JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, XXX-XXX (2023)
DOI: 10.6688/JISE.20220X_38(X).00XX

Human Activity Recognition with Long Short-Term Memory Network Based on Spatio-temporal Features

MING-FONG TSAI* AND SHIH-YUNG CHENG

Department of Electronic Engineering, National United University, Miaoli 360302, Taiwan
Corresponding: mingfongtsai@gmail.com

Current frameworks for human motion recognition rely on the identification of crucial points in the human skeleton, taking into account the spatio-temporal variation between consecutive images. However, existing frameworks are not capable of effectively training models to identify varying levels of a single motion. This paper introduces a novel approach to improving the accuracy of human movement recognition by utilising the average movement rate of the backbone as a deep learning spatio-temporal feature. Keypoint information captured using OpenPose human skeleton recognition technology is used to calculate the average backbone movement rate for adjacent keypoints. The Spatial Temporal Graph Convolutional Networks (ST-GCN) framework for human movement recognition is employed to train models to identify basic types of movement, and the Long Short-Term Memory (LSTM) framework is used to train models to identify advanced movement levels using the average movement rate features of the backbone. The ST-GCN and LSTM models are integrated to obtain the overall human motion and recognition results at the motion level. The method proposed in this paper is compared with human movement recognition models in the related literature and a performance analysis is carried out. In terms of accuracy, our approach outperforms the ST-GCN and LSTM models by at least 4% on a human squatting motion dataset, and outperforms the LSTM and Spatial Temporal Variation Graph Convolutional Networks (STV-GCN) models by at least 8% on a dataset of humans walking with different emotions.

Keywords: Long Short-Term Memory Network; Spatio-temporal Features; Human Activity Recognition;

1. INTRODUCTION

With the rapid advancements in deep learning, artificial intelligence and image processing and analysis technologies, image recognition applications have progressed from the recognition of single face images to continuous images of human movement. Human movement recognition applications involving the analysis and evaluation of specific movements and emotional gestures have been the focus of recent developments. Human movement recognition can be utilised to assess whether an individual's body conforms to the demands of specific sports movements, or can be used to provide advice and suggestions to help athletes improve their skills and abilities, and to prevent sports-related injuries [1-2]. An additional application of human movement recognition technology involves assessing whether an individual is in a negative emotional state by detecting his or her body movements, amplitudes of oscillation and speed of movement. This enables managers to provide appropriate assistance and guidance to prevent unfortunate events, or to maintain the safety of the community [3-7]. A human motion recognition framework utilises deep learning techniques to learn the features of the spatio-temporal variation in the key points of the human skeleton between successive images. For example, the point-movement rate of a keypoint is used to learn and analyse deep learning features with a focus on the changes in human posture under specific types of motion or emotional

situations [8-13]. The related literature [14] contains an open dataset which has been used as a basis for the classification of human deep squatting movements; the authors used a 3D keypoint detection technique to obtain the keypoint information of the human skeleton, and calculated the Euclidean distance between all the key points. In this way, they obtained a symmetric distance matrix, which was flattened into a vector matrix information, and then used a one-dimensional convolutional CNN for classification. However, this approach cannot efficiently a train model to identify different levels of the same squatting movement of the human body. In another study, an open dataset of human walking emotion gestures was designed and built, and a complete classification was provided. A Long Short-Term Memory (LSTM) human movement recognition framework was used to train a model to identify the emotions associated with walking movements [15]. However, this model was unable to efficiently recognise different levels of motions in humans undergoing the same walking movement.

The Spatial Temporal Variation Graph Convolutional Networks (STV-GCN) [16] human movement recognition framework was proposed to solve this problem. It was used to train a basic model of movement types, and the KNN machine learning classification algorithm was used for training and identification of the movement level based on the keypoint displacement motion rate of the human skeleton. The Spatial Temporal Graph Convolutional Networks (ST-GCN) human motion recognition and KNN motion level classification models were integrated to obtain recognition results for both the motion itself and the level of activity. However, this scheme does not provide efficient model training and recognition for the temporality of motion-level features. In this paper, we use the average motion rate of the backbone as a feature to enhance the learning of spatio-temporal features and improve the accuracy of human motion recognition. To achieve this, OpenPose[17] human skeleton keypoint recognition technology is employed to capture and compute the average movement rate of the backbone between adjacent frames. Next, the ST-GCN human movement recognition framework is applied to train a basic model of movement types. The LSTM human movement recognition framework is supplemented with the average movement rate of the backbone as a feature for the training of advanced movement level models. Finally, the ST-GCN and LSTM models are integrated and the model stacking technique [18]-[20] is applied to enhance the accuracy of human motion recognition. This research paper is structured as follows: we present a review of the relevant literature in Section 2; a description of the proposed backbone spatio-temporal features for deep learning of human action recognition is given in Section 3; our experimental methodology and performance analysis are introduced in Section 4; and our conclusions and suggestions for future work are presented in Section 5.

