Emotion Recognition Based on Body and Context Fusion in the Wild

Yibo Huang
Sichuan University
Chengdu, China
2018222050141@stu.scu.edu.cn

Hongqian Wen
Sichuan University
Chengdu, China
2020222050177@stu.scu.edu.cn

Linbo Qing
Sichuan University
Chengdu, China
qing_lb@scu.edu.cn
{Corresponding Author}

Rulong Jin
Sichuan University
Chengdu, China
jinrulong@stu.scu.edu.cn

Leiming Xiao
Sichuan University
Chengdu, China
2020222050192@stu.scu.edu.cn

Abstract

Emotion recognition in-the-wild under uncontrolled conditions is a challenge, because facial expression is often blurred or even missing in the public space, while the previous visual emotion recognition researches have mainly focused on facial expression. In this paper we present a learning-based algorithm for emotion recognition by utilizing posture and context information, aiming to realize emotion recognition based on video in the wild. The network is designed in a three-branch architecture, including three feature streams: body, skeleton and context streams. The three streams are then fused to predict dimensional emotion representation, valence, arousal, and dominance. In addition, a new Body and Context Emotions Dataset (BCEmotion) is captured in the wild and labeled to support the related research, to tackle the lack of datasets based on public space video including complete individuals with face blurs and occlusions. With the BCEmotion dataset, we trained the proposed model that jointly analyses body and context of videos to realize emotion recognition in the wild. Experimental results show that proposed method effectively integrates emotional information expressed by body and context, and has good generalization ability and applicability in public space video data.

1. Introduction

Emotion recognition has attracted researchers' attention in recent years, which aims to recognize how the person feels. Perceiving the emotions of people is a vital ability of humans in daily life. If computers have this ability to perceive and analyze human emotions and intentions, they will play an important role in various applications. For example, intelligent robots that can recognize emotions will bring better interactive experience for people. Medical assistance system with emotion recognition model can help assess mental disorders such as anxiety and depression. Emotional monitoring in airports, subways, parks and other places with large traffic will help identify potential threats and deal with emergencies in a timely manner. However, there are still many challenges to accurately recognize emotions in the absence of facial expressions.

Facial expression can directly reflect emotional state of people, traditional research of human emotion based on vision mainly focuses on the face[20]. However, different from the cropped and aligned facial images in the dataset, in the real environment, distance, posture, occlusion and other factors will have a great impact on the recognition of facial expression, which may lead to the blurring or disappearance of facial features. Similarly, the features that are difficult to obtain in public space such as audio and physiological signals[6] are not suitable for our research.

Psychological research reveals the importance of body posture in emotional expression[22]. Darwin first described the relationship between body language and emotion. In addition to the face, the body also includes the head, hands and feet, trunk and other parts. Among them, hand is an important part to show body language information[24]. The head can also reveal information about emotions. For example, when people are interested in something, they tilt their head close to it. The trunk also shows directional cues, straight or back, stiff or bent, all are clues to the emotion state.

In addition to the characteristic extracted from the person directly, context also plays a very crucial role in the understanding of the perceived emotion[23]. Many researches have shown the importance of considering the context for recognizing people's emotions, proving that context can provide additional emotional clues[17, 18, 23, 29]. Especially in the public space, the background is complex and contains abundant information, making good use of context information can better identify emotions.

In order to realize the application of emotion recognition closer to life, we use the body posture and context which are easily acquired in public space and contain a large number of emotional features to analyze. We present a learning-based network for emotion recognition, which is designed in a three-branch architecture, including three feature streams: body, skeleton and context streams. The three streams are then fused to predict dimensional emotion representation, valence, arousal, and dominance. In addition, due to the lack of video dataset including background and complete individuals in the existing public emotion recognition datasets, we build a novel dataset, BCEmotion dataset, by collecting videos captured in multiple real-world settings of people and annotating the ground-truth continuous emotion dimensions. The contributions in this paper can be summarized in the following three aspects:

1) A learning-based multimodal emotion recognition algorithm is proposed, which is designed in a three-branch architecture, including three feature streams: body, skeleton and context streams. In order to make emotion recognition systems work for real-life scenarios, we chose the body posture and context for analysis.

