TNTC: TWO-STREAM NETWORK WITH TRANSFORMER-BASED COMPLEMENTARITY FOR GAIT-BASED EMOTION RECOGNITION

Chuanfei Hu[1,2] *Weijie Sheng*[1,2] *Bo Dong*[3] *Xinde Li*[1,2‡]

[1] Key Laboratory of Measurement and Control of CSE Ministry of Education, School of Automation, Southeast University, Nanjing, China

[2] Nanjing Center for Applied Mathematics, Nanjing, China

[3] College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, China

{cfhu, wjsheng, xindeli}@seu.edu.cn, bodong.cv@gmail.com

ABSTRACT

Recognizing the human emotion automatically from visual characteristics plays a vital role in many intelligent applications. Recently, gait-based emotion recognition, especially gait skeletons-based characteristic, has attracted much attention, while many available methods have been proposed gradually. The popular pipeline is to first extract affective features from joint skeletons, and then aggregate the skeleton joint and affective features as the feature vector for classifying the emotion. However, the aggregation procedure of these emerged methods might be rigid, resulting in insufficiently exploiting the complementary relationship between skeleton joint and affective features. Meanwhile, the long range dependencies in both spatial and temporal domains of the gait sequence are scarcely considered. To address these issues, we propose a novel two-stream network with transformer-based complementarity, termed as TNTC. Skeleton joint and affective features are encoded into two individual images as the inputs of two streams, respectively. A new transformer-based complementarity module (TCM) is proposed to bridge the complementarity between two streams hierarchically via capturing long range dependencies. Experimental results demonstrate that TNTC outperforms state-of-the-art methods on the latest dataset in terms of accuracy.

Index Terms— Gait-based emotion recognition, complementarity, convolutional neural network, transformer

1. INTRODUCTION

Human emotion recognition, based on visual cues, has been widely applied in various intelligence applications, such as video surveillance [1], behavior prediction [2, 3], robot navigation [4] and human-machine interaction [5]. Facial expression is one of the most predominant visual cues [6] to be used for recognizing the human emotions including anger, disgust, happiness, sadness, fear and other combinations. However, facial expressions may be unreliable in complex situations, for instance, imitation expressions [7] and concealed expressions [8]. Therefore, recent researches gradually focus on the other visual cues of human to perceive the emotions, such as a gait of human in a walking [9, 10].

Recent efforts have been made towards improving gait-based emotion recognition [11, 12, 13, 14, 15], which can be categorized as sequence-based, graph-based and image-based methods. The paradigm of sequence-based methods is to construct a sequence deep model, such as Gate Recurrent Unit (GRU) and Long Short-term Memory (LSTM), based on skeleton sequences to predict the emotion [11, 12]. The insight of graph based-methods is to utilize Spatial Temporal Graph Convolutional Network (ST-GCN) to represent the inherent relationship between joints, since the skeleton is naturally structured as a graph in non-Euclidean geometric space [13, 14]. The image-based methods cast the sequence classification as an image classification via encoding the skeleton sequences, while Convolutional Neural Network (CNN) is constructed to extract hierarchical features for recognizing the emotions [15]. Although these methods achieve promising results in human emotion recognition, there are two major drawbacks. Firstly, the aggregation of joints and affective features might be rigid, resulting in exploiting the complementary information insufficiently. Furthermore, long range dependencies in both spatial and temporal domains are ignored, which are important to depict the implicit relationships between skeleton joints for human poses [16].

To address the above issues, we focus on the image-based method, and a novel two-stream network with transformer-based complementarity, termed as TNTC, is proposed for recognizing the human emotions. We argue that spatial and temporal information of the skeleton sequences can be effectively extracted via CNN in the image domain. Meanwhile, transformers are utilized to handle the problem of capturing long range dependencies via the self-attention mechanism whose effectiveness has been verified in many computer vision tasks [17, 18, 19, 20]. Specifically, skeleton joints are first encoded into an image, while affective features are converted to another image via a novel encoded method. Two individual

‡ Corresponding author: Xinde Li (xindeli@seu.edu.cn).

