APPLIED MACHINE LEARNING SYSTEM ELEC0134 22/23 REPORT

SN: 22157928

ABSTRACT

In this report, four image classification tasks are investigated: gender detection, smile detection, face shape recognition and eye color recognition. Several approaches were tested on all four tasks, both Deep Learning (DL) approaches, with a CNN architecture, and approaches that used more classical machine learning and computer vision algorithms. The experimental results obtained were satisfactory on all four tasks and showed, where they were employed, the higher performance of the DL approaches.
[1]

***Index Terms*—** Machine learning, deep learning, computer vision, image classification, transfer learning, smile detection, gender detection, face shape recognition, eye color recognition

1. INTRODUCTION

Computer vision (CV) is a field of study that involves the development of algorithms and technologies for interpreting and understanding images and video data. Even though the discipline of CV has just lately become extremely popular, it is by no means a young one. First groundbreaking studies on the subject were carried out as early as the 1950s and 1960s [1, 2]. In the 1970s and 1980s, researchers began to investigate advanced tasks such as feature extraction, pattern recognition, and image understanding, until two groundbreaking studies introduced the concept of Convolutional Neural Network (CNN) [3, 4]. From that moment, many other studies brought advancements to the field [5, 6, 7]. In 2005 the PASCAL VOC project [8] was launched followed by ImageNet [9] in 2009. In 2012, probably the most groundbreaking moment in the history of CV, the architecture AlexNet [10] obtained considerably better results than the previous state-of-the-art, achieving an error rate of 16.4%. From that moment, the development of powerful and efficient neural networks has been a major driving force behind the recent advancements in computer vision. Today, computer vision is a rapidly growing field, with a wide range of applications in areas such as robotics, self-driving cars, medical imaging, and surveillance systems.

[1]The link of the GitHub project is provided: https://github.com/williamdevena/AMLS$_assignment22_23$

1.1. Tasks description

Image classification is one of the most important tasks in the field of CV and can be defined as the task of assigning a label or class to an entire image. This work concerns four specific image classification tasks: gender detection, emotion detection (smile detection), face shape recognition and eye color recognition. The first two are binary tasks and concern classifying the gender of the person shown in the image (male or female) and predicting whether the person is smiling or not. The other two tasks, on the other hand, are multiclass and concern classifying the shape of the face into one of the five predetermined classes and classifying the color of the eyes (iris) into one of the five predetermined colors. The four tasks are performed on two datasets: the first two binary tasks are performed on the images of the CelebA dataset [11], a celebrity image dataset, whereas the other two on the images of the Cartoon Set. The former is composed of 5000 images for the training stage and 1000 for the testing, while the latter of 10000 for training and 2500 for testing. The Fig. **??** shows two examples.

2. LITERATURE SURVEY

2.1. Smile detection

A thorough analysis of the smile detection literature has revealed the usage of a variety of strategies, including both more traditional machine learning methods without the use of Neural Networks (NN) and Deep Learning approaches, which are mostly represented by CNNs. The authors of [12] train a CNN using pairs of images and two supervisory signals: a recognition signal (the label of the input sample) and an expression verification signal. The former is used to classify the image into smiling or non-smiling and is used in combination with a cross entropy loss, whereas the latter is used to limit variations of features extracted from an image with a similar expression while encouraging to vary the feature on images with a different expression, and is used in combination with a loss function based on the L2 norm. In [13], a CNN was employed as a feature extractor for high level features, which were then used to fit several ML classification algorithms, with the results demonstrating that SVM was the most effective technique. The authors of [14] also make use of a CNN for the final classification, but preceded by a face detector algorithm used to crop the face from the image. Similar

work has been done by the the authors of [15], that used transfer learning to improve the performance of the well known CNN architecture VGG-16 initializing it with the weights of the VGG-face model [16] trained on 2.6M images. In [17] the authors after pre-processing the images with Histogram Equalization, use AdaBoost algorithm to perform face detection and optical flow technique to predict the positions of the left and right corners of the mouth. As a last step, a threshold method based on the distance between the two corners is used to perform the classification. Finally, in [18] several feature extraction techniques have been tested in combination with SVM and GentleBoost algorithms, including Gabor Energy Filters, Box Filters, Edge Orientation Features and LBP (Local Binary Patterns).