2. RELATED WORK

The Spatial Temporal Graph Convolutional Networks (ST-GCN) [21] human motion recognition framework was proposed for the training and identification of time series data on the keypoints of the human skeleton from continuous images. It learns the correlations between the keypoints of the skeleton over time and space from continuous images, allowing for training and identification by a motion recognition model. The ST-GCN framework uses attention models to learn the critical features of the human skeleton. In the training process, feature extraction is performed using graph convolution and time convolution network layers. The trained model is formed from a combination of nine ST-

GCN units, each of which applies the attention-increasing mechanism of ResNet. The overall ST-GCN human action recognition framework takes human skeleton keypoint information as input to the standardisation layer, and the standardised keypoint information is then passed to the nine ST-GCN feature training modules. A dropout function is applied to avoid overfitting problems, and SoftMax is used for final classification through the pooling layer. The above classification model uses a stochastic gradient descent method for feature learning, in which the learning rate is reduced at a specific time.

The ST-GCN human action recognition framework has been found to be proficient in terms of learning image features and effectively classifying various human actions; however, it focuses on learning image features, and cannot efficiently train models to recognise different levels of the same human action. A combination of the LSTM and random forest algorithms has been proposed for the training of human motion recognition models. A public dataset of humans walking while expressing different emotions was designed for use in training models to identify four types of human emotions: happy, angry, sad and neutral. In this study, RGB videos of emotional walking movements were extracted in the form of 3D poses for use as features. The LSTM technique was used to learn the deep features of human walking and emotional movements, and information about joint angles, areas and distances was combined for classification of these features using a random forest algorithm. However, this recognition framework based on the LSTM and random forest algorithms could not recognise different levels of the same human motion.

Another study presented an STV-GCN human action recognition framework that integrated the ST-GCN and KNN algorithms to create a model that could recognise different levels of the same human motion. This framework used PoseNet [22] technology to acquire human skeleton keypoint information and calculated the change in the keypoint displacements between consecutive image frames, using the skeleton keypoint information to create a speed level classification model. This model used the maximum variation in the displacement of the keypoints of the human skeleton as a feature for training of the model. In another approach, an STV-GCN motion recognition framework used a ST-GCN model to classify human walking speeds and emotions into four categories: happy, angry, sad and fearful. The KNN algorithm was used to classify the speed of motion into three categories: fast, medium and slow. However, whether this involves the classification of different human motion categories or hierarchical classification within the same category, the issue of action timing sequential changes must be taken into consideration, as the KNN algorithm cannot effectively learn the correlation features between timing sequential data in classification.

3. LONG SHORT-TERM MEMORY NETWORK BASED ON SPATIO-TEMPORAL FEATURES FOR HUMAN ACTIVITY RECOGNITION

3.1 System Overview

The present study focuses on improving the accuracy of human motion recognition for different degrees of the same human motion. To this end, we utilise the average rate-of-motion feature of the plural backbone to enhance the learning capability of the model in terms of spatio-temporal features. The architecture of our system is shown in Figure 1. We use OpenPose skeleton keypoint recognition technology to capture keypoint information and the ST-GCN human motion recognition framework to train a model to identify basic categories of motion. Since the ST-GCN human motion recognition framework cannot

identify small differences between the same type of motion, we use the average movement rate of the plural backbone as a deep spatio-temporal feature, which allows the LSTM recognition framework to train and identify models at the advanced motion level. Finally, we integrate the results of the ST-GCN and LSTM recognition models using Support Vector Machine (SVM) with model stacking to enable the network to recognise both the type and the level of human movement.

