Multimedia Tools and Applications (2023) 82:11321–11336
https://doi.org/10.1007/s11042-022-13361-6

1226: DEEP-PATTERNS EMOTION RECOGNITION IN THE WILD

Emotion recognition by web-shaped model

Paola Barra[1] · **Luigi De Maio**[2] · **Silvio Barra**[3]

Received: 6 December 2021 / Revised: 14 March 2022 / Accepted: 3 June 2022 /
Published online: 16 August 2022

Abstract
Emotions recognition is widely applied for many tasks in different fields, from human-computer and human-robot interaction to learning platforms. Also, it can be used as an intrinsic approach for face recognition tasks, in which an expression-independent face classifier is developed. Most approaches face the problem by designing deeper and deeper neural networks that consider an expression as a still image or, in some cases, a sequence of consecutive frames depicting the temporal component of the expression. However, these suffer the training phase's computational burden, which can take hours or days to be completed. In this work, a Web Shaped Model is proposed, which consists of a geometrical approach for extracting discriminant features from a face, depicting the characteristics of an expression. The model does not need to be trained since it is applied on a face and centred on the nose tip, resulting in image size and face size independence. Experiments on publicly available datasets show that this approach reaches comparable and even better results than those obtained applying DNN-based approaches.

Keywords Emotion recognition · Expression recognition · Biometrics

1 Introduction

Emotion is a state of mind in which a person finds himself as an effect of a positive, negative, or neutral event. It is a psychological state common to all people, often unconscious, conditioned by individual aspects such as mood, health, disposition, temperament [9]. Recognizing an emotion means having information about what a person has seen, suffered, or heard.

Observing an emotion is helpful in many fields such as marketing, security, medicine, human-computer interaction (HCI), and automotive. It is possible to extract information of

✉ Silvio Barra
silvio.barra@unina.it

Paola Barra
barra@di.uniroma1.it

[1] Computer Science Department, Sapienza University of Rome, Rome, Italy

[2] Computer Science Department, University of Salerno, Fisciano, Italy

[3] ITEE, University of Naples "Federico II", Naples, Italy

a person's emotional state, such as stress in case of dangerous situations [34]; or is possible to find out if a person is behaving suspiciously from the way they walk [18]. In digital marketing through text, emotion detection, and sentiment analysis allow us to measure the degree of satisfaction of a product on brand tweets [3]. In medicine, through the analysis of the electrocardiogram (ECG), it is possible to identify and evaluate human emotions [19]. Also, in HCI field, emotion analysis is a field in full development, thus helping the machine to address a task in light of the emotion of the user [40].

For a long time, studies have mainly focused on seven basic emotions: *anger, contempt, disgust, fear, sadness, happy, surprise*, such as Ekman in [14], and subsequently was also introduced *neutral*. Indications of these can be found in facial expressions, which is the subject of this study. Within the wide field of face biometrics, expressions are one of the issues known to deform the face structure, complicating the recognition of a subject by increasing the intra-class variation. Therefore, it is common practice to extract facial muscular movements in order to develop an expression independent face classifier, [44]. Muscles change the shape of the principal components of the face, as for the lips and eyes (high intra-class variations). Different people have different faces, but they can have the same expression (low inter-class variations), how discussed by Ekman [15]. In [11] it was studied how people look at a face in a neutral emotional state and if the way is recognizable.

In light of this, it has become a very flourishing field of computer vision in the recognition of the expressions and emotions that transpire from a face [45]. Emotions analysis experiments were carried out on some reference datasets by re-adapting a geometric technique, the Web-Shaped Model (WSM) designed for pose estimation [6]. In work mentioned above, the coding of the input face is compared with the models of the poses to determine which pose the face belongs to. The encoding is obtained by applying a spider web model on the face centred on the nose tip; this model adapts to any shape and size of the input face.

Main contributions of this work are as follows:

- we propose an approach which does not need any image preprocessing phase and neither a training phase for computing the facial expression;
- the model is invariant to the dimension of the face and outputs the same feature vector independently from the size of the face;
- The model only needs 0.118 seconds to compute the feature vector and in this time is comprised also the time needed for the detection of the landmarks, which is an activity computed by an off-the-shelf tool that we have used for research purposes;
- even though no training phase is actually needed, the results on CK+ dataset are comparable with those at the state of the art.

The remainder of the paper is organized as follows: Section 2 analyzes and discusses the related methods for emotion recognition based on the use of geometric models or facial landmarks. Section 3 the WSM approach is presented. In Section 4 the experiments are shown, and the results are discussed. Section 5 concludes the paper.

2 Related works

Recognizing emotions from an expression means knowing how to annotate and discriminate face-related features, which generally refer to geometric or aesthetic aspects of the face itself [12]. Efficient Facial Expression Recognition (FER) methods are based on the search of a model which satisfies the fundamental requirements related to the expression: maximizing inter-class variability while minimizing intra-class variability.

The model must meet the following criteria: similar expressions from various subjects are highly similar, while distinct expressions from the same subject are very distinguishable. Table 1 shows a brief comparison between the considered related works, along with the reference. The typical FER procedure passes through three primary stages: *face detection, feature extraction, and classification*. The following types of criteria for the generation of expression models can be considered, in general: (i) texture; (ii) geometry; or (iii) mixed.

i texture is obtained by analyzing the face in its entirety;
ii geometric models are applied by considering the landmark points on the face and the geometric distance between them. Moreover, the landmark points can be joined in shape, considering a particular region of the face (i.e., the triangle formed by the tip of the nose and the corners of the mouth);
iii mixed models are formed by computing texture information in particular regions of the face.

