Classifying and Visualizing Emotions with Emotional DAN

Ivona Tautkutė

Polish-Japanese Academy of Information Technology

Tooploox

Warsaw, Poland

s16352@pjwstk.edu.pl

Tomasz Trzciński

Warsaw University of Technology

Tooploox

Warsaw, Poland

t.trzcinski@ii.pw.edu.pl

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) 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%. Furthermore, we visualize image regions analyzed by the network when making a decision and the results indicate that our EmotionalDAN model is able to correctly identify facial landmarks responsible for expressing the emotions.

Keywords: machine learning, emotion recognition, facial expression recognition
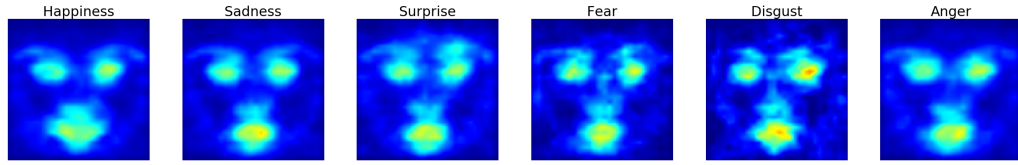
Figure 1.   Mean visualisations of Conv 4a layer activations in EmotionalDAN network for emotion classification. Even though during training model has only information about emotion label, spatial information about regions containing information relevant to expressed emotions are correctly captured. Face regions around eyes and mouth influence model's decision the most for all emotion classes. Eyebrow location and strength of activations near nose (e.g Disgust) also show some discriminative information.

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 of 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 or directly applying Facial Action Coding System (FACS) [1], there is an intrinsic difficulty embedded in the problem of human emotion classification.

While many Facial Expression Recognition (FER) systems already exist [2, 3, 4, 5, 6, 7, 8], the problem is far from being solved, in particular for expressions that are easily confused when judged without context (e.g *fear* and *surprise*). Considering facial expression recognition and face alignment are coherently related to each other, they should be beneficial for each other if putting them in a joint framework, e.g facial expression recognition has served as an auxiliary task to enhance landmark localication [9]. However, in literature it is rare to see such joint study of the two tasks. We therefore propose to use a state-of-the-art facial landmark detection model – Deep Alignment Network (DAN) [10] – 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+ [11] and ISED [12].

Furthermore, we perform an additional study that aims to interpret decisions made by model. We visualize with gradient-weighted class activation mapping (Grad-CAM [13]) image regions responsible for predicting the concept. Results show that our proposed architecture correctly focuses at the most important for emotion classification regions of the face.

The remainder of this work is organized in the following manner. In Sec. 2 we discuss related work in facial expression recognition. In Sec. 3 we present our approach and introduce EmotionalDAN

model. In Sec. 4 we present the datasets used for evaluation, explain in detail how our experiments are performed and present the results compared against baselines. Sec. 5 presents a set of experiments explaining visually the behaviour of our EmotionalDAN model with respect to the facial landmarks. Sec. 6 illustrates real life application of our proposed model. Finally, Sec. 7 concludes the paper.

2.   Related work

It is a common standard to taxonomize human facial movements with Facial Action Coding System published by Ekman and Friesen [1] in 1978 that describes facial expressions by action units (AUs) based on the anatomy of human face. Out of 30 AUs describing independent movements of the face muscles 12 are related to muscle contractions of upper face and 18 of the lower face. FACS system describes all visible facial muscle movements, and not just those presumed to be related to emotion or any other human state. Although FACS system has been widely used by behavioral scientists and allows for explicit definition of facial expression, it is a tedious task to code by hand.

There has been some work on automatic action units of facial movements system detection. Some methods work on local face patches and explore co-ocurrence of AUs in multilabel setting [14] or combine local models by training a different classifier per face region [15]. More recently, attention module has been used for AU detection in weakly superwised manner [16, 17]. Recognizing action units can directly help analyze facial expression [18]. However even though predicting emotion by detecting presence of AUs provides full transparency of the model, such methods are strictly limited to AU definitions. It has been observed that humans tend to express emotions in a wide variety of facial muscles independent of AUs.