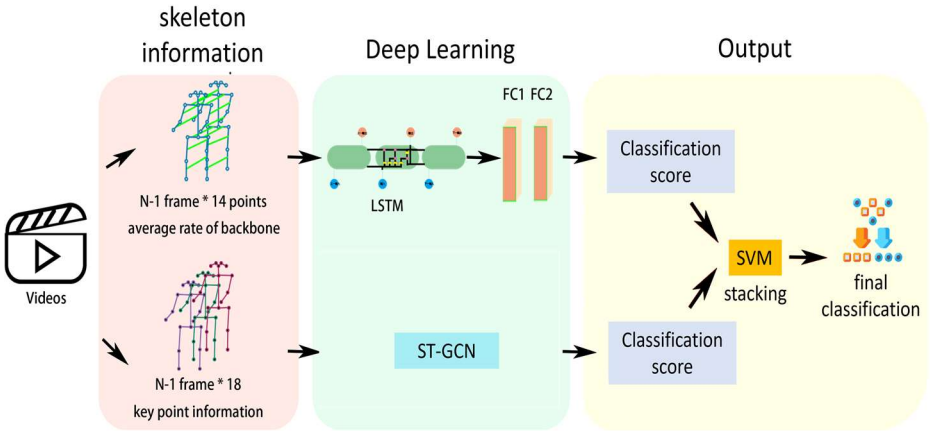

Fig. 1: System architecture

Skeleton point	Body part	Skeleton point	Body part
0	Nose	10	Right knee
1	Neck	11	Right ankle
2	Right shoulder	12	Left hip
3	Right elbow	13	Left knee
4	Right wrist	14	Left ankle
5	Left shoulder	15	Right eye
6	Left elbow	16	Left eye
7	Left wrist	17	Right ear
8	Pelvis	18	Left ear
9	Right hip		

Fig. 2: Human skeleton keypoints used by OpenPose to capture information

In this study, continuous motion videos were cut and processed to obtain separate frames, and OpenPose human skeleton keypoint recognition technology was used to identify location information on 19 keypoints. As shown in Figure 2, these were the right and left eyes, right and left ears, nose, neck, right and left shoulders, right and left elbows, right and left wrists, right and left hips, pelvis, right and left knees, and right and left ankles. These key points of the human skeleton were uniquely numbered for subsequent use in training and identification of human movements. A total of 18 points after removing the pelvis keypoint were used for training as features of a basic motion category model using an ST-GCN human movement recognition framework. We calculate the average backbone movement rate between adjacent frames for keypoints 0 to 14. The complex backbone average motion rate calculated from the above keypoints is used as a feature, and is passed to the LSTM recognition framework for advanced model training and recognition. In this paper, we use SVM to perform stacking to achieve integration of the ST-GCN and LSTM models. For classification categories that involve different levels of movement, the ST-

GCN framework is used to classify the actions, whereas for classification categories that involve similar levels of movement, the LSTM framework is used for action classification. This method is used to obtain a human motion recognition system that includes both basic categories of motion and the levels of activity.

3.2 Deep Learning Using Spatio-temporal Features

In this study, we use the average movement rate of the plural backbone as a deep learning spatio-temporal feature to improve the recognition accuracy of the overall human motion and level of movement. Pseudocode for the algorithm used to calculate the average movement rate of the plural backbone is shown in Figure 3. The required keypoint-related functions are shown in Figures 4 to 6. When the keypoint information is obtained by Openpose, it is important to avoid misjudgement of the keypoints of the human skeleton in certain continuous images, which could result in training and identification errors by the human motion recognition model. Hence, when keypoints in continuous images exceed a reasonable offset range (outliers), we calculate the average of these keypoints in the previous and next frames as a replacement, to reduce misjudgements that can affect model training and recognition.

Deep learning algorithm for plural backbone average movement rate

```
Input: video
Output: final_classification
Requires: function calculate_neck_keypoint()
Requires: function calculate_pelvis_keypoint()
Requires: function calculate_other_keypoints_keypoint()
Requires: Human skeleton key point extraction technology openpose
Requires: ST-GCN model
Requires: LSTM model
Requires: SVM model
begin:
    keypoints = openpose(video)
    if the value is missing or incorrect:
        correct_keypoints = Correct using the previous and next frame data
    if the current key is a neck key:
        A = calculate_neck_keypoint(correct_keypoints)
    if the current key is a pelvis key:
        B = calculate_pelvis_keypoint(correct_keypoints)
    if the current key is other key:
        C = calculate_other_keypoints_keypoint(correct_keypoints)
    bone_data = combined_data(A, B, C)
    ST_GCN_score = ST-GCN(keypoints)
    LSTM_score = LSTM(keypoints)
    stacked_data = stack_data(LSTM_score, ST_GCN_score)
    final_classification = SVM(stacked_data)
return final_classification
```

