Sensors and Materials, Vol. 32, No. 7 (2020) 2329–2341
MYU Tokyo

S & M 2262

Classification of Restlessness Level by Deep Learning of Visual Geometry Group Convolution Neural Network with Acoustic Speech and Visual Face Sensor Data for Smart Care Applications

Ing-Jr Ding* and Nai-Wei Zheng

Department of Electrical Engineering, National Formosa University, Huwei Township, Yunlin 632, Taiwan, ROC

(Received June 27, 2019; accepted June 1, 2020)

Keywords: restlessness classification, VGG-16 CNN, VGG-19 CNN, acoustic speech, visual face

Recently, acoustic speech recognition and visual face identification have become mature techniques widely used in real-life applications. However, human cognitive recognition issues such as human emotion classification are still a major challenge. In this study, restlessness level recognition using a deep learning scheme of the Visual Geometry Group (VGG) convolution neural network (CNN) with input acoustic speech and visual face sensor data is presented for home care applications. The well-known Microsoft Kinect device is employed with a red–green–blue sensor and an array of microphones to acquire facial expression and vocal variation data, respectively. Both VGG-16 and VGG-19 CNN deep learning models are used to evaluate the effectiveness of restlessness level classification in three different data modality inputs: acoustic speech observations alone, visual face observations alone, and combined speech and face observations. Experimental results on categorizing nine defined restlessness levels demonstrate the effectiveness of the presented approach. A specific group with problems of restlessness can benefit from the immediate care that can be provided intelligently by using the system proposed in this study.

1. Introduction

The efficient utilization of sensors will bring much intelligent assistance into our everyday lives. Visual-based camera sensors have been effectively used to capture facial data for face recognition.[1,2] Acoustics-based speech recognition with a microphone sensor has also been seen in many applications such as the voice-command control of devices.[3–5] In recent years, 3D image sensors [also called RGBD sensors when integrated with a red–green–blue (RGB) camera sensor] have been adopted to acquire the depth information of skeletons or hand shapes during human actions.[6,7] Although pattern recognition techniques with the support of sensors have led to the realization of many convenient applications, cognitive behavior recognition such as human emotion recognition is still rarely carried out in real-life applications. To promote the use of sensor devices that can obtain biometric feature data from people to recognize their

*Corresponding author: e-mail: eugen.ding@gmail.com
https://doi.org/10.18494/SAM.2020.2881

ISSN 0914-4935

emotions, a deep learning strategy of using a Visual Geometry Group (VGG) convolution neural network (CNN) with acoustic speech and visual face sensor data for classifying the restlessness level is presented in this work. By using this emotion recognition system, occurrences of unexpected or dangerous events will be significantly decreased, and people experiencing restlessness will be able to receive immediate care in home, office, laboratory, and other indoor environments.

Early studies on human emotion recognition explored its feasibility and effectiveness, and most were aimed at the use of only one modality of data (acoustic or visual signals) for developing an emotion recognition system.[8–11] In Refs. 8 and 9, a 3D face model with visual 3D space location data was employed to determine facial expressions. The development of speech emotion recognition systems using acoustic information derived from affective speech data and semantic labels was reported in Refs. 10 and 11. Recently, model-based classification techniques with a certain degree of learning have been explored for analyzing human biometric features to realize deep-layered human behavior cognition tasks including human emotion recognition through affective computing.[12–15] Although such emotion recognition approaches using constructed affective or attention models have been reported, deep-learning-based strategies using the well-known CNN or recurrent neural network (RNN) models for the deep feature extraction of input sensor data to classify human emotions have rarely been explored. Most works related to CNN- or RNN-based deep learning models achieved intuitive recognition without affection or cognition computing, such as vehicle type classification,[16] medical image segmentation,[17] speech recognition,[18] and speech reconstruction.[19]

Different from the above studies on emotion recognition and deep-learning-model-based applications, here, we employ the well-known Microsoft Kinect device equipped with an RGB sensor and an array of microphones to sense and acquire facial expressions and vocal variation data, respectively, to perform emotion recognition. Two different popular VGG-CNN deep learning models, VGG-16 and VGG-19 CNNs, are used to establish models for classifying restlessness levels using three different sensing data modality inputs: acoustic speech observations alone, visual face observations alone, and combined speech and face observations. Compared with the above related studies, the main advantages and contributions of the presented approach are as follows.

(1) In contrast to conventional emotion recognition applications, the presented approach to emotion recognition focuses on the classification of restlessness levels. A categorization design considering nine different degrees of restlessness, each of which corresponds to specific acoustic voice and visual face variations, is proposed.
(2) We effectively extend CNN deep learning model applications to the area of continuous-time emotion recognition using three different types of input sensor data modality: acoustics-based vocal data, vision-based facial data, and combined vocal and facial sensor data.
(3) The emotion recognition performance is compared between the popular VGG-16 and VGG-19 CNN models using three different input sensor data modalities.
(4) By focusing specifically on the classification of the restlessness level, the group with restlessness problems can receive immediate smart care.

