A Novel Model For Emotion Detection From Facial Muscles Activity

Bagheri, Elahe; Bagheri, Azam; Gomez Esteban, Pablo; Vanderborght, Bram

Published in:
Robot 2019: Fourth Iberian Robotics Conference

Publication date:
2019

A Novel Model For Emotion Detection From Facial Muscles Activity

Elahe Bagheri[1], Azam Bagheri[2], Pablo G. Esteban[1], and Bram Vanderborght[1]

[1]Robotics and Multibody Mechanics Research Group, Vrije Universiteit Brussel and Flanders Make, Brussels, Belgium
[2]Electrical Engineering Group, Chalmers University of Technology, Gothenburg, Sweden
elahe.bagheri,pablo.gomez.esteban,bram.vanderborght@vub.be
bazam@chalmers.se

Abstract. Considering human's emotion in different applications and systems has received substantial attention over the last three decades. The traditional approach for emotion detection is to first extract different features and then apply a classifier, like SVM, to find the true class. However, recently proposed Deep Learning based models outperform traditional machine learning approaches without requirement of a separate feature extraction phase.
This paper proposes a novel deep learning based facial emotion detection model, which uses facial muscles activities as raw input to recognize the type of the expressed emotion in the real time. To this end, we first use OpenFace to extract the activation values of the facial muscles, which are then presented to a Stacked Auto Encoder (SAE) as feature set. Afterward, the SAE returns the best combination of muscles in describing a particular emotion, these extracted features at the end are applied to a Softmax layer in order to fulfill multi classification task.
The proposed model has been applied to the CK+, MMI and RADVESS datasets and achieved respectively average accuracies of 95.63%, 95.58%, and 84.91% for emotion type detection in six classes, which outperforms state-of-the-art algorithms.

Keywords: Facial Emotion Recognition, Facial Muscles Activity, Stacked Auto Encoder, Facial Action Units

1 INTRODUCTION

For the last two decades we have been started to use different smart devices and applications in our daily life, robots are going to be used in our shops [**?**], schools and hospitals [**?**]. However, lots of them lose their favor by losing novelty effect. Haag et al [**?**] argued the communication between humans and systems can improve by considering emotions as an additional interaction modality. Meanwhile researchers showed the systems which recognize and respond to human's emotions are more caring, likable, supportive and trustworthy [**?**]. Hence, recognizing human's emotion became an important topic to study.
Emotion detection is the ability to recognize another's affective state, which typically involves the integration and analysis of expressions through different modalities, like

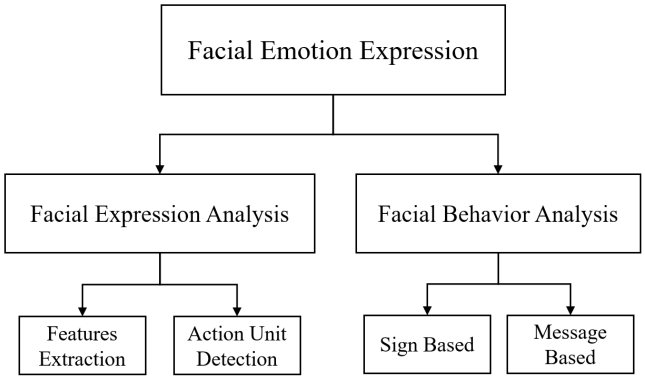

Fig. 1: General Facial Emotion Recognition approaches.

facial expression, speech, body movements and gestures [**?**]. Since 55% of human emotions are conveyed by facial expression [**?**], Facial Emotion Recognition (FER) is the most investigated method for human emotion recognition task.

FER contains two main parts, facial expression analysis and facial behavior analysis, as shown in Figure **??**. Facial expression analysis carried out via two main approaches, feature extraction and Action Unit (AU) detection. The feature extraction approaches proceed on by detecting face region and facial components, e.g., eyebrows, eyes, nose and mouth from an input image. Then two different types of features are extracted: *geometric* and *appearance* features. Geometric features represent the positions of salient points of the face, e.g., ends of the eyes, end of the nose, mouth and the shape of the facial components, while appearance features represent the text variations of the face, e.g., color, edge density, crinkles, and wrinkles [**?**]. Finally, the pre-trained machine learning classifier attempts to classify the given face as portraying one emotion [**?**].

