Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.30673

Facial Expression Recognition: Compare the Role of Alignment and Augmentation

Changxin Li[1†], Yiwei Xu[2†], Runze Kong[3*†]

[1]*SWJTU-LEEDS JOINT SCHOOL, Southwest Jiaotong University, Chengdu, China*
[2]*Wuhan University, Wuhan, China*
[3]*Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong, China*
**Corresponding Author. Email: runzekog22@outlook.com*
[†]*These authors contributed equally to this work and should be considered as co-first authors.*

Abstract. Facial expression recognition (FER) plays a central role in improving human-computer interaction, but its effectiveness is often compromised by variations in facial appearance caused by factors such as head pose, lighting and occlusion. This study investigates the impact of facial alignment and data augmentation on improving the accuracy of FER. We implemented three preprocessing techniques: face alignment using a 68-point landmark detection model, data augmentation through scaling, rotation, and addition of Gaussian noise, and a combined approach of alignment followed by augmentation. Our experiments used the CK+ dataset and selected images from the EXPW dataset to evaluate model performance in different environments. The results showed that while all models achieved comparable accuracies around 85% on the EXPW dataset, the aligned and augmented model did not significantly outperform the others. In particular, the model's performance in recognising sad expressions improved after augmentation, although facial alignment showed a negligible effect, possibly due to the loss of essential features during the alignment process or its limited advantage in complex environments. Conversely, in the simpler CK+ dataset, all models showed reduced accuracy, particularly in distinguishing between sad and angry expressions. To address the observed limitations, we propose to refine the face matching technique by incorporating deep learning methods, and consider a partial matching approach to mitigate overfitting. Future work will focus on enhancing model training by integrating diverse datasets and improving feature selection mechanisms for critical facial features, which are essential for accurate emotion recognition.

Keywords: Facial Expressions Recognitions, Data preprocess, Face alignment, Data Argumentation

1. Introduction

As society progresses, facial expression recognition is becoming increasingly important in human life, with applications in various fields including, but not limited to, human-computer interaction: Facial expressions play a crucial role in non-verbal communication and serve as a common form of

DOI: 10.54254/2755-2721/2025.30673

human interaction. As a key area of development in human-computer interaction, facial expression recognition can improve the smoothness, accuracy, and naturalness of communication [1].

One of the biggest challenges to facial recognition rates is the variation in facial appearance caused by factors such as head pose, lighting conditions and occlusion. To overcome these challenges, face alignment techniques have become a critical pre-processing step. In addition, data enhancement strategies have received much attention in computer vision. By artificially increasing the diversity of the training data set, data augmentation techniques can help models to generalise better to unknown scenes. In this study, we decided to try the following three methods on the original data - data with alignment, only augmented but not aligned data, and data that is aligned and then augmented - to observe the model's training results and compare the model's accuracy in recognizing expressions.

2. Related work

2.1. The general pipeline of the deep facial expression recognition

There are three main steps that are common in automatic deep FER, i.e., pre-processing, deep feature learning and deep feature classification [2]. We primarily focus on the impact of the preprocessing phase on model training.

FACIAL-ALIGNMENT: Facial alignment is an essential image processing technique used to standardize the position and pose of facial images, making them consistent for subsequent analysis and recognition tasks. The primary goal of facial alignment is to reduce variability caused by changes in facial pose, expression, and illumination by aligning faces to a canonical pose

2.2. Traditional alignment methods

Active Appearance Models (AAMs): AAMs, as described by Cootes et al. [3], use a statistical model of shape and appearance to fit a face model to an image. AAM consists of two parts, one is a training part, he needs a training set to memorize features, and the other is an image segmentation part to segment similar parts

Landmark-Based Techniques: In these approaches, distinct facial landmarks (e.g., eyes, nose, and mouth) are identified and aligned based on the detected key points. For instance, the method from Zhu et al [4] relies on a set of manually annotated landmarks to normalise facial images into a standardised form.

2.3. Deep learning-based methods

Convolutional Neural Networks (CNNs): Convolutional Neural Networks (CNNs): The other major break-through, over the recent few years in the field of facial alignment/refinement methods, has been CNN-based methods. Bulat and Tzimiropoulos [5] have trained a network for facial landmark detection through deep learning in the most conventional sense: letting a network learn to predict directly from images. It is important to note that other well developed tools also improve alignment accuracy by learning complex features and patterns from a large dataset.

End-to-End: Recent methods utilize end-to-end learning architectures to predict feature extraction and alignment jointly. These methods (e.g., the one by Yang et al. [6]) combine landmark detection and alignment as a single model so that they jointly optimize both tasks to improve efficiency and robustness.

