Label-less Learning for Emotion Cognition

Min Chen, *Senior Member, IEEE*, and Yixue Hao

***Abstract*— In this paper, we propose a label-less learning for emotion cognition (LLEC) to achieve the utilization of a large amount of unlabeled data. We first inspect the unlabeled data from two perspectives, i.e., the feature layer and the decision layer. By utilizing the similarity model and the entropy model, this paper presents a hybrid label-less learning that can automatically label data without human intervention. Then, we design an enhanced hybrid label-less learning to purify the automatic labeled data. To further improve the accuracy of emotion detection model and increase the utilization of unlabeled data, we apply enhanced hybrid label-less learning for multimodal unlabeled emotion data. Finally, we build a real-world test bed to evaluate the LLEC algorithm. The experimental results show that the LLEC algorithm can improve the accuracy of emotion detection significantly.**

***Index Terms*— Deep learning, emotion detection, label-less learning, multimodal emotion cognition.**

I. Introduction

With the development of the smartphones, Internet of Things (IoT), and cloud computing, more and more people have been spending a lot of time on interacting with the machines. Human–machine interaction has become a nonnegligible part of our life. To realize a friendlier and more natural human–machine interaction, the machine should be able to understand user's emotion. Therefore, emotion detection plays a vital role in the human–machine interaction [1]. In view of emotion recognition, an important issue is the acquisition of emotional data. With the advances on networking technology, the emotional data acquisition via an online social network or a content provider, such as text data for Twitter sentiment analysis and YouTube-based video emotional data, has become more and more convenient. In general, the emotion data include the facial expression, speech, text, physiological signals, and user's behavioral indicators [2], [3].

The speech and facial expression are the two most accessible data modalities [4]. By the use of speech emotional data, previous work typically recognizes emotion via extracting the prosodic features, acoustic features, and voice quality features. On the other hand, in the view of facial expression data, emotion detection is realized by extracting the appearance features and geometrical features of a face [5]. In addition, some works attempt to recognize the emotion by multimodal data [6].

However, most of the above works are based on the manual feature extraction and cannot reflect the complex nonlinear relevance among facial expression, speech, and human's emotional intentions accurately [7]. With the development of artificial intelligence (AI) [8], [9], the deep learning models have achieved an outstanding performance in the feature extraction in the aspects of image and speech [10], [11]. Recent works mainly attempt to recognize the emotion using the deep learning model.

The existing emotion detection deep models are mostly based on a trained deep network model (e.g., Alex network [12], and so on), followed by a fine tuning of the decision based on a provisioning of the intermittent labeled emotion data set. The growth of emotion cognition intelligence would reach a plateau without such provisioning of labeled data. Thus, the challenge of applying a deep learning model into emotion detection is the lack of a large-scale labeled data set [13]. In fact, although a large amount of emotion data can be acquired through IoT technology whenever and wherever possible, only a small part of emotion data is labeled. Therefore, labeling the collected massive emotion data set is a challenging issue.

Typically, there are two categories of emotion data set labeling, i.e., manual labeling and automatic labeling. Manual labeling is the most intuitive method. With the development of crowdsourcing, volunteers can be recruited to label the unlabeled data through a well-designed gaming interface. However, the disadvantage of this scheme is opportunistic and uncontrollable though some cost is saved by participatory sensing. On the other hand, automatic labeling can be divided into two types: the automatic labeling based on either different domains or the same domain. The automatic labeling based on different domains is mainly the automatic labeling based on transfer learning [14].

In this paper, we pay attention to the automatic labeling based on the same domain. The automatic labeling based on the same domain includes active learning [15] and semisupervised learning [16]. The active learning first screens the massive unlabeled data by predicting the data uncertainty and then performs the manual labeling. This method can reduce the intensity of human labor but manual intervention is still needed. The semisupervised learning (such as self-training [17]) can conduct the automatic labeling of data. However, the intrinsical feature of error accumulation hinders its application scope. Therefore, the above methods cannot be directly adopted for the purpose of the label-less emotion cognition.

Manuscript received July 27, 2018; revised January 21, 2019 and May 28, 2019; accepted July 11, 2019. Date of publication August 13, 2019; date of current version July 7, 2020. This work was supported by the National Key Research and Development Program of China under Grant 2017YFE0123600. The work of Dr. Y. Hao was supported in part by the National Natural Science Foundation of China under Grant 61 802 138 and in part by the China Postdoctoral Science Foundation under Grant 2018M632859 and Grant 2019T120657. *(Corresponding author: Yixue Hao.)*

The authors are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: minchen2012@hust.edu.cn; yixuehao @ieee.org).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2929071

To solve the aforementioned challenges, this paper proposes an emotion detection algorithm based on automatic labeling by using the multimodality of emotion data, named the label-less learning for emotion cognition (LLEC). The main ideas are to add the automatically labeled data with high confidence to the train set by selecting the massive unlabeled emotion data and further enhance the accuracy of emotion detection. Specifically, we first consider the unlabeled data from two perspectives, i.e., the feature layer and the decision layer. By utilizing the similarity and entropy models, a hybrid automatic label strategy is proposed to handle unlabeled data. To further enhance data confidence, we screen the automatic label data again. Finally, we build a test bed and verify the validity of the proposed LLEC algorithm.

In summary, the main contributions of this paper are as follows.

1) For a large number of unlabeled data, we propose a new LLEC algorithm to explore the value of unlabeled data from the decision layer and the feature layer.
2) A hybrid automatic labeling strategy is presented to purify the automatically labeled data to increase the correctness of automatic labels, and thus preventing error propagation.
3) We verify the proposed algorithm by a real-world test bed of emotion detection and interaction. The experimental results indicate that the proposed algorithm outperforms other algorithms in terms of accuracy of emotion detection.

The remainder of this paper is organized as follows. The LLEC is introduced in Section II. The design issues of LLEC are given in Section III. Section IV presents the experimental setup and results for the LLEC. Finally, Section V concludes this paper.

II. Label-Less Learning for Emotion Cognition

In this section, we first introduce emotion detection and emotion cognition. Then, the learning methods relevant to label-less learning is presented. Finally, the proposed LLEC is explained in detail.

A. From Emotion Detection to Emotion Cognition

In this paper, we use the multimodal data to recognize emotion in terms of speech data and facial expressions. The traditional method of emotion detection is to train the labeled data by machine learning and finally derive user's emotion label, as shown in the left-hand side of Fig. 1. In this figure, taking AlexNet deep convolution neural network (DCNN) as an example, we present emotion detection based on deep learning. The emotion detection based on the DCNN architecture has the ability to reach the higher emotion detection accuracy. Since the architecture not only automatically extracts the features of the speech data and facial data by means of deep learning but also combines the two modal data sets [12], [18].

However, emotion detection based on deep learning requires a large number of labeled data. To enable the deep learning to exhibit more effectiveness in emotion detection, it is mandatory to reduce its dependence on the large-scale labeled data.

TABLE I
Difference Between Emotion Detection and Emotion Cognitive

Concept	Description
Emotion detection	Emotion detection is a one-time classification problem based on the labeled data. It uses one modal or multimodal emotional data as input, and obtain the emotional classification through traditional machine learning or deep learning algorithm.
Emotion cognition	Emotion cognition is a self-learning algorithm for emotion using labeled and unlabeled data. It takes one model or multimodal of labeled emotional data as initial input, uses deep learning and other algorithms to continually learn the unlabeled emotional data, and achieves higher accuracy of emotional classification.