Fig. 3: Deep learning algorithm for plural backbone average movement rate

The skeleton keypoint information from the continuous images described above was used for training of our ST-GCN human movement recognition model, and the spatio-temporal features of the average movement rate of the skeleton between the adjacent keypoints were calculated to train the LSTM recognition model. Finally, using model stacking technology, the outputs of the ST-GCN and LSTM are stacked, and are used as the input to the SVM model for final classification. Different models may give varying results in different situations, and model stacking technology can help to combine the advantages of several models. By combining the outputs of ST-GCN and LSTM through SVM, the best classification results can be obtained. The average spatial and temporal characteristics of the backbone movement rate were calculated based on the critical points of the skeleton adjacent to the neck, as shown in Figure 4. The average rate of backbone movement based on these key points (neck and left shoulder, neck and right shoulder, and neck and pelvis) were calculated anteriorly and posteriorly, respectively. These distances were then used to find the rate of movement of these four keypoints between frames. The spatial and temporal characteristics of the average movement rate of the backbone between the keypoints of the neck and left shoulder were obtained by averaging the movement rates of consecutive images of the neck and left shoulder. The same process was performed for the neck and right shoulder and the neck and pelvis, to give the spatio-temporal characteristics of the average movement rates of these keypoints.

Calculation of the spatio-temporal characteristics of the average movement rate of the backbone based on the keypoints adjacent to the neck

```
Input: keypoints
Output: Backbone speed
Requires: Left shoulder keypoint LS
Requires: Right shoulder keypoint RS
Requires: Neck keypoint N
Requires: pelvis keypoint P
def defcalculate_neck_keypoint():
    if the current key is a neck key:
        LS_bone = [(LS_{i+1} - LS_i) + (N_{i+1} - N_i)]/2
        RS_bone = [(RS_{i+1} - RS_i) + (N_{i+1} - N_i)]/2
        P_bone = [(P_{i+1} - P_i) + (N_{i+1} - N_i)]/2
return LS_bone, RS_bone, P_bone
```

Fig. 4: Calculation of the spatio-temporal characteristics of the average movement rate of the backbone based on the keypoints adjacent to the neck

The average spatial and temporal characteristics of the backbone movement rate between the key points of the human skeleton adjacent to the hips are calculated as shown in Figure 5. The average rate of backbone movement based on the key points of the human skeleton, hip and left hip and hip and right hip, was calculated anteriorly and posteriorly respectively. We then calculated the Euclidean distance between consecutive frames for three key points: the pelvis and the left and right hips. These distance values were used to find the movement rates of these three key points between consecutive frames, and the movement rates of the pelvis and left hip between consecutive frames were averaged to obtain the average spatial and temporal characteristics of the rate of movement of the backbone between these key points. The same process was applied to the pelvis and the

right hip to obtain the spatio-temporal characteristics of the average rate of movement of the backbone based on these keypoints.

Calculation of the spatio-temporal characteristics of the average movement rate of the backbone based on the key points adjacent to the pelvis

1 Input: *keypoints*
2 Output: *Backbone speed*
3 Requires: Left hip keypoint *LH*
4 Requires: Right hip keypoint *RH*
5 Requires: pelvis keypoint *P*
6 def **defcalculate_pelvis_keypoint()**:
7 if the current key is a pelvis key:
8 $LH\_bone = [(LH_{i+1} - LH_i) + (P_{i+1} - P_i)]/2$
9 $RH\_bone = [(RH_{i+1} - RH_i) + (P_{i+1} - P_i)]/2$
10 **return** *LH_bone*, *RH_bone*

Fig. 5: Calculation of the spatio-temporal characteristics of the average movement rate of the backbone based on the key points adjacent to the pelvis

Calculation of the spatial and temporal characteristics of the average backbone movement rate based on other adjacent key points

1 Input: *keypoints*
2 Output: *Backbone speed*
3 def **defcalculate_other_keypoints_keypoint()**:
4 if the current key is other key:
5 $num\_bone = \{[(k_{i+1} - k_i) + [(k+1)_{i+1} - (k+1)_i]\}/2$
6 **return** *num_bone*

Fig. 6: Calculation of the spatial and temporal characteristics of the average backbone movement rate based on other adjacent key points

The spatio-temporal characteristics of the average backbone movement rate between other adjacent keypoints are shown in Figure 6. The Euclidean distance was calculated between consecutive frames for the non-neck and hip keypoints and the higher numbered keypoints. These distance values were used to represent the rate of movement of the keypoints between three consecutive frames. The inter-frame movement rates of consecutive images for these above two skeleton keypoints were averaged to obtain the spatial and temporal characteristics of the average backbone movement rates between the two keypoints. Since the four keypoints of the human skeleton, including the right wrist, left wrist, right ankle and left ankle, and the next number keypoints of the human skeleton are not backbones, the above processing action is excluded. In summary, a total of 14 average rates of motion of the backbone were obtained in this way and used to enhance the learning of spatio-temporal features.

