electronics

MDPI

Article

Emotion-Recognition Algorithm Based on Weight-Adaptive Thought of Audio and Video

Yongjian Cheng *, Dongmei Zhou, Siqi Wang and Luhan Wen

School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu 610059, China; zhoum@stu.cdut.edu.cn (D.Z.); 2021021106@stu.cdut.edu.cn (S.W.); wenluhan@stu.cdut.edu.cn (L.W.)
* Correspondence: 2021021102@stu.cdut.edu.cn

Abstract: Emotion recognition commonly relies on single-modal recognition methods, such as voice and video signals, which demonstrate a good practicability and universality in some scenarios. Nevertheless, as emotion-recognition application scenarios continue to expand and the data volume surges, single-modal emotion recognition proves insufficient to meet people's needs for accuracy and comprehensiveness when the amount of data reaches a certain scale. Thus, this paper proposes the application of multimodal thought to enhance emotion-recognition accuracy and conducts corresponding data preprocessing on the selected dataset. Appropriate models are constructed for both audio and video modalities: for the audio-modality emotion-recognition task, this paper adopts the "time-distributed CNNs + LSTMs" model construction scheme; for the video-modality emotion-recognition task, the "DeepID V3 + Xception architecture" model construction scheme is selected. Furthermore, each model construction scheme undergoes experimental verification and comparison with existing emotion-recognition algorithms. Finally, this paper attempts late fusion and proposes and implements a late-fusion method based on the idea of weight adaptation. The experimental results demonstrate the superiority of the multimodal fusion algorithm proposed in this paper. When compared to the single-modal emotion-recognition algorithm, the accuracy of recognition is increased by almost 4%, reaching 84.33%.

Keywords: multimodal; time-distributed CNNs; LSTM; DeepID V3; Xception

Citation: Cheng, Y.; Zhou, D.; Wang, S.; Wen, L. Emotion-Recognition Algorithm Based on Weight-Adaptive Thought of Audio and Video. *Electronics* **2023**, *12*, 2548. https://doi.org/10.3390/electronics12112548

Academic Editor: Ricardo Santos

Received: 9 April 2023
Revised: 31 May 2023
Accepted: 2 June 2023
Published: 5 June 2023

1. Introduction

Affective computing is a comprehensive research and technical field that involves various disciplines and applications, promoting research in the field. It mainly focuses on human–computer interaction and related issues to achieve emotional communication between humans and computers in a friendly environment. Affective computing has a wide range of applications in areas such as art, business, education, finance, medicine, and security. In 1997, Picard of the MIT Media Lab predicted about 50 possible application scenarios for affective computing in the book *Affective Computing* [1], which has significant research value. Currently, researchers are focused mainly on single-modal emotional computing, such as text semantics and sentiment analysis, speech transcription, and emotion recognition, and facial expression recognition. However, single-modal analysis has limitations, as many factors affect human emotions, and their internal connections are complex and changeable. Therefore, unimodal analysis cannot fully reflect human emotions. Consequently, researchers are turning to bimodal and multimodal analysis. Multimodal deep learning—a model with a high generalization ability and good recognition performance—is developed from multimodal machine learning, mainly employing deep-learning methods to address problems in the multimodal field, such as the low recognition rate and poor robustness of single-modal affective computing. Multimodal analysis can leverage the correlation and independence between different modalities to fully exploit the potential of each modality's information and increase the confidence level in emotion recognition.

This paper explores the problem of multimodal emotion recognition using deep learning and proposes two model design schemes for audio and video modalities via the application of multimodal thought. Moreover, improvements have been made to multimodal fusion-related technologies, and multimodal fusion has been conducted at the decision-making level.

The main research work comprises three aspects:

Firstly, for audio-modal emotion recognition, this paper studies audio-signal preprocessing methods and feature-extraction algorithms. The logarithmic-mel spectrogram feature [2,3] is used, and the "time-distributed CNNs + LSTMs" network [4,5] is designed to make full use of the audio time domain information. Compared with other common network model training datasets, it is found that the algorithm proposed in this paper has a certain degree of improvement in accuracy compared with existing common algorithms, and the model training speed is also significantly improved.

Secondly, for video-modal emotion recognition, this paper studies image-preprocessing methods and feature-extraction algorithms. The HOG feature-extraction technology [6,7] is used in the image preprocessing, and the main framework of the video-modal emotion feature-extraction network is constructed by using the optimized Xception system [8,9]. At the same time, the DeepID V3 network [10,11] is introduced to realize the extraction of facial feature points, so that the extracted video emotion features are more comprehensive and effective. This paper proves that the proposed algorithm improves the accuracy rate by 5% compared with the existing common algorithms.

Lastly, for multimodal fusion [12], it is particularly important to choose an appropriate fusion method. Common fusion methods do not have absolute advantages. In actual tasks, various factors need to be considered comprehensively. This paper mainly studies the multimodal deep learning emotion-recognition algorithm and its application. According to the progress of previous research work and the needs of later system expansion, as well as considering the asynchronous nature of the dataset used in this study, the late fusion method—which is more flexible in modality expansion—is chosen to carry out research on multimodal emotion recognition. The American psychologist Mehrabian proposed a formula: Emotional information expression during communication = 7% speech + 38% human voice + 55% facial expression [13]. It can be seen that the information contained in the voice data and facial-expression data during human communication accounts for 93% of the expression of emotional information in communication. Based on the above theoretical basis, this paper abandons the poorly performing text mode, comprehensively considers the emotional expressiveness of each mode and the need for subsequent mode expansion, and finally adopts a late fusion decision-making method based on the idea of weight self-adaptation to realize the decision-level fusion of audio and video modalities. In short, the model and method proposed in this paper have achieved a good performance on multiple datasets, providing new ideas and methods for the in-depth exploration of multimodal emotion-recognition problems.

2. Related Work

With the development and application of the field of affective computing and deep-learning technology, researchers began to pay attention to the research of multimodal affective computing. In 2017, researchers from the University of Stirling (School of Natural Sciences) and Nanyang Technological University (Temasek Laboratories) in Singapore conducted a collaborative study. Soujanya Poria conducted the first comprehensive literature review on the different fields of affective computing [14]. On the basis of describing the results of various single-factor impact analysis, the existing methods of information fusion under different modes are outlined. In this article, the researchers review the basic stages of the multimodal emotion-recognition framework for the first time. The available benchmark datasets are first discussed, followed by an overview of recent advances in audio, video, and text-based emotion-recognition research. The article addresses findings by other researchers that multimodal classifiers far outperform unimodal classifiers. Furthermore,

deep learning has clear advantages in multimodal tasks. Therefore, future research on multimodal fusion emotional computing combined with deep learning ideas will be an important research direction in this field.

The model proposed by researchers such as Deepak Kumar Jain is based on a single deep convolutional neural network [15], which contains convolutional layers and deep residual blocks. Image labels for all faces are first set up for training. Second, the images are passed through the proposed DNN model. The contribution is to classify each image into one of six facial emotion categories. Balaji Balasubramanian et al. present a dataset and algorithm for facial emotion recognition [16]. This algorithm ranges from simple support vector machines (SVM) to complex convolutional neural networks (CNN). These algorithms are explained through fundamental research papers and applied to the FER task. Dhwani Mehta et al. focus on identifying emotional intensity using machine-learning algorithms in a comparative study [17]. The algorithms used in the comparative study are Gabor filters, the histogram of oriented gradients (HOG), and local binary patterns (LBP) for feature extraction. For classification, support vector machines (SVM), random forests (RF), and nearest neighbors (kNN) are used. The study implements emotion recognition and intensity estimation for each recognized emotion. Yang Liu et al. conduct the first investigation of the graph-based FAA method [18]. The results of the team's findings can serve as a reference for future research in this area and summarize the performance comparison of state-of-the-art graph-based FAA methods, discussing the challenges and potential directions for future development.

Multimodal affective computing has been continuously improved by rapid development and has solved many of the problems raised previously, but new challenges have been raised by researchers one after another. Therefore, there are still many problems in this field of research, which urgently need to be solved. The core issues of the current research are how to efficiently extract effective features in multimodal datasets, eliminate redundant and invalid interference information, and achieve effective multimodal fusion, improving classification accuracy, and optimizing system performance. In order to solve these core problems in the field of multimodal affective computing, researchers are paying more attention to the application of deep-learning algorithms and the design of a more complete multimodal deep-learning model. The research on the combination of multimodal affective computing and deep learning has achieved great progress and remarkable results.