In most approaches, facial images are preprocessed to generate a new, more synthetic, and meaningful texture: the so-called template [49]. Hence, the new texture is given as input to a feature extractor and then is examined by a classifier. About the texture model, some

Table 1 Comparison between related works

Method	Pros	Cons
Geometric Landmark [4, 24]	The approach is quite simple since is based on the position and location of well known points on the face of the subject. Successive phases can eventually analyze relations among the points and do some further analysis.	In cases the landmark detection approach fails, successive steps cannot be executed. Verify whether the landmark detection approach went well is not an easy task
Geometric Graph NN [31, 32]	When applied to a set of facial landmarks, GNNs build the graph right on the basis structure of the face. The model results to be highly reliable.	Still, GNNs strongly rely on the previous landmark detection phase. Also, there could be computational delays due to the complex structure of the model, specially in the weight update phases. Finally, the same graph can have different spatial representations and this can lead to biases while building the model.
Texture [1, 7, 13, 22, 29, 39, 39, 42, 48]	Easy and fast to compute. These feature extractor works quite well in very controlled condition.	The main disadvantage of such approaches is that they highly suffer PIE-O issues and therefore the intraclass variability tents to be quite variable
Mixed and [17, 24, 50]		These approaches model the expression in other ways: as an example, [50] considers the expression in its temporal component therefore analyzes it buy using a spatiotemporal network. Fan et al. [17], instead applies an attention mechanism for storing local appearance of the expression. The more the complexity and the final accuracy, the higher the computational cost.

classifiers have proved to be very effective in FER, such as Support Vector Machine (SVM), Artificial Neural Network (ANN), or Random Forest (RF) [7, 8, 23, 37]. The actual state of the art approaches the emotion recognition topic from three different perspectives:

1. emotion recognition from speech [20], EEG [46] and other sources [41];
2. emotion recognition from multiple sources [5, 33];
3. emotion recognition from facial expression [26, 27, 51];

Within the third point, expression classification has been faced from different points of view: many approaches have dealt with facial feature extraction by applying local models which have proven their efficiency in the past like Local Binary Pattern (LBP) [42], Gabor wavelets [7], Principal Component's analysis (PCA) [1, 22], Local Directional Number pattern (LDN) [39], Histograms of Oriented Gradients (HOG) [13] and Haarclassifier in [48]. Some of these methods use the resulting new texture and leave the total interpretation of the data to the classifier. Others stack the features of salient areas or points in a vector to classify it, as experimented with the Gabor wavelets [29], and the LDN in [39]. On the one hand, these extractors consider all intrinsic characteristics of an image in its entirety; on the other hand, they do not consider any semantic reasoning about the area the features are extracted from. In the work reported by Minaee and Abdolrashidi [30] the areas helpful for FER have been identified. They explain that every emotion involves a different face area using a subtractive and iterative experimentation technique and a Deep Neural Network (DNN).

Among the geometric methods, we divide them into two categories: *Graph Neural Networks-based (GNN-based)* and *Landmark-based*. There are several articles in which facial landmarks have been used as geometric traits and fed to a classifier for performance emotion recognition (Landmark-based) [4]. In [32] the authors used facial landmarks as vertices of a direct graph, which represents the input for a Directed Graph Neural Network (GNN-based). The work in [38] analyzes facial landmarks to recognize both facial expression and affective speech; this study was tested on the SAVER dataset in which expression recognition is faced by analyzing local features changing through consecutive frames.

In the study of emotion recognition in computer vision, the face is first detected, and then the most suitable facial features are extracted. These features are generally referred to as geometric features or aesthetic features and are used to encode emotion [12].

Most recent approaches focused on developing Deep Neural Network architectures, in charge of extracting relevant features according to the data given in the training phase; these approaches have shown the highest recognition rates. In order to balance the dataset and enlarge the number of samples, GAN-based data augmentation approaches have been considered. In [36] the authors have explored several data augmentation techniques in the FER field, showing that choosing an efficient data augmentation approach can improve the experimental results up to 30%.

Most recent approaches for expression recognition are mainly based on the designing and development of more and more complex neural networks architectures. In [50], the authors have proposed DeRL (De-Expression Residue Learning), in which the model generates, for every image in input, the corresponding face with a neutral expression. The informative content of the expression is saved in the intermediate layers of the network; this expression is used for classifying the expression, given a dataset labelled with seven main expressions.

Many approaches also have focused on training attention mechanisms, which focus on the regions of a face that give the most significant support for recognizing an expression. This approach has been used in [28] in which the authors use LBP for local region feature extraction and passing the vector to a CNN. A more complex case is shown in [17], in

which a Deeply Supervised Attention Network (DSAN) is developed, consisting of blocks that memorize the local appearance and textures of facial regions.

Since an expression can also be seen as a temporal sequence of frames that compose the expression development on a face, some approaches have been designed to analyze the temporal sequence itself. In [21], as an example, a Deep Joint Spatiotemporal Network (DJSTN) is developed, which exploits the 3D convolution to involve the temporal component of an expression in the extraction of features.

