symmetry

MDPI

Article

Multimodal Emotion Recognition Using the Symmetric S-ELM-LUPI Paradigm

Lingzhi Yang [1,2], Xiaojuan Ban [1,*], Michele Mukeshimana [3] and Zhe Chen [4]

[1] School of Computer and Communication Engineering, Beijing Key Laboratory of Knowledge Engineering for Materials Science, University of Science and Technology Beijing, Beijing 10083, China; yanglingzhi@citicsteel.com
[2] Citic Pacific Special Steel Holdings Qingdao Special Iron and Steel Co., Ltd., Qingdao 266000, China
[3] Faculty of Engineering Sciences, University of Burundi, P.O. Box 1550 Bujumbura, Burundi; Mukeshimana@ustb.edu.cn
[4] Qingdao Hisense Group Co., Ltd., Qingdao 266000, China; chenzhe@hisense.com
* Correspondence: banxj@ustb.edu.cn; Tel.: +86-158-0658-0782

Received: 21 February 2019; Accepted: 30 March 2019; Published: 4 April 2019

Abstract: Multimodal emotion recognition has become one of the new research fields of human-machine interaction. This paper focuses on feature extraction and data fusion in audio-visual emotion recognition, aiming at improving recognition effect and saving storage space. A semi-serial fusion symmetric method is proposed to fuse the audio and visual patterns of emotional recognition, and a method of Symmetric S-ELM-LUPI is adopted (Symmetric Sparse Extreme Learning Machine-Learning Using Privileged Information). The method inherits the generalized high speed of the Extreme Learning Machine, and combines this with the acceleration in the recognition process by the Learning Using Privileged Information and the memory saving of the Sparse Extreme Learning Machine. It is a learning method, which improves the traditional learning methods of examples and targets only. It introduces the role of a teacher in providing additional information to enhance the recognition (test) without complicating the learning process. The proposed method is tested on publicly available datasets and yields promising results. This method regards one pattern as the standard information source, while the other pattern as the privileged information source. Each mode can be treated as privileged information for another mode. The results show that this method is appropriate for multi-modal emotion recognition. For hundreds of samples, the execution time is less than one percent seconds. The sparsity of the proposed method has the advantage of storing memory economy. Compared with other machine learning methods, this method is more accurate and stable.

Keywords: multimodal emotion recognition; symmetric S-ELM-LUPI paradigm; human-machine interaction

1. Introduction

The development of interactive technologies has enabled humans to communicate with computing devices in a more natural way. An important part of natural interaction is "emotion", which takes a key role in the natural interaction of people. Emotional recognition research is a challenging task, we will help improve the interaction between computers and people if we try our best to do emotional recognition research, such as automatic counseling system, automatic answering machine. We can help patients, the elderly and the disable in a real sense, if we can identify the user's emotion effectively.

The expression of human emotions is multi-channel, and humans infer the meaning of expression through different clues, such as facial expressions, sound features, speech content, posture language of gestures and gestures. There are also physical reactions of the body, such as heart beat rhythm,

heart blood pressure, and brain function. And physical examination equipment is required to do these tests.

At present, the computer has the ability of recording and processing user's input information. The input information includes voice information, video information, bio-electronic information, text information and so on. However, the recognition ability of computers is lower than people's recognition ability certainly. One of the main problems is that how to express the true intentions with the people's multiple performance. So it is especially important to identify people's emotions through multiple channels.

2. Related Work

Multimodal emotion recognition refers to the recognition of emotions through a combination of two or more modal information. Humans have a variety of ways to express emotions and combine multiple sources of information to recognize the emotions of others. Computers try to combine information sources in multiple ways to achieve the level of human emotion recognition.

Poria S., Chaturvedi I., Cambria E. et al. [1] proposed a temporal Convolutional Neural Network to extract features from visual and text modalities. Tzirakis P., Trigeorgis G., Nicolaou M.A. et al. [2] proposed an emotion recognition system using auditory and visual modalities. They utilize a Convolutional Neural Network to extract features from the speech, while for the visual modality, a deep residual network of 50 layers. Latha G.C.P. and Priya M.M. [3] also use Convolution Neural Network Model with multiple signal processing features

Huang Y., Yang J., Liao P. et al. [4] proposed two multimodal fusion methods between brain and peripheral signals for emotion recognition. The input signals are electroencephalogram and facial expression. Torres-Valencia C. et al. [5] proposed SVM-based feature selection methods for emotion recognition from multimodal data. Chan W.L., Song K.Y., Jeong J. et al. [6] proposed convolutional attention networks for multimodal emotion recognition from speech and text data.

Learning Using Privileged Information (LUPI) paradigm has been largely applied with Support Vector Machine plus (SVM+) algorithm. Feyereisl et al. [7] has worked on the importance and incorporation of privileged information in cluster analysis and their method has improved the clustering performance. Ji et al. [8] have proposed a multi-task multi-class by learning using privileged information on support vector machines. In their work, they have obtained improved results for multitask multiclass problems. Liu et al. [9] have empirically demonstrated an improvement of v-Support Vector for Classification and Regression by using the privileged information to solve practical problems in the experiments. Recently, Wang et al. [10] recognized audience's emotion from EEG signals with the help of the stimulus videos, and tagged the video's emotions with the aid of Electro Encephalogram (EEG) signals. Their implicit fusion has performed comparatively, or even better than the methods based on explicit fusion. In all the aforementioned-experiments, the most used algorithm is the SVM+ algorithm (Algorithm 1). In these experiments, the complexity relating to SVM parameterization considerably increased the training time. As an alternative solution, the exploitation of privileged information has been extended to the Extreme Learning Machine (ELM) method which is faster and has less parameterization.

The research work of the above scholars is to improve the classification accuracy by changing the extraction features and trying different information fusion models. How to establish a real-time and stable automatic identification system is an important problem that needs to be solved in multi-modal emotion recognition design. The system automatically detects, models, and generates natural interactions. The research focus of this paper is feature extraction and data fusion to achieve the purpose of improving classification accuracy and shortening recognition time.

In summary, multimodality is a typical feature of emotional expression. The system with multi-modal emotion automatic recognition mainly involves information fusion technology. Its main idea is to fuse raw data. However, there are many finished results in terms of literal meaning, because it is easier to implement than the fusion of data and features. In contrast, the way data and feature fusion

are identified is more accurate. In the field of multimodal emotion recognition, many achievements have been made, but natural, immediate and accurate emotional interaction is still an elusive goal. The advances in automatic feature extraction techniques support the study of new machine learning algorithms such as standing by vector machines and extreme learning machines. The research results have enhancement to promote the realization of natural instant emotional interaction.

3. The New Method of Symmetric S-ELM-LUPI (Symmetric Sparse Extreme Learning Machine-Learning Using Privileged Information) Using for Multimodal Emotion Recognition

During learning using the privileged information, the training set is composed of triplets, i.e., standards variables X, privileged variables X*, and their corresponding label Y, but the testing set comprehends standard variables X and the labels only. During the learning process of ELM, the vector X of the standard information is mapped into the hidden-layer feature space by h(x) and the vector X* of the privileged information is mapped into the hidden-layer correcting space by h*(x*). The two kernel functions h(x) and h*(x*), can be different, or the same. Figure 1 exemplified the single hidden layer feedforward neural networks (SFLNs) representation including the LUPI paradigm.

