Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

A Study on Facial Landmark Detection and Image Processing

Xiyi Xiong *

Department of Information Engineering, University of Sussex, Brighton, England

* Corresponding Author Email: xx206@sussex.ac.uk

Abstract. This research focuses on developing a robust face alignment system crucial for applications in facial recognition, expression analysis, and augmented reality. Utilizing a dataset of 2811 training images and 554 test images each annotated with 44 facial landmarks, the study explores several models including Convolutional Neural Networks (CNN), Cascade Regression, and Residual Networks (ResNet). Through rigorous experimentation, it was observed that while CNNs and Cascade Regression models provided substantial accuracy, the ResNet model displayed superior performance, especially in complex scenarios involving diverse expressions and occlusions. Additionally, the project also developed a simple algorithm for lip and eye color modification, further broadening the applicability of the system in image processing. Results indicate that while traditional models like CNN and Cascade Regression perform adequately, ResNet offers superior accuracy and robustness, particularly in challenging conditions. This research confirms the effectiveness of ResNet in enhancing face alignment technologies and suggests potential areas for future improvements in the field.

Keywords: Face Alignment, ResNet, Convolutional Neural Networks, Cascade Regression, Facial Landmark Detection.

1. Introduction

All Face alignment, which involves locating facial landmarks in images, is a crucial task in computer vision. It has wide applications in face recognition, expression analysis, and augmented reality. In the early 1990s, facial feature point detection, as a key technology in facial recognition, attracted the attention of scholars in the field of computer vision and was widely studied. Especially in the past decade, significant progress has been made. Nowadays, accurate and efficient facial feature point detection has become an important prerequisite and foundation for industrial implementation in natural scenes using technologies such as facial synthesis, head pose estimation, facial 3D reconstruction, and virtual reality.

The cascade regression method is one of the mainstream methods for dealing with facial feature point detection problems and has achieved great success in handling pose changes, complex expressions, and other aspects [1-3]. The cascaded regression method mainly uses multi-layer discriminant regression functions to directly map the facial appearance to the displacement space of feature points and then corrects the preset initial position from coarse to fine through the appearance information, ultimately obtaining the accurate position of facial feature points [4].

Since Geoffrey Hinton et al. proposed Deep Learning in 2006, it has gained popularity in many fields of computer vision [5, 6]. In 2012, Hinton's research team achieved great success in training deep learning models for image classification on the large image dataset ImageNet [7, 8]. Since then, researchers have gradually started using deep learning methods to solve the problem of facial feature point detection and have achieved very good results [9]. Convolutional networks can effectively extract features from various dimensions of images without the need for feature design and can be well applied in classification and regression problems.

The convolutional neural network is a deep learning algorithm that can obtain input images, assign weights and biases to various objects in the image, and distinguish them from other classification algorithms. The preprocessing requirements of convolutional neural networks are much lower than those of other networks. In the initial method, the filter was manually designed, and with sufficient

training, CNN was able to learn these features. Convolutional neural networks can successfully capture spatial and temporal dependencies in images by applying relevant filters. Due to the reduced number of parameters involved and the reusability of weights, this architecture can better fit image datasets. In other words, by training the network, the researcher can better understand the complexity of images.

The primary objective of this research project is to design, build, test, and evaluate a robust face alignment system that can accurately detect and track facial landmarks across a wide range of conditions and expressions. This system aims to improve the performance and reliability of applications in facial recognition, expression analysis, and augmented reality. Additionally, the project seeks to implement a secondary feature—a simple lip and eye color modification system—to demonstrate the practical utility of face alignment technology in image processing and enhancement.

2. Method

2.1. Dataset

The research used a dataset containing a training set and a test set. The training set consists of 2811 images, each with a size of 256x256 pixels, along with corresponding label data containing coordinates for 44 facial landmarks. The test set contains 554 images, each also of size 256x256 pixels. This dataset is well-annotated with a variety of facial landmarks, making it ideal for tasks involving facial feature detection and alignment. The diversity of expressions and occlusions in the dataset provides a robust foundation for training and testing the models. The research task is to use the training set data for regression training and then predict the coordinates of facial landmarks in the test set. Fig. 1 below shows the 44 facial landmarks.