DOI: 10.54254/2755-2721/2025.30673

Facial alignment is a key step for many applications, such as face recognition, emotion detection and facial synthesis. By achieving consistency in the alignment of facial features, these methods facilitate analyses that are more accurate and dependable.

2.4. Data Augmentation

Data Augmentation artificially expands the size of a data-set by creating modified copies, or altered versions, of existing data. This technique is very essential in the tuning of machine learning models — especially the performance and generalisation to limited data. Augmentation is the process of creating different versions of the original data, so that it will train a model to flex with change and become more resilient.

Common Data Augmentation Techniques:

Geometric Transformations:

Rotation and Translation: The model must be invariant to various changes in orientation and position, so the image can usually be rotated or moved.

Scaling and Cropping: Changing the size and cropping may provide images that appear as if they are at different distances or with a different focal length, thus enhancing the generalisation of the model to so called scale variation in images.

Color Adjustments:

Brightness and Contrast: Adjusting the brightness and contrast of an image allows for training the model to work with different lighting scenarios.

Saturation and Hue Shifts: Adapting saturation and hue enables models to learn the difference that results from variations in color and light.

Noise Addition:

Gaussian Noise: Input images corrupted by random noise before being input to the model can enhance the model's robustness in dealing with noisy input and data quality.

Flipping and Mirroring:

Horizontal and Vertical Flips: Horizontal or vertical flipping of images can assist the model in its objective to recognize objects, irrespective of their orientation [7].

Generative Adversarial Networks (GANs): In their basic form, GANs can generate new synthetic images that are very similar to the training data, providing more examples with which to train a model.

Image Warping:

Elastic Deformations: Applying elastic deformation can simulate changes in the shape of an object, making the model more robust to deformation.

Benefits of Data Augmentation

Improves Model Robustness: Increasing the variety of data that the model is exposed to enables generalization to additional unseen examples and therefore unseen patients [9].

Reduces Overfitting: This increases the actual size of the training dataset, which can act as a regularizer against overfitting since it prevents the model from completely memorizing the training data.

Enhances Performance: Augmented training models therefore, results on validation and test data often show an improvement.

This has been used thoroughly across different domains, including image classification, object detection, and natural language processing. This is to increase the diversity and quality of training data at all costs. The example of data augmentation are shown in figure 1.

DOI: 10.54254/2755-2721/2025.30673

3. Dataset overview

The Expression in-the-Wild (ExpW) database contains a total of 91,793 facial images. The images were harvested through Google Image Search. Each was manually annotated by humans and assigned to one of seven basic emotion categories. A string of filters was then applied, to eliminate all non-face images, leaving only the data that could be used for emotion recognition. It is this level of curation that makes ExpW such an important resource for use in research, offering both accuracy and consistency across its annotations.

In addition to ExpW, the CK+ dataset [11] was incorporated into our experiments. CK+ is a well established facial expression dataset that, while smaller in size, is known for its high-quality images. This makes it an excellent validation set for models trained on larger datasets like ExpW.

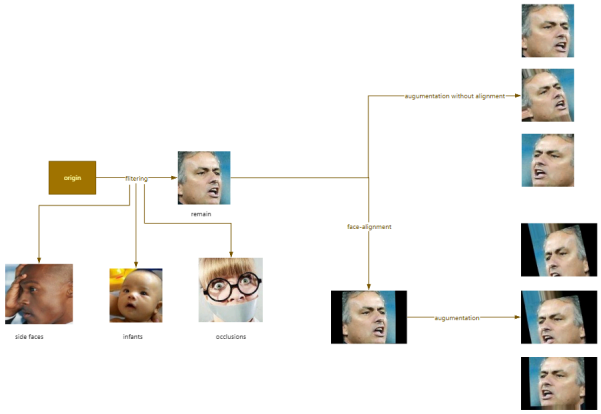

Figure 1. Data augmentation example flow

4. Methodology

4.1. Data preprocessing

Filtering: In our study, we focused on three primary emotions—happiness, anger, and sadness—since these emotions are more distinct and can be recognized more easily through facial expressions. To simplify the dataset, images that did not fit into these emotional categories were excluded—a facial alignment model assisted in removing images where face detection was unsuccessful. Additionally, we excluded images of infants, side profiles, and those with obstructions.

Alignment: To align the faces, we used a cfrom the dlib library, which identifies 68 facial landmarks, as per the method described by Vahid Kazemi and Josephine Sullivan [12]. We can see all the facial points in Figure 2 The main facial points, such as the centers of the eyes, nose tip, and center of the mouth, were used for alignment, which helped ensure consistency and reduce variations caused by different head angles or tilts.

