IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 23, 2020, accepted October 3, 2020, date of publication October 7, 2020, date of current version October 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029288

Multistep Deep System for Multimodal Emotion Detection With Invalid Data in the Internet of Things

MINJIA LI[1], LUN XIE[1], ZEPING LV[2], JUAN LI[3], AND ZHILIANG WANG[1]
[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]Rehabilitation Hospital, National Rehabilitation Auxiliary Center, Beijing 100176, China
[3]Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

Corresponding author: Lun Xie (xielun@ustb.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC2001700, in part by the National Natural Science Foundation of China (Normal Project) under Grant 61672093, and in part by the Beijing Municipal Natural Science Foundation under Grant L192005.

ABSTRACT The Internet of Things (IoT) technologies such as interconnection and edge computing help emotion recognition to be applied in healthcare, smart education, etc. However, the acquisition and transmission processes may have some situations, such as lost signals and serious interference noise caused by motion, which affect the quality of the received data and limit the performance of IoT emotion detection. We collectively refer to these as invalid data. A multi-step deep (MSD) system is proposed to reliably detect multimodal emotion by the collected records containing invalid data. Semantic compatibility and continuity are utilized to filter out the invalid data. The feature from invalid modal data is replaced through the imputation method to compensate for the impact of invalid data on emotion detection. In this way, the proposed system can automatically process invalid data and improve the recognition performance. Furthermore, considering the spatiotemporal information, the features of video and physiological signals are extracted by specific deep neural networks in the MSD system. The simulation experiments are conducted on a public multimodal database, and the performance of the MSD system measured by the unweighted average recall is better than that of the traditional system. The promising results observed in the experiments verify the potential influence of the proposed system in practical IoT applications.

INDEX TERMS Internet of Things, multimodal emotion detection, invalid data, multi-step deep (MSD) system, deep neural networks.

I. INTRODUCTION

The ability to perceive human emotions can be used to provide more personalized interactive products. Due to the massive data and powerful computational capacity, the fast growth of Internet of Things (IoT) technologies helps to realize the possibility of real-time human emotion detection or perception in different scenarios. There have been many novel studies combining the IoT and affective computing to provide various emotion detection frameworks for different applications, such as healthcare services [1], battlefield environments [2] and smart homes [3]. These studies treat emotion detection as a part of the overall framework and tend to explain the feasibility of communication and interaction processes.

The IoT based on cloud servers can turn the monitoring of real-time emotional states using big data into a reality. However, there are still some limitations in practical applications. IoT data consist of a continuous stream of data derived from terminal sensors. The data contain information pertaining to the physical states of human beings, such as their physiological signals and facial expressions. Therefore, the reliability of sensors would affect data quality. In addition, Transmission protocols and network environments also affect synchronization and the quality of IoT data. Considering the existence of the above problems, the quality of data in the IoT is not as good as the modal data collected by the lab. Specifically, we use the phrase ''invalid data'' to refer not only to the missing data, but also to the data whose extracted feature does not

The associate editor coordinating the review of this manuscript and approving it for publication was Eyhab Al-Masri.

represent the current emotional state, such as received images that are captured to the wrong location instead of the human face. The work of Azimi et al [4] showed that the performance of emotion detection using the IoT could be vulnerable to the inconsistency and incompleteness of invalid data. Invalid data may lead to invalid features that do not have the ability to represent the latent emotional semantic information; and in a further detection model, the invalid features cannot be mapped to the right labels. Hence, the detection performance can be affected by invalid data.

Moreover, most studies on multimodal emotion detection focus on the audiovisual method. However, in some situations, such as medical monitoring, audio signals are not long-standing and perform as noncontinuous signals. Physiological signals can provide information regarding the intensity and quality of an individual's internal state [5]. Some peripheral physiological signals derived from wearable noninvasive devices make real-time stress monitoring a reality for humanity [6]. In addition, image data contain facial expression information, which is the most intuitive emotional representation that can further help to understand the emotional state of an individual. Therefore, the combination of peripheral physiological signals and expression information can improve the recognition performance due to the complementarity of multimodal information [7].

Therefore, we can summarize the two problems to be solved in this paper as follows:

1. How can the emotion detection performance be protected from invalid data?

2. How can the information of multimodal data be effectively utilized in the presence of invalid data?

Various works have addressed each of these problems separately (see section 2). Most work focused on improving the sensor quality and network architecture to prevent the generation of invalid data, which requires higher costs and is less selective for application scenarios. We build a multi-step deep (MSD) system that addresses the above two problems simultaneously. The spatiotemporal, semantic information and multimodal complementary information are utilized together to reduce the impact of invalid data and further improve the emotion detection performance.

In order for the MSD system to protect the emotion detection performance from invalid data, the methodological difficulty is how to filter out invalid features automatically. We construct the discriminative module considering the semantic compatibility and continuity instead of the traditional methods addressing outliers [8], [9]. These traditional methods measure the similarity between the features from invalid data and the features from other effective data, and the data with small similarities are treated as outliers. However, the computation of this similarity can considerably increase the computing and time costs, which is unrealistic for real-time practical applications. In addition, the computation of the similarity between high-dimensional multimodal features can lead to the famous ''curse of dimensionality'' problem. By contrast, utilizing the latent semantic information can avoid the calculation of excessive dimensions. Furthermore, potential semantic information encompasses the relationship between modalities' features and the corresponding emotional information, enabling the indirect measurement of the similarity between different modal features. Hence, we evaluate the compatibility between modalities and labels in the latent semantic spaces, and the discriminative module filters out the invalid features considering the semantic difference and the modal continuity. In principle, the discriminative module can be extended to various combinations of modalities that can be preconverted into 1-dimensional feature vectors. With the detected invalid features being discarded, the new features are produced by the traditional imputation method and used as an alternative to invalid data. The new combination of features is input to the final detection model.

To utilize sufficient information of multimodal data, the spatiotemporal feature method by 3-dimensional convolution neural networks (C3Ds) and deep belief networks (DBNs) is introduced and converts all raw data into 1-dimensional vectors, which are involved in the discrimination, compensation, and detection processes. The main contributions of this work are threefold:

1. We established an MSD system for multimodal emotion detection with records containing invalid data.

2. A method of real-time multimodal emotion detection based on the combination of peripheral physiological signals and video was implemented.

3. We conducted the experiment on the Remote Collaborative and Affective Interactions (RECOLA) [10] database to imitate emotion detection with temporarily invalid data in the IoT and further investigated the effectiveness of the proposed system.

The remainder of this paper is organized as follows. Section II presents the related work, Section III describes our proposed system, Section IV illustrates our experimental results, Section V provided the discussion and Section VI concludes the paper.

II. RELATED WORK

A. AFFECTIVE COMPUTING IN THE IOT

Several existing works combining the IoT and affective computing focus on four detailed aspects: data collection, the network architecture, the detection model and the interaction framework in a specific scene.

The emergence of servers and various sensors in the IoT is the cornerstone of emotion detection. Hui and Sherratt [11] used common wearable biosensors to predict human emotion and addressed the problems of limited processing power, size constraints and battery capacities existing in the selection of embedded sensors using a lightweight methodology. Chen *et al.* [12] achieved emotion sensing through the collection of ECG signals by smart clothing and the behavioral data by smartphones. Kim *et al.* [13] innovatively provided an idea to create a virtual emotion barrier that used a wireless signal and its reflection for emotion detection.

Since the collected data must be transmitted to the detection model through communication, communication quality is also one of the main factors affecting the results of emotion detection. The emerging 5G technology helps the implementation of big data-oriented wireless technologies. Hossain and Muhammad [1] proposed a framework using 5G technology for personalized and seamless emotion-aware healthcare services. The network architecture affects the quality of the received data, and a reasonably flexible architecture can use edge computing [14] to speed up the computational response. For example, Hao *et al.* [15] introduced a smart-edge- Computation, Caching, and Communication (CoCaCo) algorithm to reduce the computation delay in an affective interaction as the amount of computing task data and the number of concurrent users increase in a real environment.