The AU detection methods, however, are independent of facial appearance and analyse facial muscles movements by tracking AUs. Each AU indicates fundamental movements of a single or a group of muscles [1]. Through facial expression of different emotions, different combinations of AUs are activated. Ekman [**?**] defined the Facial Action Coding System (FACS), which encodes the movements of AUs to describe human facial movements and converts the detected AUs to the corresponding emotion. An important advantage of the AU detection methods is that they remove the need of analysing complex high-dimensional features [**?**].

Facial behavior analysis is the other way to perform FER. Cohn et al [**?**] proposed two conceptual approaches for studying the facial behavior: "message-based" approach and "sign-based" approach. Message-based approaches categorize facial behaviors as the meaning of expressions and are widely used by psychologists. Message-based methods can be divided into discrete categorical and continuous dimensional methods. Discrete categorical methods assign an expression to one of pre-defined prototypical categories, including six basic emotions proposed by Ekman [**?**] like anger, disgust, fear, happiness, sadness, and surprise, while continuous emotional methods describe each facial

[1] `https://imotions.com/blog/facial-action-coding-system/`

expression by continuous axes, such as arousal and valence [**?**].
Sign-based approaches, however, describe facial actions regardless of their meaning, and different expressions are classified based on the activated AUs [**?**]. Indeed, sign-based approaches are similar to AU detection approaches.
Since sign-based algorithms are trained to detect activated AUs in a given image or video to recognize the emotion, the sign-based FER problems can be transformed into the problem of activated AU detection [**?**]. Hence, applying a proper toolkit, like OpenFace [**?**] the activation values of facial AUs can be obtained and used for model training for emotion detection. However, as Du et al [**?**] showed, determining the exact combination of activated AUs in each emotion is difficult. Thereby, the main contribution of this study is finding the most pivotal activated AUs in each emotion. To this end, we developed a Stacked Auto Encoder (SAE) deep network on the statues of 15 facial AUs to extract the high-order features of the input data that is not possible to obtain by humans. Given automatic extracted features, we added a Softmax layer to full-fill the classification task.
The remain of this paper is structured as follows: Section **??** presents a review on previous work. The proposed model is illustrated in Section **??**. Section **??** demonstrates the experimental results. Finally, Section **??** concludes this paper.

2 RELATED WORK

Originally classical machine learning algorithms such as Bayesian Networks [**?**], Gaussian Mixture Models [**?**], Hidden Markov Models [**?**], and Neural Networks [**?**] have been applied to detect expressed facial emotions. The quality of the training data, e.g., image resolution, face view angle and also the way emotions are labeled, strongly influences the results of the training algorithm and is the main obstacle for classical FER algorithms.
In contrast, promising results of neural network methods and deep learning (DL) based approaches in comparison with classical machine learning algorithms, caused to propose numerous DL based FER methods in the research community. Emergence of deep learning as a general end to end learning approach dispels handcraft feature detection problem too [**?**].
There are two approaches in FER, one which does not use the input's temporal information so called frame-based, and the other, which uses the temporal information of images and is known as sequence-based. The input in frame-based approaches is an image without a reference frame, while the input in sequence-based approach is a sequence of one or more frames [**?**]. Since our proposed model categorize as frame based, in this section we focus on the state-of-the-art algorithms of the frame-based methods.
Pitaloka et al [**?**] used a Convolutional Neural Network (CNN) based method to recognize 6 basic emotions. The proposed method comprises of 5 layers including two sets of convolution layer, two max-pooling layers and a fully connected layer for classification. After pre-processing, the input image is fed to the first convolution layer to extract features like edges, corners and shapes. The output image then is passed to the first max-pooling layer to reduce the image size. The compact image then is sent to the second convolution layer to obtain higher order features and afterward is passed to the

second max-pooling layer to reduce the final output size. The fully connected layer at the end, classifies the output image into one of the six basic emotions. However, the performance of the proposed algorithm decreases when the dimension of images is increased regarding to the complexity of the high dimensional images.
Liu et al [**?**] proposed a sign-based deep neural network architecture called AU-aware Deep Networks (AUDN) in order to investigate the effect of AUs in emotion recognition. The proposed AUDN includes three sequential modules. In first module a convolution layer stacked by a max-pooling layer generates a complete representation of all expression-specific appearance variations. Then in the second module, an AU-aware receptive field layer searches the subsets of the over-complete representation to find the best simulating of the combination of the AUs. The third module consists of multilayer Restricted Boltzmann Machines (RBM) to learn hierarchical features. Once the features obtain, a linear SVM classifier is applied to recognize the six basic emotions. However, AU-aware layers, in second module, are not able to detect all FACS in images.
Although different state-of-the-art algorithms are proposed in the field of FER, emotion detection has remained a challenging problem in computer vision. In this study, we propose a new SAE-based model to cope with the challenge of the FER in two steps. In the first step, the proposed SAE aims at extracting the most pivotal AUs and in the next step these extracted AUs are applied to the categorical Softmax classifier to detect six basic emotions. Next section details the proposed model.