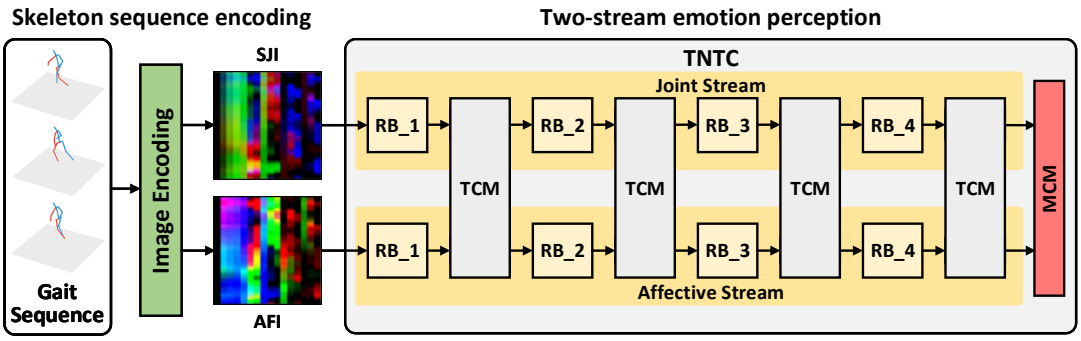

Fig. 1. The overall framework of the proposed method. "RB_id" denotes the "id"-th residual block of ResNet-34.

CNNs are built to extract the hierarchical representations separately, and then transformer-based complementarity module (TCM) is proposed to bridge the complementarity between two streams in each feature level via capturing long range dependencies. The probabilities of human emotions are predicted by the multi-layer perceptron (MLP)-based classification module (MCM) at the end of the method. In summary, the main contributions are as follows:

- We propose a novel method for gait-based human emotion recognition, learning the deep features from images encoded by skeleton joints and affective features. To the best of our knowledge, we are among the first to represent affective features as an image for gait-based human emotion recognition.
- Transformer-based complementarity module (TCM) is exploited to complement the information between skeleton joints and affective features effectively via capturing long range dependencies.

2. METHODOLOGY

The overall framework of our proposed method consists of skeleton sequence encoding and two-stream emotion perception, as illustrated in Fig. 1. Skeleton joints and affective features are first constructed via a skeleton sequence of gait, and encoded into skeleton joint image (SJI) and affective feature image (AFI), respectively. Next, TNTC based on CNNs is modeled to extract hierarchical features whose complementary information are replenished cross TCMs. MCM is conducted at the end of the network to identify the emotion category.

2.1. Skeleton sequence encoding

Skeleton joint image (SJI). The motivation of encoding the skeleton sequence as an image is to take full advantage of CNN to compactly extract the local spatial-temporal features. Inspired by [15], we encode the 3D coordinates of joints as three channel of the image. Specifically, a skeleton sequence is given as follows:

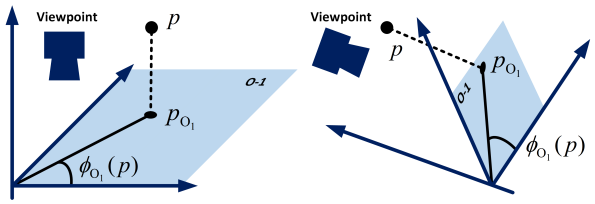

Fig. 2. The visualization of invariant $\phi_{\mathbf{o}_i}(p)$ under arbitrary viewpoints.

$$S = \left\{ P^t \in \mathbb{R}^{N_s \times D_s} \middle| t = 1, 2, \ldots, T; D_s = 3 \right\}, \quad (1)$$

where P^t presents coordinates of skeleton joints in the t-th frame. D_s denotes the dimension of coordinates, and the number of joints is denoted as N_s.

Then, we arrange P^t according to the order of frames, and SJI is encoded as follows:

$$\mathbf{M_J} = \left[P^1, P^2, \ldots, P^T \right], \quad (2)$$

where $\mathbf{M_J} \in \mathbb{R}^{T \times N_s \times D_s}$, and $[\cdot]$ is an operation to concatenate the coordinates of joints in the temporal dimension.

Affective feature image (AFI). Besides the skeleton joints, affective features, such as posture and movement, convey the emotion information in the gaits [9] which are essential to involve the prediction for affective state of the subject. Here, we merely consider the joint angles as the affective features, and discard other characteristics, such as distances and velocities of joints. The reason is that the formulations of such characteristics can be approximated as special cases of a low-order combination which are implicit in the convolution operations of joint stream.