In other words, it is required to maximize the utilization of unlabeled data without human intervention. Thus, we propose the LLEC. Furthermore, as shown in Table I, we introduce the differences between emotion detection and emotion cognition. From the table, we can obtain that the emotion cognition includes emotion detection.

B. Evolution of Label-Less Learning

Typically, the emotional big data have the following two main characteristics.

1) It includes a large amount of unlabeled data. To address this issue, existing works mainly adopt active learning, positive and unlabeled learning (i.e., PU learning), and self-training.
2) It exhibits the feature of multimodality. In order to match such feature, multimodal data learning is introduced, such as the cotraining [19].

Based on the above features, we give the specific introduction of the following four algorithms (i.e., active learning, PU learning, self-training, and cotraining) using unlabeled data.

Active learning is able to interactively query a user to obtain the desired output for new data [15]. The learning process is shown as follows. First, the labeled data set is denoted by L, and let U represent the unlabeled data set. Then, a subset C of a data set U is founded by selecting the unlabeled data with the most abundant information using the information on data set L, making the labeling request to the experts. Finally, after the experts label a data set C, it is added to data set L. Such iteration can be continued with the labor-intensive work by the experts.

According to the above analysis, the active learning typically needs manual intervention. However, a user's emotion status is very sensitive to human intervention, which is a must for the traditional labeling process. Human intervention may also cause the emotion data contaminated. Thus, label-less learning without human intervention is critical to an emotion detection system in terms of sustainability. For this reason, the active learning is not applicable to the emotion cognition.

PU learning refers to the training data set consisting of a small amount of labeled positive data and a large amount of unlabeled data [20]. Furthermore, PU learning is a machine

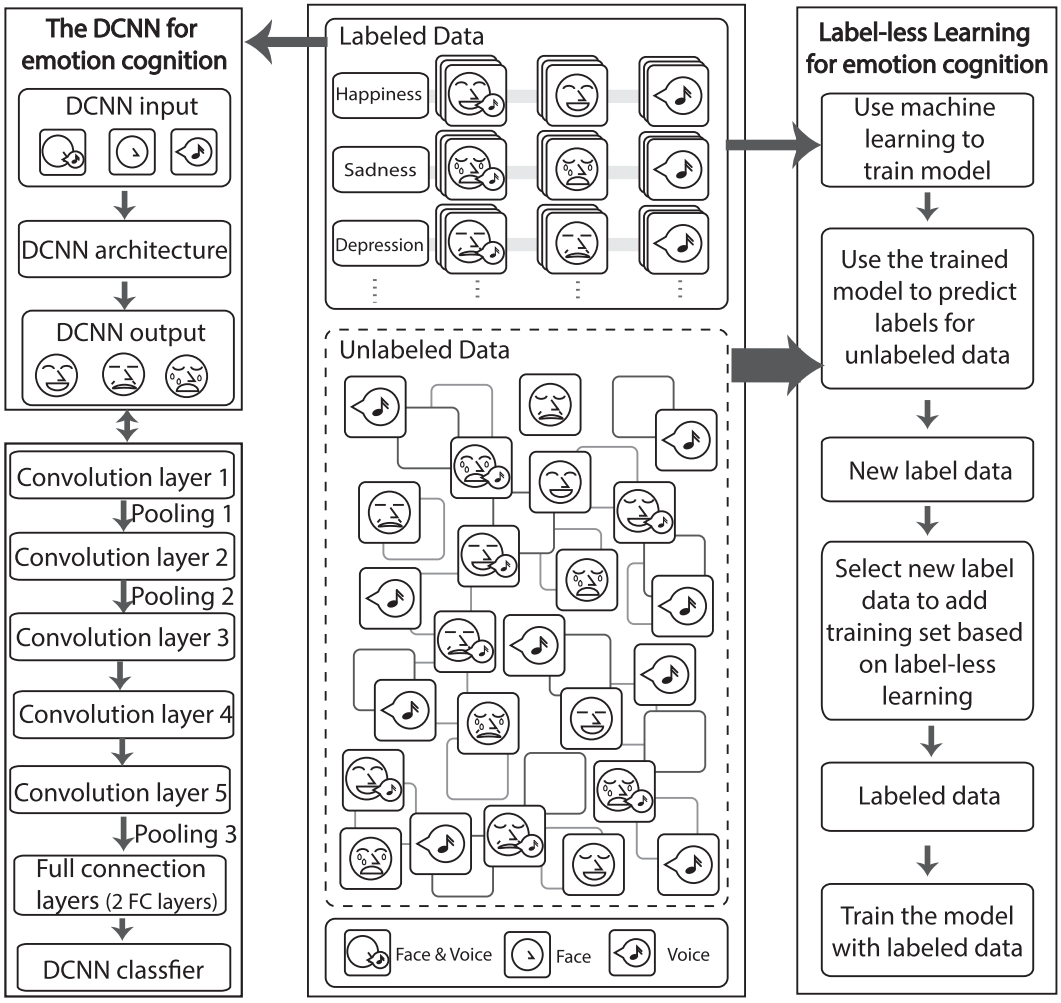


Fig. 1. Emotion detection with DCNN and label-less learning.

learning method that classified the unlabeled data into two categories, i.e., positive and negative. The purpose of PU learning is to utilize a small amount of labeled positive data and a large amount of unlabeled data to construct the classifier and predict whether the unlabeled data are positive or not. However, in the emotion detection, due to the diversity of the emotion data labels, the PU learning algorithm cannot be directly applied to the emotion data set with a large amount of unlabeled data.

Self-training is a common algorithm for inductive semi-supervised learning [16], [17], [21], and it can automatically label the unlabeled data without manual intervention. The main idea of self-training is as follows. First, the existing label data (i.e., initial training set) are used to train and form the classifier, and the unlabeled data are labeled by the classifier. Then, the unlabeled data with high confidence are added to the initial training set, and the data with predicted labels are added to the original training set and then deleted from the unlabeled data set. Finally, the updated data set is utilized to retrain the classifier. This process is repeated until the predefined goals are met. However, the above method generally fails to judge whether the automatic label is correct. After adding the wrong automatically labeled data to the training set, the error accumulation and propagation are caused. Thus, self-training cannot be directly applied to label the emotion data.

Cotraining is a common semisupervised learning from the multi-view perspective [22]. It refers to utilizing the multi-view information (such as multimodal data) of the same object to realize the data learning. For instance, in the case of emotion data, the emotion can be recognized from two modalities, the facial expression and speech data of a user. Specifically, in this paper, we consider the data set only with two modalities. Namely, we first develop two classification models F_1 and F_2 that are in accordance with two labeled data sets (X_1, Y_1) and (X_2, Y_2), respectively, where X_1 represents the facial expression data, X_2 represents the speech data, and Y_1 and Y_2 represent the corresponding labels. Then, these two classification models are utilized to label the unlabeled data. Finally, the high-confidence data predicted by F_1 classifier are added to the X_2 training data set, and the high-confidence data predicted by F_2 classifier are added to the X_1 training data set. In addition, the newly labeled data set is deleted from the unlabeled data set. This process is repeated until the predefined iterations are met. Cotraining can validate each other among multimodal data and can enhance the confidence level of the new labeled data to some extent. However, selection of unlabeled data has the limitation [23].