The work of Alam *et al.* [16] focused on the detection model part. They constructed an affective state mining framework using a distributed CNN-based module to recognize human emotions through biosignals. However, only the modality of the biosignal was utilized, and there was a lack of consideration for the applicability of transmitted modal data.

Eriksson *et al.* [17] designed personal emotion-tracking applications to conduct emotional evaluations and support emotion-sensing IoT systems. Chen *et al.* [18] proposed a prototype system of Smart Home 2.0 to achieve the affective interactions between householders and greens. Lin *et al.* [2] focused on obtaining soldiers' emotions in a battlefield environment based on an emotion-aware system model. These diverse studies have confirmed that the combination of the Internet of Things and affective computing has broad application scenarios, and the practicality and pertinence of these IoT systems improves as the theory and algorithms are further implemented.

Although the emphases of these studies are different, almost all of the works aim to improve the emotion detection performance in the IoT from different aspects. In addition, existing research is more inclined to improve the reliability of the data sources to improve information acquisition and transmission in the system construction. However, there is little research work on the processing of invalid data existing in collected records. Due to the complexity and uncontrollability of the real communication environment, the goal of our paper is to contribute to the processing of these data to make the results more robust, which helps to further extend the emotion detection applications in a variety of scenarios.

B. EMOTION FRAMEWORK

Definition and assessment are the cornerstones of emotion recognition. The definition of emotional state originated from Charles Darwin's study of the evolution of emotion, who treated the emotions as separate discrete entities, or modules [19]. At present, the most popular discrete emotion model is the six emotional states proposed by Ekman and Paul [20], which are often used in the research of facial expression recognition. Since discrete categories may not fully reflect the complexity of the emotional state, another main theory is to use multiple dimensions to label emotions. Dimensional model aims to avoid the limitations of discrete models, and allow more flexible definition of emotional states as points in a multidimensional space. The two most commonly used dimensions are valence and arousal [21]. The former is related to whether the emotions are positive or not, while the latter measures how calming or exciting the subject is.

The subjective measurement of emotive responses is one of the main factors affecting the performance of emotion recognition. Due to heavy difficulties and inherent ambiguities in emotional displays, the ground truth is generally unreliable. Therefore, the training set is of poor quality, which further leads to the limited generalization ability and overfitting. Therefore, many studies on public databases have tried to reduce the subjectivity of label measurement, so as to ensure that there is a clear probability distribution relationship between samples and labels. After reviewing the literatures on multi-modal affective databases, the three mainstream approaches aimed at reducing subjectivity can be summarized: bagging, sorting and data cleaning. In RECOLA database, 6 assistants received a document including a short list of some well identified emotional cues to perform the annotation task. The database experiment followed the idea that as many assistants as possible are arranged to annotate the same sample to obtain the final ground truth by voting or averaging. Similar to model averaging, one of the common strategies in machine learning, this approach can reduce variance by increasing randomness. Annotations of LIRIS-ACCEDE [22] database are performed via crowdsourcing under a pairwise comparison protocol, thereby ensuring that the annotations are fully consistent. eNterface database [23] discards the samples in which no emotion is clearly recognized. This data cleaning approach can ensure that the database only contains the samples carrying obvious emotional displays.

The dimensional space can be operationalized as a regression task or as a classification task which discretizes the continuous space into several regions. Regression tasks tend to evaluate the similarity between ground-truth and emotional predictions, and its output is on the metric space. Classification tasks' output is qualitative. The difference between the two tasks makes the emotion recognition performance of different tasks impossible to fairly measure.

The three mainstream emotion calculation approaches include knowledge-based methods, statistical learning methods, and hybrid methods [24]. The knowledge-based method realizes the rapid mapping between keywords and emotional labels by constructing rules. Obviously, this method is difficult to work with modal data that contains complex semantics or has complex forms. The most popular method at present is the statistical method. This method uses machine learning and deep learning algorithms and has a strong learning ability to construct mapping from various forms of data onto emotional labels. For example, Zhang *et al.* [25] provided a machine learning framework, called dynamic difficulty

awareness training (DDAT), to utilize the difficulties in learning to promote the performance of emotion prediction model. The hybrid method, which combines the statistical learning method and knowledge-based method, can predict emotion and detect polarity. For example, Chaturvedi *et al.* [26] combined deep convolutional neural networks and fuzzy logic models to construct a convolutional fuzzy sentiment classifier (CSFC) to predict the degree of specific emotions.

C. MULTIMODAL EMOTION DETECTION

At present, multimodality emotion detection by statistical learning approaches usually uses the fusion method. It tends to deal with the feature of each modality separately and affects the unified recognition effect by combining different layers. The fusion method is divided into feature-layer fusion and decision-layer fusion. The former aggregates each modal feature and then inputs them together into the recognition model [27]–[29]; the latter trains each modal feature separately and outputs an n-type probability distribution as the input to the decision-making layer [30].

The classical work [31] from Johannes *et al.* considered the situation of temporarily unavailable modalities caused by some missing data. They obtained the final recognition results from 11 decision fusion categories through combinations of different quantities of subresults derived from the respective modality to ignore the influence of the missing data. However, the contribution from the implicit relationship between different modalities was not evaluated. To overcome this drawback, Du *et al.* [32] proposed a novel multiview deep generative framework that computed the existing shared latent variables between generative networks trained separately from multimodal data. In addition, the holistic scheme of generating missing data can be treated as a specialized missing data imputation task. However, there are still some limitations. First, the framework did not consider the conditions under which multimodal data were invalid except missing data. Second, only the latent variables between modalities are measured. This means that the multimodal data at each moment are processed independently, regardless of the temporal continuity of the modality including emotional information.

In practical IoT applications, invalid data also affect the detection process. Therefore, in our work, we not only propose improved methods working around the aforementioned two limitations, but we also implement an emotion detection experiment with invalid data existing in the test set based on multimodal data from video and physiological signals.

III. METHODOLOGY

The complete emotional IoT framework involves four aspects: (1) Data Acquisition; (2) Data Storage; (3) Transmission; and (4) Application Process, such as multimodal emotion detection that is researched in this article. Data acquisition involves mobile sensors, fixtures, wearable sensors, Local Area Network (LAN) and other components connected to remote or edge clouds. Various LAN structures, such as Body Network, can be built in this aspect to obtain a variety of raw multimodal data according to the required distance between sensors and human bodies. Data storage provides a high-performance and robust data storage for a vast amount of time series data from sensors, so as to ensure efficient and fast access to data for machine learning and deep learning computations. Transmission is related to the communication infrastructures as a bridge between Local and Cloud. As this study mainly focuses on data processing of emotion detection based on video and physiological signals, the rest aspects of IoT framework can solely rely on the methods described in relevant works, such as [1], [11], [33].

In this section, we present the problem setup and describe the MSD system we tested. Figure 1 shows a brief overview of the system. The system has four parts: (1) Feature Extraction; (2) Compatibility Measurement; (3) Discriminative Module; and (4) Compensation and Detection.

A. PROBLEM SETUP

Emotion detection is mainly divided into two steps, namely feature extraction and classification. The feature extraction process is to map the preprocessed data into feature vectors. Assuming that the input is $\boldsymbol{x}$, the feature extraction process is defined as:

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}) \tag{1}$$

where $f(\cdot)$ denotes the model of feature extraction, and $\dot{\boldsymbol{x}}$ is the corresponding feature. The classification process, with the input $\dot{\boldsymbol{x}}$, is given by:

$$y = model(\dot{\boldsymbol{x}}) \tag{2}$$

where $model(\cdot)$ denotes the model for classification and y is the predicted label. For the feature fusion method of multi-modal emotion detection, the input $\dot{\boldsymbol{x}}$ is a vector of multiple modal features, which are connected in series.