2.2. Face shape classification

In literature, the face shape detection task has been tackled with several types of methodologies. In [19] a pre-processing stage using CLAHE and Gabor Energy Filters is followed by the feature extraction performed with an invariant moments technique and finally a probabilistic neural network is used to perform the classification. The authors of [20] make use of a region similarity method and a correlation method to estimate the similarity between the shape of the face and geometric shapes. In [21] several feature extraction stages are applied before using a NN to perform the classification. In particular, a first face detection stage is applied using SVM trained on HOG (histogram of oriented gradients) features, followed by ERT (Ensemble Regression Trees) that performs landmark feature extraction. Subsequently, the images that have been cropped using the face detection, are used as training samples for the Inception-V3 convolutional neural network. Finally, the high level features extracted by the CNN are combined with the landmark features and fed into three FC (fully connected) layers that perform the final classification. Furthermore, in both [22] and [23] the SVM algorithm is used on landmark features, using the RBF kernel. Finally, in [24] the authors try to perform face shape classification using 3D data.

2.3. Gender detection

In gender detection literature, the approaches found were mainly DL-based. The authors of [25, 26, 27] used a similar approach, composed of a face detection stage to crop the image and a classification stage performed by a CNN. Similar is also [28] where the authors use the images directly fed into the CNN without a face detection stage. To further limit the risk of overfitting, the author used data augmentation techniques. Furthermore, also more classical CV approaches were found, like [29]. In particular, the first step of their approach is a face detection, followed by a pre-processing stage, where they made use of HE (Histogram Equalization), subsequently a feature extraction performed with LBP and DCT (Discrete Cosine Transform) and finally a classification stage where they used the manhattan distance calculated between the sample of interest and all the other samples in the training set.

2.4. Eye color classification

With an extensive review of the eye color recognition literature it has been found that very few if none works have used DL methods. Preliminary attempts of eye (iris) color detection have been made in [30, 31]. Further advancements, have been introduced by [32] that used Viola and Jones algorithm [6] to segment the iris and Gaussian Mixture Model to classify its color. Furthermore, both the authors of [33, 34] used SVM to perform the final classification step. In particular, the former made use of LBP and BSIF (Binary Statistical Images Features) as feature extraction techniques, while the latter used hue, saturation and value of the images.

3. DESCRIPTION OF MODELS

3.1. Face detection and Landmarks prediction approaches

This section describes two approaches that have been tested across several of the four tasks. The first approach, referred to as dynamic cropping, is composed of a face detection and landmarks prediction stage, a cropping stage and a final classification step performed on the cropped image (Fig. 1). In particular, the following is a detailed scheme of the overall approach:

- Face detection and landmark prediction stage. This stage was performed using a combination of the feature descriptor HOG (Histogram of Oriented Gradients) and the classification algorithm SVM (Additional details in the Additional materials).

- The cropping stage uses the predicted landmarks to extract the ROI (Region of Interest) from the entire image. In particular, it defines and crops a rectangle that contains the ROI plus a predetermined margin. Two different approaches were used regarding the definition of the ROI: we select only the region that contains the mouth (in the smile detection and gender detection task) or we select the region that contains the eyes (only for the gender detection task).

- To have all the input samples of the same dimensions, the cropped image is resized to predetermined dimensions and transformed into a flat format, resulting in a (HW)x3 feature matrix, where H and W are the width and the height of the resized image and 3 is the number of channels of an RGB image.

- Finally the image is passed through the classification step.