Deep approaches can obtain very high results from one side. However, the computational burden of the training phase is quite high, and the setting of the hyperparameters may be long and cumbersome.

This consideration led to developing a geometric approach to creating a feature template that summarizes a facial expression. In this work, a well-known model is applied for detecting facial landmarks on the face and based on these, regions are created, and the related characteristics are extracted. The work followed these rationales and was inspired by a technique already tested for head pose estimation involving the Web Shape Model (WSM) [6], described in the next section. This technique considers the location of the facial landmarks relating to an expression, concerning a subdivision into sectors in the shape of a web positioned on the face and centred on the nose.

3 Web-shaped model

The WSM method previously mentioned is a new technique as a feature extractor of face pose estimation. It uses a geometric pattern like a shape of a spider web drawn on a face. This model is fixed and identifies the areas of the face where facial landmarks are present. Previously, the WSM was used for head pose estimation; in that case, the authors experimented with it to recognize emotion from a face to check if different individuals have similar expressions when they feel the same emotion.

The Web-shaped model (WSM) method presented was created for the head pose estimation. WSM method is described in Fig. 1:

a) input face;
b) landmark detection on the face in input, using the Kazemi-Sullivan algorithm [25] ;
c) drawing of a virtual spider web on the identified face which will create the encoding of the face.
d) associate each spider-web encoding with an emotion.

Facial landmarks are points represented by coordinates (x, y) on the image of a face. The 68 facial landmarks, Fig. 1 b), precisely indicate the position of the eyes, nose, mouth, jaw, and eyebrows. The spider-web is built on the face in input, Fig. 1 a); it is so-called because the shape resembles a spider's web. The spider-web is formed by concentric circles and rays (slices), Fig. 1 c), built with the centre on the tip of the nose (represented by landmark 33). The outermost circle of the spider-web has a long radius from the tip of the nose to the farthest landmark, which, depending on the pose of the face, can be the eyebrow, chin, jaw or other. This peculiarity allows the model to be invariant to the dimension of the face, given the fact that if we apply a scaling factor (greater or lower than 1), the model does not change and as a consequence the feature vector in output remains the same as well.

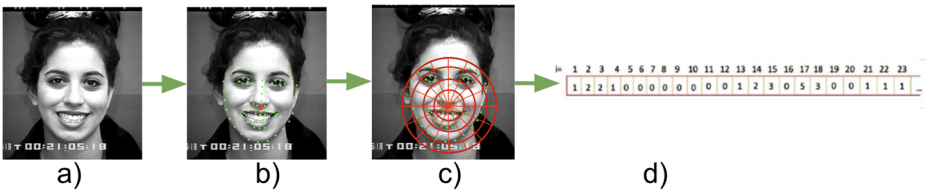

Fig. 1 The phases of encoding using WSM: face input, extraction of the 68 face landmarks, design of the spider-web and output of the resulting encoding

Different configurations of spider-web were used in the following experiments, changing the number of concentric circles and slices. Finally, the spider-web creates sectors, each of which contains the number of landmarks in it; the array with the number of landmarks for each sector represents the spider-web encoding, Fig. 1 d).

Each spider-web encoding is associated with an emotion. Once the spider-web encoding has been obtained from an input image, it is compared with the encodings present in the database; an emotion is associated with it based on the classification algorithm used.

4 Experimental section

4.1 Dataset: CK+

They use the definition of primary emotion by Ekman and Friesen [14] who separated emotion into eight classes, Fig. 2.

4.2 Dataset: KDEF

The Karolinska Directed Emotional Faces[1] (KDEF)[10] is a dataset of 4900 pictures of human facial expressions. The dataset contains 70 individuals displaying 7 different emotional expressions 5 different head pose, Fig. 3.

4.3 Mixed dataset and data augmentation

The frontal images of the datasets CK+ and KDEF were merged to increase the number of samples and evaluate a set of heterogeneous data. Since the method is of geometric type, the heterogeneous samples have been placed in the common seven classes: anger, neutrality, fear, disgust, happiness, sadness and surprise. The elements of each class of images were further doubled by applying horizontal flipping; vertically, it would make little sense; so the dataset has increased data. The horizontal flip operation ensures class preservation; the emotion does not change but change the web encoding: points are placed in different sectors. Comparing the Figs. 4a) and b) it is possible to see a case, specially in the eye and eyebrow region.

[1] https://www.kdef.se/

Fig. 2 Eight examples of facial expressions from the CK+ dataset

4.4 Different model configurations

Different configurations of the spider-web have been created. In the case of [6], the best configuration is the spider-web 4x3 that is 4 concentric circles and 14 sectors, 12 slices in total; this configuration was found to be the best in the case of head pose classification, but it is not necessarily the same in the case of emotion detection. Similarly, the experiments are done for different web configurations by changing the number of rays and equidistant concentric circles. For a complete picture, experiments were carried out by varying all combinations of concentric circles (from 3 to 8) and of slices per quadrant (from 3 to 16).

