scale image recognition, ICLR, San Diego, California, U.S.A, 7 May 2015
24. Build Your Own Convolution Neural Network in 5 mins. Available online: towardsdatascience.com/build-your-own-convolution-neural-network-in-5-mins-4217c2cf964f (accessed on 11 April 2021)
25. Keras Optimizers. Available online: keras.io/optimizers (accessed on 7 May 2021)
26. CapsNet-Keras. Available online: https://github.com/XifengGuo/CapsNet-Keras (accessed on 16 May 2021)
27. Confusion Matrix for Your Multi-Class Machine Learning Model. Available online: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826 (accessed on 1 August 2021)