2. Restlessness Classification Systems Using VGG-16 and VGG-19 CNN Models with Acoustic Speech and Visual Face Sensing Data

In our work, the Microsoft Kinect sensor device is employed as the main sensor to acquire visual and acoustic sensing data.[20] As depicted in Fig. 1, the Microsoft Kinect sensor device contains a camera image sensor and a four-microphone array, which can be effectively used to obtain visual face expression and acoustic speech raw observation data, respectively.

Note that a fast Fourier transform (FFT) is performed on part of the acoustic speech raw observation data to convert it from time-based data to frequency-based data in the form of a standard image before inputting the acoustic sensor data into a deep learning model.

2.1 Restlessness level designs using Microsoft Kinect sensor data of vocal speech and facial expression observations

Generally, the specific group with the problem of restlessness frequently exhibits the so-called "restlessness behavior." Restlessness behaviors exhibited by the specific group can be primarily divided into two different abnormal behavior categorizations, "restlessness with physical violence or assault" and "restlessness with verbal violence or assault."

This study is focused on the categorization of restlessness with verbal violence or assault. The restlessness level is first defined and designed to include restlessness events that may be exhibited in real life. Table 1 shows the presented restlessness levels. As shown in Table 1, there are a total of nine restlessness levels defined in this work (Level-1, Level-2, ..., Level-9), each of which indicates a different degree of restlessness exhibited by the specific group. For a person with restlessness problems, the Mandarin utterances 'Hey,' 'Ohh,' and 'Ahh' significantly indicate panic and flurry. Therefore, these three types of restlessness speech are mainly considered in the design of the restlessness levels in this work. In fact, restlessness with verbal violence or assault apparently contains vocal variations and facial expressions of the specific group. As can be seen in Table 1, in each defined restlessness level, acoustic vocal variations and visual facial expressions are taken into account simultaneously. For example, the

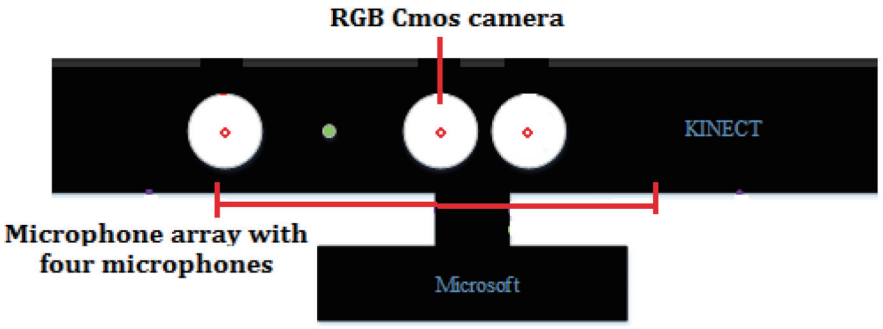


Fig. 1. (Color online) System using Microsoft Kinect with both a CMOS camera sensor and an array of four microphones to acquire facial expression and acoustic speech sensing data simultaneously to classify the restlessness level.

Table 1
Nine levels of restlessness (from slight to great) with associated vocal variations and facial expressions.

Index	Emotion	
	Definitions of nine levels of restlessness (from slight to great)	
	Acoustic vocal variations	Visual facial expressions
Level-1	Uttering 'Hey' with tight lips (low volume)	Eyebrows sagging slightly
Level-2	Uttering 'Ohh' with tight lips (low volume)	Slightly gathered eyebrows and sharp eyes
Level-3	Uttering 'Ahh' with tight lips (low volume)	Gathered eyebrows and sharp and staring eyes
Level-4	Uttering 'Hey' with tight lips (moderate volume)	Eyebrows sagging slightly
Level-5	Uttering 'Ohh' with tight lips (moderate volume)	Slightly gathered eyebrows and sharp eyes
Level-6	Uttering 'Ahh' with tight lips (moderate volume)	Gathered eyebrows and sharp and staring eyes
Level-7	Uttering 'Hey' with tight lips (high volume)	Eyebrows sagging slightly
Level-8	Uttering 'Ohh' with tight lips (high volume)	Slightly gathered eyebrows and sharp eyes
Level-9	Uttering 'Ahh' with tight lips (high volume)	Gathered eyebrows and sharp and staring eyes

Level-1 categorization denotes the smallest degree of restlessness behavior and mainly includes the vocalization of 'Hey' with tight lips at a low volume and the facial expression of eyebrows sagging slightly, whereas the greatest degree of restlessness is the Level-9 categorization, where both the acoustic speech utterance of 'Ahh' with tight lips at a high volume and visual facial expressions of gathered eyebrows and sharp and staring eyes are involved. The other seven categorization levels of restlessness, Level-2, Level-3, Level-4, Level-5, Level-6, Level-7, and Level-8, represent different degrees of both visual facial expressions and acoustic speech of the subject. As shown in Table 1, the restlessness categories are defined by three different acoustic vocal utterances at three different volumes along with visual facial expressions with three different eyebrow motions and three different eye expressions. Note that the above-mentioned restlessness categories are in accordance with the social psychology of emotion[21] and general emotion behavior cognition in common daily life.