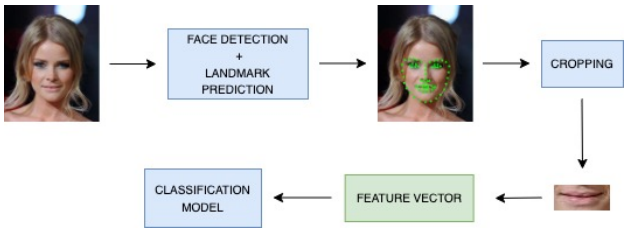


Fig. 1: High level view of the first face detection and landmark prediction approach. In particular, the example shows a specific case where the ROI is the mouth (approach used for the smile detection task).

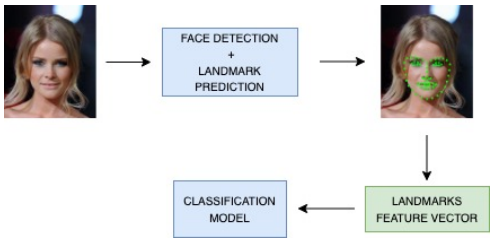


Fig. 2: High level view of the second face detection and landmark prediction approach.

The second approach (Fig. 2), is composed of the same face detection and landmarks prediction stage of the first approach, while using the landmarks in a different way. In particular, the landmarks, represented by a 68x2 feature vector, are used directly as input samples for the classification step. Each landmark is represented as (x, y) where x and y are the row and the column of the detected landmark. Regarding the final classification step, two models have been used: KNN and SVM. In particular, for the SVM algorithm, the RBF (Radial Basis Function) kernel has been used as it resulted in the best performing and also the most present in the literature reviewed ([34, 22, 23]). The rationale behind the first approach is that, by selecting and cropping the ROI we are able to both perform dimensionality reduction, reducing the computational cost and reducing the number of misleading features, and selecting a part of the image that is highly informative and predictive of the classes. Similar rationale is the one behind the second approach, where the features extracted are the coordinates of the landmarks. In particular, in this case, the dimensionality reduction is even more aggressive than the first approach and the features selected are in an intrinsic way a representation of the shape of the face and of its different parts, which can also be considered a high level and predictive features of the classes. These two approaches, thanks to their adaptive nature, that comes from the possibility of changing the ROI and selecting a subset of the 68 landmark features, represent a common framework used across all four tasks.

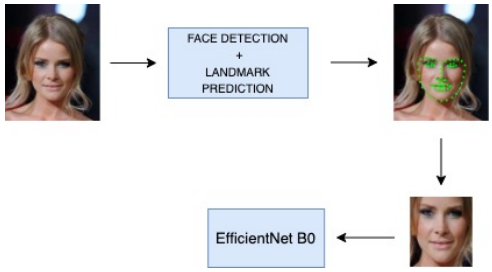


Fig. 3: Illustration of the CNN approach used on the gender detection task.

3.2. Task A1: Gender detection

For this task, the approaches tested were several. In particular, both approaches described in the Section 3.1 were tested. Furthermore, following the trends found in the literature review, a DL approach was used to improve the performance. Regarding the face detection and landmark prediction approaches, the first (Fig. 1) was used in both versions, that is defining the ROI as the area that contains the mouth and as the area that contains the eyes. The logic in this case, behind using only a part of the image, while removing highly informative features, was understanding and comparing the predictive power of the selected ones (mouth in one case and eyes in the other). The comparison will be shown in the Section 5. Regarding the second approach (Fig. 2), all the 68 landmark features were used (136 features in total in a flatten format). Finally, a DL method, based on a CNN and Transfer Learning, was used. In particular, the EfficientNet [35] CNN architecture was used and the initialization of the model was performed using weights that were pretrained on the ImageNet dataset [9] (more details on transfer learning in the additional materials). The CNN is preceded by the "dynamic cropping" approach described in the Section 3.1 (Fig. 1) that in this case is used to crop the face. The overall approach is illustrated in the Fig. 3.