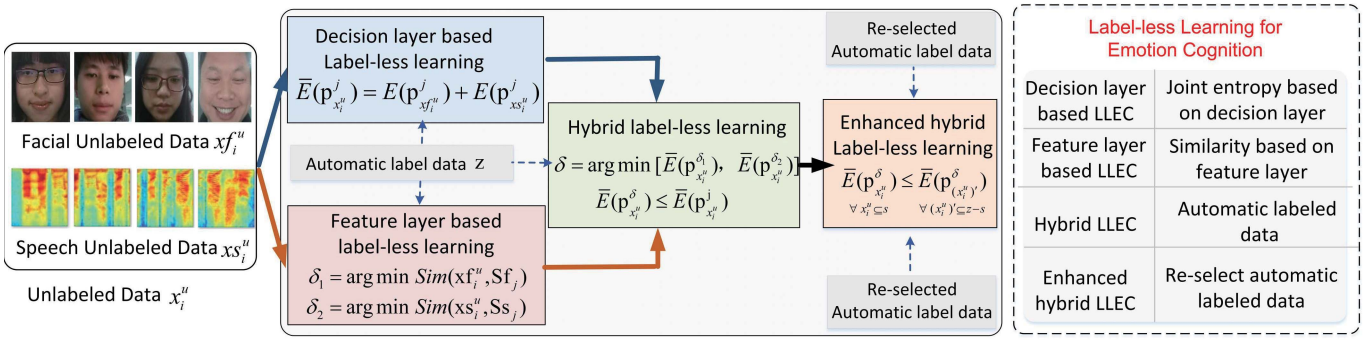


Fig. 2. Illustration of LLEC.

C. Introduction to Label-Less Learning for Emotion Cognition

To use the unlabeled data better, we propose the LLEC. The LLEC denotes a learning method for recognizing a large amount of unlabeled multimodal emotion data, which is intended to enhance the emotion detection accuracy. Specifically, the label-less learning considers the unlabeled data from two perspectives: the feature layer and the decision layer. By utilizing the similarity model and the entropy model, we propose an automatic labeling strategy for unlabeled data through a hybrid model. To further enhance the confidence of selected data, we propose an enhanced hybrid model to select the automatically labeled data.

The specific LLEC flow shown in the right-hand side of Fig. 1 can be divided into the following five steps.

1) Define and acquire the emotion-related data, such as facial expression and speech.
2) In view of a small amount of labeled multimodal data, use the machine learning to train the model and enhance the emotion detection accuracy as much as possible.
3) For the unlabeled data, based on the similarity model and the entropy model, the hybrid model is used to label the unlabeled data, and the data are then added to the original labeled data set.
4) To further enhance the confidence of the added data, the newly added data are selected.
5) Retraining the model with the new labeled data.

In conclusion, the LLEC is used to label the unlabeled data and add the automatically labeled data to the training model. Thus, the label-less learning is achieved, and the automatically labeled data are obtained so as to reduce the dependence on label data and wasting of human and material resources during labeling. However, in the LLEC, it is necessary to solve two challenging problems: 1) how to label the unlabeled data and add it to the training set and 2) in the case of the multimodal data, how to realize the mutual authentication of a multimodal data set.

III. Methods of Label-Less Learning for Emotion Cognition

In this section, we explain the LLEC method in detail, as shown in Fig. 2. First, we train the machine learning model (e.g., deep learning model) using a small amount of labeled data. Then, the unlabeled data are labeled automatically using the hybrid LLEC model. Finally, we design the enhanced hybrid algorithm to further increase the accuracy of emotion detection and utilize the unlabeled data.

A. Label-Less Learning

In this section, we explain how the unlabeled emotion data are labeled automatically. The labeled data are $\mathbf{x}^l = (x_1^l, x_2^l, \ldots, x_n^l)$, where n is the number of labeled data sets. The unlabeled data are $\mathbf{x}^u = (x_1^u, x_2^u, \ldots, x_m^u)$, where m is the number of unlabeled data sets. We assumed that the corresponding label of the labeled data set is $\mathbf{y} = (y_{x_1^l}, y_{x_2^l}, \ldots, y_{x_n^l})$. In this article, we assume that the amount of unlabeled data is larger than the labeled data, i.e., $m > n$. We aim to label the unlabeled data set automatically and select the new labeled data to be added to the training set to enhance the emotion detection accuracy.

For this purpose, the unlabeled data are first considered from two perspectives: the feature layer and the decision layer, and then the hybrid label-less learning is utilized to label the unlabeled data automatically, and the newly added data are reselected to enhance the accuracy of labeling of the newly added data. The specific label-less learning process is shown in Fig. 2. In this paper, we set our emotional labels to discrete labels, which include anger, disgust, fear, joy, sadness, and surprise. Furthermore, the decision layer-based label-less learning and feature layer-based label-less learning need to make use of existing emotion detection models. For emotion detection model, traditional machine learning algorithms or deep convolutional neural networks can be used to extract features and recognize emotions.

1) Decision Layer-Based Label-Less Learning: In the decision layer-based label-less learning, the strategy of uncertainty prediction is adopted as a criterion for selecting new labeled data, i.e., the unlabeled data with only lower prediction uncertainty are selected. To assess the prediction uncertainty, we adopt the entropy as a measure. By using the labeled data, based on the machine learning model, we can obtain the prediction probability of unlabeled data x_i^u as

$$p_{x_i^u} = \left\{ p_{x_i^u}^1, p_{x_i^u}^2, \ldots, p_{x_i^u}^c \right\} \tag{1}$$

where $p_{x_i^u}^j$ denotes the probability of the unlabeled data x_i^u predicting into the emotional class j and c is the number

of classes. The probability entropy $E(p_{x_i^u})$ predicted by the unlabeled data x_i^u is defined by

$$E\left(p_{x_i^u}\right) = -\sum_{j=1}^{c} p_{x_i^u}^j \log\left(p_{x_i^u}^j\right). \tag{2}$$

In (2), it can be seen that at smaller entropy, the new labeled data have lower prediction uncertainty. Thus, the entropy can be regarded as a criterion of new labeled data selection in the decision layer. When utilizing the entropy and determining whether the unlabeled data are added, it is required to set the threshold value δ_E, and when entropy $E(p_{x_i^u})$ is smaller than a given threshold value δ_E, the unlabeled data x_i^u are added to the new training set, but not vice versa.

2) Feature Layer-Based Label-Less Learning: In the feature layer-based label-less learning, the label-less learning is based on the similarity model. Denote the labeled emotion data set of class j as S_j, $j = 1, 2, \ldots, c$, where c is the number of classes. For instance, S_j can be labeled as all emotion data sets of happy. We define the similarity $\mathrm{Sim}(x_i^u, S_j)$ between the unlabeled data x_i^u and the labeled data set S_j as follows:

$$\mathrm{Sim}\left(x_i^u, S_j\right) = \sum_{x_j^l \in S_j} e^{-\left\|\phi\left(x_i^u\right) - \phi\left(x_j^l\right)\right\|_2} \tag{3}$$

where $\phi(x_i^u)$ is the feature vector of the unlabeled data x_i^u and $\phi(x_j^l)$ is the feature vector of the labeled data x_j^l. $||\cdot||_2$ is two-norm operation, describing the similarity between the unlabeled and the labeled data.

In (3), we can see that the closes x_i^u are to the labeled emotion data set S_j, the smaller the value of $\mathrm{Sim}(x_i^u, S_j)$ is, and vice versa. Therefore, the similarity model can be utilized to describe the feature layer-based label-less learning. The similarity model describes the similarity (or distance) between the unlabeled data x_i^u and labeled data from the perspective of the feature layer. Similar to the entropy-based model, this strategy needs setting the threshold value δ_{Sim}. When $\mathrm{Sim}(x_i^u, S_j)$ is smaller than a given threshold value δ_{Sim}, the unlabeled data are added to the training set, but not vice versa.

