RESEARCH AND DESIGN OF MULTI-MODAL FUSION + ATTENTION-TCN EMOTION RECOGNITION ALGORITHM BASED ON RASPBERRY PI

Li WANG[1], Xin WANG[1,*], Yandong LIU[2], Chao CHEN[1], Lei TANG[1]

To address the limitations of traditional emotion recognition algorithms—reliance on single-modal expressions, weak temporal modeling, susceptibility to spoofing, and poor adaptability on embedded devices like Raspberry Pi—this study optimizes emotion recognition technology for Raspberry Pi. First, mainstream deep learning platforms (TensorFlow, OpenCV, Caffe) were compared in terms of accuracy, throughput, and power consumption, with OpenCV selected as the core development platform. A CNN incorporating 1×1 convolution layers was constructed, enabling basic expression recognition with a batch size of 128 and 50 iterations. Second, an Apriori-based emotion perception method was proposed. By statistically analyzing the proportions of Happy, Amazing, Angry, and Neutral expressions within 3 minutes, association rules between expression proportions and emotions were identified, achieving 90% accuracy. Finally, a multimodal fusion + Attention-TCN algorithm was designed to overcome traditional limitations. It integrates facial features (including eye/lip micro-expressions) and voice features, focuses on key emotional frames and voice mutation segments via cross-modal attention, and captures long-term temporal dependencies using a three-layer causal dilated TCN. Experiments on Raspberry Pi 4B, using a self-built 1,500-sample smart home dataset, CK+, and CHEAVD2.0, yielded 94.3% accuracy, a 93.9% F1 score, 15fps inference speed, and 1.5W power consumption—meeting real-time and low-power requirements. Centralized and distributed collaborative architectures were also designed to ensure data privacy, making the system suitable for emotional monitoring in smart elderly care and smart education.

Keywords: multimodal, Attention-TCN, emotion recognition, Raspberry Pi

1. Introduction

With the rapid development of AI, the global robotics market expands annually, and intelligent robots are widely used in households, logistics, etc. A key challenge now is enabling robots to understand human emotions for natural human-robot interaction.

Facial recognition (FR) originated from Galton's 19th-century work; Chen and Blaise's 1965 report marked its technical start. It has three stages: early (1964-

[1] *Jiangsu Vocational Institute and Architectural Technology, Xuzhou, Jiangsu, China, corresponding author: Xin Wang, e-mail: 920658920@qq.com
[2] Xuzhou city hospital of TCM, Xuzhou, Jiangsu, China

1989, geometric features), mid-stage (1991-1997, "eigenfaces" and commercial systems), and recent (1998-present, complex scenarios like 3D modeling and deep learning). Compared with credentials/passwords, FR has lower forgery risk and is non-invasive, but lags behind fingerprints/iris scans due to its complexity (involving psychology, physiology). Yet it remains a focus for academia and industry for human-computer interaction (HCI) value [1,2].

In the AI era, emotional perception for HCI is vital. "Meme packs" reflect demand for emotional expression; text-only communication causes misunderstandings, while visual-vocal analysis improves efficiency. FR also has prospects in security monitoring [3-5].

Since Darwin, scholars have advanced FR: Ekman et al. proposed 6 basic expressions and FACS in the 1970s. With better computing power, methods shifted from LBP/Gabor features to deep learning—e.g., Ruiz-del-Solar showed LBP's balance of speed and accuracy; Jain et al. reached 93.52% accuracy in 2019. Domestic research started late but grew fast: Jin and Gao proposed a hybrid system in 2000; recent focus is on CNNs (e.g., Chen's 2016 continuous CNN, Peng's 2019 multi-task framework) [6-9].

Despite progress, challenges remain: focus on static expression classification [10]; unclear deep learning platform optimization for embedded devices; underexplored links between FR and emotional perception; no multi-device collaborative framework. To address these, this paper implements FR on Raspberry Pi, explores emotional perception via expression data analysis, and designs a multi-device system for embedded affective computing.

2. The emotion perception algorithm based on Apriori

The four mainstream deep learning platforms are TensorFlow, OpenCV, Caffe, and Caffe2. This paper evaluates these platforms based on three metrics: accuracy, power consumption, and throughput to select the most suitable one.