3.2.1. EfficientNet architecture

The primary goal of the authors of [35] was to build a scalable architecture to satisfy all kinds of computational capacities. In particular, they developed a baseline architecture (B0 version) with 5.3M parameters, that can scale up to a version with 66M (B7 version). This scaling power is the main reason behind the choice of this architecture, because it both allowed to conduct experiments and training with the strong computational limits that characterized this work, using the B0 version, but at the same time allows in possible future works to scale to a bigger model without changing the core architecture. The baseline architecture is composed of 9 layers, 7 of which are MBConv layers, a kind of bottleneck neck layer, introduced by the authors of [36]. In particular, an MBConv layer is an Inverted Linear BottleNeck layer that uses Depth-

Wise Separable Convolution. The main differences between a classic bottleneck layer [37] and an Inverted Linear Bottleneck are the expansion that happens at the beginning of an MBConv, that increases the number of channels of the input, whereas in the bottleneck the input is compressed in a smaller number of channels (that is why the name bottleneck), and the removal of the non-linear activation (ReLU) in the last layer. The authors explain the reason behind this last choice, by empirically showing that when the input channels are less than the output ones, a non-linear activation function, despite its ability to increase representational complexity, decreases performance causing a loss of information. Furthermore, as a normal bottleneck an MBConv has a residual skip connection that adds the input to the final output. The residual block was first introduced by [37] where the authors showed that adding a skip connection to the layers of a deep neural network solves the difficulties of optimizing it during training, caused by problems like vanishing gradients. Finally, an MBConv uses Depth-Wise Separable Convolution, a particular kind of convolutional layer that separates the convolution operation in two steps, reducing significantly the number of parameters (almost by a factor of k^2, where k is the kernel size of the convolution [36]), while having, from an accuracy point of view, a low cost. In particular, the normal convolution is divided into a depthwise convolution and a 1x1 convolution called pointwise convolution: the depthwise convolution applies a single kernel to each of the channels of the input, while the pointwise convolution uses a 1x1 convolution to combine the outputs of the former step (Illustration in the additional materials).

3.3. Task A2: Smile detection

For this task, the two approaches described in Section 3.1 were used. In particular, for the first approach (Fig. 1) the ROI was defined as the area that contains the mouth, while for the second approach (Fig. 2) the landmarks used where all 68. The rationale behind these two approaches is detailed in the Section 3.1.

3.4. Task B1: Face shape recognition

For the task of face shape recognition, a similar approach to the second one described in 3.1 has been used (Fig. 2). In particular, the approach for this task has been to use a subset of the face landmarks features. The selected subset is the one that contains the landmarks of the jaw (first 18 landmarks). Furthermore, because the performance of this approach is not entirely satisfactory (more details in the Section 5), a DL approach, similar to the one used in the gender detection task was used. The rationale behind choosing a DL model, was switching to a more complex model with higher expressive power, since the simpler models failed to return satisfactory results.

3.5. Task B2: Eye color recognition

For this task, the first approach described in Section 3.1 (Fig. 1) was used. In this case, the ROI was defined as that area that contained one eye (the right one). Furthermore, in the Section 5 some important aspects will be discussed that led to not look further into developing new methods to improve the performance of the one just mentioned.

4. IMPLEMENTATION

4.1. Tools

The programming language used was Python. Additionally, several well-known libraries in the fields of ML and Data Science were used: Numpy, MatplotLib, Pandas, Scikit-learn, PyTorch, OpenCV and Albumentations are the main ones. To acquire more computational power, especially for the training of the DL models, the online platform Google Colab was used.

