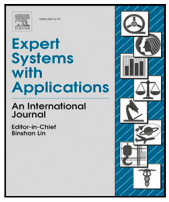
Expert Systems With Applications 237 (2024) 121419

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



# Learning facial expression and body gesture visual information for video emotion recognition

Jie Wei [a,b], Guanyu Hu [a], Xinyu Yang [a,*], Anh Tuan Luu [b], Yizhuo Dong [c]

[a] Xi'an Jiaotong University, China
[b] Nanyang Technological University, Singapore
[c] Xi'an University of Posts and Telecommunications, China

ARTICLE INFO

*Keywords:*
Video emotion recognition
Facial expression
Spatio-temporal features
Body joints
Gesture representation

ABSTRACT

Recent research has shown that facial expressions and body gestures are two significant implications in identifying human emotions. However, these studies mainly focus on contextual information of adjacent frames, and rarely explore the spatio-temporal relationships between distant or global frames. In this paper, we revisit the facial expression and body gesture emotion recognition problems, and propose to improve the performance of video emotion recognition by extracting the spatio-temporal features via further encoding temporal information. Specifically, for facial expression, we propose a super image-based spatio-temporal convolutional model (SISTCM) and a two-stream LSTM model to capture the local spatio-temporal features and learn global temporal cues of emotion changes. For body gestures, a novel representation method and an attention-based channel-wise convolutional model (ACCM) are introduced to learn key joints features and independent characteristics of each joint. Extensive experiments on five common datasets are carried out to prove the superiority of the proposed method, and the results proved learning two visual information leads to significant improvement over the existing state-of-the-art methods.

## 1. Introduction

In recent years, video has gradually replaced image and text as a new socialized way for communication. People tend to share life, hotspots, and new products through videos. Intelligent analysis of emotional state conveyed by video helps to understand the user's feelings and improves services to boost marketing competitiveness (Shukla et al., 2020). In addition, automatic emotion recognition also has a significant impact on human–computer interaction (Chowdary, Nguyen, & Hemanth, 2021; Val-Calvo, Álvarez-Sánchez, Ferrández-Vicente, & Fernández, 2020), educational practices (Wang & Shi, 2021), and intelligent vehicles (Zepf, Hernandez, Schmitt, Minker, & Picard, 2020).

In most scenarios, people express emotion through facial expressions and body gestures simultaneously (e.g., we smile and clap hands when feeling happy, and cry and twist hands when feeling sad). Ambady et al. have shown that facial expressions and body gestures appear to be the most important visual information for emotion recognition (Ambady & Rosenthal, 1992). Current research focuses on facial expression (Abdullah & Abdulazeez, 2021; Farzaneh & Qi, 2021; Li & Xu, 2020; Revina & Emmanuel, 2021; Zhu, Mao, Jia, Noi, & Tu, 2022), and there are few related to body gesture (Fu, Xue, Li, Zhang, & Cai, 2020; Noroozi et al., 2018). Some research considering both information mainly focuses on the innovation of fusion methods to improve the performance of emotion recognition (Shan, Gong, & McOwan, 2007; Yan, Zheng, Xin, & Yan, 2014; Zhang & Zhang, 2015). However, there still exist some limitations that these research did not carefully consider.

First, facial expression sequence data differs from static pictures, which contains not only spatial information but also temporal information. Although deep learning has achieved certain success in video emotion recognition, it is still unclear what is the most effective network architecture for spatio-temporal relationship learning. Deep neural networks for facial expression recognition currently exist in two groups. First, the CNN+RNN architecture is used (Huang et al., 2020; Lamba & Virmani, 2021; Liang, Liang, Yu, & Zhang, 2020), where CNN is used for spatial relationship modeling and RNN for temporal relationship modeling. However, part of the information may be lost due to independent learning spatial and temporal features, and the obtained cascaded spatio-temporal features are not optimal for emotion recognition. Second, 3D convolution (C3D) is also widely used for facial expression recognition (Kumawat, Verma, & Raman, 2019; Ma

* Corresponding author.
*E-mail addresses:* weijie_xjtu@stu.xjtu.edu.cn (J. Wei), guanyu.hu@stu.xjtu.edu.cn (G. Hu), yxyphd@mail.xjtu.edu.cn (X. Yang), anhtuan.luu@ntu.edu.sg (A.T. Luu), dyzhuo@xupt.edu.cn (Y. Dong).

https://doi.org/10.1016/j.eswa.2023.121419
Received 2 August 2022; Received in revised form 23 August 2023; Accepted 31 August 2023
Available online 9 September 2023

et al., 2019; Sun, Zhao, & Jin, 2019). C3D consists of 3D convolutional layers that perform convolution simultaneously in temporal and spatial dimensions to learn spatio-temporal features instead of CNN+RNN consists of two separate neural network structures, the CNN for spatial features and the RNN for temporal features. However, numerous parameters and lower computational efficiency limit the effectiveness and practicality of the model.

