2017 IEEE/ACS 14th International Conference on Computer Systems and Applications

Facial Expression Recognition via Deep Learning

Abir Fathallah
ENISO, Sousse university
Sousse, Tunisia
Email: abir.fathallah1803@gmail.com

Lotfi Abdi
ENISO, Sousse university
Sousse, Tunisia
Email: lotfiabdi@hotmail.com

Ali Douik
ENISO, Sousse university
Sousse, Tunisia
Email: ali.douik@enim.rnu.tn

***Abstract*—Automated Facial Expression Recognition has remained a challenging and interesting problem in computer vision. The recognition of facial expressions is difficult problem for machine learning techniques, since people can vary significantly in the way they show their expressions. Deep learning is a new area of research within machine learning method which can classify images of human faces into emotion categories using Deep Neural Networks (DNN). Convolutional neural networks (CNN) have been widely used to overcome the difficulties in facial expression classification. In this paper, we present a new architecture network based on CNN for facial expressions recognition. We fine tuned our architecture with Visual Geometry Group model (VGG) to improve results. To evaluate our architecture we tested it with many largely public databases (CK+, MUG, and RAFD). Obtained results show that the CNN approach is very effective in image expression recognition on many public databases which achieve an improvements in facial expression analysis.**

***Index Terms*—Facial Expression; Recognition; Deep Learning; CNN; Architecture; Classification.**

I. Introduction

Interaction Human-Machine (IHM) has long confined researches to develop techniques based on the use of triplet screen-keyboard-mouse. Today, it is moving towards new paradigms: the user must be able to evolve unimpeded in its natural environment; fingers,hand, face or familiar objects are seen as many devices input/output; the boundary between the physical and electronic worlds is blurring. These new forms of interaction require usually the capture of the observable behavior of the user and his environment. That is why they rely on artificial perception techniques, including computer vision. IHM is a rapidly evolving discipline. Future generations of human-machine environment will become multimodal integrating new information, from the consideration of the dynamic behavior, speech and/or facial expressions, so as to make the use of machines the most intuitive and natural as possible.

The face is the most expressive and communicative part of a human being [1], it represents a major focus in current research concerning the improvement of IHM for establishing a dialogue between the two entities.

Facial expression is a visible manifestation of a face from the state of mind (emotion, reflection), cognitive activity, physiological (fatigue, pain), personality and psychopathology of a person. The essential of facial expression information is contained in the deformation of main permanent facial features, characterized by a change visually perceptible.

Today, analysis computer assisted of face and its facial expressions is an emerging field. Emotion recognition consists in associating an emotion to face image. So the goal is to determine from a face, the internal emotional state of a person. An automatic facial expression recognition system is an important component in human machine interaction. It consists to evaluate the possibility of emotions recognition. However, this is not an easy task.

Facial expression recognition usually employs a three-stage training consisting of face Acquisition [2], facial feature extraction [3] and classifier construction [4, 5].

Recently, Many works [6, 7] demonstrated that expression recognition can benefit from collecting two stages together facial feature extraction and classifier construction.

Deep learning methods have been successfully applied to extract features and classification, in particular Convolutional Neural Networks (CNN) architectures which are biologically-inspired multi-stage one that learned automatically hierarchies of invariant features [8]. The ConvNets consist of a multi-stage processing of an input image to extract hierarchical and high-level feature representations.

Motivated by this, we present in this paper an effective approach system based on ConvNets for facial expression recognition. We proposed a new architecture which the input of the system is an image; then, we use CNN to predict the facial expression label which should be one these labels[9]: anger,happiness, sadness, disgust, surprise and neutral.

II. Related work

Many facial expression recognition methods represent faces by high-dimensional over-complete face descriptors, followed by shallow models. In this context, MLIKI et al. [10] proposed a method which is able to monitor the intensity variation of facial expression and to reduce the classification confusion between facial expressions classes. The approach uses Vector Field Convolution (VFC) method to segment facial feature contours. COTRET et al. [11] also proposed a wavelet-based face recognition method robust against face position and light variations for real-time applications. Wang et al. [12] proposed expression recognition method based on evidence theory and local texture where the facial image is divided into regions

