Understanding Abstraction in Deep CNN: An Application on Facial Emotion Recognition

Francesca Nonis[1 [0000−0002−7332−7894]], Pietro Barbiero[1 [0000−0003−3155−2564]], Giansalvo Cirrincione[2], Elena Carlotta Olivetti[1], Federica Marcolin[1 [0000−0002−4360−6905]], and Enrico Vezzetti[1 [0000−0001−8910−7020]]

[1] Politecnico di Torino, Torino, Italy
francesca.nonis@polito.it
[2] University of the South Pacific, Suva, Fiji
nimzoexin59@gmail.com

Abstract. Facial Emotion Recognition (FER) is the automatic processing of human emotions by means of facial expression analysis [1]. The most common approach exploits 3D Face Descriptors (3D-FD) [2], which derive from depth maps [3] by using mathematical operators. In recent years, Convolutional Neural Networks (CNNs) have been successfully employed in a wide range of tasks including large-scale image classification systems and to overcome the hurdles in facial expression classification. Based on previous studies, the purpose of the present work is to analyze and compare the abstraction level of 3D face descriptors with abstraction in deep CNNs. Experimental results suggest that 3D face descriptors have an abstraction level comparable with the features extracted in the fourth layer of CNN, the layer of the network having the highest correlations with emotions.

Keywords: Abstraction, CNN, Deep Learning, Explainable AI, Facial Emotion Recognition, FER

1 Introduction

1.1 Facial emotion recognition and deep learning

Facial Emotion Recognition (FER) is an active line of research in the human-computer interaction domain, due to its potential in many real-time applications, such as surveillance, security and communication. Different architectures of deep neural networks have been proposed, such as Convolutional Neural Networks (CNNs), which have been applied in several research fields, including health care [4] and cybersecurity [5]. Most of the existing algorithms exploit 2D features extracted from images to predict emotions. Albeit computational expensive, 3D feature-based approaches have produced more robust and accurate models thanks to their information supplement [6]. In recent years, CNNs have been successfully employed in large-scale image classification systems and to overcome the hurdles in facial expression classification. The first studies on 3D

FER have appeared only in the last decade, thanks to the publication of the first public databases suitable for this objective [7]. In this research field, the state of the art is currently represented by a few interesting neural-based approaches. In [8], the authors presented a novel deep fusion CNN for subject-indipendent multimodal 2D+3D FER. A 3D facial expression recognition algorithm using CNNs and landmark features/masks, exploiting 3D geometrical facial models only, has been proposed in [9]. Finally, a deep CNN model merging RGB and depth map latent representation has been designed in [10] for facial expression learning.

1.2 Understanding abstraction in deep CNN

In mathematics, abstraction refers to the process of extracting the underlying structure, properties or patterns from observations, removing case specific information, and building high-level concepts that can be profitably applied in unseen but equivalent environments [11][12]. Similarly to animals and human beings, deep neural networks process raw signals by building abstract representations that can be used to generalize to new data. The deeper the layer, the higher the abstraction level. Such abstract representations are encoded in the numeric values of the network weights. The cross-correlation operation [13][14] used in convolutional layers of deep CNNs does not change the data type provided in input. Therefore, when deep CNNs are applied to images, the abstract features extracted by the neural network can be visualized and manually analyzed by domain experts. Based on results and developments in previous studies, the purpose of the present work is to analyze and compare the abstraction level of 3D face descriptors with abstraction in deep CNNs.

2 Data

The data set used in this work was obtained from the Bosphorus 3D facial database [15]. The database contained both 3D facial expression shapes and 2D facial textures up to 54 scans in various poses, expressions and occlusion conditions. Such samples were obtained from 105 different subjects with different racial ancestries and gender (for a total of 4666 face scans). In the following, only two sets of expressions have been considered from the original database. The expressions of the first set were based on Action Units (AU) of the Facial Action Coding System (FACS) [16]. The second set, instead, was composed of facial expressions corresponding to the 6 basic emotions (happiness, surprise, fear, sadness, anger and disgust) plus the neutral expression. The resulting data set was composed of 453 images. Among them, 299 were faces having a neutral expression while the others were almost evenly split into the 6 universal emotions.