3 Proposed Model

According to the sign-based FER approaches, one way to recognize facial emotion expression is detecting the status of all individual AUs and then analyzing combinations of activated AUs. For example, if a face has been analyzed as having activated AU5, and AU26, a properly trained algorithm should classify it as expressing “surprise”. However, Du et al [**?**] showed that encoding the activated AUs into a specific emotion is difficult, if the expressed emotion is a mixture of several emotions. For instance, when some one is surprised by a good news all AUs related to both happiness and surprise can be activated, however, if be shocked of an online scam, the AUs related to sadness, anger and surprise can be activated at the same time. This ambiguity of emotion expression makes the FER a challenging task.
The SAE is able to extract higher order features and detect relations between AUs, which is not possible by human experts or conventional machine learning techniques, therefore, we used a SAE deep network to extract the most effective combinations of AUs in each emotion and used them as the feature set to train our classifier. Figure **??** shows the overall scheme of our proposed model for emotion type detection task. Also the list of the applied AUs is shown in Table **??**. The next subsections explain principles of the SAE and the architecture and methodology of the developed deep SAE for emotion type detection.

3.1 Principals of the Stacked Auto Encoder

A SAE is a deep neural network consisting of several hidden layers in which the output of each layer is imposed as input to the next layer. By inner layers higher order features,

Table 1: The list of applied Action Units and related emotions. The pivotal AUs of each emotional state obtained by proposed model are indicated by sign $*$, and the pivotal AUs obtained by [**?**] are indicated by **+**.

No.	**AU Description**	**Happiness**	**Sadness**	**Fear**	**Anger**	**Surprise**	**Disgust**
1	AU01_Inner Brow Raiser		$*$	$*$**+**		**+**	
2	AU02_Outer Brow Raiser			$*$		$*$**+**	
3	AU04_Brow Lowered		**+**	**+**	$*$**+**		
4	AU05_Upper Lid Raiser					$*$	
5	AU06_Cheek Raiser	$*$		$*$			
6	AU07_Lid Tightener				**+**		
7	AU09_Nose Wrinkle						$*$**+**
8	AU10_Upper Lip Raiser						$*$**+**
9	AU12_Lip Corner Puller	$*$**+**					
10	AU15_Lip Corner Depressor		$*$**+**				
11	AU17_Chin Raiser		$*$				**+**
12	AU20_Lip Stretcher			$*$**+**			
13	AU23_Lip Tightener				$*$		
14	AU25_Lip Part	**+**		$*$**+**		**+**	
15	AU26_Jaw Drop			$*$		$*$**+**	

i.e., those are not easily possible for humans to craft, are obtained. Equation **??** gives the encoding step for k^{th} layer.

$$a^{k+1} = F(\omega^k a^k + b^k), \tag{1}$$

where F is the activation function, e.g., sigmoid or Rectified Linear Unites (ReLU), ω and b are corresponding weight vector and bias value to the units of k^{th} layer. The decoding step is given by running the decoding stack of each AE in reverse order as shown in Equation **??**.

$$a^{n+k+1} = F(\omega^{n-k} a^{n+k} + b^{n-k}), \tag{2}$$

where a^n contains the information of interest and is the activation of the deepest layer of hidden units. Applying the input values as output values, the SAE will learn the high-order, i.e., low dimension features of input values at the layer n. This vector gives a representation of the input in term of higher order features, which can be used for classification problems by feeding a^n to a Softmax classifier. After training the SAE, the encoder part of the network is saved and the activation values of the last layer are imposed to the classification layer, which uses a Softmax activation function to include more than two classes.