Second, for body gesture information, there is insufficient attention given. The few early studies on body gesture (Camurri, Lagerlöf, & Volpe, 2003; Saha, Datta, Konar, & Janarthanan, 2014) mainly focused on designing handcrafted features. However, there are no evident mapping relationships between emotion and these features. As deep learning is rapidly evolving, studies on body gesture-based emotion recognition have shifted to designing deep models (Li, Chen, Wu, Pedrycz, & Hirota, 2021; Sun, Cao, He, & Yu, 2018). These studies built various deep networks to explore emotional cues and learn spatio-temporal features from video frames. However, tremendous computational resources and time are required to deal with original video in deep learning models. Moreover, treating all pixels with equal weights cannot highlight the crucial advantages of body joints movement in emotion understanding.

To address above problems, we take into account the characteristics of facial expressions and body gestures, and propose algorithms to improve recognition accuracy with spatio-temporal features extraction and emotion recognition model construction. For facial expression modality, we propose a super image-based spatio-temporal convolution model (SISTCM), which stacks the video frames into two super images along the width and height axes, thereby applying 2D convolution to capture the local spatio-temporal features of the facial expressions. The convolution kernels of the two super images are shared to learn local spatio-temporal features under different perspectives collaboratively. Besides, considering the progressive relationship of emotion expressions in time, a two-stream LSTM model is introduced to further learn global temporal cues. It uses clip-level emotion representations and local spatio-temporal features as input to get the final recognition result. The whole facial expression recognition framework is an end-to-end model, and this model is optimized by multi-stage supervised learning to enhance the recognition performance.

For body gesture modality, we propose a body gesture representation method based on body joints movement, in which the body gestures are represented with 25 body joints. To be specific, this method first detects the positions of key joints and saves the change information, which can retain vital gesture information to expedite the subsequent processes. Then, aggregate these joints change information in time to capture time-dependent relationship of body gestures. With this representation result, we propose an attention-based channel-wise convolutional model (ACCM) to learn joints features and recognize the emotion. The model is able to preserve the independent properties of each joint under channel-wise convolutional layers, while maximizing the advantage of key features leveraging the attention mechanism.

In the end, we explore different fusion mechanisms, which leverage the respective advantages and complementarities of two visual modalities to maximize the overall performance of emotion recognition.

Our contributions are listed as follows:

(1) We propose a novel feature extraction model SISTCM that applies 2D convolution to capture the local spatio-temporal features of the facial expressions. This model with fewer parameters compared to 3D convolution could improve training speed.
(2) We propose a two-stream LSTM model based on clip-level emotion representations and local spatio-temporal features. The model considers the global temporal relationship of emotional expression and clip-level features, and fuses them for improving the performance of facial expression recognition.
(3) We propose body gesture representation methods. The methods can capture the changes of body gestures by aggregating the key joints information in time, which retains rich temporal relationships while saving subsequent computational cost and time.
(4) We introduce an attention-based channel-wise convolutional model ACCM that can maximize the advantages of key features and preserve the independent properties of each component joint to optimize the emotion recognition performance. The experimental results on common emotion recognition datasets show significant improvement in performance compared to the state-of-the-art models.

We note that a shorter conference version of this paper appeared in Wei, Yang, and Dong (2021). Our initial conference paper only focuses on body gesture modality to recognize emotional state. This paper adds facial expression modality to obtain more complementary information, and more experimental results show the validity of the proposed methods.

The rest of this paper is organized as follows: Section 2 introduces the related work on video emotion recognition methods based on facial expression and body gesture modalities. In Sections 3 and 4, the proposed methods are described in detail. In Section 5, the emotion datasets and experimental setting are described, and the experimental results are discussed. Finally, Section 6 draws a conclusion.

## 2. Related works

### 2.1. Facial expression-based emotion recognition

Currently, 95% of video emotion recognition studies focus on facial expressions (Noroozi et al., 2018). These studies explore many feature representations and emotion recognition methods. Early works pay attention to design handcrafted features to learn spatio-temporal information. For example, Chen, Chen, Chi, and Fu (2018) proposed a new facial texture variation feature called histogram of oriented gradients from three orthogonal planes (HOG-TOP), which extends HOG feature to 3-dimensional space to compute the oriented gradients on orthogonal planes XY, XT, and YT. Zhao and Pietikainen (2007) proposed a dynamic texture feature with volume local binary patterns (VLBP). VLBP extend the LBP operator to combine the motion and appearance for feature extraction. Wang, Hou, Hu, and Ren (2017) proposed a new feature description algorithm, temporal–spatial local ternary pattern moment (TSLTPM). TSLTPM is extended to the temporal–spatial series to quantize the differences of pixel values among adjacent frames into three levels. However, handcrafted feature extraction requires much prior knowledge, and there is still a gap between the features and emotional expression.