3 Methods

The CNN utilized in the following experiments was AlexNet [17]. The input layer of the network requires RGB images having size 227×227. For such purpose,

grayscale depth maps (see Fig. 2) extracted from the Bosphorus database have been cropped, and converted to RGB images by replicating the grayscale channel.

3.1 3D face descriptors

One of the most common techniques for the analysis of human emotions using facial expressions exploits 3D Face Descriptors (3D-FD). 3D-FDs can be generated from depth maps by means of mathematical operators. In this study, the first principal curvature (k1), the shape index (S), the mean curvature (H), the curvedness (C) and a second fundamental form coefficients (f) have been used [18][2]. In Fig. 1 three geometrical 3D-FDs for happiness, sadness and surprise emotions are shown.

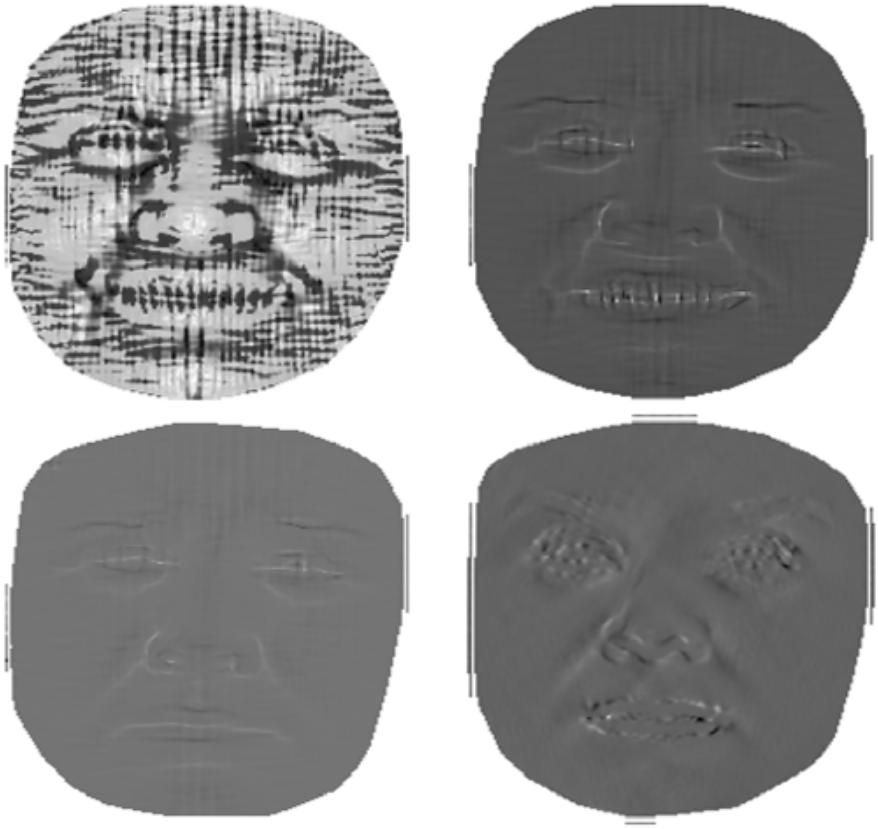

Fig. 1. Geometrical 3D Face Descriptors: S **(top left)**, k1 **(top right)**, H **(bottom left)**, and f **(bottom right)**.

3.2 Transfer learning

When the sample size is small, training a deep CNN from scratch may be time consuming as well as resulting in poor performances or overfitting. Better and satisfying performances can be easily obtained through transfer learning approaches [19]. Given the small amount of samples in the Bosphorus data set, the 3D facial emotion recognition has been performed by using a pretrained AlexNet model. The CNN was fine-tuned using the Bosphorus data set and trained for classifying images into the 7 universal emotions.

3.3 Correlation analysis

The maximum activations of the fine-tuned network were calculated. The corresponding filters were manually visualized and analyzed by domain experts to understand the abstraction process of the network. As expected [20], first layers tend to detect simple patterns like edges, while channels in deeper layers tend to focus on more complex and abstract features like nose and mouth. In order to assess the abstraction level of 3D face descriptors the Pearson correlation coefficient ρ has been used [21]. Pearson's ρ has been computed between each 3D face descriptor and CNN filter activations using three different images representing happy, sad and surprise emotions.

