ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20200423145443
Original scientific paper

Facial Landmark Based Region of Interest Localization for Deep Facial Expression Recognition

Omer Faruk SOYLEMEZ*, Burhan ERGEN

Abstract: Automated facial expression recognition has gained much attention in the last years due to growing application areas such as computer animated agents, sociable robots and human computer interaction. The realization of a reliable facial expression recognition system through machine learning is still a challenging task particularly on databases with large number of images. Convolutional Neural Network (CNN) architectures have been proposed to deal with large numbers of training data for better accuracy. For CNNs, a task related best achieving architectural structure does not exist. In addition, the representation of the input image is equivalently important as the architectural structure and the training data. Therefore, this study focuses on the performances of various CNN architectures trained by different region of interests of the same input data. Experiments are performed on three distinct CNN architectures with three different crops of the same dataset. Results show that by appropriately localizing the facial region and selecting the correct CNN architecture it is possible to boost the recognition rate from 84% to 98% while decreasing the training time for proposed CNN architectures.

Keywords: convolutional neural networks; deep learning; facial expression recognition

1 INTRODUCTION

Facial expressions are the most prominent form of non-verbal communication among mankind. They exhibit changes in social situations and help individuals to infer the feelings of others. Many studies put forth that the facial expressions portray the same feelings across different cultures [1, 2]. They are universal rather than being unique and this makes them a principal element for emotional expressiveness.

Automatic facial expression recognition (FER) has been an active field of research for the last two decades. It boosts many use cases such as human computer interaction, driver fatigue detection [3], supportive healthcare [4], social marketing [5], interactive games [6], sociable robotics [7] and face expression synthesis [8]. It is challenging due to the nature of the faces as they may be captured in unfavorable conditions. Numerous studies have been conducted in order to implement better solutions and overcome domain specific problems, such as face detection [9], face alignment [10] and facial feature point extraction [11]. Studies on facial expression recognition utilize both supervised and unsupervised machine learning algorithms in order to classify facial images according to their related classes.

In general, the feature extraction stage is the most important step in a pattern recognition system. The same applies to FER systems as their performances rely heavily on selected features. Features could either be hand selected or automatically generated. Acquiring hand selected features on large image data is time consuming and prone to errors. To avoid this, automatic feature extraction using deep learning frameworks has become a widely followed practice to avoid image specific feature extraction, especially when working with datasets that contains excessively large numbers of images.

Deep learning and its applications are on the rise since Alex Krizhevsky won the ILSVRC2012 competition with his CNN approach named as AlexNET [12]. Since then, many different neural network architectures and designs have been proposed and most of them have been used in different recognition and classification tasks such as computer vision [13], speech recognition [14], natural language processing [15] and audio emotion recognition [16].

Although automatic feature extraction does not require human intervention, there are some preliminary steps that must be taken. Preprocessing is an important aspect, as the robustness of the feature extraction is heavily dependent on it. Different kinds of preprocessing algorithms exist, each helping to improve feature extraction in other ways such as mean subtraction, normalization and PCA. Mean subtraction is one of the most common forms of preprocessing. It is achieved by subtracting the image mean from every individual feature in the data. It helps to center the feature distribution which keeps the initial activations in a reasonable range. Normalization, PCA and whitening are some other forms of preprocessing, differing in implementations, but contributing to make data more centric as mean subtraction.

Other than preprocessing, selection of a Region of Interest (ROI) is also a critical aspect for CNNs. CNNs take the whole image as an input and extract features exhaustively. Therefore, a sample that includes non-correlated data with the actual feature space would yield feature vectors that are irrelevant to this problem domain. Classification with these feature vectors would be inefficient since they may lack the true feature representation of the problem domain. It can be said that, regardless of the training data, ROI for deep learning applications should be selected carefully, neither too broad to contain unrelated information nor too restricted that may conceal the necessary representation. Defining an accurate ROI for the input images will yield a better performance for CNN based object recognition tasks.

