Comparison of Angle and Size Features with Deep Learning for Emotion Recognition

Patrick Dunau[1,5(✉)], Marco F. Huber[2,3], and Jürgen Beyerer[4,5]

[1] USU Software AG, Rüppurer Str. 1, 76131 Karlsruhe, Germany
p.dunau@usu.de
[2] Institute of Industrial Manufacturing and Management (IFF), University of Stuttgart, Stuttgart, Germany
[3] Fraunhofer IPA, Stuttgart, Germany
[4] Fraunhofer IOSB, Karlsruhe, Germany
[5] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Abstract. The robust recognition of a person's emotion from images is an important task in human-machine interaction. This task can be considered a classification problem, for which a plethora of methods exists. In this paper, the emotion recognition performance of two fundamentally different approaches is compared: classification based on hand-crafted features against deep learning. This comparison is conducted by means of well-established datasets and highlights the benefits and drawbacks of each approach.

Keywords: Emotion recognition · Classification · Deep learning

1 Introduction

The human face provides a rich display of emotions via facial expressions. The ability to read the sentiment from facial expressions significantly enhances communication, especially in the context of human machine interaction. To allow for a computer to recognise the emotion from a facial expression, it is necessary to classify the sentiment given an image of a human's face.

The field of classifying static images provides a broad variety of methods. This paper presents a comparison of two specific methodologies: the Angle and Size Featureset (ASF) in [3] and deep convolutional neural networks (CNN). The ASF represents the class of feature engineering, while the CNN is a representative of deep learning. While ASF concentrates on specific hand-crafted features to retrieve a high amount of coded information, the CNN automatically mines hidden features. This paper aims for comparing the feature extraction and decision ability of both approaches.

The remainder of this paper organises as follows: the next section formally defines the classification problem. Then each method is described thoroughly in the following Section. In Sect. 4, both classification approaches are compared by means of well-known benchmark datasets. Section 5 concludes this paper.

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 602–610, 2019.
https://doi.org/10.1007/978-3-030-13469-3_70

2 Problem Statement

In general, the classification problem consists of four vital parts: an observation, feature extraction, a classifier, and a prediction. In case of emotion recognition, the observation corresponds to the image of an emotional facial expression. Figure 1 displays the six facial expressions from the Cohn-Kanade+ (CK) dataset for the basic emotion classes *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* provided by Ekman in [4].

(a) (b) (c) (d) (e) (f)

Fig. 1. Figures (a)–(f) show the basic emotion classes anger, disgust, fear, happiness, sadness, and surprise of subject S055 from the CK dataset (©Jeffrey Cohn).

In the feature extraction step structural information is gathered that allows for classifying the emotional class. The process' output corresponds to the class estimate given in terms of a vector containing class probabilities. The whole process is depicted in Fig. 2.

Fig. 2. Process flow of the classification problem.

3 Considered Approaches

In the following, the ASF is compared to the CNN approach. First, the properties and techniques used in the ASF are described. Second, a brief introduction to the CNN is given.

3.1 Angle and Size Featureset (ASF)

ASF is described in detail in Dunau et al. [3]. The feature set constitutes on geometric and local properties. The extraction process is done in four steps: (1) face localisation, (2) facial landmark extraction, (3) estimation of ellipses for sizes computations, and (4) line constructions for angle evaluations. The face is located using a deep neural network provided by the OpenCV library. The facial landmarks are retrieved using the landmark extractor by Qu et al. [6], which yields a landmark model comprising 68 points. Based on this model, the contours of the eyes and the mouth can be retrieved reliably. For the size evaluation, ellipses are fitted to the eyes and the mouth. The size is computed

from the ratio of the major and minor semi-axes of the ellipses. The next step constructs lines from selected pairs of points and computes angles between pairs of lines.