Table 2 shows the original input PCM raw data of acoustic speech of the three different Mandarin vocalizations 'Hey', 'Ohh', and 'Ahh' uttered at different volumes for each of the nine defined levels of restlessness. Both the acoustic speech spectrum and RGB image sequences of visual facial expressions for each of the nine defined levels of restlessness are provided and listed in Table 3.

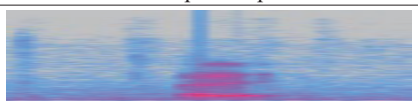
2.2 Restlessness classification by CNN deep learning models of VGGNet

The deep learning model is adopted in this work to classify restlessness. The well-known deep learning calculation approach, VGGNet CNN,[22] also known as VGG-CNN, is employed for emotion recognition computation in this study. The VGGNet-series CNN can perform satisfactory classification by adjusting the depth of the feature learning model. The typical CNN model contains both convolution and maximum pooling calculations to perform deep-learning-based feature extraction. The fully connected (FC) layered neural network is appended to the CNN model for classification computations of such deep learning feature parameters. In the VGG-CNN model configurations, six model levels, types A, A-LRN, B, C, D, and E, are involved, each of which contains different numbers of convolution and pooling calculations. In

Table 2
(Color online) PCM raw data of acoustic speech for the nine defined levels of restlessness.

Level-1	Level-2	Level-3	Level-4	Level-5	Level-6	Level-7	Level-8	Level-9
								

Table 3
(Color online) Acoustic speech spectrum and RGB image sequences of visual facial expressions for the nine defined levels of restlessness.

Index	Emotion	
	Expressions of nine different levels of restlessness (from slight to great)	
	Acoustic speech spectrum	Visual facial expressions
Level-1		
Level-2		
Level-3		
Level-4		
Level-5		
Level-6		
Level-7		
Level-8		
Level-9		

this study, VGG-CNN models D (i.e., VGG CNN-16) and E (i.e., VGG CNN-19) are employed. VGG CNN-16, which has 13 convolution layers (with a compound of five pooling layers) and three FC layers, and VGG CNN-19, which has 16 convolution layers (with a compound of five pooling layers) and three FC layers, are used to carry out restlessness level classifications in accordance with the obtained modalities of sensor data inputs, acoustic speech spectrum RGB images (see Fig. 2), and visual facial expression RGB images (see Fig. 3).

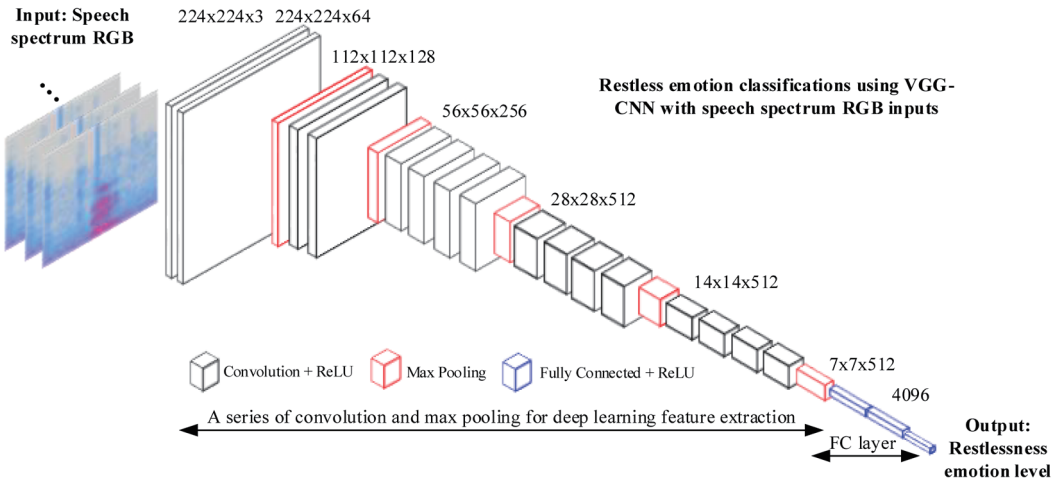


Fig. 2. (Color online) VGG-CNN deep learning model with input of a series of speech spectrum RGB image sequences for classification of nine different restlessness levels.