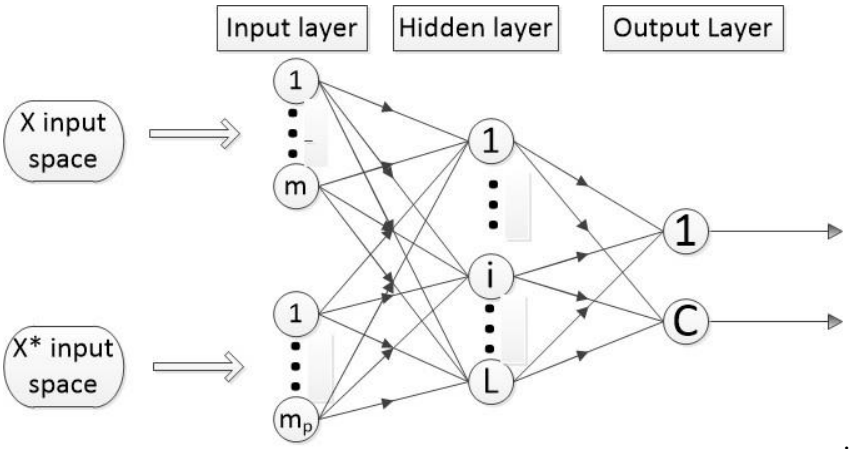


Figure 1. SFLNs with LUPI paradigm.

Figure 1 represents the flow of the processes in the ELM using privileged information with a standard space (X input space), of m-dimension; and the privileged information space (X* input space), of mp-dimension. There are L hidden nodes in the hidden layer and the C-classes. Figure 1 serves a simplified representation; the computation considers the two inputs spaces independent space to learn in parallel, and define two mapping functions h(x) and h*(x*). These functions can be the same or different. They are mapped into the same decision space.

The introduction of the LUPI model in S-ELM finds origins in the optimization-based ELM method. Based on the ELM, the slack variables are unknown to the learner. If there is an oracle who can give more information, they can be estimated by a correcting function defined by that additional information. The correcting function, which estimates the slack value, is computed as follows:

$$\xi(x) = \phi(x^*) = h^*(x^*)\beta^* \tag{1}$$

Hence, the problem of optimization of ELM becomes as follows:

$$\min_{\beta,\beta^*} L_p = \frac{1}{2}\left[\|\beta\|^2 + \gamma\|\beta^*\|^2\right] + C\sum_{i=1}^{N} h^*(x_i^*).\beta^* \tag{2}$$

$$s.t. : y_i(\beta.h(x_i)) \geq 1 - (h^*(x_i^*).\beta^*)$$

$$h^*(x_i^*).\beta^* \geq 0$$

where β^* is the correcting weight connecting the hidden node to the out-put node in the correcting space and γ is introduced for the regularization.

To minimize the above functional subject to constraints, The Lagrangian function is:

$$L_D = \frac{1}{2}\left[\|\beta\|^2 + \|\beta^*\|^2\right] + C\sum_{i=1}^{N} h^*(x_i^*).\beta^* - \sum_{i=1}^{N} \alpha_i\left[y_i(\beta.h(x_i)) - \left(1 - \left(h^*(x_i^*).\beta^*\right)\right)\right] - \sum_{i=1}^{N} u_i\left(h^*(x_i^*).\beta^*\right) \quad (3)$$

where $\alpha_i > 0$ and $\mu_i > 0$, are the Lagrangian multipliers and are non-negative values. In order to solve the above optimization problem, it is needed to find the saddle point of the Lagrangian (the minimum with respect to β, β^* and the maximum with respect to α_i and μ_i), i=1 ... N.

The KKT optimality condition of (3) are as follow:

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow \beta = \sum_{i=1}^{N} \alpha_i y_i h(x_i) \quad (4)$$

$$\frac{\partial L}{\partial \beta^*} = 0 \Rightarrow \beta^* = \frac{1}{\gamma}\left(\sum_{i=1}^{N} (\alpha_i + \mu_i - C)h^*(x_i)\right)$$

then the optimization problem becomes:

$$\min L_D(\alpha, \mu) = \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j y_i y_j K(x_i, x_j) + \frac{1}{2\gamma}\sum_{i,j=1}^{N} (\alpha_i + \mu_i - C)(\alpha_j + \mu_j - C)K^*(x_i^*, x_j^*) - \sum_{i=1}^{N} \alpha_i \quad (5)$$

$$s.t.: \sum_{i=1}^{N}(\alpha_i + \mu_i - C) = 0, \alpha_i \geq 0, \mu_i \geq 0 \quad (6)$$

where K and K* are two kernels in two different spaces namely, decision space and the correcting space. Then the decision function is computed as follows:

$$f(x) = h(x).\beta = \sum_{i=1}^{N} \alpha_i y_i h(x_i, x) \quad (7)$$

and the corresponding correcting function is:

$$\varphi(x^*) = h(x^*).\beta^* = \frac{1}{\gamma}\sum_{i=1}^{N} (\alpha_i + \mu_i - C)K^*(x_i^*, x^*) \quad (8)$$

In this computation, two kernels define similarity between two objects in different spaces (decision and correcting spaces). The decision function value depends directly on the kernel defined in the decision space. Still, it receives the contribution of the additional information knowledge through the computation of coefficient α which depends on the similarity measure in both spaces.

The proposed algorithm for Symmetric Sparse Extreme Learning Machine Learning Using Privileged Information is summarized as follows:

Input: Training set X, Privileged Information X*, L hidden number and activation functions g and g*.

Output: The prediction of the approximated function f(x).

(1) Random generation of the respective input weights W, and W*.

$$W = \begin{pmatrix} (a_1, b_1) \\ \vdots \\ (a_l, b_l) \end{pmatrix} \qquad W^* = \begin{pmatrix} (a_1^*, b_1^*) \\ \vdots \\ (a_l^*, b_l^*) \end{pmatrix} \tag{9}$$

(2) Calculate the hidden node output matrices H and H*;

$$H = G(W, X) = \begin{pmatrix} (a_1, b_1, x_1) & \dots & (a_l, b_l, x_1) \\ \vdots & \ddots & \vdots \\ (a_1, b_1, x_n) & \dots & (a_l, b_l, x_n) \end{pmatrix} \tag{10}$$

$$H^* = G^*(W^*, X^*) = \begin{pmatrix} (a_1^*, b_1^*, x_1^*) & \dots & (a_l^*, b_l^*, x_1^*) \\ \vdots & \ddots & \vdots \\ (a_1^*, b_1^*, x_n^*) & \dots & (a_l^*, b_l^*, x_n^*) \end{pmatrix} \tag{11}$$

Compute the output weight β, solving the dual expression of optimization according to the Equation (5) and constraints (6).

Compute the decision function f(x), the predictive function.

$$f(x) = h(x).\beta = \sum_{i=1}^{N} \alpha_i y_i h(x_i, x) \tag{12}$$

Algorithm 1: Symmetric S-ELM-LUPI algorithm

Input: Training set X, Privileged Information set X*, hidden nodes number L and activation functions g and g*, gamma γ, kappa κ
Output: The prediction of the approximated function f(x).