We define one training instance for a certain short period of time as:

$$T_i^r = \left\{ s_i^r, z^{(i)} \right\} \tag{3}$$

$$s_i^r = \left\{ V^{(i)}, (g_k^{(i)})_{k=1}^K \right\} \tag{4}$$

where $V^{(i)}$ denotes the ith video feature that derives n_i continuous frames and is given a ground-truth label $z^{(i)}$, where $z^{(i)} \in \{z_1, z_2..z_M\}$. There are a total of K physiological signals, and each signal's feature is described as $g_k^{(i)}$.

Similarly, the test instance is defined as:

$$T_j^{te} = \left\{ s_j^{te}, \tilde{z}^{(j)} \right\} \tag{5}$$

where $\tilde{z}^{(j)}$ denotes the predicted label corresponding to s_j^{te}. There may be features from invalid data existing in s_j^{te}. We use V^* and g^* to refer to the invalid features derived from video and physiological signals, respectively. Among them, the invalid feature suffering from missing data is treated as a null value. The main goal of the proposed system is to

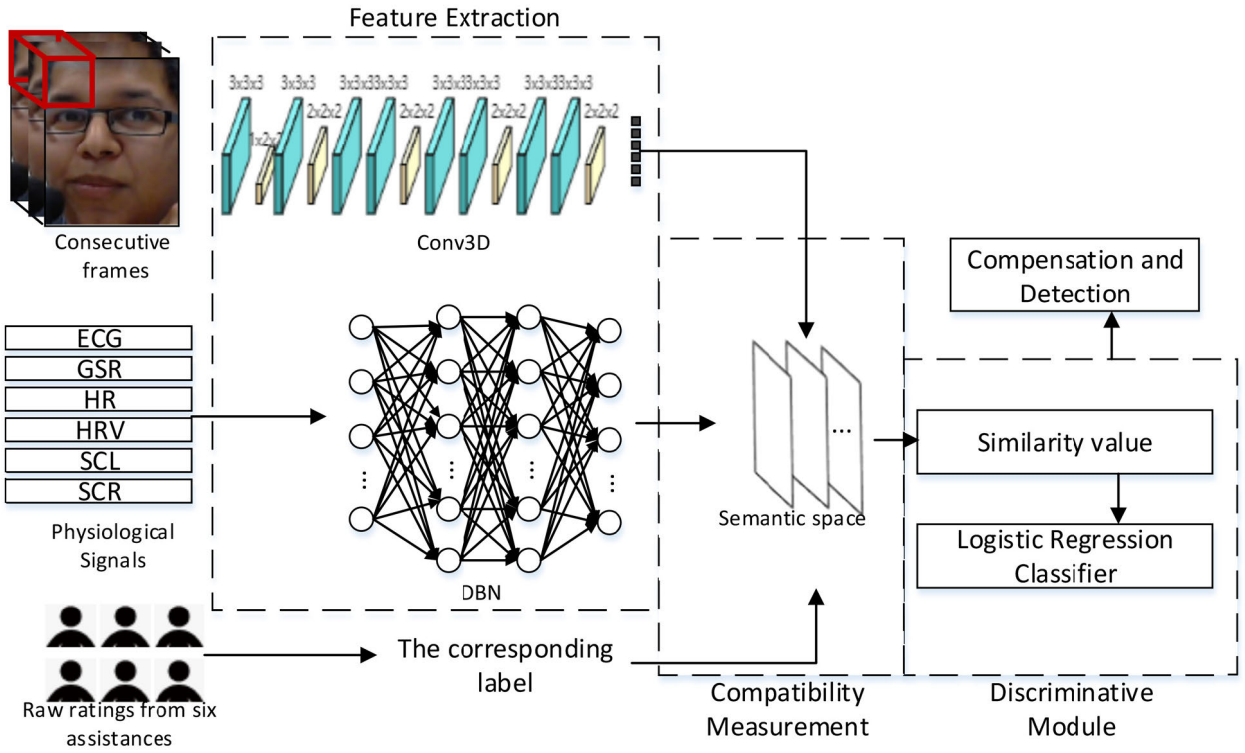


FIGURE 1. Illustration of the MSD system, which describes the feature extraction, compatibility measurement, discrimination and detection processes.

estimate $\tilde{z}^{(j)}$ given $(s_i^r)_{i=1}^N$ and s_j^{te} containing invalid features V^* or g^*.

As shown in Figure 1, we utilize the latent semantic space to evaluate the compatibility between modalities and labels. In addition, the discriminative module is constructed considering the modal continuity and semantic differences between different modalities to determine whether invalid data are present. The compensation module utilizes the imputation method to find a feature as the substitute for the invalid feature, further participating in the next classification process.

B. FEATURE EXTRACTION

DBNs and C3Ds are applied in this part for feature extraction. DBNs can be treated as a stack of multiple restricted Boltzmann machines (RBMs) and achieve an optimal solution by adopting layer-by-layer training and fine-tuning [34]. Due to its powerful unsupervised learning style in many nonlinear hidden layers, DBNs are exceptionally good at capturing implicit nonlinear information in the data structure. C3Ds have a typical 3D convolution net structure, which is more suitable for extracting temporal-spatial features than 2D nets [35]. In addition, C3Ds have the advantage of high computational efficiency due to their concise structure.

We utilized the C3D model to extract the temporal-spatial features from video. The features from multiple one-dimensional physiological signals are extracted by DBNs. The detailed process is described as follows.

1) EXPRESSION FEATURE

To reduce the influence of insignificant regions, we first capture the face region using a face detection algorithm. Considering the group application scenarios, the advanced work of Jian *et al.* [36], which can detect faces in various situations, has been applied to the RECOLA dataset. Each facial image is reconstructed with a size of 112×112.

We established the network structure and use the pretrained model "sport1m" following the works of Du *et al.* [35] and Yin *et al.* [37]. In addition, the fully connected activations are treated as the extracted features. The multiple frames are converted into one 1024 dimensional feature. We trained the C3D with 5 convolution layers, 5 max pooling layers and 1 fully connected layer. The fully connected layer has 1024 outputs. The C3D network is trained with a batch size of 20. The initial learning rate is set to 0.01 and divided by 5 after every 10 epochs.

The preallocated training samples of RECOLA are used to construct the parameters. Each training instance is created as one clip with 16 consecutive frames. Each adjacent clip has 8 overlapping frames. Since the RECOLA database's label ratings were assessed by six assistants and each frame has one independent rating, the adjacent frames may have inconsistent ratings. Averaging all six ratings was used for some works [10], but this simple operation caused imprecise gold standards. Mencattini *et al.* [38] provided a gold-standard estimation method designed to mitigate the impact

TABLE 1. The manual features.

Cha.	Feature	Description
GSR	Number of peaks	Number of peaks in resistance exceeding 100 Ω
	Amplitude of peaks	The amplitude from the saddle point to the nearest peak.
	Rise time	The time from the saddle point to the nearest peak.
ECG	IBI	Mean IBI
	Multiscale entropy (MSE)	MSE at the 5th level
HR	Statistical moments	Mean and SD
HRV	Statistical moments	Mean and SD
SCR	Number of peaks	Number of peaks in the resistance exceeding 100 Ω
	Amplitude of peaks	The amplitude from the saddle point to the nearest peak.
	Rise time	The time from the saddle point to the nearest peak.
SCL	Statistical moments	Mean and SD

of subjectivity of ratings. We calculate the mean of their result as a label for one clip. It is worth noting that the end of the video uses the last 1 to 16 frames as the last clip.

2) PHYSIOLOGICAL FEATURE

The peripheral physiological signals derived from wearable noninvasive devices can be obtained readily. Therefore, all physiological signals of RECOLA, including the galvanic skin response (GSR), electrocardiogram (ECG), heart rate (HR), heart rate variability (HRV), skin conductance response (SCR) and skin conductance level (SCL), are utilized for feature extraction.