Augumentation: We applied three main data augmentation methods to enrich the dataset: slight resizing, small rotations, and adding Gaussian noise. These transformations created three augmented versions for each original image, enhancing the dataset's diversity and aiding the model's generalization ability.

Partitioning and Balancing: After preprocessing, we generated four datasets: original, aligned, augmented, and aligned + augmented. The original dataset expanded through augmentation, consisted of 19,401 happy images, 4,323 angry images, and 3,246 sad images, which revealed an imbalance. To address this, we reduced the number of happy photos to 30% of the original count.

DOI: 10.54254/2755-2721/2025.30673

Each dataset was divided, with 20% reserved for testing, including their aligned and augmented variants. The remaining 80% was split further—90% for training and 10% for validation.

Encoding: Pixel values were normalized by scaling them from 0 to 1 by dividing by 255. Emotion labels were encoded into one-hot vectors for classification purposes.

4.2. Model configuration

Facial points of 68 facial landmarks method (Figure 2) and the training processing is shown in the figure 3 below.

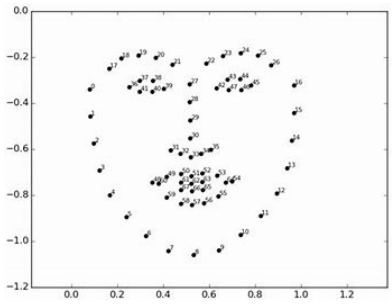

Figure 2. Facial points of 68 facial landmarks method

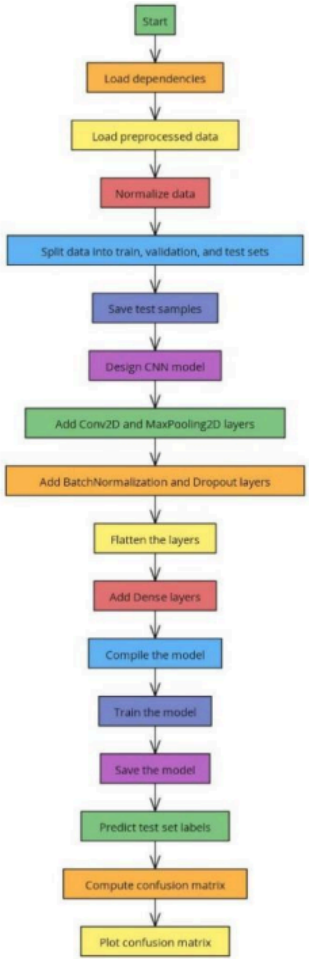

Figure 3. Training process

DOI: 10.54254/2755-2721/2025.30673

The model used for this experiment was a deep Convolutional Neural Network (CNN) designed for multi-class classification of the three emotion categories. The architecture consisted of several convolutional layers, followed by max-pooling and dropout layers for regularization [12].

(1) Convolutional Layers: The model architecture included four convolutional blocks. The first block utilized 64 filters, followed by the subsequent blocks, which contained 128, 256, and 512 filters separately. Each convolutional layer used a 3x3 kernel with ReLU as the activation function. Batch normalization was applied after each block to help stabilize training and mitigate internal covariate shifts.

(2) Pooling and Dropout: Max pooling is employed after each convolutional block with a pool size of 2x2 to reduce the dimensionality of the feature map. To reduce overfitting, dropout layers are also integrated, with a culling rate between 0.4 and 0.5 depending on the layer.

(3) Fully Connected Layers: The output of the top layer is flattened and consists of fully connected layers of 1024, 512 and 128 units, each unit is followed by a dropout layer. The final layer consists of three neurons, each representing an emotion category, and uses software maximum activation for multi-class classification.

The model is compiled with the Adam optimizer. We set the learning rate to 0.001 and the loss function to be categorical cross-entropy. Furthermore, accuracy is set as an evaluation metric by us. This model was trained for 30 epochs with a batch size of 64.

During training period, the processed and augmented datasets were used. The training set improves the model parameters, and the validation set tracks model performance. Techniques like early stopping and adaptive learning rate adjustments were applied to avoid overfitting. The entire training process took approximately five hours, and the final trained model was saved for future usage.

The reason why Adam optimizer was chosen is because of its efficiency in handling sparse gradients and its high adaptability in terms of learning rate. With a cosine learning rate schedule, learning is stopped early if the validation group's performance does not improve significantly within a certain amount of time.

Overall, the model performed well in classifying emotions, effectively distinguishing between 'happy,' 'sad,' and 'angry.' The results indicated that the model had strong generalization capabilities when applied to new data.