1. start
2. Random generation of the input weights W // For the standard training set
3. Random generation of the input weights W*. // privileged information set
4. Compute the hidden nodes output matrices H ← g(W, X) H* ← g*(W*, X*)
5. Compute α, μ variable to solve the Equation (5)
6. Compute the output weight β according to Equation (4)
7. End

4. Data Preparation and Feature Extraction

The data set created by Martin et al. [8] is used as the experimental data in this paper. This is a data set that can be downloaded for free. It contains two modes of data records: Speech and facial expressions.

There were 42 participants from 14 different nationalities. Each participant received five sentences and expressed pronunciation for six different emotions (namely happiness, fear, disgust, surprise, sadness and anger). 81% of participants were male and the others were female. All of them spoke English. The samples were selected randomly. This data set belongs to the type that induces emotional expression.

There were 1166 video sequences, the number of female videos was 264, and the others were male videos. This database has the advantage of being close to reality and acquiring easily and free. Before the experiment, the data set went through a series of pre-processing procedures.

4.1. Date Processing

The ENTERFACE ′05 audiovisual data set is a data set invented by Martin et al. The original file is in the format of zip. The unzipped folder contains 44 folders which correspond to 44 topics for recording. Each theme's folder contains six folders that correspond to six types of emotions, including anger, disgust, fear, happiness, sadness, and surprise.

In this study, each file contains audio speech and facial expressions. The main research modes include sound mode and visual mode. Before the feature extraction, the two modes have been extracted and separated. Then two new entities are obtained: One contains the visual image, and the other contains the sound signal.

4.2. Feature Extraction

Audio, vision and text are the three main types of information in the data set. During the experiment, only audiovisual-based information can be utilized. Due to the size and clarity limitations of the recorded information data, there were only 1285 different records which adapted the experimental requirements.

Visual and audio can be separated using the toolbox in MATLAB. The visual data is stored in .jpg format and the sound data is stored in .wav format.

Feature extraction is performed separately. There are three types of facial expression related features, including local binary mode (LBP), edge direction histogram (EOH) and local direction mode (LDN). The Figure 2 illustrates the separation and feature extraction operations.

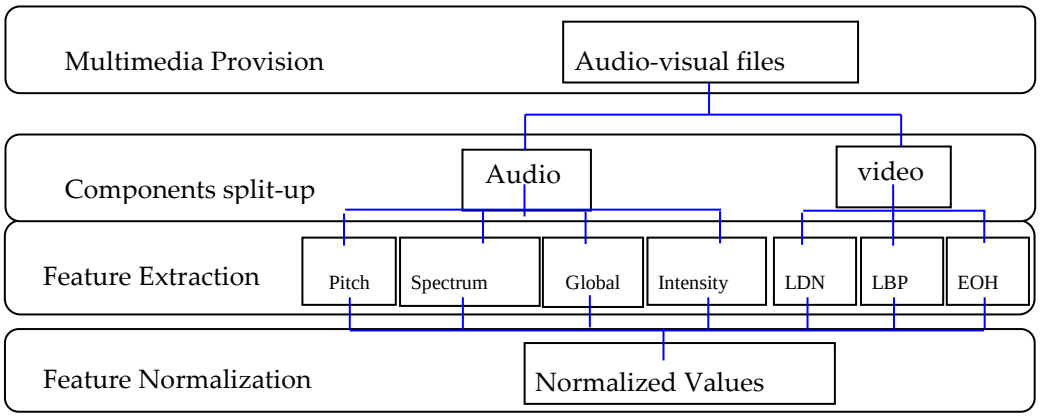


Figure 2. Files split-up and feature extraction operation.

On Figure 2, it depicts the data separation and feature extraction operations. The audio and video files represent a record of the original data set. They contain the speaker's voice and facial images. The first operation is to separate the audio component from the visual component, which will create the '.wav' audio files and the '.jpg' image files. The second operation is feature extracting of the audio and video. Finally, we can obtain the standardized features through standardizing the data. During the feature extraction process, four types of features are extracted from the audio elements including pitch-related features, global features related to sound, spectral-related features and intensity-related features. Each frame of the facial expression features must be extracting features. And then obtained feature values are averaged to construct corresponding feature vectors. Therefore, the corresponding feature vector v_j is calculated as shown in Equation (5) for each frame in a record:

$$v_j = \frac{1}{N_j} \sum_{i=1}^{N_j} a_i \tag{13}$$

Nj represents the number of frames contained in the j record a_i represents the feature value in the i frame. Therefore, the obtained values have different sizes. The values are normalized. The processed values are in the range of 0 to 1. Features number by category. the link to the standardized value component means that all extracted functions must undergo the standardization phase. The normalization is separately done file by file. The final number of features obtained in each category is represented in the Table 1.

Table 1. List of the final number of features obtained for each category.

Categories	Number
Audio	53
Facial -EOH	1764
Facial -LBP	2891
Facial -LDN	3136

5. Experimental Design

The standard information and privilege information are symmetric. The standard information is used as a modal feature set. Another modal feature set is used as privilege information, and the audio feature is treated as an entity. Multiple facial expression related features are regarded as one sample space. As the result, we got a fusion combination which is shown in Figure 3:

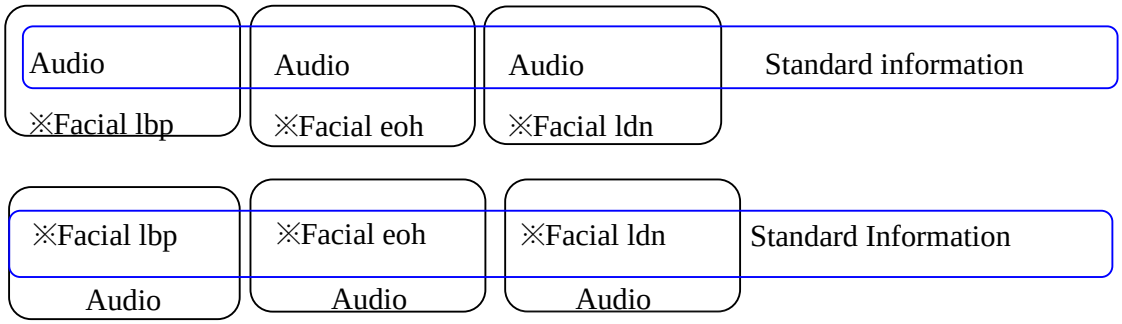


Figure 3. Representation of the combinations.

In the first group, audio information is represented at the top as a standard information set, and facial expressions are indicated at the bottom as privileged information. In the second group, the opposite is true.

This experiment has three goals:

The first goal is to test the effects of multi-modal emotion recognition in real life. Each mode can be treated as privileged information for another mode. The facial expression feature data is divided into three groups. Audio data is no longer feature separated as a data set because of the small number of features.

The second goal is to contrast the mentioned method with the others and assess applicability. The method is contrasted with machine learning based on neural network learning methods, such as extreme learning machine (ELM), sparse extreme learning machine (S-ELM), extreme learning machine-learning using privilege information (ELM -LUPI) and support vector machine (SVM). For contrasting conveniently, the selected parameters must have high similarity and comparability, such as the number of hidden nodes and other parameters.

The third goal is to improve recognition accuracy and execution efficiency. When the recognition time is as small as possible, the proposed algorithm has more chance to serve the actual application.

6. Experimental Design Analysis of Results Analysis of Results

According to the three objectives of the experiment, the analysis of the experimental results is divided into the following three parts.

6.1. Date Processing

The purpose of the first type of experiment was to assess the application of the mentioned method for multi-modal emotion recognition. Audio and facial expressions are used as standard information sets and privileged information sets respectively. In Table 2, the corresponding results are shown.