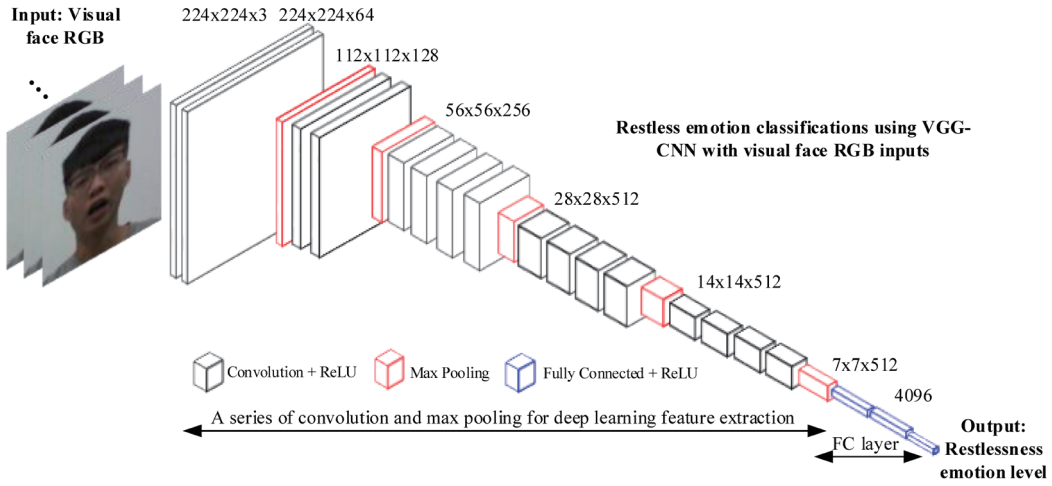


Fig. 3. (Color online) VGG-CNN deep learning model with input of a series of visual face RGB image sequences for classification of nine different restlessness levels.

As can be seen in Figs. 2 and 3, a series of sensor data inputs of the RGB image with the size of 224 by 224 can finally be classified as one of the nine different restlessness levels. Note that for the FC configuration with three processing layers, the final layer in the FC contains nine different neural nodes, each of which specifically denotes one of the defined nine restlessness levels.

2.3 Design consideration to combine acoustic speech and visual face sensing data for CNN restlessness emotion classifications

Figure 4 depicts VGG-CNN restlessness level classifications by combining the sensor modalities of the acoustic speech spectrum and visual facial expression RGB images for emotion

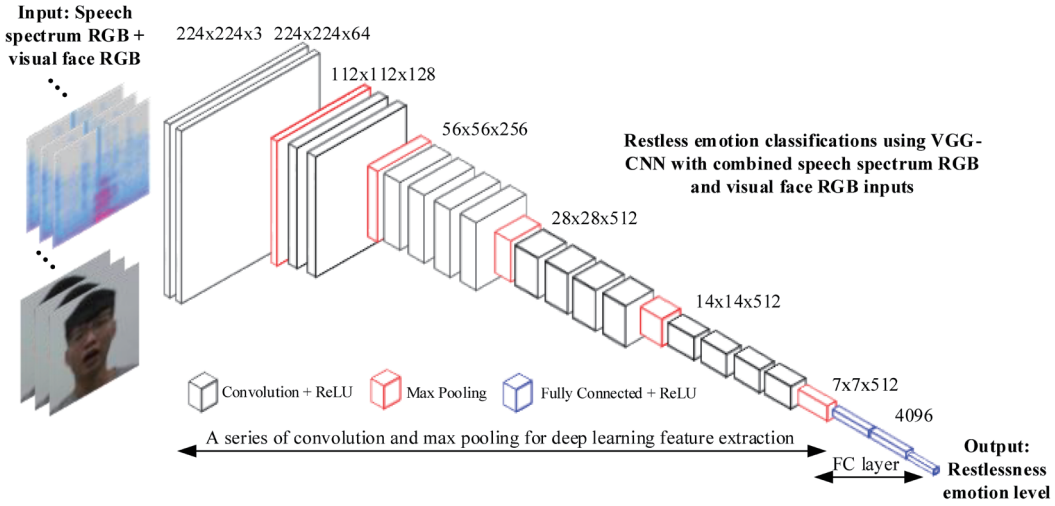


Fig. 4. (Color online) VGG-CNN deep learning model with input of a series of combined speech spectrum and visual face RGB image sequences for classification of nine different restlessness levels.

recognition performance evaluations of using speech spectrum RGB data alone or facial expression RGB data alone. As can be seen in Fig. 4, when sensor data are combined, the two different modalities of sensor data can cause a certain degree of data distribution variety in the inputs of the VGG-CNN model. It is worth noting that, compared with VGG-CNN restlessness emotion level classifications using single modality sensor data of the speech spectrum or facial expression, such data variety between the acoustic speech spectrum and visual facial expression RGB images will significantly increase or decrease the emotion classification accuracy.

In fact, for CNN deep-learning-model-based recognition applications using multiple modalities of sensor data, a general strategy is to adopt a single channel with a series of convolution and pooling calculations to extract the deep learning feature parameters by the specific single modality of sensor data. Multiple modalities of deep learning feature parameter data, each of which is derived from the original input image data specifically using convolution and pooling calculations identical to those of the CNN process, can then be fused using the FC layer scheme of the CNN process or an additional pattern recognition classifier (e.g., the typical support vector machine classifier).

3. Experimental Designs and Results

In this work, experiments on restlessness emotion recognition by CNN deep learning with inputs of acoustic speech, visual face, or combined speech and face sensor data are carried out in a laboratory environment. A total of four subjects were recruited to collect the acoustic and visual sensor data to establish the required database for recognition accuracy evaluations of the classification of nine defined restlessness levels. As mentioned previously, the widely used Kinect sensor is employed as the sensing data collector to properly acquire the required speech and face raw data of one of the four subjects in this study. The sensor parameter specifications of the Kinect sensor device are a frame rate of 30 frames per second (fps) of the RGB camera to

capture the continuous-time face variation images, and in acoustic speech data acquisitions, a sampling rate of 44.1K (i.e., 44100 samples, each of which is represented by 16 bits, acquired in one second) of the microphone array inside the Kinect senor.

