

# I Know How You Feel: Emotion Recognition with Facial Landmarks

Ivona Tautkute<sup>1,2</sup>, Tomasz Trzcinski<sup>1,3</sup> and Adam Bielski<sup>1</sup>

<sup>1</sup>Tooploox <sup>2</sup>Polish-Japanese Academy of Information Technology <sup>3</sup>Warsaw University of Technology

Ffirstname.Lastname@toopl oox.com

## Abstract

Classification of human emotions remains an important and challenging task for many computer vision algorithms, especially in the era of humanoid robots which coexist with humans in their everyday life. Currently proposed methods for emotion recognition solve this task using multi-layered convolutional networks that do not explicitly infer any facial features in the classification phase. In this work, we postulate a fundamentally different approach to solve emotion recognition task that relies on incorporating facial landmarks as a part of the classification loss function. To that end, we extend a recently proposed Deep Alignment Network (DAN), that achieves state-of-the-art results in the recent facial landmark recognition challenge, with a term related to facial features. Thanks to this simple modification, our model called EmotionalDAN is able to outperform state-of-the-art emotion classification methods on two challenging benchmark dataset by up to 5%.

## 1. Introduction

Since autonomous AI systems, such as anthropomorphic robots, start to rapidly enter our lives, their ability to understand social and emotional context of many everyday situations becomes increasingly important. One key element that allows the machines to infer this context is their ability to correctly identify human emotions, such as happiness or sorrow. This is a highly challenging task, as people express their emotions in a multitude ways, depending on their personal characteristics, e.g. people with an introvert character tend to be more secretive about their emotions, while extroverts show them more openly. Although some simplifications can be applied, for instance reducing the space of recognized emotions, there is an intrinsic difficulty embedded in the problem of human emotion classification.

Most of the currently available methods that address this problem use some variation of a deep neural network with convolutional layers. For instance [1] proposes to use a standard architecture of a convolutional neural network (CNN) with two convolutional, two subsampling and one fully connected layer. Before being processed, the image

is spatially normalized with a pre-processing step. Another method presented in [2] incorporates additional inception layers into the architecture, inspired by the Inception model [3] that achieves state-of-the-art object classification results on the ImageNet dataset [4]. Another variation of the Inception [3], uses the Inception-V3 model pretrained on the ImageNet dataset with a custom softmax layer trained specifically to classify emotions. Finally, the most recent method called EmotionNet [5] and its extension EmotionNet2 [6] builds up on the ultra-deep ResNet architecture [7] and improves the accuracy by using face detection algorithm that reduces the variance caused by a background noise.

Although all the above methods rely on the state-of-the-art deep learning architectures, they draw their inspiration mostly from the analogical models that are successfully used for object classification tasks. We believe that as a result these approaches do not exploit intrinsic characteristics of how humans express emotions, i.e. by modifying their face expression through moving the landmark features of their faces. We therefore propose to use a state-of-the-art facial landmark detection model – Deep Alignment Network (DAN) [8] – and extend it by adding a surrogate term that aims to correctly classify emotions to the neural network loss function. This simple modification allows our method, dubbed EmotionalDAN, to exploit the location of facial landmarks and incorporate this information into the classification process. By training both terms jointly, we obtain state-of-the-art results on two challenging datasets for facial emotion recognition: CK+ [9] and ISED [10].

## 2. EmotionalDAN

Our approach builds up on the Deep Alignment Network architecture [8], proposed initially for robust face alignment. The main advantage of DAN over the competing face alignment methods comes from an iterative process of adjusting the locations of facial landmarks. The iterations are incorporated into the neural network architecture, as the information about the landmark locations detected in the previous stage (layer) are transferred to the next stages through the use of facial landmark heatmaps. As

a result and contrary to the competing methods, DAN can therefore handle entire face images and not patches which leads to a significant reduction in head pose variance and improves its performance on a landmark recognition task. DAN ranked 3<sup>rd</sup> in a recent face landmark recognition challenge Menpo [11].

In this work, we hypothesize that DAN’s ability to handle images with large variation and provide robust information about facial landmarks transfers well to the task of emotion recognition. To that end, we extend the network learning task with an additional goal of estimating expressed facial emotions. We incarnate this idea by modifying the loss function with a surrogate term that addresses specifically emotion recognition task and we minimize both landmark location and emotion recognition terms jointly. The resulting loss function  $L$  can be therefore expressed as:

$$L = \alpha \cdot \frac{S - S'}{d} + \beta \cdot CE(E, E'),$$

where  $S$  is the transformed output of predicted facial landmarks using Landmark Transform and Image transform layers [8],  $E$  is the softmax output for emotion prediction.  $S'$  is the vector of ground truth landmark locations,  $d$  is the distance between the pupils of ground truth that serves as a normalization scalar and  $E'$  is the ground truth for emotion labels.  $CE$  denotes Cross-Entropy loss. We weigh the influence of the terms with  $\alpha$  and  $\beta$  coefficients and after an initial set of experiments we fix their values to  $\alpha = 0.4$  and  $\beta = 0.6$ .