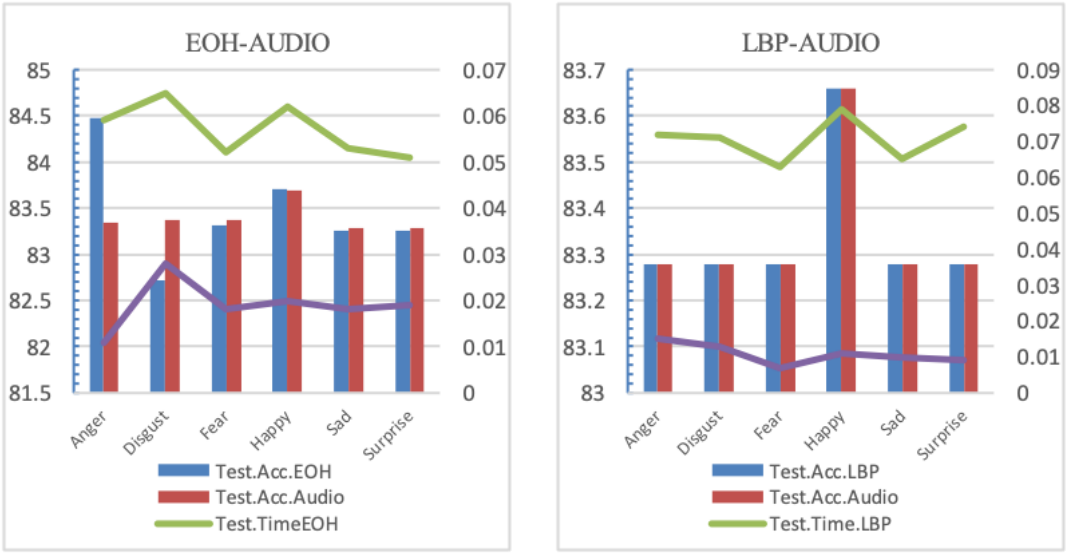
Table 2. Multimodal emotion recognition results.

Datasets		Train Time (sec.)	Test Time (sec.)	Train Accu. (%)	Test Accu. (%)
Standards	**Privileged Information**				
Audio	EOH	0.3392	0.0033	84.18	86.45
EOH	Audio	0.3405	0.0423	84.10	86.45
Audio	LBP	0.0182	0.0000	84.10	86.19
LBP	Audio	0.4512	0.0586	84.13	86.25
Audio	LDN	0.4772	0.0020	83.87	85.34
LDN	Audio	0.4896	0.0566	83.87	85.34

The results shown in Table 2 are experimental data of the mentioned method based on multimodal emotion recognition. The 'Dataset' column corresponds to a dataset that is utilized as standard or privileged information. The main datasets are audio-based data sets and visually relevant data sets (EOH, LBP and LDN). The best performance demonstrated by the experimental results corresponds to the average of the minimum execution time, and the average of the maximum recognition rates.

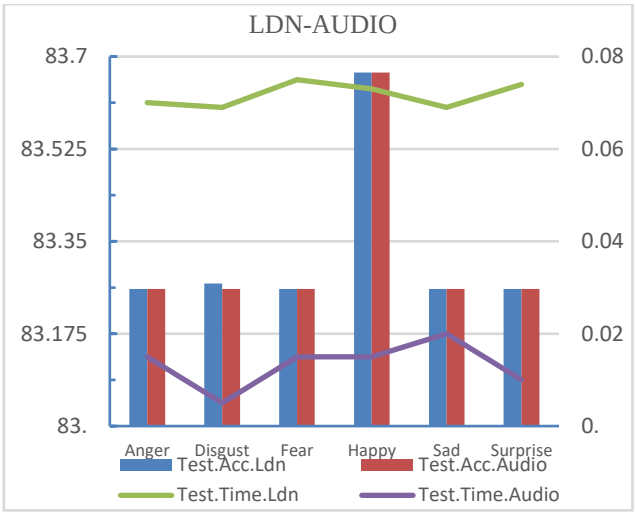
There are six different emotional states (anger, disgust, fear, happiness, sadness, and surprise). The values in the table correspond to values with a hidden node number is 1000.

The recognition accuracy and the mean of the test time of the sentiment type are shown in Figure 4.



(**a**) (**b**)

Figure 4. *Cont.*



(**c**)

Figure 4. (**a**) Recognition accuracy and testing time of EOH-AUDIO; (**b**) Recognition accuracy and testing time of LBP-AUDIO; (**c**) Recognition accuracy and testing time of LDN-AUDIOTest. Acc.EOH (LBP, LDN or Audio) indicates the accuracy of EOH (LBP, LDN or Audio) as a standard information set. Test. Time EOH (LBP, LDN or Audio) represents the test time when EOH (LBP, LDN or Audio) is used as the standard information set.

It was observed that the method is suitable for solving the problem of multi-modal emotion recognition automatic learning, and the accuracy of recognition rate is always above 80%. Because of privileged information learning (LUPI), the proposed method always improves the recognition rate in unknown sample predictions. Figure 4 depicts this improvement.

In Figure 5, The recognition rate of the training set is expressed by the accuracy rate. In Figure 4, the accuracy of the test is better than the accuracy of the training. In essence, the test set provides more information for the training set to better identify new unknown samples.

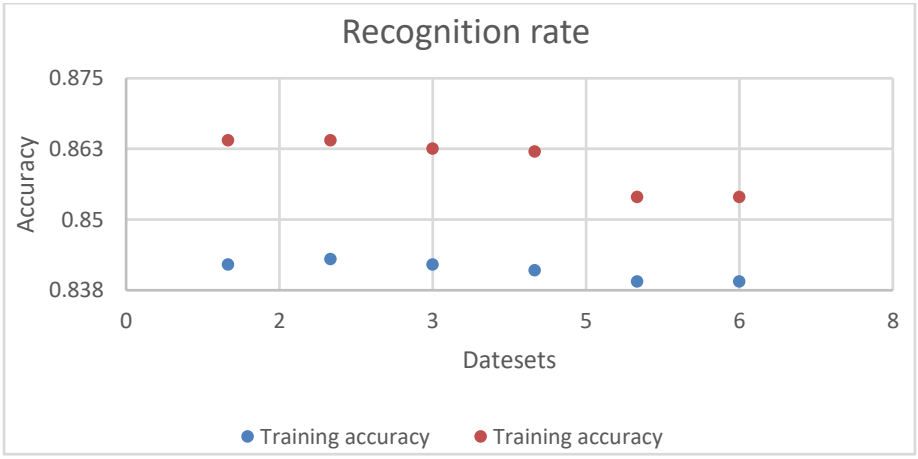


Figure 5. Recognition rate improvement representation.

6.2. Date Processing Analysis of Results Compared with Other Methods

The purpose of this set was to assess the ability of the proposed method by other methods. The results are shown in Tables 3–6.

This four tables represent the best performance of the ELM PLUS, S-ELM, ELM and SVM methods in recognition accuracy and length of execution. The first column contains two sets of data sets for the

standard set and the privilege information set. The 'audio' set represents the audio feature set, and the 'EOH / LBP / LDN' set represents the visual features.

The data show that the ability of the proposed method is better than the stability and recognition rate of other methods.

Table 3. Comparison with other methods in 'Training Accuracy'.