4.2. Datasets

The datasets used in this work are two and they are both preprocessed subsets of two existing datasets: CelebA [11] and Cartoon Set [38]. From now on, for convenience, the datasets that have been used will be referred to with the names of the original datasets. The CelebA dataset is composed of 6000 jpg images in total, 5000 for training and 1000 for testing ($\sim$ 83% training and $\sim$ 17% for testing). On the other hand, the Cartoon Set dataset is composed of 12500 png images total, 10000 for training and 2500 testing ($\sim$ 80% training and $\sim$ 20% for testing). Furthermore, a data exploration stage showed that both datasets, both in the training and the testing set, present a balance between all classes. Additionally, in the Section 4.3 it will be discussed how the splitting of the validation set, for the training of DL models, was handled. Furthermore, for convenience all intermediate steps mentioned in the approaches, before the classification step, are done offline and saved locally.

4.3. EfficientNet training

As already mentioned, for the task of gender detection, one of the approaches used a pretrained CNN architecture called EfficientNet B0. In particular, a PyTorch implementation was used. Regarding the training, a validation set was extracted from the original training set of 5000 images, resulting in a final split of 4000 images for training, 1000 for validation and 1000 for testing ($\sim$ 66% training, $\sim$ 17% for testing and $\sim$ 17% for validation). The choice of extracting the validation set from the training set was led by the fact that extracting it from the testing set would have resulted in a too unbalanced split ($\sim$ 83% training, $\sim$ 8.5% for testing and $\sim$ 8.5% for validation). The strategy followed for the stopping criterion is

early stopping, a technique that consists in stopping the iteration when the validation error increases. Because the increase of the validation error is often associated with overfitting, the early stopping technique is also one of the most simple and used forms of regularization. Regarding the hyperparameters, the learning rate was set to 0.0001, the batch size to 4, the optimization algorithm used was Adam and for the binary gender detection task, a Binary Cross Entropy Loss function was used, whereas for the multiclass face shape recognition task a Cross Entropy Loss function was used. In particular, the choice of the batch size was led both by computational limit, but also inspired by the work [39] where the authors demonstrated empirically that on several benchmark datasets like ImageNet, using small batch size achieves better training stability and generalization performance.

4.3.1. Regularization

One of the major difficulties to overcome when training a DL model, is tackling overfitting. Regularization refers to all the techniques aimed at preventing overfitting. In this work, several regularization techniques were used to prevent the model from overfitting: early stopping, L2 regularization, dropout and stochastic depth (more details in the Additional materials).

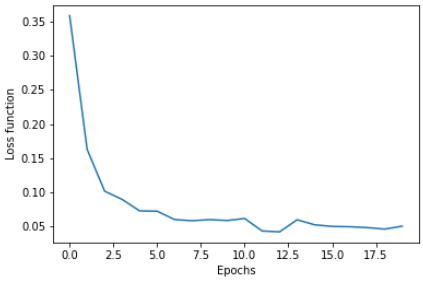
4.4. SVM and KNN

As mentioned before, for the final classification step of the two approaches described in the Section 3.1, SVM and KNN algorithms were used. Regarding the implementation, the library Scikit-learn was used. Furthermore, to select a value for the hyperparameter k of the KNN algorithm, k-fold cross validation was performed (example shown in the additional materials).

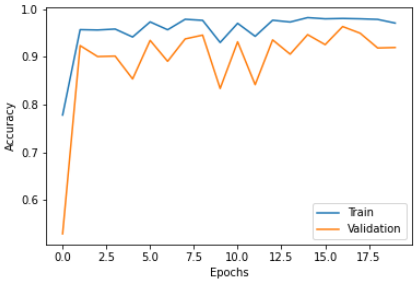
5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. CNN Training

As mentioned before, the adopted strategy was early stopping. Regarding both the gender detection and face shape recognition tasks, for computational and time reasons the training were stopped after 20 epochs. The (Fig. 4 and Fig. 5) show that the accuracies slowed down but did not show signs of overfitting, the loss functions were still decreasing, even if at a slower rate, and converging to 0, sign that the models could have still improved their performance. Both plots of Fig. 4 and Fig. 5 show evidence of the advantage that transfer learning gives. In particular, both plots show a large increase, reaching close to the final accuracy, immediately in the first epochs. This desirable behavior is attributable to the effect of the pre-training. Further evidence can be found in the additional material.