2) A 3D convolutional neural network with appropriate depth is proposed to extract the emotional information contained in the scene information, which can avoid over fitting in the training process of the network and ensure the generalization performance of the model, reducing the overall parameters of the network.

3) A new dataset BCEmotion is collected for emotion recognition in the wild. To the best of our knowledge, there exist very few datasets based on public space videos including complete individuals to support the related research. BCEmotion is a collection of video clips captured in multiple real-world. The videos have about 3961 clips annotated with emotion labels.

2. Related work

2.1. Emotion recognition based on body posture

Posture information including tilt direction, body openness and the position of arm, shoulder and head contributes to the recognition of emotional state. However, the research on emotion recognition using posture information is relatively late than facial expression recognition, and there are relevant studies in psychology and emotion calculation[4, 8, 10]. Crenn *et al.* [7] obtained low-level features from 3D skeleton sequence frames, and decomposed the features into three categories: geometric features, motion features and Fourier features, then calculated the meta features of these low-level features, and finally sent them to the classifier for classification. Ranganathan *et al.* [26] produced a multimodal data set (emoFBVP), which includes face, body movement, sound and physiological signals. Four different deep belief networks (DBN) were used to extract four features, and then fused the features to predict the final emotion. Uttaran *et al.* [2, 3] used semi-supervised network and synthetic skeletons to make full use of the data for gait emotion recognition. The research of using posture to judge emotion is still in the stage of continuous development.

2.2. Emotion recognition based on context

Before 2017, there were few researches on using context information to perceive and recognize human emotions, and the reason is that there is no appropriate scene related emotional dataset. Kosti *et al.* [16, 17] used the image data of people in uncontrolled scenes to build EMOTIC dataset, and designed a dual channel network for emotion recognition. Zhang *et al.* [29] used the context information in the image to construct an emotion map to infer the emotional state of people. Ruan *et al.* [27] proposed a new network structure, called Context-Aware Generation-Based Net (CAGBN). This network structure can consider both the whole image and the details of the target person, and has achieved good recognition results on EMOTIC dataset. Bendjoudi *et al.* [1] used Xception network[5] to extract body features and improved VGG16 to extract context features of image data, and fused the two features to recognize emotional state. Lee *et al.* [18] collected a large amount of video clips from TV shows to form a new dataset CAER, and designed a dual stream coding network to extract face and context features for emotion recognition. More and more researches have taken scene information as an important supplement for emotion recognition. After adding context features, the accuracy of individual emotion recognition has been effectively improved, which shows that context is an important part of emotion recognition.

2.3. Emotion recognition datasets

Most of the emotion recognition datasets in the past have been taken in lab-controlled environments and only focused on a single modality. For example, CK+[21], MMI[25] and Oulu-CASIA[30] are datasets that focus on the facial expressions collected in lab settings, which are different from the real situation. Some work also focuses on multi-task emotion analysis of face modality[12, 13, 14, 15]. The context-focused emotional dataset captured in public space is lacking[23], especially the video datasets containing complete individuals. EMOTIC dataset[16] is a collection of images from other datasets and networks. These images include different activities in the real environment and express rich emotional information. CAER dataset[18] is a collection of video-clips from TV shows with 7 discrete emotion annotations. However, most of the video clips focus on the actor's face and upper body, and do not contain a complete individual, which is different from the study of in this paper. GroupWalk dataset[23] consists of 45 video clips captured in multiple real-world settings of people walking in dense crowd settings. The dataset contains four categories of emotions, which are less fine-grained for the emotions of individuals in the public space. In order to meet the needs of this paper for emotional analysis of people in public space, a dynamic data set containing scene information and complete individual information in public space is needed.

3. Proposed Method

3.1. Motivation and Overview

In this section, we describe an effective framework for video emotion recognition in the wild. Limited by the external conditions in the public space scene, not all individuals can efficiently obtain the information often used for emotion recognition, such as facial expression, audio and so on. In order to study the emotional state of people in public space under general circumstances, we extract features contributing to emotions from the body posture of individuals, including key points of bones and appearance information, and the context, and then realize emotion recognition by combining the above various features. Figure 1 shows the overall framework of emotion recognition algorithm based on real life video.