With the development of convolutional neural networks and recurrent neural networks, facial expression recognition based on deep neural networks has achieved advanced results (Dong, Ji, & Mei, 2023; Kim, Kim, Roy, & Jeong, 2019; Kollias & Zafeiriou, 2020), showing their superiority in spatial and temporal feature learning. Liang et al. (2020) proposed an end-to-end architecture that uses a deep network to extract spatial features from each frame and uses a convolutional network to model temporal dynamics from a pair of consecutive frames. The framework accumulates clues from fused spatial and temporal features by a BiLSTM network for facial expression recognition. Similarly, Tang, Xie, Li, Liang, and Zhao (2022) extracted spatial features based on ResNet and proposed a bidirectional gated recurrent unit to model the time series of features. Miyoshi, Nagata, and Hashimoto (2021) proposed an enhanced ConvLSTM model that adds skip connections to spatial and temporal directions to learn broader and earlier features, and adds temporal gates to control the flow of information in the model, helping to filter out irrelevant information. Zhi et al. (2022) designed three attention modules in space, channel, and

time dimensions based on CNN backbone network to capture more important spatial and temporal information. Kumawat et al. (2019) proposed a novel Local Binary Volume Convolutional Neural Network (LBVCNN), which uses a 3D convolutional Local Binary Volume layer to capture spatio-temporal information from XY, XT, and YT planes for facial expression recognition. Wang, Ma, Xing, and Pan (2020) proposed Eulerian motion-based 3D convolution network (EM-C3D) that combines C3D with global attention module to learn rich spatiotemporal Eulerian motion features. Park, Kim, and Chilamkurti (2021) proposed a multirate-based 3D convolutional neural network, and designed minimum overlapped frames as inputs to provide more spatio-temporal information learning. Lo, Xie, Shuai, and Cheng (2020) utilized 3D ConvNets to extract AU features and applied GCN layers to discover the dependency laying between AUs. Through the above analysis, the spatio-temporal feature extraction based on deep model is particularly important for facial expression recognition. The experimental results demonstrated that the CNN-RNN framework and 3D CNN network are beneficial to improve recognition performance. However, CNN-RNN will lose part of the information because of learning the spatial and temporal features independently, and 3D CNN has too many parameters.

Based on this problem, we propose a novel spatio-temporal extraction model SISTCM. SISTCM applies 2D convolution to capture the local spatio-temporal features, and the parameters are reduced significantly compared to 3D CNN. Also, we propose a two-stream LSTM model to further learn global temporal cues based on clip-level emotion representations and local spatio-temporal features to improve the performance of emotion recognition.

### 2.2. Body gesture-based emotion recognition

Recognizing human emotions through body gestures is an important study in social psychology (Pease & Pease, 2008). Siegman et al. specified that there may exist general body movement protocols for emotion recognition (Siegman & Feldstein, 2014) ; thus, selecting appropriate movement features and classifiers is becoming the main focus. Six types of body movement features were proposed by Psaltis et al. (2016), including distances, acceleration, velocity, and so on. For different sets of features, a two-layer RBM network was trained separately, then fuse each output probability to predict the emotional states. Maret, Oberson, and Gavrilova (2018) utilized statistical analysis of the 3D human skeleton to extract arm movement features, then selected five classifiers to recognize emotions. Razzaq, Bang, Kang, and Lee (2020) proposed Mesh Distance Features and Mesh Angular features from upper body joints for emotion representation, and utilized Support Vector Machine to build the classifier. Piana, Staglianò, Odone, and Camurri (2016) selected more expressive features (for example, fluidity, impulsiveness, contraction index, et al.) from motion data of 15 joints and process these features based on dictionary learning. The learned features are used as the input of linear SVM to perform emotion recognition. The above methods design various handcrafted features for body gesture-based emotion recognition, but the feature extraction process is relatively time-consuming and complex.

Because of the outstanding results achieved by deep learning in image processing, the research of gesture emotion recognition has shifted to the deep model designation. Sun et al. (2018) presented a deep network, which combines CNN and BLSTM-RNN. CNN is used to extract spatial features, and BLSTM-RNN is used to extract high-level spatio-temporal hierarchical features. Wu, Zhang, Sun, Li, and Zhao (2022) introduced a generalized zero-shot learning framework consisting of two branches. The first learns the prototypes of body gestures based on CNN feature extraction, and the second utilizes autoencoder-based representation learning to predict emotions. Shen, Cheng, Hu, and Dong (2019) used a deep neural network to fuse optical flow features extracted by the STN and skeleton features extracted by ST-GCN. Avola, Cinque, Fagioli, Foresti, and Massaroni (2020) provided an original set of global and time-dependent features for body movement description, and investigated a framework consisting of MLP and N-stacked LSTMs to obtain a higher-level representation. Deng, Leung, Mengoni, and Li (2018) built a deep learning attention-based BLSTM, which can learn the correlations between emotions and human movements. The advantage of the model is that it can focus on the most significant information and better represent the various emotions. The above methods achieve effective performance improvement based on deep models, but there may exist too much noise of learned features from the original video directly. In addition, deep neural networks with plenty of parameters require more computational time and resources.

Therefore, in this paper, we propose a body gesture representation approach based on body joints movement. This method represents body gestures and captures the time-dependent relationship through the joint movement information. Furthermore, we build the ACCM, which uses channel-wise convolutional to preserve the independent properties of each joint and leverage attention mechanism to maximize the advantage of vital features.

## 3. Method for facial expression

The proposed facial expression-based approach for emotion recognition is introduced in this section. As shown in Fig. 1, the method includes three parts: video pre-processing, spatio-temporal features extraction, and emotion recognition. Firstly, the original video is divided into a certain number of clips, and only face part in sequences are kept through the pre-processing module. Secondly, the frames of each clip are used as the input of SISTCM to obtain local spatio-temporal features and clip-level emotion representations. Finally, the local spatio-temporal features and the clip-level emotion representations are simultaneously sent to the two-stream LSTM model to learn the global temporal relationship of facial expressions.