To construct the reliable AFI, we utilize the projection angles of joints on three planes to avoid the inconsistency caused by various viewpoints. As shown in Fig. 2, the projection angles of joints on three planes are invariant with arbitrary viewpoints obviously. Formally, the projection angle of the n-th joint $P_n^t \in \mathbb{R}^{D_s}$ on the projected planes can be computed as follows:

$$Q_n^t = \left[\phi_{\mathbf{o}_1}(P_n^t), \phi_{\mathbf{o}_2}(P_n^t), ..., \phi_{\mathbf{o}_i}(P_n^t) \right] \quad (3)$$

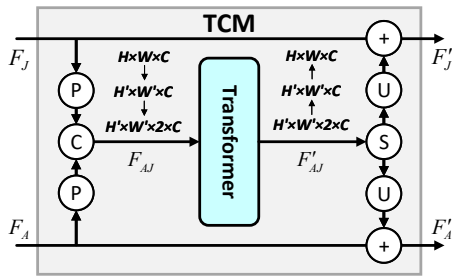

Fig. 3. The details of TCM in TNTC. "P", "C", "S", and "U" represent average pooling operation, concatenation operation, splitting operation, and upsampling operation, respectively. Element-wise summation is denoted as "+".

where $\phi_{\mathbf{o}_1}(P_n^t)$ is a function to compute a projection angle of joint P_n^t on plane O-1, and $\mathbf{o}_1$ is belonged to $\mathbf{O}$ which is a set of projection operators $\{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_i | i = 1, 2, \ldots, D_s\}$. Since P_n^t is represented via 3D coordinates, $D_s = 3$ and $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}$ can be instantiated as followed:

$$\mathbf{o}_1 = \big[e_1, e_2\big], \mathbf{o}_2 = \big[e_1, e_3\big], \mathbf{o}_3 = \big[e_2, e_3\big], \tag{4}$$

where $e_1, e_2, e_3 \in \mathbb{R}^3$ are unit vectors. For a joint $p \in \mathbb{R}^3$, the function $\phi_{\mathbf{o}_i}(p)$ can be formulated as follows:

$$\begin{aligned} p_{\mathbf{o}_i} &= p^{\mathsf{T}} \mathbf{o}_i, \\ \phi_{\mathbf{o}_i}(p) &= arctan \frac{p_{\mathbf{o}_i(2)}}{p_{\mathbf{o}_i(1)} + \epsilon}, \end{aligned} \tag{5}$$

where $p_{\mathbf{o}_i(1)}$ represents the 1-st value of the coordinate on the projected plane O-i, and ϵ is a small positive infinitesimal quantity to avoid the invalid denominator.

Finally, the angles of the skeleton joints P^t in the t-th frame can be formulated as follows:

$$Q^t = \Big[Q_1^t, Q_2^t, \ldots, Q_{N_s}^t\Big], \tag{6}$$

where $Q^t \in \mathbb{R}^{N_s \times D_s}$. AFI constructed via the projection angles can be denoted as follows:

$$\mathbf{M_A} = \Big[Q^1, Q^2, \ldots, Q^T\Big]. \tag{7}$$

where $\mathbf{M_A} \in \mathbb{R}^{T \times N_s \times D_s}$, and $[\cdot]$ is an operation to concatenate the projection angles of joints in the temporal dimension.

2.2. TNTC

Backbone. The backbone of TNTC is ResNet-34 whose capability of feature representations has been proved in many vision tasks [21, 22]. We discard the fully connected layer of ResNet-34, and retain the four level residual blocks to extract hierarchical features, as shown in Fig. 1. Assuming an input encoded map with a size of 224×224, the size and channel of output feature extracted via four level residual blocks are 7×7 and 512, respectively.

TCM. After each residual block, TCM is modeled based on a vanilla transformer architecture [17] and a series of operations, as shown in Fig. 3. The transformer architecture takes as input a sequence consisting of discrete tokens, each represented by a feature vector. The long range dependencies between feature vectors are captured, and the feature vectors supplemented by positional encoding to incorporate positional inductive biases. Specifically, given joints feature F_J and affective feature F_A extracted via residual blocks in a level, the size of F_J and F_A with $H \times W \times C$ are first reduced via average pooling operations to $H' \times W' \times C$, where $H' = H/l$, $W' = W/l$, and scale parameter l is set to different values to modify the features with a fixed size in each level, since processing features with transformer at high spatial resolutions is computationally expensive. Then, two reduced features are concatenated to $F_{AJ} \in \mathbb{R}^{H' \times W' \times 2 \times C}$ as input of transformer, which is an encoder structure completely described in [17] and its implementation details are discussed in Section 3. After capturing long range dependencies between F_J and F_A via transformer, F'_{AJ} is arranged to $H' \times W' \times 2 \times C$ and divided into two complementary information F'^c_J and F'^c_A with dimensions of $H' \times W' \times C$ for joint stream and affective stream, respectively. Bilinear interpolation is utilized as upsampling operation to resize the complementary information to the same size of F_J and F_A. Finally, element-wise summation is leveraged to aggregate the complementary information and corresponding stream, respectively.