In contrast to the original paper [3], we apply the following modifications: First, the multi-layered perception is replaced by XGBoost classifier [12], which is based on gradient boosted random forests. The XGBoost classifier turned out to be more stable given the problem at hand compared to the MLP classifier. Furthermore, the classifier is not prone to overfitting, given a small amount of data. Also, the experiments showed, that the classifier was able to separate the classes quite well. Second, the implementation of the facial landmark extraction process has been improved. In the original implementation, the computation of the similarity transformation to back-project the landmarks to the facial image was done inversely in an intermediate step, which resulted in low performance and a random choice of alignment positions. The correction of the back-projection stabilised the landmark extraction process and resulted in a significantly improved emotion recognition performance.

3.2 Convolutional Neural Network (CNN)

Deep learning poses a distinct path in machine learning. It is based on artificial neural networks and aims at problems that are easily solved by humans, e.g., emotion recognition given facial expressions. Particularly for image data, so-called deep CNNs as depicted in Fig. 3 have proofed to perform similar or even better compared to humans in some recognition tasks. These networks comprise an input and an output layer as well as several intermediate hidden layers, which can also be convolutional layers.

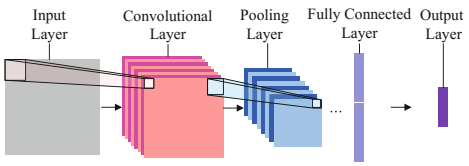

Fig. 3. An example CNN architecture from Peng et al. [14].

The layers represent so called filters. They are used to extract features from the input data to perform the classification task. In the training procedure, the weights of the connections between the layer's nodes are optimized. Additionally, each node has an activation function, which determines the output of the nodes. The activation function used for the network at hand is a Rectified Linear Unit (ReLU) activation function, which is $R(z) = \max(0, z)$.

A deep CNN comprises a vast number of nodes. Consequently, a high number of weights has to be trained, which requires a large number of training data to train the network from scratch. There is also the possibility to use a pre-trained

network. As we are concentrating on the recognition of the emotional class given input images, the training effort can be reduced by using pre-trained models from the imagenet dataset [7]. These nets were trained to discriminate 1000 classes. Given these weights also small datasets like the Cohn-Kanade+ (CK) [1] and the Oulu-Casia (OC) [2] datasets can be utilized. In doing so, training the network becomes rather a fine-tuning of the last 10 layers' and the output layer's weights.

In the literature examples of specially trained CNNs for the classification of emotional facial expressions exist: Lopes et al. [8] present a CNN with seven layers. The images were preprocessed by correcting the face orientation with the alignment of the eye centres to the horizontal axis. Then the face was cropped to eliminate background information from the images to only retrieve information being relevant to expressions. The intensity of the images was normalised to overcome lighting variations, and the images were downsampled to 32 by 32 pixels. Furthermore, data augmentation by image rotation and geometric distortions was applied to increase the dataset. In doing so, a high accuracy of 96.76 % for the CK dataset was achieved. Another example of a seven-layered CNN is presented in Liu et al. [9]. They adapted a 3D CNN to learn a deformable action parts model for dynamic facial expression analysis. Furthermore, they incorporated action parts learning to detect special facial action parts under structured spatial constraints to obtain deformable part detection maps for expression classification. As input n consecutive frames from expression videos were extracted and normalized to 64 by 64 pixels. On the CK dataset, Liu et al. reported accuracy rates up to 92 %. Mollahosseine et al. [10] approximated a sparse neural network by using inception layers for this task. The inception layers were used to incorporate another network into the main network. For preprocessing, the facial images were registered by means of aligning landmarks to an average face. Furthermore, the images were downsized to 48 by 48 pixels and data augmentation based on image cropping and flipping was applied. The resulting network achieved an accuracy of 93.2 %.

In this paper, preprocessing is kept at a minimum on purpose, as the idea of deep learning is to allow end-to-end learning, i.e., images are processed directly. Thus, we merely cropped and resized the images. No further preprocessing is performed.