### 3.1. Video pre-processing

Original videos contain various extraneous information such as background, hairstyle, clothing changes, etc. Therefore, for facial expression recognition, it is important to enhance the facial visual information through video pre-processing before learning visual features and training emotion recognition model.

In the pre-processing stage, we first extract all the frames of each video. Secondly, DBFace model[1] is used to detect and position face, then the background information, which is not related to facial expressions, is cropped, and the frames only containing facial parts are returned. DBFace is a lightweight, real-time, single-stage detector for face detection, with faster speed and higher accuracy compared with other face detection models. Finally, considering the inconsistency of each video duration and the progressive relationship of emotional expression, we divide each video into $C$ clips, and each clip consists of a certain number of frames for subsequent processing. With this division, more streamlined video can be obtained, which expands the amount of data and can effectively avoid overfitting to a certain extent.

### 3.2. Spatio-temporal features extraction

Feature representation learning is a very important module in emotion recognition tasks. Facial expression sequences contain rich spatial and temporal information, in which spatial features mainly describe the appearance of facial expressions and temporal features mainly capture emotion change clues over time. Therefore, this paper proposes SISTCM that uses deep neural network to extract deep spatio-temporal features.

***Super-image.*** Because of the similarity of adjacent video frames, we sample $T$ frames for each clip to learn local spatio-temporal relationship instead of complete frames, which can reduce the computational

[1] https://github.com/dlunion/DBFace.