4.5 Experimental setting

The experiments has been executed by using the K-nearest neighbors classifier upon a LOOCV (Leave One Out Cross Validation); the choice of the classifier has been made on an empirical basis. For both KNN and LOOCV the used implementation has been the one defined within the scikit-learn python library [35]. The used datasets have been divided into as many folds as the total number of samples. The accuracy of each experiment is defined by the (1), where i is the $i-th$ experiment. The classifier has been trained on all data except one which has been used for testing; this was done as many times as the number of samples in the dataset. So, all images were used as tests, and the final accuracy is the average of the

Fig. 3 Seven examples of facial expressions from the KDEF dataset

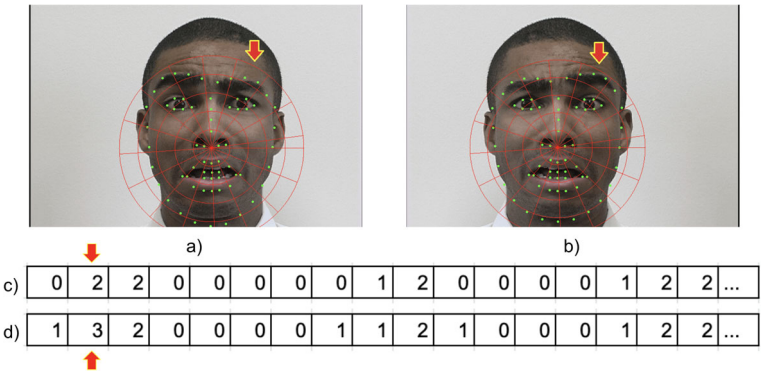

Fig. 4 Flipped images from CK+ dataset. a) the original, b) the flipped, c) coding vector of the original, d) coding vector of the flipped

accuracies of the experiments performed, (2), where N is the number of experiments.

$$Accuracy_i = \frac{|Correctly\ recognized\ expression\ samples_i|}{|expression\ samples_i|} \tag{1}$$

$$Accuracy = \frac{\sum_{i=1}^{N} Accuracy_i}{N} \tag{2}$$

The best spider-web configurations are calculated in 5x14 for CK+(8cls), 4x15 for CK+(7cls), and 7x12 for CK+ ∪ KDEF, Fig. 10 .

Given the strong imbalance of the datasets (CK+, specially), these were balanced by several augmentation strategies, as stated in Section 4.3: for CK+, KDEF and CK+ ∪ KDEF 108, 272 and 413 new samples have been generated respectively. The popularity of the datasets are shown in Table 2. As stated above, the data were divided into training and test sets using the Leave One Out Cross Validation technique.

Table 2 Number of samples for each class of each dataset

Class	CK+	KDEF	CK+ ∪ KDEF
Anger	267	260	527
Disgust	345	272	617
Fear	148	278	426
Happy	411	278	689
Neutral	1177	277	1454
Sadness	137	276	413
Surprise	466	279	745
Contempt	108	-	-
Total	3059	1920	4871

4.6 Results discussion

All the samples were used in both the testing and training phases. Compared to a random split 70% train and 30% test set, this type of subdivision allows to repeat the experiments and always get the same result. Therefore, we can state the the experiments are fully reproducible. Figures 5, 6, 7, 8 and 9 show the average accuracy obtained for each of the experiments; the color is used to indicate which experiments better results have been obtained: the lighter the color the higher the accuracy. The experiments on the CK+ dataset has been carried out by considering in turn the all 8 classes (anger, disgust, fear, happy, neutral, sadness, contempt and surprise,), 7 classes (all except neutral) or 6 classes (all except neutral and contempt); the choice has been done in order to obtain a fair comparison against the state of the art existing papers. As shown in Fig. 5 the configuration 5x14 has obtained 91.5% of accuracy in the classification of all 8 emotions of the CK+ dataset, the image of this web configuration is represented in Fig. 10 (top-left).

In Fig. 6 is shown that the configuration 8x15 obtains 96% accuracy in the classification of 7 emotion classes of the CK+ dataset, the image of this web configuration is represented in the Fig. 10 (top-centre).

In contrast, the configuration 8x14 obtains 98% accuracy in the classification of 6 emotions of the CK+ dataset, as can be seen in Fig. 7, the image of this web configuration is represented in the Fig. 10 (top-right).

For the KDEF dataset, the best configuration consists of 8x12 and achieves 67.6% accuracy, Fig. 8; the image of this web configuration is represented in the Fig 10 (bottom-left).

For the union of the CK+ and KDEF datasets, the configuration 7x12 obtains the accuracy of 75%, Fig. 9, the image of this web configuration is represented in the Fig. 10 (bottom-right).

Table 3 shows in detail the accuracy obtained for each expression class; the trend is similar even for different datasets; “sadness” is always the emotion classified with the lowest accuracy; “happy” is always the emotion classified with the highest accuracy. Furthermore, the average accuracy is higher for the CK+ dataset, which correctly classifies 6 and 7 classes. In both cases, the neutral class is highly prone to be misclassified, given the fact that the tool tents to place the singular sample into a well specific class rather than classifying it as neutral.