In the establishment of the sensor data database, each of the four subjects is requested to perform the actions of the nine defined levels of restlessness. Each of the collected action sensor data includes both acoustic speech observations with vocal and visual observations of facial expressions. For each specified restlessness level, the subject is requested to perform 50 actions, one half (i.e., 25 actions) of which is chosen as the training data of the VGG-16 or VGG-19 CNN deep learning model, and the other half is used as test data for recognition accuracy evaluations of the constructed CNN restlessness classification models. A total of 1800 acoustic and visual actions for the nine variations of restlessness conditions are observed in this work, 450 observations (50 observations included in one categorization of 9 restlessness levels) for each of the four subjects.

Tables 4–7 show recognition performance characteristics of restlessness level classifications using VGG-CNN (VGG-16 or VGG-19 CNN) deep learning models with input data sensor modalities of "face RGB", "speech spectrum RGB", and "combined face and speech spectrum RGB" of subject-1, subject-2, subject-3, and subject-4, respectively. Table 8 shows the average classification performance of restlessness level recognition for these four subjects. As seen in Tables 4–7, for each of the four subjects, both training and validation rates of VGG-16 and VGG-

Table 4
Recognition accuracy for VGG-16 and VGG-19 CNN models with different modalities of sensor data (subject-1).

Sensor data	CNN model					
	VGG-16 CNN model			VGG-19 CNN model		
	Training	Validation	Test	Training	Validation	Test
Face RGB	100	100	60.74	100	99.94	59.68
Speech spectrum RGB	100	100	92	100	100	91.11
Face RGB + Speech spectrum RGB	100	100	74.87	100	100	79.17

(Unit: %)

Table 5
Recognition accuracy for VGG-16 and VGG-19 CNN models with different modalities of sensor data (subject-2).

Sensor data	CNN model					
	VGG-16 CNN model			VGG-19 CNN model		
	Training	Validation	Test	Training	Validation	Test
Face RGB	100	100	66.23	100	100	55.81
Speech spectrum RGB	100	100	77.76	100	100	75.10
Face RGB + Speech spectrum RGB	95.37	95.24	58.95	95.45	95.82	66.85

(Unit: %)

Table 6
Recognition accuracy for VGG-16 and VGG-19 CNN models with different modalities of sensor data (subject-3).

Sensor data	CNN model					
	VGG-16 CNN model			VGG-19 CNN model		
	Training	Validation	Test	Training	Validation	Test
Face RGB	100	100	48.92	100	100	42.28
Speech spectrum RGB	99.78	100	84.44	99.60	99.06	83.56
Face RGB + Speech spectrum RGB	99.88	99.97	63.91	99.83	100	65.44

(Unit: %)

Table 7
Recognition accuracy for VGG-16 and VGG-19 CNN models with different modalities of sensor data (subject-4).

Sensor data	CNN model					
	VGG-16 CNN model			VGG-19 CNN model		
	Training	Validation	Test	Training	Validation	Test
Face RGB	100	100	33.17	100	100	43.94
Speech spectrum RGB	100	100	94.67	100	100	92.89
Face RGB + Speech spectrum RGB	100	100	69.22	100	100	73.24

(Unit: %)

Table 8
Average recognition accuracy for VGG-16 and VGG-19 CNN models with three different input sensor data modalities (average of four subjects).

Sensor data	CNN model					
	VGG-16 CNN model			VGG-19 CNN model		
	Training	Validation	Test	Training	Validation	Test
Face RGB	100	100	52.27	100	99.98	50.43
Speech spectrum RGB	99.94	100	87.22	99.90	99.76	85.66
Face RGB + Speech spectrum RGB	98.81	98.80	66.74	98.82	98.96	71.18

(Unit: %)

19 restlessness level models using facial expression or vocal voice sensing approach 100%, that is, almost perfect recognition, indicating the excellent internal parameter convergences of these two types of VGG-CNN models when utilizing the established training data database. The sensor modality of acoustic speech data shows a higher performance in classifying restlessness levels for both VGG-CNN models. VGG-CNN deep learning models that use the input modality of face RGB data result in substandard performance in some situations, with average test performance values of only 52.27 and 50.43% with VGG-16 and VGG-19 CNN models, respectively (see Table 8). Compared with VGG-CNN restlessness level recognition using visual face data, vocal variations derived from the acoustic speech sensor data result in highly superior recognition accuracy. As shown in Table 8, the average recognition accuracies of 87.22 and 85.66% for VGG-16 and VGG-19 models, respectively, can be achieved utilizing the acoustic speech voice sensor data for the categorization of restlessness levels.