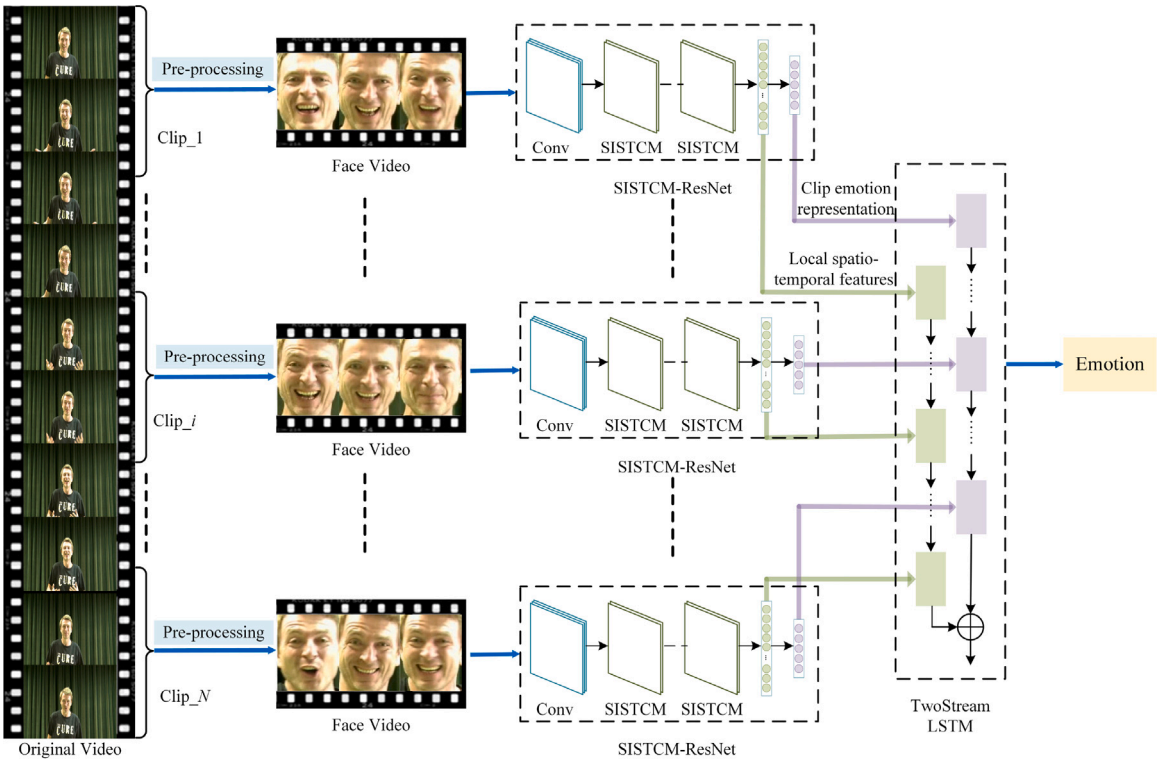



**Fig. 1.** The framework of facial expression-based emotion recognition system.

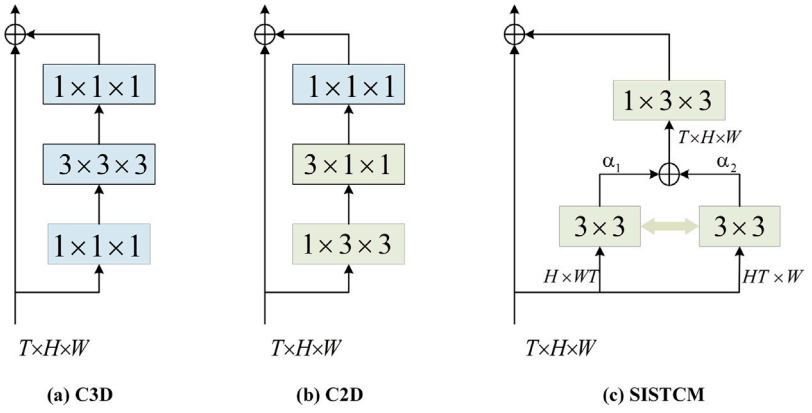



**Fig. 2.** Structure comparison of SISTCM and other models.

complexity. Inspired by the work of Kumawat et al. (2019), the video sequence can be understood as a stack of the XY plane along the T-axis, XT plane along the $Y$-axis, as well as the YT plane along the $X$-axis. Therefore, we stack the $T$ frames along the H and W dimensions to obtain two super images $H \times WT$ and $HT \times W$. After this processing, the super image not only contains the spatial information of single frame, but also contains the temporal dependence information between consecutive frames.

***Super image-based spatio-temporal convolutional model (SISTCM).*** Fig. 2 shows the model structure comparison of SISTCM and others for extracting spatio-temporal features. Among them, the C3D model uses $3 \times 3 \times 3$ 3D convolution to jointly extract spatio-temporal features, while the C2D model first uses $1 \times 3 \times 3$ 2D convolution to extract spatial features, then uses $3 \times 1 \times 1$ 2D convolution to extract temporal features. Different from the C3D and C2D models, our proposed SISTCM only uses $3 \times 3$ 2D convolution to learn spatio-temporal features of the facial expressions. Each clip with the shape of $H \times W \times T$ is used as the input of SISTCM. Firstly, SISTCM converts the input into two super images $H \times WT$ and $HT \times W$. Secondly, $3 \times 3$ 2D convolutions are performed on super images, and the convolution kernels are shared to collaboratively learn local spatio-temporal features under different perspectives, which can reduce the parameters and avoid overfitting. Significantly, there are two reasons why the convolution kernels can be share. (1) The visual information in the two super images is the same, except the viewing angle, so the visual appearance can be considered compatible. (2) The convolution kernels in the deep neural network are essentially redundant without pruning and can be fully utilized for spatio-temporal feature learning. Finally, the obtained two spatio-temporal feature maps $X_H$ and $X_W$ are transformed into the original form $H \times W \times T$, and weighted fused as the final result. The weights are learned through the attention mechanism. The two feature maps are connected, and then a fully-connected layer and a Softmax layer are used to calculate weight:

$$\alpha = Softmax[W_\alpha(X_H, X_W)]$$

**Table 1**
The structure of SISTCM-ResNet18.

| Layer | Output size | Filter | Stride | Padding |
|---|---|---|---|---|
| input | K × 224 × 224 | / | / | / |
| Conv1 | K × 112 × 112 | 7 × 7 | 2 | 3 |
| Maxpool | K × 56 × 56 | 3 × 3 | 2 | 1 |
| SISTCM 1 | K × 56 × 56 | [3 × 3,64 3 × 3,64] × 2 | 1 1 | 1 |
| SISTCM 2 | K × 28 × 28 | [3 × 3,128 3 × 3,128] × 2 | 2/1 1 | 1 |
| SISTCM 3 | K × 14 × 14 | [3 × 3,256 3 × 3,256] × 2 | 2/1 1 | 1 |
| SISTCM 4 | K × 7 × 7 | [3 × 3,512 3 × 3,512] × 2 | 2/1 1 | 1 |
| AdaptiveAvgPool | 512 | 1 × 1 × 1 | / | / |
| FC | ClassNum | 512 × ClassNum | / | / |

where $W_\alpha$ is a learnable parameter. SISTCM can joint learn spatio-temporal features compared with C2D, while can reduces model parameters compared with C3D. It bridges the gap between C2D and C3D and maximizes its advantages, so that the compactness of C2D and the representation ability of C3D can be retained.

***Local spatio-temporal feature extraction and clip-level emotion representation.*** Based on the ResNet18 and SISTCM model, we built a deep convolutional neural network for local spatio-temporal feature extraction and clip-level emotion representation. Table 1 shows the model structure of SISTCM-ResNet18. It needs to be noted that there are differences between SISTCM-ResNet18 and vanilla ResNet18. Firstly, all video frames are processed by 3 × 3 2D Convolutional layer and Maxpool layer. Secondly, replace the original BasicBlock with SISTCM, and the 2D global average pooling is adjusted to 3D global average pooling. At this time, local spatio-temporal features can be obtained. Finally, we use a fully-connected layer (FC) for sentiment classification to obtain the clip-level emotion representation.

### 3.3. Two-stream LSTM model

After SISTCM-ResNet18, we obtain a local spatio-temporal feature sequence and a clip-level emotion representation sequence, respectively. The both only contains clip-level facial expression information and lacks video-level temporal dependency. Meanwhile, emotional expression is not sudden and transient, it has a certain progressive relationship with time. The emotional state of the previous moment will have an impact on the subsequent moment. With above insight, we propose a two-stream LSTM model to learn global temporal cues for facial expression recognition. The model structure of the two-stream LSTM is shown in Fig. 3. It takes the local spatio-temporal feature sequence as the feature stream and clip-level emotion representations as the emotion stream. The feature stream performs emotion recognition and obtains emotion vector $E_1$ from local spatio-temporal features, while the emotion stream is trained to obtain emotion vector $E_2$ from clip-level emotion representations. Then the final video emotional state $E$ is obtained by fusing both emotion vector.