3.2 SAE Architecture for Emotion Type Detection

Table **??** shows the architecture of the proposed SAE for emotion detection. We used OpenFace to extract the AU values of training data. The activation values of 15 different AUs, in both regression and binary scale, are presented by a 30×1 vector and imposed

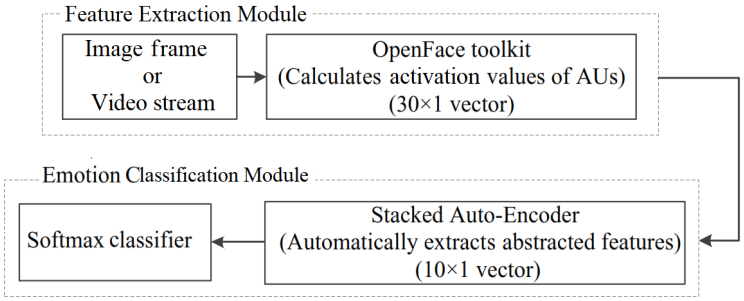

Fig. 2: OpenFace is able to read both images and videos and returns the activation value of different AUs. Passing AU values to SAE, abstracted features are obtained, which by feeding to a Softmax classifier the type of emotion, which AUs are showing is obtained.

Table 2: The architecture of the applied SAE neural network for six class classification task.

	Layer	Input size	Output size
Encoder	Dens layer (ReLU)	30×1	70×1
	Dens layer (ReLU)	70×1	70×1
	Dens layer (ReLU)	70×1	70×1
Decoder	Dens layer (ReLU)	70×1	70×1
	Dens layer (ReLU)	70×1	70×1
	Dens layer (sigmoid)	70×1	10×1
Fully Connected	Classifier (Softmax)	10×1	6×1

as an input to the developed SAE, where the regression values of AUs are normalized between 0 to 1. The extracted features from SAE are in the shape of a 10×1 vector, which are applied to the Softmax layer. Imposing the abstracted features vector into the Softmax layer, an original input data is classified into one of the six different basic emotion classes.

By applying 2D grid search, the hyper parameters of the SAE, e.g., learning rate and dropout are selected optimally in the learning process. The number of epochs and batch size are set as 200 and in the fully connected layer, “Adam” is used as optimizer and “Softmax” is used as supervised categorical classifier. Reconstructing the obtained features from SAE revealed the most pivotal AUs in each emotion as shown in Table **??**.

4 VERIFICATION AND RESULTS

To verify the accuracy of the proposed model we applied it to three well-known datasets and compared obtained results with the results from two state-of-the-art methods, which showed convincing performance on these datasets. Table **??** summarizes two baselines. One of the baseline methods used a convolutional neural network, while the other used a SVM method. Since the training and testing approach and the used datasets are different for baselines, we defined different experiments to be in align with compared method.

Table 3: State-of-the-art algorithms in FER over CK+ and MMI datasets.

Authors	Approach	Emotions	Feature extraction	Model	Datasets	Train and Test	Accuracy (%)
Hasani et al [?]	Sequenced	7,6	AAM	CNN, CRF	CK+, MMI	80% train,10% test	93.0, 78.6
Zhao et al[?]	Sequenced	6	LBP, Gabor multiorientation	SVM	CK+, MMI	Leave one out	93.9, 71.9

Table 4: Number of samples for each emotion class in two datasets.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total
CK+	45	59	25	69	28	83	309
MMI	32	28	28	42	32	41	203

However, as both baselines used confusion matrix to show the accuracy of their model, we also showed the accuracy of our model by confusion matrix. In following we first review the applied datasets, then the comparison between proposed model and baselines are discussed through different experiments. For easiness of read in next subsections happiness, sadness, fear, anger, disgust, and surprise are indicated by H, Sa, F, A, D, and Su respectively.

4.1 Data Bases

The extended Cohn-Kanade database (CK+)[**?**], contains 593 frontal face poses images of 123 subjects ranging from 18 to 50 years old. However, only 327 sequences from 118 subjects have labels.
MMI, contains 203 video sequences, including different head poses and subtle expressions of 19 participants with ages ranging from 19 to 62 years old [**?**].
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [**?**], contains frontal face poses videos from 12 female and 12 male, all north American actor and actress, expressing six basic emotions, calm, and neutral.
While the sequences of all datasets start with the neutral state frame and end at the apex of the target emotion [**?**] removed the beginning frames and [**?**] labeled them as neutral, thereby we also removed the beginning frames of both datasets. Table **??** shows the number of each expression class in CK+ and MMI datasets in our experiments.

4.2 Experiment A: Recognition rate on CK+ dataset for 6 emotion classes

The first experiment is conducted on the CK+ dataset. The best accuracy over CK+ for the six emotion recognition presented by Zhao et al [**?**], which obtained by leave-one-out cross validation strategy. Hence, we also applied leave-one-out method to verify the accuracy of the proposed model. Table **??** shows the comparison between confusion matrices of proposed model and results reported in [**?**] over CK+ dataset.
Proposed model outperforms the baseline for 3 classes out of 6 classes, i.e., anger, sadness, and surprise, while baseline has higher accuracy rate for detection disgust emotion. The model could detect all samples of happiness and surprise, i.e., 100% accuracy.