3.4 Symbolic regression

In order to extend the correlation analysis to more complex models, *symbolic regression* [22][23] has been exploited. Symbolic regression is a multi-objective regression analysis for exploring the space of mathematical expressions to find optimal models in terms of accuracy and simplicity. In this experimental setting, symbolic regression was used to assess the abstraction level of 3D face descriptors. Symbolic regression has the advantage of returning human-readable models, that can later be interpreted and explained. For this task, the commercial evolutionary-based software Eureqa Formulize[3] was employed. The software has been used to find mathematical expressions involving CNN filter activations (obtained from images representing happy, sad and surprise emotions) which were highly correlated with 3D face descriptors.

4 Experiments

Given the small number of samples, the data set has been augmented using geometric transformations, such as random reflection in the left-right direction, uniform scaling, and vertical and horizontal translation [24]. In order to assess the network performance, a cross-validation procedure has been applied to the fine-tuning process [25] and a random set of images were selected for the final blind test. The network reached a validation accuracy of 82.67% and a blind test accuracy of 82.09%.

[3] Eureqa Formulize is developed by Nutonian, Inc. https://www.nutonian.com/products/eureqa/

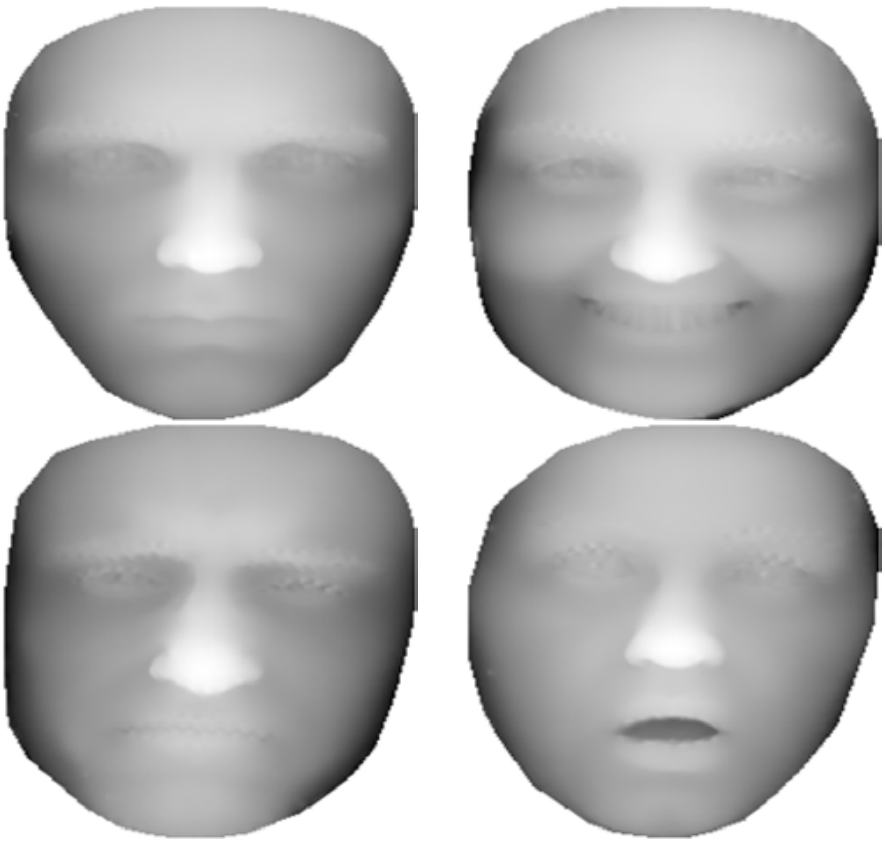

Fig. 2. Grayscale depth maps: neutral expression **(top left)**, happiness **(top right)**, sadness **(bottom left)**, and surprise **(bottom right)** emotions.

4.1 Correlations between filter activations and emotions

Having trained the network for emotion classification, the CNN has been fed with three images representing happy, sad and surprise emotions. The resulting activations of the convolutional layers have been statistically analyzed. Fig. 3 shows the filter with the maximum activation in the fourth (*conv4*) and fifth (*conv5*) convolutional layers for sad and surprise emotions. The selected filter of the fifth layer highlights image areas having strong correlations with emotion patterns, like the mouth and the wrinkles under the eyes. Besides, the most active filter of the fourth layer does not seem to detect human-recognizable patterns. Table 1 shows the highest correlations found in the last two convolutional layers between single filters and emotion images. Analogous results using symbolic regression are presented in Table 2. As expected, symbolic regression generated models having higher correlations with emotions by merging and weighting the contribute of different filters.