Concretely, the network structure is mainly composed of three sub network structures, which are AS-GCN[19], 3D-ResNet101[11] and Plain-3D for extracting emotional features from skeleton key points information. The input of AS-GCN network is normalized skeleton data of 16 consecutive frames. Thanks to the characteristics of residual structure, 3D-ResNet101 has very good feature extraction ability, which is very suitable for emotion feature extraction of individual appearance sequence. For scene feature extraction, we design a convolutional neural network Plain-3D with moderate depth, which is composed of five convolution layers and five pooling layers. The features extracted from the above three sub network structures will be fused in the full connection layer, and then classified to get the corresponding emotions.

3.2. Feature Extraction Module

GCN In more and more applications, data is generated from non-Euclidean domains and represented as graph structures with complex relationships, which brings great challenges to existing machine learning algorithms. In addition, for non-Euclidean data such as skeleton key point sequence, the information on the timeline also contains many characteristics related to behavior and emotion. Therefore, in order to obtain the spatial and temporal features of the skeleton key point sequence, Li *et al.* [19] used AS-GCN to construct a general skeleton map to capture more abundant dependencies between joint points. Specifically, a data-driven approach is used to infer Actional Links (A-links) to capture potential dependencies between any connections. Structural Links (S-links) not only considers the nodes connected with the central node, but also considers multiple nodes close to the central node, so as to extract more abundant features, as shown in Figure 1. Based on A-links and S-links, Actional-Structural Graph Convolution (ASGC) for extracting spatial features is formed. In order to obtain the temporal features between frames, AS-GCN uses one layer of Temporal Convolution (T-CN), which extracts the temporal features of each joint independently while sharing the weight of each joint. ASGC and T-CN are combined to form AS-GCN block, which can be used to extract the emotion information contained in the skeleton key point sequence, so as to improve the accuracy of emotion recognition.

3D-ResNet With the deepening of convolution neural network, gradient vanishing, gradient explosion and degradation of recognition accuracy will follow. ResNet effectively solves the above problems through residual structure. However, the convolution kernel of ResNet is two-dimensional and cannot be applied to temporal tasks. Hara *et al.* [11] used 3D convolution and 3D pooling to extend the existing ResNet network structure, and got 3D-ResNet network structure to adapt to more abundant feature extraction tasks. 3D-ResNet takes 16 consecutive RGB images as input, and the specific input dimension is $112 \times 112 \times 3 \times 16$. In the process of network reasoning, the pooling operation will down sample the data in the time dimension, and learn the spatiotemporal characteristics of the input data. The public space video emotion recognition algorithm designed in this paper uses 3D-ResNet to extract the emotion information from the individual appearance sequence.

Plain-3D In order to extract the emotional information contained in the context, this paper designs a 3D convolutional neural network Plain-3D with moderate depth, as shown in Figure 1. Plain-3D consists of five convolution layers, five pooling layers and three fully connected layers. The size of convolution core is $3 \times 3 \times 3$, which is inspired by the construction method of VGG network, replacing the relatively large convolution core with a small convolution core such as 3×3. There are two advantages, the first is to reduce the amount of parameters of the whole network, the second is to add the ReLU[9] activation function between multiple convolution operations and introduce more nonlinear factors to make the extracted features more discriminative. When there are multiple individuals in a scene, there will be multiple individuals sharing similar context in-