Datasets		SYMMETRIC S-ELM-LUPI	ELM PLUS	S-ELM	ELM BASIC	SVM
Feature	**Privileged Information**	**Train. Accu. (%)**	**Train. Accu. (%)**	**Train. Accu. (%)**	**Train. Accu. (%)**	**Train. Accu. (%)**
EOH	Audio	84.15	100+	91.11	74.50	95.27
Audio	EOH	84.13	-	91.10	75.15	95.24
LBP	Audio	84.12	100+	89.28	68.27	93.16
Audio	LBP	84.16	-	89.53	68.69	93.13
LDN	Audio	83.80	100+	89.64	69.22	96.03
Audio	LDN	83.81	-	89.62	69.29	96.09

Table 4. Comparison with other methods in 'Testing Accuracy'.

Datasets		SYMMETRIC S-ELM-LUPI	ELM PLUS	S-ELM	ELM BASIC	SVM
Feature	**Privileged Information**	**Test. Accu. (%)**	**Test. Accu. (%)**	**Test. Accu. (%)**	**Test. Accu. (%)**	**Test. Accu. (%)**
EOH	Audio	86.42	58.173+	86.446	60.10*	61.76
Audio	EOH	86.43	-			61.67
LBP	Audio	86.25	54.599+	86.865	64.68*	66.64
Audio	LBP	86.18	-			66.49
LDN	Audio	85.35	50.240+	85.356	55.35*	62.79
Audio	LDN	85.31				62.43

+,* Special record at L=104.

Table 5. Comparation with other methods in 'Training Time'.

Datasets		SYMMETRIC S-ELM-LUPI	ELM PLUS	S-ELM	ELM BASIC	SVM
Feature	**Privileged Information**	**Train. Time (sec)**	**Train. Time (sec)**	**Train. Time (sec)**	**Train. Time (sec)**	**Train. Time (sec)**
EOH	Audio	0.3411	4.5165	0.9023	0.6045	103.8934
Audio	EOH	0.3389	3.6657	0.9132	0.5665	108.3962
LBP	Audio	0.4534	4.9947	1.1987	0.6732	43.7973
Audio	LBP	0.0179	5.2152	1.2154	0.6743	52.3634
LDN	Audio	0.4883	4.8047	1.2494	0.7198	61.6862
Audio	LDN	0.4787	4.7758	1.2342	0.7356	59.4054

Table 6. Comparison with other method in 'Recognition Time'.

Datasets		SYMMETRIC S-ELM-LUPI	ELM PLUS	S-ELM	ELM BASIC	SVM
Feature	**Privileged Information**	**Test. Time (sec)**	**Test. Time (sec)**	**Test. Time (sec)**	**Test. Time (sec)**	**Test. Time (sec)**
EOH	Audio	0.0445	0.084555	0.0345	0.1021	2.9384
Audio	EOH	0.0076	0.0011	0.0432	0.0843	2.8145
LBP	Audio	0.0582	0.1276	0.0696	0.0656	5.8265
Audio	LBP	0.0000	0.0145	0.0787	0.0634	5.0938
LDN	Audio	0.0556	0.1032	0.0834	0.1423	6.5984
Audio	LDN	0.0068	0.0000	0.0687	0.1467	7.7163

Figure 6 shows a comparison of the identification accuracy changes for different methods. The precision is expressed in the number of hidden nodes. The mentioned method has better generalization and stability than other methods.

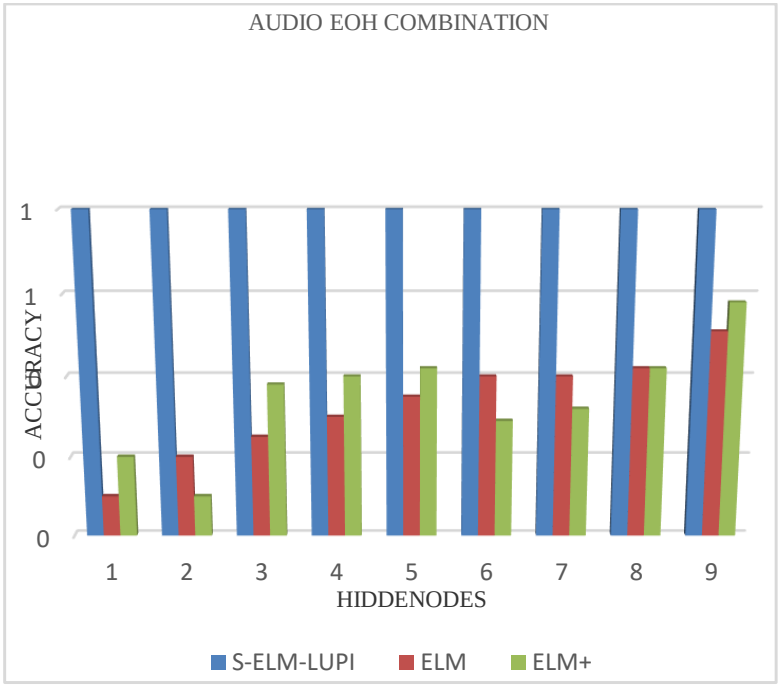


Figure 6. The stability of the training accuracy on EOH-Audio.

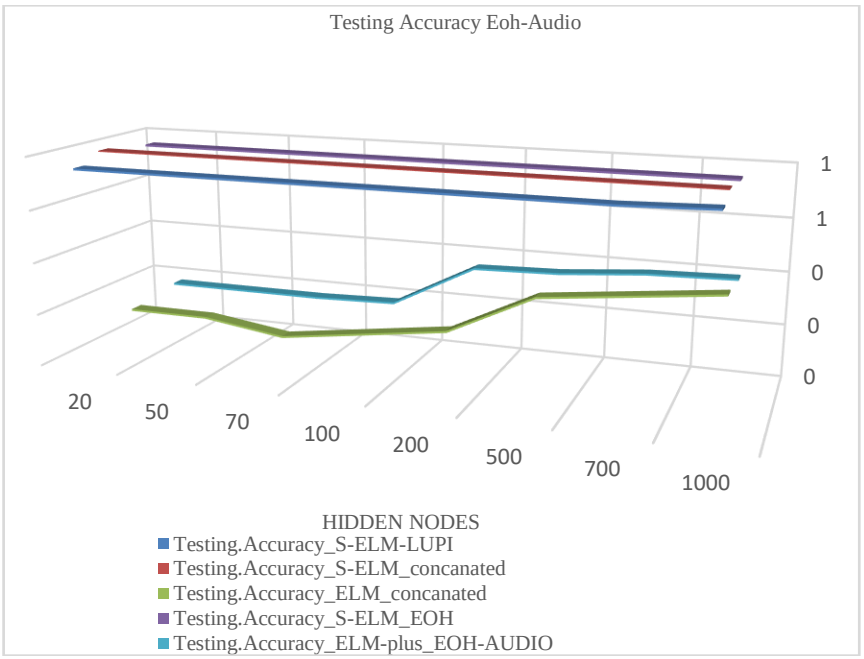
6.3. Analysis of Improved Result Analysis

The purpose of the third type of experiment is to evaluate improved content related to the performance of the proposed method. SYMMETRIC S-ELM-LUPI inherits the main advantages of the original method in the process of introducing the method. Firstlu, the method inherits the advantage of the Extreme Learning Machine method in fast calculations. Secondly, it gains the advantage of saving memory from the sparse limit learning machine. Finally, it achieved an increase in the recognition rate because of using the LUPI.

