IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 23, 2020, accepted January 11, 2021, date of publication January 18, 2021, date of current version January 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052246

Spatial Temporal Variation Graph Convolutional Networks (STV-GCN) for Skeleton-Based Emotional Action Recognition

MING-FONG TSAI AND CHIUNG-HUNG CHEN
Department of Electronic Engineering, National United University, Miaoli 360001, Taiwan
Corresponding author: Ming-Fong Tsai (mingfongtsai@gmail.com)

ABSTRACT The main core purpose of artificial emotional intelligence is to recognize human emotions. Technologies such as facial, semantic, or brainwave recognition applications have been widely proposed. However, the abovementioned recognition techniques for emotional features require a large number of training samples to obtain high accuracy. Human behaviour pattern can be trained and recognized by the continuous movement of the Spatial Temporal Graph Convolution Network (ST-GCN). However, this technology does not distinguish between the speed of delicate emotions, and the speed of human behaviour and delicate changes of emotions cannot be effectively distinguished. This research paper proposes Spatial Temporal Variation Convolutional Network training for human emotion recognition, using skeleton detection technology to calculate the degree of skeleton point change between consecutive actions and using the nearest neighbour algorithm to classify speed levels and train the ST-GCN recognition model to obtain the emotional state. Application of the speed change recognition ability of the Spatial Temporal Variation Graph Convolution Network (STV-GCN) to artificial emotional intelligence calculation makes it possible to efficiently recognize the delicate actions of happy, sad, fear, and angry in human behaviour. The STV-GCN technology proposed in this paper is compared with ST-GCN and can effectively improve the recognition accuracy by more than 50%.

INDEX TERMS Artificial emotional intelligence, spatial temporal graph convolution network, human skeleton joint point.

I. INTRODUCTION

Realization of human-emotion recognition applications in open fields such as transportation systems and metropolitan squares can avoid possible dangerous conflicts. To avoid the occurrence of regrets, the system can actively identify a specific area that has emotional conditions such as anger or sadness and actively notify the manager of the specific area to despatch staff to assist and deal with it. Recognition of human emotions can be done through changes in human facial features or delicate continuous movements [1]–[5]. At present, deep-learning image, speech, and brain-wave recognition technologies still need to work hard to recognize the delicate changes in human emotions. Since delicate changes of human facial features such as happy, angry, fear, and sad emotions require a large amount of data to be collected for image recognition models to train the neural network in deep learning, the training data are a series of images with exaggerated expressions to facilitate the inference process and reach a conclusion. With regard to the differences in expression of happy, angry, fear, and sad emotions through human language, semantic speech analysis needs to solve the problems of cultural variation and sound source noise filtering to obtain reliable identification results. Emotional brainwave analysis [6] requires sophisticated equipment to perform the recognition function, but its construction cost is relatively high and the effectiveness is not good. Consideration of continuous human body movements in artificial emotional intelligence training and recognition has been gradually discussed by research scholars. Relevant literature [7] uses continuous walking actions related to human emotions for data set construction and recognition model training, but currently there is no behavioural distinction regarding the speed of delicate emotional movements, which results in ineffective

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

recognition of the movement speed and delicate changes of human behaviour.

This research paper proposes the use of a Spatial Temporal Variation Graph Convolutional Network (STV-GCN) training for human emotion recognition, skeleton detection technology to calculate the degree of skeleton point changes between specific consecutive actions, the k-nearest neighbour (KNN) algorithm for speed level classification, and the Spatial Temporal Graph Convolutional Network (ST-GCN) [8]–[13] for emotional state classification. This research paper aims at the recognition and classification of the emotional state of the continuous movement speed of the human body. PoseNet skeleton posture point detection technology is used to extract the information of various human joint points and the amount of skeleton joint point variation between consecutive actions is calculated. Based on the above proposed method, the speed levels of human behaviour can be effectively distinguished, and this method is applied to human emotional state recognition to distinguish between the more rapid (fast) and delicate actions such as anger and tardy (slow) human emotional states such as sadness. The system recognizes the human emotional state after classifying the action speed level using the KNN algorithm. Since ST-GCN can obtain reliable posture recognition ability after training through distinguishing between movement speed levels, it can perform effective classification of extreme and general happiness, extreme and general anger, and extreme and general sadness. The remaining chapters of this research paper are arranged as follows. Chapter 2 presents related research to explore the feasibility and reliability of using human body language to recognize emotions and briefly introduces related technologies such as ST-GCN, PoseNet, and KNN. Chapter 3 presents the KNN algorithm for classifying action speed levels and the recognition and classification method of STV-GCN. Chapter 4 describes the experimental parameter setting, processing, and data analysis of results, and Chapter 5 provides the conclusions and future prospects.