3) Hybrid Label-Less Learning: The entropy-based measure and the similarity measure provide the method of automatic labeling of unlabeled data and adding it to the training set in the decision layer and the feature layer, respectively. However, the disadvantage of these two schemes is a need for manual setting of the thresholds δ_E and δ_{Sim}. In the following, based on the entropy-based measure and similarity measure, the hybrid label-less learning is introduced. The specific process of hybrid label-less learning is as follows. First, the most similar class j to the unlabeled data x_i^u is determined in accordance with the similarity measure and denoted as δ. The class δ is defined as

$$\delta = \underset{j}{\operatorname{argmin}} \quad \mathrm{Sim}\left(x_i^u, S_j\right). \tag{4}$$

Then, according to the minimum entropy principle, it is decided whether to add the unlabeled data x_i^u and its label δ to the origin training set. It is judged whether the entropy of $p_{x_i^u}^{\delta}$ is smaller than that of $p_{x_i^u}^j$, $j \in \{1, 2, \ldots, c\}$, $j \neq \delta$, as shown in the following:

$$p_{x_i^u}^{\delta} \log\left(p_{x_i^u}^{\delta}\right) \geqslant \underset{j \in \{1,2,\ldots,c\}, j \neq \delta}{p_{x_i^u}^j \log\left(p_{x_i^u}^j\right)}. \tag{5}$$

Thus, hybrid label-less learning is obtained. In (4) and (5), we can see that the hybrid label-less learning can overcome the manual setting threshold value, maximize the utilization of labeled data and unlabeled data in the feature layer and decision layer, and enhance the confidence of the unlabeled data while adding it to the training set.

4) Enhanced Hybrid Label-Less Learning: The hybrid label-less learning can enhance the confidence of added data to a certain degree. However, if deeming the added unlabeled data based on the hybrid label-less learning strategy as being fully credible and retraining the model as the new training data set, the error accumulation of the training model may be caused. This is because the hybrid label-less learning may be labeled wrongly, leading to a noise of added data set and an training error.

To overcome the above problems, we propose the enhanced hybrid label-less learning. In other words, when the automatically labeled data based on the hybrid label-less learning is added to the training set, the newly added data should be reassessed, rather than confiding the newly labeled data. Specifically, the assessment algorithm is as follows. First, the automatically labeled data set based on the hybrid label-less learning is labeled as $\mathbf{z}$. Then, in the enhanced hybrid label-less learning, the k automatically labeled data sets are authenticated, and the data added to the enhanced hybrid label-less learning each iteration each class is denoted as $\mathbf{s}_j$. Thus, $|\mathbf{s}| = k$, where $|\cdot|$ is the number of elements. The enhanced hybrid label-less learning strategy is as follows. For any samples $x_i^u \subseteq \mathbf{s}$ and samples $(x_i^u)' \subseteq (\mathbf{z}_j - \mathbf{s}_j)$, data added to the enhanced hybrid label-less learning at each iteration should meet the following condition:

$$\underset{\forall x_i^u \subseteq s^j}{E\left(p_{x_i^u}\right)} \leq \underset{\forall \left(x_i^u\right)' \subseteq (z^j - s^j)}{E\left(p_{\left(x_i^u\right)'}\right)} \qquad j \in \{1, 2, \ldots, c\}. \tag{6}$$

In other words, the wrongly labeled data are corrected by reassessment. During the specific experimental process, to keep the class balance, the same amount of data are added to each class at each iteration. In addition, the data selected at each iteration are arranged in the incremental order.

B. Method of Label-Less Learning for Emotion Cognition

In the emotion cognition, according to the discussion presented in Section II, the speech and facial expression are two important modalities for emotion detection. The multimodal emotion detection based on speech and facial expression can not only utilize the information of speech and facial expression but can also utilize the information of other parties. Thus, the LLEC based on the multimodal data can enhance the emotion detection accuracy. However, the automatically labeled multimodal data are more complex, because the labels given by different modal data may be inconsistent. Thus, the labeling of the unlabeled data is a challenging problem.

The method of LLEC is as follows. We assume the facial expression data set is **xf**, and the speech data set is **xs**. Then, we can denote the labeled multimodal data set as $\mathbf{x}^l = (\mathbf{xf}^l, \mathbf{xs}^l)$, and the unlabeled multimodal data set as $\mathbf{x}^u = (\mathbf{xf}^u, \mathbf{xs}^u)$. Using the hybrid label-less learning, the unlabeled multimodal data are automatically labeled. First, we denote the most similar classifications of facial expression unlabeled data xf_i^u and speech unlabeled data xs_i^u as δ_1 and δ_2, respectively, and they are defined as

$$\delta_1 = \underset{j}{\operatorname{argmin}} \ \operatorname{Sim}\left(xf_i^u, Sf_j\right) \tag{7}$$

$$\delta_2 = \underset{j}{\operatorname{argmin}} \ \operatorname{Sim}\left(xs_i^u, Ss_j\right). \tag{8}$$

Then, we adopt the minimum joint entropy strategy for the automatic labeling of the unlabeled multimodal data. Given j as the classification of unlabeled multimodal data xf_i^u and xs_i^u, the joint entropy can be defined by

$$\bar{E}\left(p_{x_i^u}^j\right) = -p_{xf_i^u}^j \log\left(p_{xf_i^u}^j\right) - p_{xs_i^u}^j \log\left(p_{xs_i^u}^j\right). \tag{9}$$

Based on the similarity measurement, the most similar classifications of the unlabeled multimodal data xf_i^u, xs_i^u can be obtained as δ_1 and δ_2, respectively. According to the measurement of a minimum joint entropy, we can obtain the most possible classification of the unlabeled multimodal data xf_i^u, xs_i^u, which is denoted as δ, and defined by

$$\delta = \operatorname{argmin}\left[\bar{E}\left(P_{x_i^u}^{\delta_1}\right), \bar{E}\left(P_{x_i^u}^{\delta_2}\right)\right]. \tag{10}$$

Finally, we decide whether the unlabeled multimodal data xf_i^u, xs_i^u and their classification label, δ, should be added to the training set or not. It is important to evaluate whether the joint entropy of $p_{xf_i^u}^{\delta}$ and $p_{xs_i^u}^{\delta}$ is smaller than the joint entropy of other classifications

$$\bar{E}\left(p_{x_i^u}^{\delta}\right) \leqslant \bar{E}\left(p_{x_i^u}^j\right), \quad j \in \{1, 2, \ldots, c\}, \ j \neq \delta. \tag{11}$$

Using the above strategy, we can not only avoid the prediction conflict arising out of the multimodal data but also increase the correctness of automatic labeling data selection.

Furthermore, similar to the enhanced hybrid label-less learning, the new added data are reselected. According to (6), the reselection criterion for class j of each iteration is given by

$$\underset{\forall x_i^u \subseteq s^j}{\bar{E}\left(p_{x_i^u}^{\delta}\right)} \leq \underset{\forall x_i^u \subseteq (z^j - s^j)}{\bar{E}\left(p_{\left(x_i^u\right)'}\right)} \tag{12}$$

where the sample $x_i^u \subseteq \mathbf{s}$ and $\mathbf{s}$ denotes the added unlabeled multimodal data after the enhanced hybrid label-less learning. $(x_i^u)' \subseteq \mathbf{z} - \mathbf{s}$, $\mathbf{z}$ denotes the added unlabeled multimodal data after the hybrid label-less learning. Based on the above methods, after the unlabeled multimodal data are selectively added to the train set, they are further purified. The specific solving algorithm is shown in Algorithm 1.

