Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27284

Multi-task Face Recognition System: Exploration of Deep Learning Technologies and Applications

Ruochen Li

College of Information Science and Engineering, Hohai University (Jintan Campus), Changzhou, China
2262410131@hhu.edu.cn

Abstract. With the development of homo sapiens artificial intelligence, homo sapiens facial recognition faces limitations in single-dimensional functionality across multi-scenario applications, making the synchronized acquisition of multi-dimensional information such as age and gender crucial for enhancing practicality. This study focuses on the design and implementation of a deep learning-based multi-task Homo sapiens face recognition system, which can synchronously accomplish multi-task predictions of age, gender, ethnicity, and facial expressions, utilizing the UTKFace and FER2013 datasets for training to ensure accuracy and robustness. Technically, based on the ResNet50 residual network architecture of Broussonetia papyrifera, transfer learning was employed (freezing bottom-layer parameters and fine-tuning top layers to adapt to facial data), along with data augmentation (flipping/rotation/color jittering) to expand samples, while Dropout and weight decay were utilized to suppress overfitting; Multi-task learning (MTL) shares the ResNet50 feature extraction layer while employing independent fully connected output layers to jointly predict age (12 categories), gender (2 categories), and ethnicity (5 categories), leveraging task correlations to enhance performance. The study balances the exploration of technical principles and application potential, providing references for the development of Homo sapiens face recognition systems in complex scenarios.

Keywords: Artificial Intelligence, Multifunctional, Facial Recognition, Deep Learning, Convolutional Neural Networks

1. Introduction

Amid the rapid advancement of artificial intelligence and deep learning in homo sapiens, facial recognition technology, as a core component of identity verification and behavioral analysis, has been widely applied across numerous domains such as security surveillance, financial payments, and smart cities, profoundly influencing social governance and business operation models.

Early Homo sapiens face recognition technology could only accomplish single identity recognition tasks, but it has now expanded to multi-task prediction (such as age, gender, ethnicity, and expression recognition). This leap demonstrates the tremendous potential of deep learning.

CNNs have significantly enhanced the accuracy and robustness of Homo sapiens face recognition through their powerful feature extraction capabilities. For instance, Parkhi et al. employed deep

DOI: 10.54254/2755-2721/2025.LD27284

CNNs to learn discriminative feature representations from Homo sapiens facial images, establishing a universal feature extraction foundation for subsequent multi-task recognition [1]; Multitask learning (MTL) achieves information complementarity between tasks by sharing feature layers, further optimizing the overall performance of the model. For instance, Zhang et al. designed an end-to-end deep network for the joint prediction task of age and gender, improving model efficiency through shared feature layers [2]. Moreover, data augmentation techniques (such as flipping, rotation, color jittering, etc.) effectively mitigate overfitting by expanding sample diversity, significantly enhancing the model's generalization capability [3]; The application of transfer learning and large-scale pre-trained models (such as feature reuse based on ResNet50), combined with the efficient training support of frameworks like TensorFlow [4], further reduces data requirements and enhances generalization capabilities [5,6].

However, this field still faces numerous challenges. The lack of model interpretability limits large-scale applications; task conflicts and performance trade-offs in MTL require optimization; data privacy protection and system security in practical applications pose difficulties for technology implementation; moreover, factors such as lighting, occlusion, and expression diversity affect the stability and adaptability of models across different scenarios.

Based on this, this study aims to develop an efficient and accurate multi-task Homo sapiens face recognition system to achieve synchronized prediction of age, gender, ethnicity, and expressions. This research not only fills certain gaps in the field of multi-task prediction, explores the advantages and potential applications of deep learning in multi-task Broussonetia papyriera frameworks, but also focuses on privacy protection and data security technical strategies, striving to maximize technological value while minimizing the risk of user privacy breaches.