II. RELATED WORK

A. HUMAN EMOTION EXPRESSION

1) DEEP LEARNING CLASSIFICATION

Emotion recognition technology has been widely discussed and researched in recent years. In work by [14], continuous frame emotion recognition has been carried out relying on two neural networks. The authors proposed to extract single frame images features through Convolutional Neural Network (CNN) and its characteristics using Recurrent Neural Network (RNN) to do the time connection, and analysed the contribution of each neural network component to the overall performance of the system. A related research work [15] combined CNN and Long Short-Term Memory (LSTM) network, and proposed a method of emotion-related feature extraction with context awareness. Since the features extracted relying on CNN are limited to two-dimensional data, it is unreliable for posture recognition, but the above two methods use RNN and LSTM to concatenate the features extracted in the time dimension. Therefore, the above two methods can effectively perform feature processing on continuous facial expression states or voice messages. However, the data set used in the above two works is RECOLA, which contains visual and physiological data. The physiological data of electrocardiograms and electrical skin activity data are difficult to obtain and are not suitable for open field or long-term emotional monitoring. Related literature [7] provides the Emotion Walk data set, which uses the three human emotions of sadness, anger, and happiness to walk patterns representing different emotions of body movements. That research was based on the deep features of the Long Short-Term Memory Network and the Random Forest classifier to train and recognize walking patterns associated with human emotions. Since the subjects express their emotional state and the person concerned may not be able to perceive specific emotions, this research paper focuses on the emotional state perceived by the observer rather than the emotion described by the subject. Although the easiest way to detect human emotions is by facial image recognition technology, the pattern of body behaviour is also a very important technology for artificial emotional intelligence to recognize emotions. Relevant literature [7] proves that training in and recognition of human walking style and walking posture help the perception of human emotions.

2) KNN CLASSIFICATION

The KNN algorithm uses supervised machine learning technology to measure the similarity of the Euclidean distance or angle cosine distance for different feature points. Related literature [16] uses the Berlin and Hindi emotion databases to classify and train the recognition model and the KNN algorithm to classify different voice messages to obtain emotion classification results. Related literature [17] classification of emotional states of speech spectral features of Mandarin through the KNN algorithm. The related literature [18] uses the nearest neighbour algorithm and local binary mode for image feature extraction and classification to obtain the facial feature emotional state. It can be seen that many researches have used the KNN algorithm to classify emotional features. No matter whether the classification is based on the speech spectrum or image features, the algorithm has excellent performance.

B. HUMAN EMOTION RECOGNITION

Related literature [12] proposed the ST-GCN identification model as a topological structure graph convolutional neural network that combines time and space and is connected by complex points and complex edges. Compared with the traditional Convolutional Neural Network architecture, the Graph Convolution Network (GCN) is not limited to the convolution of Euclidean geometry, as the features that can be extracted include social networks, three-dimensional graphics, and skeleton models that exceed the limitations of time and space, as shown in the human skeleton GCN architecture in Figure 1.

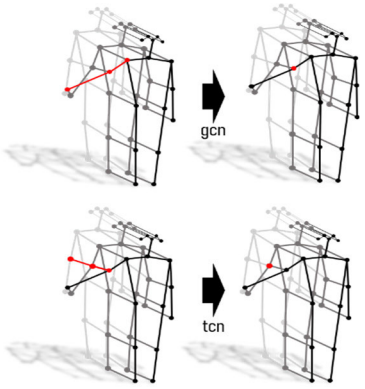


FIGURE 1. ST-GCN concept.

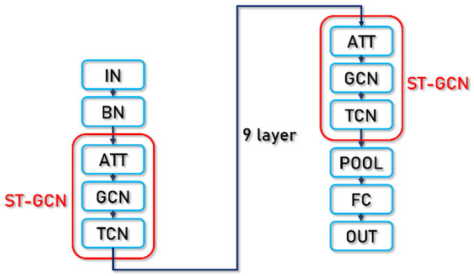


FIGURE 2. ST-GCN architecture.

ST-GCN uses OpenPose to extract key point information about the skeleton and performs convolution operations to extract features for recognition of human action poses. The aforementioned human skeleton key points are the image space dimension, and the continuous human actions are the image time dimension. The continuous movement changes of the human skeleton key points in the space and time dimensions are integrated for feature training and recognition classification. The custom network structure of the ST-GCN is divided into three parts as shown in Figure 2. The first part is the input layer, which is mainly for the normalization of the batch of training data to enable convergence of the method and reduces the range of accuracy changes. The second part is the ST-GCN layer, which contains three functions: the attention model function, the GCN function, and the Temporal Convolution Network (TCN) function. The attention model function allows the network architecture to pay more attention to important features and point information, and the TCN and GCN functions perform topological convolution processing for the time dimensions and spatial dimensions respectively. The ST-GCN structure performs nine ST-GCN layer actions as a whole to extract more detailed feature information. The third part is the pooling layer and the fully connected layer, which realize the classification of the continuous motion state of the human body.