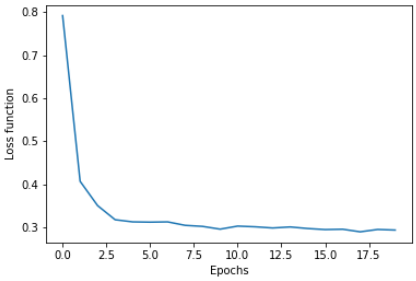


(a) Plot of the Binary Cross Entropy loss function.

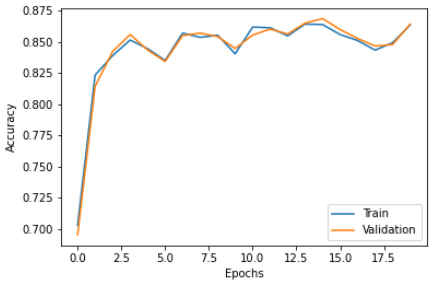


(b) Plot of the training accuracy and of the validation accuracy.

Fig. 4: Plots of the EfficientNet training on the gender detection task.



(a) Plot of the Cross Entropy loss function.



(b) Plot of the training accuracy and of the validation accuracy.

Fig. 5: Plots of the EfficientNet training on the face shape recognition task.

5.2. Results and Analysis

For the sake of evaluating the approaches the accuracy metric was used, which is calculated dividing the number of correct predictions by the total number of testing samples. Both Table 1 and Table 2 show that the higher accuracy is obtained by the DL approaches, in particular in the face shape recognition task, where models with lower expressive power, like SVM and KNN, failed to return satisfactory results. The reason can be attributed to the fact that DL models have a higher capacity compared to more classical machine learning models and to the fact that with DL models, the feature extraction stage is implicit in the training and the feature extracted are more complex and of higher level than the handcrafted ones. Furthermore, an interesting detail that the results show is that, from both only the eyes and only the mouth, the approaches used were to able to classify the gender with satisfactory performance.

Table 1: Results of the gender detection task.

Method	Acc
Baseline (KNN)	0.75
Dynamic cropping (mouth) + SVM	0.87
Dynamic cropping (mouth) + KNN	0.8
Dynamic cropping (eyes) + SVM	0.92
Dynamic cropping (eyes) + KNN	0.88
Dynamic cropping (face) + EfficientNet	**0.942**
Landmark features + SVM	0.84
Landmark features + KNN	0.78

Table 2: Results of the face shape recognition task.

Method	Acc
Landmark features (all) + SVM	0.29
Landmark features (all) + KNN	0.45
Landmark features (jaw) + SVM	0.34
Landmark features (jaw) + KNN	0.53
Dynamic cropping (face) + EfficientNet	**0.86**

The analysis of the intermediate output of the face detection and landmarks prediction (called for simplicity dynamic cropping), revealed that in some cases the stage fails to detect landmarks. Because the resulting output in these cases is just a blank image, the classification step becomes impractical. To investigate how much of the testing error was caused by this, the dynamic cropping approach was compared to a static cropping approach where there images were statically cropped in a predetermined area. In particular, this last approach was possible because in the dataset the faces are all centred and all of the same scale, but it would not have been possible in a real-world dataset or in another dataset where the faces are not aligned or of different scales, whereas the dynamic cropping approach has more generalizing power. The Table 3 shows how the static cropping obtains 0.84 of accuracy compared to the 0.71 of the dynamic approach. This gives evidence that by improving the face detection and landmark prediction stage, the overall approach would certainly improve its performance. Furthermore, a careful analysis of the Cartoon Set dataset, revealed that roughly 15-20% of the cartoons in the images present dark glasses that prevent both humans and models from predicting the eye color, setting an upper limit to the accuracy of 0.8-0.85, hence eventual improvements would be just a matter of random lucky guesses. Further evidence is shown in the additional material.