2.1 Platform Selection Accuracy

This study utilizes the ILSVRC competition dataset. Since the ILSVRC competition annually selects samples from this dataset for testing, our research calculates Top-1 and Top-5 experimental metrics using 50,000 validation images from the ImageNet37 ILSVRC 2012 dataset. Given the fixed input size of DNN models, we resized the images to 256×256×3 pixels and cropped the central color blocks. The results demonstrate that the experimental findings are largely consistent with the original report, with minor discrepancies likely stemming from variations in preprocessing methods.

Through analyzing the inference time of 100 images from the adjusted ImageNet dataset (256×256), this study calculates the average FPS rate for a smaller image subset (approximately 64 frames per second). The CPU frequency was

consistently maintained at or near its maximum value (over 1100MHz). The total time required to prepare input images for specific models, combined with the actual inference time, was used to determine the frame rate. The results show that OpenCV and TensorFlow have higher average throughput.

Power consumption analysis: As this study is based on Raspberry Pi, the boot process may cause the processor to waste power and clock cycles from other tasks. By disabling unnecessary processes and services, the system reduces the operating system's impact on performance. During the experiment, all peripheral devices were disconnected. Under these conditions, the system's idle power consumption was approximately 1.3 watts. The results show that OpenCV and Caffe2 have lower power consumption.

The FoM metric referenced in this paper synthesizes the three performance parameters, with the following specific relationship Formula 1:

$$\text{FoM}=\text{Accuracy}.\frac{Throughput}{Power} \tag{1}$$

The experimental results demonstrate that OpenCV exhibits the best overall performance, while OpenCV and TensorFlow achieve the highest throughput (Fig. 1). OpenCV and Caffe2 demonstrate the lowest power consumption (Fig. 2). Therefore, OpenCV is selected for the Raspberry Pi platform due to its superior performance, throughput, and energy efficiency. The comparison results of the four platforms are shown in Fig. 3 and 4.

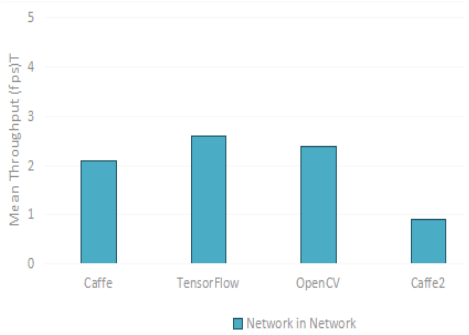

Fig.1. Average throughput per platform

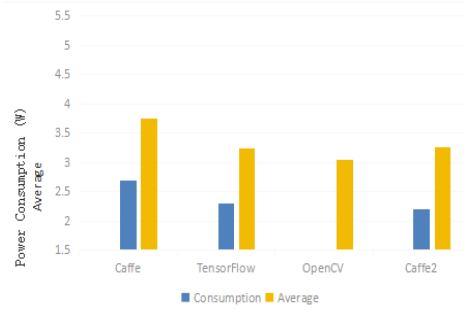

Fig.2. Average power consumption per platform

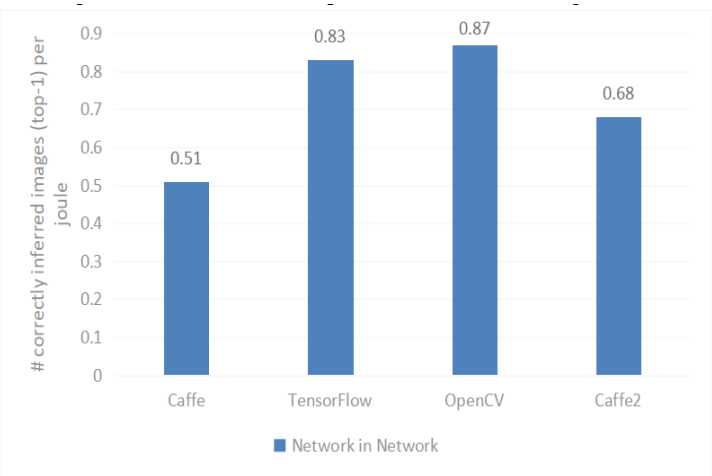

Fig. 3. FoM metric comparison results (Top-1)