3.3 Human Activity Recognition

In this study, we use the average movement rate of the plural backbone network as a spatio-temporal feature for deep learning to train the LSTM human motion recognition model, as shown in Figure 7. The LSTM model can learn and remember the correlations between

time series data. Through the use of LSTM, the network can effectively learn time correlations and yield good accuracy. Since our ST-GCN recognition framework can effectively recognise different motion poses, supplemented by using the backbone average motion rate proposed in this paper as features used to train the LSTM human motion recognition model, i.e., different levels of motion can be effectively recognised. We use OpenPose recognition technology to capture the skeleton keypoint information from N consecutive frames K_ij, where i is the image frame and j is the keypoint number. The Euclidean distance of each skeleton keypoint between consecutive frames is calculated as shown in Equation 1. The rate of movement of the human skeleton keypoints between consecutive image frames is then obtained. As shown in Equation 2, a total of (N−1) frames multiplied by the average movement rates of 14 backbones can be obtained, and this information is used as spatio-temporal features for deep learning. This enhances the learning of spatio-temporal features during training of the LSTM human movement recognition model. The method proposed in this paper enhances the learning of spatio-temporal features, allowing the model to learn the degree of temporal variation of the human backbone between successive image frames, as shown in Figure 8. After stacking the prediction results output by the ST-GCN network and the prediction results output by the LSTM network, we will use the stacked output results as the input data of SVM, and extract ST-GCN and LSTM through SVM classification. Their respective advantages are combined to obtain better recognition results.

$$D_{ij} = K_{(i-1)j} - K_{ij} \tag{1}$$

$$S_{ij} = D_{ij} - D_{adjacentkeypoint} \tag{2}$$

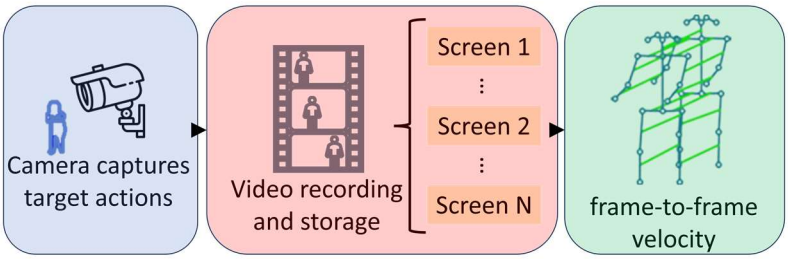

Fig.7: Deep learning from spatio-temporal characteristics

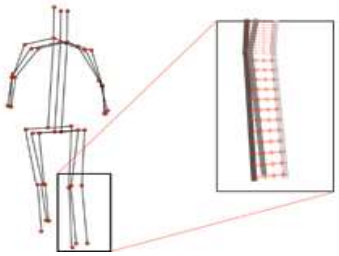

Fig. 8: Temporal changes in the human keypoints between consecutive image frames

4. EXPERIMENTAL RESULTS

In this paper, two datasets are used to evaluate the classification accuracy of our model, in terms of both the motion itself and the level of activity. The first of these datasets contains videos of humans carrying out a squatting motion, which is used to train and identify one correct and four incorrect squat motion recognition models. The second dataset is used with the human emotion motion dataset to train and identify four emotion and three speed motion recognition models. Finally, the accuracy of our system is compared with that of ST-GCN, LSTM and STV-GCN. The system tested here was built on a CPU i9 10900, a GPU of RTX 3070, with the Windows 10 operating system. For the ST-GCN human action recognition framework, we set the number of epochs to 300, the batch size to 32, the test batch size to 32, and the base learning rate to 0.1, and the stochastic gradient descent algorithm was used as an optimiser. During training, the learning rate of the framework was multiplied by 0.1 after every 30 epochs, and the learning rate was decreased up to five times. For the LSTM human action recognition framework, we set the number of epochs to 300, the batch size to 64, and the base learning rate to 0.001, and the Adam algorithm was used as a framework optimiser. For the stacked model, we use the SVM from the scikit-learn suite. Its core uses poly and the value of C is set to one.