In this study, we aimed to investigate the impact of ROI determination on the classification performance of CNNs designed for FER. We propose that effective ROI selection can increase the performance of CNNs for facial expression recognition task. CK+ database, one of the most utilized databases among researchers for FER, have been used to this purpose [17]. Active Appearance Model (AAM) has been used to extract facial landmarks points and some predefined self-created metrics are used to crop

the ROI from the full facial image [18]. By scaling these metrics, several different ROIs are achieved, which later are used as input into CNNs. Three distinct CNN structures are used, each implemented in a different architecture - Proposed sequential model, a GoogLeNet architecture "Inceptionv3" model [19] and a Residual Network "ResNET50" model [20].

By training each of these models with the mentioned ROIs, we presented the impact of ROI selection on the performance of CNNs. In brief, the main contributions of this work are:

- A study to evaluate the classification performances of different CNN architectures across different facial ROIs,
- The effect of k-fold training and presentation order in CNNs.

The rest of this paper is organized as follows: Section 2 describes the historical development of the CNNs and how it is adopted by FER tasks with some of the noteworthy works mentioned. Section 3 describes our proposed approach. In Section 4, the results of carried experiments are shown and discussed. Finally, in Section 5, we conclude our study along with some guides for the future works.

2 RELATED WORK

Researches on FER are active for nearly two decades. There have been many studies conducted on this topic and nearly all feature extraction and classification methods that are valid and applicable for this domain have taken their part. We do not intend to give an in-depth explanation about these studies, so we deem suitable to present surveys for further information in past work on this topic [21].

Most of the former algorithms and methods for computer science did not get their deserved attention at the first time they emerged. For instance, first paper on Support Vector Type algorithms was mentioned at 1963 though they did not attract much attention at that time [22]. It became popular and achieved high applicability in 1990s. Since then, many different variations of SVMs are proposed [23]. The same goes for the CNNs, with first application of convolution introduced by Fukushima [24]. It gets a little more recognized by Lecun et al. [25] in the way used today. Due to the unavailability of the computing resources to process inputs with higher resolutions and training deeper networks at that time, CNNs were tagged as impractical. With the pioneering work of Alex Krizhevsky [12], it is revealed that exploitation of task parallelization of today's GPGPUs made training of deeper CNNs feasible unlike in the past. Nowadays almost every problem domain that is suitable for automatic feature extraction has its own type of CNN implementation.

As mentioned above, many studies on human facial expressions recognition also made use of CNNs. Dating back to late 2013's, we would like to give some brief information about the ones that we consider as noteworthy. Besides these, some of the works with their contributions to the field and their recognition rates are given in Table 1.

Lopes et al. [26] offer an approach to overcome one of the difficulties that hinder the performance of CNNs - Overfitting. Overfitting mostly occurs when training a complex model without sufficient amount of training samples. Accuracy decreases when classifying unseen samples, since model memorizes the training samples instead of creating a generalizable one. Overfitting could be overcome by means of simpler networks, dropout regularization, L1 and L2 regularization, batch normalization, increased sample size, data augmentation or early stopping when validation error increases for several epochs. The authors generated synthetic face samples to increase the number of training samples. They achieved this by applying artificial rotations, translations and skewing onto original images. It is stated that for every single image, 70 artificial samples are constructed. They also applied pre-processing before feeding images into training such as cropping, rotation and intensity normalization. The offer of CNN consists of 5 layers achieving 98.92% and 98.80% for 6-expression and 7-expression recognition respectively.

Liu et al. [27] proposed an action unit driven deep network (AUDN), inspired by the psychological theory that expressions can be decomposed into multiple facial Action Units (AUs). They have built a convolutional layer and a max-pooling layer to learn the Micro-Action-Pattern (MAP) representation, which can explicitly depict local appearance variations caused by facial expressions. Then a feature grouping is applied to simulate larger receptive fields by combining correlated MAPs adaptively, aiming to generate more abstract mid-level semantics. Finally, a multi-layer learning process is employed in each receptive field respectively to construct group-wise sub-networks for higher-level representations. The pipeline they provide achieves 93.70% accuracy on CK+ database by using a cross-validation approach and without subject overlap between training and test sets.