Table 1. Highest correlations between single filters and emotions.

Descriptor	Conv4			Conv5		
	Happy	Sadness	Surprise	Happy	Sadness	Surprise
C	0.7482	0,7223	-0,7008	-0,5322	0,5205	0,4722
f	0,7430	0,7313	-0,7007	0,5507	0,5301	0,4754
H	**0,7464**	**0,7355**	**-0,7027**	**0,5773**	0,5311	**0,4929**
k1	0,7450	0,7338	-0,7030	0,5724	**0,5333**	0,4906
S	-0,5229	-0,5283	-0,4863	0,4538	0,3642	0,3896

Table 2. Highest correlations between Conv4 filters and emotions using symbolic regression.

Descriptor	Happy	Sadness	Surprise
C	0.8327	0,8149	-0,7958
f	0,8222	0,8201	-0,8077
H	0,8406	0,8265	**-0,8184**
k1	**0,8498**	**0,8329**	-0,8078
S	-0,6578	-0,6329	-0,6307

4.2 Correlations between filter activations and 3D face descriptors

A similar correlation analysis has been performed between filter activations and 3D face descriptors. Pearson's ρ increased from the input to the output layer of the CNN culminating in the fourth convolutional layer (see Fig. 4). On the

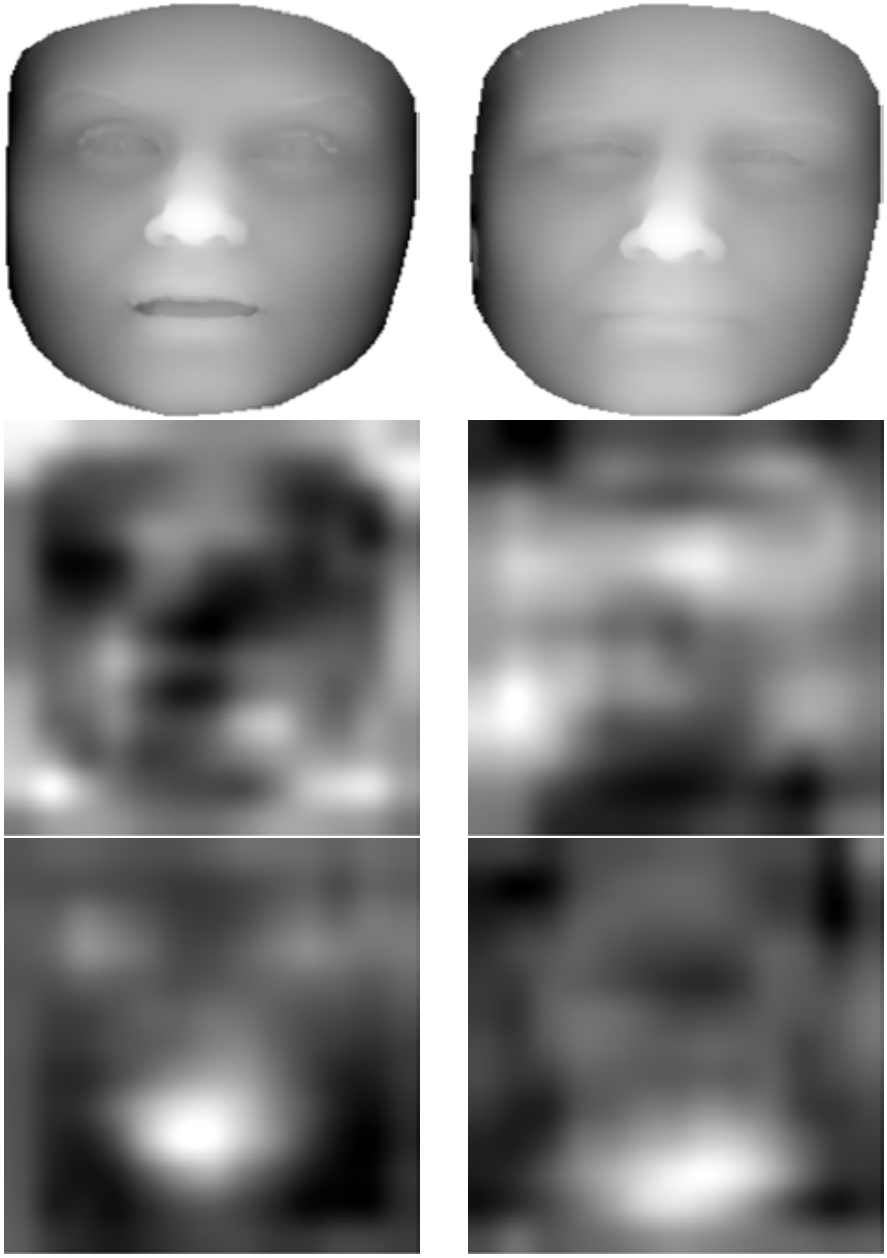

Fig. 3. Channel with the largest activation: *conv4* **(middle)** and *conv5* **(bottom)**, compared to the original image **(top)**.

contrary, in the fifth layer correlations between filter activations and descriptors dropped 4. This result suggests that 3D face descriptors correspond to an abstraction level comparable with the fourth layer of the network.

4.3 Abstraction level of 3D face descriptors

The fourth layer of the network was the one having the highest correlations with emotions. Besides, the above experiments show how 3D face descriptors correspond to a similar abstraction level. However, both for the CNN and from a human point of view the fifth layer is the most useful for emotion classification (compare filters in Fig. 3). These results support the hypothesis that CNNs have a superior level of abstraction with respect to 3D face descriptors. Such superior level may play a key role in transforming features that are highly correlated with emotions (as conv4 filters and descriptors) into useful classification patterns.