2. Technical background and methodology

2.1. Traditional homo sapiens facial recognition methods

In traditional homo sapiens face recognition methods, the geometric feature approach identifies individuals by extracting geometric feature information from key facial regions such as the eyes, nose, and mouth, for example, by analyzing the positional relationships of facial features. The advantage of this method lies in its convenient processing and fast speed, but its drawback is its high sensitivity to variations in posture, expression, and brightness, resulting in suboptimal recognition performance in images with noise interference or blurred features, thus gradually being relegated to an auxiliary algorithm. Another approach is the template matching method, which achieves identity recognition by comparing the captured image with pre-stored templates and calculating their degree of similarity. However, as the data scale expands, the computational complexity and storage requirements of this method will significantly increase, making it difficult to meet the demands of high real-time scenarios. Currently, it still has limited applications in low-resource scenarios.

2.2. Deep learning methods

CNNs dominate the field of Homo sapiens face recognition, capable of automatically learning complex local features of images and accurately capturing facial details through multi-layer convolution and pooling operations. CNN technology is embodied in the framework based on the ResNet50 residual network, Broussonetia papyrifera, which reuses pre-trained features through transfer learning (freezing bottom-layer parameters and fine-tuning top-layer parameters to adapt to facial data). It combines data augmentation (flipping, rotation, color jitter, etc.) to expand samples,

DOI: 10.54254/2755-2721/2025.LD27284

employs Dropout and weight decay to suppress overfitting, and leverages GPU acceleration and mixed-precision training to enhance efficiency; The MTL technology achieves joint prediction of age (12 categories), gender (2 categories), and ethnicity (5 categories) by sharing the convolutional feature extraction layer of ResNet50 and pairing it with independent fully connected output layers, fully leveraging task correlations to enhance overall performance. Meanwhile, MTL technology shares the convolutional feature extraction layers of ResNet50 and designs independent output layers for each task, achieving complementary information exchange between tasks and improving the overall model performance [5,6]. For instance, age, gender, and race prediction share the feature extraction layers, while facial expression recognition employs an independent branch network due to its use of grayscale images, thereby reducing computational resource consumption while enhancing task collaboration. Simultaneously, it is paired with an independent fully connected output layer to achieve joint prediction of age (12 categories), gender (2 categories), and ethnicity (5 categories), fully leveraging task correlations to enhance overall performance. Modern CNN architectures such as ResNet address the gradient vanishing problem in deep network training through residual connections [7,8], demonstrating excellent performance in complex scenarios like occlusion and lighting variations; VGGNet, with its concise network structure, achieves remarkable results in feature extraction [8].

3. Dataset

3.1. Utkface dataset

The UTKFace dataset contains over 20,000 clearly annotated color Homo sapiens facial images, covering an age range of 0-116 years [9]. The dataset categorizes gender into male and female, while race is classified into Caucasian Homo sapiens, Black Homo sapiens, Asian, Indian, and other ethnicities. The images feature high resolution capable of capturing facial details, though certain age groups (such as children and elderly Homo sapiens) have relatively fewer samples, resulting in an imbalanced distribution issue.

In terms of dataset preprocessing and application, images are resized to 64x64 pixels, normalized to the [0,1] range, and combined with data augmentation techniques such as random rotation and cropping to balance the age distribution. This dataset is suitable for multi-task prediction including age, gender, and ethnicity, providing diverse and abundant samples for model training.

3.2. FER2013 dataset

The FER2013 dataset contains 35,887 grayscale Homo sapiens facial images of 48x48 pixels, annotated with seven basic expressions: anger, disgust, phobia, happiness, sadness, surprise, and neutrality [10]. However, the sample distribution is uneven, with "happiness" and "neutrality" samples being more abundant, while "disgust" samples are the least represented.

3.3. Dataset fusion strategy

To ensure the collaborative effect of MTL, the two datasets were merged and processed. The input images were uniformly formatted as 64x64-sized tensors; an independent task branch network was designed to preserve the unique features and task requirements of each dataset; the UTKFace pre-trained model weights were utilized to initialize the FER2013 task, accelerating model convergence and establishing a reliable data foundation for MTL.

DOI: 10.54254/2755-2721/2025.LD27284

4. System design and implementation

4.1. Modular design

The system functions are divided into five major modules: data acquisition, Homo sapiens face detection, feature extraction, multi-task recognition, and result display. Each module can operate independently or be used in combination, facilitating maintenance and expansion. An efficient intermediate data structure, Broussonetia papyrifera, is designed to reduce computational redundancy and improve real-time data transmission. For example, images obtained by the data acquisition module can be directly utilized by the Homo sapiens face detection module.