3. Materials and Methods

3.1. Audio-Modal Model Construction

3.1.1. Implementation Process of Audio-Modality Emotion-Recognition Model

For the accurate detection of emotional flashpoints during audio processing and analysis, appropriate audio segmentation is necessary. Furthermore, other preprocessing operations [2,19] are essential to enable efficient emotional feature extraction and obtain the expected emotion-recognition classification model through training. This paper presents an audio-modality emotion-recognition network based on a time-distributed convolutional neural network and long short-term-memory network [20,21]. The network leverages time and frequency domain information to extract audio signals and incorporates contextual correlation into model training. Figure 1 below illustrates the overall model training and prediction process:

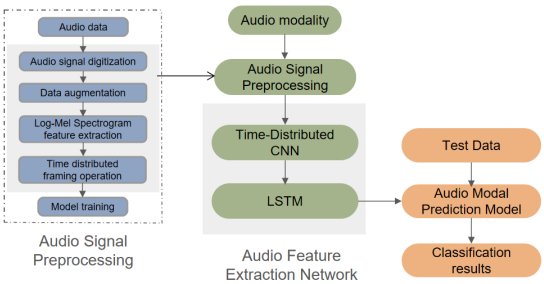


Figure 1. Flow chart of audio-modal emotion recognition based on a time-distributed CNN + LSTM.

The audio feature-extraction network consists of a stack of time-distributed CNN and LSTM network [4], which is different from the traditional CNN [22]. The CNN network introduces the time-distributed wrapper [23], and the neural network with this design is referred to as a time-distributed convolutional neural network. The time-distributed layer is applied to each input time slice, and the input data dimensionality contains at least three dimensions. In experiments, audio samples have been transformed into a dataset of shape (N,128,384), where each sample contains 128 frames that have undergone time-distributed framing operations, and each frame has 384 feature vectors. The fully connected layer is then applied to each of these 128 frames independently using the time-distributed wrapper, which keeps the parameter weights of the shared layer and total number of parameters constant. Additionally, time-distributed wrappers are also applied to convolutional layers, batch normalization layers, activation function layers, pooling layers, and dropout layers, allowing feature maps of different layers to share parameter weights, which is an important implication of its application.

The primary idea of the time-distributed convolutional neural network is to use a rolling window on the log-mel spectrogram, with a preset window size and moving step. Each window is used as an input for a convolutional neural network consisting of four local featur- learning blocks [24] to extract shallow information about audio samples. Subsequently, the output of each convolutional neural network unit is flattened to obtain the output vector of the time-distributed convolutional neural network. These vectors are then fed into a recurrent neural network containing two LSTM units to learn long-term contextual dependencies in samples and extract deeper features of audio samples. Finally, a fully connected layer with a Softmax activation function [25] is applied to predict emotions in the audio samples. This is the overall design of the audio-modality feature.

Figure 2 illustrates the schematic diagram of the network structure.

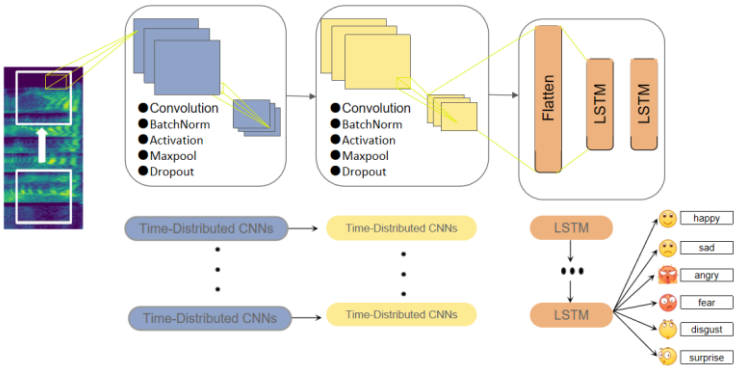


Figure 2. Schematic diagram of the audio feature-extraction network structure.

The above figure elaborates on the process of inputting log-mel spectrogram features [2] into the shallow feature-extraction network comprising four time-distributed CNNs to extract shallow features of the audio samples. The output is then passed to the deep feature-extraction network, consisting of two LSTM units, which mines the long-term contextual relationships and extracts the deep features. This model construction is referred to as "time-distributed CNNs + LSTMs".

3.1.2. Audio-Modality Dataset

The Ryerson Audio–Visual Database of Emotional Speech and Song: RAVDESS [26]

The RAVDESS dataset, also known as the Ryerson affective language and song audio–visual dataset, was released in 2013. It is a large-scale audio emotion dataset developed by the SMART Lab team. The dataset is collected from 24 professional actors with pure North American pronunciation, including 12 male and 12 female professional actors. Each actor records several speech audios and song audios with emotional labels. The research in this paper only uses speech audio samples in the dataset. A total of 24 professional

actors participate in the speech audio collection. The number of collections per person is 60, and the total number of samples is 1440. The composition of emotional labels is shown in Table 1 for details:

Table 1. RAVDESS dataset sentiment label distribution table.

Emotional Label	Happy	Sad	Angry	Fear	Disgust	Surprise	Calm	Neutral	Total
Male	96	96	96	96	96	96	96	48	720
Female	96	96	96	96	96	96	96	48	720
total	192	192	192	192	192	192	192	96	1440

Each of the seven emotions in the Table 1 above: happy, sad, angry, fear, disgust, surprise, and calm, is distinguished by two emotional intensities: normal and strong. The number of valid samples verified for model training for the above seven emotions is 1344. There is also a neutral emotion, which is not differentiated by intensity. It should be noted that because we focus on the task of sentiment analysis in speech, only samples of the first six emotions are used for training the audio-modal emotion-recognition model in this paper.

The Interactive Emotional Dyadic Motion Capture Database: IEMOCAP [27]

The IEMOCAP dataset is a multimodal dataset for emotion recognition, including audio, video, text, and action data. The dataset includes sessions from 10 actors, each character exhibiting multiple emotions in different environments. In order to compare the accuracy of the six classifications, this article uses the audio part of the IEMOCAP dataset for verification and comparison testing. The audio part excerpts six emotional classifications, which are: happy, sad, angry, fear, disgust, and surprise. Each of these six emotional classifications has a corresponding speech sample, a total of 1083 audio clip samples. In addition, the audio part of the IEMOCAP dataset also contains other emotional classifications, such as neutral and others. Among them, each sample is an audio file with an emotion label, marking the emotional state expressed by the audio file, which can be used for model training, testing, and evaluation of emotion classification tasks. In addition, there are certain other labels and attributes related to emotion, such as speech features, speech quality, and emotional intensity, etc., which can help researchers to better understand and analyze the audio part of the IEMOCAP dataset. The emotional labels distribution table of the IEMOCAP dataset used is shown in Table 2.

Table 2. IEMOCAP dataset sentiment label distribution.

Emotional Label	Happy	Sad	Angry	Fear	Disgust	Surprise	Total
Number of sample points	188	218	214	200	134	129	1083

3.2. Construction of Video-Modality Model

3.2.1. Implementation Process of Video-Modality Emotion-Recognition Model

Facial expressions are often utilized to construct features for video-modality emotion-recognition tasks. Human expression provides an intuitive reflection of psychological emotions, making facial expressions one of the crucial modalities in affective computing research. Many researchers focus on facial expression and rely on chosen features to play a significant role in final model performance.

For the task of video-modality emotion recognition, this paper presents the "DeepID V3 + Xception architecture" model construction scheme: First, video-modal data undergo image preprocessing [28,29] and HOG feature extraction [6,7]; then, the residual network design [30,31] is introduced, along with the design of the network structure and the adjustment of the relevant optimization strategies based on the working principles and

roles of DeepID V3 [10,11] and the Xception system in the video feature-extraction network. Figure 3 depicts the training and prediction flowchart of the video-modality emotion-recognition model. The extended Cohn–Kanade dataset [32] undergoes preprocessing operations, to extract valid frames from the video sequences, and subsequently passes them into the neural network for deeper feature extraction. The final classification prediction model is obtained through training.

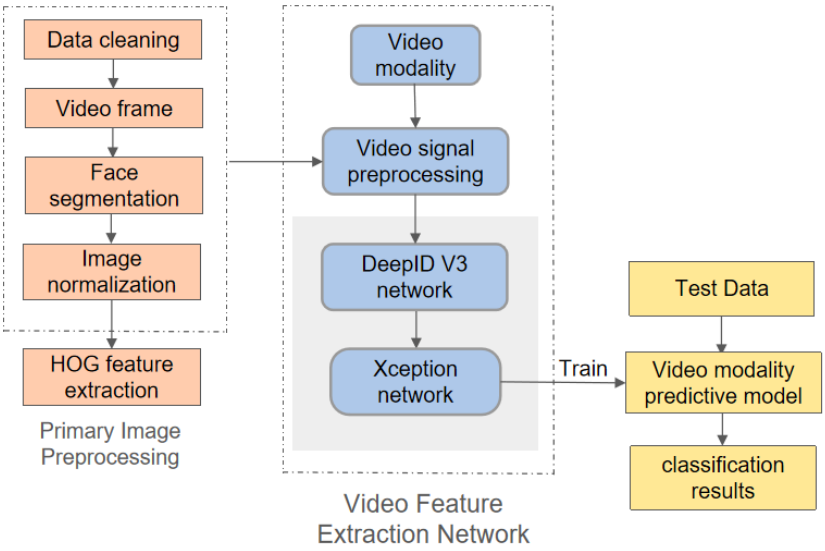


Figure 3. Flowchart of video-modal emotion recognition based on DeepID V3 + Xception architecture.