Corresponding to the time of the video clip, each physiological signal window size of RECOLA is 160. The DBN is constructed with hidden layers sized [160, 1000, 500,110] and fully connected layers sized [110,50,45]. The training process contains pretraining and fine-tuning. The training set is utilized for the two training stages, and the learning rate is set to 0.0001. Each physiological signal trains its individual DBN feature extraction model, and a trained DBN produces 110 features from the topmost hidden layer. Although the fine-tuning process does not change the parameter information of the hidden layer, it can be used as an obvious measure of the hidden layer.

The DBN features and manual features are both utilized in this paper. The manual features are listed in Table 1.

C. COMPATIBILITY MEASUREMENT

The primary focus of this subsection is to encode the multimodal data via the relationship between modalities and labels. We learn the maps from feature set s_i and corresponding label $z^{(i)}$ to a shared semantic space R^D. The compatibility is obtained between different modalities' features and the label set by applying the WSABIE algorithm [39] to extracted video feature $V^{(i)}$ and each physiological feature $g_l^{(i)}$, which computes $K+1$ similarities $f_q^{(i)}(1), f_q^{(i)}(2) \ldots, f_q^{(i)}(K+1)$ for label z_q, where z_q the qth class, and $q = 1, 2, ..M$. Each similarity between the qth label and the video feature of the ith sample is defined as:

$$f_q^{(i)}(\bullet) = W_q^T \bullet \Theta(V^{(i)}) \tag{6}$$

where W_q denotes the qth column of the label embedding matrix, and $\Theta(V^{(i)})$ represents the linear projection from the feature space to R^D. Similarly, the linear transformation of physiological features is $\Theta(g_k^{(i)})$, and we use $modality_l$ to represent the lth modality.

The preallocated training set is utilized to update the parameters online via stochastic gradient descent (SGD) while the trained parameters are locked in the testing process. Several sets of label embeddings and linear conversions are trained based on the modalities. This means that each modality corresponds to an independent semantic space. Therefore, we obtain compatibility indicators between different modalities and each label.

Since this is a supervised learning process, the emotional information hidden in the modal features is mined to determine the greatest similarity to the corresponding emotional label. Compared to complete effective data, invalid data with insufficient emotional information will likely be converted to outliers in the semantic space, producing a small similarity value with the corresponding label. There may be a case where although the feature data are extracted from the data with the noise or the missing portion, the similarity with the corresponding label is still large. This is because this low quality raw data has enough modal emotional information and hence has less impact on further processing. Besides, the online learning method of WSABIE is applicable to massive data sets. Unlike the traditional classification methods [40], [41], it obviously outperforms them in terms of time complexity and memory requirements.

D. DISCRIMINATIVE MODULE

Inspired by traditional imputation methods that deal with missing data, since the modalities' performance varies gradually, we assume that the modality of the previous moment has a certain similarity with the current one in a short period of time. From another perspective, since the emotional labels represent the semantic information related to the linear projection from feature, the label distribution is computed using each modality's compatibility measurement, which is equivalent to the indirect measurement of the difference between the modalities. Therefore, the discriminative module is designed

to utilize the combination for the semantic compatibility and temporal continuity of the modality to filter out the invalid features.

First, for each s^{te} in the test set, the objective function for discriminating true invalid data is defined as:

$$g(l,t) + b_l < 0 \tag{7}$$

where $g(l,t)$ denotes the comprehensive score of the lth modality at time t, and b_l is the discriminative threshold. We compute the sum of two factors as a comprehensive score:

$$g(l,t) = \rho_{l,1}\phi(l,t) + \rho_{l,2}\chi(l) \tag{8}$$

where $\rho_{l,1}$ and $\rho_{l,2}$ are the scaling parameters. The first term is the temporal similarity of the modality while the other term measures the semantic difference of each modality from the others. The former uses the cosine similarity in R^D between the linear mappings from modalities' feature at the previous time and current time:

$$\phi(l,t) = \frac{\Theta\left(modality_l^{t-1}\right)^T \Theta\left(modality_l^t\right)}{\left\|\Theta\left(modality_l^{t-1}\right)\right\| \left\|\Theta\left(modality_l^t\right)\right\|} \tag{9}$$

Following the works of [42], [43], we learn the semantic difference by using the maximum normed residual test theory, assuming that the set of labels has an approximately normal distribution. Each modality's semantic difference is measured by the $\chi(l)$ that is obtained based on a two-sided test by applying Eqs. (10)-(12)

$$\chi(l) = \gamma_l - Cr \tag{10}$$

$$\gamma_l = \frac{|q_l - \text{mean}(Q)|}{\text{std}(Q)} \tag{11}$$

$$Cr = \sqrt{\frac{K}{K+1}} * \sqrt{\frac{\tau^2}{K+1+\tau^2}} \tag{12}$$

where $Q = \left\{q_l = q | \max_q f_q(l), l = 1, 2..K+1\right\}$, and τ denotes the critical value of the t distribution, which has a significance level of $\alpha/(2K+2)$ and K-1 degrees of freedom.

Recall that the structure of Eq. (7) contains the fixed parameter $\phi(l,t)$ and the binary inputs $\chi(l)$ to determine the final discriminative results (True or False). This function can be treated as a linear classifier with the binary inputs as the feature of the classifier, $\rho_{l,1}$ and $\rho_{l,2}$ as the weights and b_l as the bias. This avoids manually tuning the parameters. We train the logistic regression classifier (LRC) and the obtained parameters are used for the discriminative process.

E. COMPENSATION AND DETECTION

The detected invalid modal feature is treated as a missing feature at this stage and is not involved in the classification calculation. Considering the correlation between multimodalities, we use the feature fusion method for classification. However, the fusion method needs to ensure that the input features have the same dimension. Therefore, imputation is utilized to compensate for the missing feature. For temporary invalid data, replacing the discarded feature with the nearest neighboring feature with emotional semantics can be the easiest and most effective method. For invalid features that are continuously produced for a long time, autoencoder-based methods [32], [44], [45] can be integrated into the MSD system to generate new features/views using the remaining features/views.

Due to the powerful approximations of the fully connected network (FCN), the fusion feature is input into the FCN to reconstruct the nonlinear relationship between the modalities for recognition. The sizes of the fully connected layers are [1698, 800, 400, 200].

We also use support vector machine (SVM), known as the best shallow classifier, as the multimodal classifier. The radial basis function (RBF) kernel projects the input features into a higher-dimension feature space and helps to process nonlinear separable samples by constructing the hyperplane. Therefore, the SVM-RBF can be effective for multimodal emotion classification.

IV. EXPERIMENT AND RESULTS

A. DATASETS AND SETTINGS

1) RECOLA DATASET

The RECOLA database has been widely utilized for multimodal emotion recognition and has been provided for The Audio/Visual Emotion Challenge and Workshop (AVEC) since 2015 [46]–[48]. The dataset was created using remote collaborative tasks and the audiovisual and physiological signals of the subjects' spontaneous and natural interactions were recoded. The training and development parts, each containing 9 subexperiments, were allocated in advance for AVEC 2018 by balancing the influence caused by the different ages, genders and native languages of the subjects. Moreover, there are six assistants giving the continuous ratings of the subjects' valence and arousal based on each frame derived from the five-minute recorded video. The total number of clip samples for our experiment is 8397 (for training)+8411 (for testing). Each instance contains 160 × 6 physiological data and 16 frames sized 112 × 112.