2161-5330/17 $31.00 © 2017 IEEE
DOI 10.1109/AICCSA.2017.124

with important recognition features, and the Local Binary Patterns (LBP) textural features of the regions are extracted.
Recent works showed how deep learning models could be applied on facial expression recognition, in [13] the authors present a novel Boosted Deep Belief Network (BDBN) for performing the three training stages iteratively in a unified loopy framework. Burkert et al. [14] proposed a CNN architecture for facial expression recognition. Mollahosseini [15] also proposed a deep neural network architecture to address the facial expression recognition problem across multiple well-known standard face datasets. In [16] Zhang et al. proposed a novel deep neural network: DNN-driven facial feature learning model which employs several layers to characterize the corresponding relationship between the SIFT feature vectors and their corresponding high-level semantic information. The authors in [17] extract from facial landmarks fixed number of SIFT features used as an input matrix to CNN architecture. The matrix size is X×Y where X is the number of SIFT features, and Y is the size of each feature. Moreover, in [18] Sun et al. proposed a mixture of SIFT and deep convolution, the authors used Partial least squares regression and linear SVM to train these features. Then, the output from all classifiers are combined with fusion network. In [19] a deep learning technique is adopted. The idea is to combine two models, in the first deep network temporal appearance features from image sequences are extracted. Meanwhile temporal geometry features from temporal facial landmark points are extracted by the second deep network. The authors in [20] proposed a novel technique of taking Inter Vector Angles (IVA) as geometric features, which proved to be scale invariant and person independent. A feature redundancy-reduced convolutional neural network (FRR-CNN) is presented in [21].

III. Proposed Approach

Our purpose is to ameliorate the accuracy of facial expression classification by using a new CNN architecture. As deep networks need a big database for the trainig, we combine many databases to get a final one.
As a first step, After preparing the database we fixed the batch size input of CNN architecture to 165 × 165 then we trained the architecture with fine tuning by Visual Geometry Group (VGG) model to generate the first model.
In second step to improve the classification we repeat the training of our CNN architecture but the fine tuning here is achieved with the obtained first model, and finally we get our final model as shown in Figure 1.

A. Network architecture

Deep Neural Networks (DNN) are models inspired of the human brain, and particularly its ability to extract structures (patterns) from raw data. From raw data input, deep learning models operate a large number of successive transformations to discover representations increasingly abstract of such data. The operated transformations are combinations of linear and nonlinear operations. These transformations are used to represent the data at different levels abstraction. The most popular

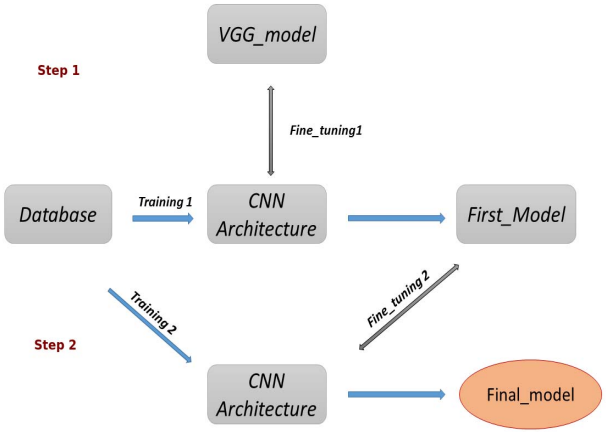

Fig. 1. Proposed approach.

image processing structure of DNN is CNN [8] which is constructed by three main processing layers: Convolutional Layer, Pooling Layer and Fully Connected Layer.
The adopted architecture of the convolutional neural network is given in Figure 2.
The proposed deep ConvNet contain four convolutional layers (with three max-pooling layers) to extract features hierarchically, followed by the fully-connected layer and the softmax output layer indicating 6 expression classes. The input of the network is 165×165 × k for all patches, where k = 3 for color patches and k = 1 for gray patches. The output is one of the 6 expression classes.

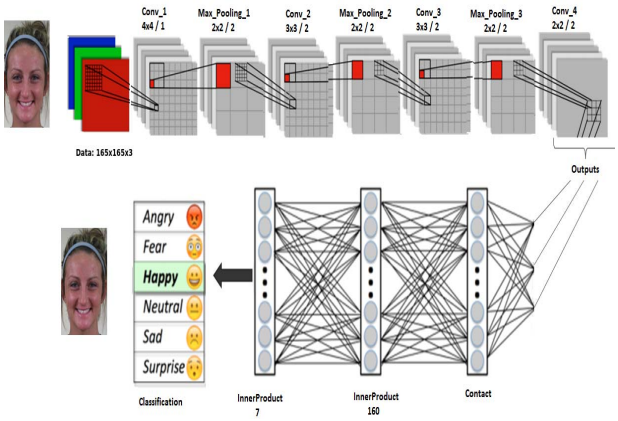

Fig. 2. Architecture of the convolutional neural network.