3.2.2. Video Feature-Extraction Network

After completing relevant image preprocessing operations on the video sequences within the dataset, experimental samples undergo transformation into multidimensional arrays with a higher processing capability for neural networks. In deep learning, the neural network design can impact model performance. Hence, this paper utilizes the Xception system network [8,9] as the primary video-modality emotion-recognition network framework, with the DeepID V3 network [10,11] serving as the front-end facial feature-extraction network. Together, they form the video feature-extraction network, known as the "DeepID V3 + Xception architecture" scheme. The use of related networks is detailed below.

Facial Feature-Extraction Network: DeepID V3

DeepID3 proposes two deeper neural network architectures inspired by the VGG network and GoogLeNet [33,34], respectively, named DeepID3 net1 and DeepID3 net2. The structure is illustrated in Figure 4.

Within DeepID3 net1, each pooling layer is preceded by two consecutive convolutional layers. Compared to the traditional VGG network [35], DeepID3 net1 has been optimized in several ways: (i) supervisory signals are added to multiple fully connected layers branched from the middle pooling layer, enhancing the neural network's ability to learn mid-level features and allowing for easier training and optimization of deep neural networks; (ii) the top two convolutional layers are replaced with two local-connection layers, utilizing unshared weight parameters to create more expressive features while also reducing feature size.

In DeepID3 net2, the shallow network structure resembles DeepID3 net1, with a pooling layer inserted after every two consecutive convolutional layers. The subsequent deep network feature-extraction stage introduces the inception structure, where three consecutive inception layers are stacked before the third pooling layer, and two consecutive inception layers before the fourth pooling layer. A joint identity verification supervision signal is additionally included on top of the fully connected layer.

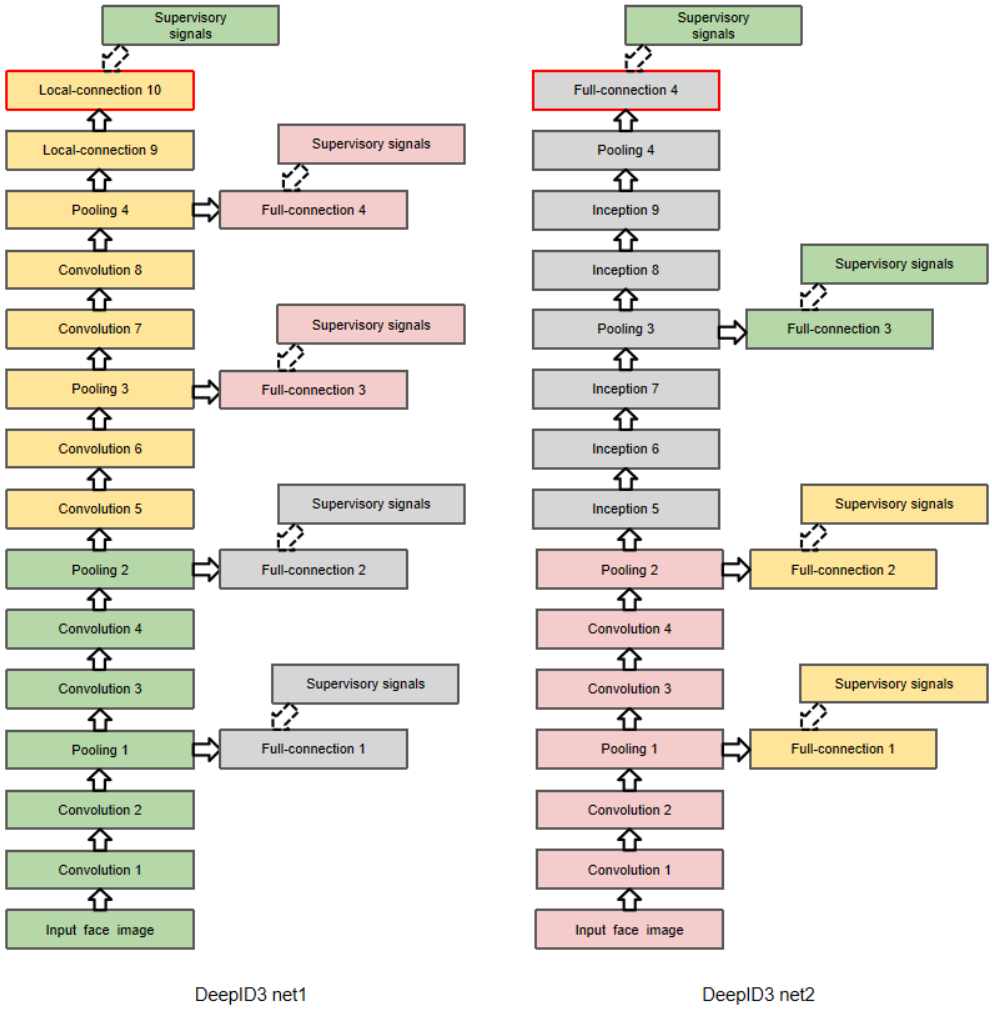


Figure 4. DeepID3 two-network architecture diagram.

This paper employs DeepID3 net1 as the facial feature-extraction network, enhanced using two optimizations: (i) a ReLU activation function [36] is applied to all neural network layers except the pooling layer, reducing the network's overall computational load; (ii) a dropout layer is introduced following the final feature-extraction layer. Although this design increases network depth, the improved network's size is smaller than that of the VGG network and GoogLeNet. After image preprocessing, feature extraction is performed on samples using this improved DeepID3 net1 network, resulting in a class activation map, as shown in Figure 5. The figure displays activated pixels on the final layer. One can observe the emotion "happy" linked to pixels surrounding the eyes and mouth, while the emotions "angry" and "sad" seem to be linked to pixels near the eyebrows. Of course, these visualizations merely generate a rough perception of facial features, and building a complete neural network model is ultimately necessary to determine the relationship between emotion types and features.

Xception System Network

Xception [8,9] is a novel deep convolutional neural network architecture inspired by the Inception structure. The Inception structure [37,38] sits between traditional convolution and depthwise separable convolution operations in convolutional neural networks. Depth-separable convolution comprises two steps: channel-by-channel convolution and point-by-point convolution, which can be interpreted as the Inception structure with the largest number of towers, which allows Xception to make use of deep separable convolution instead of Inception structures. In image classification, Xception outperforms

Inception V3 [39]. Therefore, this paper adopts Xception as the primary video-modality emotion-recognition model framework, with targeted fine-tuning and optimization to form a video-modal feature-extraction network together with the previously mentioned DeepID V3 network.

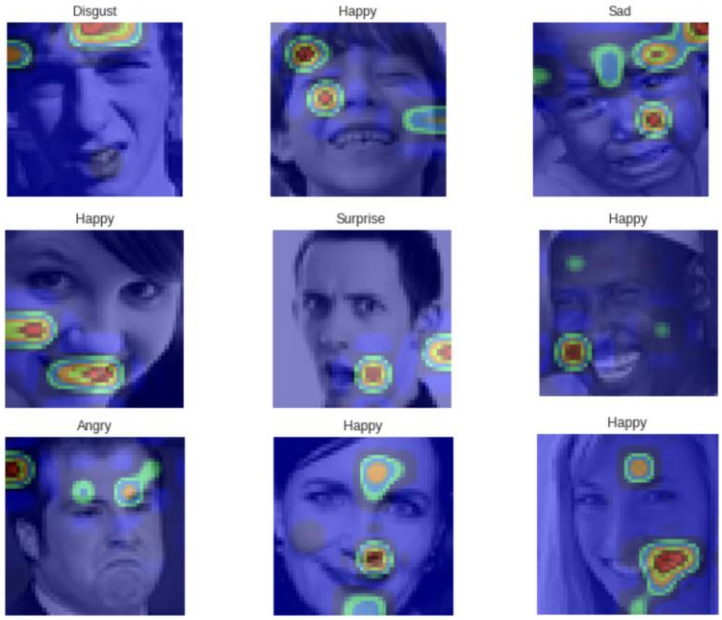


Figure 5. Class activation map of several emotions.

Xception System Network Structure

The Xception system proposes a convolutional neural network architecture based on depth-separable convolutional layers, distinct from the Inception system. Xception hypothesizes that the mappings of cross-channel correlations and spatial correlations in feature maps of convolutional neural networks can be fully decoupled. This assumption stems from those proposed by the Inception architecture, thereby naming Xception as "extreme Inception".

Figure 6 illustrates the network structure diagram of the Xception system. The network comprises 36 convolutional layers, including classical convolutional layers and depthwise separable convolutional layers that form the fundamental feature-extraction component. The last fully connected layer and logistic regression layer [40] of the network are optional, depending on task requirements. For our video-modality emotion-recognition task, these layers are necessary to achieve image classification.