Bone-based action recognition methods have recently received increasing attention due to their powerful and superior performance. This section briefly reviews some neural networks based on skeletal action recognition. In [19], the authors use dual-stream RNN to model space and time separately. The temporal RNN modelling uses a stacked RNN that can be used to process variable-length sequences, and according to human kinematics. A hierarchical RNN is designed, and a graph traversal method for accessing joint points in the sequence based on the adjacency relationship is proposed in the spatial modelling. From this research, it can be seen that RNN can be used as a skeleton-based action recognition method. Because GCN is highly adaptable to the topological structure of the human skeleton, there are many related research works that use GCN as a neural network for skeleton action recognition. In [20], the authors proposed an adaptive adjacency matrix that can be optimized together with other parameters during the training process, and a dual-stream training strategy that separates the skeleton and joint points for training, so that the identification results of this study have a good accuracy. Work by [21] included the design of multiple adjacency matrix generation strategies, and used the Cross-entropy Evolution strategy with Importance-Mixing (CEIM) search method proposed in this research to select strategies in the Neural Architecture Search (NAS) space, so that the neural network uses different adjacency matrix generation strategies in different training layers.

III. STV-GCN

This research paper proposes STV-GCN training for human emotion and action recognition. Skeleton detection technology is used to calculate the degree of change in speed of the skeleton point between specific consecutive actions, the KNN is used for speed classification, and the ST-GCN recognition model is used for emotional posture training and classification. The training model of the continuous skeleton point of the specific action after distinguishing the action speed level can obtain the ability to recognize delicate postures of humans.

A. SYSTEM OVERVIEW

As shown in Figure 3, a continuous motion video of a human walking with three emotions will be captured by the camera and then entered into the system for frame cutting, and PoseNet will be used to extract the key point information of the human skeleton. The key point numbers of the control human body are shown in Figure 4. Numbers 3 to 5 are points for the left arm, numbers 6 to 8 are points for the right arm, numbers 9 to 11 are points for the left leg, numbers 12 to 14 are points for the right leg, and the rest of the numbers are all points on the face and neck. After extracting the maximum variation feature for the key points of the human skeleton, the key point information of the human skeleton is subjected to the ST-GCN to identify different human emotional poses, and the KNN is used to classify the movement speed levels of different human emotional postures to determine the degree of emotion.

B. SYSTEM ARCHITECTURE

The system architecture is shown in Figure 5. The continuous motion video is cut into frames and then enters the speed

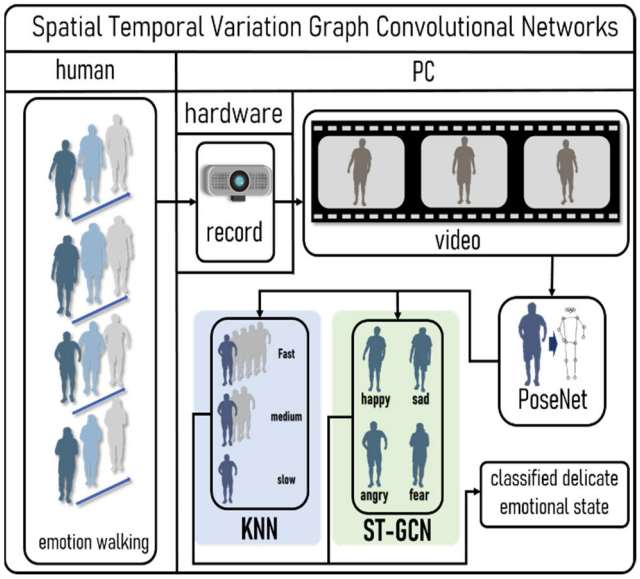
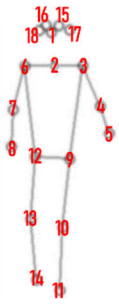


FIGURE 3. System overview.



Number of joint point	Corresponding joint point	Number of joint point	Corresponding joint point
1	Nose	10	Left knee
2	Middle point between shoulders	11	Left ankle
3	Left shoulder	12	Right hip
4	Left elbow	13	Right knee
5	Left wrist	14	Right ankle
6	Right shoulder	15	Left eye
7	Right elbow	16	Right eye
8	Right wrist	17	Left ear
9	Left hip	18	Right ear

FIGURE 4. Human joint points.

levels and emotional posture network architecture. The speed levels classification model takes the maximum variation of the key points of the human skeleton as the feature of the KNN algorithm training and distinguishes it into three speed classes: fast, medium, and slow. The emotional posture classification model normalizes the key points of the human skeleton in batches, and its purpose is to allow the algorithm to converge and reduce the range of accuracy changes. Then the attention model is used to focus on the important topological point edge features during neural network training. Then features are extracted through the GCN and TCN layers. The GCN is the convolution of the spatial dimension of the topological graph. The actual method is to take the edge as the weight to calculate the weighted average of the node features. The TCN is the same as the topological graph structure but acts on the spatial dimension. Finally, the pooling layer and the fully connected layer so that the neural network architecture can accurately classify human emotional postures.