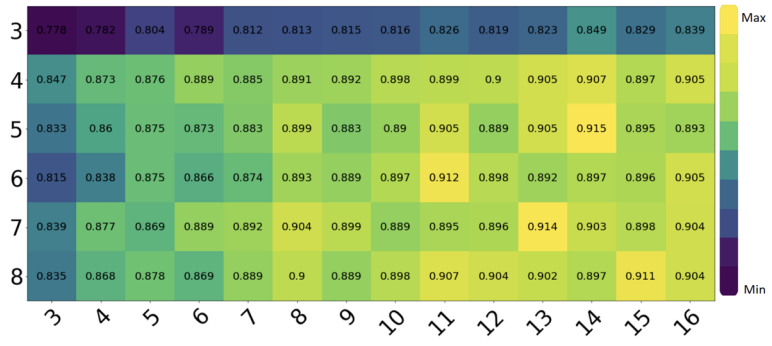

Fig. 5 Accuracies for CK+ dataset (8 classes). X axes = circles. Y axes = slices for quadrant

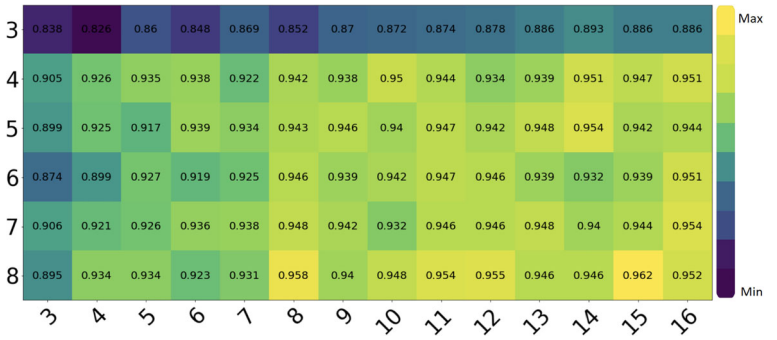

Fig. 6 Accuracies for CK+ dataset (7 classes). X axes = circles. Y axes = slices for quadrant

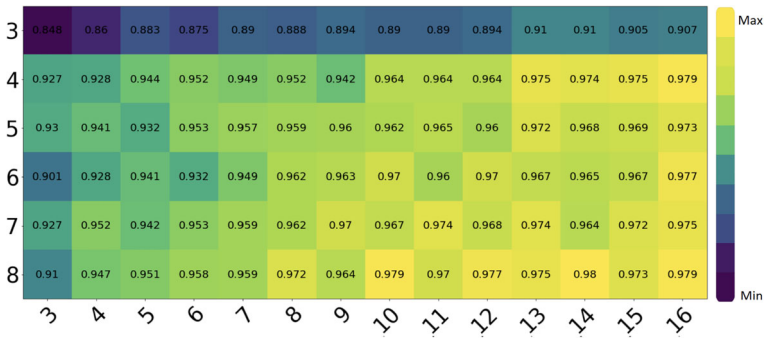

Fig. 7 Accuracies for CK+ dataset (6 classes). X axes = circles. Y axes = slices for quadrant

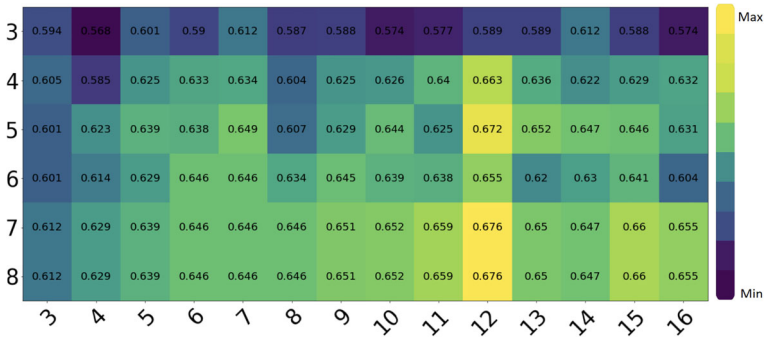

Fig. 8 Accuracies for KDEF dataset. X axes = circles. Y axes = slices for quadrant

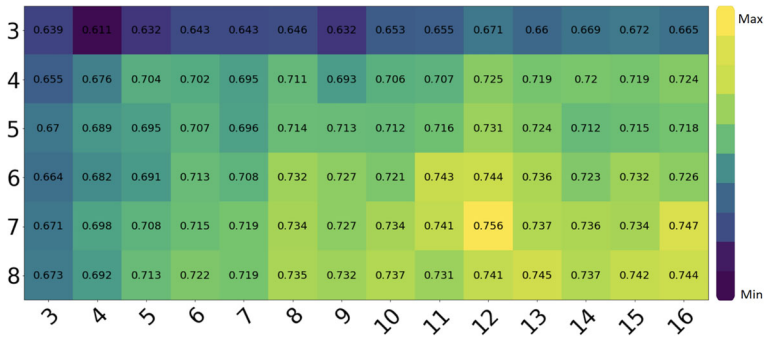

Fig. 9 Accuracies for CK+ ∪ KDEF dataset. X axes = circles. Y axes = slices for quadrant

Table 4 shows the comparative results for the CK+ dataset, which is currently the most widely used in the state of the art. One of the main advantage of the proposed approach is the capability of working with different numbers of classes and still obtaining high accuracies (above 90%) perfectly comparable to the state of the art methods. Unfortunately for KDEF dataset, whose results are shown in Table 5, does not allow to obtain good results, mainly due a twofold reason: firstly, the number of samples is quite low with respect to the one in CK+. Secondly, the images in KDEF consist of simulated expressions and thus, in many cases, these are extremely emphasized. As a consequence, the results in Table 3, related to the last column, have been strongly affected by the bad results obtained in KDEF.