Most of the recently proposed methods for automatic facial expression recognition are Deep Learning based methods and have proven to be more successful at emotion prediction than handcrafted features [19, 4, 8]. They commonly use some variation of a deep neural network with convolutional layers. With their broad spectrum of applications to various computer vision tasks, convolutional neural networks (CNN) have also been successful at recognizing emotions. For instance [19] propose to use a standard architecture of a CNN with two convolutional, two subsamping and one fully connected layer. Before being processed, the image is spatially normalized with a pre-processing step. Their model achieves state-of-the-art accuracy on CK+ [11] database of 97.81%. Some modified versions of this approach also include different numbers of layers (e.g five convolutional layers).

A number of methods is inspired by Inception model [20] that achieves state-of-the-art object classification results on the ImageNet dataset [21]. Inception layers provide an approximation of sparse networks hence are often applied to emotion recognition problem [7]. Ranging from simple transfer learning approaches where Inception-V3 model pretrained on ImageNet [21] is used with custom softmax classification layer [2] to custom architectures with Inception layers [6]. In another example [3] propose a deep neural network architecture consisting of two convolutional layers each followed by max pooling and then four Inception layers.

Another method called EmotionNet [4] and its extension EmotionNet2 [5] builds up on the ultra-deep ResNet architecture [22] 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. Moreover, vast majority of published methods is evaluated within the same database that the model was trained for with no cross-database comparison. While such accuracy results might be impressive they often lack the ability to generalize to different shooting conditions (lightning, angles, image quality) or subjects of different ethnic backgrounds.

To combine best of the two worlds, we propose a method that is not restricted by limitations of FACS system but by building up on facial landmarks concept can provide insights on how the decisions are made.

3. EmotionalDAN

Our approach builds up on the Deep Alignment Network architecture [10], initially proposed 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 instead of patches which leads to a significant reduction in head pose variance and improves its performance on a landmark recognition task. DAN ranked 3rd in a recent face landmark recognition challenge Menpo [23].
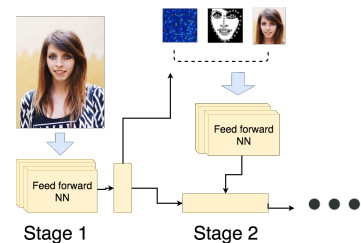

Figure 2. Information about landmark location estimates, landmark heatmaps and mean image are transferred across stages. Landmark location estimates are refined in the proceeding stage.

Originally, DAN was inspired by the Cascade Shape Regression framework and similarly it starts with initial estimate of face shape which is refined after following iterations. In DAN, each iteration is represented with a single stage of deep neural network. During each stage (iteration) features are extracted from entire image instead of local images patches (in contrast to CSR).

Training is composed of consecutive stages where single stage consists of feed-forward neural network and connection layers generating input for next stage. Each stage takes three types of inputs: input image aligned with the canonical shape, features image generated from dense layer of the previous stage and landmarks heatmap. Therefore output at each DAN stage is defined as:

$$S_t = T_t^{-1}(S_{t-1}) + \Delta S_t, \tag{1}$$

where ΔS_t is the landmarks output at stage t and T_t is the transform that is used to warp the input image to canonical pose.

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 $\mathcal{L}$ can be therefore expressed as:

$$\mathcal{L} = \alpha \cdot \frac{\parallel S_t - S^* \parallel}{d} - \beta \cdot E^* \cdot log(E_t),\qquad(2)$$

where S_t is the transformed output of predicted facial landmarks at stage t, 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. We weigh the influence of the terms with α and β coefficients.

We present the final version of our network in the table 1. It was originally inspired by network used in ImageNet ILSVRC competition (2014) [24] and contains four convolutional layer pairs followed by pooling layers. Top layers of the network consist of one common fully connected layer and two separate fully connected layers for landmark and emotion features.

4.   Experiments

In this section we perform quantitative evaluation of our model against published baselines as well as present an overview of datasets used for training and testing.

4.1.   Datasets

We include datasets that are made available to the public (upon request) and present a high variety of subjects' ethnicity. All compared models are trained on AffectNet [25] and evaluated cross-database on remaining test sets.

AffectNet [25] is by far the largest available database for facial expression. It contains more than 1,000,000 facial images from the Internet collected by querying major search engines with emotion related keywords. About half of the retrieved images were manually annotated for the presence of seven main facial expressions. 7 000 images from AffectNet database are set aside for validation and test sets.

CK+ [11] includes both posed and non-posed (spontaneous) expressions. 123 subjects are photographed in 6 prototypic emotions. For our analysis we only include images with validated emotion labels.

JAFFE [26] The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects.