The overall Xception system comprises three parts: entry flow, middle flow, and exit flow. A total of 36 convolutional layers construct 14 modules, where all modules, except the first and last, have linear residual connections. This design simplifies the neural network's learning process, improving gradient propagation efficiency and training speed. In contrast with learning original features, the network with residual connections learns feature differences. That is, each residual block within the network need not learn complex functional relationships but instead simpler ones, reducing the task's learning difficulty. The specific formula is as follows:

$$F(x) = H(x) - x \tag{1}$$

In the formula, H(X) represents a nonlinear transformation of a deep neural network, and X is the input of the network. F(X) is the residual calculated by the residual block, which represents the error of the network. By adding the input X and the residual, the output H(X) of the network is obtained, that is, H(X) = X + F(X). The residual block can help the model to reduce the learning difficulty of the task. This is because the residual

block can help the model converge faster during training, and it can also avoid problems such as gradient disappearance and gradient explosion. In addition to the aforementioned benefits, residual connections in the network can partially mitigate the undesirable effects of neural network degradation while simultaneously enhancing generalization ability. In essence, the Xception system is a linear stack of depth-separable convolutional layers utilizing residual connections. This design simplifies researchers' ability to define and adjust network parameters for practical applications and delivers promising results across various fields.

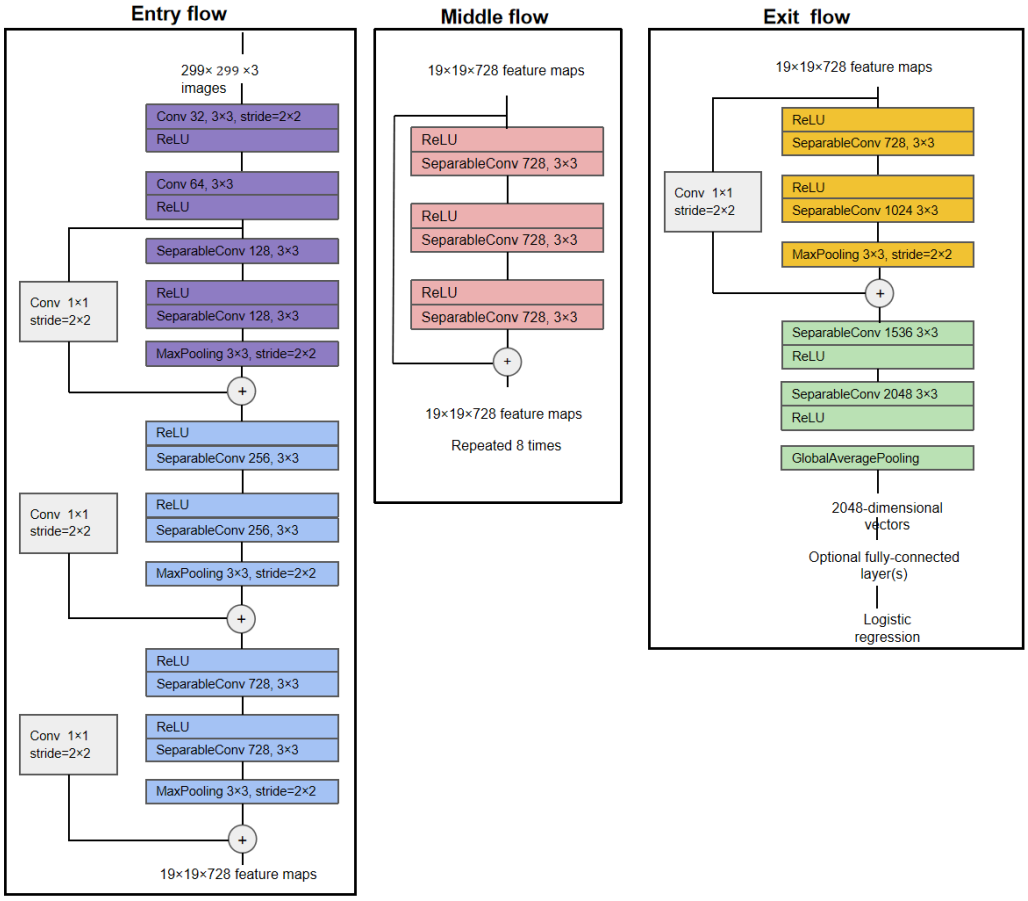


Figure 6. Network structure diagram of the Xception system.

During model training using the Xception system, the features that have been extracted by DeepID V3 are first fed into the entry stream for processing. These processed features are then passed through eight stacked intermediate flows before entering the exit flow to extract feature vectors. Subsequently, these feature vectors pass through the fully connected layer and logistic regression layer to conclude the overall model training process.

3.2.3. Video-Modality Dataset: The Extended Cohn–Kanade [32]

The extended Cohn–Kanade (CK+) dataset was released in 2010. It is a video sequence dataset developed by Cohn, Kanade, and others and applied to the field of emotion recognition. This dataset has been optimized and refined to address three limitations of the Cohn–Kanade dataset released by the same team in 2000, which is used as the video modality in the multimodal dataset in the study of this paper. The data are collected from 123 subjects, a total of 593 video sequence samples containing human facial expressions, including seven emotions, namely happy, sad, angry, fear, disgust, surprise, and neutral.

The expression of facial emotions is complex. The expression of specific emotions must have a process from brewing to eruption, and then to fading. No emotion appears

suddenly and without warning. The expression of emotions is affected by many factors. In this paper, 593 video sequences in the CK+ dataset are retrieved by a manual FACS encoder, and the number of occurrences of the seven emotions is counted by the peak frame at the moment of the emotional outburst in the video sequence. The distribution of the seven emotions in the 593 video sequences is shown in Table 3. It should be noted that this paper only uses the samples of the first six emotions for video-modality emotion-recognition model training.

Table 3. Distribution table of seven emotional labels in the CK+ dataset.

Emotional Label	Happy	Sad	Angry	Fear	Disgust	Surprise	Neutral
The number of occurrences	69	28	46	25	59	83	146

3.3. Implementation Ideas of Multimodal Fusion

3.3.1. Overview of Multimodal Fusion Methods and Ideas

Multimodal fusion technology has garnered extensive attention from researchers due to its convenience for various multimedia analysis tasks. In the case of multimodal emotion recognition, the integration of multiple modalities, their related characteristics, or intermediate decisions is referred to as multimodal fusion. While multimodal fusion overcomes the limitations of incomplete emotional features in the single modality, it also introduces new problems. The benefits of multimodal fusion come with certain costs and complexities during analysis, which are caused by the different characteristics involved in multimodal fusion. Thus, selecting an appropriate fusion strategy has become the core issue of multimodal fusion. Only by choosing an appropriate fusion strategy, comprehensively considering the implementation cost and model performance, and finding a balance point can the advantages of multimodal fusion be brought into play, such that good results can be obtained in emotion-recognition tasks.

Multimodal fusion methods can be categorized from multiple perspectives, including early-fusion methods, late-fusion methods [41], and hybrid-fusion methods [42,43].

(1) Early-fusion methods include data-layer fusion methods and feature-layer fusion methods. In the data-layer fusion method, the original data of different modalities are merged, and the classifier directly classifies them. The feature-layer fusion method entails fusion at the feature layer after the feature extraction of each modal dataset. For instance, multimodal feature fusion methods have been designed based on wavelet transform and PCA [44,45].

(2) The late-fusion method, also known as the decision-making level fusion method, accomplishes multimodal fusion in the late stage of the multimodal emotion-recognition process. Each modality can be trained using a different model, each mode is independent of each other in the stage before the decision-making level. Based on the characteristics of each modality and the actual research needs, the optimal model suitable for different modes can be selected, and finally, fusion is realized at the decision-making level.

(3) The hybrid-fusion method combines the advantages of both early- and late-fusion methods by fusing some of the modal information together in the data preprocessing stage while feeding the other modal information into different classifiers for learning and prediction during training and testing to achieve better fusion results. A flow chart of this method is shown in Figure 7.

Hybrid-fusion methods encompass both feature-level fusion and decision-level fusion methods. This method attempts to combine the respective advantages of the previous two fusion methods, The hybrid-fusion method combines the advantages of both early- and late-fusion methods by fusing some of the modal information together in the data preprocessing stage while feeding the other modal information into different classifiers for learning and prediction during training and testing to achieve better fusion results. In multimodal emotion-recognition research, the aforementioned three methods have their

individual merits and demerits. It is essential to choose an appropriate fusion method based on different research contexts.

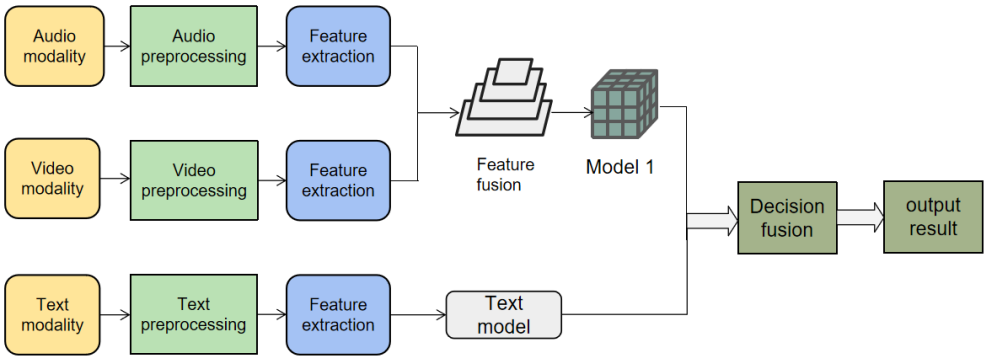


Figure 7. Flow chart of multimodal emotion recognition based on the hybrid-fusion method.