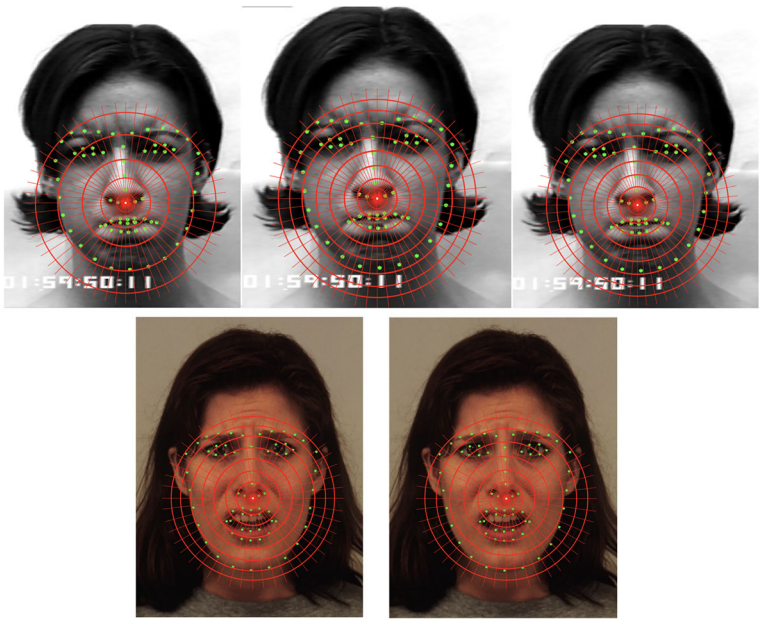

Fig. 10 The WSM configurations with the highest accuracy: 5x14 (top-left), 8x15 (top-center), 8x14 (top-right), 8x12 (bottom-left), 7x12 (bottom-right)

Table 3 Detailed results of emotion recognition accuracy

Class	CK+(6)	CK+(7)	CK+(8)	KDEF	CK+ ∪ KDEF
Anger	0.96	0.96	0.93	0.55	0.72
Disgust	0.93	0.94	0.91	0.52	0.73
Fear	0.94	0.95	0.95	0.47	0.70
Happy	0.96	0.98	0.98	0.89	0.94
Neutral	-	-	0.68	0.64	0.65
Sadness	0.91	0.89	0.87	0.46	0.69
Surprise	0.99	0.99	0.98	0.80	0.86
Contempt	-	1	0.99	-	-
Average	0.98	0.96	0.915	0.676	0.756

On the column headers in round brackets are the number of classes

Table 4 Accuracy comparison of CK+ and existing methods in percentage

Method	Acc. (6cls)	Acc. (7cls)	Acc. (8cls)
Pu et al. [37]	-	0.96	-
Khorrami et al. [23]	0.98	-	0.96
Ramirez Rivera et al. [39]	-	0.89	-
Minaee and Abdolrashidi [30]	-	0.98	-
Álvarez et al. [4]	-	-	0.88
Ngoc et al. [32]	-	0.96	-
Yang et al. [50]	-	0.97	-
Li et al. [28]	-	0.98	-
Fan et al. [17]	0.98	-	-
Umer et al. [47]	-	0.97	-
Proposed	0.98	0.96	0.91

Table 5 Accuracy comparison of KDEF and existing methods

Method	Accuracy (%)
Sun et al. [43]	0.82
Umer et al. [47]	0.83
Eng et al. [16]	0.81
Akhand et al. [2]	0.99
Proposed	0.68

The average time on 5000 images is 0.112 seconds for image, of which: 0.102 seconds for landmark detection (for 640x490 resolution images) and 0.015 seconds for the construction of the spider web. The time taken by the Kazemi-Sullivan algorithm for landmark detectors depends on the size of the image. The hardware involved in the computation of the processing time is the following: Apple MacBook Pro with 2.9GHz IntelCore I9. No GPU has been used.

5 Conclusions and limitations

This research analyzed a method for recognizing emotions on a face based on a geometric approach that analyzes landmark points through a virtual spider web on the face. Emotion coding is classified using a K-nearest neighbour classifier on the CK+ and KDEF datasets. Different configurations were used for the web-shaped structure to find the optimal one for our purposes. The method used does not need training because it studies the position of the facial reference points on the web, and this structure adapts to any size of the face or kind of face image. Despite the lack of training, this method brings competitive results with the state of the art. Furthermore, a large margin for improvement in the future could be brought about by the use of data augmentation techniques and the use of a different landmark detector. The approach, even if very simple in its structure and definition, still allows to obtain very good results comparable with the state of the art. On the basis of the conducted experiments, the results show that without a heavy training phase, the results are highly comparable with the state of the art specially when the CK+ dataset is used. The approach has been proposed in the past by this laboratory for head pose estimation issues, still behaving quite well. This means that with a single pass over a head sample, both head pose and facial expression may be classified. There are several limitations which showed up in this research specially as regards two factors: (i) the neutral expression is hardly recognizable since the approach itself prefers classifying a single sample within one of the classes which show a well defined expression. Geometrically speaking, this happens because the cluster related to the neutral expression tents to have some outliers in the training phase which deform the cluster topology, thus intersecting with other clusters; (ii) The KDEF shows results which are definitely different from those obtained with CK+ and this is mainly due to the lower popularity of the dataset.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Declarations All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Paola Barra], [Luigi De Maio] and [Silvio Barra]. The first draft of the manuscript was written by [Paola Barra] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