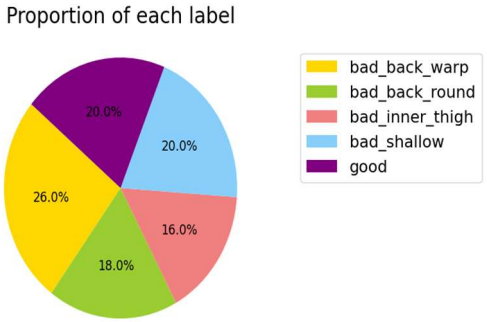

Fig. 9: Proportion of sample labels for the human squatting dataset

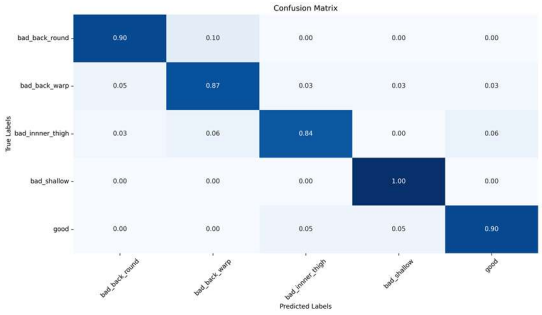

Fig. 10: Confusion matrix for recognition of squatting movements

The first experiment involved the human squat dataset, which was used to identify human actions consisting of the same movement but at different levels. The proportion of samples in each category is shown in Figure 9. This dataset contained 3,856 video files, each 10 s in length, and each video showed a human target performing between one and

five squats. Our system used OpenPose to extract continuous human skeleton keypoint information from each video for 30 frames multiplied by 10 s. In this dataset, the items were divided into a total of five categories, which were labelled 'Good' for perfectly correct squats, 'Bad shallow' for insufficient squat depth, 'Bad inner thigh' for inner thigh swing during the squat, and 'Bad back warp' and 'Bad back round' for incorrect states of back flexion. In this paper, for the human deep squatting motion dataset, the plural average backbone motion rate was used as a feature for learning and recognition, and deep learning was performed for human motion recognition of the same motion but different motion levels. The confusion matrix of the experimental results is shown in Figure 10. The results were sufficient to verify that the proposed method could effectively improve the recognition accuracy for different levels of the same movement. For the perfectly correct deep squat, squat, inner thigh swing and two incorrect backbends, the accuracy rate exceeded 80% for all five categories, with an average accuracy of 90.81%. As shown in Figure 11, we compared the proposed method with ST-GCN and LSTM in terms of recognition accuracy, and the results were 90.81%, 86.5% and 82.0% for the same movements at different levels of activity, respectively. Our experimental results also show that using the average motion rate of the plural backbone as a feature for deep learning can effectively improve the accuracy of human action recognition by at least 4%.

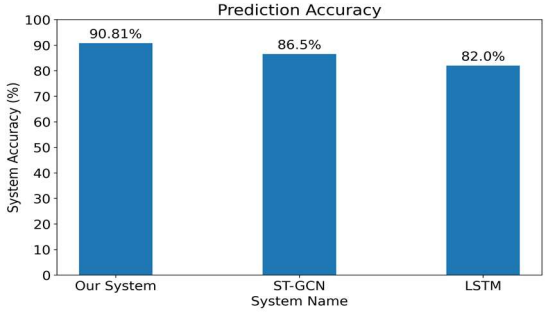

Fig. 11: Comparison of performance on the squatting database

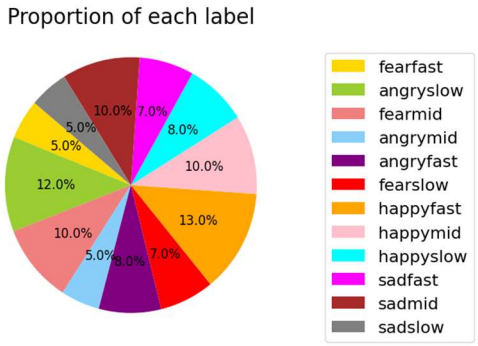

Fig. 12: Proportion of labelled samples in the human emotion activity dataset

The second experiment in this paper used a human emotional action dataset to identify different levels of the same action. The dataset was divided into four emotions and three action speeds, giving a total of 12 labels indicating each emotion and action speed.