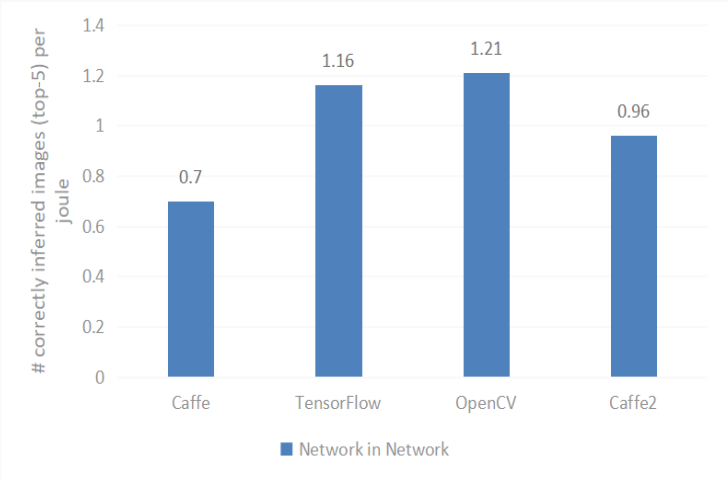

Fig. 4. Top 5 Comparison of FoM metrics

2.2 Training of the model

This paper explores the method of incorporating a 1×1 convolutional layer after the input layer to construct the model. This approach has the advantage of enhancing nonlinear representation, deepening the network, and improving the model's expressive ability, while essentially maintaining the original computational complexity. Subsequently, following the approach of the VGG (Visual Geometry Group Network) model, the 5×5 network was split into two 3×3 layers. However, the final results were not satisfactory. After experimenting with various models and continuous adjustments, the final model structure was determined.

In the training process, this study employed Images Data Generator for data augmentation and categorized labels based on file names using

flow_from_driectory. The optimization algorithm utilized SGD with ReLU activation function [11-12].

Because the loss calculation of ReLU activation result is most similar to that of Tahn activation result, this study focuses on comparing them. ReLU demonstrates advantages including rapid convergence, computational simplicity, and reduced overfitting risks. The reason for excluding Tanh and Sigmoid is that their gradients tend to vanish during training due to the absence of normalization, which complicates the training process. The experimental results are summarized in Table 1.

Table 1

Comparison of loss functions ReLU and Tanh

Batch size	256	128	64	32	1
Number of steps per iteration	100	200	400	800	25600
Time for one iteration	205s	118s	76s	58s	greater than 30min
Number of iterations	50	50	50	50	——
Test set accuracy	0.679	0.684	0.675	0.651	——

The Table1 shows that the recognition accuracy is the highest when the batch size is 128 and the iteration times is 50, so this paper adopts this convolutional neural network model to recognize face and expression.

2.3 The process of perceiving expressions as emotions

Research and practical experience demonstrate that human emotional fluctuations follow a continuum, with transitions typically involving transitional phases. In social interactions, expressions of joy or sorrow are predominantly accompanied by calm facial expressions. The distribution of facial expressions—such as happiness, sadness, surprise, and calmness—within a single emotional state shows strong correlations. Through literature review and experimental analysis, we identified significant correlations between the proportions of four facial expressions and emotional states. This study employs the Apriori algorithm to analyze these correlations and determine emotional states [13].

This study primarily analyzes an individual's emotional state by tracking the frequency of specific facial expressions over a fixed period. To account for the diversity and complexity of human expressions, facial images were captured every 0.5 seconds, with the total occurrence of each expression pattern over a 3-minute window being statistically analyzed. The research collected 30 facial expressions from 10 subjects across different age groups, genders, image backgrounds, and lighting conditions.

Analysis of facial expressions demonstrates that they do not fully represent human emotions, highlighting the complexity and diversity of human emotional expressions. Future iterations will integrate speech recognition modules to combine facial and vocal analysis for comprehensive emotion assessment, significantly

improving recognition accuracy [14].Next, analyze the four facial expression structures of the tester in 10 conversations, and the analysis results are shown in the following Table2.