4 Experiments

For the comparison, we used a VGG-16 network by Simonyan and Zisserman [11] from the Keras library for Python with the TensorFlow backend. As described earlier the network was pre-trained given the imagenet weights. For the problem at hand, the output layer of the network was truncated and replaced by a softmax layer with six output nodes. To avoid overfitting, we equipped a dropout layer before the output layer. A validation set containing 10 percent of the training data was used to regularize the neural network. The ASF was used with the XGBoost classifier. The preprocessing for both models consists of cropping the

facial region and resizing to 224 by 224 pixels. Additionally, the brightness was altered by applying an exponent of 0.5 to the gray values to brighten dark regions of the face.

The six emotional classes given by Ekman [4] were used, as defined in Sect. 2. The data for the experiments was retrieved from the CK and the OC datasets. The CK dataset consists of image sequences ranging from the neutral facial expression to the full emotional expression. The OC dataset is constructed similarly, but the image resolution is lower. From both datasets the last three images of each subject and emotion were used, as these images show the full emotional expression. This increased the number of training samples. For CK this resulted in 1329 and for OC in 1440 images. For the combined test we mixed the images from both datasets and added the neutral images from each participant of the datasets. This resulted in 2769 input images.

We used a computer powered by an Intel Core i7 CPU with 16 GB of RAM, equipped with a NVIDIA Quadro 1000 GPU with four GB of video RAM. For comparison, we computed the metrics recall, precision, f_1-score, and accuracy. A cross-validation scheme with 10 folds was applied to compute the class individual metrics. First, we present the experiment on each dataset individually. Second, we discuss the results on the combined dataset.

4.1 Experiment 1: Cohn-Kanade+ Dataset

In Table 1 the results for the CK dataset are listed, where the tolerances correspond to one-sigma values. The results retrieved by the VGG-16 network are already good, but compared to Lopes et al., Liu et al., Mollahosseini et al., and Liu et al. [8–10,13] superior accuracies on this dataset can be achieved for CNNs. Their improvements in performance result from significantly higher efforts spend on image preprocessing as stated above. The results given from the ASF are significantly higher compared to the VGG-16 and comparable to the state of the art. Also, compared to [3] the improved accuracies obtained by the ASF can be explained from an updated version of the landmark detector by Qu et al. [6] as well as by replacing the MLP classifier with XGBoost.

4.2 Experiment 2: Oulu-Casia Dataset

The experiment on the OC dataset is conducted similarly to the previous one. Table 2 shows the corresponding results, which in case of the VGG-16 network are lower compared to the first experiment. This can be explained by the lower resolution of the pictures contained in the OC dataset. Furthermore, the data is corrupted by high noise, which leads to vanishing facial features. Nevertheless, the ASF performs better compared to VGG-16.

4.3 Experiment 3: Compound Test

The final experiment stacks all pictures from both datasets, i.e., the images showing full emotions from the CK and OC datasets. Table 3 gives the results

Table 1. Experimental results for the CK dataset.

Emotion	VGG-16			ASF		
	Recall	Precision	F_1-Score	Recall	Precision	F_1-Score
Anger	0.51 ± 0.09	0.98 ± 0.04	0.66 ± 0.08	0.97 ± 0.04	0.96 ± 0.05	0.97 ± 0.04
Disgust	0.88 ± 0.07	0.95 ± 0.06	0.91 ± 0.03	0.99 ± 0.02	0.97 ± 0.04	0.98 ± 0.02
Fear	0.98 ± 0.02	0.77 ± 0.02	0.86 ± 0.02	0.98 ± 0.03	0.97 ± 0.03	0.97 ± 0.02
Happiness	0.92 ± 0.04	0.97 ± 0.03	0.94 ± 0.03	0.99 ± 0.02	1.00 ± 0.01	0.99 ± 0.01
Sadness	0.98 ± 0.02	0.79 ± 0.02	0.87 ± 0.01	0.94 ± 0.04	0.96 ± 0.03	0.95 ± 0.03
Surprise	0.98 ± 0.02	0.97 ± 0.02	0.97 ± 0.01	0.97 ± 0.02	0.99 ± 0.02	0.98 ± 0.02
Average	0.88 ± 0.18	0.90 ± 0.10	0.87 ± 0.11	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.01

Table 2. Experimental results on the OC dataset.