Algorithm 1 LLEC Algorithm

Require:
The number of labeled data, n;
The number of unlabeled data, m;
Multimodal labeled data, $x^l = (xf^l, xs^l)$;
Multimodel unlabeled data, $x^u = (xf^u, xs^u)$;
Ensure:
New label data, x^l;

```
% Hybrid label-less learning
for i = 1 to m do
  % obtain the δ1 and δ2 based on the feature layer;
  δ1 = argmin_j Sim(xf_i^u, Sf_j);
  δ2 = argmin_j Sim(xs_i^u, Ss_j);
  % obtain the δ based on the decision layer;
  δ = argmin[Ē(P_{x_i^u}^{δ1}), Ē(P_{x_i^u}^{δ2})]
  if Ē(P_{x_i^u}^{δ}) ≤ Ē(P_{x_i^u}^{j}) for j ∈ {1,2,...,c}, j ≠ δ then
    z = z + {x_i^u, δ}
  end if
end for
% Enhanced hybrid label-less learning
I = ⌊size(z)/k⌋
for i = 1 to I do
  for j = 1 to c do
    Copy s^j from z^j, size(s^j) = k
    if Ē(p_{x_i^u}^{δ}) ≤ Ē(p_{(x_i^j)'}) for ∀x_i^u ⊆ s^j, ∀x_i^u ⊆ (z^j − s^j) then
      x^l = x^l + {x_i^u, j}
    end if
  end for
end for
```

C. Emotion Detection Based on the Edge Cloud

In the real emotion detection system, due to the network bandwidth, it is not practical to offload all the unlabeled data to the cloud that cause the long delay. Thus, in this paper, we use the edge cloud to reduce the time delay [24]. This is because the edge cloud computing can provide users with short-delay and high-performance computing services by deploying computing servers at the edge of the network, to meet users' requirements for delay-sensitive tasks [25], [26].

Specifically, we design the LLEC system using the edge cloud, as shown in Fig. 3. The details are as follows. First, the collected unlabeled multimodal data are offload to the edge cloud through a cellular network or Wi-Fi. On the edge cloud, based on label-less learning, the unlabeled data are labeled and selected. The communication delay is much lower, because the edge cloud is close to users' devices. Second, the edge cloud offloads the preliminary selected data to the remote cloud for processing. Through the preliminary selecting of unlabeled data on the edge cloud, the amount of data are greatly reduced and the data include only the most useful information. Thus, the accuracy of emotion detection in the remote cloud is also increasing.

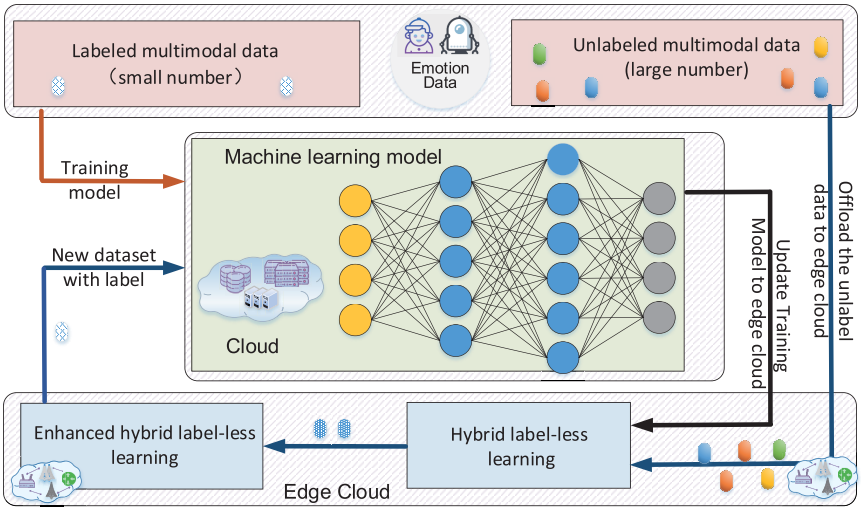


Fig. 3. LLEC method using the edge cloud.

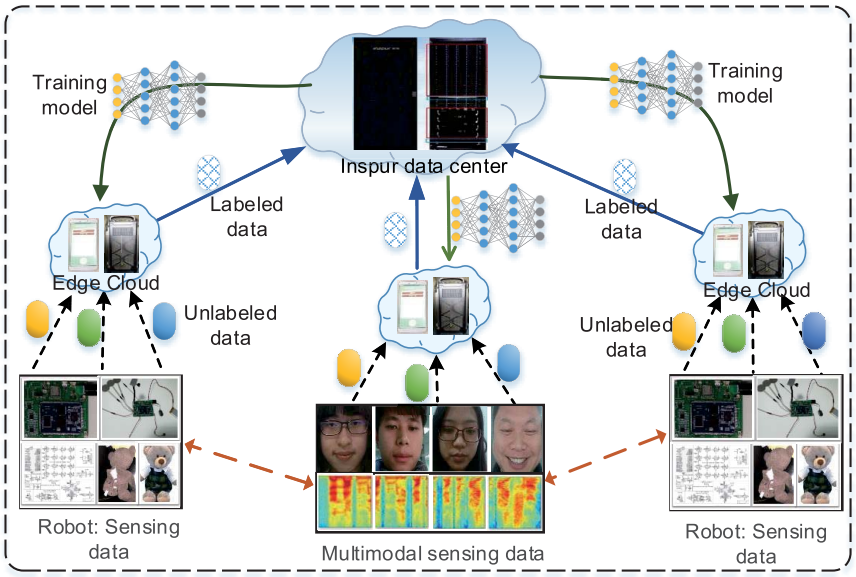


Fig. 4. Test bed of label-less learning for emotion cognition using edge cloud.

IV. Label-Less Learning for Emotion Cognition System

In this section, we design the robot-based LLEC system and apply the label-less learning method for the emotion detection and interaction. We first explain the test bed and data collection (including speech and facial expression) and then provide the result and analysis from two perspectives of emotion detection accuracy and communication delay.

A. System Test Bed

1) Robot-Based Emotion Detection and Interaction System: The robot-based emotion detection system test bed is as shown in Fig. 4. The system includes the robot with sensing multimodal emotional data, edge cloud, and inspur data center [27]. This test bed can detect user's emotion. The emotion cognition is as follows. First, the labeled speech and facial expression data are used to train the machine learning model. In this experiment, we adopt AlexNet DCNN as the machine learning model. Then, the trained DCNN model is employed to predict the label, in view of unlabeled speech and facial expression data. Finally, the unlabeled data are added to the new model in accordance with the proposed LLEC algorithm. Thus, a small amount of labeled data are needed.