Liu et al. [28] proposed a deep learning framework called Boosted Deep Belief Network (BDBN) for FER which works as follows: an initial feature representation is learned through a bottom-up unsupervised feature learning (BU-UFL) process for image patches. Then, a subset of weak learners is selected by boosting and is fine-tuned jointly in a boosted top-down supervised feature strengthen (BTD-SFS) process. Then these two processes run alternatively until convergence. As the learning continues, the discriminative ability of the strong classifiers and the weak learners increase thus granting better classification accuracy in the later stages. Their model achieves 96.7% for 6-class expression recognition on CK+ database.

Following studies are also performed on CK+ database. Fan and Tjahjadi [29] proposed Spatial-temporal framework based on histogram of gradients and optical flow achieving 83.70% for 6-class expression recognition. Gu et al. [30] developed a radial encoding strategy for efficiently down sampling Gabor filter outputs and a new classifier combination method by extracting information from local classifiers. This approach achieved 91.51% for 6-class expression recognition. Zhong et al. [31] proposed a two-stage multi-task sparse learning (MTSL) framework to efficiently locate the discriminative patches that discloses the expressions. MTSL framework achieved 93.30% accuracy for 6-class expression classification task. Liu et al. [32] proposed a manifold modeling of videos based on a proposed mid-level representation, i.e. expressionlets. Their approach achieved 94.19% on 6-class expression recognition task.

In this study, a facial landmark-based ROI localization for deep FER task is presented. Aside from other methods that only utilize a single crop of the face, three different ROIs for every single facial image have been evaluated. Extensive training benchmarks have been conducted on three different CNN architectures with three ROIs in order to determine their efficiencies. 98.04% accuracy for FER on CK+ dataset is the current state of the art without utilizing extensive data augmentation techniques. This study presents important evidence on the impact of ROI localization for FER tasks with CNNs.

3 PROPOSED WORK

This study consists of three stages and within each a different problem is addressed. At the initial data set preparation stage, images are acquired from the CK+ database, facial landmark points are detected with an AAM and region of interests of distinct sizes are cropped with the aid of mentioned landmarks. At the network preparation stage, CNNs with different architectures are built. The final training and testing stage cover training and testing of these networks with the batches of images created at the first stage. After each training epoch, CNNs have been validated with validation sets.

3.1 Dataset Preparation

A visualization for the dataset preparation steps is presented in Fig. 1. Initially, Viola-Jones [33] facial cascade classifiers have been used in order to detect the faces in the given input images. After successfully determining the facial area, AAM is used to detect landmark points on images. iBUG68 [34] annotations were used to conduct AAM and the facial landmark points are acquired as shown in Fig. 1c. Since landmarks lack eye center information, the centroid of quadruple of points that encompass each eye is appointed as respective eye center. The inter-pupillary distance, the distance between two eye centers, is calculated by drawing a line between two eye centers. Unit distance α (semi inter-pupillary distance) for each image is acquired by dividing inter-pupillary distance with two. Unit distance is represented in Fig. 1d. To focus on the same ROI in each image, an axis point had to be defined. To this end, the midpoint of the interpupillary line that connects two eye centres is appointed as the axis. As a sequel to axis definition, two ROIs are cropped from each image. The original input image and the cropped regions are named as ROI scale 0, 1 and 2 respectively. ROI scale 0 covers the original input image. It includes irrelevant data for classification such as dress, timestamps and background. It is 640 × 490 pixels in size. ROI scale 1 encompasses the whole facial structure including forehead, eyes, nose, ears, mouth and chin. It spans approximately 260 × 350 pixels. ROI scale 2 encompasses the most prominent elements of the face that portrays the expressions; mouth, eyes and nose. It spans approximately 160 × 200 pixels.

The cropping process is performed as shown in Fig. 1e. Semi interpupillary distance α is calculated for each image and then used to define the boundaries of the ROIs for the pertained image. Coefficients of the α to enclose the mentioned region of interests are also presented in Fig. 1e.

3.2 Network Preparation

One of the cases that we would like to examine in this study are the performances of CNNs with different architectures on the same task. A simple sequential model, a GoogLeNet architecture "Inceptionv3" model and a Residual Network "ResNET50" model are proposed in this context.

The proposed simple sequential model is a lightweight network that is composed of 9 layers. The architecture of the proposed model could be seen in Tab. 2. It includes three pairs of "conv + maxpool" layers followed by 3 consecutive dense layers with the last one having a softmax classifier for 6 classes.