- The authors did not receive support from any organization for the submitted work.
- No funding was received to assist with the preparation of this manuscript.
- No funding was received for conducting this study.
- No funds, grants, or other support was received.
- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Abdulrahman M, Gwadabe TR, Abdu FJ, Eleyan A (2014) Gabor wavelet transform based facial expression recognition using pca and lbp. In: 2014 22nd Signal processing and communications applications conference (SIU), pp 2265–2268. https://doi.org/10.1109/SIU.2014.6830717
2. Akhand MAH, Roy S, Siddique N, Kamal MAS, Shimamura T (2021) Facial emotion recognition using transfer learning in the deep cnn. Electronics, 10. https://doi.org/10.3390/electronics10091036
3. Al-Hajjar D, Syed AZ (2015) Applying sentiment and emotion analysis on brand tweets for digital marketing. In: 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), pp 1–6. https://doi.org/10.1109/AEECT.2015.7360592
4. Álvarez VM, Sánchez CN, Gutiérrez S, Domínguez-Soberanes J, Velázquez R (2018) Facial emotion recognition: A comparison of different landmark-based classifiers. In: 2018 International conference on research in intelligent and computing in engineering (RICE), pp 1–4. https://doi.org/10.1109/RICE.2018.8509048
5. Ayata D, Yaslan Y, Kamasak ME (2020) Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. J Med Biol Eng 40:149–157
6. Barra P, Barra S, Bisogni C, De Marsico M, Nappi M (2020) Web-shaped model for head pose estimation: An approach for best exemplar selection. IEEE Trans Image Process 29:5457–5468. https://doi.org/10.1109/TIP.2020.2984373
7. Bartlett M, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: 2005 IEEE Computer society conference on computer vision and pattern recognition (CVPR'05), vol 2, pp 568–573. https://doi.org/10.1109/CVPR.2005.297
8. Bettadapura V (2012) Face expression recognition and analysis: The state of the art. arXiv:1203.6722
9. Cabanac M (2002) What is emotion? Behav Processes 60:69–83. https://doi.org/10.1016/S0376-6357(02)00078-5
10. Calvo M, Lindqvist D (2008) Facial expressions of emotion (kdef): Identification under different display-duration conditions. Behav Res 40:109–115. https://doi.org/10.3758/BRM.40.1.109
11. Cantoni V, Porta M, De Maio L, Distasi R, Nappi M (2012) Towards a novel technique for identification based on eye tracking. In: 2012 IEEE Workshop on biometric measurements and systems for security and medical applications (BIOMS) proceedings, pp 1–4. https://doi.org/10.1109/BIOMS.2012.6345780
12. Cohn JF, Ambadar Z, Ekman P (2007) Observer-based measurement of facial expression with the facial action coding system. vol 1. Oxford University Press, New York
13. Dahmane M, Meunier J (2011) Emotion recognition using dynamic grid-based hog features. In: 2011 IEEE International conference on automatic face gesture recognition (FG), pp 884–888. https://doi.org/10.1109/FG.2011.5771368
14. Ekman P (1971) Universals and cultural differences in facial expressions of emotion. Neb Symp Motiv 19:207–283
15. Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200. https://doi.org/10.1080/02699939208411068
16. Eng SK, Ali H, Cheah AY, Chong YF (2019) Facial expression recognition in JAFFE and KDEF datasets using histogram of oriented gradients and support vector machine. IOP Conf Ser: Mater Sci Eng 705:012031. https://doi.org/10.1088/1757-899x/705/1/012031
17. Fan Y, Li V, Lam JC (2020) Facial expression recognition with deeply-supervised attention network. IEEE Transactions on Affective Computing, p 1–1. https://doi.org/10.1109/TAFFC.2020.2988264
18. Freire-Obregón D, Castrillón-Santana M, Barra P, Bisogni C, Nappi M (2020) An attention recurrent model for human cooperation detection. Comput Vis Image Underst 102991:197–198. https://doi.org/10.1016/j.cviu.2020.102991