### 3. Experiments

To evaluate the performance of the proposed EmotionalDAN method, we compute classification accuracy for the emotion recognition task, using several benchmark datasets, described in the next section. As our baselines, we use methods described in the introduction of this work, namely Convolutional Neural Network (CNN) [1] with 2 and 5 convolutional layers, Inception-V3 [2] and EmotionNet 2 [6].

When available, we use original implementations of the competing methods. For methods that are not made public, we implement them in Keras. Our EmotionalDAN is based on Tensorflow implementation of DAN [8].

#### 3.1. Datasets

To train all the evaluated methods we use AffectNet [12] - the largest available database for facial expression that contains over 1,000,000 face images collected from the Internet through emotion keyword querying. About half of the retrieved images were manually annotated for the presence of seven main facial expressions and 68 facial landmarks locations. This part of the dataset is used to train our methods.

For testing, we use CK+ [9], JAFFE [13] and [10] datasets with face images of over 180 individuals of different genders and ethnic background.

To provide a more holistic comparison of the methods, we split the emotions annotated in the test datasets into two scales: seven-grade scale with happy, sad, angry, surprised, disgust, fear and neutral emotions, and a three-grade scale with positive, negative and neutral emotions. This way we obtain two complimentary evaluation sets with various amount of bias introduced by confusion of labelers and other confounding factors.

#### 3.2. Results

Tables 1 and 2 show the results of the evaluation of our EmotionalDAN method and the competing approaches. Although the accuracy varies between the tested datasets, our approach outperforms the competitors by a large factor of up to 5% on two out of three benchmark datasets, namely on CK+ and ISED. The performance of our method is inferior to convolutional neural networks on the JAFFE dataset, although the accuracy values obtained on this dataset are generally lower than the competitors. We believe that this may be the result of a more challenging image acquisition conditions. Furthermore, our results show that convolutional neural networks achieve competitive results when compared with other methods despite their simplistic architecture.

### 4. Application

We implement our emotion recognition model as a part of the in-car analytics system to be deployed in autonomous cars. Figure 1 shows the results obtained by the camera installed inside a car. As autonomous car operation can potentially be influenced by emotions of the passengers (e.g. fear of speed expressed on passenger’s face could signal the need for speed reduction), this is an excellent playground for our method to show its full potential. Although alternative applications are possible, we believe that this use case showcases the capabilities of our method and can serve as an interesting input to the driving system, typically focused on the exterior views from outside the car.

### 5. Conclusion

In this paper, we overview a work-in-progress method for emotion recognition that allows to exploit facial landmarks. Although the results computed on the JAFFE dataset show that there is still place for improvement, we believe that this approach has a strong potential to outperform currently proposed methods. In future work, we will therefore focus on improving our method by using attention mechanism on facial landmarks and experiment with additional loss function terms. We also plan to investigate other applications of our method, e.g. in the context of autistic children with incapacities related to emotion recognition.

	CK +	JAFFE	ISED
CNN (2)	0.628	0.484	0.516
CNN (5)	0.728	<b>0.502</b>	0.593
Inception-V3	0.304	0.268	0.479
EmotionNet 2	0.204	0.249	0.21
EmotionalDAN	<b>0.736</b>	0.465	<b>0.62</b>

Table 1. Cross-database accuracy results compared for different model architectures and seven emotion categories. All models are trained on AffectNet database. Face detection is applied as a pre-processing step on all test sets for all methods.

	CK +	JAFFE	ISED
CNN (2)	0.819	0.525	0.814
CNN (5)	0.92	<b>0.765</b>	0.867
Inception-V3	0.582	0.536	0.673
EmotionNet 2	0.478	0.497	0.587
EmotionalDAN	<b>0.921</b>	0.634	<b>0.896</b>

Table 2. Cross-database accuracy results compared for different model architectures and three emotion categories - positive, negative and neutral.

Figure 1. Our emotion recognition model in passenger detection system for autonomous cars. Emotion recognition is performed on detected facial regions

## References

- [1] A. T. Lopes, E. de Aguiar, and T. Oliveira-Santos, "A facial expression recognition system using convolutional networks," In SIBGRAPI, 2015-. **1, 2**
- [2] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," In IEEE Winter Conference on Applications of Computer Vision (WACV), 2016. **1, 2**
- [3] X.-L. Xia, C. Xu, and B. Nan, "Facial expression recognition based on tensorflow platform," In ITM Web of Conferences, 2017. **1**
- [4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in CVPR09, 2009. **1**
- [5] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," In CVPR, 2016. **1**
- [6] B.Kennedy and A. Balint, "Emotionnet2." <https://github.com/co60ca/EmotionNet>. **1, 2**
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015. **1**
- [8] M.Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," In CVPRW, 2017. **1, 2**
- [9] 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 CVPRW, 2010. **1, 2**
- [10] S. L. Happy, P. Patnaik, A. Routray, and R. Guha, "The indian spontaneous expression database for emotion recognition," IEEE Transactions on Affective Computing, 2017. **1, 2**
- [11] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017. **2**
- [12] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, 2017. **2**
- [13] Lyons, Akamatsu, Kamachi, and Gyoba, "The japanese female facial expressions database." <http://www.kasrl.org/jaffe.html>. **2**