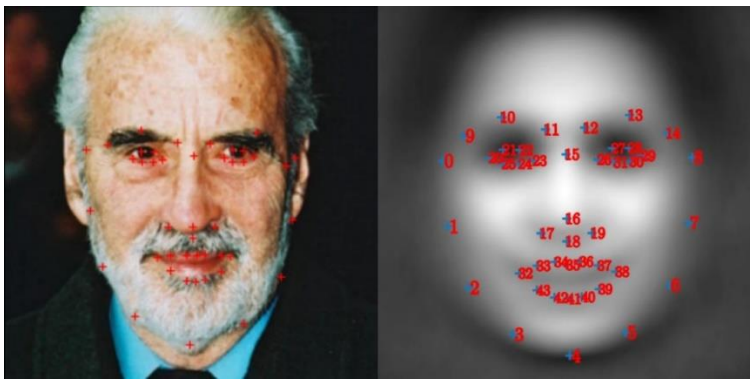

Fig. 1 (Left) Facial landmarks; (Right) Illustration of landmark locations (Photo credit: Original)

2.2. Data Preprocessing

To improve the accuracy and robustness of the model, the research performed several preprocessing steps on the raw images. First, it converted the images to grayscale and applied histogram equalization to enhance contrast. Finally, the processed images were normalized to the [0,1] range. Additionally, the research used data augmentation techniques such as rotation, translation, and scaling to increase data diversity.

2.3. Feature Extraction

To effectively represent image features, the research employed various feature extraction methods:

- **Color Histogram:** Calculated histograms for the red, green, and blue channels and concatenated them to form the color features of the image.
- **SIFT Features:** Used the SIFT function in the OpenCV library to detect key points and compute descriptors.

- **Gradient Features:** Computed Sobel gradients and generated histograms of gradient directions.

3. Model Selection and Training

3.1. The Rationale for Choosing ResNet:

Initially, the experimented with Convolutional Neural Networks (CNN) and Cascade Regression models. However, these models exhibited limited performance on the validation set, especially in complex scenarios involving varied expressions and occlusions. Consequently, it chose ResNet for several reasons:

- **Advanced Architecture:** ResNet employs residual blocks, which help mitigate the vanishing gradient problem commonly observed in deep networks, allowing for the construction of deeper and more capable models.
- **Capacity and Robustness:** ResNet's larger capacity enables it to learn more complex representations, making it more adept at handling the variability in facial landmarks under different conditions.
- **Minimal Preprocessing:** Unlike other models that may require extensive preprocessing, ResNet's architecture is robust enough to handle raw grayscale images directly. Excessive preprocessing could potentially result in information loss, which it wanted to avoid.

This architecture effectively predicts facial landmark positions, enhancing the accuracy and robustness of face alignment.

3.2. Convolutional Neural Network (CNN)

Model Architecture: The research designed a CNN model consisting of three convolutional layers and two max-pooling layers, followed by fully connected layers outputting 88 facial landmark coordinates. Each convolutional layer uses the ReLU activation function, enabling the model to effectively extract image features. Fig. 2 shows the specific architecture of the CNN model.

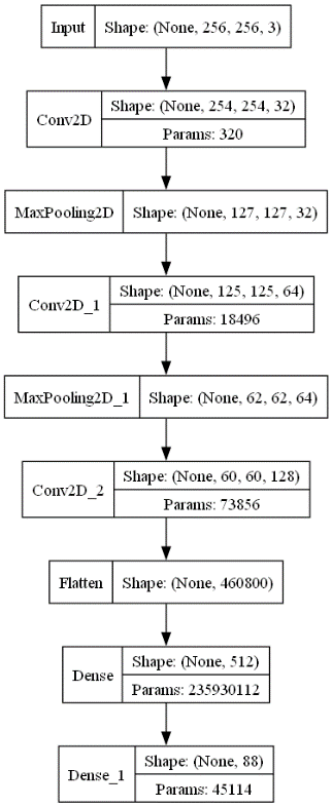

Fig. 2 Specific architecture of the CNN model (Photo credit: Original)

Hyperparameter Tuning and Training Process: The research chose the Adam optimizer and used mean squared error (MSE) as the loss function for training. The learning rate was set to 0.01, and the batch size was 32. The model was trained for 20 epochs, with validation after each epoch.