2) Emotion Data Collection and Detection: In this paper, we use two emotional data: first, we use the existing multimodal emotion data set (i.e., enterface05 data set). Second, we use the data set collected by the WebChat. For the multimodal emotion data, we first adopt the enterface05 data set [28]. The enterface05 data set includes 1290 sections of videos, including 43 different speakers and speech contents in English. The set included six basic emotions, i.e., anger, disgust, fear, joy, sadness, and surprise. The audio sample

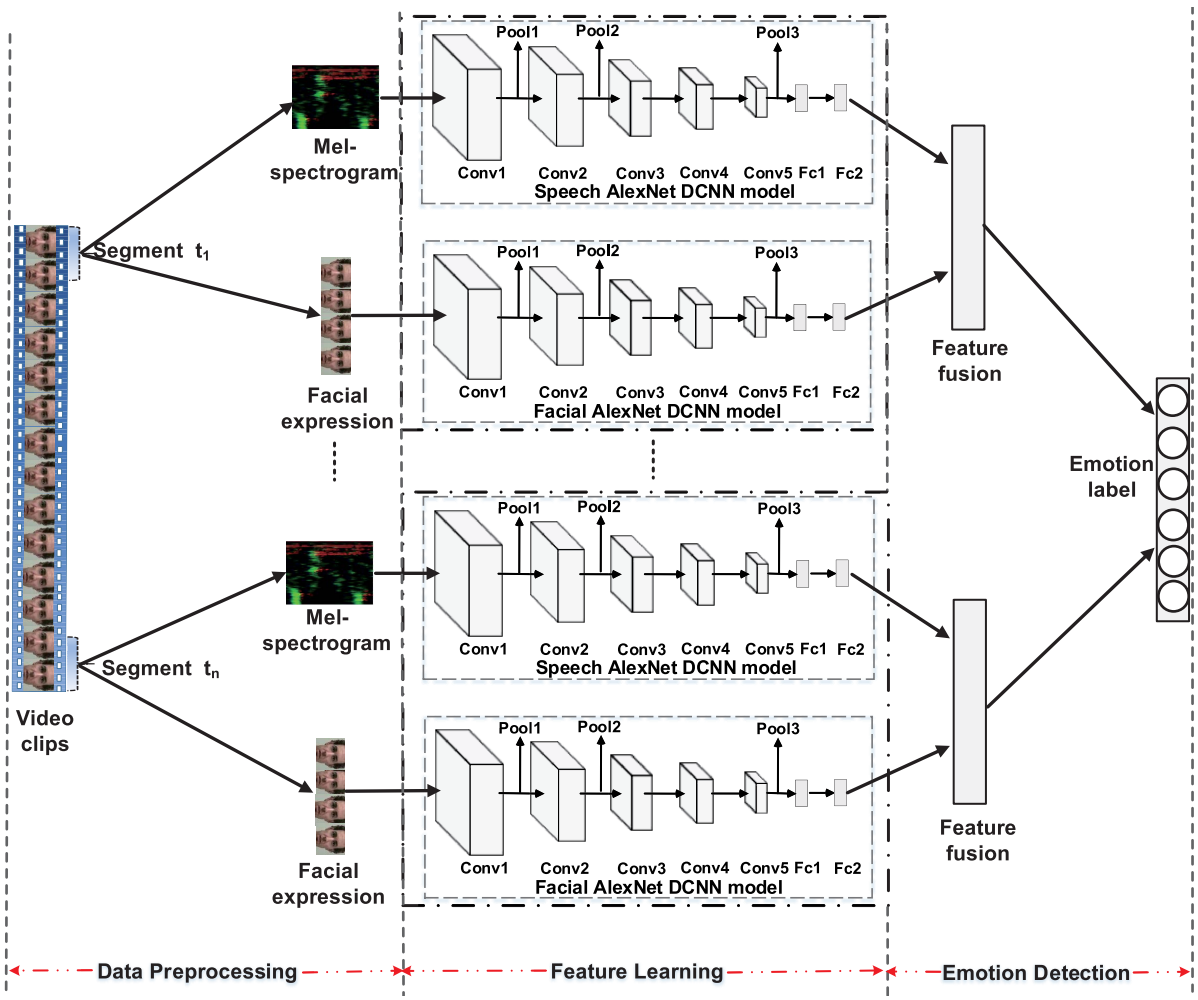


Fig. 5. Emotion detection using DCNN model.

rate is 48 kHz, and the dual audio track is of 16-bit precision. The emotions are performed by the participants. The video frame frequency is 25 frames/s, and the image size is $368 \times 240 \times 3$ pixel.

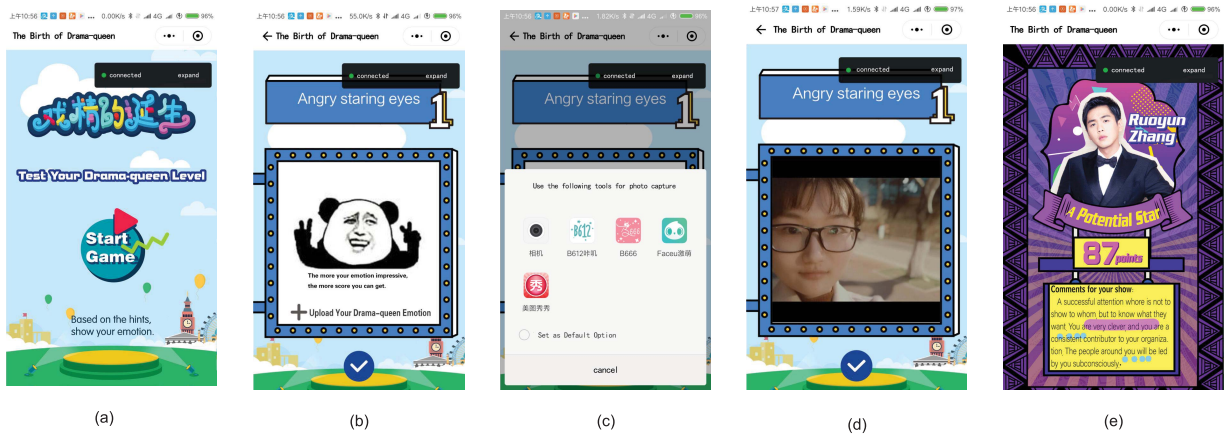
For the above video data, we first derive the speech and facial expression from video. Then, we perform data preprocessing on speech and facial data, respectively. To be specific, for speech data, we convert a speech into a 2-D spectrum (i.e., an image) using the method in [6], and finally, we can turn the speech signal into the log Mel-spectrogram with the size of $64 \times 64 \times 3$ containing three channels (static, δ, and $\delta - \delta$), which is the input of the DCNN. In the case of facial expressions, similar to the speech emotions and method in [6], we can obtain the facial expression. In this experiment, we use an average of face expression sequence and regarded it as an input of the DCNN.

Then, for the emotion cognition model, we first use AlexNet DCNN to extract the features of the above processed speech and facial data, next combine the extracted features, and, finally, emotion label will be acquired, as shown in Fig. 5. The AlexNet DCNN architecture includes five convolution layers (Conv1-Conv2-Conv3-Conv4-Conv5), two maximum pooling layers (Pool1-Pool2), and two fully connected layers. In this paper, the DCNN is trained by the stochastic gradient descent (SGD) training method. Owing to a small volume of speech and facial expression emotion data, the initial network is trained on the large-scale ImageNet data set. The specific training process is as follows. First, the AlexNet network parameters are initialized, and then the network parameters are fine-adjusted by the enterface05 data set.

Furthermore, we collect the facial expression labeled data using a WebChat application, as shown in Fig. 6. In Fig. 6, it can be seen that the WeChat includes five parts, game home page, prompt page, opening the camera and tasking the photos, facial expression collection, and game score. Using WeChat, we collected users' facial expression and corresponding labels. Specifically, we derive the facial expressions data collected by Wechat and process them using the method of facial expression in Fig. 5.