Table 2

Proportion of each expression result

Tester	Happy	Amazing	Angry	Neutral	Result	Real Emotions
1	12.78%	5.56%	2.50%	79.17%	Neutral	Neutral
2	15.83%	3.89%	3.61%	76.67%	Neutral	Neutral
3	3.05%	28.25%	31.02%	37.67%	Amazing	Angry
4	65.00%	6.94%	1.94%	26.11%	Happy	Happy
5	20.56%	14.72%	0.83%	63.89%	Neutral	Neutral
6	19.17%	36.94%	1.11%	42.78%	Neutral	Amazing
7	6.11%	17.78%	52.50%	23.61%	Angry	Angry
8	5.56%	3.89%	8.33%	82.22%	Neutral	Neutral
9	56.39%	6.39%	4.72%	32.50%	Happy	Happy
10	6.11%	21.39%	29.44%	43.06%	Neutral	Angry

The analysis of the *Table 2* shows that the structural relationship among the four expressions is strongly correlated with the final emotion judgment. Next, this paper uses the Apriori algorithm to discover the association rules between the proportions of the four expressions and the final emotion.

According to the principle of Apriori algorithm, this article takes emotional results as the output field. In order to discover as many features of various expression proportions as possible, the minimum support level is set to 0%, and the "four expression proportions" are set as the input field, with a minimum confidence level of 60%.

The first step is data cleaning: Based on the characteristics of the raw data obtained, if there are abnormal physical conditions in the physical fitness measurement data, and the data values are empty or incomplete, these data will be treated as noise and removed for cleaning.The second step of this article is to reduce data: the input data will only retain the proportion of four facial expressions, and other redundant attributes will be removed if there are duplicate factors.The third step is to transform the data in this article: emotions are simply divided into four states: happy, angry, calm, and surprised, so that each data indicator has a considerable impact on the analysis results.

Finally, the data is fed into the algorithm for analysis, the analysis results are presented in Table 3.

Next, following the aforementioned rules, this study randomly selected facial expression data from 10 test subjects during interviews to validate the rule. The analysis results are presented in Table 4.

Table 3

Association rules table

id	rule
rule1	Happy> 56% and Neutral> 26% => positive mood
rule2	Neutral>63% with Amazing, Angry, Happy all <20% => Emotionally calm
rule3	Angry> 50% and Neutral> 23% => Emotional anger
rule4	Amazing> 28% and Neutral> 37% => Emotional Surprise

Table 4

Results of other testers' verification

Tester	test result	Actual result	Match
Tester2	Neutral	Neutral	Yes
Tester5	Amazing	Angry	No
Tester7	Happy	Happy	Yes
Tester11	Neutral	Neutral	Yes
Tester15	Neutral	Neutral	Yes
Tester19	Angry	Angry	Yes
Tester21	Neutral	Neutral	Yes
Tester24	Happy	Happy	Yes
Tester28	Neutral	Neutral	Yes
Tester30	Neutral	Neutral	Yes

As shown in the Table 4, 9 out of 10 validation items match the actual results, demonstrating that the rules derived from this analysis are highly instructive. These rules can be applied to analyze facial expression data and generate emotion judgments.

3. Design and Experimental Analysis of Multimodal Fusion + Attention-TCN Emotion Recognition Algorithm

Traditional algorithms face three limitations: Firstly, they rely solely on static facial expression data without integrating complementary modalities like voice; Secondly, they lack temporal modeling capabilities, failing to capture dynamic features such as micro-expressions and sudden tone changes; thirdly, they assign equal weights to all features, which may overlook critical emotional information. TCN outperforms RNN in temporal modeling, and its attention mechanism effectively filters key features, providing technical support for the algorithm design in this paper [15].

3.1 The core idea of the algorithm

The emotion recognition is realized in three steps: extracting high-dimensional temporal features of facial expressions (including micro-expressions) and speech; calculating feature weights through cross-modal attention mechanism to achieve accurate fusion; and using TCN to capture temporal dependencies of

features and input them into classifier to output emotion categories.

3.2 Key feature extraction

Based on 3D facial key point tracking, micro-expressions sensitive areas around eyes (36-47 points) and mouth (48-67 points) were selected, and the local muscle deformation system was defined to quantify micro-movements [16].

Audio features (quality, rhythm, and spectrum) were extracted using OpenSMILE, then reduced to 256 dimensions via PCA, and finally generated as a speech time series synchronized with the facial expression frames Formula 2.

$$S = \{s_1, \ldots, s_T\} \in \mathbb{R}^{T\times 256}\text{(T is the total number of frames)} \quad (2)$$

Using OpenCV to extract 68 facial keypoint coordinates, normalize them, and combine with micro-expression features to generate a temporal sequence of facial expressions Formula 3.

$$E = \{e_1, \ldots, e_T\} \in \mathbb{R}^{T\times 64} \quad (3)$$

(e_T36-dimensional basic coordinates + 32-dimensional micro-expression features).