5 Conclusions

The main purpose of this work was to analyze the differences between the abstraction level of 3D face descriptors with abstraction in deep CNNs. For this purpose a pre-trained deep CNN was fine-tuned on the Bosphorus data set. Correlation analyzes have been performed both between filter activations and universal emotions, and between filter activations and 3D face descriptors. Experimental results suggested that 3D face descriptors correspond to an abstraction level comparable with the features extracted in fourth layer of the CNN. However, both for the network and from a human point of view the most useful features for emotion recognition correspond to the fifth layer activations. Such features may play a key role in transforming features that are highly correlated with emotions into useful classification patterns. Future steps consist of continuing and deepening the activation and correlation analyses to better understand abstraction in deep CNN.

References

1. Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.
2. Enrico Vezzetti and Federica Marcolin. Geometrical descriptors for human face morphological analysis and recognition. *Robotics and Autonomous Systems*, 60:928–939, 06 2012.
3. Francesca Nonis, Nicole Dagnes, Federica Marcolin, and Enrico Vezzetti. 3d approaches and challenges in facial expression recognition algorithmsa literature review. *Applied Sciences*, 9(18):3904, 2019.
4. Cosimo Ieracitano, Nadia Mammone, Alessia Bramanti, Amir Hussain, and Francesco C Morabito. A convolutional neural network approach for classification of dementia stages based on 2d-spectral representation of eeg recordings. *Neurocomputing*, 323:96–107, 2019.