There is no absolute advantage for any fusion method. In actual tasks, various factors need to be comprehensively considered. This paper primarily explores the multimodal deep learning emotion-recognition algorithm and its applications. Based on the progress of previous research work and the need for later system expansion, the late-fusion method—which is more flexible in terms of modal expansion—is chosen to carry out this research on multimodal emotion recognition, considering the asynchronous nature of the dataset used in the study. Late fusion of audio and video modalities is attempted in late fusion to verify the effectiveness and feasibility of multimodal fusion.

3.3.2. Late-Fusion Method Based on Mean Thought

By utilizing the average value idea [46], the output results of both the audio- and video-modality models are averaged to acquire the final output results. The implementation process is illustrated in Figure 8 below.

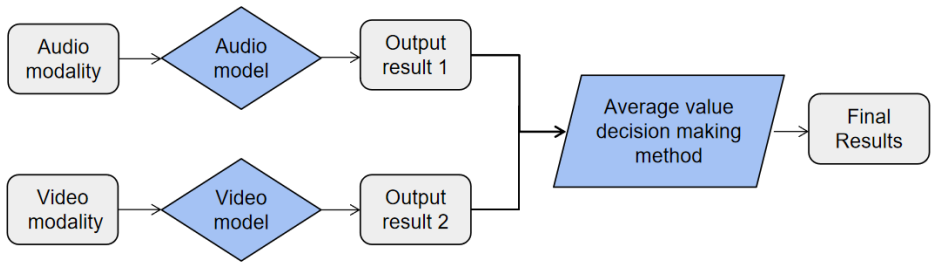


Figure 8. Flowchart of implementing the late-fusion method based on the average value idea.

In making the mean thought decision, firstly, the test data are input and the respective outputs are predicted using the prediction models obtained above for the audio modality, as well as the video-modality emotion recognition. Then, using the mean-value idea, the paper assumes that the reflections of the modalities on the expression of emotions have the same importance level, and the final output is obtained. The decision-level fusion of audio and video modalities is realized via the average decision method, as validated through experiments. The experimental results are presented in Table 4 below.

As shown in the table above, after the decision-making layer fusion, the accuracy rate of the six-category emotion-recognition task improves as compared to the single video modality, with an average accuracy rate of 0.7166. However, this decision-making method does not consider the varying importance levels of each modality's response to emotional expression. Thus, to address this challenge, this paper proposes a weight-adaptation-based late fusion decision-making method.

Table 4. Confusion matrix of late-fusion experiment results based on the idea of averages.

Unit: %	Happy	Sad	Angry	Fear	Disgust	Surprise
happy	66	11	21	2	0	0
sad	0	65	12	13	10	0
angry	0	13	79	8	0	0
fear	3	15	0	75	7	0
disgust	0	0	17	13	68	2
surprise	0	11	0	5	7	77

3.3.3. Late-Fusion Method Based on the Weight-Adaptive Idea

Considering the differences in the importance levels of the two modalities of audio and video in reflecting emotional expressions, this paper proposes a late-fusion decision method based on the idea of weight adaptation. When making decisions, the two modalities should be assigned different weights to better match the process of real human emotion expressions. Drawing inspiration from related research, this paper introduces a late-fusion approach based on a weight-adaptive concept; when making decisions, the method takes full account of individual variability and adaptively adjusts the respective weights according to the two modalities of the input information. Compared to the mean-based late-fusion method, the method proposed in this paper has the following two characteristics: (i) it takes into account the influence of individual variability on the experimental results; (ii) it adopts a weighted adaptive decision-making method.

The specific implementation process of the algorithm is demonstrated in Figure 9 below.

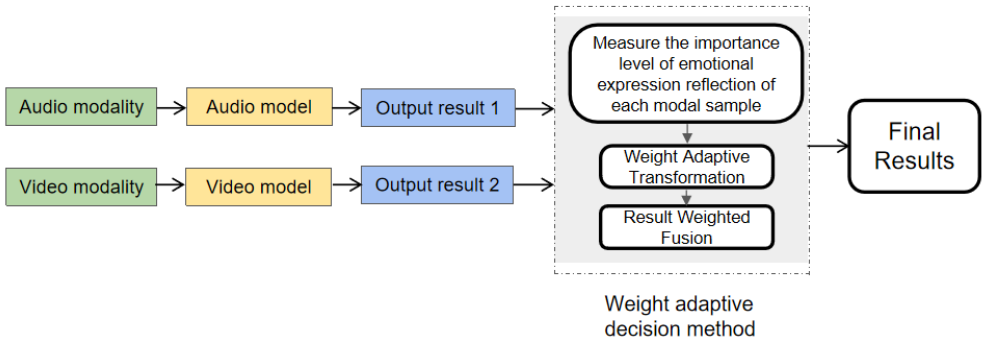


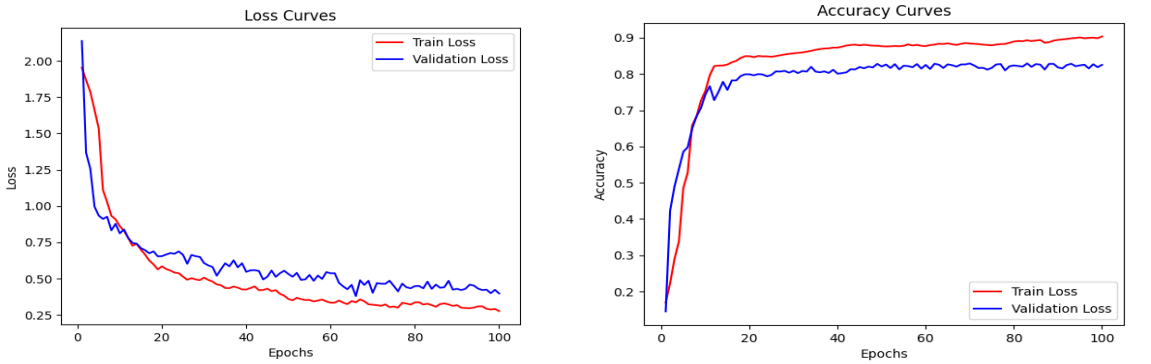
Figure 9. Flowchart of implementing late-fusion method based on the weight self-adaptive idea.

When employing weight-adaptive decision making, the process mainly includes the three steps shown in Figure 10 above. This paper introduces a parameter δ_j, which is used to indicate the importance level of each modality's response to emotional expression, where j denotes the modality serial number, and $j = 1, 2, \ldots, J$. Each sample corresponds to δ_j value in the decision-making phase, and the adaptive operation of weight distribution is realized according to the size of the value. The specific formula is as follows:

$$P = \{p_i | i = 1, 2, \ldots, I\} \tag{2}$$

$$\delta_j = 1 - d\left(P, P_{\frac{1}{I}}\right) \tag{3}$$

Among them, the vector P is composed of the predicted probability of each category label of the test sample, and I denotes the number of emotion types to be predicted, which is six in the research task of this paper. d represents the Euclidean distance, which is used to find the Euclidean distance between two vectors [47].



(**a**)Training set and validation set loss curve. (**b**)Training set and validation set accuracy curve.

Figure 10. Loss function curves and accuracy curves on the training set and validation set.

Through the above formula, the δ_j parameter value of each test sample is calculated, which can be used to measure the importance level of each modality's response to emotional expression. Next, the sigmoid function [48] is used to normalize the parameter δ_j and the parameter value is scaled to the interval [0, 1]. It can be expressed as:

$$\mu_j = 1 - \frac{1}{1 + e^{-a(\delta_j - b)}} \tag{4}$$

Through Formula (4), the adaptive operation of the weight is realized, and the importance level of any sample to the emotional expression is mapped to the weight value μ_i. For any sample of the test input, Formula (4) is used to realize the adaptive allocation operation of the weight. Next, according to the weight-adaptation results, the fusion of prediction probabilities can be realized to obtain the final fusion prediction probabilities. The specific formulae are as follows:

$$P_{last} = \{p_{last\_i} | i = 1, 2, \ldots, I\} \tag{5}$$

$$p_{last\_i} = \sum_{j=1}^{J} \frac{\mu_j}{\sum_{m=1}^{M} \mu_m} p_{i\_j} \tag{6}$$

After the decision fusion of weight self-adaptive thinking, the final prediction probability output vector P_{last} is obtained, and the emotion type corresponding to the maximum value of the probability in this vector is the emotion type finally predicted by the algorithm. In Formula (6) above, p_{i-j} represents the j-th mode and the predicted probability values for the i-th emotion type derived from predictions made by the corresponding modal emotion-recognition model.