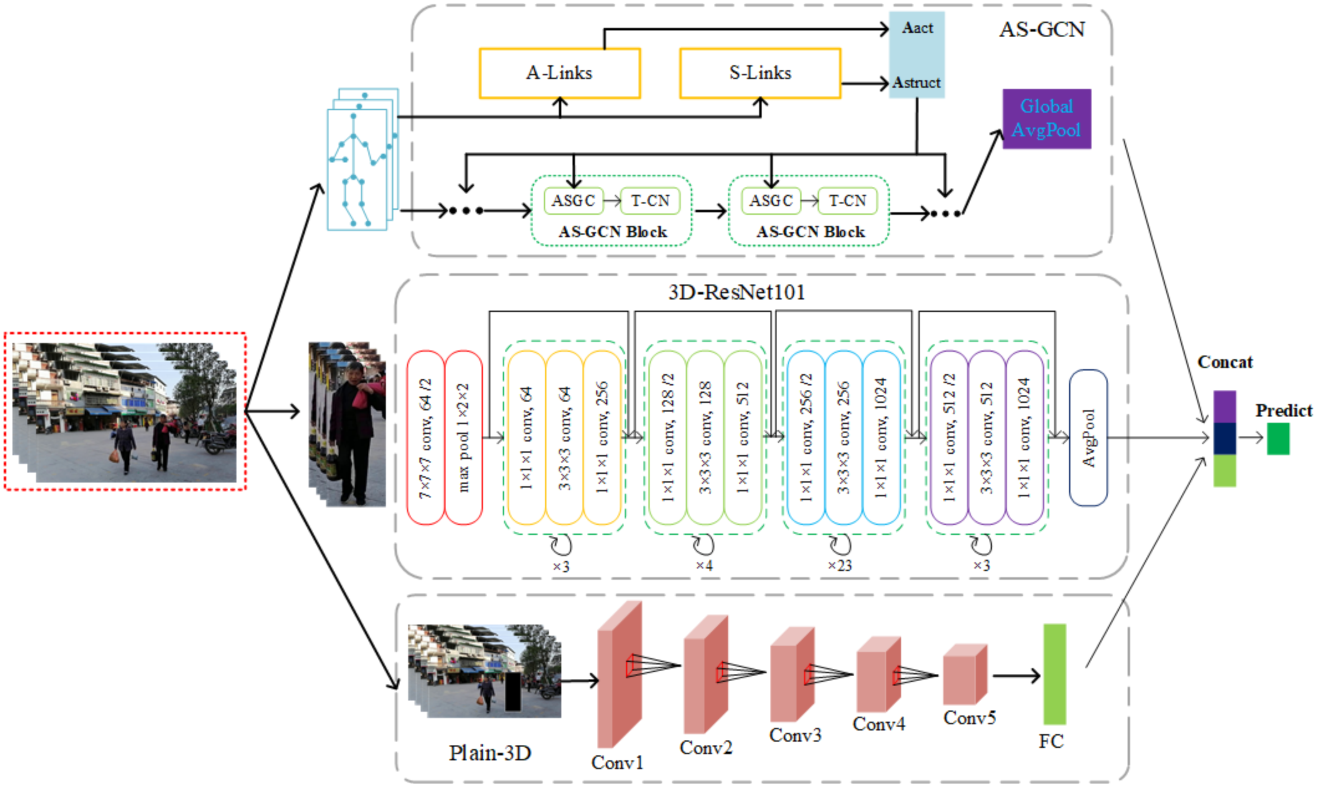


Figure 1. The overall framework of emotion recognition algorithm based on public space video, consisting of three sub network structures.

formation. In order to avoid the network overfitting in the training process and ensure the generalization performance of the model, the network structure of Plain-3D is not very deep.

4. BCEmotion Dataset

4.1. Data Source

There is a lack of datasets based on public space videos including complete individuals, so a new Body and Context Emotions Dataset (BCEmotion) is captured in real public space and labeled to support the related research. The examples of dataset are shown in Figure 2.

The data sources in the dataset established in this paper mainly include two aspects: first, the data is shot on the spot by video equipment; second, the video data that meets the conditions (real world, complete body, enough time) is obtained on the Internet. In the process of field shooting, more than 15 researchers filmed 10-minute videos at different locations in nine cities (no researchers were in the videos). The original data was filtered and edited, and the edited data contained a total of 3961 individual video clips, each of which contained 16 frames of images.