The proportion of samples in each category is shown in Figure 12. There were 360 video files in the dataset, each of which was 1 s long and showed one human target walking a fixed distance with a specific emotion. Using OpenPose's capture technique, continuous human skeleton keypoint information was extracted from each video, and a total of 29 frames of human skeleton keypoint information for 1 s were obtained. In this experiment, plural skeleton average motion rate features were used to enhance the learning from a human emotional action dataset. At the same time, we use 10 videos for each small label as samples for testing. A confusion matrix of the results from our system in terms of recognition of human actions at different levels is shown in Figure 13. From these results, it can be seen that the proposed method can effectively improve the recognition accuracy of different levels of the same action, with an average accuracy of 83.33%. The recognition accuracy of our proposed method was compared with that of LSTM and STV-GCN, and the results are shown in Figure 14. The recognition accuracies for the same action but different levels of activity were 83.33%, 70.0% and 75.0%, respectively. From these results, it can be observed that using the plural backbone average motion rate as a feature for deep learning improves the accuracy of human motion recognition by at least 8%.

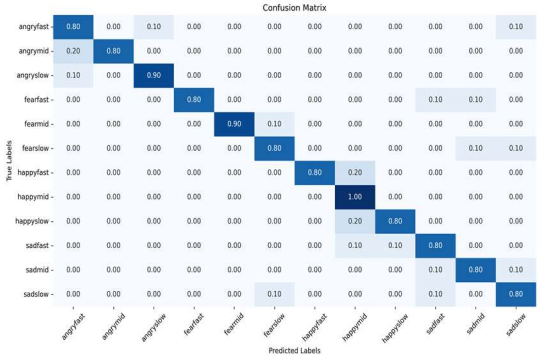

Fig. 13: Confusion Matrix for Human Emotional Action Recognition

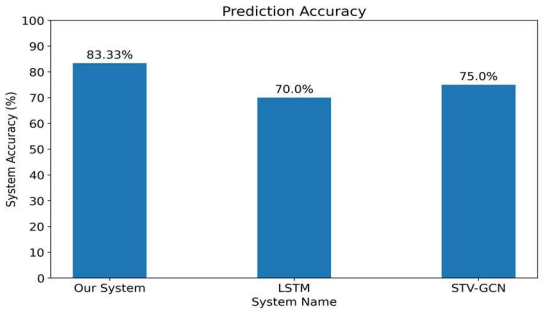

Fig. 14: Comparison of performance for human emotional action recognition

5. CONCLUSIONS

To solve the problem in which current human motion recognition frameworks cannot efficiently train models to identify different levels of the same motion, the average movement rate of the backbone was implemented as a deep learning spatio-temporal feature to improve the overall accuracy of human motion recognition. Our network was

trained to identify basic types of movement and advanced movement levels using ST-GCN and LSTM human movement recognition frameworks. The results from these two models were integrated and stacked to obtain predictions for the final category of motion and the level of activity. A performance analysis of the proposed method with ST-GCN, LSTM and STV-GCN human motion recognition models showed that on a dataset of videos of squatting motions, our scheme outperformed the ST-GCN and LSTM models in terms of accuracy by at least 4%, and on a dataset containing videos of emotional walking, it outperformed LSTM and STV-GCN by at least 8%.