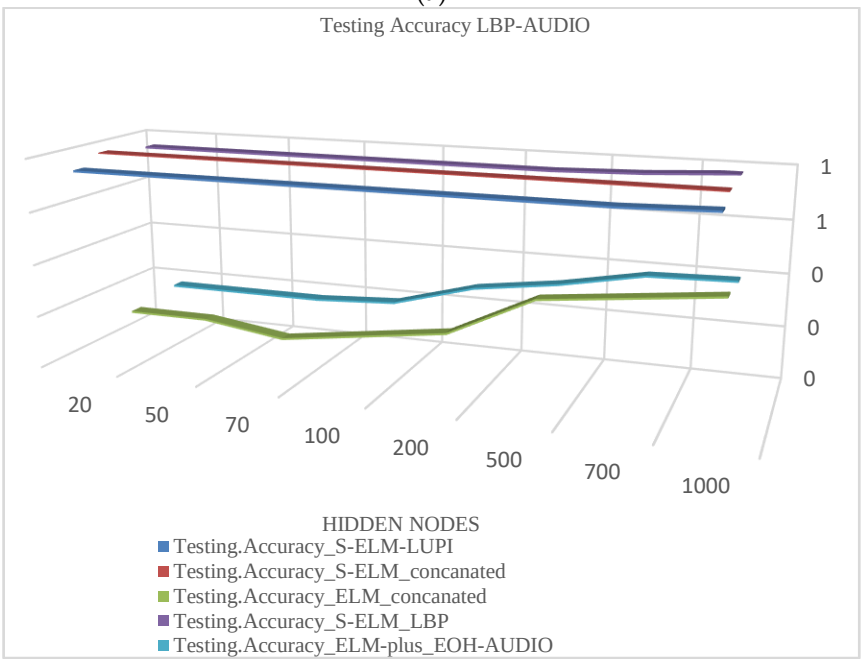
In the improvement analysis of recognition accuracy, the mentioned method is contrasted with the results of single-mode experiments.

Figure 6 illustrates the comparison of the proposed method recognition rate with the corresponding single mode recognition rate.

Figure 7 shows the result of the recognition accuracy change follow the number of hidden nodes. The EOH feature of the audio feature is added in Figure 7a; the LBP feature of the audio feature is added in Figure 7b; the LDN feature of the audio feature is added in Figure 7c. Visually relevant features (EOH, LBP and LDN) were used as standard information. The x-axis represents the hidden node and the y-axis represents the accuracy of the recognition

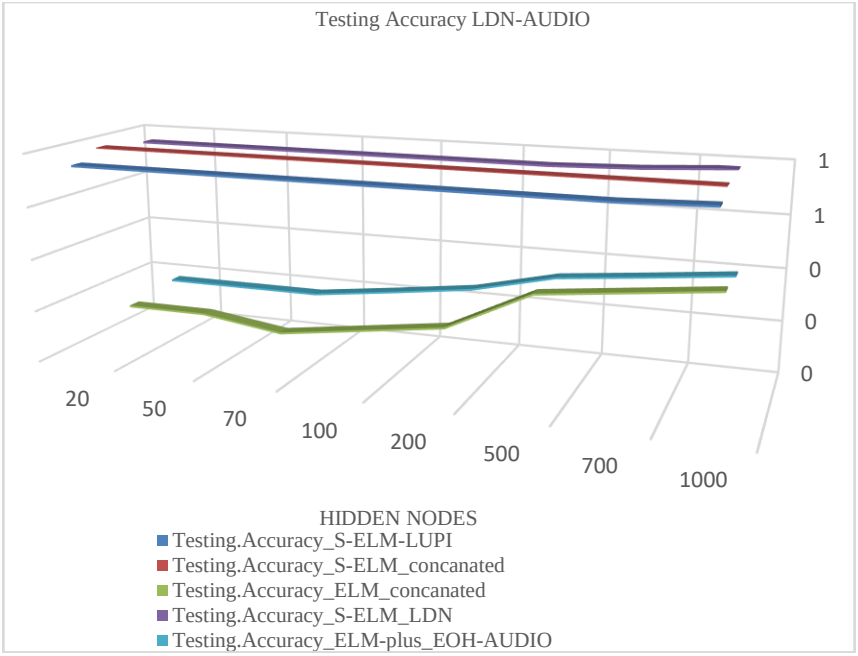


(**a**)



(**b**)

Figure 7. *Cont*.

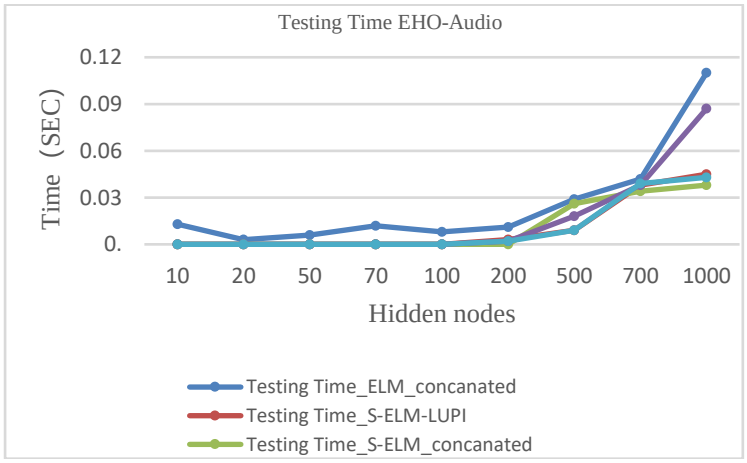


(**c**)

Figure 7. (**a**) Recognition improvement for adding EOH feature; (**b**) Recognition improvement for adding LBF feature; (**c**) Recognition improvement for adding LDN feature.

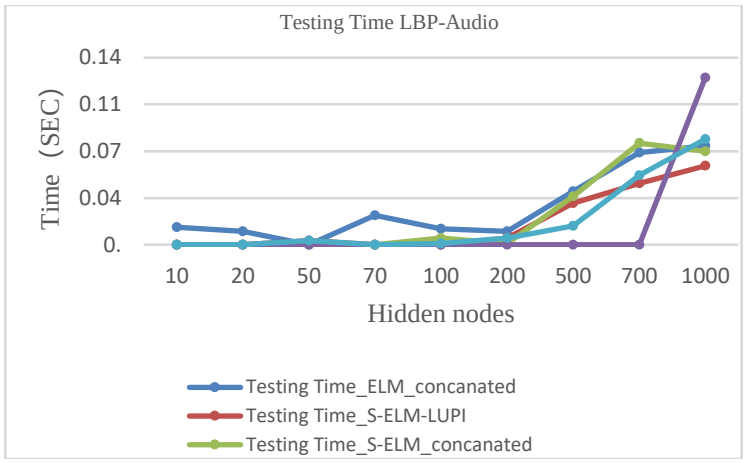
When comparing the accuracy of the SYMMETRIC S-ELM-LUPI method with the other four methods, the facial visual information is used as a standard information set, and the voice is used as a privileged information set.

As a result, it can be found that the proposed method is superior to the single-modal method in the recognition of the multi-mode emotion recognition problem, and the execution time is also different. Figure 8 shows the corresponding results.