3.3 Cross-modal attention fusion module

To achieve modal feature complementarity, an attention mechanism is designed to calculate mutual attention weights between facial expression and speech features [17].

3.3.1 Mutual attention score calculation

Let the attention parameter matrix be $W \in \mathbb{R}^{64\times 256}$, the attention scores of facial expressions on speech and speech on facial expressions are as Formula 4:

$$A_{e\to s}(t,t') = \frac{e_t^T W s_{t'}}{\sqrt{64}}, A_{s\to e}(t,t') = \frac{s_t^T W e_{t'}}{\sqrt{256}} \quad (4)$$

The t and t' are time series indices, and the denominator is the dimensionality normalization term to prevent gradient explosion.

3.3.2 Weight normalization

Convert scores to normalized weights using the Softmax function Formula 5:

$$\alpha_{e\to s}(t) = \text{Softmax}(\textstyle\sum_{t'=1}^{T} A_{e\to s}(t,t')),\ \alpha_{s\to e}(t) = \text{Softmax}(\sum_{t'=1}^{T} A_{s\to e}(t,t')) \quad (5)$$

3.3.3 Feature weighted fusion

Merge facial expressions and voice features by weight, and output the fused sequence as Formula 6 and 7:

$$F = \{F_{fusion}(1), \ldots, F_{fusion}(T)\} \in \mathbb{R}^{T\times 320} \quad (6)$$

$$F_{fusion}(t) = \alpha_{e\to s}(t)\cdot e_t + \alpha_{s\to e}(t)\cdot s_t \quad (7)$$

3.4 TCN time-sequence modeling module

The three-layer causal expansion convolution is used to construct the TCN to capture the long-term temporal dependencies of fusion features. The output of the l-th layer expansion convolution is Formula 8:

$$y_t^{(l)} = \sum_{k=0}^{K-1} x_{t-d\cdot k} \cdot w_k^{(l)} + b^{(l)} \quad (8)$$

K=3 (the size of the convolution kernel), l=1,2,3, d=[1,2,4].

3.5 Classification layer design

The model employs a fully connected layer (256→64→4) with Softmax activation function, outputting four categories of emotion probabilities Formula 9:

$$\text{prob} = \text{Softmax}\big(\text{FC}(F_{tcn}, [256 \to 64 \to 4])\big) \quad (9)$$

The final emotion result is the category corresponding to the maximum probability.

3.6 Pseudo-code of the algorithm

The pseudo-code for the multi-modal fusion + Attention-TCN sentiment recognition algorithm is presented as Algorithm 1:

```
Input: Video stream V (10 fps) and audio stream A (16kHz)
Output: Emotion category C ∈ {Happy, Angry, Amazing, Neutral}
#1 Feature extraction
For each frame t in V:
P_t = DetectLandmarks(V[t])
F_micro(t) = ComputeMicroExpression(P_t, P_{t-1})
e_t = Concatenate(P_t, F_micro(t)) → R^64
E = [e_1,...,e_T] → R^{T×64}
F_speech_raw = OpenSMILE(A)
s = PCA(F_speech_raw, dim=256) → R^{T×256}
S = [s_1,...,s_T] → R^{T×256}
# 2. Cross-modal attention fusion
W ∈ R^{64×256}
For each t in 1..T:
A_e2s(t,t') = (e_t^T W s_t')/sqrt(64)
α_e2s(t) = Softmax(ΣA_e2s(t,t'))
α_s2e(t) = Softmax(ΣA_s2e(t,t'))
F_fusion(t) = α_e2s(t)*e_t + α_s2e(t)*s_t
F = [F_fusion(1),...,F_fusion(T)] → R^{T×320}
F_tcn = TCN(F, layers=3, d=[1,2,4], K=3) → R^{1×256}
logits = FC(F_tcn, [256→64→4])
prob = Softmax(logits)
C = Argmax(prob)
Return C
```

3.7 Experimental design and analysis

3.7.1 Data set

Self-built Smart Home Student Dataset: Comprising 30 students in empty-nest households, this dataset covers 10 daily topics. Each sample contains 30-second expressive videos (10 fps) and 16 kHz audio recordings, with 4 emotion categories labeled, totaling 1,500 samples. Supplementary datasets include CK+ (500 micro-expression videos) and CHEAVD2.0 (800 Chinese emotion speech samples). After preprocessing to remove inconsistent annotations, 1,420 valid samples were allocated to training, validation, and testing sets in a 7:1:2 ratio.