REFERENCES

1. H. Song, X. Han, C. Montenegro-Marin and S. krishnamoorthy, Secure Prediction and Assessment of Sports Injuries using Deep Learning based Convolutional Neural Network, Springer Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 3399-3410, 2021.
2. C. Dindorf, E. Bartaguiz, F. Gassmann and M. Frohlich, Conceptual Structure and Current Trends in Artificial Intelligence, Machine Learning, and Deep Learning Research in Sports: A Bibliometric Review, MDPI International Journal of Environmental Research and Public Health, vol. 20, no. 1, pp. 1-23, 2023.
3. M. Tsai and C. Chen, Enhancing the Accuracy of a Human Emotion Recognition Method Using Spatial Temporal Graph Convolutional Networks, Springer Multimedia Tools and Applications Journal, vol. 82, pp. 11285-11303, 2023.
4. A. Buerkle, W. Eaton, N. Lohse, T. Bamber and P. Ferreira, EEG based Arm Movement Intention Recognition towards Enhanced Safety in Symbiotic Human-Robot Collaboration, Elsevier Robotics and Computer-Integrated Manufacturing Journal, vol. 70, pp. 1-9, 2021.
5. W. Taylor, S. Shah, K. Dashtipour, A. Zahid, Q. Abbasi and M. Imran, An Intelligent Non-Invasive Real-Time Human Activity Recognition System for Next-Generation Healthcare, MDPI Sensors Journal, vol. 20, no. 9, pp. 1-20, 2020.
6. A. Dzedzickis, A. Kaklauskas and V. Bucinskas, Human Emotion Recognition: Review of Sensors and Methods, MDPI Sensors Journal, vol. 20, no. 3, pp. 1-40, 2020.
7. Y. Liu, X. Jiang, T. Sun and K. Xu, 3D Gait Recognition Based on a CNN-LSTM Network with the Fusion of SkeGEI and DA Features, IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 1-8, 2019.
8. J. Kim, K. Lee, J. Kim and S. Hong, Patient Identification based on Physical Rehabilitation Movements using Skeleton Data, IEEE International Conference on Information and Communication Technology Convergence, pp. 1572-1574, 2021.
9. J. Liu, N. Akhtar and A. Mian, Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition, IEEE Conference on Computer Vision and Pattern Recognition, pp. 10-19, 2017.
10. T. Ahmad, L. Jin, L. Lin and G. Tang, Skeleton-based Action Recognition using Sparse Spatio-temporal GCN with Edge Effective Resistance, Elsevier Neurocomputing Journal, vol. 423, pp. 389-398, 2021.
11. M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang and Q. Tian, Multiscale Spatio-Temporal Graph Neural Networks for 3D Skeleton-Based Motion Prediction, IEEE Transactions on Image Processing Journal, vol. 30, pp. 7760-7775, 2021.
12. Y. Li, R. Xia, X. Liu and Q. Huang, Learning Shape-Motion Representations from Geometric Algebra Spatio-Temporal Model for Skeleton-Based Action Recognition, IEEE International Conference on Multimedia and Expo, pp. 1066-1071, 2019.

13. F. Sardari, A. Paiement, S. Hannuna and M. Mirmehdi, VI-Net - View-Invariant Quality of Human Movement Assessment, MDPI Sensors Journal, vol. 20, no. 18, pp. 1-15, 2020.
14. R. Ogata, E. Simo-Serra, S. Iizuka and H. Ishikawa, Temporal Distance Matrices for Squat Classification, IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2533-2542, 2019.
15. T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera and D. Manocha, Identifying Emotions from Walking Using Affective and Deep Features, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-15, 2019.
16. M. Tsai and C. Chen, Spatial Temporal Variation Graph Convolutional Networks (STV-GCN) for Skeleton-Based Emotional Action Recognition, IEEE Access Journal, vol. 9, pp. 13870-13877, 2021.
17. Z. Cao, T. Simon, S. Wei and Y. Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291-7299, 2017.
18. B. Pavlyshenko, Using Stacking Approaches for Machine Learning Models, IEEE International Conference on Data Stream Mining and Processing, pp. 255-258, 2018.
19. N. Kourentzes, D. Barrow and S. Crone, Neural Network Ensemble Operators for Time Series Forecasting, Elsevier Expert Systems with Applications, vol. 41, no. 9, pp. 4235-4244, 2014.
20. L. Yu, S. Wang and K. Lai, Forecasting Crude Oil Price with an EMD-based Neural Network Ensemble Learning Paradigm, Elsevier Energy Economics, vol. 30, no. 5, pp. 2623-2635, 2008.
21. S. Yan, Y. Xiong and D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition, Association for the Advance of Artificial Intelligence, pp. 7444–7452, 2018.
22. A. Kendall, M. Grimes and R. Cipolla, PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization, IEEE International Conference on Computer Vision, pp. 2938-2946, 2015.

Biography

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 a Full Professor with the Department of Electronic Engineering, National United University, Taiwan. His current research interests include Artificial Industrial Internet of Things Technologies, Artificial Intelligence, Machine Learning, Deep Learning, Vehicular Communications and Multimedia Communications.

Shih-Yung Cheng received his Bachelor's degree from the Department of Electronic Engineering, National United University, Taiwan. He is currently pursuing further studies in the Department of Electronic Engineering at National United University, Taiwan. His current research interests include Artificial Internet of Things, Artificial Intelligence, Smart Manufacturing Application, Machine Learning, and Deep Learning.