ISED [12] Indian Spontaneous Expression Database. Near frontal face video was recorded for 50 participants while watching emotional video clips. The novel experiment design induced spontaneous emotions among the participants and simultaneously gathered their self ratings of experienced emotion. For evaluation individual frames from recorded videos are used.

Table 1.   Structure of the feed-forward part of EmotionalDAN network stage with multiple outputs. Dimensions of the last fully connected layer before emotion classification depend on the number of emotion classes used in training.

Name	Input shape	Output shape	Kernel
conv1a	224×224×1	224×224×64	3×3,1,1
conv1b	224×224×64	224×224×64	3×3,64,1
pool1	224×224×64	112×112×64	2×2,1,2
conv2a	112×112×64	112×112×128	3×3,64,1
conv2b	112×112×128	112×112×128	3×3,128,1
pool2	112×112×128	56×56×128	2×2,1,2
conv3a	56×56×128	56×56×256	3×3,128,1
conv3b	56×56×256	56×56×256	3×3,256,1
pool3	56×56×256	28×28×256	2×2,1,2
conv4a	28×28×256	28×28×512	3×3,256,1
conv4b	28×28×512	28×28×512	3×3,512,1
pool4	28×28×512	14×14×512	2×2,1,2
fc1	14×14×512	1×1×256	-
fc2_landmark	1×1×256	1×1×136	-
fc2_emotion	1×1×256	1×1×{3,7}	-

4.2. Datasets preparation and training

To allow for fair comparison we follow an unified approach for all datasets and methods.

While some datasets come with ground-truth information about bounding boxes of present faces (AffectNet), most test sets do not contain such information. Face regions often account only for small part of the image with a lot of unnecessary background (ISED). To address this issue, we extracted regions of interest with face detection algorithm Multi-task CNN [27]. For $< 2\%$ of test images where algorithm failed to recognize a face, full images were used. All test images were resized to 224×224, converted to black and white, and normalized by mean and standard deviation of the training set.

Training procedure is independent for three and seven emotion classes. For simplified emotions we perform a mapping of original ground truth labels where *fear, sadness, disgust* and *anger* are mapped to *negative* emotion, *happiness* and *contempt* to *positive* emotion and *neutral* class is kept without change. In this case we do not include images labeled with *surprise* as this emotion might have both positive and negative connotations.

Similarly to [10], training of EmotionalDAN is performed sequentially - first stage is trained until validation error stops improving. Afterwards, second stage is added and trained. After an initial set of experiments we set α and β coefficients to $\alpha = 0.4$ and $\beta = 0.6$. To prevent overfitting we add dropout layers with $p = 0.5$ for all stages after pooling layers. To improve training procedure, we

Figure 3. Visualisation of single stage of EmotionalDAN. There are two independent fully connected layers for the task of facial landmarks localization and emotion prediction. Size of the second one depends on number of emotion classes in the model.

use cyclical learning rate with triangular policy where learning rate varies between 0.0001 (*base_lr*) and 0.05 (*max_lr*). Code with EmotionalDAN acrhitecture and training details is publicly available in GitHub repository[1].

4.3. Results

Tables 2 and 3 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
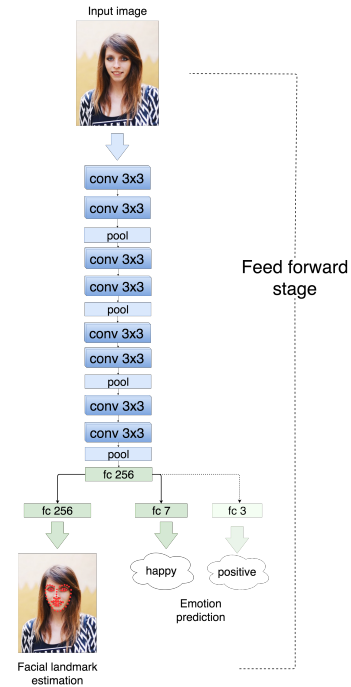
---

[1]https://github.com/IvonaTau/emotionaldan

Table 2. 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 preprocessing step on all test sets for all methods.

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

Table 3. Cross-database accuracy results compared for different model architectures and three emotion categories - positive, negative and neutral.

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

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.

Qualitative results are presented in Fig.4 and show examples of correct and incorrect predictions for each testset.

5. Visual Explanations

In this section we present visual explanations for emotion classification with EmotionalDAN.