Emotion	VGG-16			ASF		
	Recall	Precision	F_1-Score	Recall	Precision	F_1-Score
Anger	0.23 ± 0.19	0.87 ± 0.17	0.31 ± 0.17	0.71 ± 0.11	0.81 ± 0.20	0.72 ± 0.08
Disgust	0.65 ± 0.11	0.49 ± 0.07	0.55 ± 0.04	0.79 ± 0.16	0.61 ± 0.26	0.63 ± 0.09
Fear	0.52 ± 0.05	0.43 ± 0.12	0.46 ± 0.08	0.68 ± 0.06	0.92 ± 0.03	0.78 ± 0.04
Happiness	0.50 ± 0.11	0.58 ± 0.18	0.51 ± 0.06	0.74 ± 0.06	0.95 ± 0.03	0.83 ± 0.04
Sadness	0.10 ± 0.10	0.48 ± 0.26	0.16 ± 0.15	0.77 ± 0.12	0.82 ± 0.19	0.76 ± 0.09
Surprise	0.96 ± 0.31	0.57 ± 0.18	0.69 ± 0.13	0.71 ± 0.04	0.96 ± 0.03	0.82 ± 0.03
Average	0.49 ± 0.31	0.57 ± 0.16	0.45 ± 0.19	0.73 ± 0.04	0.84 ± 0.12	0.76 ± 0.07

Table 3. Experimental results on the mixed dataset.

Emotion	VGG-16			ASF		
	Recall	Precision	F_1-Score	Recall	Precision	F_1-Score
Anger	0.33 ± 0.11	0.82 ± 0.10	0.46 ± 0.10	0.78 ± 0.07	0.76 ± 0.12	0.76 ± 0.05
Disgust	0.67 ± 0.11	0.74 ± 0.11	0.69 ± 0.06	0.83 ± 0.06	0.69 ± 0.17	0.74 ± 0.08
Fear	0.38 ± 0.10	0.75 ± 0.06	0.49 ± 0.08	0.75 ± 0.06	0.89 ± 0.15	0.76 ± 0.07
Happiness	0.70 ± 0.10	0.87 ± 0.03	0.77 ± 0.06	0.85 ± 0.02	0.93 ± 0.03	0.89 ± 0.02
Sadness	0.83 ± 0.06	0.76 ± 0.06	0.79 ± 0.04	0.79 ± 0.03	0.88 ± 0.03	0.83 ± 0.03
Surprise	1.00 ± 0.00	0.49 ± 0.09	0.66 ± 0.07	0.83 ± 0.03	0.96 ± 0.02	0.89 ± 0.01
Average	0.65 ± 0.26	0.74 ± 0.13	0.64 ± 0.14	0.81 ± 0.03	0.84 ± 0.10	0.81 ± 0.06

on the mixed dataset. Again ASF outperforms VGG-16 significantly. The results show a high capability of generalization for the ASF model. Furthermore, the higher amount of training data stabilizes the VGG-16 model. This implies that an even higher amount of training data would improve the CNNs performance.

4.4 Experiments Summary

Table 4 lists the accuracies obtained for all three experiments using the ASF with a XGBoost model and the VGG-16 CNN model. Additionally, the results on the CK dataset reported by Lopes et al., Liu et al., Mollahosseini et al., and Liu et al. [8–10,13] are provided. Liu et al. [13] also gives results on the OC dataset. The results obtained by the ASF are close to or even better than the state of the art.

Table 4. Accuracy results for all tests on the CK and OC datasets.