2) EXPERIMENTAL SETTINGS

First, the feature extraction and detection models were trained using a raw training set. A binary classification of two emotional dimensions, valence and arousal, was provided to detect positive and negative emotion. The experimental results of the raw data were verified on the test set and were compared with the results of the state-of-the-art approach. In the detection module, the FCN was trained using an initial supervised learning rate of 0.005, which was divided by 5 after every 10 epochs. The parameters of the SVM-RBF were set to C = 0.5 and $\gamma = 0.5$ and were optimized by grid search cross-validation.

In the supervised learning process of the compatibility measurement and discriminative module, the performance with too few classes relies heavily on the quality

of the database and the complexity of semantic embedding algorithms. More classes can reflect the cohesiveness of multimodal semantic embeddings. Therefore, the RECOLA's ratings of valence and arousal were divided into five intervals to the screen modalities with outlier semantics at this stage to prevent the discriminative module from taking too long. We added artificially generated invalid data to the raw training set as the subtraining set and the subtest set at this stage.

For the detection process with invalid data, the previously trained models at various stages were assembled to detect emotions. We presented the comparison results between the MSD system and the traditional system.

B. RESULTS

1) CLASSIFIER PERFORMANCE

Table 2 and Table 3 present the classification performance on the RECOLA database using the SVM-RBF and FCN classifiers. We provide the detection results for the video, the combination of peripheral physiological signals, and the combination of all modalities. They show a comprehensive evaluation measured by the accuracy, unweighted average recall (UAR) and F1 score. We can draw some conclusions from the observations as follows. The FCN usually achieves better performance than the SVM-RBF except for the results of the physiological signals. The results reported for different combinations show that the detection using physiological signals and video in the FCN outperforms the others (UAR of valence = 57.9% and UAR of arousal = 61.9%). The combination's performance of physiological signals is generally slightly below the results for the video. Therefore, the performance of the MSD system is superior using the visual-physiological multimodality. In addition, the comparison of the two tables shows that the performance for arousal is more prominent than the performance for valence.

TABLE 2. The detection performance(%) of the raw test set by different classifiers (Valence).

	Video	Physiological signals	Physiology +video
SVM-RBF(classifier)			
Accuracy	58.5	52.1	58.5
UAR	58.5	51.0	59.0
F1 score	58.0	51.5	58.4
FCN(classifier)			
Accuracy	58.4	54.1	60.0
UAR	59.0	50.1	**59.7**
F1 score	61.0	49.3	57.2

Regarding the comparison between different emotion classification systems on the same publicly available database, Table 4 shows the UAR (%) from our proposed system and from some retrievable state-of-the-art works. The work of Neumann et al [49] developed a system to detect emotion recognition using speech. The authors' best model achieved

TABLE 3. The detection performance(%) of the raw test set by different classifiers (Arousal).

	Video	Physiological signals	Physiology +video
SVM-RBF(classifier)			
Accuracy	64.7	54.3	63.5
UAR	59.0	51.3	59.6
F1 score	58.5	50.5	59.4
FCN(classifier)			
Accuracy	54.3	47.3	59.1
UAR	59.8	55.8	**61.9**
F1 score	50.5	48.4	51.6

TABLE 4. The UAR(%) comparison between the system using Recola database (Negative/Positive).

Method	Modality	Valence	Arousal
State-of-the-art approach			
Attentive convolutional neural network (ACNN)[49]	Audio	52.3	60.77
L2-regularized support vector classification[50]	Video	50.6	51.8
	Audio, Video	50.5	51.7
Our method			
C3D	Video	59.0	59.8
DBN+DBN	Physiology	50.1	55.8
(C3D+DBN)+FCN	Video, physiology	**59.7**	**61.9**

UAR (Arousal) = 60.77% and UAR (Valence) = 52.3% using RECOLA's samples. Kantharaju *et al.* [50] used facial action units (FAUs) and audiovisual signals to classify negative and positive emotions. Their emotion detection samples were filtered based on laughter episodes. We find that our proposed system based on video and physiology performs best on valence (59.7%) and arousal (61.9%). This indicates that the combination of video and peripheral physiological signals may be more effective for multimodal emotion recognition than the audiovisual method.

2) DISCRIMINATIVE MODULE

The temporarily invalid data were randomly produced to train the discriminative module and test the MSD system performance. Face detection is often affected by occlusion and other factors, causing background images that do not contain faces to be captured. In order to ensure the comparability and reproducibility of the experiment, we used the following method based on a public database to simulate invalid data in the IoT:

(1) The invalid data of video was created by capturing partial images from raw frames randomly; For physiological invalid data, non-stationary noise was added into the raw signals by the form of multiplication and addition. In detail, a random number (1-10) conforming to the Gaussian distribution was generated, and then the number was added or

multiplied with the original number. The obtained signal was compressed into an interval [0,1].

(2) Invalid data was selected by k-Nearest Neighbors (kNN). The extracted feature of data was the input of kNN. The invalid data closest to the opposite class of the original data were chosen to replace these original data, as shown in Figure 2. The value of k is set to 100.

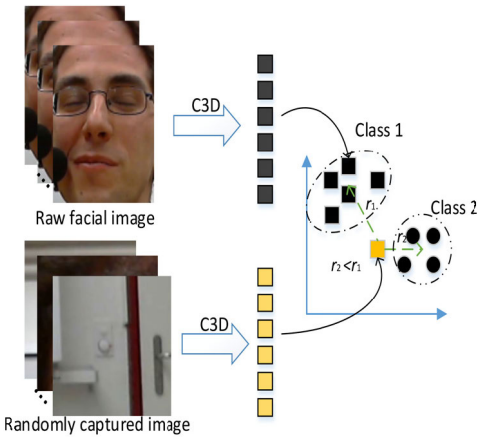


FIGURE 2. The process of generating invalid visual data. The black dots represent the raw features and the yellow dots represent the feature extracted from the generated invalid data.

The t-Distributed Stochastic Neighbor Embedding (tSNE), with the perplexity value of 30, was utilized to visualize an example of clustering results of features extracted from both raw facial image and invalid image. These image's emotional label was negative valence. As shown in Figure 3, a map was constructed in which the distribution difference between the classes of features extracted from invalid data and raw data is obvious. Invalid data were generated by separate modalities. Each modal data containing invalid data and original data were used to compute $\phi(l,t)$ and $\chi(l)$ to train the LRC. The total number of training samples and test samples was 141288.

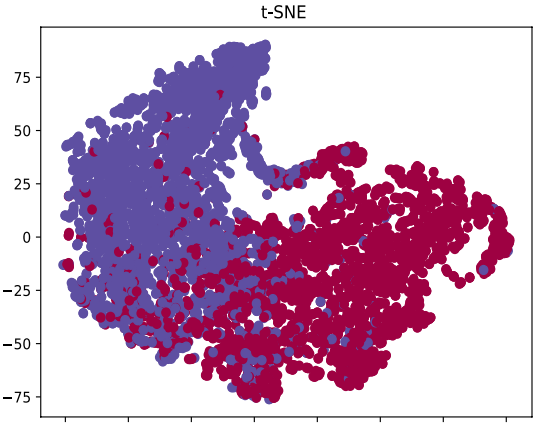


FIGURE 3. Visualization of high-dimensional features using tSNE.

The confusion matrix of the discriminative module is displayed in Figure 4. Regarding the procedure for identifying invalid data, the results of the recall (valence: 85.5%, arousal: 83.9%) show that the method is effective at detecting invalid data. Interestingly, the number of FNs (false negatives), which represent the number of times that data are incorrectly identified as invalid data, are slightly above the number of TNs (true negatives). This may be caused by the subjectivity of the emotional ratings' evaluation.

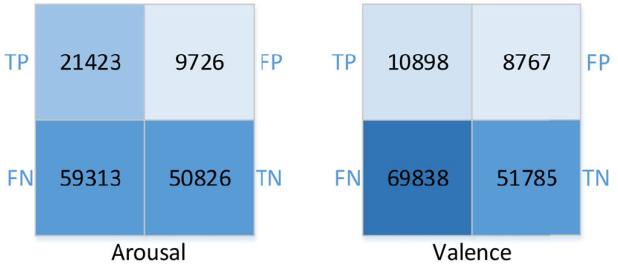


FIGURE 4. The confusion matrix of discriminative module. Left: Arousal. Right: Valence.