### 3.4. Multi-stage supervision

The network proposed for facial expression recognition is an end-to-end network, and the spatio-temporal feature extraction and the emotion recognition modules are trained together. We note that traditional network is mainly focused on feature representation enhancement while neglecting a careful consideration of supervised knowledge. Also, it only supervises the final output layer to train and optimize the entire model. For our proposed model, the recognized emotion of each stage in the entire recognition process should be as consistent with the label as possible. Therefore, we propose a multi-stage supervised learning approach that provides supervision not only on the final output layer, but at each intermediate output layer of the model, allowing the overall model to be well trained.

Since the clip-level emotion representations can be obtained after the SISTCM-ResNet18 module, L1 Cross-Entropy-Loss is calculated between the clip-level emotion representations and the labels in order to extract the emotion-related features. L1 loss can be written as: $L_1 = -\sum log(p_c)$, where $p_c$ is the estimated probability for the *c-th* example. Similarly, in the two-stream LSTM module, Cross-Entropy-Loss is used to supervise the emotion vectors of feature stream and emotion stream, denoted as L2 loss and L3 loss. Therefore, the final loss is summarized as: $L = L_1 + \lambda L_2 + \mu L_3$, where $\lambda$ and $\mu$ are equilibrium coefficient.

## 4. Method for body gesture

In this section, we introduce the proposed body gesture-based method for emotion recognition. As show in Fig. 4, the method consists of three steps: body joints marking, body gesture representation, and emotion recognition. Firstly, the position of each joint in each video frame is marked. Secondly, we use different methods to represent the changes of body joints. Finally, the body gesture representation is sent to the ACCM[2] to further learn features and recognize emotions.

### 4.1. Body joints marked

The position information of body key joints (e.g., torso, hands, and head) is enough to construct body gestures representation in video frames (Pease & Pease, 2008). Followed this line of reasoning, we only use the position information of key joints to reduce the computational resource. The position data of the whole body in video can be obtained using the OpenPose method (Cao, Simon, Wei, & Sheikh, 2017). In addition, the position data of the whole body can be accessed through the motion capture system Kinect or Xsens.

In this paper, we select 25 body joints as key joints. $(x_t^i, y_t^i)$ represents the obtained position coordinates of joint $i$ in frame $t$. The sparse matrix $I_t^i$ is constructed to represent the position information of each joint, in which the value of $(x_t^i, y_t^i)$ is set to 1, and the value of the others is set to 0. To ensure that the matrix size of each video is consistent, all matrixes are rescaled by setting the $x$ and $y$ coordinates within 64. Therefore, we obtain $T$ description images $I_t = (I_t^1, I_t^2, \ldots, I_t^{25})$ of 25 channels with a 64 × 64 resolution for each video, where $T$ denotes the number of video frames.

### 4.2. Body gesture representation

Based on the description images of each video, two different methods are proposed to represent the changes of body gestures.

#### 4.2.1. Body gesture representation without timeline

Considering that the essence of body gestures changes is the stacking of video frames in time, we encode the temporal relationship to present the changes of body gesture. The relationship $W(t)$ is established to assign weights to the description images, in which different weights mean images at different moments, and the same weight means images at the same moments. For different situations, non-linear

[2] This ACCM is ACCNN of the conference paper. We change the name ACCNN to ACCM just to be consistent with the name SISTCM (face recognition model).