Table 5: Experiment A, confusion matrices for six emotion classification over the CK+ dataset, validated by leave-one-out technique.

(a) The result of the baseline [**?**].

	A	**D**	**F**	**H**	**Sa**	**Su**	**Acc (%)**
A	**42**	0	0	0	3	0	93.3
D	0	**59**	0	0	0	0	100
F	0	0	**22**	0	2	0	88
H	0	0	0	**69**	0	0	100
Sa	3	0	1	0	**24**	0	85.7
Su	0	0	2	1	0	**80**	96.4
Avg							**93.9**

(b) The result of the proposed model.

	A	**D**	**F**	**H**	**Sa**	**Su**	**Acc (%)**
A	**43**	0	0	0	2	0	**95.5**
D	1	**58**	0	0	0	0	98.3
F	0	1	**22**	0	0	1	88
H	0	0	0	**69**	0	0	100
Sa	1	0	0	0	**26**	1	**92**
Su	0	0	0	0	0	**83**	**100**
Avg							**95.63**

Table 6: Experiment A, six emotions classification over the MMI dataset.

(a) Accuracy comparison with 4 other baselines.

Research	**Method**	**Acc (%)**
Zhao et al [**?**]	SVM	71.92
Hasani et al[**?**]	CNN CRF	78.67
Proposed model	SAE	**95.58**

(b) Confusion matrix of the proposed model validated by 10 fold cross validation.

	A	**D**	**F**	**H**	**Sa**	**Su**
A	**93.75**	3.12	0.0	0.0	3.12	0.0
D	3.22	**96.77**	0.0	0.0	0.0	0.0
F	0.0	0.0	**90.00**	0.0	3.33	6.66
H	0.0	0.0	0.0	**100**	0.0	0.0
Sa	0.0	0.0	0.0	0.0	**100**	0.0
Su	0.0	0.0	6.97	0.0	0.0	**93.02**

The lowest accuracy is for recognising the fear class samples as 88% with misclassifying one sample as disgust and one sample as surprise. Overall the average accuracy of the proposed model is higher than baseline, i.e., 95.63% compared to 93.9%.

4.3 Experiment B: Recognition rate on MMI dataset for 6 emotion classes

The second experiment performed on MMI dataset for which [**?**] obtained the best performance over it. We applied 10-fold cross validation to report our results, because Hasani et al[**?**] used 5-fold cross validation and Zhao et al [**?**] used 10-fold cross validation for verification. Table **??**.a shows that the proposed model outperforms both baselines significantly, i.e., 95.6% compared to 78.67% and 71.92%.
Since confusion matrices of baselines are not provided in main references, we compared the overall obtained accuracy. However, the confusion matrix of the proposed model is shown in Table **??**.b. Analysing Table **??**.b, the best accuracy is obtained for happiness and sadness with the accuracy of 100% and the lowest accuracy obtained for fear class with accuracy of 90%.

Table 7: Experiment C, confusion matrices for six emotions over the RAVDESS dataset.

(a) Comparison of different baselines over RAVDESS dataset for six emotion classification.

	A	D	F	H	Sa	Su	Avg%
SAE	**86.9**	90.2	**83.9**	90.9	**82.6**	75.0	**84.91**
1NN	80.5	83.9	73.6	**91.0**	75.5	78.2	80.45
2NN	83.2	81.9	72.0	89.5	74.1	78.1	79.8
CART	83.6	73.9	72.4	80.2	63.1	74.5	74.61
MLP	69.7	**100**	51.6	79.8	57.5	**86.7**	74.21

(b) Confusion matrix of the proposed model over the RAVDESS dataset.

	A	D	F	H	Sa	Su
A	**86.97**	3.10	3.19	0.00	2.46	4.26
D	2.33	**90.29**	2.26	0.00	4.00	0.61
F	3.54	3.30	**83.94**	0.00	5.02	4.17
H	0.9	0.46	2.4	**90.91**	0.73	4.58
Sa	3.31	6.40	6.06	0.00	**82.67**	1.53
Su	10.70	3.07	7.86	0.00	3.33	**75.00**

4.4 Experiment C: Recognition rate on RAVDESS dataset for 6 emotion classes

For further validation, we tested the proposed model on RAVDESS dataset [**?**]. While RAVDESS contains both facial and speech data, it is mostly used for speech emotion recognition and, to our best knowledge, is not used for FER, hence to confirm our results, we used Weka [**?**] to obtain the accuracy of four well-known classical machine learning models including K-nearest neighbors, e.g., 1NN and 2NN, Multilayer perceptron (MLP) with learning rate of 0.3, and decision tree (M5P). The batch-size for all models set as 200.