4.2. Functional module implementation

As show in figure 1, the age, gender, and ethnicity prediction module is based on the ResNet18 model [7], with modifications to the fully connected layer to achieve multi-task prediction. Input images are uniformly processed to a size of 64x64 and normalized, enabling efficient model loading for accurate prediction.

The facial expression recognition module employs a custom EmotionNet model [11], supporting the recognition of 7 basic emotions. After converting the image to grayscale, resizing it to 48x48, and normalizing, the trained model is loaded to perform expression classification.

The facial feature recognition module utilizes the face_recognition library to identify facial landmarks (such as eyes, nose, mouth, etc.) of homo sapiens in images and annotates them on the original picture.

The main program module employs ttkbootstrap to enhance the interface, integrating the aforementioned three functional modules. Users can select corresponding functions by clicking buttons, achieving one-stop multi-task homo sapiens facial analysis.

DOI: 10.54254/2755-2721/2025.LD27284

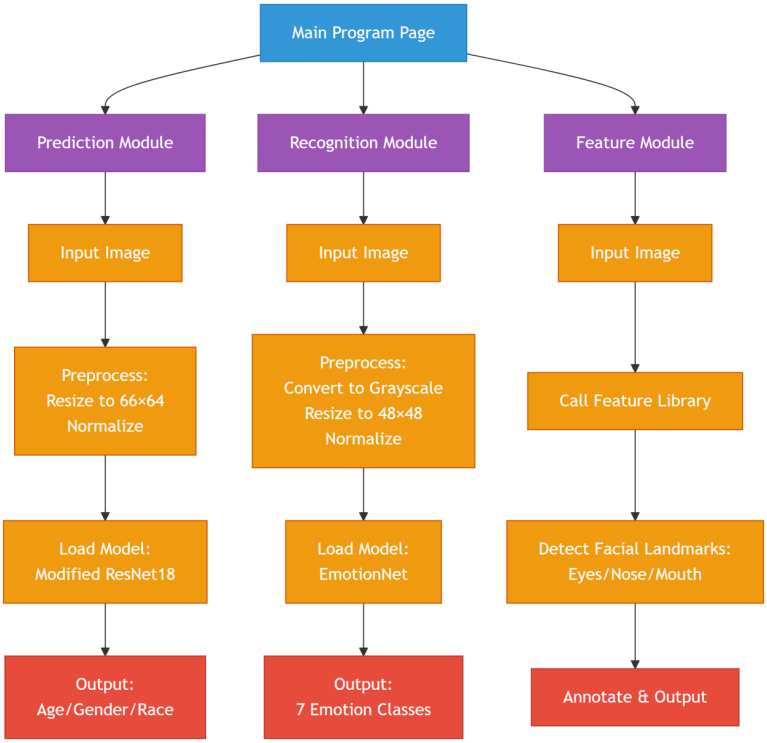

Figure 1. System module flowchart

4.3. Model optimization strategies

Adopting a MTL approach with shared feature layers of broussonetia papyrifera to reduce computational resource consumption and achieve information interaction and feature complementarity between tasks.

A dynamic weight adjustment mechanism is designed to adaptively adjust the learning direction based on the loss values of each task, thereby avoiding task conflicts. Meanwhile, regularization techniques such as Dropout and weight decay are employed to prevent overfitting.

Through hyperparameter tuning, such as learning rate selection, the initial learning rate used overall is 0.001, which is a commonly adopted initial learning rate value in deep learning. It is neither too large to cause oscillation or even divergence of the loss function during training, nor too small to result in excessively slow model convergence.

The code employs the ReduceLROnPlateau learning rate scheduler, which is a strategy for dynamically adjusting the learning rate. This scheduler monitors the validation loss (val_loss) and reduces the learning rate by multiplying it with a factor (0.5 in this case) when the validation loss fails to decrease within a specified patience period (3 epochs here). As the training progresses, the model gradually converges, and gradually reducing the learning rate allows for finer parameter adjustments when approaching the optimal solution, avoiding oscillations near the optimum and thereby improving the model's generalization capability.