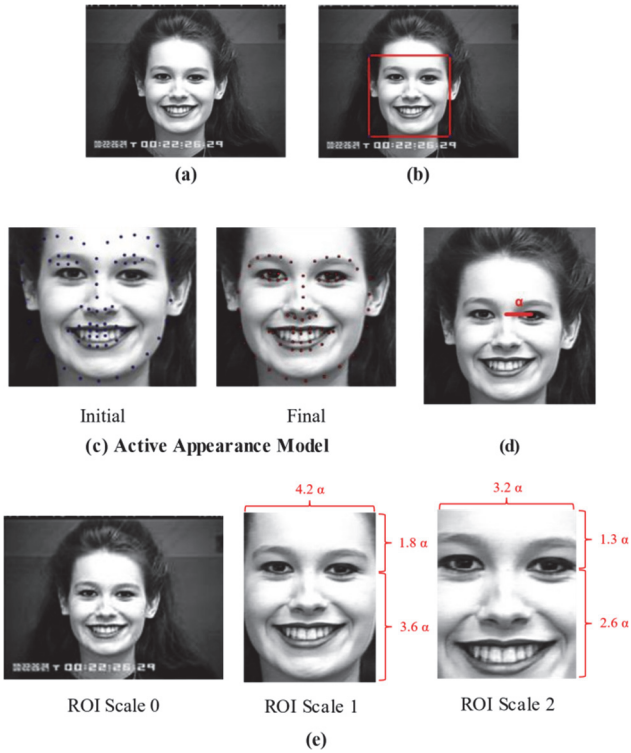

Figure 1 An overview of the dataset preparation process. (a) Input image; (b) Face Detection; (c) AAM; (d) Semi Interpupillary Distance; (e) Selected ROIs

Table 1 Network architecture of the proposed model

Layer	Input	Filter/Stride	Output Size
conv0	I (200 × 200 × 1)	(3 × 3)/1	200 × 200 × 32
maxpool0	conv0	(2 × 2)/2	100 × 100 × 32
conv1	maxpool0	(3 × 3)/1	100 × 100 × 64
maxpool1	conv1	(2 × 2/2	50 × 50 × 64
conv2	maxpool1	(3 × 3)/1	50 × 50 × 128
maxpool2	conv2	(2 × 2)/2	25 × 25 × 128
[flatten, fc0]	maxpool2	-	512
fc1	fc0	-	512
softmax	fc1	-	6

The GoogleNet architecture (also known as Inceptionv3) proposed by Szegedy et al. [19] is a deep learning architecture that implements tiny CNN modules inside a larger network. It utilizes a set of smaller convolutions instead of a bigger convolution hence increasing speed by maintaining a lower number of parameters. The top 1000 class softmax layer on vanilla Inceptionv3 is replaced by a 6 class softmax layer.

The final CNN we make use of is a ResNet50 model [20]. ResNet proposes identity shortcut connections to overcome the vanishing gradient problem. Resnet50 has

been found as a suitable candidate, since it brings a different resolution to a major problem. The top layer of Resnet50 has been replaced with a 6 class softmax layer as in Inceptionv3.

3.3 Dataset Preparation

The final part of this work includes training and validation of created networks with ROIs acquired at the first stage. As mentioned previously, 3 subsets of the CK+ database have been formed so that each encompasses a different ROI. By conducting 3 training sessions on each of the 3 subsets, a total of 9 training sessions were held in our study.

4 EXPERIMENTS AND DISCUSSION

The experiments were carried on CK+ database which is composed of 123 subjects. In order to create an accurate dataset for the experiment, the initial frame for each sequence is appointed as neutral image and the last 3 frames of the same sequence are appointed as the tagged expression. The following expressions are collected from subjects: Angry, Disgust, Fear, Happy, Sadness, Surprise and Neutral. Fig. 2 demonstrates a subject and his/her derived images for dataset creation.

Images are normalized before the training. Rescaling is carried out by dividing each pixel value by 255. This made inputs lie in the range of [0, 1] rather than [0, 255], thus helping networks to converge faster. Feature wise center is held by setting the pixel values mean to zero across the whole dataset. By doing this, it is ensured that illumination differences amongst the images are normalized. To further normalize the data, feature wise standard normalization is held by diving each pixel value with the standard deviation of the dataset.