CNN units are described below:
Convolutional layer: A convolution layer C^i (i network layer) is parameterized by its number N of convolution cards $M_j^i (j \in \{1, ..., N\})$, the size of the convolution kernels $K_x \times K_y$ (often square), and the connection diagram in the previous layer L^{i-1}. Each convolution card M_j^i is the result of a convolution sum of cards previous layer M_j^{i-1} by its respective convolution kernel. In the case of a fully connected

TABLE I
NETWORK CONFIGURATION.

Layer type	**Size/Stride**	**Output** Dropout Probability
Data	165×165	-
Convolution_1	4×4 / 1	20
Max_Pooling_1	2×2 / 2	-
Convolution_2	3×3 / 2	40
Max_Pooling_2	2×2 / 2	-
Convolution_3	3×3 / 2	60
Max_Pooling_3	2×2 / 2	-
Convolution_4	2×2 / 2	80
Fully Connected	-	160 Dropout=0.5
Fully Connected	-	7

card to the cards of the previous layer, the result is calculated by the equation 1.

$$\sum_{n=1}^{N} M_j^i = \phi\left(b_j^i + \sum_{n=1}^{N} M_n^{i-1} * k_n^i\right) \quad (1)$$

where * is the convolution operator.

Pooling layer: In the classical architectures of convolutional neural networks, convolution layers are followed by sub-sampling layers. A sub-sampling layer reduces the size of cards, and introduces invariance to (low) rotations and translations can appear as input. The output of max-pooling layer is given by the maximum activation value, in the input layer for different regions of size $K_x \times K_y$ non-overlapping. Similarly to a convolution layer, a bias is added and the result is passed to the transfer function ϕ defined above.

Fully connected layer: After several max pooling and convolutional layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset. Table I present network configuration, the patch size od data input is 165×165. Then, the convolutional and Max_Pooling layers are chosen with different kernel sizes (4×4, 3×3, 2×2) and different strides (1, 2).

B. Fine tuning

In the step of Fine tuning, we chose to use VGG model[22]. For fine-tuning with VGG, we used Caffe [23] framework to get the model weights. VGG model was trained on the large CASIA WebFace dataset [24] and the Static Facial Expressions in the Wild (SFEW) dataset, which is a smaller database of labeled facial emotions and it has been developed by selecting frames from Acted Facial Expressions In The Wild (AFEW: is a dynamic temporal facial expressions data corpus consisting of close to real world environment extracted from movies)[25].

IV. EXPERIMENTAL RESULTS

Several databases were used to train and evaluate the proposed architecture.

TABLE II
DATABASES DESCRIPTION.

Base		Images Number	Pose
Learning base	Ck+	8000	1
	KDEF	4900	5
	MUG	21000	1
	RAFD	1400	1
Test base	Ck+	150	1
	RaFD	120	1
	MUG	3006	1

A. Databases

To train our CNN architecture, we used many databases and we standardize the size of all images to 224 × 242 pixels.

The CK+ database [26] The CK+ database presented by 327 expression sequences. From each image sequence, we selected only the last four frames to save the most clear expression. Meanwhile the first frames from each sequence are collected for neutral expression.

We use also KDEF database is a set of totally 4900 pictures of human facial expressions of emotion. This database proposed 7 facial expressions in different positions. We choose to use the images of this database in the training step but only with 3 positions.

The Radboud Faces Database (RaFD) [27] is a set of pictures of 67 models (including Caucasian males and females, Caucasian children, both boys and girls, and Moroccan Dutch males) displaying 8 emotional expressions.

MUG database [28] consists of image sequences of 86 subjects performing facial expressions. In the database participated 35 women and 51 men all of Caucasian origin between 20 and 35 years of age. Each image was saved with a jpg format, 896896 pixels.

By this way, we built an experimental dataset with a total of 37000 images. 330000 images are used in training step, and the rest of images are used as test images. As shown in Table II, for learnig base we uses Ck+, KDEF, RaFD and MUG datasets while for test base we tested only with Ck+, RaFD and MUG datasets.