Table 3: Results of the eye color recognition task.

Method	Acc
Baseline (KNN)	0.25
Dynamic cropping + SVM	**0.71**
Dynamic cropping + KNN	0.67
Static cropping + SVM	**0.84**

This upper bound of the accuracy, combined with the accuracy of the static cropping approach, shows that improving the face detection and landmarks prediction stage could increase the accuracy of the overall approach, up to the upper bound just discussed. Furthermore, the Table 4 shows the results of all the approaches tested on the smile detection task. In particular, the one with the highest accuracy is illustrated in the Fig. 2. Additionally, similar to the aspect discussed previously, about the Table 1, this one also shows an interesting aspect that regards the different parts of the face and how the approaches manage to classify the smiling of a person by just seeing the eyes with a certain accuracy (0.75).

Table 4: Results of the smile detection task.

Method	Acc
Baseline (KNN)	0.65
Dynamic cropping (mouth) + SVM	0.88
Dynamic cropping (mouth) + KNN	0.85
Dynamic cropping (eyes) + SVM	0.75
Dynamic cropping (eyes) + KNN	0.66
Landmark features + SVM	0.89
Landmark features + KNN	**0.9**

6. CONCLUSIONS

Regarding the gender detection and face recognition tasks, caused by computational and time limits, in both cases the training was stopped before reaching the stopping criteria. Further developments, apart from increasing the number of

epochs, could include several training strategies, like decreasing the learning rate during training or increasing the batch size during training [40]. On the smile detection task, our approach returned satisfactory results (accuracy of 0.9). For further improvements, as all the approaches that use it, the attention should be directed at improving the face detection and landmark prediction stage. In particular, a possible route to explore is using DL models. A similar thing is in the eye color recognition task where, as already mentioned in the Section 5, the performance has an upper bound that has already been reached by the static cropping approach. Because the static approach, contrary to the dynamic one, has not the ability to generalize to other datasets, further improvements should be directed into improving the face detection and landmark prediction stage to bring the performance of the dynamic approach up to the one of the static.

7. REFERENCES

[1] David H Hubel and Torsten N Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574, 1959.

[2] Lawrence G Roberts, *Machine perception of three-dimensional solids*, Ph.D. thesis, Massachusetts Institute of Technology, 1963.

[3] Kunihiko Fukushima and Sei Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.

[4] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[5] David G Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*. Ieee, 1999, vol. 2, pp. 1150–1157.

[6] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Ieee, 2001, vol. 1, pp. I–I.

[7] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE conference on computer vision and pattern recognition*. Ieee, 2008, pp. 1–8.

[8] Mark Everingham, Andrew Zisserman, Christopher KI Williams, Luc Van Gool, Moray Allan, Christopher M Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorkó, et al., "The 2005 pascal visual object classes challenge," in *Machine Learning Challenges Workshop*. Springer, 2005, pp. 117–176.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[11] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "From facial parts responses to face detection: A deep learning approach," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3676–3684.

[12] Kaihao Zhang, Yongzhen Huang, Hong Wu, and Liang Wang, "Facial smile detection based on deep learning features," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 534–538.

[13] Junkai Chen, Qihao Ou, Zheru Chi, and Hong Fu, "Smile detection in the wild with deep convolutional neural networks," *Machine vision and applications*, vol. 28, no. 1, pp. 173–183, 2017.

[14] Simone Bianco, Luigi Celona, and Raimondo Schettini, "Robust smile detection using convolutional neural networks," *Journal of Electronic Imaging*, vol. 25, no. 6, pp. 063002, 2016.

[15] Xin Guo, Luisa Polania, and Kenneth Barner, "Smile detection in the wild based on transfer learning," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 679–686.