3) MULTIMODAL EMOTION DETECTION WITH INVALIDDATA

The proportion of unimodal or two modal invalid data in the test set varied from 20% to 60%. The two experiments, using the discriminative module (DM) or not using this module (NDM), were repeated 10 times for each proportion group.

As shown in Figure 5, the recall results of invalid data, averaged over 10 independent runs, are generally above 80%. It is seen that the invalid data detection of valence universally outperforms arousal. The arousal results for invalid data contained in ECG or HR data are not ideal.

The emotion detection results obtained on the invalid data are shown in Figure 6. The evaluation is measured by the average UAR. It is shown that the performance of multimodal emotion detection with invalid data using the discriminative module generally has better performance than that using the NDM.

The left part of Figure 6 shows the comparison between different modalities. The detection performance with invalid data present in physiological signals seems to be stable, which varies slightly for the NDM or DM. It is proven that the MSD system did not significantly improve this case since the video modality can well represent the emotional information and the invalid data present in physiological unimodalities such as EDA does not have much impact on the detection performance. Comparing the two conditions (DM and NDM), the performance of detection with 60% invalid data present in video is the most improved in the DM. Comparing valence and arousal, the arousal results in the DM reported for the proportion groups of the physiological modalities are slightly better than those in the NDM. However, in the valence results of physiological signals, the DM does not improve the emotion detection performance.

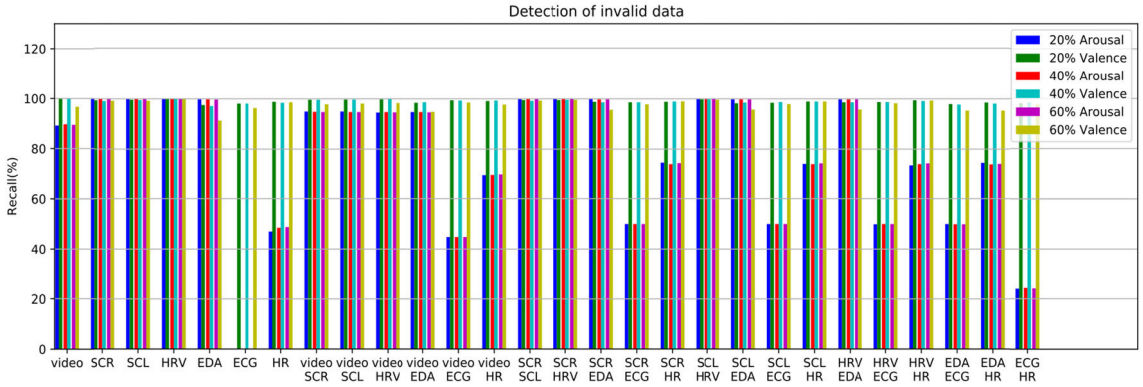


FIGURE 5. The performance (recall %) of the discriminative module in processing multimodal emotion detection.

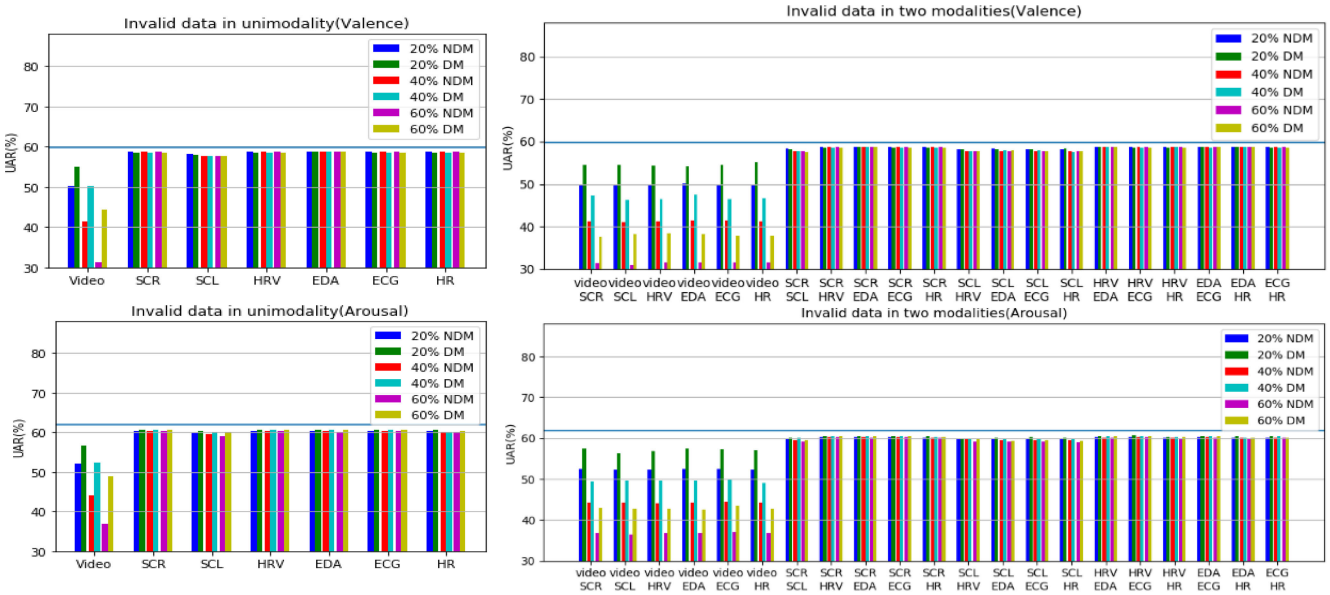


FIGURE 6. The UAR (%) of multimodal emotion detection with invalid data. The blue horizontal lines represent the UAR results of the raw test set. Upper Left: Invalid data only exist in one unimodality, corresponding to the abscissa (valence). Upper Right: Invalid data exist in two modalities (valence). Bottom Left: Invalid data only exist in one modality (arousal). Bottom Right: Invalid data exist in two modalities (arousal).

The right part of Figure 6 shows the multimodal emotion detection performance of the two modalities containing the invalid data simultaneously. We can observe that the performance of the two modalities is slightly below the performance of one modality. The UAR of the video condition in the DM significantly surpasses that in the corresponding NDM. The above descriptions show the effectiveness of the MSD system, especially in the case of invalid video data.

4) RUN TIME IN FIELD STUDY

The computer running the MSD contained a 3.6 GHz, Intel(R) Core(TM) i7-4790U, 8 GB of RAM, and an 8 GB NVIDIA GeForce GTX 1080; and the program platform used was Spyder. The programming language is Python 2.7, the background development framework was Pytorch. The feature extraction of physiological signals and video was processed in parallel.

TABLE 5. The time performance (ms) of multimodal emotion detection.

	NDM	DM
	SVM-RBF (classifier)	
Mean	24.69	39.96
SD	7.15	7.09
	FCN (classifier)	
Mean	204.87	220.93
SD	17.48	17.50

Table 5 shows the time performance of the proposed system, revealing the practicality of the MSD. It was measured

TABLE 6. Comparison of the recent multimodal emotion detection research.