Model	CK	OC	CK and OC
ASF	0.97 ± 0.01	0.73 ± 0.03	0.81 ± 0.01
VGG16	0.89 ± 0.02	0.49 ± 0.06	0.66 ± 0.05
Lopes et al. [8]	0.97	–	–
Liu et al. [9]	0.92	–	–
Mollahosseini et al. [10]	0.93	–	–
Liu et al. [13]	0.95	0.79	–

With regard to practical applications, it is neccessary to also discuss the runtime of both methods. According to Table 5, the training of a CNN consumes a very high amount of time, while a standard classifier can be trained in a glimpse compared to that. The total runtime comprises the times spent on preprocessing, feature extraction, and prediction. Preprocessing is similar for both methods, but the runtime performing the prediction deviates significantly. Feature extraction in a CNN is performed implicitly by traversing the network and thus, contributes to the prediction time. For XGBoost instead, the actual prediction is very fast, but a lot of time has to be spent for explicit feature extraction. However, it is worth mentioning that VGG-16 benefits from GPU support for all processing steps but preprocessing, while ASF merely runs on CPU. This explains the high runtime for feature extraction, which relies on a CNN for face detection. Thus, the ASF would benefit from a GPU. Furthermore, the high runtime for feature extraction is also caused by the facial feature point detector, as it is repeatedly extracting SIFT features from the input image for every internal regression step.

Table 5. Processing times for the ASF with an XGBoost model and the VGG-16 model. Preprocessing, Feature Extraction, and Prediction correspond to Classification.

Model	Training	Preproc	Feat. Extraction	Prediction	Classification (Total)
VGG-16	1987.8 ms	34.8 ms	–	25.3 ms	60.1 ms
ASF	40 ms	37.8 ms	121 ms	0.08 ms	158.88 ms

5 Conclusions

This paper compares deep learning, represented by the VGG16 model, against the ASF feature set for the emotion recognition problem from facial images. The state of the art indicates that deep learning is capable of reaching high recognition performances similar to ASF, but this paper shows that pure end-to-end learning obtains mediocre results. Very good results can be obtained with very high effort spent on image preprocessing, resulting in a high demand for computational resources as well as a high amount of training data. Contrarily, the ASF is a much more simple model compared to a CNN. This fact provides several advantages: First, the ASF provides an interpretable model. Second, significantly fewer data is needed. Consequently, the ASF has significantly lower computational demands for training. The currently high computational demand for feature extraction can be reduced by utilizing GPUs and by replacing the facial feature point detector.

Currently, the ASF offers a robust feature set for emotion classification based on static imagery. In the future, the model will be extended to handle dynamic data. The explicit exploitation of crafted features efficiently enables the tracking of changes in facial expressions. From that, dynamic analysis and detection of facial expressions can be performed.

References

1. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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 - Workshop Proceedings, pp. 94–101, San Francisco, USA (2010)
2. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikäinen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**, 607–619 (2011)
3. Dunau, P., Bonny, M., Huber, M.F., Beyerer, J.: Reduced feature set for emotion recognition based on angle and size information. In: Strand, M., Dillmann, R., Menegatti, E., Ghidoni, S. (eds.) IAS 2018. AISC, vol. 867, pp. 585–596. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-01370-7_46
4. Ekman, P.: Basic emotions. In: Handbook of Cognition and Emotion, pp. 45–60. Wiley (1999)
5. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
6. Qu, C., Gao, H., Monari, E., Beyerer, J., Thiran, J.P.: Towards robust cascaded regression for face alignment in the wild. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1–9 (2015)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009)
8. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit. **61**, 610–628 (2017)

9. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 143–157. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16817-3_10
10. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10 (2016)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, Volume abs/1409.1556, arXiv (2014)
12. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. CoRR, Volume abs/1603.02754, arXiv (2016)
13. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets via universal manifold model for dynamic facial. CoRR, Volume abs/1511.05204, arXiv (2015)
14. Peng, M., Wang, C., Chen, T., Liu, G.: NIRFaceNet: a convolutional neural network for near-infrared face identification. Information **7**(4), Article no. 61 (2016)