We designed a subject-independent experiment, i.e., the dataset is partitioned into two subsets for train and validation such that 18 subjects (9 female and 9 male), i.e., 75% of the total dataset considered as training set and the other 6 subjects (3 female and 3 male), i.e., 25% of the total dataset considered as test set.

Table **??**.a shows the comparison between proposed model (SAE) with four other classical machine learning approaches. The proposed model outperforms baselines in three classes of anger, fear and sadness out of six classes. The best performance for happiness achieved by 1NN and for disgust and surprise by MLP. The overall average accuracy of the proposed SAE based model is 84.91% which outperforms all other baselines. The confusion matrix of the proposed model on the test dataset is shown in Table **??**.b. Also, the confusion matrices of provided baselines over RAVDESS are shown through Table **??**.

5 CONCLUSION

Since one facial expression might have an ambiguity or similarity to some other basic emotions, precise Facial Emotion Recognition is a challenging task. To find the best features for recognizing different emotions we used Stacked Auto Encoder, which is able to find high order features, which are not possible to craft by humans. The provided raw input data for SAE is the activation value of AUs, the final output, i.e., feature set, is the combination of most pivotal AUs for each basic emotion. The obtained feature set then is imposed to a Softmax classifier layer to find 6 basic emotions.

Table 8: Experiment C, confusion matrices for six emotion classification over the RAVDESS dataset for 4 different baselines.

(a) Confusion matrix of 1NN model.

	A	D	F	H	Sa	Su
A	**80.55**	4.33	3.71	0.00	4.15	7.24
D	4.49	**83.98**	2.66	0.00	5.18	3.67
F	4.78	4.38	**73.59**	0.00	6.55	10.68
H	1.00	0.00	1.23	**91.00**	1.72	4.15
Sa	5.84	6.44	3.60	0.00	**75.54**	8.55
Su	6.69	4.31	6.01	0.00	4.79	**78.17**

(b) Confusion matrix of 2NN model.

	A	D	F	H	Sa	Su
A	**83.20**	3.10	4.82	0.00	3.50	5.35
D	6.28	**81.92**	2.66	0.00	4.45	4.66
F	7.16	2.96	**71.99**	0.00	7.46	10.39
H	1.50	0.00	3.47	**89.55**	2.06	3.32
Sa	8.01	4.47	4.82	0.00	**74.07**	8.59
Su	7.09	3.49	7.27	0.00	4.02	**78.09**

(c) Confusion matrix of decision tree.

	A	D	F	H	Sa	Su
A	**83.63**	3.09	3.5	0.00	2.66	7.09
D	5.6	**73.90**	2.8	0.00	1.23	16.49
F	8.78	5.36	**72.45**	0.00	0.00	12.56
H	5.16	0.22	5.97	**80.22**	6.18	2.21
Sa	15.45	3.60	10.47	0.00	**63.12**	7.33
Su	13.50	1.5	8.01	0.00	2.37	**74.54**

(d) Confusion matrix of MLP.

	A	D	F	H	Sa	Su
A	**69.73**	5.99	0.0	0.0	0.0	24.28
D	0.0	**100**	0.0	0.0	0.0	0.0
F	37.25	2.63	**51.64**	0.0	0.0	8.47
H	5.71	0.26	7.33	**79.81**	6.34	5.27
Sa	9.29	2.71	28.18	0.0	**57.49**	2.32
Su	0.0	13.32	0.0	0.0	0.0	**86.68**

The proposed method is compared with several key methods and the experiments' results show that it is capable to outperform all rival methods. The proposed method achieves average accuracy of 95.63%, 95.55% and 84.91% for CK+, MMI and RAVDESS datasets respectively. Overall the best accuracy obtained for classifying happiness, while the worst result obtained for classifying fear.
In future work, we will apply the proposed method to classify more emotion classes. Also other features like head pose and gaze direction will be investigated to improve the accuracy of the proposed model. Furthermore, we will apply the proposed SAE to find the intensity of the detected emotion.

ACKNOWLEDGMENT

The work leading to these results has received funding from the European Commission 7th Framework Program as a part of the DREAM project, grant no. 611391 and the ICON project ROBO-CURE.