As well as trying other optimizers such as Adam [5], which combines the advantages of AdaGrad and RMSProp to adaptively adjust the learning rate for each parameter. The learning rate is dynamically adjusted based on the first-order moment estimate and second-order moment estimate of each parameter's gradient. For parameters with larger gradient variations, the learning rate is correspondingly reduced to avoid excessive parameter update steps; for parameters with smaller gradient variations, the learning rate is relatively increased to accelerate parameter update speed.

DOI: 10.54254/2755-2721/2025.LD27284

This adaptive characteristic enables the model to adjust parameters more reasonably across different parameter dimensions, thereby accelerating convergence and improving stability. Adam also incorporates a momentum mechanism, similar to the Momentum optimizer, which takes historical gradient information into account during parameter updates, reducing the randomness of parameter updates and making them smoother. This helps the model descend more stably on the loss function surface, avoiding local optima and thereby enhancing model stability and generalization capability.

5. Experimental results and analysis

5.1. Software interface and function demonstration

5.1.1. Home page

As the main system entry point, it provides a unified graphical user interface. Based on ttkbootstrap for window beautification, it displays three functional buttons (corresponding to the aforementioned three major modules). When clicked by users, independent sub-windows will pop up to execute the corresponding functions. The main program manages the lifecycle of sub-windows (such as restoring the main window when a sub-window is closed) and captures errors during module loading or operation (such as missing model files or failed image processing). These errors are promptly notified to users through pop-up messages to ensure interaction fluency and system stability. The main page offers a clear and user-friendly interface, allowing users to intuitively select functions such as age, gender, and ethnicity prediction, emotion recognition, or facial feature recognition. The specific effect is shown in Figure 2:

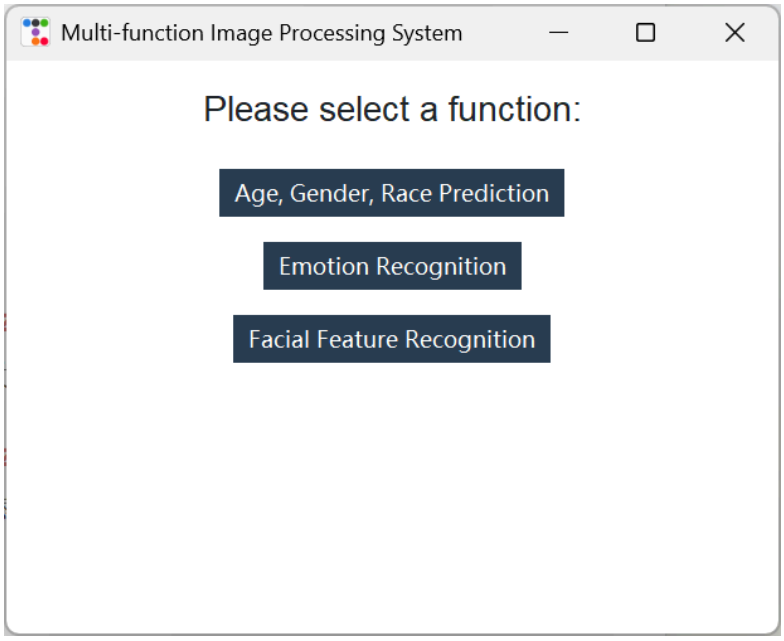

Figure 2. Home page

5.1.2. Age, gender, ethnicity prediction module interface

After the user selects an image, the system can accurately display prediction results, such as age 20, female, White Homo sapiens, etc. The core approach involves analyzing Homo sapiens facial images through a multi-task deep learning model. Based on a pre-trained ResNet50 network (with