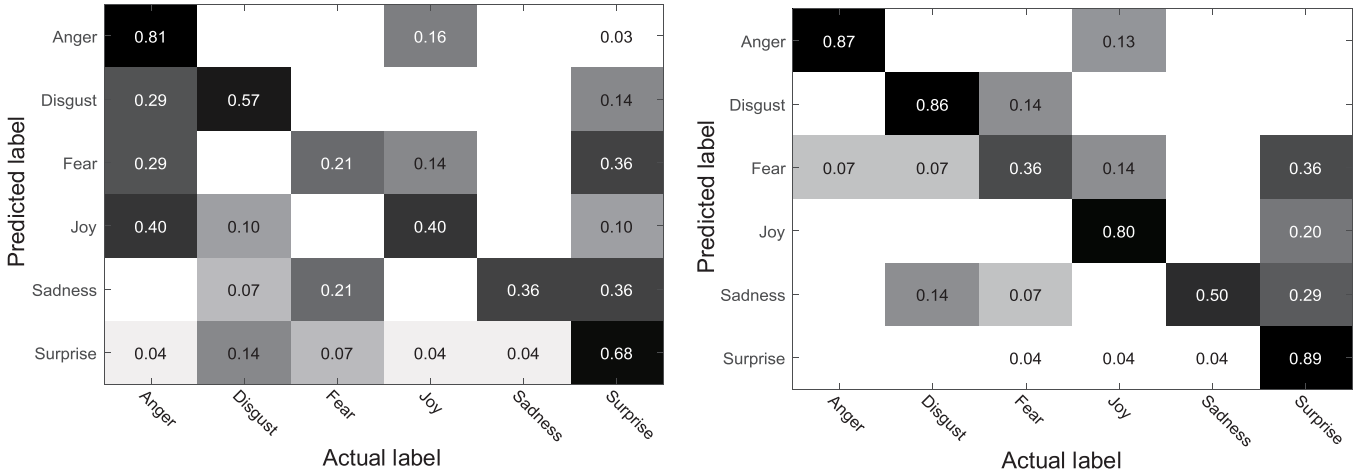
B. Experimental Results and Analysis

Based on the LLEC system test bed, we give the experimental results and analysis from the following three aspects, i.e., emotion detection with labeled data and unlabeled data, the evaluation of LLEC method, and average time delay. The first and second are using the collected emotion data. The third is using the robot-based emotion detection and interaction system.



(a) (b) (c) (d) (e)

Fig. 6. Facial Emotion data collection. (a) Game home page. (b) Prompt page. (c) Opening the camera and taking the photos. (d) Facial expression collection. (e) Game score.



(a) (b)

Fig. 7. Experiment results. (a) Confusion matrix of the original labeled data set. (b) Confusion matrix after adding the unlabeled data to the data set using LLEC.

1) Emotion Detection With Labeled Data and Unlabeled Data: In this experiment, we randomly divide the data instance in the enterface05 data set into a training set and a test set. The ratio of the training set to the test set is 3:1. For the data in the training set, we randomly divide it into labeled data and unlabeled data, where the ratio of labeled data to unlabeled data is 1:2. To be specific, labeled data are used to train the initialization model, and the unlabeled data are labeled based on the initialization model and the LLEC algorithm. The test set is utilized to evaluate the accuracy of the model.

The emotion detection accuracy using only the labeled data is presented in Fig. 7(a), where the x-axis is the actual emotional label, and the y-axis represents the predicted emotional label. The diagonal line in Fig. 7(a) represents the probability of emotion detection accuracy, and other parts show the probability of emotion detection error. In Fig. 7(a), it can be seen that the accuracy of angry detection is 0.81. The probability of recognizing anger as happiness is 0.16 and the probability of recognizing anger as neutral emotion is 0.03. The emotion detection accuracy after using the LLEC algorithm is presented in Fig. 7(b), wherein it can be seen that the accuracy of emotion detection is improved. For instance, the detection accuracy of angry is 0.84, which is a better result than when the LLEC is not used. Thus, through the use of unlabeled data, we improve the accuracy of emotion detection.

2) Evaluation of LLEC Method: We give the comparison of emotion detection accuracy between the following four LLEC algorithms: decision layer-based LLEC, feature layer-based LLEC, hybrid LLEC, and enhanced hybrid LLEC. In this experiment, we first train the model with 323 labeled data in the training set of the enterface05 data set. Then, in each time, we add 100 unlabeled data instances from enterface05 data set and WeChat data set into training set, and use the four algorithms mentioned above to label the unlabeled data. The new labeled data are added to the new training set, and the accuracy of emotion detection is based on the DCNN algorithm. The result is shown in Fig. 8. The x-axis represents the number of unlabeled data instances, and the y-axis represents the average accuracy of the six emotional classes.

From Fig. 8, we can conclude that enhanced hybrid LLEC has the highest emotion detection accuracy. Hybrid LLEC, decision layer-based LLEC, and feature layer-based

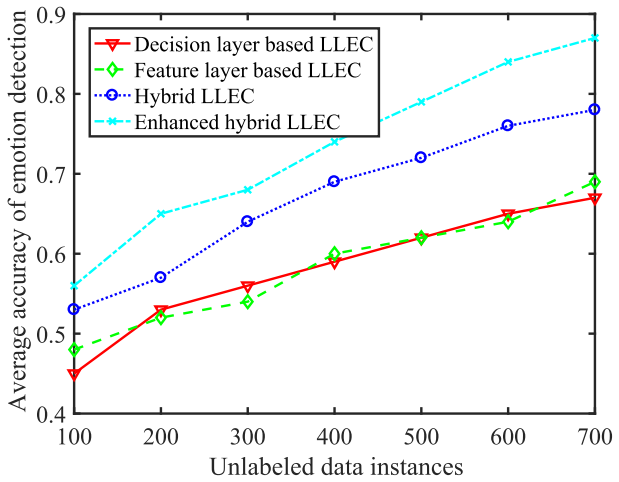


Fig. 8. Emotion detection accuracy comparison among decision layer-based LLEC, feature layer-based LLEC, hybrid LLEC, and enhanced hybrid LLEC under different numbers of unlabeled data instances.

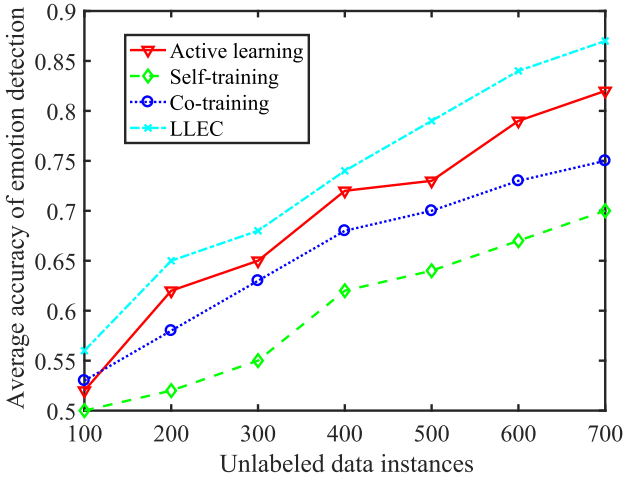


Fig. 9. Emotion detection accuracy comparison among active learning, self-training, cotraining, and LLEC under different numbers of unlabeled data instances.