1) Performances of the proposed method: To verify the effectiveness of our proposed approach, we opted for a validation on standard databases MUG, Rafd and CK+. Figure 3 illustrate the confusion matrix on CK+ database. We can observe that our proposed method achieves the best perfermance on only 5 database classes (Disgust, Happy,Neutral, Sad and Surprise) with recognition rate 100%. However recognition rate of Angry emotion is still difficult (96%) because of the database characteristics. Our model excelled with classifying the CK+ dataset with recognition rate of 99.33%. Figure 4 shows the confusion matrix on MUG database we achieved only 87.65% as recognition rate, the differences between different emotions among the subjects is very subtle. Also, MUG images are in

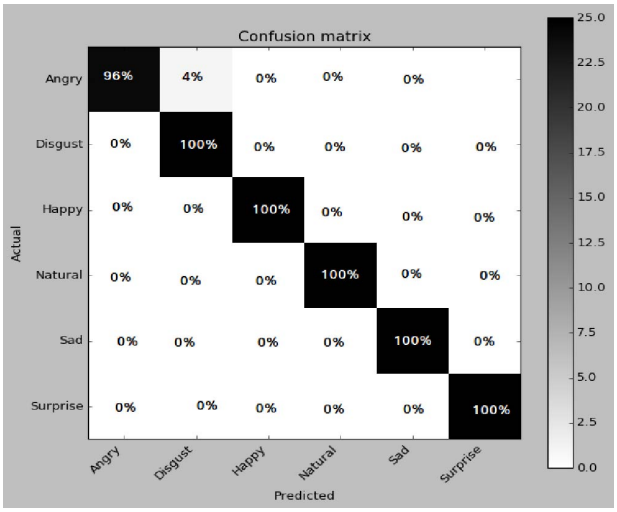

Fig. 3. Confusion matrix of proposed method on CK+ database.

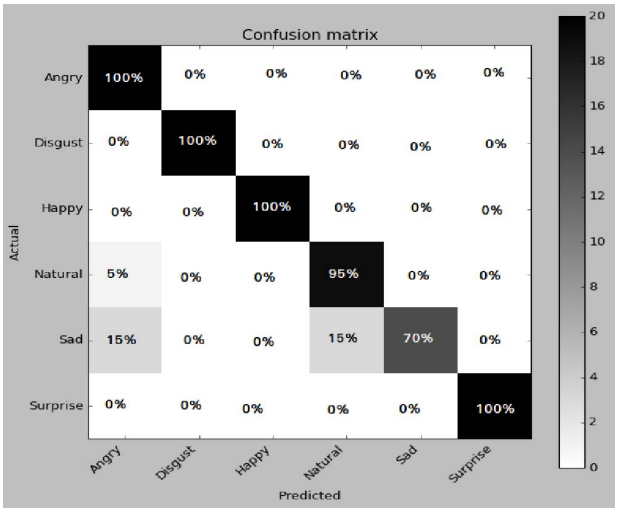

Fig. 5. Confusion matrix of proposed method on RaFD database.

grey-scale format, whereas our model work is optimized for RGB images. The lack of RGB format can exacerbate the ability of the network to make distinction between important features and background elements. Our model is also evaluated with RaFD and we achieved a accuracy rate of 93.33% as shown in Figure5. In this database we observe some incorrect classification for Sad and Neutral emotions and specially the Sad expression due to the unclarity features in this class.

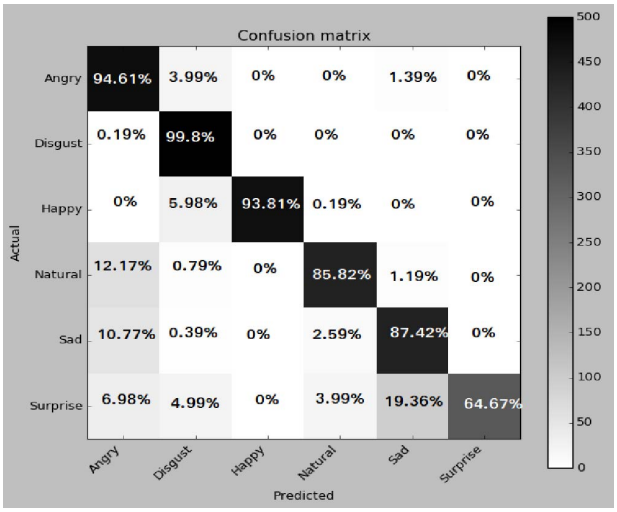

Fig. 4. Confusion matrix of proposed method on MUG database.

B. Method comparison

To get accuracy close to 99% in training step, it took about 4 days to finish the training (300000 iterations). To accelerate the learning process with parallelization, we used CAFFE on ubuntu 14.04 LTS and a GeForce GT 525M GPU, which has 2GB of memory. we tried an architecture with 3, 5 and 6 hidden layers but our chosen architecture present the best accuracy of features extraction and emotions classification. To demonstrate the perfermance of our method we compared it with other methods tested on Ck+ database. As shown in Figure 6, our method outperformed all the methods in comparison. This demonstrated that the features learned and selected through our method contain more discriminative information for facial expression recognition.