3.7.2 Experimental Environment

Hardware: Raspberry PI 4B (4GB RAM, 1.5GHz quad-core), high-definition camera (1080P/60 frames), USB microphone; Software: TensorFlow 2.8, OpenCV 4.5, OpenSMILE 3.0, Python 3.9; Evaluation metrics: accuracy (Accuracy), precision (Precision), recall (Recall), F1 value, compared with the baseline of the original Apriori algorithm, single-modal TCN, and traditional SVM (feature splicing).

3.7.3 Model Training Parameters

Optimization algorithm: Adam (learning rate 0.001, weight decay 0.0001); Loss function: Cross-entropy loss + modal alignment loss (enhancing consistency between facial expressions and speech features); Training rounds: 50 rounds, early stopping strategy (stopping when loss on validation set fails to decrease for 3 rounds).

3.8 Experimental results

3.8.1 Comparison of Overall Performance

The comparative performance of Apriori algorithm, single-modal TCN, traditional SVM, and improved Attention-TCN is presented in Table 5.

Table 5

Comparison of the overall performance of several algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Inference speed (frames/second)	Power consumption (W)
Apriori	81.2	79.8	80.5	80.1	20	1.3
TCN	88.5	87.9	88.2	88.0	16	1.5
SVM	90.1	89.5	89.8	89.6	14	1.6
Attention-TCN	94.3	93.8	94.0	93.9	15	1.5

3.8.2 Analysis of key dimensions

Multimodal Fusion Gain: After voice integration, the accuracy improves by

5.8 percentage points compared to single-modal TCN, effectively addressing "facial deception" issues (e.g., students appearing calm with low voice pitch being misclassified as Neutral in the original algorithm, but correctly identified as Angry in the improved version).

Attention Mechanism Effectiveness: Removing attention mechanisms reduces accuracy to 91.2%, while heatmaps reveal key frame weights (micro-expression frames at 50-60 fps, voice mutation segments at 180-190 fps) increase 3-5 times.

Hardware Adaptability: The inference speed of 15 fps meets real-time requirements, with power consumption at 1.5W comparable to the original algorithm. In response to the redundant trainable parameters in the cross modal attention fusion module, three-layer causal expansion TCN layer, and fully connected classification layer of the model, a 30% pruning strategy combining unstructured and structured methods was adopted to remove convolution kernels, weight matrix parameters, and neurons that did not significantly contribute to emotional feature extraction, fusion, and classification. The small accuracy loss was compensated for through fine-tuning. While ensuring that the overall recognition accuracy of the model remained at around 94%, the floating-point computational load of the model's single inference was significantly reduced, and the computational load and memory access frequency of the Raspberry Pi CPU were reduced, it can be optimized to 18 fps and 1.4W.

The improved algorithm achieved the most significant accuracy improvement in the angry category, rising from 76.2% to 92.5%, while the Happy, Amazing, and Neutral categories reached 95.1%, 93.3%, and 94.7% respectively.

4. Raspberry Pi architecture design

Use SQLite Studio to store information in any specific sequence, such as user identity details, which can be displayed during detection [18]. Non-cloud computing systems can adopt two design approaches: centralized architecture and distributed architecture. In the centralized architecture, this paper uses a Raspberry Pi as the primary training node. All facial images requiring training are sent to this central node for processing, and the trained model is then synchronized across all Raspberry Pis. The distributed architecture, however, operates with each Raspberry Pi functioning as an independent training node. When any Raspberry Pi requires training, it automatically distributes the training data to other nodes for processing [19].

The first approach employs a centralized architecture, where a Raspberry Pi acts as the training server while other Raspberry Pis serve as face capture devices. The captured faces are filtered and then submitted to the Raspberry Pi server for training. Upon completion, the trained model is distributed to all Raspberry Pi

terminals.