C. ACTION SPEED LEVELS CLASSIFICATION

The ST-GCN technology can efficiently identify and classify two or more different human continuous actions but cannot

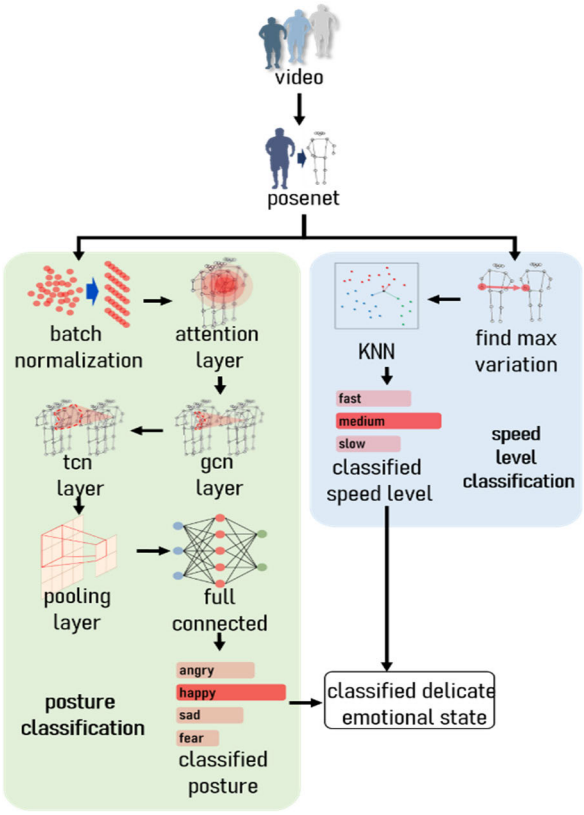


FIGURE 5. System architecture.

identify and classify the speed of continuous actions, such as fast walking and slow walking. If the speed cannot be identified and classified, human emotions will not be recognized. For example, fast actions are angry, nervous, or happy emotional states and slow actions are fearful or sad emotional states. In view of the fact that the speed difference in the continuous movement of the human body can be judged according to the degree of skeleton point variation between consecutive pictures, this research paper calculates the coordinate variation of each skeleton joint point of a specific movement between consecutive action frames. Each skeleton joint point only retains the maximum amount of variation in all consecutive frames of a specific action as the KNN training sample to obtain the identification model as shown in Figure 6.

The system separates the input videos into training and test data in the PoseNet stage, and the training data are cut into frames after labelling different speed levels, while the test data are cut into frames directly. The purpose of frame cutting is to enable PoseNet to capture key point information of the human skeleton. The system performs the calculation of the variation of the key points of the human skeleton in the two-dimensional plane in the Find Max Variation stage. The x-axis is used as the reference to find the displacement of all the key points of the human skeleton in all frames, and then the y-axis is used as the reference to find the displacement of all the key points of the human skeleton. The y-axis human skeleton key

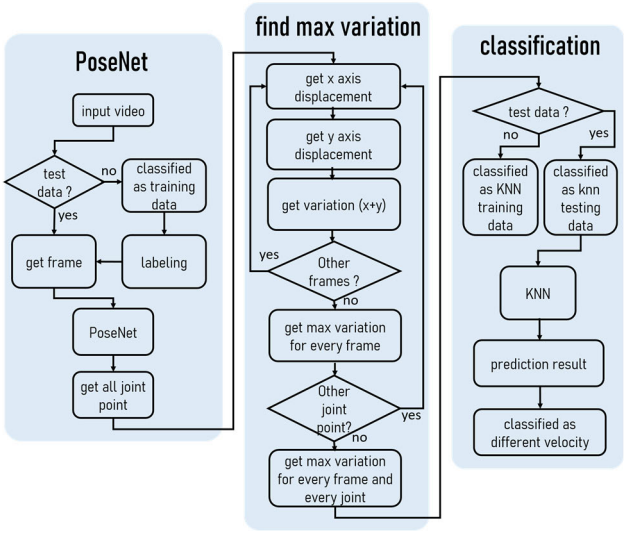


FIGURE 6. Flow chart of speed level classification algorithm.

point displacement between all frames and the x-axis human skeleton key point displacement between all frames are added one by one, and the maximum value of the variation between all frames is obtained as the KNN classification feature to perform speed step recognition training.