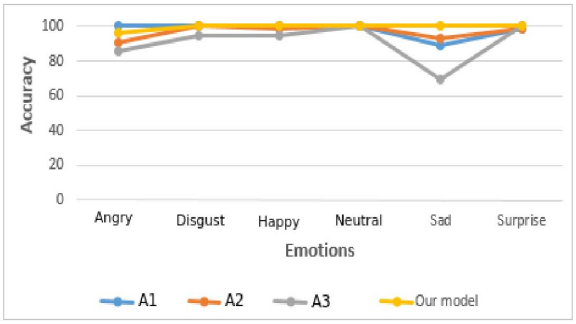

Fig. 6. Performance comparison on the CK+ database, A1[20] , A2 [19], A3 [21].

Our Deep learning method achieves the best performance on CK+ database. The deep learning work developed by [19] achieved the second best performance of 96.93% accuracy on Ck+. The third best performance of 95.34% accuracy is presented by [20]. The work proposed at the same time of this paper achieved the last best performance of 71.04% accuracy on Ck+.

V. Conclusion and Discussion

A part of automatic analysis of facial expression field is presented in this paper. We proposed a new deep neural network architecture for facial expression recognition. The proposed network consists of four convolutional layers, th first three layers are followed by max pooling and the last one

is followed by fully connected layer. It takes facial images as the input and classifies them into either of the six facial expressions: angry, disgust, happy, neutral, sad and surprise. The proposed architecture obtained is evaluated with MUG, RAFD and Ck+ databases. Results and recognition rates prove that our method outperforms the state-of-the-art methods.
For this project, we trained the model with images which the face was in one position. In the futurework, we would like to extend our model to different face positions. This will allow us to investigate the efficacy of pre-trained models such as VGGNet for facial emotion recognition.

REFERENCES

[1] R. G. Harper, A. N. Wiens, and J. D. Matarazzo, *Nonverbal communication: The state of the art.* John Wiley & Sons, 1978.

[2] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[3] S. M. Lajevardi and M. Lech, "Facial expression recognition from image sequences using optimized feature selection," in *Image and Vision Computing New Zealand, 2008. IVCNZ 2008. 23rd International Conference*. IEEE, 2008, pp. 1–6.

[4] A. Ben-Hur and J. Weston, "A users guide to support vector machines," *Data mining techniques for the life sciences*, pp. 223–239, 2010.

[5] R. Samad and H. Sawada, "Extraction of the minimum number of gabor wavelet parameters for the recognition of natural facial expressions," *Artificial Life and Robotics*, vol. 16, no. 1, pp. 21–31, 2011.

[6] J. Susskind, V. Mnih, G. Hinton *et al.*, "On deep generative models with applications to recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2857–2864.

[7] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," *Computer Vision–ECCV 2012*, pp. 808–822, 2012.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *California Mental Health Research Digest*, vol. 8, no. 4, pp. 151–158, 1970.

[10] H. Mliki, N. Fourati, M. Hammami, and H. Ben-Abdallah, "Data mining-based facial expressions recognition system." in *SCAI*, 2013, pp. 185–194.

[11] P. Cotret, S. Chevobbe, and M. Darouich, "Embedded wavelet-based face recognition under variable position," in *SPIE/IS&T Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 94 000A–94 000A.

[12] W. Wang, F. Chang, Y. Liu, and X. Wu, "Expression recognition method based on evidence theory and local texture," *Multimedia Tools and Applications*, pp. 1–15, 2016.

[13] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[14] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

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

[16] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[17] ——, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[18] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, vol. 25, no. 6, pp. 061 407–061 407, 2016.

[19] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2983–2991.

[20] R. Islam, K. Ahuja, S. Karmakar, and F. Barbhuiya, "Sention: A framework for sensing facial expressions," *arXiv preprint arXiv:1608.04489*, 2016.

[21] S. Xie and H. Hu, "Facial expression recognition with frr-cnn," *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.

[22] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[23] (2016, september) Caffe: Deep learning framework by the bvlc. http://caffe.berkeleyvision.org/.

[24] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[25] A. Dhall *et al.*, "Collecting large, richly annotated facial-expression databases from movies," 2012.

[26] 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

[27] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus,

S. T. Hawk, and A. van Knippenberg, “Presentation and validation of the radboud faces database,” *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[28] N. Aifanti, C. Papachristou, and A. Delopoulos, “The mug facial expression database,” in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*. IEEE, 2010, pp. 1–4.