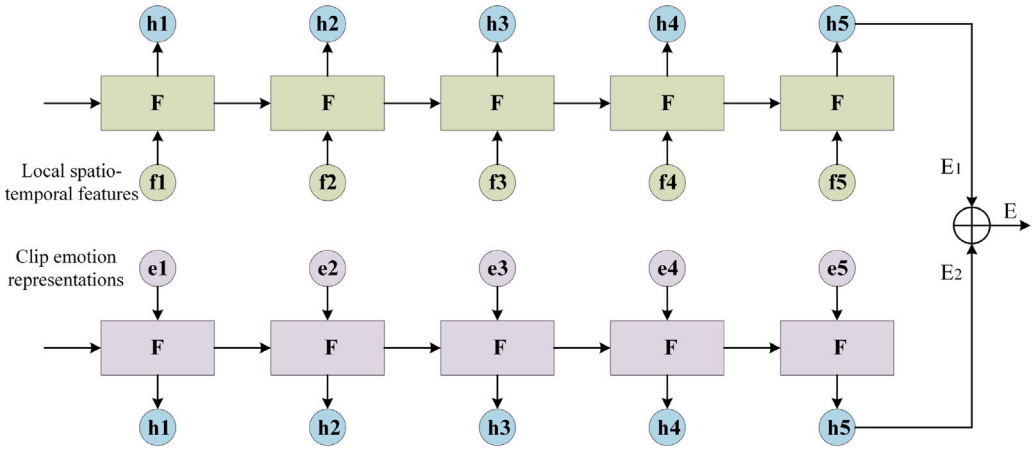



**Fig. 3.** The model structure of the two-stream LSTM.

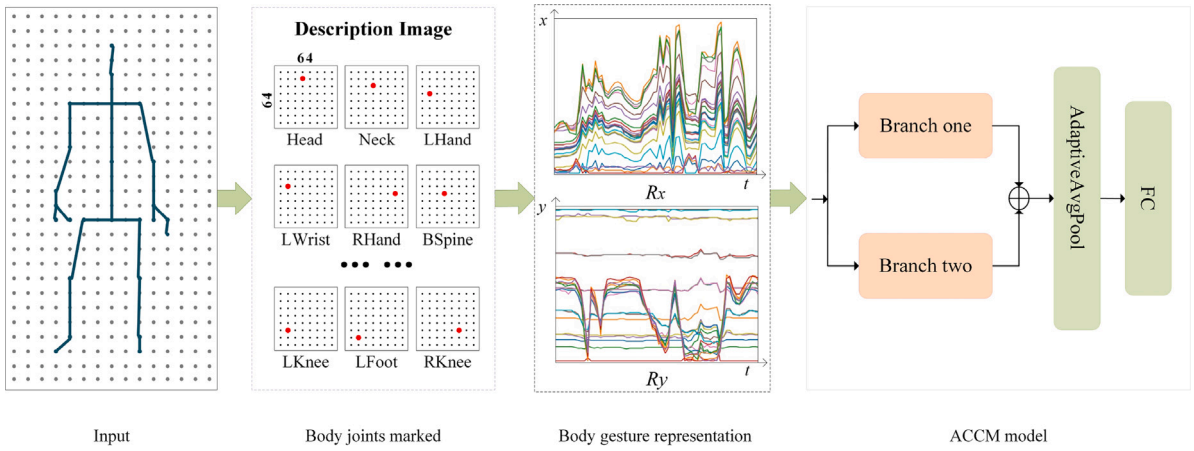



**Fig. 4.** The framework of body gesture-based emotion recognition.

relationship $W(t) = \frac{1}{T-1}(t^2 - t)$ and linear relationship $W(t) = \frac{T}{T-1}(t - 1)$ can be used to establish the relationship. Specifically, firstly, we assign different weights $W(t)$ to each description image of different time, and then obtain the representation $G_t = I_t \cdot W(t)$. Secondly, the representations at different times are aggregated. Finally, body gesture representation is obtained $G = \sum G_t$, in which $G_t = (G_t^1, G_t^2, \ldots, G_t^{25})$.

#### 4.2.2. Body gesture representation with timeline

There may exist some problems of body gesture representation with the above method in some situations. For instance, arms swinging from bottom to top or swinging from top to bottom repetitively, using above method may lead to obtaining the same result of gesture representation due to the aggregate operation. With this consideration, to make up for the above method's deficiency, the trajectory of each body joint is used to represent the movements.

Following the position of each joint over time, the trajectory of body gesture movement can be constructed. Two coordinate systems are established between the $T$-axis and $X$-axis, and between the $T$-axis and $Y$-axis, separately. The diagram of trajectories located in two coordinates will be formed to completely represent the change of body gesture. Specifically, the position coordinate of each joint $i$ in each frame $t$ is represented as $(x_i^t, y_i^t)$, and split it into two vectors $R_x^i$ and $R_y^i$, where $R_x^i = (x_i^1, x_i^2, \ldots, x_i^T)$ and $R_y^i = (y_i^1, y_i^2, \ldots, y_i^T)$ indicates the two trajectories formed by the position of the joint $i$ over time. At last, we aggregate the trajectories of all 25 joints to form the representation $R = (R_x^1, R_y^1, R_x^2, R_y^2, \ldots, R_x^{25}, R_y^{25})$, which represents the body gesture.

### 4.3. ACCM

#### 4.3.1. Model structure

After obtaining the change representation, the ACCM is built to recognize emotion. Fig. 5 shows the model structure of ACCM, which includes two branches. Specifically, one of the branches consists of two blocks, in which each block contains a convolutional layer, BatchNorm layer, and ReLU layer. Another branch includes a channel-wise convolutional layer, attention layer, and above two blocks. The two branches are executed independently and then aggregated, followed by an AdaptiveAvgPooling layer, a fully-connected layer, and a Softmax layer.

ACCM uses the obtained change representations of body gestures as the input, and outputs the final emotion label. Compared with the original video frame, the representations of body gestures preserve more simplified information. Thus, we propose ACCM with shallow layers and without pretraining. It is worth noticing that the parameters of ACCM are about $5 \times 10^2$k, where the ResNet18 are about $3 \times 10^4$k.