After normalization, images are rescaled via bilinear interpolation to meet the requirements of the Input layer. 200 × 200 pixels are determined as a fixed image size for the input layer of each network and images are either upscaled or downscaled to comply with it.

Apart from their classification performances, CNNs also differ from each other by the time they take to train. Average training time in seconds for an epoch for each network along with the corresponding scale is given in Tab. 3. In every training run, Simple Net exhibited the lowest training time amongst all 3 CNNs. This is due to the simplicity of its design. Inceptionv3 came as second while Resnet-50 displayed the worst result.

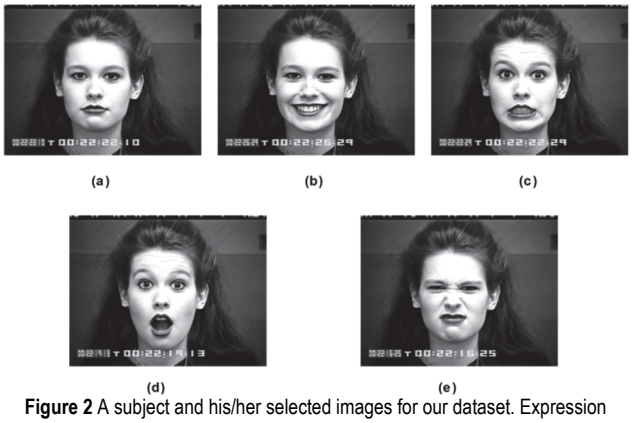

Figure 2 A subject and his/her selected images for our dataset. Expression sequences for this subject are: (a) Neutral; (b) Happy; (c) Fear; (d) Surprise; (e) Disgust

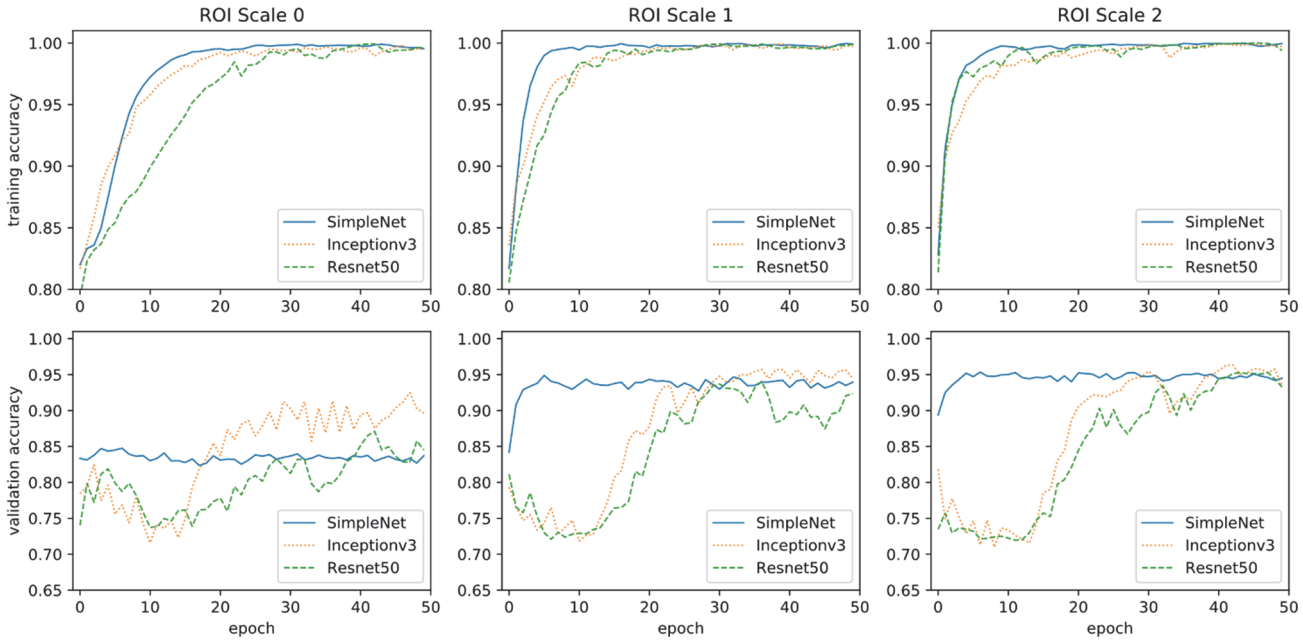

Figure 3 Training and validation accuracy averages of 8 folds for each CNN with respect to all ROI scales