4. Experiment and Results

4.1. Training and Evaluation of Audio-Modal Emotion-Recognition Models

For audio mode, this paper employs the "time-distributed CNNs + LSTMs" approach and conducts 100 rounds of training on the model by continuously tuning parameters and executing other operations. In order to avoid overfitting and improve the generalization ability of the model, the dataset is divided into a training set and a test set in the ratio of 8:2, and the cross-validation method is used to assist in adjusting the model parameters, resulting in a more stable and better performing model. The differences in loss function values and accuracy values are compared between the training and validation sets at the end of the 1st training round and at the end of the 100th training round, as shown in Table 5.

Table 5. Comparison table of loss function value and accuracy value on the audio-modality training set and verification set.

Number of Training Rounds	Training Set Loss	Training Set Accuracy	Validation Loss	Validation Set Accuracy
round 1	1.9524	0.1696	2.1361	0.1450
...	...	...	...	...
round 100	0.2774	0.9036	0.3986	0.8254

Based on the results in Table 5 above, it is evident that after 100 rounds of training, the model's loss function value on the training set decreases from 1.9524 to 0.2774, and the accuracy rate increases from 0.1696 to 0.9036. The loss function value on the validation set also decreases from 2.1361 to 0.3986, with the accuracy improving from 0.1450 to 0.8254. This demonstrates that while maintaining appropriate model complexity, the generalization ability of the model has been successfully improved. In order to better illustrate the change trends in the two datasets during the training process, Figure 10 displays the loss function value and accuracy value curves for the audio-modality training and validation sets for rounds 1–100.

Upon completion of the training process, the "time-distributed CNNs + LSTMs"-based audio-modality emotion-recognition model is obtained. The model file can then be used for prediction, and the detailed parameters of each layer within the network are recorded. The network parameters are listed in Table 6.

Table 6. Names and parameters of each layer in the audio-modal emotion-recognition network.

CNNs with Time-Distributed Layer:			
Conv2D_1: filters = 64 kernel_size = 3, 3 strides = 1, 1 padding = same activation = linear	Conv2D_2: filters = 64 kernel_size = 3, 3 strides = 1, 1 padding = same activation = linear	Conv2D_3: filters = 128 kernel_size = 3, 3 strides = 1, 1 padding = same activation = linear	Conv2D_4: filters = 128 kernel_size = 3, 3 strides = 1, 1 padding = same activation = linear
BatchNorm_1: axis = 3 momentum = 0.99 epsilon = 0.001	BatchNorm_2: axis = 3 momentum = 0.99 epsilon = 0.001	BatchNorm_3: axis = 3 momentum = 0.99 epsilon = 0.001	BatchNorm_4: axis = 3 momentum = 0.99 epsilon = 0.001
Activation_1: activation = elu	Activation_2: activation = elu	Activation_3: activation = elu	Activation_4: activation = elu
MaxPool_1: pool_size = 2, 2 padding = same strides = 2, 2	MaxPool_2: pool_size = 4, 4 padding = same strides = 4, 4	MaxPool_3: pool_size = 4, 4 padding = same strides = 4, 4	MaxPool_4: pool_size = 4, 4 padding = same strides = 4, 4
Dropout_1: rate = 0.2	Dropout_2: rate = 0.2	Dropout_3: rate = 0.2	Dropout_4: rate = 0.2
	Flatten: input = (None, 5, 1, 1, 128) output = (None, 5, 128)		
LSTM_1: units = 256 activation = tanh dropout = 0.2	LSTM_2: units = 256 activation = tanh dropout = 0.2		
Dense: units = 7 activation = softmax			

After training, this paper successfully constructs an audio-modal emotion-recognition model based on the "time-distributed CNNs + LSTMs" scheme and records the detailed

parameters of each layer in the model. In the test phase, the performance of the model was evaluated using the RAVDESS dataset; six emotions were classified and predicted; and the "time-distributed CNNs + LSTMs" scheme was combined with the "SVM on global statistical features" [49] program and the "hybrid LSTM-transformer model" [50] in a comparative experiment. The specific effects are shown in Table 7 below.

Table 7. Accuracy of RAVDESS dataset in various network models.

Model Scheme	Accuracy
SVM on Global Statistical Features	68.3%
Hybrid LSTM-Transformer Model	75.6%
Time-Distributed CNNs + LSTMs	80.4%

The results show that the combination of the "time-distributed CNNs + LSTMs" network and the log-mel spectrogram features of audio samples, compared with the traditional SVM combined with low-level statistical features, significantly improves the performance of the model. A six-classification accuracy 80.4% is achieved, a 12% improvement over the traditional scheme. It is also 4.8% more accurate than the well-performing hybrid LSTM-transformer model network model.

In order to further verify the generalization ability of the model, the IEMOCAP dataset with a larger data volume and audio duration is selected to be verified on the two model networks in the verification stage. This article excerpts the six emotions in the audio part of the IEMOCAP dataset to classify and predict based on the scheme of this article, calculates the accuracy of the six categories, and compares the verification results of the IEMOCAP dataset on the attention-oriented parallel CNN encoders network [51]. The specific effects are shown in Table 8 below.

Table 8. Accuracy of IEMOCAP dataset in various network models.

Model Scheme	Accuracy
Attention-Oriented Parallel CNN Encoders	71.11%
Time-Distributed CNNs + LSTMs	73.82%

The results show a slight decrease in accuracy when faced with the more complex IEMOCAP dataset compared to the RAVDESS dataset, probably due to the greater complexity of the IEMOCAP dataset and the long sample fragment times. However, it is also 2.5% more accurate than the attention-oriented parallel CNN encoders with a good performance.

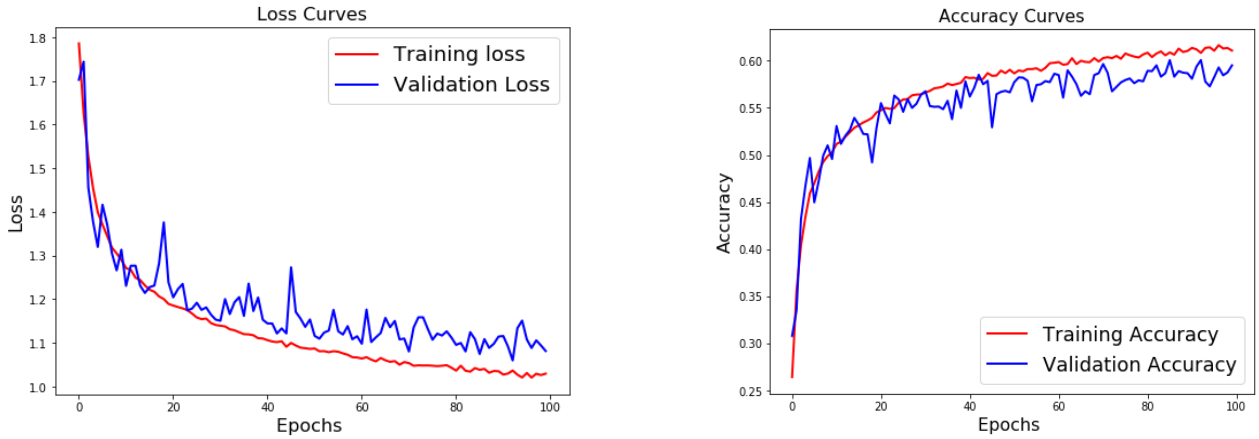
In addition, the model complexity is successfully controlled within a reasonable range, and it can be quickly deployed for sentiment prediction. Our method meets the criteria for practical application, and in the field of emotion computing, using audio to analyze human emotions has greater advantages. Therefore, audio modalities should be given higher voting weights when fusing models.

4.2. Training and Evaluation of Video-Modality Emotion-Recognition Model

For the video mode, this paper adopts the "DeepID V3 + Xception architecture" scheme, and conducts 100 rounds of training on the experimental platform. The cross-validation method is also adopted for training. As mentioned earlier, the Xception structure has excellent working principles and features. In order to further improve the model performance of Xception in emotion-recognition tasks, optimization strategies such as data augmentation, early stopping, learning rate decay, L2 regularization, and class weight balance are optimized and adjusted.

After adopting relevant optimization strategies, the loss function value drops from 1.7515 to 1.0031, and the accuracy rate increases from 0.2968 to 0.6453 on the training set during the whole training process. At the same time, on the validation set, the loss function

value drops from 2.3387 to 1.0921, and the accuracy rate increases from 0.2644 to 0.5965. The 1–100 round loss function value change curve of the video-modality training set and the verification set and the numerical change curve of the accuracy rate are shown in Figure 11.



(**a**)Loss function curves on the video-modality training set and verification set.

(**b**)Accuracy curves on the video-modality training set and verification set.

Figure 11. Loss function curves and accuracy curves on the video-modality training set and verification set.

In addition to the aforementioned analysis, this paper also conducts comparative experiments with several schemes proposed by other researchers. These include SVM on HOG features, SVM on facial landmarks features, SVM on facial landmarks and HOG features, SVM on sliding window landmarks and HOG, and Inception architecture. The experiment employs the SVM classifier and utilizes methods such as HOG features, face feature point features, and their combined features and sliding window to conduct the experiments.