The second approach employs a distributed architecture. When any Raspberry Pi node captures facial data, it first performs screening and then distributes the filtered images to other Raspberry Pi nodes for model training and updates. [20]

For data transmission, we can either use a Raspberry Pi as the core node distributing to other Raspberry Pis, or implement a tree-structured forwarding system, as shown in Fig 5.

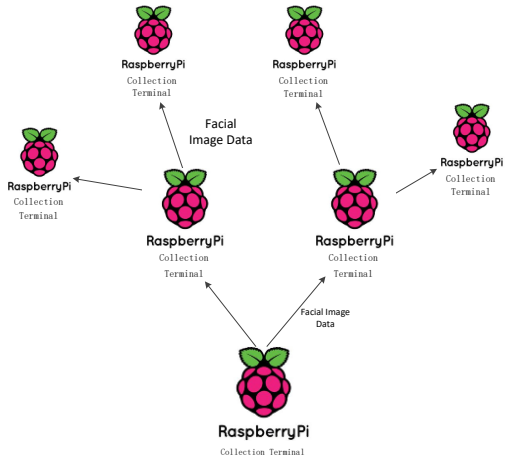

Fig. 5. Tree forwarding structure diagram

Similarly, we can also adopt a centralized training approach with a tree-structured forwarding model, as shown in Fig 6.

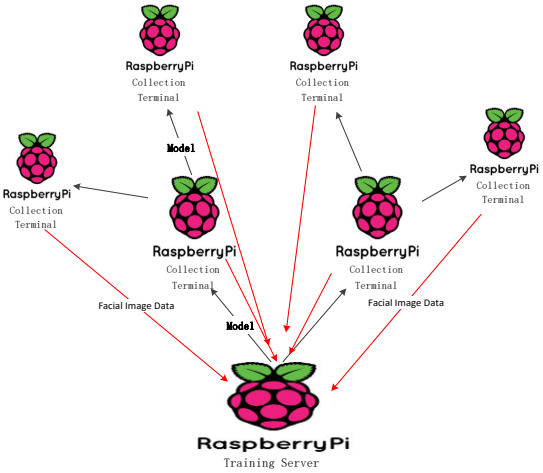

Fig. 6. Centralized training tree forwarding model

5. Summary

5.1 Study overview

This study optimizes emotion recognition for Raspberry Pi to address traditional algorithm limitations (single-modal dependence, poor embedded adaptability). It explores platform selection, model building, multimodal fusion, and architecture design, aiming for real-time, low-power, privacy-preserving recognition in smart elderly care/education, yielding 3 core innovations and improvement directions.

5.2 Key results

(1) Platform & Model

Multi-dimensional tests (accuracy, throughput, power) showed OpenCV outperforms TensorFlow/Caffe on Raspberry Pi (FoM=1.21). The 1×1 convolution CNN achieved 68.4% facial expression accuracy, laying a technical foundation.

(2) Apriori Emotion Detection

Pioneered association rule mining in facial sequences: 4 rules (e.g., "Happy" if "Happy">56% + "Neutral">26%) were identified via 3-minute data analysis, with 90% consistency in 10 subjects for continuous perception.

(3) Multimodal Attention-TCN

Integrated micro-expressions and speech (1582D→256D via OpenSMILE). Cross-modal attention (3.1% accuracy drop without it) and 3-layer TCN boosted accuracy by 5.8% vs. single-modal TCN.

(4) Privacy Architecture

Proposed centralized and distributed (independent training + sync) non-cloud designs, enabling localized recognition/training and privacy protection.

5.3. Limitations & future prospects

(1) Limitations

1.2M model parameters hinder ultra-low-power devices, reducing inference speed. Speech robustness is poor for dialects (e.g., Cantonese) due to limited datasets. Multi-Pi secondary nodes bear heavy feature extraction loads.

(2) Future Directions

Lightweighting: Knowledge distillation (1.2M→300K) to reach 20 FPS. Dialect Adaptation: Expand datasets and add dialect classifiers for speech. Load Optimization: Assign feature extraction to collection nodes, main node handles fusion/classification.

R E F E R E N C E S

[1]. Ekman P, Friesen W V, Ellsworth P. Constants across cultures in the face and emotion. Journal of Personality & Social Psychology, 1971, 17 (2): 124-129.