Equation (1) represents the extraction of the change vector between frames, where f represents the current frame, and $x_{(f)}$ represents the position of the joint point on the x-axis in the frame f. Similarly, $y_{(f)}$ is the position of the joint point on the y-axis in the frame f, and then the joint point changes of the two axes are added to get $\vec{V}$, that is, the joint point change vector between two frames. Formula (2) represents the largest joint point change vector in the entire video data, where F is the total number of frames of the entire video. The method of extracting the largest change vector is to perform the result of Equation (1) F times, store the result in the set, and then take the maximum value for the set. Formula (3) indicates that the members in the feature set are to extract the maximum joint point change vector of Formula (2) on all 18 joint points on the human body.

$$\vec{V} = (x_{(f+1)} - x_{(f)}) + (y_{(f+1)} - y_{(f)}) \quad (1)$$

$$\vec{V}max = max\{\vec{V}_i \mid 1 \le i < F, i \in Z\} \quad (2)$$

$$feature = \{\vec{V}max_j \mid 1 \le j \le 18, j \in Z\} \quad (3)$$

D. RECOGNITION DELICATE EMOTIONAL STATE

This research paper proposes the STV-GCN according to the classification of speed levels to determine the three delicate states of human emotions as fast, moderate, and slow. The human continuous motion skeleton joints extracted by PoseNet point information are used by ST-GCN for classification training to obtain the characteristic recognition model of skeleton key point variation with speed differences unique to human emotions and delicate emotion state recognition

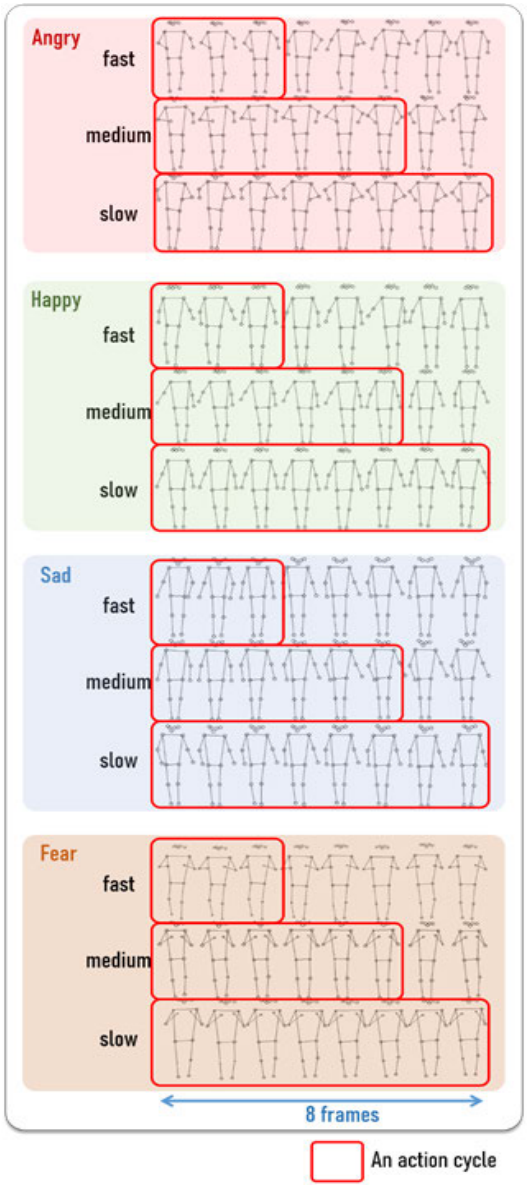


FIGURE 7. Recognition of delicate emotions.

process. As shown in Figure 7, the four emotional states of anger, happiness, sadness and fear all have different skeleton postures; for example, in the angry action, the elbows are bent and swing back and forth; in the happy posture, the hands are relaxed and have a large swing; in the fear posture, both elbows are bent tightly and the hands are close to the mouth and in the sad posture, the head is down and the swing of both hands is small. Each emotion can be divided into three speed levels, and each speed level has a different variation between frames according to its speed. If a fixed frame number is used as the unit, the action period included in each speed also has a difference in length. Therefore, the action speed classification model defines three levels of fast, moderate, and slow speed, and the emotional action posture prediction model defines four states of anger, happiness, sadness, and

fear. The emotional state can be analysed in a more detailed manner according to the speed level of its action. For example, the degree of anger in ''fast anger'' is more serious than that in ''slow anger''.