In the comparison of restlessness emotion recognition performance between VGG-16 and VGG-19 CNN models observed in Tables 4–8, VGG-16 CNN performs slightly better than VGG-19 CNN for both visual face and acoustic voice sensor data. As mentioned in Sect. 2, in the deep learning feature extraction of input sensor data, VGG-19 performs more convolution calculations than VGG-16. In the special case of restlessness level recognition using recorded continuous-time acoustic and visual sensor data from the four subjects in this study, VGG-16, with a slightly smaller computation load for convolution procedures, unexpectedly surpasses VGG-19.

Finally, with the combination of acoustic speech and visual face sensing data as input data for VGG-CNN restlessness classification calculations, as observed from the experimental results of average recognition accuracy shown in Table 8, 66.74 and 71.18% accuracies can be achieved when using VGG-16 and VGG-19 CNN models, respectively. Both of these values are significantly higher than those of VGG-CNN models with the input of visual facial expression data alone. Experimental results in Table 8 show that by supplementing acoustic speech raw

data to the input of VGG-CNN calculations, the accuracy of classification based on visual facial expression could be increased by 14.47% (from 52.27 to 66.74%) for VGG-16 CNN and 20.75% (from 50.43 to 71.18%) for VGG-19 CNN.

To clearly observe the variations in loss values and convergence conditions during VGG-CNN model training, recognition accuracy and loss curves in all 60 training epochs are also plotted (see VGG-16 CNN training conditions in Figs. 5–7 and VGG-19 CNN training conditions in Figs. 8–10). From these results of VGG-16 and VGG-19 CNN model

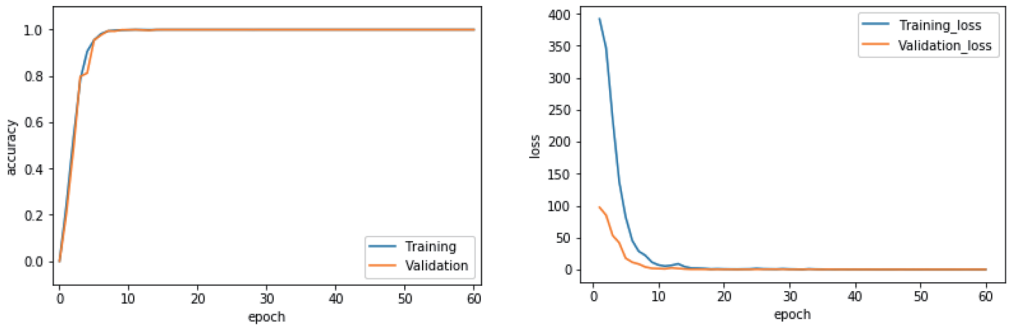


Fig. 5. (Color online) Training and validation accuracy curves, and training and validation loss curves for "VGG-16 CNN" model training with 60 epochs using "face RGB sensor data" (observations of subject-1).

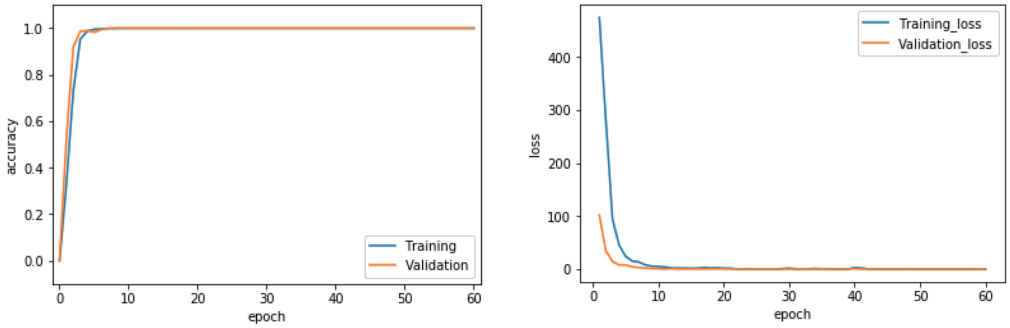


Fig. 6. (Color online) Training and validation accuracy curves, and training and validation loss curves for "VGG-16 CNN" model training with 60 epochs using "speech spectrum RGB sensor data" (observations of subject-1).

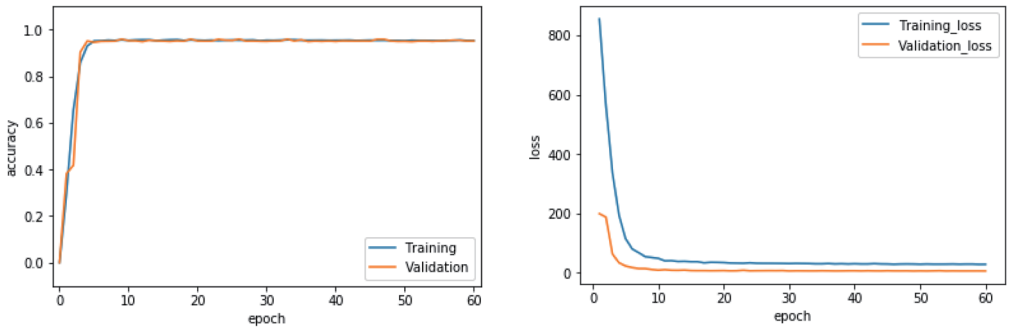


Fig. 7. (Color online) Training and validation accuracy curves, and training and validation loss curves for "VGG-16 CNN" model training with 60 epochs using "combined face and speech spectrum RGB sensor data" (observations of subject-1).