MCM. At the end of joint and affective streams, joint and affective features are reduced to a size of $1 \times 1 \times 512$ via average pooling, and are combined by element-wise addition as the final feature vector with 512-dimension. The feature vector is fed to MLP with two hidden layers and softmax function to classify the emotion of the skeleton sequence.

3. EXPERIMENTS

3.1. Experimental Setup

Dataset. We evaluate the proposed method on Emotion-Gait [13] dataset, which consists of 2177 real gait sequences separately annotated into one of four emotion categories including happy, sad, angry, or neutral. The gait is defined as the 16 point pose model, and the steps of gait sequences are maintained via duplication to 240 which is the maximum length of gait sequence in the dataset.

Evaluation protocol. We employ 5-fold cross-validation to evaluate the proposed method, where the sample numbers of each category are divided in the same ratio among the folds. Accuracy is adopted as the metric defined as follows:

$$Accuracy = T/S, \tag{8}$$

where T denotes to the number of successfully classified gait sequences with corresponding categories, and S denotes the number of test samples. The average accuracy of 5-fold cross-validation is recorded along its standard deviation.

Table 1. Comparison of our method with the state-of-the-art on Emotion-Gait. The best result of accuracy is highlighted in **bold**.

Method		Accuracy %
Graph-based	STEP [13]	77.65(0.87)
	G-GCSN [14]	80.31(0.92)
Sequence-based	LSTM (Vanilla) [12]	75.38(0.98)
	TEW [11]	81.89(0.69)
Image-based	ProxEmo [15]	80.33(0.85)
	TNTC (**Ours**)	**85.97(0.75)**

Table 2. Ablation analysis of our method on Emotion-Gait.

Method					Accuracy %
TNTC w/o TCMs (Baseline)					81.53(0.88)
Baseline w/o joint stream					79.52(0.69)
Baseline w/o affective stream					80.71(0.71)
Levels	1	2	3	4	
Baseline with TCMs	✓				83.27(0.79)
	✓	✓			84.09(0.91)
	✓	✓	✓		84.13(0.85)
	✓	✓	✓	✓	85.97(0.75)

Implementation Details. The experiments are conducted on a work station with an NVIDIA RTX 2080Ti GPU. The proposed method is implemented based on PyTorch deep learning framework. The main hyperparameters consist of the network and training stage. For the hyperparameters of TCMs in the network, we stack 2 transformers and 4 attention heads for each TCM. The dimensions of feature embedding are 64, 128, 256, and 512 for corresponding TCMs, while the values of each level scale parameter l are 8, 4, 2, and 1. In the training stage, stochastic gradient descent (SGD) is used to optimize the learnable parameters with a momentum of 0.9 and a weight decay of 5e-4. The total epochs are 300 and initial learning rate is 1e-3, where the decay ratio is 0.1 every 75 epochs. SJI and AFI are resized into 224×224 via bilinear interpolation, and the size of mini-batches is 64.

3.2. Comparisons with the State-of-the-art

We compare our method with 5 state-of-the-art methods lately reported on Emotion-Gait including sequence-based [11, 12], graph-based [13, 14], and image-based [15] methods. To come up with a fair comparison, we reproduce these methods by public codes and report the experimental results with the same evaluation protocol, as listed in Tab. 1. We can observe that the proposed method achieves a superior performance than all the other methods.

3.3. Ablation studies

Effectiveness of two-stream architecture. To clarify the effectiveness of two-stream architecture, we separately train joint stream and affective stream as two independent networks. Meanwhile, we construct TNTC without TCMs as a baseline network. As shown in Tab. 2, the performances of individual streams are obviously lower than the baseline. The results demonstrate the effectiveness of two-stream architecture which implies the complementary relationship between joint and affective features.

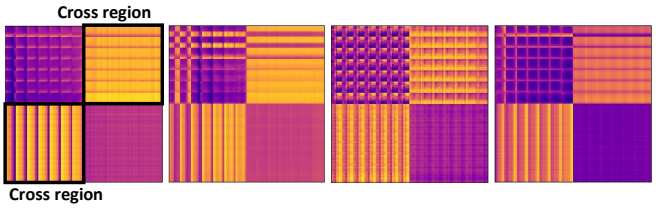

Fig. 4. The visual interpretation of complementary information. The average attention maps in TCMs of different instances are ordered by four categories from happy, sad, angry to neutral. The higher heat values in the cross regions (**bold** squares) represent qualitatively that more complementary information between two streams can focused by TCM.

Effectiveness of TCMs. To confirm the effectiveness of TCMs, we gradually insert TCM at each level based on the baseline, and report the performances of the networks in Tab. 2, respectively. It can be observed that the accuracy of the baseline is improved as the number of TCM increases. Furthermore, we plot the attention maps of TCMs to reveal the capability of representing complementary information. As shown in Fig. 4, the high scores of the attention maps focus on the cross regions between skeleton joints and affective features, which interpret visually the complementary information between two streams represented via TCMs.

