2018 14th Symposium on Neural Networks and Applications (NEUREL)
Belgrade, Serbia, November 20-21, 2018

Real-time Algorithms for Facial Emotion Recognition: A Comparison of Different Approaches

Aneta Kartali, Miloš Roglić, Marko Barjaktarović, Milica Đurić-Jovičić, and Milica M. Janković, *Member, IEEE*

***Abstract* — Emotion recognition has application in various fields such as medicine (rehabilitation, therapy, counseling, etc.), e-learning, entertainment, emotion monitoring, marketing, law. Different algorithms for emotion recognition include feature extraction and classification based on physiological signals, facial expressions, body movements. In this paper, we present a comparison of five different approaches for real-time emotion recognition of four basic emotions (happiness, sadness, anger and fear) from facial images. We have compared three deep-learning approaches based on convolutional neural networks (CNN) and two conventional approaches for classification of Histogram of Oriented Gradients (HOG) features: 1) AlexNet CNN, 2) commercial Affdex CNN solution, 3) custom made FER-CNN, 4) Support Vector Machine (SVM) of HOG features, 5) Multilayer Perceptron (MLP) artificial neural network of HOG features. The result of real-time testing of five different algorithms on the group of eight volunteers is presented.**

***Keywords* — convolutional neural network, emotion recognition, facial expression, Multilayer Perceptron, Support Vector Machine.**

I. Introduction

In recent years, there has been a growing interest for development of accurate and reliable computer algorithms for emotion recognition based on facial features acquired from camera. Facial expression is one of the most important features of human emotion recognition [1]. Nowadays, automated facial expression recognition has a large variety of applications, such as data-driven animation, neuromarketing, interactive games, sociable robotics and many other human-computer interaction systems [2]. However, the recognition of facial expressions is not an easy problem for machine learning methods, since people can vary significantly in the way they show their expressions.

Facial features, such as eyes, brows, nose, mouth and chin, can be labeled in a face image and create facial feature points [3]. These facial features can be detected in an image through the process of fitting a predefined set of facial feature points into a face image which is called Facial Feature Point Detection (FFPD) [4]. Face detection is the first and primary step in facial feature localization. This step helps removing non-facial information from the image and makes the FFPD process more efficient and accurate.

Facial expression recognition systems can work with static images [5-7] or with dynamic image sequences [8-11]. In static-based methods, a feature vector comprises information about the current input image only. Sequence based methods use temporal information of images to recognize the expression captured from one or more frames. Automated systems for facial expression recognition receive the expected input (static image or image sequence) and typically give as output one of the basic expressions (anger, sadness, happiness and fear), while some systems also recognize the neutral expression, surprise and disgust.

In the literature, different approaches were performed for solving the emotion recognition based on facial expression problem: machine learning approaches like Support Vector Machine (SVC), Linear Discriminant Analysis (LDA), AdaBoost methods applied for hand crafted features extraction [12] and neural network training where deep-learning approaches are preferred [13].

Despite efforts made in developing various methods for feature extraction for emotion recognition, existing approaches traditionally lack generalizability when applied to unseen images or those that are captured in outdoor settings [14]. Most of the existing approaches are based on engineered features such as Histogram of Oriented Gradients (HOG) [15], Local Binary Pattern Histogram (LBPH) [16], and Gabor [17] where the classifier's hyperparameters are tuned to give best recognition accuracies across a single database, or a small

This research was partly supported by the Ministry for Education, Science and Technology Development of Serbia, Belgrade, Serbia (OS175016, 17013).

Aneta Kartali is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e mail: anetakartali@gmail.com).

Miloš Roglić is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e mail: milos.roglic@gmail.com).

Marko Barjaktarović is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e-mail: mbarjaktarovic@etf.rs).

Milica Djurić-Jovičić is with the Innovation Center, University of Belgrade - School of Electrical Engineering, Serbia (e-mail: milica.djuric@etf.rs).

Milica M. Janković is with the School of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 11120 Belgrade, Serbia (e mail: piperski@etf.rs).

978-1-5386-6974-7/18/$31.00 ©2018 IEEE

collection of similar databases. Nevertheless, the results are not significant when they are applied to novel data.

Deep learning (DL) methods provide learning of the input data representations at different levels of abstraction, whereas higher representational levels provide features that are more important for differentiation and classification [LeCun, Yann et al]. Implementation of DL methods, such as convolutional neural networks (CNN), has significantly contributed to improvement of results for different image recognition problems [LeCun, Yann], including the recognition of facial expressions [18].

In this paper, we present a comparison of five different methods for real-time emotion recognition of four basic emotions (happiness, sadness, anger and fear) from static facial images. We have compared three deep-learning approaches based on convolutional neural networks (CNN) and two conventional approaches for classification of Histogram of Oriented Gradients (HOG) features: 1) AlexNet CNN, 2) commercial Affdex CNN solution, 3) custom made FER-CNN, 4) Support Vector Machine (SVM) of HOG features, 5) Multilayer Perceptron (MLP) artificial neural network (ANN) of HOG features.