19. Guo HW, Huang YS, Chien JC, Shieh JS (2015) Short-term analysis of heart rate variability for emotion recognition via a wearable ecg device. In: 2015 International conference on intelligent informatics and biomedical sciences (ICIIBMS), pp 262–265. https://doi.org/10.1109/ICIIBMS.2015.7439542
20. Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control 59:101894
21. Jeong D, Kim B-G, Dong SY (2020) Deep joint spatiotemporal network (djstn) for efficient facial expression recognition. Sensors 20:1936. https://doi.org/10.3390/s20071936
22. Juanjuan C, Zheng Z, Han S, Gang Z (2010) Facial expression recognition based on pca reconstruction. In: 2010 5th International conference on computer science education, pp 195–198. https://doi.org/10.1109/ICCSE.2010.5593658
23. Khorrami P, Paine TL, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition?. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). p 19–27. https://doi.org/10.1109/ICCVW.2015.12
24. Kalyan Kumar V, Suja P, Tripathi S (2016) Emotion recognition from facial expressions for 4d videos using geometric approach. In: Advances in Signal Processing and Intelligent Recognition Systems, Springer, pp 3–14
25. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on computer vision and pattern recognition, pp 1867–1874. https://doi.org/10.1109/CVPR.2014.241
26. Khaireddin Y, Chen Z (2021) Facial emotion recognition: State of the art performance on fer2013. arXiv:2105.03588
27. Kwon S et al (2021) Att-net: Enhanced emotion recognition system using lightweight self-attention module. Appl Soft Comput 102:107101
28. Li J, Jin K, Zhou D, Kubota N, Ju Z (2020) Attention mechanism-based cnn for facial expression recognition. Neurocomputing 411:340–350. https://doi.org/10.1016/j.neucom.2020.06.014
29. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings third ieee international conference on automatic face and gesture recognition, pp 200–205. https://doi.org/10.1109/AFGR.1998.670949
30. Minaee S, Abdolrashidi A (2021) Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors (Basel, Switzerland) 21. https://doi.org/10.3390/s21093046
31. Nayak S, Routray A, Sarma M, Uttarkabat S (2022) Gnn based embedded framework for consumer affect recognition using thermal facial rois. IEEE Consumer Electronics Magazine
32. Ngoc QT, Lee S, Song BC (2020) Facial landmark-based emotion recognition via directed graph neural network. Electronics, 9. https://doi.org/10.3390/electronics9050764
33. Park CY, Cha N, Kang S, Kim A, Khandoker AH, Hadjileontiadis L, Oh A, Jeong Y, Lee U (2020) K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci Data 7:1–16
34. Partila P, Tovarek J, Rozhon J, Jalowiczor J (2019) Human stress detection from the speech in danger situation. In: S. S. Agaian, V. K. Asari, S. P. DelMarco (Eds.), Mobile Multimedia/Image Processing, Security, and Applications 2019, volume 10993, International Society for Optics and Photonics, publisher SPIE, pp 179–185. https://doi.org/10.1117/12.2521405
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 1:2825–2830
36. Porcu S, Floris A, Atzori L (2020) Evaluation of data augmentation techniques for facial expression recognition systems. Electronics 9. https://doi.org/10.3390/electronics9111892
37. Pu X, Fan X, Chen X, Ji L, Zhou Z (2015) Facial expression recognition from image sequences using twofold random forest classifier. Neurocomputing 168:1173–1180. https://doi.org/10.1016/j.neucom.2015.05.005
38. Rahdari F, Rashedi E, Eftekhari M (2019) A multimodal emotion recognition system using facial landmark analysis. Iran J Sci Technol Trans Electr Eng 43:171–189. https://doi.org/10.1007/s40998-018-0142-9
39. Ramirez Rivera A, Rojas Castillo J, Oksam Chae O (2013) Local directional number pattern for face analysis: Face and expression recognition. IEEE Trans Image Process 22:1740–1752. https://doi.org/10.1109/TIP.2012.2235848
40. Schuller B, Reiter S, Muller R, Al-Hames M, Lang M, Rigoll G (2005) Speaker independent speech emotion recognition by ensemble classification. In: 2005 IEEE international conference on multimedia and expo, pp 864–867. https://doi.org/10.1109/ICME.2005.1521560
41. Sepúlveda A, Castillo F, Palma C, Rodriguez-Fernandez M (2021) Emotion recognition from ecg signals using wavelet scattering and machine learning. Appl Sci 11:4945

42. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: A comprehensive study. Image Vis Comput 27:803–816. https://doi.org/10.1016/j.imavis.2008.08.005
43. Sun Z, Hu ZP, Wang M, Zhao SH (2017) Discriminative feature learning-based pixel difference representation for facial expression recognition. IET Comput Vis 11:675–682. https://doi.org/10.1049/iet-cvi.2016.0505
44. Tan H-C, Zhang Y-J (2008) Expression-independent face recognition based on higher-order singular value decomposition. In: 2008 International conference on machine learning and cybernetics, vol 5, pp 2846–2851. https://doi.org/10.1109/ICMLC.2008.4620893
45. Tian Y-I, Kanade T, Cohn J (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23:97–115. https://doi.org/10.1109/34.908962
46. Torres EP, Torres EA, Hernández-Álvarez M, Yoo SG (2020) Eeg-based bci emotion recognition: A survey. Sensors 20:5083
47. Umer S, Rout R, Pero C, Nappi M (2021) Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-020-02845-8
48. Whitehill J, Omlin C (2006) Haar features for facs au recognition. In: 7th International conference on automatic face and gesture recognition (FGR06), pp 5–101. https://doi.org/10.1109/FGR.2006.61
49. Xie X, Lam K-M (2009) Facial expression recognition based on shape and texture. Pattern Recognit 42:1003–1011
50. Yang H, Ciftci U, Yin L (2018) Facial expression recognition by de-expression residue learning. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 2168–2177. https://doi.org/10.1109/CVPR.2018.00231
51. Zhang K, Li Y, Wang J, Cambria E, Li X (2021) Real-time video emotion recognition based on reinforcement learning and domain knowledge. IEEE Transactions on Circuits and Systems for Video Technology