4. CONCLUSION

In this paper, we propose a novel method for gait-based human emotion recognition, modeled via a two-stream network with transformer-based complementarity (TNTC). Skeleton joints and affective features of gait sequence are encoded into images as the inputs of the two-stream architecture. Meanwhile, the importance of complementary information between two streams is revealed, which can be represented effectively via the proposed transformer-based complementarity module (TCM). Experimental results demonstrate that the proposed method achieves the superior performance over state-of-the-art methods on the latest gait-based emotion dataset.

5. REFERENCES

[1] Nasim Hajati and Amin Rezaeizadeh, "A wearable pedestrian localization and gait identification system using kalman filtered inertial data," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–8, 2021.

[2] Sylmarie Dávila-Montero, Jocelyn Alisa Dana-Lê, Gary Bente, Angela T. Hall, and Andrew J. Mason, "Review and challenges of technologies for real-time human behavior monitoring," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 1, pp. 2–28, 2021.

[3] Zhe Wang, Yongxiong Wang, Chuanfei Hu, Zhong Yin, and Yu Song, "Transformers for eeg-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors J.*, pp. 1–1, 2022.

[4] Aniket Bera, Tanmay Randhavane, and Dinesh Manocha, "Modelling multi-channel emotions using facial expression and trajectory cues for improving socially-aware robot navigation," in *CVPRW*, 2019, pp. 257–266.

[5] Rui Zhang, Zhenyu Wang, Zhenhua Huang, Li Li, and Mengdan Zheng, "Predicting emotion reactions for human–computer conversation: A variational approach," *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 4, pp. 279–287, 2021.

[6] Shan Li and Weihong Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2020.

[7] Laurie Mondillon, Paula M Niedenthal, Sandrine Gil, and Sylvie Droit-Volet, "Imitation of in-group versus out-group members' facial expressions of anger: A test with a time perception task," *Social neuroscience*, vol. 2, no. 3-4, pp. 223–237, 2007.

[8] Stephen Porter and Leanne Ten Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychological science*, vol. 19, no. 5, pp. 508–514, 2008.

[9] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Yi Guo, Victor Leung, Jun Cheng, and Bin Hu, "Emotion recognition from gait analyses: Current research and future directions," *arXiv:2003.11461*, 2020.

[10] Weijie Sheng and Xinde Li, "Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network," *Lect Notes Comput Sc*, vol. 114, pp. 107868, 2021.

[11] Uttaran Bhattacharya, Christian Roncal, Trisha Mittal, Rohan Chandra, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha, "Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping," in *ECCV*, 2020, pp. 145–163.

[12] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha, "Identifying emotions from walking using affective and deep features," *arXiv:1906.11884*, 2019.

[13] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," *AAAI*, vol. 34, no. 02, pp. 1342–1350, Apr. 2020.

[14] Yuan Zhuang, Lanfen Lin, Ruofeng Tong, Jiaqing Liu, Yutaro Iwamoto, and Yen-Wei Chen, "G-gcsn: Global graph convolution shrinkage network for emotion perception from gait," in *ACCVW*, 2020, pp. 46–57.

[15] Venkatraman Narayanan, Bala Murali Manoghar, Vishnu Sashank Dorbala, Dinesh Manocha, and Aniket Bera, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," in *IROS*, 2020, pp. 8200–8207.

[16] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li, "Learning trajectory dependencies for human motion prediction," in *ICCV*, 2019, pp. 9488–9496.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[18] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang, "Transreid: Transformer-based object re-identification," *arXiv:2102.04378*, 2021.

[19] Yangyundou Wang, Hao Wang, Yiming Li, Chuanfei Hu, Hui Yang, and Min Gu, "High-accuracy, direct aberration determination using self-attention-armed deep convolutional neural networks," *J. Microsc.*, vol. n/a, no. n/a, 2022.

[20] Yangyundou Wang, Zhaosu Lin, Hao Wang, Chuanfei Hu, Hui Yang, and Min Gu, "High-generalization deep sparse pattern reconstruction: feature extraction of speckles using self-attention armed convolutional neural networks," *Opt. Express*, vol. 29, no. 22, pp. 35702–35711, Oct 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, June 2016, pp. 770–778.

[22] C. Hu and Y. Wang, "An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images," *IEEE Trans. Ind. Electron.*, vol. 67, no. 12, pp. 10922–10930, 2020.