II. Method

A. Conventional approaches

Emotion recognition algorithms based on conventional approaches include: 1) facial landmark detection (eyes, brows, nose, mouth and chin) and face extraction, 2) feature extraction and classification.

1) Facial landmark detection and face extraction

Facial landmark extraction was performed on monochromatic (gray-scale) images using open source *OpenFace* toolkit [19] for *Matlab* (Mathworks, USA). Facial landmark detection was performed using the generic algorithm in *OpenFace*, introduced by Yu et al. [20]. This algorithm is based on *Constrained Local Neural Field* (CLNF) and *Constrained Local Model* (CLM) models. To remove non-facial information from the image, a binary mask was created by using a convex hull that surrounds facial landmarks (Fig. 1b) and then applied to extract the face (Fig. 1c,d).

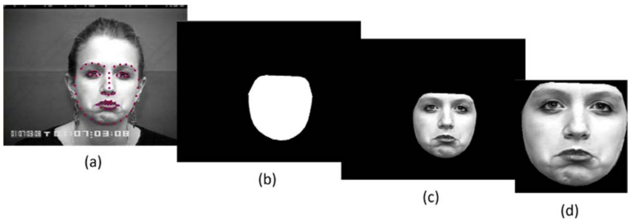

Fig. 1. Face extraction procedure: (a) facial landmark detection by *OpenFace*; (b) binary mask creation; (c) removing non-facial information by masking; (d) extracted face.

2) Feature extraction and classification

SVM and MLP ANN classification methods are applied to differ emotions based on HOG features from different face regions. Two publicly available facial expression databases have been used for training: Extended Cohn-Kanade Dataset (CK+) [21] and Karolinska Directed Emotional Faces (KDEF) [22].

For the SVM classifier, HOG features extraction was performed using facial landmarks returned by *OpenFace* toolkit, from eyes with eyebrows region and mouth region in face extracted images, as it is suggested by Pao [23].

For MLP ANN classifier, HOG features extraction was performed using facial landmarks returned by *OpenFace* toolkit as well, from regions surrounding major 36 facial landmarks as it was suggested by Lee et al. [24].

B. Deep learning approaches

1) AlexNet CNN

We have used a pre-trained AlexNet CNN [25], which has five convolutional layers and three fully connected layers. To prevent the risk of overfitting, which is unavoidable when training the network from scratch, *Transfer learning* training method was applied to fine tune the CNN's already learned parameters to solve a specific classification problem. AlexNet was fine-tuned as described by Gui et al. [26] with a few changes. The last fully connected layer was replaced and adopted to have 4 instead of 1000 outputs to classify four basic emotions: happiness, sadness, anger and fear. Two fully connected layers before the last one were replaced as well and the initial weight values of these replaced neuron layers were randomly chosen applying a zero mean Gaussian distribution with standard deviation 0.01. Their learning rate factor was set to 20. We should point out that during training, the whole network was being trained, not just the replaced layers. Therefore, it was necessary to set the global learning rate to a small value. The global learning rate was set to 10^{-4}, while the learning rate drop factor was set to 0.1. This way, fast learning of replaced, fully connected layers was enabled, while the learning rate of transferred layers was set to be very slow, as it is of interest that the parameters of these layers remain practically unchanged. The network was iteratively changed in mini-batches of 35 images, applying the gradient decent algorithm with momentum, which value was set to 0.9. Additionally, overfitting prevention was performed by data augmentation, which consisted of generating random image patches and their horizontal reflections. First the images from training datasets [21], [22] were converted to RGB images and resized to match the AlexNet input layer dimensions. We extracted five 227x227x3 random patches from 256x256x3 images and their horizontal reflections. In that way 10 images were generated from one image, so the total number of available images was enlarged 10 times. For the fine tuning of AlexNet, 70% of image dataset was used, while the remaining 30% was used for testing.

2) FER-CNN

FER-CNN is a deep learning approach proposed by a group of students [27] from TU Delft with three convolutional layers and one fully connected layer. The first layer is a convolutional layer with 44x44x64 feature map as output. After local contrast normalization (each of 64 feature maps is independently normalized) max pooling reduces feature maps to 22x22, following by the second convolutional layer with 18x18x64 feature map as output. Then another max pooling and convolutional layer are applied, following with a fully connected layer. Network outputs are obtained by *Softmax* function. This approach is a slightly modified idea presented by Gudi [28], introducing second max pooling layer between second and third convolutional layer. That modification reduces the number of parameters in overall architecture, dropping in performances by less than 1%.