Table 2 Confusion matrix for CK+ database on 6-expressions with the best achieving ROI (scale 2)

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	43	-	-	1	-	1
Disgust	1	58	-	-	-	-
Fear	-	1	23	1	-	-
Happy	-	-	-	69	-	-
Sadness	-	1	1	-	25	-
Surprise	-	-	-	-	-	83

Table 3 Average validation accuracies for each ROI scale

Net/ROI	ROI 0	ROI 1	ROI 2
Simple Net	0.8447	0.8945	0.9706
Resnet-50	0.8945	0.9609	0.9714
Inceptionv3	0.9611	0.9743	0.9804

Table 4 Average training time in seconds for an epoch for each network along with the ROI scale

	ROI Scale 0	ROI Scale 1	ROI Scale 2
Simple Net	3.4	1.3	1.1
Inceptionv3	5.9	5.2	5.2
Resnet-50	8.1	8.1	8.1

A confusion matrix for 6 class facial expression recognition on CK+ database with the best achieving ROI (which is scale 2) is given in Tab. 2. Happy and surprise expressions are easily distinguished by the network. This is because a smiling mouth with shape "V" and a surprised mouth with shape "O" are discriminative enough to separate these expressions from the others. However, other expressions are occasionally confused with others due to some underlying resemblances such as narrow eyes for disgust and anger and closed mouth for sadness and disgust.

Training and validation run of networks were performed with 8-fold cross validation. To prevent data leak, the dataset is divided into 8 equal parts without subject overlap. For each run, 7 parts were selected as training samples and one part was left out for validation. Each network has been trained for 50 epochs. Training and validation accuracy averages of 8 folds for mentioned CNNs with respect to all ROI scales are given in Fig. 3. Average validation accuracies of the best runs for the selected folds are given in Tab. 3. It can be inferred from Tab. 3 that, when the ROI is selected as narrow as possible, which in our case encloses just the eyes and the mouth duo (ROI scale 2), CNNs nearly perform identically. However, this is not the case for the ROI scales 0 and 1, as shown in Table 3, more specialized networks achieve better than their simpler counterparts. In both ROI scales 0 and 1, Inceptionv3 achieves better average validation accuracy than Resnet-50 and Simple Net, and Resnet-50 achieves better average validation accuracy then Simple Net. This emphasizes that deeper and complex networks can handle variations more robust than simpler ones. This phenomenon can be visualized by taking Figure 4 into consideration. It can be seen that validation accuracy for Simple Net has been on a plateau for the early time of the training which is a sign of stuck on local minima and can be dealt with either adding more data or setting a more complex network.

Tab. 4 shows that larger ROIs are bringing up more computational workload. As ROI grows in size, training time for both Simple Net and Inceptionv3 increases. However, this is not the case for Resnet as it takes nearly the same time to train the model for every ROI. This surge in training time is due to the broader ROIs having much more details than the smaller ones. To elaborate, ROI scale 2 has fewer variations in surface patches than ROI scale 0, since ROI scale 2 encompasses a narrower facial display while ROI scale 0 has also some background information like hair, forehead and background details. This excessive information makes training harder and hampers the discriminative ability of the networks.

A comparison with other studies conducting FER on CK+ database is given in Tab. 5. To conduct a fair evaluation amongst studies, two types of classifiers are listed, namely $C_{n\text{class}}$ and C_{bin}. In $C_{n\text{class}}$ classification, a single classifier is used for classifying inputs among n number of classes. On the other hand, in C_{bin} classification, one binary classifier for each expression is used to conduct a one-versus-all classification, resulting in n number of classifiers. Most of the time, C_{bin} performs better than $C_{n\text{class}}$ but also has a linear complexity depending on the number of classes. Our proposed method outperformed the rest with 98.04% accuracy for 6 class FER. This was achieved without using any data augmentation, ensemble of various CNN models or using the last layer output of CNN as a feature vector for other classifiers such as SVM.