Finally, this paper also uses the Inception architecture network to conduct experiments. Table 9 presents a comparison of the experimental results of the scheme proposed in this paper and the abovementioned comparison schemes.

Table 9. Performance comparison between DeepID V3 + Xception architecture and other scheme models.

Model Scheme	Accuracy
SVM on HOG Features	32.8%
SVM on Facial Landmarks Features	46.4%
SVM on Facial Landmarks and HOG Features	47.5%
SVM on Sliding Window Landmarks and HOG	24.6%
Inception Architecture	59.5%
DeepID V3 + Xception Architecture	64.5%

After comparison, the DeepID V3 + Xception architecture is tested on the CK+ dataset using a combination of HOG features and facial feature points. The results show that the model achieves an accuracy of 64.5%, which is a 5% improvement over the Inception architecture. Furthermore, the model size is only 15 MB. It should be noted that the proposed scheme in this paper has much room for improvement in the video-modal emotion-recognition task due to the quality of the dataset, among other reasons. Therefore, when implementing model fusion, video modalities should be assigned larger weights to further improve the model performance.

4.3. Modal Experiment Results Comparison and Evaluation

The late-fusion decision method based on the idea of weight adaptation proposed in this paper is verified through experiments. The experimental results are presented in Table 10 below.

Table 10. Confusion matrix of late-fusion experiment results based on a weight-adaptive idea.

Unit: %	Happy	Sad	Angry	Fear	Disgust	Surprise
happy	82	7	11	0	0	0
sad	0	85	8	5	2	0
angry	0	9	84	7	0	0
fear	3	3	0	88	6	0
disgust	0	0	9	8	83	0
surprise	0	6	0	4	6	84

The above table reveals that after incorporating individual differences and the significance level of each modality in emotional expression, the model's performance improves substantially, with the average accuracy rate of six classifications reaching 0.8433. We compare the results obtained by all the algorithms proposed in this paper in Figure 12.

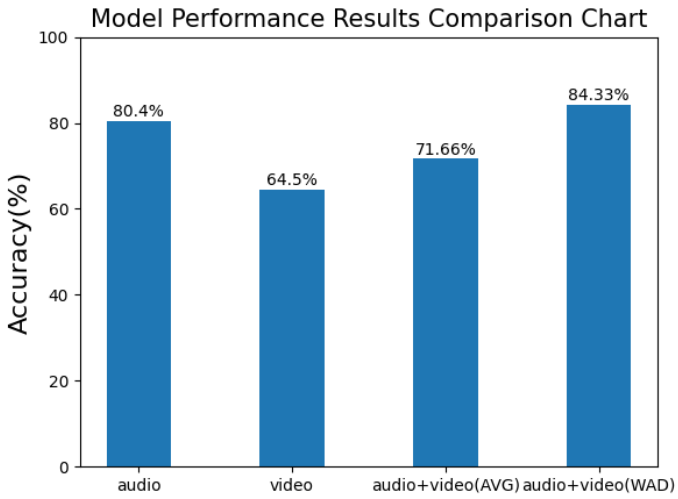


Figure 12. Comparison chart of model performance results in this paper.

The above figure reveals that in the single-modal emotion-recognition task, the accuracy of the algorithm presented in this paper is as follows: 80.40% for the audio mode and 64.50% for the video mode. The results indicate that the proposed algorithm utilizing audio modality performs better.

This paper further explores the multimodality late-fusion method by integrating audio and video modalities. Two late-fusion techniques based on average and weight-adaptive ideas are designed and used separately to predict the accuracy of the emotion-recognition model. As per the experimental results, the latter algorithm yields a prediction accuracy of 84.33%, surpassing the prediction accuracy of all single-mode models. Compared to the best single-modal model, the accuracy rate has improved by approximately 4%; findings which demonstrate the feasibility and effectiveness of multimodal fusion in emotion-recognition tasks.

In order to better reflect and verify the advantages of the algorithm proposed in this paper, the model training results obtained in this paper are compared with the experimental results of deep convolutional neural networks [52] and an audio–visual and emotion-centered network [53]. The experimental results are shown in Table 11.

Table 11. Performance comparison table between the scheme of this paper and other multimodal schemes.

Model Scheme	Accuracy
Deep Convolutional Neural Networks	77.64%
Audio–Visual and Emotion-Centered Network	81.28%
Our Scheme	84.33%

After the experimental results, it can be seen that the experimental results of multimodal emotion recognition in this chapter are the highest, and the accuracy rate of the model used reaches 84.33%, which is better than other network models, reflecting the superior performance of the algorithm proposed in this paper.

5. Conclusions

This paper delves into the direction of emotion recognition in the field of emotion computing through a detailed study of multimodal emotion recognition using audio and video data and deep-learning-based methods. Initially, we describe the construction process of the emotion-recognition models for each modality. In order to validate the model construction schemes proposed in this paper, experimental verification is conducted by comparing the model results with those obtained by previous researchers.

In order to achieve this, the paper introduces the comparative experimental methodology and analyzes and evaluates the outcomes. Finally, we provide a brief overview of the three fusion methods, i.e., early fusion, late fusion, and hybrid fusion, and realize the late fusion of the audio and video modalities based on two distinct ideas. The specific contributions and innovations of this paper are as follows:

(1) For the task of audio-modal emotion recognition, this paper determines the model construction scheme of "time-distributed CNNs + LSTMs". Firstly, the audio-signal preprocessing is carried out on the audio-modality data, and then the log-mel spectrogram feature extraction is carried out on the audio sequence. Next, this paper performs a time-distributed framing operation on the data sample to adapt it to the subsequent network. Through model training and related comparative experiments, it reflects and proves the superior performance of the network model.

(2) For the video-modal emotion-recognition task, the model construction scheme of "DeepID V3 + Xception architecture" is determined. In the experiment, the image preprocessing of the video-modality data is first performed, and then the process of HOG feature extraction is performed. Finally, the role of DeepID V3 and the Xception system in the video feature-extraction network is introduced. At the same time, the residual network design is introduced, and the design of the network structure and the adjustment of related optimization strategies are carried out.

(3) In order to verify the audio- and video-modality emotion-recognition model construction scheme proposed in this paper, experimental verification is carried out, and the model construction scheme for each modality is compared with the existing common emotion-recognition algorithms. The results show that the emotion-recognition accuracy of the two modalities is increased by 12% and 5%, respectively, confirming the advantages and performance of the proposed algorithm. A late-fusion method based on the idea of weight self-adaptation is also attempted, which confirms the advantages of the multimodal fusion algorithm; the recognition accuracy is improved by nearly 4% on the basis of the optimal single-modal emotion-recognition algorithm proposed in this paper when compared with other multimodal network models, proving the superiority of the algorithm proposed herein.

In order to improve the recognition accuracy of the proposed multimodal model, this paper chooses to focus on the accuracy of emotion recognition, mainly by comparing the experimental results of different schemes and the accuracy of different datasets to verify the reliability and generalization of the scheme. It does not include other accuracy metrics for evaluating performance reliability, such as G-Mean, precision, recall, F1 value,

Matthews correlation coefficient (MCC), and the area under the precision-recall curve (PR AUC). Although the accuracy index cannot fully reflect the performance of our model, we believe that the accuracy rate is still one of the most basic and simplest evaluation indicators. In many cases, the accuracy rate is still one of the main indicators for evaluating the performance of classifiers and remains of great reference value. In addition, it is helpful to our work when we have insufficient information. At the same time, this paper compares the results of different datasets and other network model schemes, proving the effectiveness and accuracy of the scheme proposed in this paper resulting in improvement to varying degrees.

Future prospects: The research on the multimodal emotion-recognition algorithm in this paper is based on the two modalities of audio and video. In view of the shortcomings of certain existing algorithms, corresponding improvement strategies and the algorithm scheme of this paper are proposed. After experimental verification, the validity and feasibility of the algorithm in this paper are proved. However, the accuracy rate is only one of the main indicators to evaluate the performance of the classifier. Although the reliability of the scheme is verified through comparative experiments in this paper, in follow-up research, we should consider further evaluation indicators to verify the reliability of the model. At the same time, there are many problems that have not yet been covered in the research of this paper. There are still many problems in the field of multimodal emotion recognition which urgently require further study by researchers. Research on multimodal effective fusion is the core issue in the field of multimodal emotion recognition. Realizing the effective fusion of multimodal information has always been a popular research direction in this field. The fusion results determine the upper limit of the performance of emotion recognition. Therefore, breakthroughs in multimodal fusion methods will give a huge boost to the development of the field and, at the same time, pose a huge challenge to researchers, who will focus on innovations in multimodal fusion methods in the next phase of research. In the next stage of research, we will focus on the innovation of multimodal fusion methods.

Emotion-recognition technology has a wide range of applications in the field of driver safety. In subsequent in-depth research, the driver's emotion will be identified based on the emotion-recognition algorithm proposed in this paper, the attention mechanism will be used to weight the integrated feature vector, and the emotion-classification results of multiple modalities will be integrated to obtain the final emotion-recognition results, resulting in the integrated emotion state of the driver. Then, early warning tips will be given according to the driver's emotion classification, thus improving driver safety and security and promoting the development of intelligent transportation.