5.1. Grad-CAM

To gain more insights from our model, we produce visual explanations for classification decisions using a popular gradient-based localization technique Grad-CAM [13]. This approach produces a coarse localization map of gradients flowing into the final convolution layer of arbitrary CNN architecture. Neurons in upper convolutional layers look for class-specific semantic information in the image, information that is lost in proceeding fully connected layers.

Specifically, class-discriminative localization map Grad-CAM $L_{Grad-CAM}^{C} \in \mathbb{R}^{u \times v}$ of width u and height v for class c is obtained by first computing the gradient of the score for class c with respect

CK+ JAFFE ISED

Label: Surprise
Predicition: Surprise

Label: Happy
Prediction: Happy

Label: Happy
Prediction: Happy

Label: Disgust
Prediction: Angry

Label: Sad
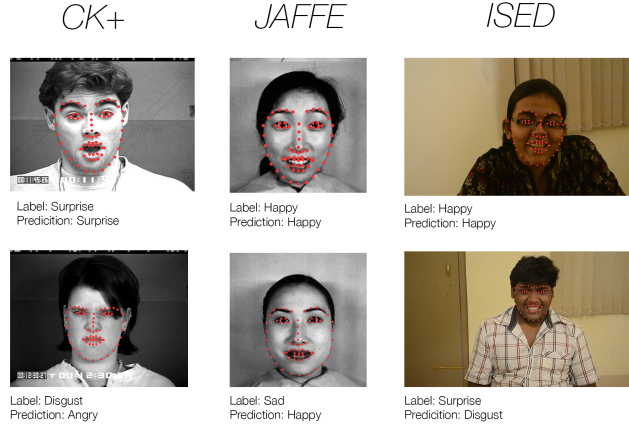Prediction: Happy

Label: Surprise
Predicition: Disgust

Figure 4. Mapping of EmotionDAN predictions to original images from evaluated test sets. The top row shows examples of correct predictions while the bottom one illustrates classification errors. Most of the errors happen when ambiguous emotions are expre

to feature maps A^k of convolutional layer. Then, neuron importance weights are obtained by global-average-pooling these gradients flowing back:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \tag{3}$$

Finally, weighted combination of forward activation maps is followed by ReLU:

$$L_{Grad-CAM}^C = ReLU \left(\sum_k \alpha_k^c A^k \right). \tag{4}$$

We present a visualization of Grad-CAM activations for final convolutional layer (Conv 4a) of EmotionalDAN in Fig. 5. Most of the decisions related to emotion classification is judged based on the regions surrounding mouth, eyes, nose and brows. The model correctly identifies those regions although no prior information about what defines a given emotion was known to the network.

5.2. Grad-CAM activation analysis per emotion label

We perform a detailed analysis of Grad-CAM activations for different emotion labels. We take a subset of AffectNet test set I^{test} such that we only have images with faces facing forward. More formally, we restrict conditions on eye corner locations by taking the following subset of test images:

$$I^{front} = \left\{ I \in I^{test} \quad s.t \quad \|(x,y)_{left\ eye} - (x_l, y_l)\| < \epsilon \wedge \|(x,y)_{right\ eye} - (x_r, y_r)\| < \epsilon \right\} \tag{5}$$
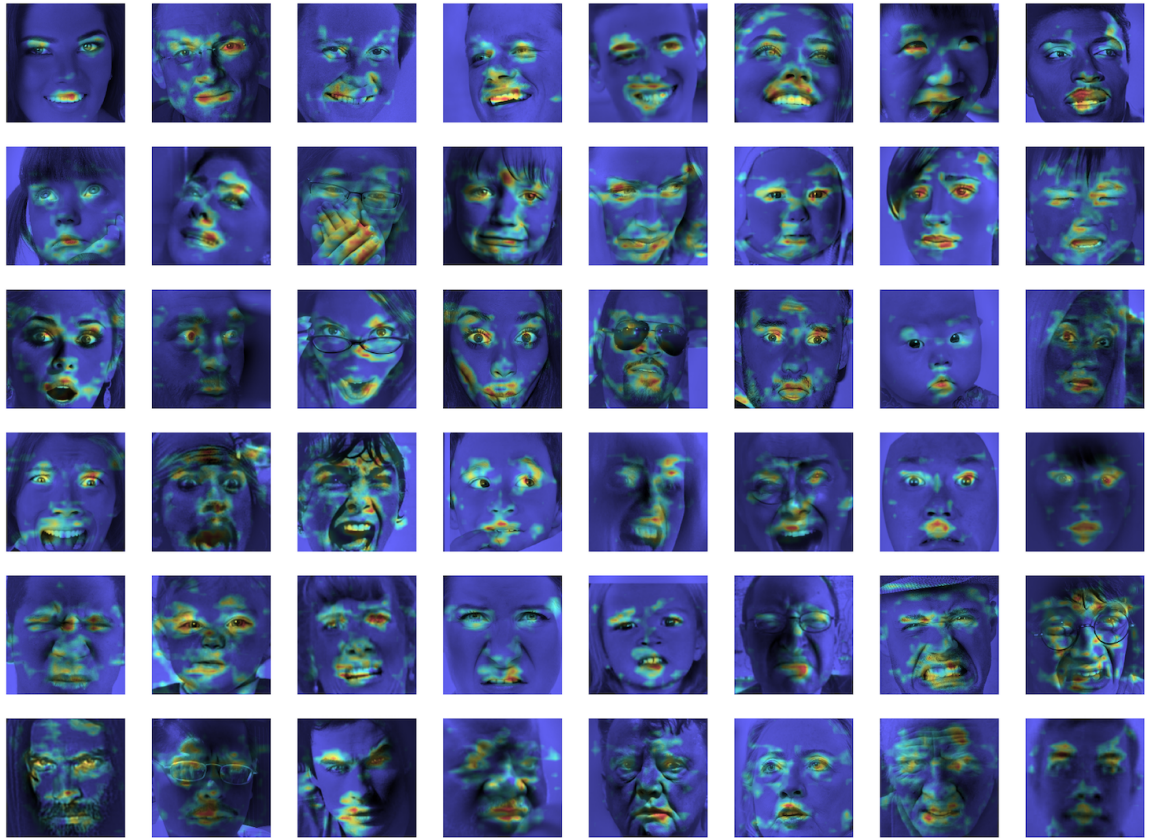
Figure 5.   Grad-CAM visual explanations for emotion classification on test set from AffectNet database. Each row represents random sample of images for emotion label (happiness, sadness, surprise, fear, disgust, anger). EmotionalDAN is able to capture important information from face regions close to eyes, brows or mouth.

where (x_l, y_l) are mean coordinates of left eye left corner and (x_r, y_r) are mean coordinates of right eye right corner.

For each emotion class C:

$$I^C = \left\{ I \in I^{front} \quad s.t \quad y(I) = C \right\} \tag{6}$$

For each set I^C we calculate mean localization map:

$$\overline{L}_{Grad-CAM}^C = \frac{1}{N_C} \sum_{i \in I^C} L_{Grad-CAM}^{(i)}$$

where N_C is the cardinality of I^C. Figure 6 shows heatmaps of mean localization maps for two last convolutional layers in EmotionalDAN architecture: Conv4a and Conv4b. Penultimate convolutional layer Conv4a shows a more focused activation of most important face regions - mouth and eyes.

Going further, we extract most activated regions in each mean localization map and compare them to Emotional Facial Action Coding System (EMFACS [28]). To that end, we use AU descriptions to relate them to facial landmarks. Overview is presented in Table 4. Due to transient nature of AUs, the relationship between AU and related facial landmarks is not a strong one. It however indicates points of interest where information about expressed emotion should be located. For example, *Happiness* is documented as presence of AU6 and AU12, where AU6 is *Cheek Raiser* and AU12 is *Lip Corner Puller* [28]. Hence, to detect happiness one should focus its attention on face regions close to cheeks and lips.

To verify this approach, for each emotion we retrieve top k activated landmarks, where k is the total number of related landmarks for given emotion from Table 4. Visualization of mean activated landmarks is presented in Figure 6. We then calculate overlap between most activated landmarks in our model and landmarks from Table 4. We present detailed results in Table 5. Layer Conv 4a gives slightly closer results to AU related landmarks than Conv 4b.

6. Application

We implement our emotion recognition model as a part of the in-car analytics system to be deployed in autonomous cars. Figure 7 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.