IV. EFFECTIVENESS EVALUATION AND EXPERIMENTAL RESULTS

A. ACTION SPEED LEVEL CLASSIFICATION

1) DATASETS AND DATA PREPROCESSING

The performance evaluation in this chapter defines three different speed levels for the action of putting the hands up. The training data set is based on beats per minute (bpm) as the speed standard, and the fastest action speed step is 150 bpm; that is, a complete action takes 1.6 seconds. The established speed steps are 150 bpm, 90 bpm (2.68 s/act), and 30 bpm (8.0 s/act) action speed levels and define the speed range of the first level as the slowest and the third level as the fastest. Twenty movies are used for the training of the KNN classification model for the three speed levels. The pre-processing of the data set involves performing PoseNet skeleton joint point detection of all the video data, and performing the feature extraction method of formulas (1) to (3) on the joint data of each video for the maximum joint point change vector. Taking this experiment as an example, there is a total of 20 video data for KNN training. After each video passes through PoseNet, it will get 100 frames multiplied by 18 joint points of training data, and each joint point of the whole video, after performing the calculation of formula (1) individually, will obtain 99 change vectors between 100 frames, after which we extract the maximum value of formula (2) from these 99 change vectors. The extraction result will be 1 maximum joint point change vector. Finally, through formula (3), all 18 joint points are repeated to perform the above actions to obtain the maximum joint point change vector feature set of a video.

2) EXPERIMENTAL RESULTS

Since the ST-GCN architecture cannot identify and classify the same human emotional actions at different speeds, this problem can be solved efficiently by using the action speed classification algorithm proposed in this research paper. The experimental environment parameters define three different speeds for the same action, which are the fast, moderate, and slow speeds of putting up the hands in the video. The above three videos are used as ST-GCN training samples to obtain a classification model as shown in Figure 8. The ST-GCN identification model obtained from the experimental results cannot accurately classify the test samples of the above three different speeds because the slow action provides more continuous details of the action, as shown in Figure 9, and the other two actions cannot be accurately classified by the ST-GCN architecture. The STV-GCN proposed in this research paper classifies human emotions as fast, moderate, and slow according to the classification of speed levels. The characteristics of the maximum variation of the key point

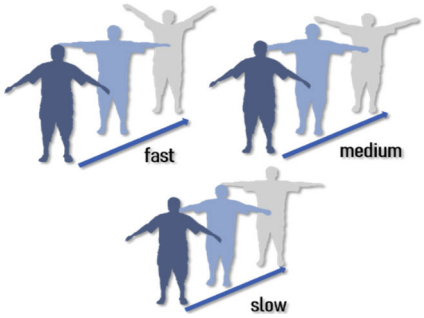


FIGURE 8. Action of putting hands up.

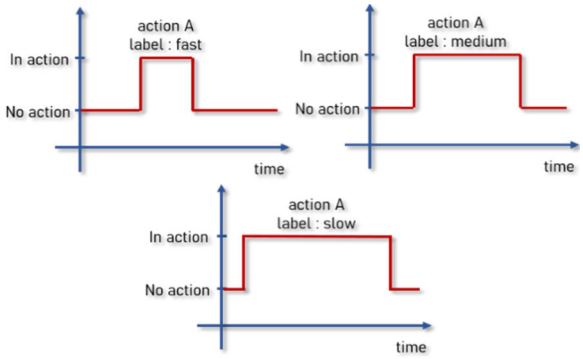


FIGURE 9. Speed difference.

TABLE 1. Classification results of speed levels in the same action.

Level of speed	*Beats per minute (bpm)*	*Second per action (s/act)*	*Accuracy*
3	150	1.60	73.34%
2	90	2.68	86.67%
1	30	8.00	93.34%

vector of the human skeleton for different speed levels are shown in Figure 10. The experimental results in this chapter are classified and tested with 15 films for different speed levels. The recognition accuracies of speed levels 1, 2, and 3 are 93.34, 86.67, and 73.34%, respectively. The abovementioned overall average accuracy is 84.45%, as shown in Table 1.

B. RECOGNITION OF DELICATE EMOTIONAL STATE

1) DATASETS AND IMPLEMENTATION DETAILS

The effectiveness evaluation in this chapter defines three emotion markers of anger, happiness, sadness, and fear. Each emotion label includes three speed levels: fast, moderate, and slow, and speed levels of 240, 187, and 150 bpm, respectively. The abovementioned represent a total of 360 video training sample data sets. The experimental environment of the system settings of the ST-GCN is as follows: the CPU is an i7 7700, the GPU is an RTX 2080 Ti, and the operating system is Windows 10, while the training parameters are as follows:

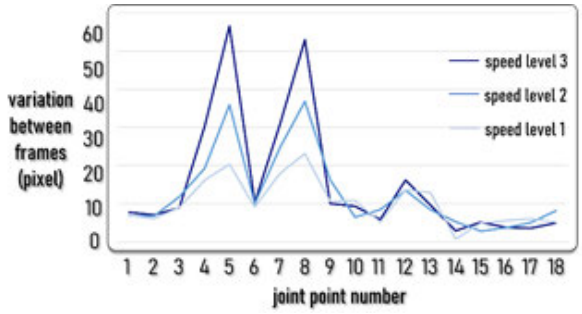


FIGURE 10. Maximum variation feature of key point vector.