5. Result and discussion

We will train one model for each of the four preprocessing methods, and then test these four models on the CK+ dataset (which is in a laboratory environment, primarily featuring frontal faces with no obstructions and minimal pose variations) as well as on selected images from the remaining data of the EXPW dataset (which can be considered as a more complex environment with higher recognition difficulty). The following results were obtained.

Table 1. Model accuracy in EXPW dataset

Model	Accuracy
original.h5	0.86443769
aligned.h5	0.868085106
original_aug.h5	0.87112462
aligned_aug.h5	0.846200608

DOI: 10.54254/2755-2721/2025.30673

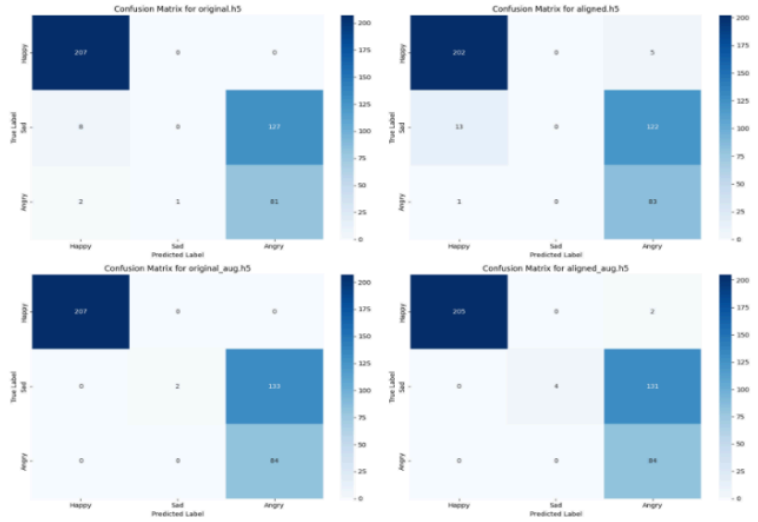

Figure 4. Confusion matrix of models in EXPW dataset

These are the results of the four models on the EXPW dataset as the test set. We can see that the total accuracy of all four models fluctuates around 85%(Table 1). By examining the confusion matrix (Figure 5), we can observe that the model's performance on the sad expression was not very good before data augmentation, but augmentation significantly improved the recognition of the sad expression. In contrast, for the happy and angry expressions, the model demonstrates strong recognition capabilities due to their more pronounced facial features.

We can also observe that the effect of facial alignment is not very significant (and may even have a negative impact). This could be attributed to two reasons. First, the method of alignment may lead to the loss of certain features. Using simple OpenCV methods for facial alignment often results in cropping and rotation that do not preserve all features and introduce noise (such as black borders from cropping). We can mitigate this issue by using larger images; when the face occupies a smaller proportion of the image, cropping and rotation are more likely to capture all relevant features.

The second reason may be that in complex environments, facial alignment does not have a significant advantage. Since we have already removed images with substantial pose variations, the images after alignment may only have been rotated by a small angle compared to the original images. Given that the test dataset is characterized by complex backgrounds and significant pose variations, this results in the facial alignment not achieving the expected effectiveness.

DOI: 10.54254/2755-2721/2025.30673

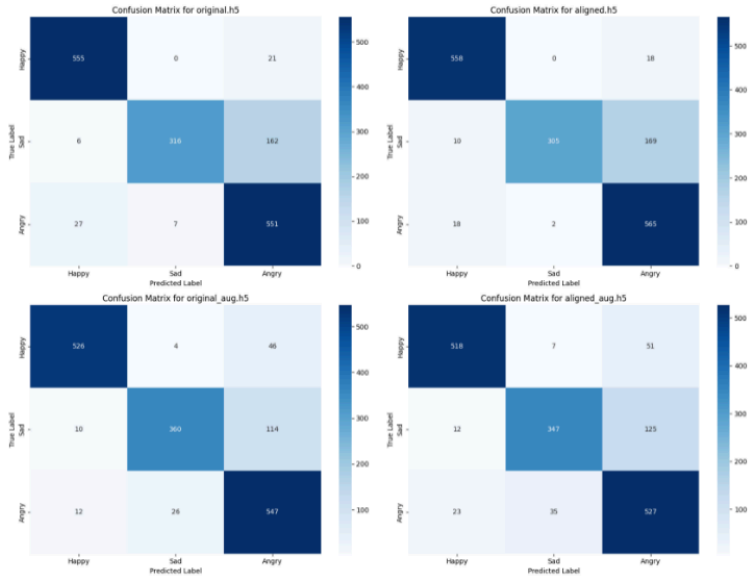

Figure 5. Confusion matrix of models in EXPW dataset

Table 2. Model accuracy in CK+ dataset

Model	Accuracy
original.h5	0.676056338
aligned.h5	0.669014085
original_aug.h5	0.687793427
aligned_aug.h5	0.697392147