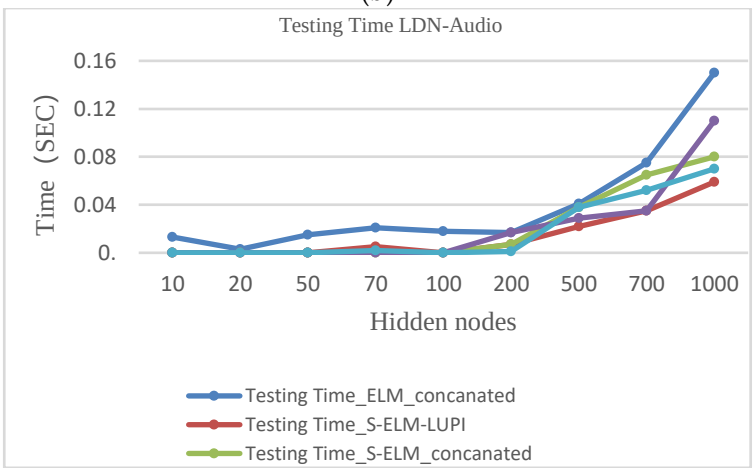


(**a**)

Figure 8. *Cont*.



(**b**)



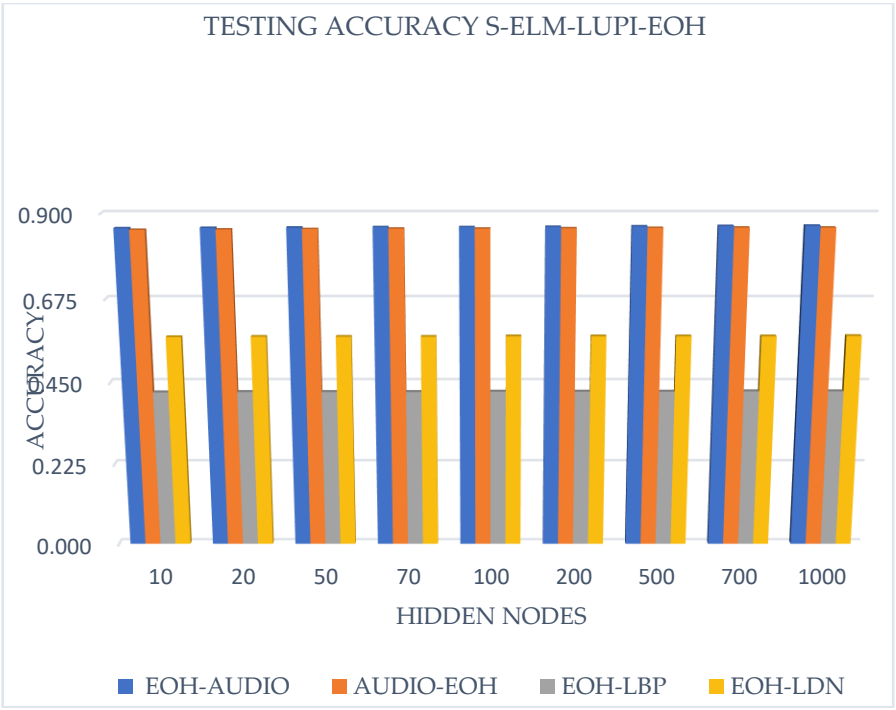
(**c**)

Figure 8. (**a**) The comparison of testing time EHO-Audio of different methods; (**b**) The comparison of testing time LBP-Audio execution time of different methods; (**c**) The comparison of testing time LDN-Audio execution time of different methods.

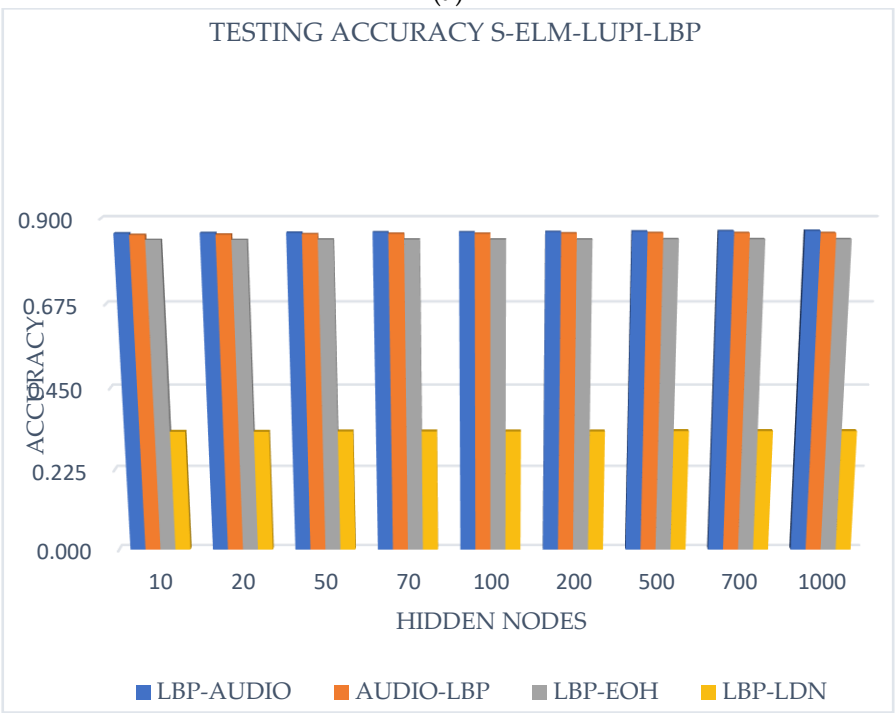
It is seen very clearly that this method is faster than other methods. The number of test set dimension is reduced, and the core computing features and rapid generalization capabilities of the ELM are utilized, so that the mentioned method can solve real-time problems.

The results shown in Figure 8 illustrate that the execution time depends on the size of the test set. The traditional method utilizes the same dimension in the training and test set; the method can effectively reduce the training time because the privilege information dimension is small.

Concerning the study on the modality's individual contribution, the recognition is compared based on the coupling point of view. The corresponding results are represented on Figure 9.

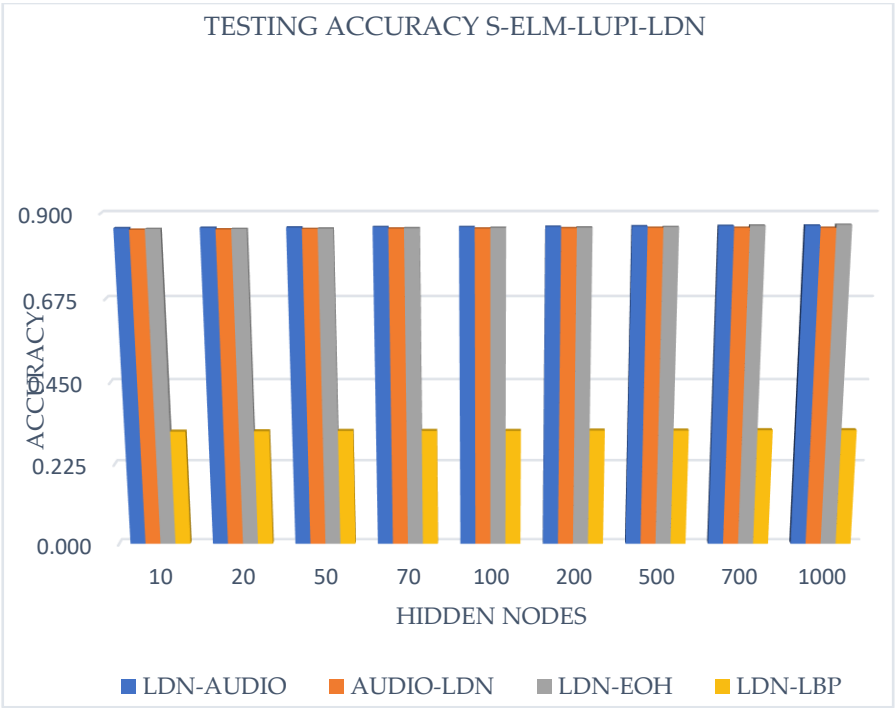


(**a**)