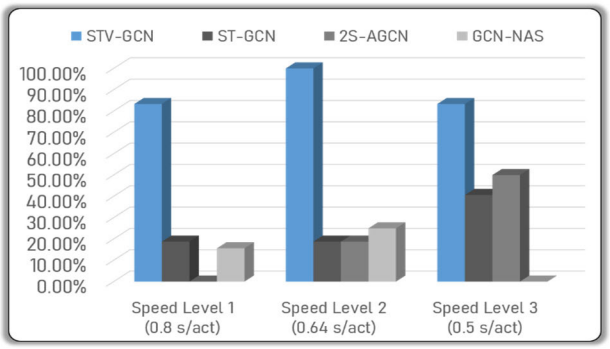


FIGURE 11. Classification results of speed levels for different emotions.

TABLE 2. The parameters of the training network.

Spatial kernel size	*Temporal kernel size*	*Strides*	*Activation function*	*Padding*	*Number of epoch*
3	9	1	ReLU	4	10,000

an epoch is 100,000 steps, a batch size is 5, and a test batch size is 300. The spatial kernel size is 3, the temporal kernel size is 9, and except for the 4th and 7th temporal convolution layers which are set to 2,the other strides are set to 1, while the activation function uses the ReLU function and the padding is set to 4, as shown in Table 2.

2) EXPERIMENTAL RESULTS

The experimental results in this chapter are classified and tested with four movies for different delicate emotions. In the speed classification stage, the recognition accuracies of speed levels 1, 2, and 3 are 83.33, 100, and 83.33% respectively. The abovementioned overall average accuracy of STV-GCN is 88.89%, as shown in Figure 11. In the speed classification stage, the accuracy of each emotion of the ST-GCN [12] is the sum of the three speeds of the same emotion. Taking this experiment as an example, the accuracy of anger is the sum of anger-fast, anger-medium, and anger-slow. Because it cannot distinguish between the speed of the action, the total accuracy

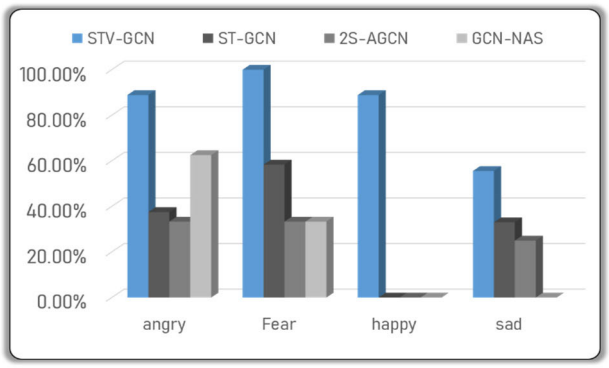


FIGURE 12. Recognition result of action states for different emotions.

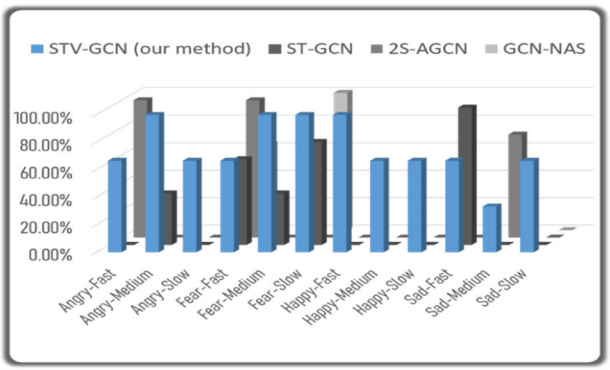


FIGURE 13. Recognition result of delicate emotional states.

of the speed classification part of the ST-GCN is only 26.04%. In the speed classification stage, the accuracy of each emotion of the 2S-AGCN [20] is the sum of the three speeds of the same emotion. Taking this experiment as an example, the accuracy of anger is the sum of anger-fast, anger-medium, and anger-slow. Because it cannot distinguish between the speed of the action, the total accuracy of the speed classification part of the 2S-AGCN is only 22.92%. In the speed classification stage, the accuracy of each emotion of the GCN-NAS [21] is the sum of the three speeds of the same emotion. Taking this experiment as an example, the accuracy of anger is the sum of anger-fast, anger-medium, and anger-slow. Because it cannot distinguish between the speed of the action, the total accuracy of the speed classification part of the GCN-NAS is only 13.54%. Since STV-GCN performs KNN classification for the change of each frame of the action, the total accuracy of the evaluation results in this part is 88.89%.