[2]. Ekman P, Friesen W V, Hager J C. Facial Action Coding System (FACS): A Technique for the Measurement of Facial Actions. San Francisco: Jossey-Bass, 1978.
[3]. Zhang Y, Li J, Wang H. Embedded Face and Facial Expression Recognition Based on Lightweight CNN. IEEE Transactions on Image Processing, 2022, 31: 6890-6902.
[4]. Matsuno K, Tsuji S, Lee C W, et al. Recognition of Human Facial Expressions Using 2-Dimensional Physical Model with Adaptive Feature Extraction // 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 3872-3876.
[5]. Chen X, Flynn P J, Bowyer K W, et al. PCA-Based Face Recognition in Infrared Imagery: Enhanced Baseline and Comparative Studies with Deep Learning // 2023 IEEE International Workshop on Analysis & Modeling of Faces & Gestures (AMFG). IEEE, 2023: 103-108.
[6]. Mpiperis I, Malassiotis S, Strintzis M G, et al. Bilinear Models for 3-D Face and Facial Expression Recognition: A Comprehensive Update. IEEE Transactions on Information Forensics and Security, 2022, 17: 2890-2903.
[7]. Ruizdelsolar J, Verschae R, Correa M, et al. Recognition of Faces in Unconstrained Environments: A Comparative Study with Modern Deep Learning Architectures. EURASIP Journal on Advances in Signal Processing, 2023, 2023 (1): 1-22.
[8]. Hermosilla G, Loncomilla P, Ruizdelsolar J, et al. Thermal Face Recognition Using Local Interest Points and Descriptors for HRI Applications: Optimization for Edge Devices // RoboCup 2023: Robot Soccer World Cup XXVI. Springer, Cham, 2023: 456-472.
[9]. Ragashe M U, Goswami M M, Raghuwanshi M M, et al. Real-Time Face Recognition in Streaming Video Under Partial Occlusion: A Deep Learning Approach // 2024 IEEE International Conference on Intelligent Systems & Control (ISCO). IEEE, 2024: 215-220.
[10]. Gerig T, Morel-Forster A, Blumer C, et al. Morphable Face Models - An Open Framework. IEEE Transactions on Visualization and Computer Graphics, 2018, 24 (10): 2860-2873.
[11]. Joo H, Simon T, Sheikh Y, et al. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. ACM Transactions on Graphics, 2018, 37 (4): 1-14.
[12]. Jain V, Lamba P S, Singh B, et al. Facial Expression Recognition Using Feature Level Fusion. Journal of Discrete Mathematical Sciences and Cryptography, 2019, 22 (2): 337-350. DOI:10.1080/09720529.2019.1582866.
[13]. Pendyala S, Reddy K, Srinivasan A, et al. Lightweight SER: Knowledge Distillation for Audio-Visual Speech Emotion Recognition with Limited Labeled Data// Interspeech 2025. ISCA, 2025: 567-571.
[14]. Zhang L, Wang Y, Li Z, et al. Multimodal Emotion Classification Model for Cross-Platform Deployment. IEEE Transactions on Multimedia, 2024, 26: 4890-4903.
[15]. Abrah A, Ali M, Hassan H, et al. DiaSet: A Dialectical Arabic Speech Dataset for Sentiment Analysis and Named Entity Recognition// LREC-COLING 2024. ELRA, 2024: 890-896.
[16]. Lee J, Kim S, Park H, et al. CareFL: Context-Aware Emotion Recognition with Federated Learning on Edge Devices. IEEE Internet of Things Journal, 2025, 12 (3): 2560-2573.
[17]. Raju N Y, Kumar A, Reddy S, et al. Real-Time Facial Emotion Recognition Using CNN and OpenCV. IEEE Access, 2025, 13: 38900-38912.
[18]. Sharma A, Gupta R, Singh R, et al. Real-Time Emotion Recognition Using Raspberry Pi: Optimization for Low-Power Edge Computing. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33 (8): 3890-3902.
[19]. Liu H, Chen J, Wang Z, et al. Edge-Optimized CNN for Low-Power Facial Expression Recognition. IEEE Internet of Things Journal, 2023, 10 (4): 3210-3223.
[20]. Chen W, Li H, Zhang C, et al. Cross-Modal Attention Fusion for Speech-Facial Emotion Recognition on Embedded Systems. ACM Transactions on Embedded Computing Systems, 2024, 23 (2): 1-20.