The input to the FER-CNN is the extracted face found using standard Viola-Jones algorithm [29]. This algorithm for face detection is fast, accurate and it is considered as golden standard in image processing community. The algorithm is based on haar-cascade and it utilizes integral image to achieve real time performances. FERC-2013 dataset was used for training [30].

3) *Affdex CNN*

Affdex SDK consists of API that can be used to detect seven emotion metrics, 20 facial expression metrics, 13 emojis, and four appearance metrics [31]. We used SDK distributed as a Unity package for Windows. Affdex emotion metrics are trained and tested on very difficult datasets. The training set was comprised of hundreds of thousands of facial frames, from more than 3.2 million facial videos. This data is from more than 75 countries, representing real-world, spontaneous facial expressions, made under challenging conditions, such as varying lighting, different head movements, and variances in facial features due to ethnicity, age, gender, facial hair and glasses. As it can be noticed, this facial expression analysis tool is far more complex than previous algorithms proposed in this paper. However, it was used in order to get an insight in how much can hand-crafted emotion recognition methods be brought closer to commercial one.

III. Results

A. *Experiment description*

Real-time testing was performed on 8 volunteers (3 male and 5 female) of age 22.75±4.62. All subjects have signed a written consent for participation. The testing was done during daylight, in a room with additional lighting. Camera Nikon D5100 was used, while the subjects' distance from the camera was 50 cm. In order to compare results of different approaches, it would be necessary to synchronously test all proposed algorithms in real-time with the same input data. As it was too complex, we acquired a video of volunteers and used recorded videos as inputs to all algorithms in the same way that real-time input data would be used. Videos of subjects' faces were recorded at a frame rate of 24 frames per second (fps) for 160 seconds, during which they had to express four emotions: happiness, sadness, anger and fear, cyclic, five times each. The specified frame rate was chosen as it was previously determined that all algorithms work without delay in real-time at this frame rate (*offline* frame rate was greater than or equal to 24 fps).

On a screen there was an indicator showing the emotion that had to be expressed. Every emotion indicator was displayed for 5 s and a pause between them was 3 s. During the emotion indication, five frames were recorded (one per second). Based on these frames every tested algorithm gives five predictions which were then averaged to have one prediction given by each algorithm for every emotion. To have valid testing results, two observers checked facial images and only those that really matched the reference emotion indicated on the screen were considered.

B. *Confusion matrices*

Results of real-time testing are shown in the form of confusion matrices. Table 1 and 2 shows the results SVM and MLP classifiers, respectively.

Table I: Confusion matrix for SVM classification

		Predicted output			
		happiness	***sadness***	***anger***	***fear***
Desired output	***happiness***	75.86%	3.45%	20.69%	0%
	sadness	3.03%	48.48%	48.48%	0%
	anger	3.23%	0%	96.77%	0%
	fear	0%	28.57%	71.43%	0%

Table II: Confusion matrix for MLP classification

		Predicted output			
		happiness	***sadness***	***anger***	***fear***
Desired output	***happiness***	89.66%	10.34%	0%	0%
	sadness	6.06%	81.82%	12.12%	0%
	anger	6.45%	67.74%	22.58%	3.23%
	fear	35.71%	35.71%	28.57%	0%

In Table 3 confusion matrix for AlexNet CNN is shown, Table 4 shows confusion matrix for FER-CNN approach and Table 5 present results of commercial Affdex CNN solution.

Table III: Confusion matrix for AlexNet CNN

		Predicted output			
		happiness	***sadness***	***anger***	***fear***
Desired output	***happiness***	86.21%	0.00%	13.79%	0%
	sadness	0%	69.70%	30.30%	0%
	anger	0%	3.23%	96.77%	0%
	fear	21.43%	21.43%	28.57%	28.57%

Table IV: Confusion matrix for FER-CNN

		Predicted output			
		happiness	***sadness***	***anger***	***fear***
Desired output	***happiness***	62.07%	31.03%	6.90%	0%
	sadness	0%	81.82%	18.18%	0%
	anger	0%	61.29%	32.26%	6.45%
	fear	0%	50.00%	21.43%	28.57%

Table V: Confusion matrix for Affdex

		Predicted output			
		happiness	***sadness***	***anger***	***fear***
Desired output	***happiness***	96.55%	3.45%	0%	0%
	sadness	3.03%	84.85%	9.09%	3.03%
	anger	0%	29.03%	70.97%	0%
	fear	0%	7.14%	0%	92.86%

Overall accuracies of all tested algorithms are shown in Table 6. Affdex CNN performs with the highest accuracy of 85.05%, followed by AlexNet, with accuracy of 76.64%.