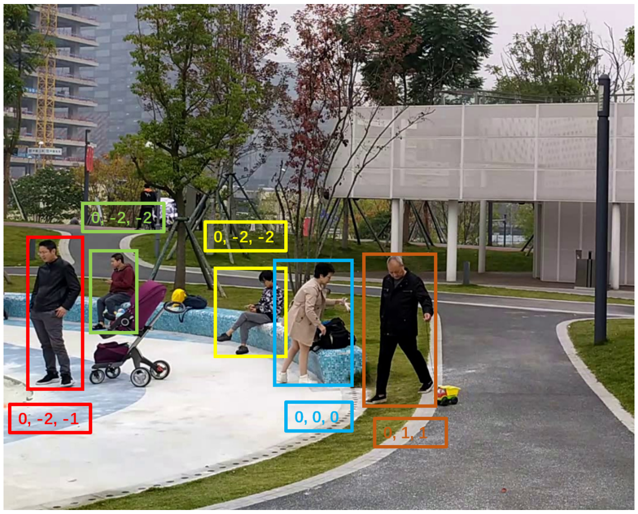
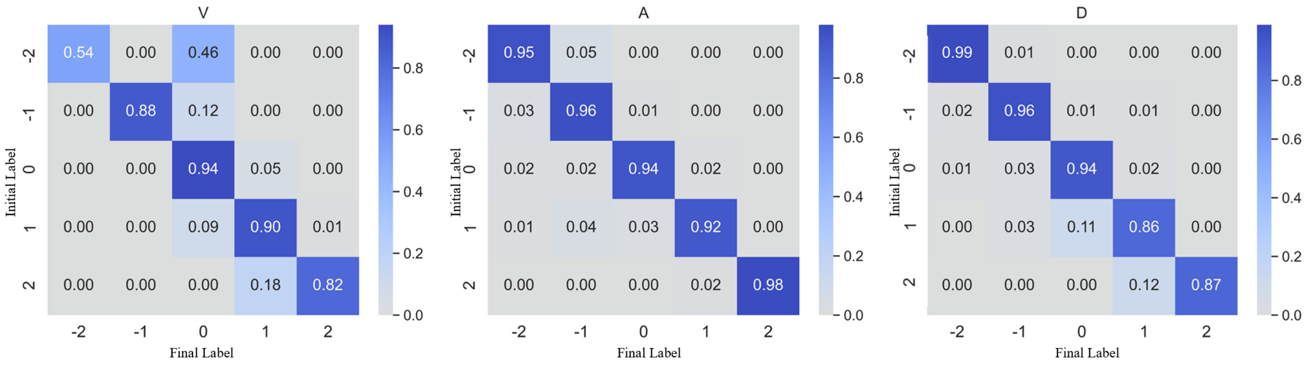


Figure 2. Some examples of BCEmotion dataset. The three dimensions from left to right are V, A and D.

4.2. Label Production

In the real wild scene, people will control and restrain their emotions, for example, emotions like fear and disgust are less. Dimensional emotion model is more suitable to

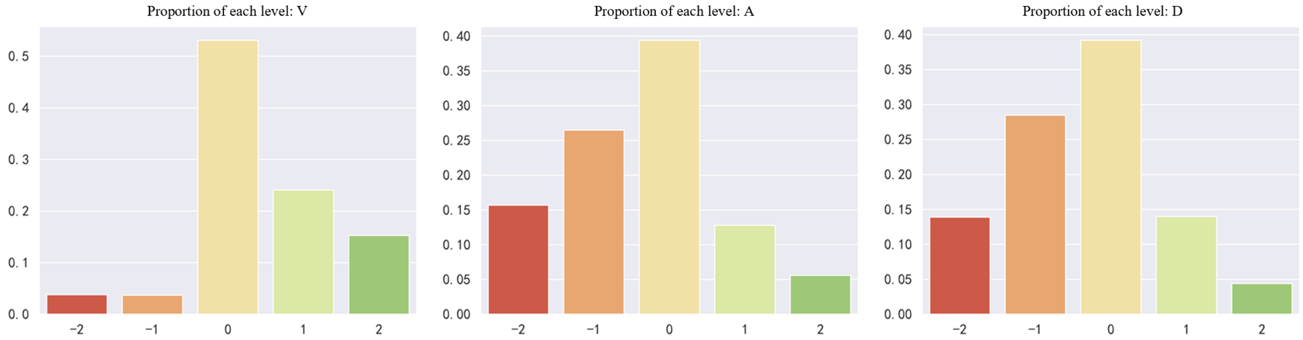


(a)

(b)

(c)

Figure 3. The comparison matrix of initial and final labels of dimensional emotion, (a) (b) and (c) correspond to the comparison matrix of dimensions V, A and D respectively.