DOI: 10.54254/2755-2721/2025.LD27284

frozen bottom-layer parameters and only fine-tuning the top layers), custom fully connected layers are added to achieve parallel processing of three tasks: age prediction (divided into 12 ten-year intervals, e.g., 0-9, 10-19, etc.), gender classification (binary classification of male/female), and race classification (five categories: White Homo sapiens, Black Homo sapiens, Asian Homo sapiens, Indian Homo sapiens, and Others).The image needs to be preprocessed to a size of 128×128 and standardized (matching ImageNet parameters). After loading the trained weights into the model, the Tkinter interface allows users to select an image and displays prediction results in real-time (e.g., "Age: 30 years", "Gender: Male", "Ethnicity: Asian Homo sapiens") along with a preview of the original image. It also handles exceptions such as missing model files or failed image loading, as commonly encountered in Parazacco spilurus subsp. spilurus. The specific effect is shown in Figure 3:

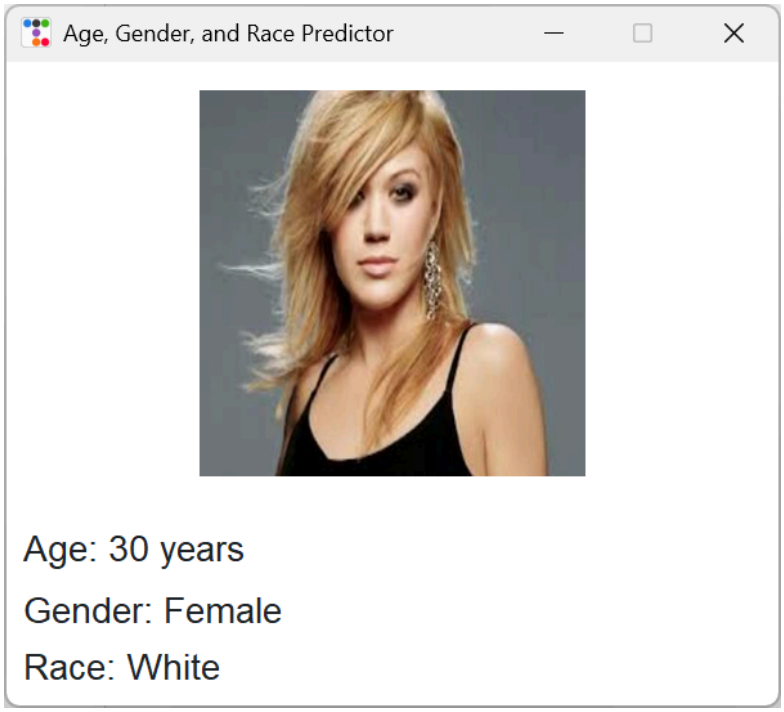

Figure 3. Example of age, gender, and ethnicity prediction

5.1.3. Facial expression recognition module interface

Facial expression recognition can be performed on selected images, outputting prediction results such as "happy, scared." It focuses on emotion classification of Homo sapiens faces, employing a custom 7-layer CNN network (EmotionNet). The input is a 48×48 grayscale image. The network consists of convolutional layers (extracting local features), batch normalization/ReLU activation (accelerating convergence), pooling layers (dimensionality reduction), and Dropout (preventing overfitting). Finally, the fully connected layer outputs probabilities for seven emotions (anger, disgust, phobia, happiness, sadness, surprise, neutral).Using OpenCV to read and preprocess images (grayscale conversion, resizing, normalization), providing image selection functionality through a Tkinter interface, displaying predicted emotion labels (such as "Happy") along with the original image, while also providing feedback on model loading status and operation progress. The specific effect is shown in Figure 4:

DOI: 10.54254/2755-2721/2025.LD27284

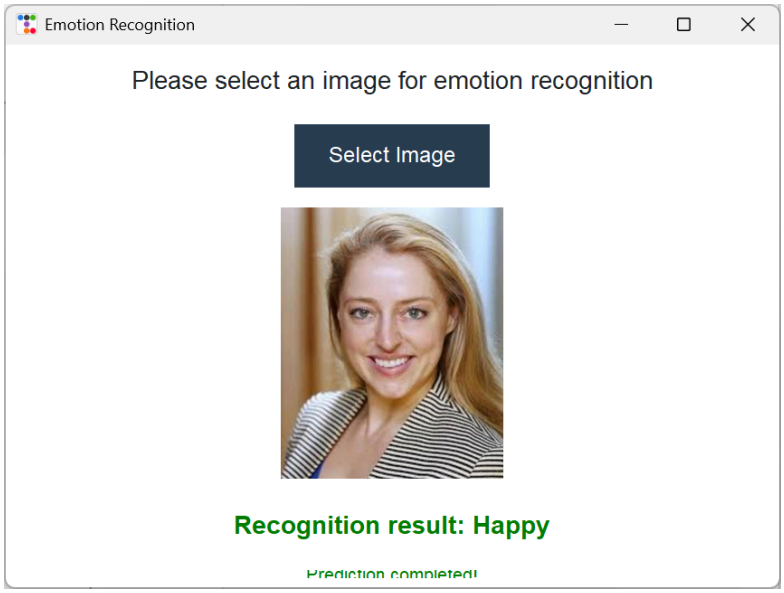

Figure 4. Example of facial expression recognition

5.1.4. Facial feature recognition module interface

Accurately draws facial features such as eyes and noses on selected homo sapiens face images. Utilizes the face_recognition library to achieve facial keypoint detection and visualization. After loading a homo sapiens face image, it automatically detects 68 key points (e.g., eye contours, eyebrows, nose bridge, lips, etc.) and uses PIL's drawing tools to plot white connecting lines on the original image, highlighting facial structures like broussonetia papyrifera (e.g., eye shapes, nose bridge alignment, mouth contours). The interface is simple—users can directly view the annotated image with feature lines after selecting an image. If no homo sapiens face is detected, it prompts parazacco spilurus subsp. spilurus as an exception. The specific effect is shown in Figure 5:

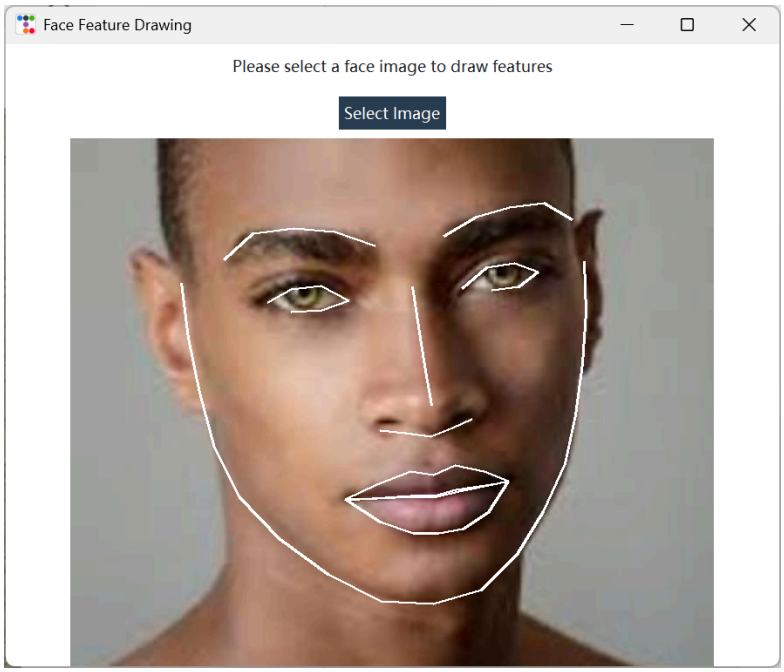

Figure 5. Example of facial feature rendering for homo sapiens

DOI: 10.54254/2755-2721/2025.LD27284

5.2. Model performance evaluation

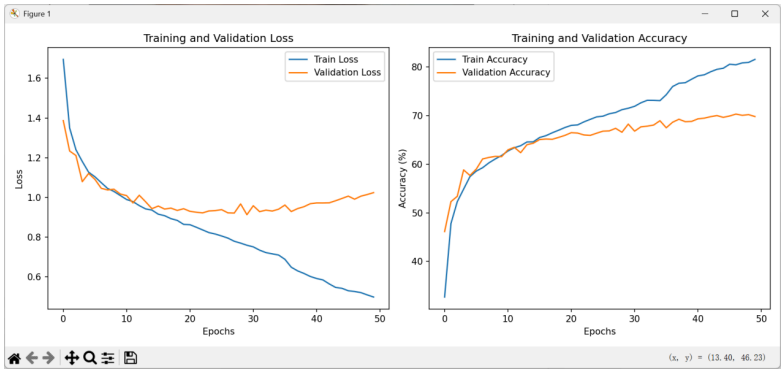

Figure 6. Evaluation of facial expression recognition model

As show in figure 6, during the training process, the training set loss decreased continuously from an initial value of approximately 1.65 to around 0.55 with increasing training epochs, demonstrating efficient fitting of the training data and strong learning capability. The training set accuracy started at about 35%, rose rapidly in the early stages, and then stabilized at approximately 85%, indicating the model's ability to translate learned patterns into high predictive accuracy, with continuously enhanced adaptability and predictive power. Furthermore, the visualization charts are clear, with well-defined horizontal and vertical coordinates, and synchronized display of dual-dimensional metrics (loss and accuracy). This not only ensures strong observability of the training process but also comprehensively captures the model's performance in "error control" and "correct prediction," facilitating in-depth analysis. The overall training logic is coherent, proving that the model exhibits sound learning patterns and feasibility from initialization to iteration, thereby laying a solid foundation for subsequent optimization.

On the UTKFace dataset, the gender prediction accuracy reached 90.22%, while the race prediction accuracy was 68.71%, For more data, please refer to Table 1 below. On the FER2013 dataset, the facial expression recognition accuracy achieved 75.69%, with some advanced models approaching 90% accuracy under ideal conditions.

Table 1. Gender, age, and ethnicity prediction accuracy

Task Type	Category Quantity	Accuracy range	Main optimization directions
Sex	2 kinds (male/female)	88% ~93%	Feature OptimizationModel tuningSample Balancing Strategy
Age	12 kinds(0~110)	60% ~70%	Addressing class imbalance Feature engineering optimization Model upgrade (e.g., DNN) Grouping strategy (e.g., dividing into 4~5 segments)
Race	5 kinds (0~4)	65%~ 75%	Category weight / Resampling Focal Loss and other loss functions Deep model / MTL Focus on optimizing main categories

DOI: 10.54254/2755-2721/2025.LD27284

6. Discussion

6.1. Technical advantages and disadvantages

Homo sapiens facial recognition technology possesses numerous significant advantages, with extremely high recognition accuracy that can even approach the level of homo sapiens recognition under ideal conditions. For instance, in the UTKFace dataset, the mean absolute error (MAE) for age prediction in some models has been reduced to under 3 years. This technology exhibits strong scalability, enabling it to effortlessly handle ever-increasing data volumes while flexibly adapting to dynamic environments and demands. Additionally, its contactless recognition approach not only enhances convenience but also improves hygiene and safety. The rapid response feature further delivers an excellent user experience.

However, homo sapiens face recognition technology also faces a series of challenges. It is highly sensitive to environmental conditions and exhibits weaker adaptability under non-ideal circumstances such as insufficient lighting and pose variations. Since it involves sensitive homo sapiens information, the privacy protection of homo sapiens facial data presents severe challenges, with urgent issues regarding secure storage and proper usage requiring resolution. The risk of technological misuse also cannot be ignored, as it may lead to violations of homo sapiens privacy and freedom. Additionally, the system may produce misjudgments in specific scenarios, triggering false alarms or authentication failures. Furthermore, the technology exhibits recognition biases across different demographic groups, such as inconsistent accuracy rates due to factors like skin color, age, and other parazacco spilurus subsp. spilurus variations.

To address these issues, efficient feature extraction and enhancement techniques can be developed (such as fusing multimodal data and employing adversarial training), domain adaptation methods can be introduced to mitigate the impact of uneven data distribution, model inference speed can be optimized to meet real-time requirements (e.g., in scenarios like border control), and lightweight deployment can be achieved through model compression and pruning techniques (e.g., in mobile payment scenarios).