LLEC have little difference. This is because decision layer-based LLEC and feature layer-based LLEC give the labels of unlabeled data at the decision layer and the feature layer, respectively, which may cause a wrong label. Hybrid LLEC can make full use of the decision layer and the feature layer, so the utilization of unlabeled data is further enhanced. Enhanced hybrid LLEC not only makes full use of the hybrid LLEC but also can further filter the newly added label data; thus, the accuracy of its emotional detection is the highest.

Furthermore, we compare the enhanced hybrid LLEC algorithm proposed in this paper with active learning, self-training, and cotraining, and give the accuracy of emotion detection under different numbers of unlabeled data. From Fig. 9, we can see that the LLEC algorithm proposed in this paper is the best. This is because the proposed LLEC has two advantages: 1) it can validate emotion data from two modalities of face and image and 2) it can further filter the unlabeled data of automatic labeling to validate the accuracy of the label data.

3) Average Delay Comparison: We give the average delay of the LLEC algorithm in the following two situations, i.e., LLEC in edge cloud and LLEC in the cloud.

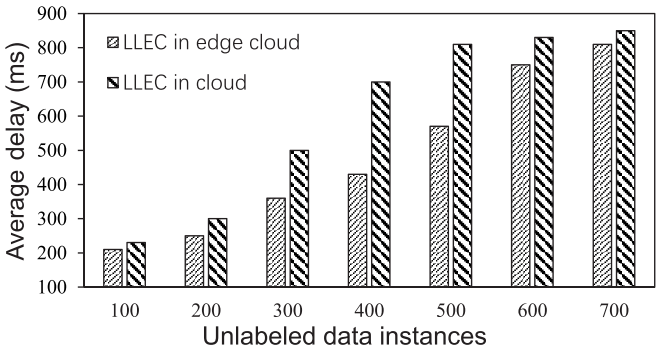


Fig. 10. Time average delay comparison of LLEC algorithm in edge cloud and cloud.

1) *LLEC in edge cloud:* as shown in Fig. 3, the LLEC algorithm runs in the edge cloud, and the DCNN algorithm runs in the cloud. Specifically, we regard robot as the sensing node, small server as the edge cloud, and regard inspur data center as the remote cloud.
2) *LLEC in cloud:* the LLEC and DCNN algorithms run in the cloud. We give the average delay of 10 times running of LLEC in the edge cloud and LEC in the cloud under different numbers of unlabeled data. From Fig. 10, we can obtain that the average delay of LLEC in edge cloud is lower than that of LLEC in cloud. This is because LLEC in edge cloud can filter unlabeled data, thus reducing transmission delay of unlabeled data.

V. Conclusion

Due to a large amount of unlabeled data in emotion detection, we propose a new method for emotion cognition called the LLEC. The proposed LLEC first trains the neural network model by using a small amount of multimodal labeled data. Then, it labels the unlabeled data automatically and adds it to the training set using the enhanced hybrid label-less learning, to further improve the model detection accuracy. The proposed method is validated by the real test bed. The experimental results show that the LLEC algorithm can improve the emotion detection accuracy significantly. In our future work, we will consider reducing the complexity of the LLEC algorithm and further improve the accuracy of emotion detection by using deep reinforcement learning [29] for emotion detection.

References

[1] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, Mar. 2019.
[2] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, Feb. 2018.
[3] B. Shi, X. Bai, W. Liu, and J. Wang, "Face alignment with deep regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 183–194, Jan. 2016.
[4] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 491–497.
[5] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1496–1509, Jun. 2017.

[6] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.
[7] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[8] X. Yang and H. He, "Self-learning robust optimal control for continuous-time nonlinear systems with mismatched disturbances," *Neural Netw.*, vol. 99, pp. 19–30, Mar. 2018.
[9] M. Chen, Y. Hao, H. Gharavi, and V. Leung, "Cognitive information measurements: A new perspective," *Inf. Sci.*, 2019. doi: 10.1016/j.ins.2019.07.046.
[10] G. Yang, H. He, and Q. Chen, "Emotion-semantic-enhanced neural network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 531–543, Mar. 2019.
[11] L.-W. Kim, "DeepX: Deep learning accelerator for restricted boltzmann machine artificial neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1441–1453, May 2018.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 2012, pp. 1097–1105.
[13] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.
[14] B. Sun, Q. Ma, S. Zhang, K. Liu, and Y. Liu, "iSelf: Towards cold-start emotion labeling using transfer learning with smartphones," *ACM Trans. Sensor Netw.*, vol. 13, no. 4, p. 30, 2017.
[15] Z. Zhao and X. Ma, "Active learning for speech emotion recognition using conditional random fields," in *Proc. 14th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput. (SNPD)*, Jul. 2013, pp. 127–131.
[16] Q. Gao, Y. Huang, X. Gao, W. Shen, and H. Zhang, "A novel semi-supervised learning for face recognition," *Neurocomputing*, vol. 152, pp. 69–76, Mar. 2015.
[17] D. Wu *et al.*, "Self-training semi-supervised classification based on density peaks of data," *Neurocomputing*, vol. 275, pp. 180–191, Jan. 2017.
[18] Y. Pang, M. Sun, X. Jiang, and X. Li, "Convolution in convolution for network in network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1587–1597, May 2018.
[19] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.
[20] H. Shi, S. Pan, J. Yang, and C. Gong, "Positive and unlabeled learning via loss decomposition and centroid estimation," in *Proc. IJCAI*, Jul. 2018, pp. 2689–2695.
[21] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller, "Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning," *IEEE Access*, vol. 6, pp. 22196–22209, 2018.
[22] W. Zhan and M.-L. Zhang, "Inductive semi-supervised multi-label learning with co-training," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1305–1314.
[23] M. Chen, Y. Hao, K. Lin, L. Hu, and Z. Yuan, "Label-less learning for traffic control in an edge network," *IEEE Netw.*, vol. 32, no. 6, pp. 8–14, Nov./Dec. 2018.
[24] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. IEEE 10th Int. Conf. Intell. Syst. Control (ISCO)*, Coimbatore, India, Jan. 2016, pp. 1–8.
[25] Y. Sun, S. Zhou, and J. Xu, "EMM: Energy-aware mobility management for mobile edge computing in ultra dense networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2637–2646, Nov. 2017.
[26] L. Tong, Y. Li, and W. Gao, "A hierarchical edge cloud architecture for mobile computing," in *Proc. 35th IEEE Conf. Comput. Commun.*, Apr. 2016, pp. 399–400.
[27] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "AIWAC: Affective interaction through wearable computing and cloud technology," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 20–27, Feb. 2015.
[28] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng.*, Apr. 2006, p. 8.
[29] H. He and X. Zhong, "Learning without external reward," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 48–54, Aug. 2018.

Min Chen (SM'09) has been a Full Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China, since 2012. His Google Scholars Citations reached more than 17,500 with an h-index of 65. His current research interests include Internet of Things (IoT) sensing, 5G networks, healthcare big data, and cognitive computing.

Prof. Chen was a recipient of the IEEE Communications Society Fred W. Ellersick Prize in 2017 and the IEEE Jack Neubauer Memorial Award in 2019. He is the Chair of the IEEE Computer Society STC on Big Data. He was recognized by Clarivate Analytics as a Highly Cited Researcher in 2018.

Yixue Hao received the B.E. degree from Henan University, Kaifeng, China, in 2013, and the Ph.D. degree in computer science from the Huazhong University of Science and Technology (HUST), Wuhan, China, 2017.

He is currently a Post-Doctoral Scholar with the School of Computer Science and Technology, HUST. His current research interests includes 5G network, Internet of Things, edge caching, and mobile edge computing.