(a)

(b)

(c)

Figure 4. Sample proportions of different levels in dimensional emotion, (a) (b) and (c) correspond to V, A and D respectively.

evaluate the emotional state of individuals in the scene of public space. Specifically, this paper divides three different dimensions of VAD into five levels, respectively representing the measurement of the state degree of each dimension of individual emotion. There are altogether 14 staff participating in the marking work, among which 9 are male and 5 are female. Before the marking work starts, the marking staff will undergo unified training. A total of three taggers annotate each video clip, and the final tag adopts the majority rule. The initial labels obtained after the first round of labeling are statistically compared with the final labels obtained after the third round of labeling. The comparison matrix is shown in Figure 3, the overall consistency is high.

4.3. Data Statistics

In this paper, the samples of different levels in each dimension of VAD are statistically analyzed and displayed in a bar chart, as shown in Figure 4. As can be seen from the figure, no matter for dimension V, A or D, the sample proportion of grade 0 is the largest. V, especially dimensions level 0 samples of more than 0.5, and the dimension of 2, 1 sample proportion is relatively small, the whole data reflects the imbalance. This is related to people's restraint on their emotions in the wild. Individuals in the public space are more inclined to display their neutral and positive emotions while suppressing their relatively negative emotions. Therefore, the statistical results of data are consistent with the reality.

4.4. Privacy Protection

When obtaining the videos, we took it into account and complied with relevant laws and privacy protection policy. Meanwhile, we also concentrate on the general practice about how to release a dataset. For example, WoodScape[28] was released with original data and a license agreement which enforces the users to strictly adhere to the General Data Protection Regulation(GDPR). We will take a similar approach as well. The dataset with the annotations for training and testing will be made freely available to academic and non-profit organizations for non-commercial, scientific use under the premise of privacy protection.

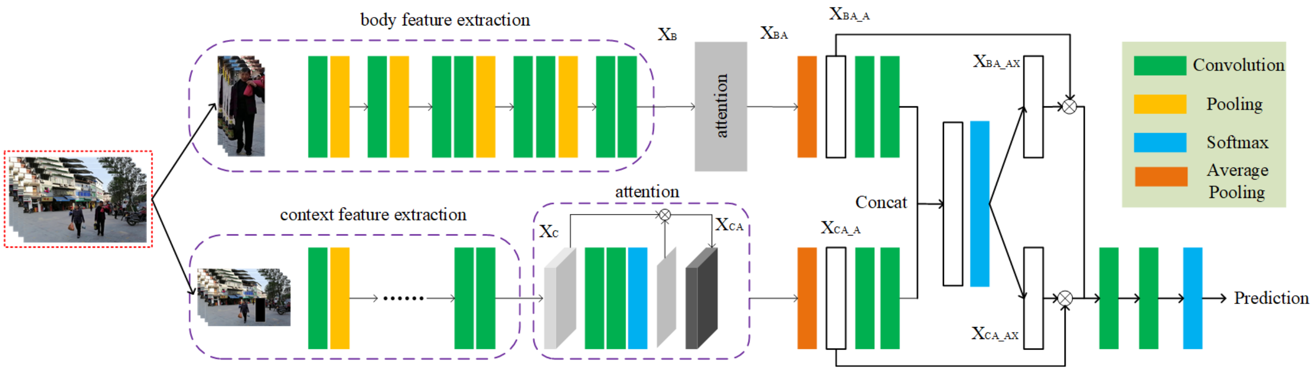


Figure 5. The overall framework of baseline network, consisting of two-stream networks and fusion network.

5. Experiments

5.1. Experimental Settings

Model is implemented with PyTorch library. Model training is divided into two stages. Firstly, the AS-GCN is trained, and the Aact and Astruct models are obtained by unsupervised learning. The models are trained for 1000 epochs on NVIDIA RTX 2080ti, the batch size is 64, and the Adam optimization algorithm is used. The learning rate is fixed at 0.0005. Then the Aact and Astruct models are used to complete the training process of AS-GCN. Each sample will first generate the corresponding data through the previously trained model, and then use Adam optimization algorithm to train on NVIDIA RTX 2080ti. The initial learning rate is 0.1, with a total of 100 epochs. Every 20 training epochs will reduce the learning rate to 0.1 times of the original.