Much work has been conducted on FER with CNNs, but they did not present detailed information about the effects of localization. We carried out this study, in order to provide a detailed explanation about localization of the ROI for deep learning applications. We have selected FER as a problem domain, but it does not mean necessarily the findings of this work are limited to itself. It could be used as a baseline for any classification task with CNNs that includes similar input images with extractable ROIs by the help of some other algorithms.

Table 5 Comparison with other methods

Method	Classifier	6 expressions
Fan & Tjahjadi [29]	$C_{n\text{class}}$	83.70
	C_{bin}	-
Gu et al. [30]	$C_{n\text{class}}$	91.51
	C_{bin}	-
Zhong et al. [31]	$C_{n\text{class}}$	93.30
	C_{bin}	-
AUDN[27]	$C_{n\text{class}}$	93.70
	C_{bin}	-
Liu et al. [32]	$C_{n\text{class}}$	94.19
	C_{bin}	-
BDBN [28]	$C_{n\text{class}}$	-
	C_{bin}	96.70
Lopes et al. [26]	$C_{n\text{class}}$	96.76
	C_{bin}	98.92
Proposed	$C_{n\text{class}}$	98.04
	C_{bin}	-

5 CONCLUSIONS

The effect of localization of the region of interest for deep learning applications is investigated in this study. FER task has been selected as the problem domain. CK+ facial expression database along with 3 different CNNs are used in this study. We have achieved 98.04% accuracy on CK+ database, which is the state of the art without using any image augmentation. Results of the study can be summarized as follows:

- Narrowing the ROI for facial images improves CNN performance,
- CNNs with complex architectures yield better performance when ROI includes irrelevant information,
- Using larger ROIs increases training time.

The success of our proposed algorithm is due to effective selection of ROIs. As indicated earlier, feature extraction by CNNs is an exhaustive process. This means that features are derived from the entire image without having any prior knowledge whether they are relevant or

irrelevant for classification. Based on this, every feature that is derived has a chance of affecting the classification performance of the model. In this study, we have narrowed the ROI gradually, from the point where both irrelevant and relevant features are present to the point where only the relevant features are present. Increase of invalidation accuracies underlines that this study supports our "effective ROI selection can increase the performance of CNNs for facial expression recognition task" proposal and objective given at the introduction of this study.

As for future work, we will apply our methodology to other facial expression databases. We would also like to train a network in one database and conduct validation on another one to present how successful our methodology is in cross-database experiments.

6 REFERENCES

[1] Darwin, C. (1872). *The expression of the emotions in man and animals.* John Murray. https://doi.org/10.1037/10001-000

[2] Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., Le Compte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology*, *53*(4), 712-717. https://doi.org/10.1037/0022-3514.53.4.712

[3] Ramzan, M., Khan, H. U., Awan, S. M., Ismail, A., Ilyas, M., & Mahmood, A. (2019). A Survey on State-of-the-Art Drowsiness Detection Techniques. *IEEE Access*, *7*, 61904-61919. https://doi.org/10.1109/ACCESS.2019.2914373

[4] Muhammad, G., Alsulaiman, M., Amin, S. U., Ghoneim, A., & Alhamid, M. F. (2017). A Facial-Expression Monitoring System for Improved Healthcare in Smart Cities. *IEEE Access*, *5*, 10871-10881. https://doi.org/10.1109/ACCESS.2017.2712788

[5] Barreto, A. M. (2017). Application of facial expression studies on the field of marketing. *Emotional Expression: The Brain and the Face*, *June*, 163-189.

[6] Blom, P. M., Bakkes, S., Tan, C. T., Whiteson, S., Roijers, D., Valenti, R., & Gevers, T. (2014). Towards personalised gaming via facial expression recognition. *Proceedings of the 10th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2014*, 30-36.