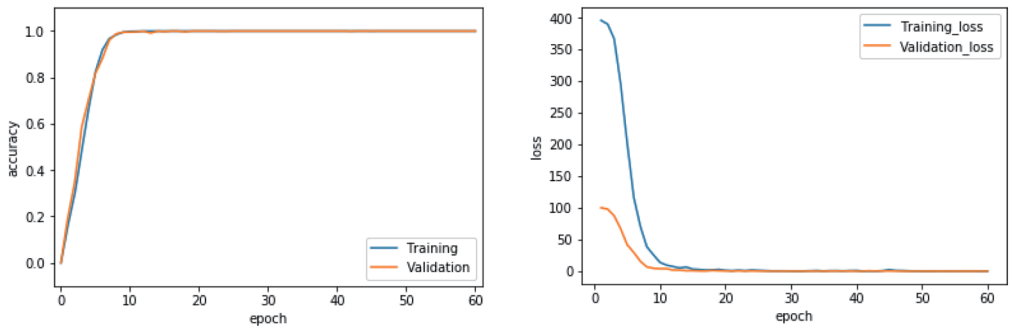


Fig. 8. (Color online) Training and validation accuracy curves, and training and validation loss curves for "VGG-19 CNN" model training with 60 epochs using "face RGB sensor data" (observations of subject-1).

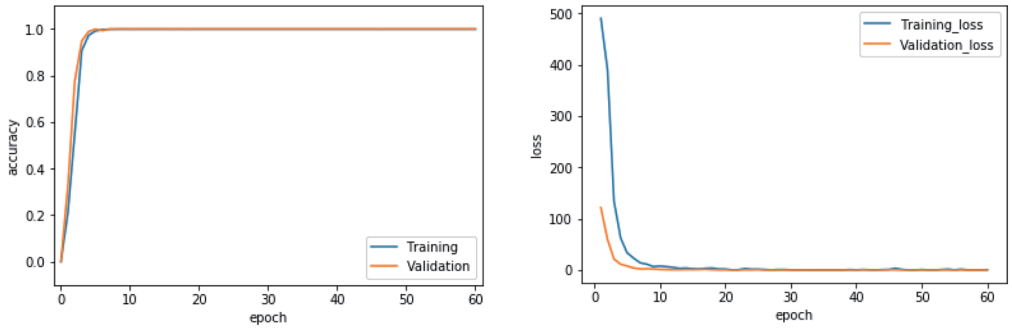


Fig. 9. (Color online) Training and validation accuracy curves, and training and validation loss curves for "VGG-19 CNN" model training with 60 epochs using "speech spectrum RGB sensor data" (observations of subject-1).

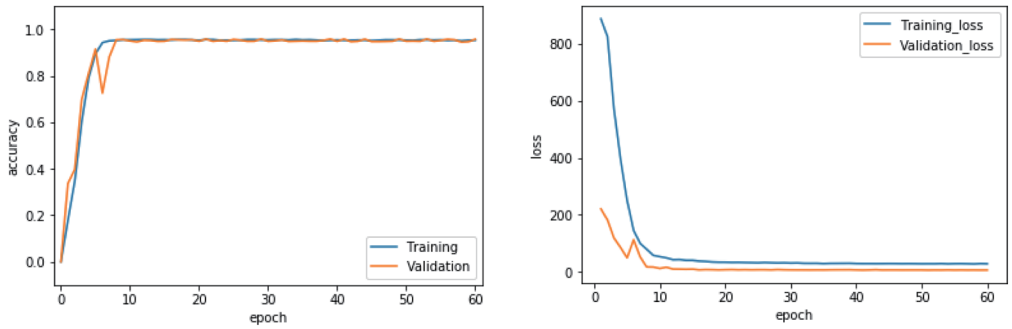


Fig. 10. (Color online) Training and validation accuracy curves, and training and validation loss curves for "VGG-19 CNN" model training with 60 epochs using "combined face and speech spectrum RGB sensor data" (observations of subject-1).

training, convergence with low and seemly invariant loss is seen to be reached before the tenth training epoch.

4. Discussion

As seen from the experimental results presented in Sect. 3, the modality of speech spectrum RGB sensor data exhibits the highest recognition performance for both VGG CNN-16 and

CNN-19 deep learning recognition schemes. Compared with utilizing speech spectrum RGB, deep learning emotion recognition using the face RGB data alone yields a dissatisfactory recognition result. Speech spectrum RGB data are apparently superior to face RGB data in terms of recognition accuracy because speech spectrum RGB data have much greater modality data characteristics variations among the nine defined categorizations of restlessness behaviors. The speech spectrum RGB image acquired from raw acoustic voice data includes distinguishable data separation features derived from utterance data of three different vocal sounds at three different volumes, as mentioned previously, which can be used to perform the categorization of restlessness behavior. On the other hand, the characteristics variations of the facial expression RGB data of the nine defined restlessness behaviors are obtained mainly from facial image data of three different eyebrow motions with three different eye expressions. Characteristics variations of the face RGB data obtained from only slight eyebrow motions and eye actions will neither be significant nor provide sufficient pattern separation information for VGG-CNN model learning and recognition. When using datasets of combined speech spectrum RGB and face RGB data, the presence of the face RGB data in the speech spectrum RGB dataset will apparently degrade the emotion categorization performance of VGG CNN recognition using only the acoustic sensor data. This performance reduction issue can be resolved by using the multilevel CNN scheme where two separate CNNs are independently used for extracting facial expression and vocal utterance features; this will be explored in a future study.