Before the start of the second stage of training, the model generated by AS-GCN training is added to the complete algorithm as a pretraining model. It is worth noting that in order to avoid the influence of context information and appearance information on AS-GCN pretraining model in the process of back propagation, the parameters of AS-GCN pretraining model are frozen in the second stage of training. The initial learning rate is 0.0001. Every 20 epochs of training will reduce the learning rate to 0.1 times of the original. The optimizer uses Adam. In this paper, the accuracy of five-level classification of dimensional emotions is used to measure the performance of the corresponding algorithm.

5.2. Qualitative Results

In this paper, the baseline network was designed by referring to the attention module in the CAER-Net[18]. The baseline network uses the selected region of bounding box in the dataset and the context information after the bounding box region was erased as the input of the two channels. The overall framework of baseline network structure is shown in Figure 5, which is mainly composed of four modules, including body feature extraction module, context feature extraction module, attention module, and mixed attention module.

Algorithm	V	A	D
Baseline	0.7103	0.6729	0.7477
Ours	**0.7259**	**0.7103**	**0.8100**

Table 1. Quantitative evaluation of the proposed network in comparison to baseline network on the BCEmotion dataset.

The experimental results of baseline network and our algorithm are shown in Table 1. Compared with the baseline method, our algorithm uses AS-GCN to extract the emotional features of skeleton key points sequence, which greatly improves the accuracy of dimension A and dimension D. The main reason is that dimension A and dimension D are the individual's activation degree and the individual's degree of control over the scene or others respectively, which are closely related to the body information, while the skeleton key information reflects the body posture information. As an individual's happiness level, dimension V is more closely related to human facial information, while skeleton key information is completely unable to reflect human facial conditions, so the improvement of dimension V is limited.

The confusion matrix of our algorithm is shown in Figure 6. As can be seen from Figure 6 (a), the recognition accuracy of samples with level 0 in dimension V is higher than 0.9, while the recognition rate of samples with other levels is relatively low, and they are easy to be confused with samples of adjacent levels. The main reason may be that in public space, the scene is more complex and the face information is blocked. There will be problems in the process of label making and feature extraction, which leads to a high false recognition rate. Figure 6 (b) is the confusion matrix of dimension A. It can be seen from the figure that no matter which level of data is, it is easy to be confused

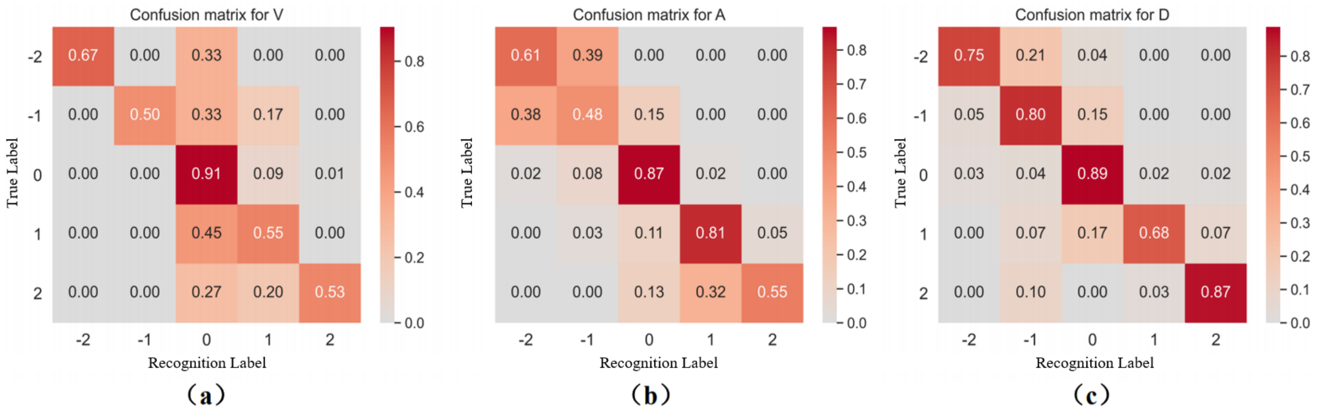


Figure 6. The confusion matrix of experimental results of the proposed network, (a) (b) and (c) correspond to the confusion matrix of dimensions V, A and D respectively.