TABLE VI: TOTAL ACCURACIES OF ALL TESTED ALGORITHMS

Facial emotion recognition algorithm	*Total accuracy [%]*
Affdex CNN	85.05
Fine-tuned AlexNet CNN	76.64
SVM classification of HOG features	63.55
MLP classification of HOG features	56.07
FER-CNN	55.14

IV. CONCLUSION

We have presented a pilot study for real-time testing of conventional and deep learning approaches in facial emotion recognition. Preliminary results show better generalization power and better performance in real-time application of fine-tuned AlexNet CNN and Affdex CNN than SVM and MLP approaches. Commercial Affdex CNN has overall superior accuracy, but AlexNet and SVM had better "anger" recognition (96.77% vs. 70.97%). FER-CNN had the lowest overall accuracy but high accuracy for "sadness", comparable with Affdex CNN result (81.82% vs. 84.85%). In further research we will test this fact in a larger group of volunteers and for more than four emotions.

REFERENCES

[1] Y. Wu, H. Liu and H. Zha, "Modeling facial expression space for recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS 2005)*, 2005, pp. 1968-1973.

[2] A. T. Lopes, E. de Aguiar, A. F. De Souza and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610-628, 2017.

[3] J. H. Yu, K. E. Ko and K. B. Sim, "Facial point classifier using convolution neural network and cascade facial point detector," *Journal of Institute of Control, Robotics and Systems*, vol. 22, no. 3, pp. 241-246, 2016.

[4] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Computer vision and pattern recognition (CVPR),* 2012, pp. 2578-2585.

[5] C. Shan, S. Gong and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803-816, 2009.

[6] M. Liu, S. Li, S. Shan and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126-136, 2015.

[7] G. Ali, M. A. Iqbal and T. S. Choi, "Boosted NNE collections for multicultural facial expression recognition," *Pattern Recognition*, vol. 55, pp. 14-27, 2016.

[8] Y. H. Byeon and K. C. Kwak, "Facial expression recognition using 3d convolutional neural network," *International journal of advanced computer science and applications*, vol. 5, no. 12, pp. 107-112, 2014.

[9] J. J. J. Lien, T. Kanade, J. Cohn and C. Li, "Detection, tracking, and classification of action units in facial expression," *Journal of Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131-146, 2000.

[10] X. Fan and T. Tjahjadi, "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences," *Pattern Recognition*, vol. 48, no. 11, pp. 3407–3416, 2015.

[11] W. Zhang, Y. Zhang, L. Ma, J. Guan and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognition*, vol. 48, no. 10, pp. 3191– 3202, 2015.

[12] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," in *IEEE International Conference on Systems, Man and Cybernetics*, 2004, pp. 592-597.

[13] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Winter conference on Applications of Computer Vision (WACV),* 2016, pp. 1-10.

[14] C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition," *Pattern recognition and image analysis*, vol. 24, no. 1, pp. 124–132, 2014.

[15] O. Déniz, G. Bueno, J. Salido and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598-1603, 2011.

[16] B. Yang, and S. Chen, "A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image," *Neurocomputing*, vol. 120, pp. 365-379, 2013.

[17] C. Liu, and H. Wechsler, "Gabor feature-based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467-476, 2002.

[18] LeCun, Yann, Yoshua Bengio and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.

[19] T. Baltrušaitis, P. Robinson and L. P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Winter conference on Applications of Computer Vision (WACV),* 2016, pp. 1-10.

[20] X. Yu, J. Huang, S. Zhang, W. Yan and D.N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1944-1951.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW),* 2010, pp. 94-101.

[22] M. G. Calvo and D. Lundqvist, "Facial expressions of emotion (KDEF): Identification under different display-duration conditions," *Behavior research methods*, vol. 40, no. 1, pp. 109-115, 2008.

[23] J. Pao, "Emotion Detection Through Facial Feature Recognition," Stanford University, Stanford, California, Tech. Report Project_Autumn_1617, 2016.

[24] K. W. Lee, T. H. Kim, S. H. Kim, S. H. Lee and H. S. Lee, "Facial Expression Recognition using Dual Stage MLP with Subset Pre-Training," *Indian Journal of Science and Technology*, vol. 8, no. 25, pp. 1-7, 2015.

[25] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[26] L. Gui, T. Baltrušaitis and L. P. Morency, "Curriculum Learning for Facial Expression Recognition," in *Automatic Face & Gesture Recognition (FG 2017),* 2017, pp. 505-511.

[27] E. Correa, A. Jonker, M. Ozo and R. Stolk, "Emotion Recognition using Deep Convolutional Neural Networks," TU Delft, Delft, Holland, Tech. Report IN4015, 2016.

[28] A. Gudi. "Recognizing semantic features in faces using deep learning", M.S. thesis, Cornell University, 2015.

[29] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 1-9.

[30] Challenges in representation learning: Facial expression recognition challenge. [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge. Last access in September 2018.

[31] Affectiva's emotion AI. [Online]. Available: https://www.affectiva.com/. Last access in September 2018.