[7] Zhang, L., Jiang, M., Farid, D., & Hossain, M. A. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, *40*(13), 5160-5168. https://doi.org/10.1016/j.eswa.2013.03.016

[8] Zhou, Y. & Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, *2018-Janua*, 370-376. https://doi.org/10.1109/ACII.2017.8273626

[9] Srivastava, A., Mane, S., Shah, A., Shrivastava, N., & Thakare, B. (2017). A survey of face detection algorithms. *Proceedings of the International Conference on Inventive Systems and Control, ICISC 2017*, 1-4. https://doi.org/10.1109/ICISC.2017.8068607

[10] Jin, X. & Tan, X. (2017). Face alignment in-the-wild: A Survey. *Computer Vision and Image Understanding*, *162*, 1-22. https://doi.org/10.1016/j.cviu.2017.08.008

[11] Wang, N., Gao, X., Tao, D., Yang, H., & Li, X. (2018). Facial feature point detection: A comprehensive survey. *Neurocomputing*, *275*, 50-65. https://doi.org/10.1016/j.neucom.2017.05.013

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90. https://doi.org/10.1145/3065386

[13] Altenberger, F. & Lenz, C. (2018). *A Non-Technical Survey on Deep Convolutional Neural Network Architectures*.

[14] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *22*(10), 1533-1545. https://doi.org/10.1109/TASLP.2014.2339736

[15] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [Review Article]. *IEEE Computational Intelligence Magazine*, *13*(3), 55-75. https://doi.org/10.1109/MCI.2018.2840738

[16] Lim, W., Jang, D., & Lee, T. (2017). Speech emotion recognition using convolutional and Recurrent Neural Networks. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*. https://doi.org/10.1109/APSIPA.2016.7820699

[17] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, 94-101. https://doi.org/10.1109/CVPRW.2010.5543262

[18] Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active apperance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6). https://doi.org/https://doi.org/10.1109/34.927467

[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *abs/1512.0*, 770-778. https://doi.org/10.1109/CVPR.2016.90

[21] Sandbach, G., Zafeiriou, S., Pantic, M., & Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, *30*(10), 683-697. https://doi.org/10.1016/j.imavis.2012.06.005

[22] Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, *24*, 774-780.

[23] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92*, 144-152. https://doi.org/10.1145/130385.130401

[24] Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position. *Encyclopedia of Machine Learning & Data Mining*, 658-665.

[25] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2323. https://doi.org/10.1109/5.726791

[26] Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition*, *61*, 610-628. https://doi.org/10.1016/j.patcog.2016.07.026

[27] Liu, M., Li, S., Shan, S., & Chen, X. (2015). AU-inspired Deep Networks for Facial Expression Feature Learning. *Neurocomputing*, *159*(1), 126-136. https://doi.org/10.1016/j.neucom.2015.02.011

[28] Liu, P., Han, S., Meng, Z., & Tong, Y. (2014). Facial

expression recognition via a boosted deep belief network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1805-1812. https://doi.org/10.1109/CVPR.2014.233

[29] Fan, X. & Tjahjadi, T. (2015). A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition*, *48*(11), 3407-3416. https://doi.org/10.1016/j.patcog.2015.04.025

[30] Gu, W., Xiang, C., Venkatesh, Y. V., Huang, D., & Lin, H. (2012). Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, *45*(1), 80-91. https://doi.org/10.1016/j.patcog.2011.05.006

[31] Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., & Metaxas, D. N. (2012). Learning active facial patches for expression analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2562-2569. https://doi.org/10.1109/CVPR.2012.6247974

[32] Liu, M., Shan, S., Wang, R., & Chen, X. (2014). Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1749-1756. https://doi.org/10.1109/CVPR.2014.226

[33] Viola, P.,& Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I-511-I-518. https://doi.org/10.1109/CVPR.2001.990517

[34] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2015). 300 Faces in the Wild Challenge: database and results. *Image and Vision Computing*, *47*, 3-18. https://doi.org/10.1016/j.imavis.2016.01.002

Contact information:

Omer Faruk SOYLEMEZ
(Corresponding author)
Dicle University, Faculty of Engineering,
Department of Computer Engineering,
Diyarbakir, Turkey
E-mail: osoylemez@dicle.edu.tr

Burhan ERGEN
Firat University, Faculty of Engineering,
Department of Computer Engineering,
Elazig, Turkey
E-mail: bergen@firat.edu.tr