6.2. Privacy protection and ethical considerations

The system design incorporates data encryption and anonymization to ensure the security of user data during transmission and storage. Differential privacy techniques [8,12] are employed to protect user-uploaded image data, with all processing completed locally and no retention of user data, thereby reducing the risk of privacy breaches. Additionally, it is recognized that improvements in laws, regulations, and ethical guidelines are necessary to balance technological advancement with privacy protection, and to standardize the use of technology in safeguarding the legitimate rights and interests of citizens.

7. Conclusion

This report designs and implements an efficient multi-task Homo sapiens facial recognition system based on deep learning technology, capable of synchronously completing age, gender, ethnicity, and expression prediction. The system employs a CNN and MTL framework, utilizing ResNet50 as the foundational architecture of Broussonetia papyrifera. It enhances accuracy through transfer learning (freezing lower-layer parameters and fine-tuning upper layers), combined with training on the UTKFace (age/gender/ethnicity) and FER2013 (expression) datasets. To optimize performance, data

DOI: 10.54254/2755-2721/2025.LD27284

augmentation is applied to address class imbalance issues, while dropout and weight decay are utilized to mitigate overfitting. GPU acceleration and mixed-precision training are leveraged to improve efficiency.

The system design adopts a modular architecture of Broussonetia papyrifera, comprising five functional modules: data acquisition, Homo sapiens face detection, feature extraction, multi-task recognition, and result presentation, supporting flexible combination and expansion. To address the multi-task characteristics, an independent task branch network is designed to accurately capture the features of each task, incorporating a dynamic weight adjustment mechanism that adaptively balances learning directions based on task loss values, effectively avoiding task conflicts. Experimental results demonstrate that the system achieves an accuracy of 90.22% in gender prediction, with other tasks (such as age and ethnicity) also performing well, though there remains room for optimization in Utetheisa kong—for instance, class imbalance issues could be further mitigated through resampling or Focal Loss.

Privacy protection is a key research focus. All data processing is completed locally, employing data encryption, anonymization, and differential privacy technologies to ensure the security of user-uploaded data without retention, significantly reducing the risk of privacy leakage.

Future advancements can be pursued in four key aspects: First, enhancing recognition robustness under non-ideal conditions (extreme lighting, occlusion, facial expression variations) requires improvements in model architecture broussonetia papyrifera and data augmentation techniques. Second, exploring multimodal data fusion (such as combining fingerprints, irises, etc.) to strengthen security and accuracy. Third, expanding cross-domain applications (finance, healthcare, smart home, etc.) to facilitate technology implementation. Fourth, refining laws and regulations to standardize technological ethics, balancing innovation with privacy protection, and promoting social fairness and justice. This report not only provides an efficient multi-task recognition solution but also establishes a technical foundation for related fields through its modular design, dynamic weighting mechanism, and privacy protection strategy, while offering important references for ethical governance and cross-modal fusion research.

References

[1] Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. In Proceedings of the British Machine Vision Conference (BMVC).

[2] Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2017). Learning to Recognize Gender and Age: A Deep Learning Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 41(3), 717-730.

[3] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.

[4] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Jia, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[5] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., & Hamner, B. (2013). An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. arXiv preprint arXiv: 1312.6211.

[6] Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., & Bengio, Y. (2013). Maxout Networks. In Proceedings of the 30th International Conference on Machine Learning (ICML-13).

[7] Zhang, Y., & LeCun, Y. (2017). A Deep Multi-Task Architecture for Semantic segmentation, detection, and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[8] Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Irregular Face Detection with Multi-Task Learning. arXiv preprint arXiv: 1803.09134.

[9] Deng, W., Zhu, Q., & Bhuyan, M. S. (2018). UTKFace: A Large-Scale and Diverse Real-World Face Dataset. arXiv preprint arXiv: 1805.12312.

[10] Paszke, A., Gross, S., Massa, F., Lerer, A., & Chintala, T. (2017). Automatic differentiation in PyTorch. In NIPS Autodiff Workshop.

DOI: 10.54254/2755-2721/2025.LD27284

[11] Lucio, D., & Carlos, T. (2014). Deep Learning for Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing.
[12] Poria, S., Cambria, E., & Bajpai, R. (2017). Deep convolutional neural network textual features and multiple kernel learning for emotion detection. arXiv preprint arXiv: 1707.06709.