[16] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," 2015.

[17] Yu-Hao Huang and Chiou-Shann Fuh, "Face detection and smile detection," in *Proceedings of IPPR Conference on Computer Vision, Graphics, and Image Processing (CVGIP)*, 2009.

[18] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan, "Toward practical smile detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 2106–2111, 2009.

[19] Romi Fadillah Rahmat, Muhammad Dian Syahputra, Ulfi Andayani, and Tifani Zata Lini, "Probabilistic neural network and invariant moments for men face shape classification," in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing, 2018, vol. 420, p. 012095.

[20] NK Bansode and PK Sinha, "Face shape classification based on region similarity, correlation and fractal dimensions," *International Journal of Computer Science Issues (IJCSI)*, vol. 13, no. 1, pp. 24, 2016.

[21] Theiab Alzahrani, Waleed Al-Nuaimy, and Baidaa Al-Bander, "Hybrid feature learning and engineering based approach for face shape classification," in *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*. IEEE, 2019, pp. 1–4.

[22] Mohamed Hossam, Ahmed Ashraf Afify, Mohamed Rady, Michael Nabil, Kareem Moussa, Retaj Yousri, and M Saeed Darweesh, "A comparative study of different face shape classification techniques," in *2021 International Conference on Electronic Engineering (ICEEM)*. IEEE, 2021, pp. 1–6.

[23] Wisuwat Sunhem and Kitsuchart Pasupa, "An approach to face shape classification for hairstyle recommendation," in *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 2016, pp. 390–394.

[24] Pornthep Sarakon, Theekapun Charoenpong, and Supiya Charoensiriwath, "Face shape classification from 3d human data by using svm," in *The 7th 2014 Biomedical Engineering International Conference*. IEEE, 2014, pp. 1–5.

[25] B Abirami, TS Subashini, and V Mahavaishnavi, "Gender and age prediction from real time facial images using cnn," *Materials Today: Proceedings*, vol. 33, pp. 4708–4712, 2020.

[26] Abdullah M Abu Nada, Eman Alajrami, Ahmed A Al-Saqqa, and Samy S Abu-Naser, "Age and gender prediction and validation through single user images using cnn," 2020.

[27] Insha Rafique, Awais Hamid, Sheraz Naseer, Muhammad Asad, Muhammad Awais, and Talha Yasir, "Age and gender prediction using deep convolutional neural networks," in *2019 International conference on innovative computing (ICIC)*. IEEE, 2019, pp. 1–6.

[28] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 34–42.

[29] Emon Kumar Dey, Mohsin Khan, and Md Haider Ali, *Computer vision based gender detection from facial image*, Citeseer, 2013.

[30] M Melgosa, MJ Rivas, L Gomez, and E Hita, "Towards a colorimetric characterization of the human iris," *Ophthalmic and Physiological Optics*, vol. 20, no. 3, pp. 252–260, 2000.

[31] Shaohua Fan, Charles Dyer, and Larry Hubbard, "Quantification and correction of iris color," Tech. Rep., University of Wisconsin-Madison Department of Computer Sciences, 2003.

[32] Antitza Dantcheva, Nesli Erdogmus, and Jean-Luc Dugelay, "On the reliability of eye color as a soft biometric trait," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2011, pp. 227–231.

[33] Denton Bobeldyk and Arun Ross, "Predicting eye color from near infrared iris images," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 104–110.

[34] Camelia Florea, Mihaela Moldovan, Mihaela Gordan, Aurel Vlaicu, and Radu Orghidan, "Eye color classification for makeup improvement," in *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2012, pp. 55–62.

[35] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] Google, "Cartoon set," `https://google.github.io/cartoonset/`, (accessed January 11, 2021).

[39] Dominic Masters and Carlo Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 2018.

[40] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le, "Don't decay the learning rate, increase the batch size," *arXiv preprint arXiv:1711.00489*, 2017.