The presented restlessness emotion level classification system with two different types of VGG-CNN deep learning models in this work will be helpful for people with emotion problems. The presented system can be applied in real-life scenarios to classify people with restlessness problems, such as children with hyperactivity, elderly persons who live alone, people with certain afflictions (Parkinson's disease, aftereffects of stroke, and disabilities that hinder normal actions, for example), and people with frequent depression, enabling them to receive appropriate care in a timely manner. From the viewpoint of "smart care" applications, the proposed restlessness level categorization system that provides dynamical continuous emotion recognition will offer immediate emotion monitoring that can reduce the frequency of unexpected and risky events.

5. Conclusions

We presented a restlessness level recognition approach of using deep-learning-based VGG-CNN models with three different modalities of sensor data inputs: acoustic speech observations, visual face observations, and combined speech and face observations. Two popular VGG-CNN models, CNN-16 and CNN-19, were employed to evaluate the emotion recognition accuracy when using the three different types of input sensor data. Experiments on classifying the nine restlessness levels, each of which is defined by specific vocal utterance and facial expression data, of four subjects showed that CNN-16 with acoustic speech data input yields the highest average test recognition accuracy of 87.22%, and the supplemental incorporation of acoustic speech observations to the input data of the visual-facial-expression-based CNN emotion recognition model significantly increased the recognition rate.

Acknowledgments

This research is partially supported by the Ministry of Science and Technology (MOST) in Taiwan under Grant MOST 108-2221-E-150-037.

References

1 G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh: Proc. 2013 IEEE 6th Int. Conf. Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, USA, (Oct. 2013).
2 R. Min, N. Kose, and J.-L. Dugelay: IEEE Trans. Syst. Man Cybern. **44** (2014) 1534.
3 I. J. Ding and S. K. Lin: IEEE Access **5** (2017) 4154.
4 I. J. Ding and J. Y. Shi: Multimeida Tools Appl. **76** (2017) 25297.
5 I. J. Ding and J. T. Liu: Comput. Electr. Eng. **49** (2016) 173.
6 I. J. Ding and Z. G. Wu: IEEE Sens. J. **19** (2019) 8432.
7 I. J. Ding and Y. J. Chang: Neurocomputing **262** (2017) 108.
8 H. Chen, J. Li, F. Zhang, Y. Li, and H. Wang: Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA (2015).
9 R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp: IEEE Trans. Multimedia **12** (2010) 544.
10 C. H. Wu and W. B. Liang: IEEE Trans. Affective Comput. **2** (2011) 10.
11 Harouni, A. Karargyris, M. Negahdar, D. Beymer, and T. S. Mahmood: Proc. 2018 IEEE 15th Int. Symposium on Biomedical Imaging (ISBI) (2018) 872–876.
12 E. Acar, F. Hopfgartner, and S. Albayrak: Proc. 2015 13th Int. Workshop on Content-Based Multimedia Indexing (CBMI) (2015) 1–6.
13 J. J. Deng, C. H. C. Leung, P. Mengoni, and Y. Li: Proc. 2018 IEEE First Int. Conf. Artificial Intelligence and Knowledge Engineering (AIKE) (2018) 249–253.
14 F. Lingenfelser, J. Wagner, J. Deng, R. Brueckner, B. Schuller, and E. André: IEEE Trans. Affective Comput. **9** (2018) 410.
15 N. Churamani, P. Barros, E. Strahl, and S. Wermter: Proc. 2018 Int. Joint Conf. Neural Networks (IJCNN) (2018) 1–8.
16 B. Selbes and M. Sert, Proc. 2017 14th IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS) (2017) 1–6.
17 Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li: Proc. 2018 IEEE 15th Int. Symp. Biomedical Imaging (ISBI) (2018) 903–907.
18 W. Feng, N. Guan, Y. Li, X. Zhang, and Z. Luo: Proc. 2017 Int. Joint Conf. Neural Networks (IJCNN) (2017) 681–688.
19 J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth: IEEE/ACM Trans. Audio Speech Language Process. **25** (2017) 2362.
20 I. Tashev: IEEE Signal Process. Mag. **30** (2013) 129.
21 W. G. Parrott: Emotions in Social Psychology: Essential Readings (Psychology Press, UK, 2001).
22 K. Simonyan and A. Zisserman: Proc. Int. Conf. Learning Representations (ICLR), San Diego, CA (2015).