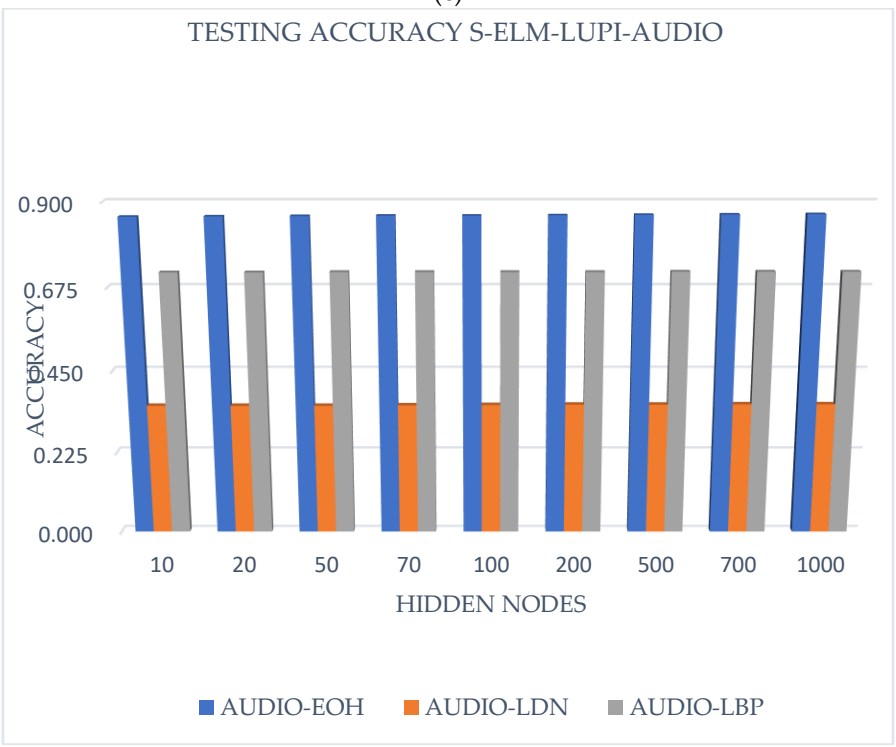


(**b**)

Figure 9. *Cont.*



(**c**)



(**d**)

Figure 9. (**a**) Modality contribution comparison of EOH; (**b**) Modality contribution comparison of LBP; (**c**) Modality contribution comparison of LDN; (**d**) Modality contribution comparison of Audio.

Figure 9, represents the comparison of the different modality's contribution to each other, using the proposed method. The title of each graph signifies the considered standard information source set, i.e., EOH in 9a, LBP in 9b, LDN in 9c, and Audio in 9d. The 'Testing Accuracy EOH(LBP/LDN/AUDIO)-AUDIO (EOH/LBP/LDN)', means that the first set is the standard information source set and the second is the 'Additional (Privileged) Information' source set. In observed results, the use of different

modality (audio and visual) gives better results than the unimodality using multiple features type. In fact, the use of multiple features helps to collect more information on the same data sets from different point of views, but the multiple modalities combination procures more useful information; to make a distinction between emotional states. Therefore, the use of features from different modalities is better applied with this proposed method than the use of multiple features from one same modality.

7. Conclusions

A new method of emotion recognition based on multi-modality is to use the sparse limit learning machine and using the symmetric privileged information learning in this paper. Symmetric S-ELM-LUPI Paradigm has passed the tests performed on the data set. The symmetry used in this paper refers to the symmetry of the method. This method regards one pattern as the standard information source, while the other pattern as the privileged information source, each mode can be treated as privileged information for another mode, rather than the symmetry of the simple data level, such as the symmetry of the data level. For example, the symmetry of data on both sides of the face. This method has been proven to be applicable to data sets based on multimodal emotions, and the correct recognition rate is over 80% at very fast execution speeds.

The experimental results prove that the method is very reasonable in multi-modal emotion recognition because of the method's stability. In fact, multimodal emotion recognition is a real-life problem that requires a very stable method in predicting. The method of this paper provides a new thought for the accurate identification of emotions in real life.

Author Contributions: Conceptualization, X.B.; methodology, L.Y.; software, L.Y.; validation, M.M. and Z.C.; formal analysis, L.Y.; investigation, L.Y.; resources, X.B.; data curation, L.Y.; Writing—Original Draft preparation, L.Y.; Writing—Review and Editing, L.Y.; visualization, L.Y.; supervision, X.B.; project administration, X.B.; funding acquisition, X.B.

Funding: This research was funded by The National Key Research and Development Program of China (Grant No. 2016YFB1001404) & National Natural Science Foundation of China (61873299, 61702036, 61572075).

Acknowledgments: This work was supported by Han Zhishuai, Gao Yuxing and other students in the Laboratory of Artificial Fish and Intelligent Software, Beijing University of Science and Technology.

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

References

1. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In Proceedings of the IEEE International Conference on Data Mining, Barcelona, Spain, 12–15 Decenber 2016.
2. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
3. Latha, G.C.P.; Priya, M.M. A Novel and Enhanced Facial Electromyogram Based Human Emotion Recognition Using Convolution Neural Network Model with Multidata Signal Processing Features. *J. Comput. Theor. Nanosci.* **2017**, *14*, 1572–1580. [CrossRef]
4. Huang, Y.; Yang, J.; Liao, P. Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. *Comput. Intell. Neurosci.* **2017**, *2017*, 1–8. [CrossRef] [PubMed]
5. Torres-Valencia, C.; Álvarez-López, M.; Orozco-Gutiérrez, Á. SVM-based feature selection methods for emotion recognition from multimodal data. *J. Multimodal User Interfaces* **2017**, *11*, 1–15. [CrossRef]
6. Chan, W.L.; Song, K.Y.; Jeong, J.; Choi, W.Y. Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data. *arXiv*, 2018; arXiv:1805.06606.
7. Feyereisl, J.; Aickelin, U. Privileged information for data clustering. *Inf. Sci.* **2012**, *194*, 4–23. [CrossRef]
8. Ji, Y.; Sun, S.; Lu, Y. Multitask multiclass privileged information support vector machines. *IEEE Int. Conf. Pattern Recognit.* **2013**, *2012*, 2323–2326.

9. Liu, J.; Zhu, W.-X.; Zhong, P. A New Multi-class Support Vector Algorithm Based on Privileged Information. *J. Inf. Comput. Sci.* **2013**, *10*, 443–450.
10. Wang, S.; Zhu, Y.; Yue, L.; Ji, Q. Emotion Recognition with the Help of Privileged Information. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 189–200. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).