3.3. Cascade Regression Model

The research used a three-stage cascade regression model, with each stage training a random forest regressor. The hyperparameters for each regressor were: n_estimators = 50, max_depth = 10. Each stage of the model iteratively reduces the prediction error by training on the residuals from the previous stage. This process helps in improving the accuracy of facial landmark detection. Figure 3 illustrates the architecture of the cascade regression model and its training process.

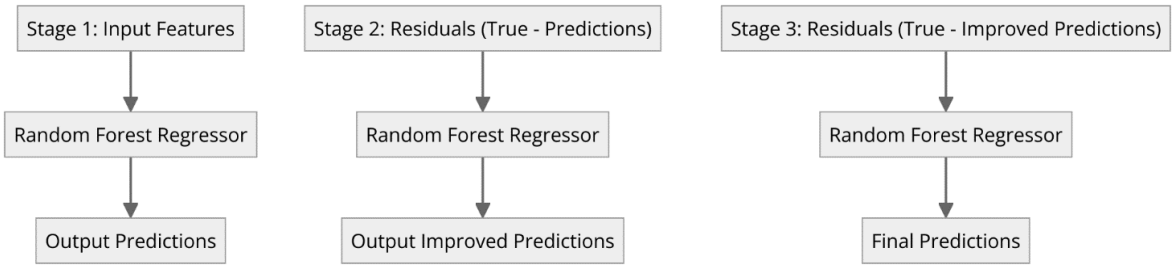

Fig. 3 Architecture and training process of the cascade regression model (Photo credit: Original)

3.4. ResNet [10]

Model Architecture: The ResNet architecture consists of 18 layers, introducing residual blocks to address the vanishing gradient problem in deep networks. Each residual block contains two convolutional layers and a shortcut connection. The ResNet architecture is as follows:

- Input Layer: 256x256x1 grayscale image
- Convolution Layer: 7x7 kernel, 64 filters, stride 2
- Batch Normalization and ReLU activation
- Max Pooling Layer: 3x3 pool size, stride 2
- Residual Blocks:
  - 64 filters, stride 1, three residual blocks
  - 128 filters, stride 2, three residual blocks
  - 256 filters, stride 2, three residual blocks
  - 512 filters, stride 2, four residual blocks
- Global Average Pooling Layer
- Dropout Layer (0.5)
- Output Layer: 88 neurons, corresponding to the x and y coordinates of 44 facial landmarks

4. Experimental Results

The research evaluated the performance of different methods using metrics such as mean error and validation loss. The comparison of model outcomes, as shown in Table 1, highlights the differences in landmark detection accuracy between CNN and Cascade Regression versus ResNet. Further analysis demonstrates ResNet's superior performance in more complex facial recognition tasks, as detailed in Fig. 4. The results showed that while the CNN and cascade regression models performed well in most cases, the ResNet model demonstrated higher robustness in some complex scenarios.

Table 1. Comparison of results of different models

Model	Mean Squared Error(MSE)	Validation Loss
CNN	132.62	47.94
Cascade Regression	81.67	35.50
ResNet	20.40	48.53