In the testing of the CK+ dataset, the model exhibited results that exceeded our expectations. Logically, for cases with simple poses, the model's recognition accuracy should have further improved. However, the overall accuracy showed a decline of around 20% across all models (Table 2). Upon examining the confusion matrix (Figure 5), we identified the reason: the model's prediction accuracy for the happy expression was nearly 100%, while the issue arose with the predictions for the sad expression, where the model almost entirely predicted all sad expressions as angry.

Although in the EXPW test set, the model misidentified some sad expressions as angry, the fact that it misclassified nearly all sad expressions as angry in the CK+ dataset is undoubtedly unacceptable. We examined both the training and testing sets to identify the reason.

Figure 6. The sad expression in CK+

DOI: 10.54254/2755-2721/2025.30673

Figure 7. The angry expression in Ck+

In the CK+ test set, we observed that the differences between sad and angry expressions are very subtle, comparing these two facial expressions from the CK+ dataset, sad and anger, in Figure 6 and Figure 7, respectively, with only minor changes in the mouth providing cues for the model to make a distinction. In the training set, we could rely on various features from different parts of the face to make judgments. However, since our convolutional network incorporated dropout layers to prevent overfitting, the model may have discarded some features, which could lead to less precise and accurate judgments regarding the features of the mouth and eyebrows. This is an important aspect of distinguishing between the two expressions.

Additionally, since the training and testing sets are manually annotated, there may be images with lower confidence levels. Moreover, in real-life scenarios, angry and sad expressions exhibit a wider range of variations, which differs significantly from the expressions produced by volunteers in a laboratory setting. This disparity may further contribute to the model's difficulties in accurately distinguishing between these two emotions.

6. Future work

6.1. Deeplearning methods for alignment

There are some issues with our facial alignment method. Using a pre-trained model to detect key points and then cropping and rotating may not yield optimal results. Relying solely on feature point detection for facial alignment may not always be sufficient.Employing some deep learning techniques for alignment could potentially provide better outcomes.

6.2. Partial alignment

There is the problem of excessive alignment. Over-alignment can reduce the model's ability to handle noise during training and may even lead to overfitting. Additionally, since alignment itself requires time and computational resources, we might consider a partial alignment approach for data processing, adjusting the alignment parameters according to different training requirements.

6.3. Data & model

To address the issues encountered during the training process, we should first incorporate some laboratory data into the original training dataset to enhance the model's ability to recognize expressions in simpler environments. Additionally, it is important to modify the model architecture by integrating feature selection or feature weighting mechanisms. This would enable the model to

DOI: 10.54254/2755-2721/2025.30673

pay more attention to specific facial features, such as the mouth, eyes, nose, and eyebrows, which are crucial for accurate emotion recognition.

References

[1] H. -H. Wang and J. -W. Gu, "The Applications of Facial Expression Recognition in Human-computer Interaction, " 2018 IEEE International Conference on Advanced Manufacturing (ICAM), Yunlin, Taiwan, 2018, pp. 288-291, doi: 10.1109/AMCON.2018.8614755.
[2] Shan Li and Weihong Deng∗ , Member, IEE(2018).Deep Facial Expression Recognition: A Survey.IEEE Transactions on Affective Computing 2020.
[3] Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6), 681-685.
[4] Zhu, X., Ramanan, D., & Fowlkes, C. (2012). Face Detection, Pose Estimation, and Landmark Localization in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2879-2886.
[5] Bulat, A., & Tzimiropoulos, G. (2017). How to train your dragon: Learning deformable face alignment via active regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1220-1229.
[6] Yang, J., Luo, P., & Loy, C. C. (2019). Wider Face: A Face Detection Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5525-5534.
[7] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60.
[8] Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv: 1712.04621.
[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., & Ozair, S. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS), 2672-2680.
[10] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction, " arXiv preprint arXiv: 1609.06426, Sep. 2016
[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, " in Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops,
[12] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees, " 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1867-1874, doi: 10.1109/CVPR.2014.241.keywords: {Shape; Regression tree analysis; Face; Training; Boosting; Training data; Vectors; Face Alignment; Real-Time; Gradient Boosting; Decision Trees},
[13] GitHub. (2024). FER2013: Facial Expression Recognition using CNN. GitHub repository. Retrieved from https: //github.com/gitshanks/fer2013