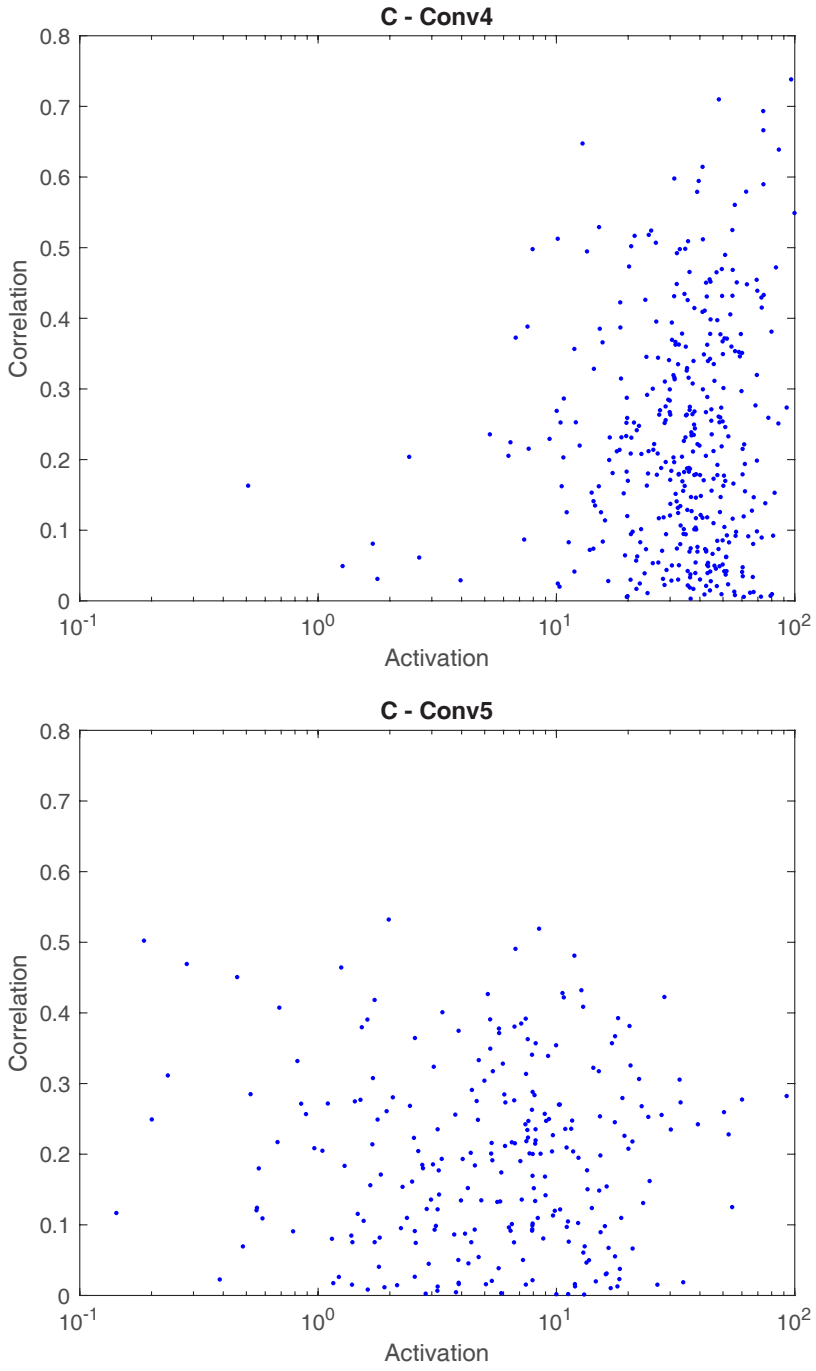

Fig. 4. Relationship between filter activations and correlation values in *conv4* **(top)** and *conv5* **(bottom)**. Correlation analysis has been performed between filter activations and descriptors.

5. Cosimo Ieracitano, Ahsan Adeel, Francesco Carlo Morabito, and Amir Hussain. A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. *Neurocomputing*, 2019.
6. Phung Huynh, Tien-Duc Tran, and Yong-Guk Kim. *Convolutional Neural Network Models for Facial Expression Recognition Using 3D-BUFE Database*, pages 441–450. 02 2016.
7. Zhixing Chen, di Huang, Yunhong Wang, and Liming Chen. Fast and light manifold cnn based 3d facial expression recognition across pose variations. pages 229–238, 10 2018.
8. Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831, 2017.
9. Huiyuan Yang and Lijun Yin. Cnn based 3d facial expression recognition using masking and landmark features. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 556–560. IEEE, 2017.
10. Oyebade K Oyedotun, Girum Demisse, Abd El Rahman Shabayek, Djamila Aouada, and Bjorn Ottersten. Facial expression recognition via joint deep learning of rgb-depth map latent representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3161–3168, 2017.
11. Bertrand Russell. *Principles of mathematics*. Routledge, 2009.
12. Pier Luigi Ferrari. Abstraction in mathematics. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435):1225–1230, 2003.
13. Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.
14. Athanasios Papoulis. *The Fourier integral and its applications*. McGraw-Hill, 1962.
15. Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.
16. E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.
17. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
18. Federica Marcolin and Enrico Vezzetti. Novel descriptors for geometrical 3d face analysis. *Multimedia Tools and Applications*, 76, 07 2016.
19. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
20. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
21. Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.
22. John R Koza and John R Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.
23. Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.
24. Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.
25. Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.