7. Conclusion

In this paper, we overview extension of our previous method [29] for emotion recognition that allows to exploit facial landmarks. Although the results computed on the JAFFE dataset show that there is still
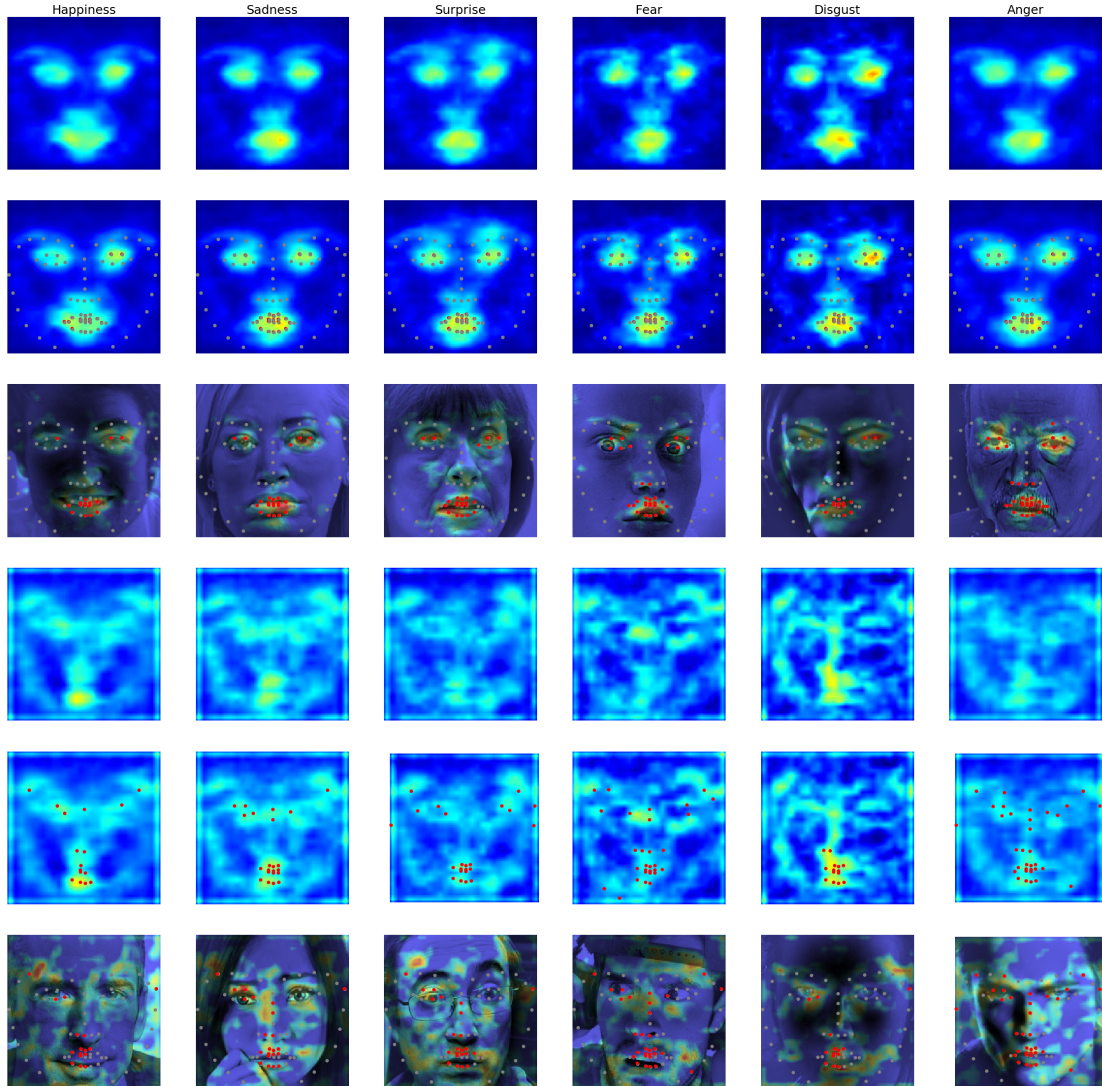
Figure 6. Generated mean Grad-CAM visualizations for each emotion label. First three rows represent heatmaps for Conv 4a layer of EmotionalDAN. The next three were generated with activations from Conv 4b layer. First and fourth rows represent mean Grad-CAM activation heatmaps. Second and fifth present most activated facial landmarks (in red). In third and sixth, images with closest Grad-CAM activations to the mean heatmap are shown.

Table 4.　Facial expression descriptions using EMFACS [28] and their relation to facial landmarks.