In the emotional state classification stage, the emotion recognition accuracies for anger, fear, happiness, and sadness are 88.89, 100, 88.89, and 55.56%, respectively. The abovementioned overall average accuracy of STV-GCN is 83.34%, as shown in Figure 12. Because it cannot distinguish between the speed of the action, the total accuracy of the speed classification part of the ST-GCN is only 32.21%. Because it cannot distinguish between the speed of the action, the total accuracy of the speed classification part of the 2S-AGCN is only 22.92%. Because it cannot distinguish

between the speed of the action, the total accuracy of the speed classification part of the GCN-NAS is only 23.96%. Considering the detection results of the nine states of delicate emotions, the average accuracy rate of STV-GCN is 75.00%, as shown in Figure 13. In the emotional state and speed classification stage, the ST-GCN is directly divided into nine types of labels for training and classification. Because it has more labels and considering the effect of movement speed, the total accuracy rate is only 26.04%. Because it has more labels and considering the effect of movement speed, the total accuracy rate of 2S-AGCN is only 22.92%. Because it has more labels and considering the effect of movement speed, the total accuracy rate of GCN-NAS is only 13.54%. The STV-GCN has a total accuracy rate of 75.00% after the reference results of the speed classification stage and the emotional state classification stage.

V. CONCLUSION AND FUTURE WORK

This research paper proposes STV-GCN training for human emotion recognition, using skeleton detection technology to calculate the degree of skeleton point change between consecutive actions and using the nearest neighbour algorithm to classify speed levels and train the ST-GCN recognition model to obtain the emotional state. The proposed STV-GCN is an artificial emotional intelligence calculation with fine motion speed change recognition. The recognition accuracy of the system in different speed steps of the same action reaches 88.89%, and the recognition accuracy of emotional states reaches 83.34%. The STV-GCN obtains a better accuracy rate than the ST-GCN in the speed classification stage and the emotional state classification stage by more than 50%.

REFERENCES

[1] C. Kashyap and P. Vishnu, "Facial emotion recognition," *Int. J. Eng. Future Technol.*, vol. 7, no. 7, pp. 18–29, 2016.

[2] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, pp. 401–421, 2018.

[3] Y.-D. Zhang, Z.-J. Yang, H.-M. Lu, X.-X. Zhou, P. Phillips, Q.-M. Liu, and S.-H. Wang, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.

[4] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun. 2016.

[5] X. Ma, W. Lin, D. Huang, M. Dong, and H. Li, "Facial emotion recognition," in *Proc. IEEE Int. Conf. Signal Image Process.*, Dec. 2017, pp. 77–81.

[6] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[7] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "Identifying emotions from walking using affective and deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2019, pp. 1–15.

[8] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3634–3640.

[9] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3289–3299.

[10] G. Xiao, R. Wang, C. Zhang, and A. Ni, "Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks," *Multimedia Tools Appl.*, vol. 79, pp. 1–19, Mar. 2020.

[11] Q. Zhong, C. Zheng, and H. Zhang, "Research on discriminative skeleton-based action recognition in spatiotemporal fusion and human-robot interaction," *Complexity*, vol. 2020, pp. 1–10, Aug. 2020.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7444–7452.

[13] A. Sun and Y. Huang, "A traffic balance scheme of group emotion recognition by using the service function chain," *Int. J. Commun. Syst.*, vol. 32, no. 14, pp. 1–15, 2019.

[14] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 619–623.

[15] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.

[16] A. Bombatkar, G. Bhoyar, K. Morjani, S. Gautam, and V. Gupta, "Emotion recognition using speech processing using k-nearest neighbor algorithm," *Int. J. Eng. Res. Appl.*, vol. 3, no. 2, pp. 68–71, 2014.

[17] T. Pao, Y. Chen, J. Yeh, Y. Cheng, and Y. Lin, "A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech," in *Proc. Int. Conf. Intell. Comput.*, 2007, pp. 997–1005.

[18] G. Panchal and K. Pushpalatha, "A local binary pattern based facial expression recognition using K-nearest neighbor (KNN) search," *Int. J. Eng. Res. Technol.*, vol. 6, no. 5, pp. 525–530, 2017.

[19] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–10.

[20] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[21] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2020, pp. 1–9.

MING-FONG TSAI received the Ph.D. degree from the Department of Electrical Engineering, Institute of Computer and Communication Engineering, National Cheng Kung University, Taiwan. He is currently an Assistant Professor with the Department of Electronic Engineering, National United University, Taiwan. His current research interests include the Internet of Things, mechanism and deep learning technology, vehicular communications, and multimedia communications.

CHIUNG-HUNG CHEN received the B.S. degree from the Department of Electronic Engineering, National United University, Taiwan. His current research interests include Artificial Internet of Things, machine learning, and deep learning technology.