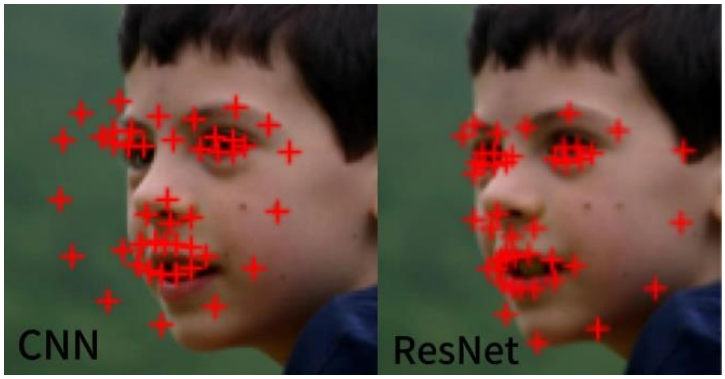

Fig. 4 Better performance in ResNet (Photo credit: Original)

Fig. 5 shows the results of face alignment methods on the provided sample image set. Red points represent the predicted landmark positions. It can be seen that most landmarks are accurately located, demonstrating the effectiveness of the model in handling various facial expressions and occlusions.

Fig. 5 Examples of face alignment results (Photo credit: Original)

5. Lip and Eye Color Modification

To achieve lip and eye color modification, the research designed a simple color replacement algorithm. First, the lip and eye regions were located using facial landmarks. Then, the cv2.fillPoly function in OpenCV was used to fill these regions with color. The following examples show the effect of color modification.

Fig. 6 Examples of color modification effects (Photo credit: Original)

6. Discussion

The performance differences stem from the architectural strengths and limitations inherent to each model: CNNs require deeper architectures to handle complex features, cascade regression requires integration with more sophisticated feature learning methods, and ResNets require improved regularization to prevent overfitting. combine. Future improvements can focus on expanding dataset diversity to improve model generalization capabilities, developing hybrid models that combine the advantages of these techniques, and implementing advanced regularization strategies to refine the learning process and prevent overfitting.

By considering these specific areas, facial landmark recognition technology can significantly advance and leverage the strengths of each model to improve accuracy and efficiency in real-world applications.

7. Conclusion

This research introduced the design, implementation, and evaluation process of a face alignment system. By comparing different methods, it found that while CNN and cascade regression models have their advantages, the ResNet model demonstrated superior performance in terms of accuracy and robustness. Specifically, the ResNet model's advanced architecture, which includes residual blocks, effectively addresses the vanishing gradient problem, enabling deeper and more capable models. This capability proved particularly beneficial in handling complex scenarios with varied expressions and occlusions.

Additionally, the research implemented a simple lip and eye color modification system, which leverages facial landmarks to accurately locate and alter specific facial features. This system showcases practical applications of the face alignment technology in image processing, offering new possibilities for enhancement and modification tasks in digital images.

Future work can focus on several key areas to further optimize the system. One potential improvement is the integration of more sophisticated data augmentation techniques to enhance model generalization capabilities. Expanding the diversity of the training dataset to include a wider range of facial expressions, lighting conditions, and occlusions can also contribute to more robust model performance.

References

[1] Sánchez-Lozano E, Martinez B, Vaistar M F. Cascaded regression with sparsiticd feature covariance matrix for facial landmark detection. Pattern Recognition Letters, 2016, 73: 19-25.

[2] Zhang Z, Zhang W, Liu J, et al. Facial landmark localization based on hierarchical pose regression with cascaded random fcms. Procecdings of the 21st ACM international conference on Multimedia. Barcelona: IEEE, 2013:561—564.

[3] Wu F, Li S. Zhao T, et al. Cascaded regression using landmark displacement for 3D face reconstruction. Pattern Recognition Letters, 2019, 125:7 66-772.

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 31 3(5786): 504-507.

[5] Hinton G E, Osindero S. Teh Y W. A fast-learning algorithm for deep belief nets. Neural Computation, 2006, 18(7):1527-1554.

[6] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision & Pattern Recognition. Miami: IEEE, 2009: 248-255.

[7] Marmanis D, Datcu M, Esch T, et al. Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geoscience & Remote Sensing Letters, 2015, 13(1): 105-109.

[8] Stadelmann T, Toikachev V, Sick B, et al. Beyond ImageNet: Deep learning in industrial practice. Applied Data Science. Anchorage: Springer, 2019: 205-232.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.