	Dataset	Annotators	Modality	Assessment indicators (Valence/ Arousal)	Task
Nguyen et al.[51]	eNterface	Self-assessment	Audio, visual	Average accuracy(90.85)	Classification
Zhang et al.[52]	LIRIS-ACCEDE	3 annotators using the pairwise comparisons for each video excerpt	Audio, visual	CCC(91.8/94.6)	Regression
Yi et al.[53]	LIRIS-ACCEDE		Audio, visual	Accuracy(46.22/57.40) UAR(45.63/38.20)	Classification
Av+ec 2018 baseline[48] MT-Lasso	RECOLA	6 annotators for each frame	Audio, visual, ECG, EDA, HR, HRV,SCR, SCL	CCC(57.0/58.5)	Regression
Av+ec 2018 baseline[48] Lasso	RECOLA		Audio, visual, ECG, EDA, HR, HRV,SCR, SCL	CCC(77.5/49.2)	Regression
Neumann et al.[49]	RECOLA		Audio	UAR(52.3/60.77)	Classification
Kantharaju et al.[50]	RECOLA		Visual	UAR(50.6/51.8)	Classification
Presented work	RECOLA		Visual	UAR(59.0/59.8)	Classification
Kantharaju et al.[50]	RECOLA		Audio, visual	UAR(50.5/51.7)	Classification
Presented work	RECOLA		Visual, ECG, EDA, HR, HRV,SCR, SCL	UAR(59.7/61.9)	Classification

by calculating the time to recognize emotions once and repeated 100 times. The MSD with the DM has a larger time cost than that with the NDM using either the SVM-RBF or FCN. This is primarily caused by the calculation cost of the compatibility measurement and discrimination. From the results, we can conclude that the SVM-RBF achieves the best time performance. However, in many real-time applications, the running time of the MSD with the DM using the FCN may be acceptable due to a minor delay (<0.25 s). It is noteworthy that the differences in the configuration of the operating platform could bring about an order of magnitude difference. In addition, the time consumption of face recognition and communication needs to be considered in practice.

V. DISCUSSION

In the related work, it can be seen that deep neural network are main frame of most state-of-the-art multimodal emotion recognition systems. They are applied for learning of (i) feature representations, and (ii) joint classification for multimodality. Considering the latest trend towards real-time emotion detection for spontaneous affective displays using IoT big data streams, deep learning based on spatiotemporal feature extraction is usually suitable for addressing such system challenges. In addition, most studies only consider audiovisual signals for multi-modal emotion recognition. However, in real life, audio signals need to be generated in a continuous conversation scene, which limits the implementation of audiovisual multimodal emotion recognition. Based on the daily acquisition equipment (such as wristbands) of peripheral physiological signals and the physiological mechanism process of emotions, this paper utilized the combination of video and peripheral physiological signals as the input signal of multi-modal fusion. In particular, our results show that the multimodal emotion recognition performance based on physiology and video is better than the recent audiovisual results using the same database.

As shown in Table 6, many studies on multi-modal emotion detection based on different databases have acquired obviously different results. This can, in part, be attributed to the different subjective measurement methods of emotive responses, due to inherent ambiguity of the response. In addition, a lot of studies on public affective databases have made great efforts to mitigate the impact of subjective measurement. For example, RECOLA hired 6 assistants to annotate each frame, and our proposed system can achieve UAR (Arousal) $=$ 61.9% and UAR (Valence) $=$ 59.7%. LIRIS-ACCEDE, by which Yi *et al.* [53] achieved UAR $=$ 45.63/38.20(valence/arousal) and Zhang *et al.* [52] achieved CCC $=$ 91.8/94.6 (valence/arousal), takes pairwise comparisons, rather than rating approaches. From each pair of video excerpts, three annotators have to identify the one which can convey most strongly the given emotion in terms of valence or arousal. Based on the eNterface database, Nguyen *et al.* [51] deliver the best performance of 90.85 (at the average accuracy) by implementing the idea of data cleaning. It can be seen that the database obtained through data cleaning can achieve the seemingly best performance. Nevertheless, this method can only recognize the emotional states of obvious displays, and it requires a lot of manpower for manual screening. In addition to assessment methods, different collection methods and potential data cleaning approaches are also one of the important factors that affect the recognition results. Because data aggregation involved in the learning process determines the form of the sample distribution, which can accurately represent the implicit population and is the basic guarantee for the model learning performance.

The experimental results with invalid data show that the poor quality of video data results in a significant reduction in the recognition performance, while the existence of invalid data in the physiological signals delivers less impact. These findings are attributable to the fact that the annotators' working on the RECOLA database is based on the emotional displays in the videos, leading to a strong correlation between the videos and the labels. But the correlation with the physiological signals has not been further measured. Due to this correlation, the multimodal emotion detection results with invalid data existing in the videos are obviously improved after DM processing. A conclusion can be drawn that the effectiveness of the MSD with the DM also depends on construction methods of the database.

VI. CONCLUSION AND FUTURE WORK

A multistep deep system to reliably detect multimodal emotion using collecting records containing invalid data is proposed. The proposed system includes a feature extraction and emotion detection method using peripheral physiological signals and video modalities via deep neural networks. The invalid data are filtered out in the discriminative module using the semantic compatibility and continuity. The experiments are conducted using a public database containing different proportions of invalid data. The results verify the effectiveness of the discriminative module. Besides, the performance of the MSD is compared with the state-of-the-art approach in two conditions (the records contain invalid data and do not contain invalid data), and the proposed system based on peripheral physiological signals and video significantly improves the detection performance. The promising results imply that the proposed system can be deployed in many IoT scenarios, even without the complex network structure and brilliant data acquisition facilities. This work can be extended further in the following ways:

1) In practical applications, due to the emergence of large-scale data on people, advanced face recognition technology can be utilized. For example, the segmentation of face images in complex environments such as the wild can be input into the MSD to achieve the emotion detection of large-scale crowds. Besides, wearable physiological signal devices and cameras can be used in limited application scenarios, such as hospitals and nursing homes. Our research framework can receive these preprocessed facial images and acquired physiological signals to further detect emotions.

2) Advanced view learning algorithms can be integrated into the MSD to generate new features as a substitute for the discarded features. In this way, emotion detection with long-term invalid features can be achieved. In addition, a suitable algorithm can further improve the recognition performance of the MSD.

REFERENCES

[1] M. Shamim Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2399–2406, Aug. 2018.

[2] K. Lin, F. Xia, C. Li, D. Wang, and I. Humar, "Emotion-aware system design for the battlefield environment," *Inf. Fusion*, vol. 47, pp. 102–110, May 2019.

[3] L. Y. Mano, B. S. Faiçal, L. H. V. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, G. P. R. Filho, G. T. Giancristofaro, G. Pessin, B. Krishnamachari, and J. Ueyama, "Exploiting IoT technologies for enhancing health smart homes through patient identification and emotion recognition," *Comput. Commun.*, vols. 89–90, pp. 178–190, Sep. 2016.

[4] I. Azimi, T. Pahikkala, A. M. Rahmani, H. Niela-Vilén, A. Axelin, and P. Liljeberg, "Missing data resilient decision-making for healthcare IoT through personalization: A case study on maternal health," *Future Gener. Comput. Syst.*, vol. 96, pp. 297–308, Jul. 2019.

[5] P. Gong, H. T. Ma, and Y. Wang, "Emotion recognition based on the multiple physiological signals," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Jun. 2016, pp. 140–143.

[6] M. Singh, K. Yadav, A. Kumar, H. J. Madhu, and T. Mukherjee, "Method and device for non-invasive monitoring of physiological parameters," Patent Appl. 14/886 165, Apr. 20, 2017.

[7] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2121–2129.

[8] O. Krestinskaya and A. P. James, "Facial emotion recognition using min-max similarity classifier," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2017, pp. 752–758.

[9] Sujata, M. Trivedi, and S. K. Mitra, "A modular approach for facial expression recognition using euler principal component analysis (e-PCA)," in *Proc. IEEE Appl. Signal Process. Conf. (ASPCON)*, Dec. 2018, pp. 204–208.

[10] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[11] T. Hui and R. Sherratt, "Coverage of emotion recognition for common wearable biosensors," *Biosensors*, vol. 8, no. 2, p. 30, Mar. 2018.