Author Contributions: Conceptualization, D.Z. and Y.C.; methodology, Y.C.; software, Y.C.; validation, S.W.; formal analysis, S.W.; investigation, L.W.; writing—original draft, Y.C.; writing—review and editing, D.Z.; visualization, S.W.; supervision, L.W.; project administration, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This study examines the publicly available Ryerson affective language and song audio–visual dataset (RAVDESS) and the extended Cohn–Kanade (CK+) datasets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
2. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
3. Atsavasirilert, K.; Theeramunkong, T.; Usanavasin, S.; Rugchatjaroen, A.; Boonkla, S.; Karnjana, J.; Keerativittayanun, S.; Okumura, M. A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms. In Proceedings of the 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 30 October–1 November 2019; pp. 1–4.

4. Salian, B.; Narvade, O.; Tambewagh, R.; Bharne, S. Speech Emotion Recognition using Time Distributed CNN and LSTM. *ITM Web Conf.* **2021**, *40*, 03006. [CrossRef]
5. Mao, K.; Zhang, W.; Wang, D.B.; Li, A.; Jiao, R.; Zhu, Y.; Wu, B.; Zheng, T.; Qian, L.; Lyu, W. Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN. *IEEE Trans. Affect. Comput.* **2022**. [CrossRef]
6. Kobayashi, T. BFO meets HOG: Feature extraction based on histograms of oriented pdf gradients for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 747–754. [CrossRef]
7. Albiol, A.; Monzo, D.; Martin, A.; Sastre, J.; Albiol, A. Face recognition using HOG–EBGM. *Pattern Recognit. Lett.* **2008**, *29*, 1537–1543. [CrossRef]
8. Kaiser, L.; Gomez, A.N.; Chollet, F. Depthwise separable convolutions for neural machine translation. *arXiv* **2017**, arXiv:1706.03059.
9. Poulose, A.; Reddy, C.S.; Kim, J.H.; Han, D.S. Foreground Extraction Based Facial Emotion Recognition Using Deep Learning Xception Model. In Proceedings of the 2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), Jeju Island, Republic of Korea, 17–20 August 2021; pp. 356–360. [CrossRef]
10. Sun, Y.; Liang, D.; Wang, X.; Tang, X. Deepid3: Face recognition with very deep neural networks. *arXiv* **2015**, arXiv:1502.00873.
11. Yuan, Z. Face detection and recognition based on visual attention mechanism guidance model in unrestricted posture. *Sci. Program.* **2020**, *2020*, 8861987. [CrossRef]
12. Vielzeuf, V.; Lechervy, A.; Pateux, S.; Jurie, F. Centralnet: A multilayer approach for multimodal fusion. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018. [CrossRef]
13. Mehrabian, A. *Silent Messages: Implicit Communication of Emotions and Attitudes*; Wadsworth Pub, Co.: Belmont, CA, USA, 1981.
14. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
15. Jain, D.K.; Shamsolmoali, P.; Sehdev, P. Extended deep neural network for facial emotion recognition. *Pattern Recognit. Lett.* **2019**, *120*, 69–74. [CrossRef]
16. Balasubramanian, B.; Diwan, P.; Nadar, R.; Bhatia, A. Analysis of facial emotion recognition. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 23–25 April 2019; pp. 945–949.
17. Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors* **2019**, *19*, 1897. [CrossRef]
18. Liu, Y.; Zhang, X.; Li, Y.; Zhou, J.; Li, X.; Zhao, G. Graph-based facial affect analysis: A review. *IEEE Trans. Affect. Comput.* **2022**. [CrossRef]
19. Ibrahim, Y.A.; Odiketa, J.C.; Ibiyemi, T.S. Preprocessing technique in automatic speech recognition for human computer interaction: An overview. *Ann. Comput. Sci. Ser.* **2017**, *15*, 186–191.
20. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
21. Staudemeyer, R.C.; Morris, E.R. Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv* **2019**, arXiv:1909.09586.
22. Lei, X.; Pan, H.; Huang, X. A dilated CNN model for image classification. *IEEE Access* **2019**, *7*, 124087–124095. [CrossRef]
23. Slimi, A.; Nicolas, H.; Zrigui, M. Hybrid Time Distributed CNN-Transformer for Speech Emotion Recognition. In Proceedings of the 17th International Conference on Software Technologies ICSOFT, Lisbon, Portugal, 11–13 July 2022; pp. 11–13.
24. Zhao, H.; Gao, Y.; Xiao, Y. Upgraded Attention-Based Local Feature Learning Block for Speech Emotion Recognition. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, 11–14 May 2021, Proceedings, Part II*; Springer International Publishing: Cham, Switzerland, 2021; pp. 118–130.
25. Sharma, S.; Sharma, S.; Athaiya, A. Activation functions in neural networks. *Towards Data Sci.* **2017**, *6*, 310–316. [CrossRef]
26. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
27. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
28. Liu, Z.; Luo, S.; Li, W.; Lu, J.; Wu, Y.; Sun, S.; Li, C.; Yang, L. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv* **2020**, arXiv:2011.10185.
29. Segundo, M.P.P.; Silva, L.; Bellon, O.R.P.; Queirolo, C.C. Automatic face segmentation and facial landmark detection in range images. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2010**, *40*, 1319–1330. [CrossRef]
30. Qin, J.; Huang, Y.; Wen, W. Multi-scale feature fusion residual network for single image super-resolution. *Neurocomputing* **2020**, *379*, 334–342. [CrossRef]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2016**, 770–778. [CrossRef]
32. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition–Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.

33. Tang, P.; Wang, H.; Kwong, S. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing* **2017**, *225*, 188–197. [CrossRef]
34. Yu, Z.; Dong, Y.; Cheng, J.; Sun, M.; Su, F. Research on Face Recognition Classification Based on Improved GoogleNet. *Secur. Commun. Netw.* **2022**, *2022*, 7192306. [CrossRef]
35. Gu, S.; Ding, L. A complex-valued vgg network based deep learing algorithm for image recognition. In Proceedings of the 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP), Wanzhou, China, 9–11 November 2018; pp. 340–343.
36. Agarap, A.F. Deep learning using rectified linear units (relu). *arXiv* **2018**, arXiv:1803.08375.
37. Szegedy, C.; Ioffe, S.; Vanhoucke, V. Inception-ResNet and the impact of residual connections on learning. *arXiv* **2016**, arXiv:1602.07261.
38. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2016**, 2818–2826. [CrossRef]
39. Tio, A.E. Face shape classification using inception v3. *arXiv* **2019**, arXiv:1911.07916.
40. Kang, K.; Gao, F.; Feng, J. A new multi-layer classification method based on logistic regression. In Proceedings of the 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 8–11 August 2018; pp. 1–4.
41. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs. late fusion in multimodal convolutional neural networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–6.
42. Amer, M.R.; Shields, T.; Siddiquie, B.; Tamrakar, A.; Divakaran, A.; Chai, S. Deep multimodal fusion: A hybrid approach. *Int. J. Comput. Vision* **2018**, *126*, 440–456. [CrossRef]
43. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2022**. [CrossRef]
44. Mukhedkar, M.M.; Powalkar, S.B. Fast face recognition based on Wavelet Transform on PCA. In Proceedings of the 2015 International Conference on Energy Systems and Applications, Pune, India, 30 October–1 November 2015; pp. 761–764.
45. Abdulrahman, M.; Gwadabe, T.R.; Abdu, F.J.; Eleyan, A. Gabor wavelet transform based facial expression recognition using PCA and LBP. In Proceedings of the 2014 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, Turkey, 23–25 April 2014; pp. 2265–2268.
46. Abdullah, S.M.S.A.; Ameen, S.Y.A.; Sadeeq, M.A.; Zeebaree, S. Multimodal emotion recognition using deep learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 52–58. [CrossRef]
47. Lee, L.H.; Wan, C.H.; Rajkumar, R.; Isa, D. An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization. *Appl. Intell.* **2012**, *37*, 80–99. [CrossRef]
48. Menon, A.; Mehrotra, K.; Mohan, C.K.; Ranka, S. Characterization of a class of sigmoid functions with applications to neural networks. *Neural Netw.* **1996**, *9*, 819–835. [CrossRef] [PubMed]
49. Jayalakshmi, S.; Chandrakala, S.; Nedunchelian, R. Global statistical features-based approach for acoustic event detection. *Appl. Acoust.* **2018**, *139*, 113–118. [CrossRef]
50. Andayani, F.; Theng, L.B.; Tsun, M.T.; Chua, C. Hybrid LSTM-transformer model for emotion recognition from speech audio files. *IEEE Access* **2022**, *10*, 36018–36027. [CrossRef]
51. Makhmudov, F.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047. [CrossRef]
52. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]
53. Ghaleb, E.; Popa, M.; Asteriadis, S. Multimodal and temporal perception of audio-visual cues for emotion recognition. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 552–558.