Emotion	Related AUs	AU description	Related Facial Landmarks
Happiness	6	Cheek Raiser	1,2 ,14,15
	12	Lip Corner Puller	48, 49, 53, 54, 55, 59, 60, 64
Sadness	1	Inner Brow Raiser	17, 18, 19, 20, 21
	4	Brow Lowerer	22, 23, 24, 25, 26
	15	Lip Corner Depressor	48, 49, 53, 54, 55, 59, 60, 64
Surprise	1	Inner Brow Raiser	20, 21, 22, 23
	2	Outer Brow Raiser	17, 18, 19, 24, 25, 26
	5	Upper Lid Raiser	37, 38, 39, 42, 43, 44
	26	Jaw Drop	55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67
Fear	1	Inner Brow Raiser	20, 21, 22, 23
	2	Outer Brow Raiser	17, 18, 19, 24, 25, 26
	4	Brow Lowerer	17, 18, 19, 20, 21, 22, 23, 24, 25, 26
	5, 7	Upper Lid Raiser, Lid Tightener	37, 38, 39, 42, 43, 44
	20	Lip Stretcher	48, 49, 53, 54, 55, 59, 60, 64
	26	Jaw Drop	55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67
Disgust	9	Nose Wrinkler	27, 28, 29, 30, 31, 32, 33, 34, 35
	15	Lip Corner Depressor	48, 49, 53, 54, 55, 59, 60, 64
	16	Lower Lip Depressor	48, 54, 55, 56, 57, 58, 59, 60, 64
Anger	4	Brow Lowerer	17, 18, 19, 20, 21, 22, 23, 24, 25, 26
	5, 7	Upper Lid Raiser, Lid Tightener	37, 38, 39, 42, 43, 44
	23	Lip Tightener	48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59,60, 61, 62, 63, 64, 65, 66, 67

Table 5.　Accuracy for top activated landmarks with Grad-CAM when compared to Action Units (AU) related landmarks of given emotion. Outputs of different Emotional DAN final convolutional layers are compared.

	Happy	Sad	Surprise	Fear	Disgust	Anger	Avg
Conv 4a	0.375	0.455	0.522	0.633	0.214	0.647	**0.474**
Conv 4b	0.312	0.409	0.478	0.6	0.429	0.559	0.464

Figure 7.   Our emotion recognition model in passenger detection system for autonomous cars. Emotion recognition is performed on detected facial regions.

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 incapabilities related to emotion recognition.

References

[1] P. Ekman and W. Friesen, "Facial action coding system: Investigators guide," *Consulting Psychologists Press*, 1978.

[2] X.-L. Xia, C. Xu, and B. Nan, "Facial expression recognition based on tensorflow platform," *In ITM Web of Conferences*, 2017.

[3] 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.

[4] 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.

[5] B.Kennedy and A. Balint, "Emotionnet2." https://github.com/co60ca/EmotionNet.

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

[7] B. Hasani and M. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[8] S. Kahou, V. Michalski, and K. Konda, "Recurrent neural networks for emotion recognition in video.," *In Proceedings of the ACM on International Conference on Multimodal Interaction*, 2015.

[9] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," *In CVPR*, 2018.

[10] M.Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," *In CVPRW*, 2017.

[11] 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.

[12] S. L. Happy, P. Patnaik, A. Routray, and R. Guha, "The indian spontaneous expression database for emotion recognition," *IEEE Transactions on Affective Computing*, 2017.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *In ICCV*, 2017.

[14] K. Zhao, W.-S. Chu, F. Torre, J. F. Cohn, and Z. H, "Joint patch and multi-label learning for facial action unit detection," *In CVPR*, 2015.

[15] S. Jaiswal, B. Martinez, and M. Valstar, "Learning to combine local models for facial action unit detection," *In IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[16] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," *In CoRR*, 2018.

[17] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," *In ECCV*, 2018.

[18] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[19] A. T. Lopes, E. de Aguiar, and T. Oliveira-Santos, "A facial expression recognition system using convolutional networks," *In SIBGRAPI*, 2015-.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and D. Erhan, "Going deeper with convolutions," *In CVPR*, 2015.

[21] 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.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[23] 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.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computing Research Repository*, 2014.

[25] 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.

[26] Lyons, Akamatsu, Kamachi, and Gyoba, "The japanese female facial expressions database." `http://www.kasrl.org/jaffe.html`.

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016.

[28] P. Ekman and W. Friesen, "Rationale and reliability for emfacs coders," *Unpublished*, 1982.

[29] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: Emotion recognition with facial landmarks," *In CVPRW*, 2018.