[12] M. Chen, "CP-robot: Cloud-assisted pillow robot for emotion sensing and interaction," in *Proc. Int. Conf. Ind. IoT Technol. Appl.* Cham, Switzerland: Springer, 2016, pp. 81–93.

[13] H. Kim, J. Ben-Othman, S. Cho, and L. Mokdad, "A framework for IoT-enabled virtual emotion detection in advanced smart cities," *IEEE Netw.*, vol. 33, no. 5, pp. 142–148, Sep. 2019.

[14] L. Zhou, D. Wu, Z. Dong, and X. Li, "When collaboration hugs intelligence: Content delivery over ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 91–95, Dec. 2017.

[15] Y. Hao, Y. Miao, L. Hu, M. S. Hossain, G. Muhammad, and S. U. Amin, "Smart-Edge-CoCaCo: AI-enabled smart edge with joint computation, caching, and communication in heterogeneous IoT," *IEEE Netw.*, vol. 33, no. 2, pp. 58–64, Mar. 2019.

[16] M. Golam Rabiul Alam, S. Fakhrul Abedin, S. Il Moon, A. Talukder, and C. Seon Hong, "Healthcare IoT-based affective state mining using a deep convolutional neural network," *IEEE Access*, vol. 7, pp. 75189–75202, 2019.

[17] J. Eriksson, N. L. Russo, and J. Marin, "Using the Internet of Things to support emotional health," *ICST Trans. Ambient Syst.*, vol. 5, no. 17, Mar. 2018, Art. no. 154372.

[18] M. Chen, J. Yang, X. Zhu, X. Wang, M. Liu, and J. Song, "Smart home 2.0: Innovative smart home system powered by botanical IoT and emotion detection," *Mobile Netw. Appl.*, vol. 22, no. 6, pp. 1159–1169, Dec. 2017.

[19] P. Ekman, "Darwin's contributions to our understanding of emotional expressions," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 364, no. 1535, pp. 3449–3451, Dec. 2009.

[20] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[21] L. J. Peter, "The emotion probe: Studies of motivation and attention," *Amer. Psychol.*, vol. 50, no. 5, pp. 372–385, 1995.

[22] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 43–55, Mar. 2015.

[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 8.

[24] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.

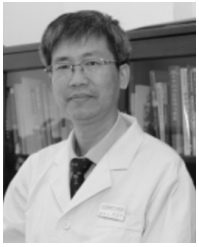
[25] Z. Zhang, J. Han, E. Coutinho, and B. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1289–1301, May 2019.
[26] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, Jul. 2019.
[27] A. Mencattini, F. Ringeval, B. Schuller, E. Martinelli, and C. Di Natale, "Continuous monitoring of emotions by a multimodal cooperative sensor system," *Procedia Eng.*, vol. 120, pp. 556–559, 2015.
[28] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, 2015, pp. 73–80.
[29] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 97–104.
[30] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Comput. Intell. Neurosci.*, vol. 2017, Sep. 2017, Art. no. 2107451.
[31] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 206–218, Oct. 2011.
[32] C. Du, C. Du, H. Wang, J. Li, W.-L. Zheng, B.-L. Lu, and H. He, "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 108–116.
[33] A. Kanawaday and A. Sane, "Machine learning for predictive maintenance of industrial machines using IoT sensor data," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 87–90.
[34] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.
[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
[36] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5060–5069.
[37] F. Yin, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 445–450.
[38] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. D. Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 314–327, Jul. 2017.
[39] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–7.
[40] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of Features," *J. Mach. Learn. Res.*, vol. 8, pp. 725–760, Apr. 2007.
[41] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2008, pp. 1–8.
[42] R. B. Jain, "A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data," *Clin. Biochem.*, vol. 43, no. 12, pp. 1030–1033, Aug. 2010.
[43] P.-T. Wilrich, "Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic," *AStA Adv. Stat. Anal.*, vol. 97, no. 1, pp. 1–10, 2013.
[44] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational neural networks," *Neural Comput.*, vol. 28, no. 2, pp. 257–285, Feb. 2016.
[45] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1405–1414.
[46] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audio/visual emotion challenge and workshop," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1335–1336.
[47] M. Valstar, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 3–10.
[48] F. Ringeval, "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proc. Audio/Visual Emotion Challenge Workshop*, 2018, pp. 3–13.
[49] M. Neumann and N. G. Thang Vu, "CRoss-lingual and multilingual speech emotion recognition on english and French," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5769–5773.
[50] R. B. Kantharaju, F. Ringeval, and L. Besacier, "Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals," in *Proc. Int. Conf. Multimodal Interact. (ICMI)*, 2018, pp. 220–228.
[51] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition," *Comput. Vis. Image Understand.*, vol. 174, pp. 33–42, Sep. 2018.
[52] G. Zhang, T. Luo, W. Pedrycz, M. A. El-Meligy, M. A. F. Sharaf, and Z. Li, "Outlier processing in multimodal emotion recognition," *IEEE Access*, vol. 8, pp. 55688–55701, 2020.
[53] Y. Yi and H. Wang, "Multi-modal learning for affective content analysis in movies," *Multimedia Tools Appl.*, vol. 78, 10, pp. 13331–13350, 2019.

MINJIA LI was born in Taiyuan, Shanxi, China, in 1992. She received the B.S. degree in information engineering from the Xi'an University of Posts and Telecommunications, in 2014 and the M.S. degree in electric information engineering from the City University of Hong Kong (Cityu). She is currently pursuing the Ph.D. degree in affective computing and pattern recognition with the University of Science and Technology Beijing, Beijing, China.

Her research interests include affective computing and pattern recognition.

LUN XIE received the M.S. and Ph.D. degrees in control theory and control engineering from the University of Science and Technology Beijing, Beijing, China, in 1998 and 2002, respectively.

He is currently a Professor and a Doctorate Supervisor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. His research interests include affective computing, robotics, the theory and application of networked control systems, and artificial intelligence. He is also a Board Member of the Chinese Association of Artificial Intelligence and the Vice Director of the Beijing Society of the Internet of Things.



ZEPING LV is currently a Chief Physician with the Rehabilitation Hospital, National Rehabilitation Auxiliary Center. His research interests include multimodal cognition and neurological rehabilitation. In the past five years, he has participated in the 683 Project of the Ministry of Science and Technology and research on the National Science and Technology Support Program. He has undertaken and completed eight provincial and ministerial-level scientific and technological research projects and natural science fund research. He won third prize of the provincial and ministerial level scientific and technological progress three times. He is also the Vice President of the China Rehabilitation Technology Transformation and Development Promotion Association, the Director of the China Rehabilitation Medicine Association, and the Standing Committee of the Cerebrovascular Disease Rehabilitation Branch.

JUAN LI received the Ph.D. degree in cognitive science from the University of Chinese Academy of Sciences, Beijing, China, in 2000.

She is currently a Professor with the Institute of Psychology, Chinese Academy of Sciences (Level 3). Her research interests include cognitive science and elderly emotion. She is also a member of the Cognitive Neuroscience Society (CNS). She has hosted and participated in more than ten projects, including the National Natural Science Foundation of China, the Science and Technology Support Project of the Ministry of Science and Technology, the Knowledge Innovation Project of the Chinese Academy of Sciences, and foreign funded projects (including those funded by the U.S. NIH, the U.K. Wellcome Trust, and the Swedish Council for Social Research).

ZHILIANG WANG received the M.S. degree in control theory and control engineering from Yanshan University, Qinhuangdao, China, in 1985, and the Ph.D. degree in control theory and control engineering from the Harbin Institute of Technology, Harbin, China, in 1989.

He is currently a Professor and a Doctorate Supervisor with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. His research interests include affective computing, robotics, the theory and application of networked control systems, and artificial intelligence. He is also a Senior Board Member of the Chinese Association of Artificial Intelligence and the Director of the Beijing Society of the Internet of Things.