Method	V	A	D
AS-GCN	0.6573	0.6012	0.6293
AS-GCN+Plain-3D	0.6791	0.6137	0.6822
3D-ResNet101	0.6978	0.6542	0.7726
3D-ResNet101+AS-GCN	0.7196	0.6573	0.7757
3D-ResNet101+Plain-3D	0.7134	0.6978	0.7819
3D-ResNet101+AS-GCN+Plain-3D	**0.7259**	**0.7103**	**0.8100**

Table 2. Ablation study of the proposed network on the BCEmotion dataset.

with the data of one or two adjacent levels. Of course, this is related to the continuity of human motion state in public space. Figure 6 (c) shows the high recognition accuracy of dimension D. In general, the algorithm in this paper has a good effect on human emotion recognition in public space scenes, which proves the effectiveness of the algorithm.

5.3. Ablation Experiments

The network structure of the algorithm in this paper is composed of three sub-network structures. In order to explore the contribution of each sub-network to the final emotion recognition result, the ablation experiments are carried out, as shown in Table 2.

As can be seen from the table, the accuracy obtained by using skeleton sequence as the data source in the experiment is lowest. The introduction of context information on the basis of skeleton information effectively improves the recognition results of each dimension, which proves the importance of context information for individual emotion recognition. Adding skeleton information and context information improves the accuracy on the premise of appearance sequence as input. Relatively speaking, the improvement of experimental results by adding skeleton information is limited, which may be because appearance sequence also expresses skeleton information to a certain extent. Finally, the experimental results are the best when the three sub-network structures are fused, which shows that various information related to human emotions in the public space can be complementary, and the accuracy of emotion recognition can be effectively improved by fusing various information to comprehensively judge the emotional state of individuals.

6. Conclusion

In this paper we present a learning-based algorithm for emotion recognition by utilizing posture and context information, aiming to realize emotion recognition based on video in real-life scenarios. The network consists of three feature streams: body, skeleton and context streams. The three streams are then fused to predict dimensional emotion, valence, arousal, and dominance. We also release BCEmotion dataset, which is captured in the wild and includes complete individuals, for emotion recognition. In the future, we will further study more flexible strategies and more scales of information to respond to the diversity of contexts and environments, in a more refined way to perceive and identify individual emotional states.

Acknowledgement

This research was supported by the National Nature Science Foundation of China under Grant 61871278 and the Sichuan Science and Technology Program under Grant 2018HH0143.

References

[1] Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, and Fadi Dornaika. Multi-label, multi-task cnn approach for context-based emotion recognition. *Information Fusion*, 2020.

[2] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1342–1350, 2020.

[3] 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 *European Conference on Computer Vision*, pages 145–163. Springer, 2020.

[4] Gabriel Castillo and Michael Neff. What do we express without knowing? emotion in gesture. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 702–710, 2019.

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[6] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

[7] Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer, and Saida Bouakaz. Body expression recognition from animated 3d skeleton. In *International Conference on 3D Imaging (IC3D)*, pages 1–7. IEEE, 2016.

[8] Beatrice De Gelder. Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3):242–249, 2006.

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[10] M Melissa Gross, Elizabeth A Crane, and Barbara L Fredrickson. Effort-shape and kinematic assessment of bodily expression of emotion during gait. *Human movement science*, 31(1):202–221, 2012.

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[12] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[13] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[14] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[15] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[16] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1667–1675, 2017.

[17] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019.

[18] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[19] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3595–3603, 2019.

[20] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[21] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[22] Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision*, 128(1):1–25, 2020.

[23] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[24] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 2018.

[25] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 5–pp. IEEE, 2005.

[26] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[27] Shulan Ruan, Kun Zhang, Yijun Wang, Hanqing Tao, Weidong He, Guangyi Lv, and Enhong Chen. Context-awar generation-based net for multi-label visual emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[28] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek Odea, and Patrick Prez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9307–9317, 2019.

[29] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019.

[30] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.