#### 4.3.2. Channel-wise convolutional layer

In the operation of traditional convolution, each kernel is multiplied by all elements of input, and the values are merged to get the result. In our representations (the input of ACCM), each channel represents body joints information, thus the channel-wise convolutional is considered to preserve the independent properties of each joint. The channel-wise convolutional layer performs separate convolution operation for each channel, where each channel of input is multiplied by the separate convolution kernel. For example, the input is $H_1 \times W_1 \times C_1$ and the kernel

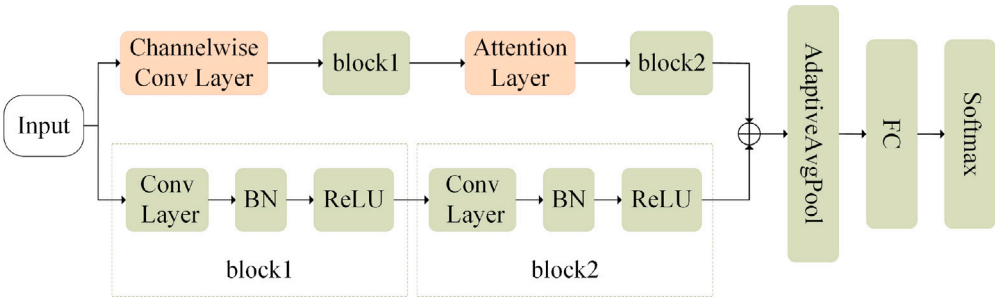



**Fig. 5.** The structure of ACCM.

is $h_1 \times w_1 \times C_1$, and the kernel's number is equal to $C_1$. The $H_1 \times W_1$ of channel $C_i$ performs a convolution operation with the $h_1 \times w_1$ of $C_i$ channel, and gets the output $H_2 \times W_2 \times C_1$.

#### 4.3.3. Attention layer

Considering that the contribution of each joint in emotion recognition may be not consistent, attention layer is added to pay more attention to key joint information. We adopt the Squeeze-and-Excitation module (Hu, Shen, & Sun, 2018) as attention layer, because each channel of the input represents body joints information. Firstly, the Squeeze operation adopts the AdaptiveAvgPool layer to realize, after that, the global features of channel level are obtained. Secondly, the Excitation operation adopts two fully-connected layers to perform. The relationship between each channel can be learned, and the weight of each channel is obtained. Finally, the original input and corresponding weight are multiplied to obtain the final result. The weights and bias are initialized by random sampling from a uniformly distributed $U[-a, a]$, in which $a = \sqrt{\frac{6}{n_{in}+n_{out}}}$, $n_{in}$ and $n_{out}$ represent the number of input and output channels, respectively.

## 5. Experimental results and analysis

### 5.1. Databases

eNTERFACE05 (Martin, Kotsia, Macq, & Pitas, 2006) is an audio-visual emotion dataset, including six emotional states: surprise, sadness, happiness, fear, disgust, and anger. The dataset contains 1,260 video samples of 42 subjects from 14 different nationalities. The resolution of the original frame is 720 × 576, and the average length of the video is 3–4 s. Each subject was asked to listen to six short stories during the collection process, and each story was used to elicit a particular emotion. The subjects were asked to act the six basic emotions, and each emotion was executed five times.

CK+ (Lucey et al., 2010) is one of the most widely used facial expression dataset. It is composed in a restricted laboratory environment, including seven emotional states: surprise, sadness, happiness, fear, disgust, contempt, and anger. The CK+ dataset contains a total of 593 video sequences collected from 123 subjects aged 18 to 30 years old, in which 327 videos have emotion labels. The frame resolution of each video is 640 × 480 and 640 × 490, and the number of frames is 10–60.

Aff-Wild2 (Kollias & Zafeiriou, 2018) is an audio-visual emotion dataset in the wild, which shows various diversity in terms of subjects' ages, appearances, and ethnicities. The dataset consists of 548 videos, including 248 in the TrainingSet and 70 in ValidationSet. It is annotated with 8 expressions annotated (neutral state, anger, disgust, fear, happiness, sadness, surprise, and other). The emotional labels are annotated for each frame, so there are emotional transitions within the same video. We split the video according to label categories to ensure consistent emotional expression in each small segment, resulting in TrainingSet with 5308 video segments and ValidationSet with 2068.

Emilya (Fourati & Pelachaud, 2014) is an audio-visual dataset that expresses emotions through body gestures under a variety of daily actions. Eleven graduate students as actors participated in recording the expression of eight emotions: shame, sadness, pride, panic fear, neutral, joy, anxiety, and anger. The eight actions are walk with an object in hands (WH), throw (Th) an object with one hand, sit down (split into sitting down (SD) and being seated (BS)), simple walk (SW), move books on a table with two hands (MB), lift (Lf), and knock at the door (KD). Each actor data is recorded separately, and the actors are asked to perform each emotion under each action four times to capture a large set of data. Due to some data loss during the acquisition process, the effective data is 8206 for study: 1022 for WH, 1006 for Th, 1038 for SD, 1038 for BS, 1025 for SW, 1031 for MB, 1019 for Lf, and 1027 for KD.

BRED (Filntisis, Efthymiou, Koutras, Potamianos, & Maragos, 2019) is a multimodal dataset that records children interacting with two robots in a laboratory setting. Thirty children were asked to express one of six emotions: surprise, sadness, happiness, fear, disgust, and anger. BRED dataset contains a total of 215 valid sequence, and the average number of frames is 72. During collection stage, children randomly select a card that represents a type of emotion. Children express their emotions after seeing the card.

ESVG (Sapiński et al., 2018) is a multimodal dataset that contains Emotional Speech, Video and Gestures. The dataset was recorded from 16 professional actors, and the available part includes 13 actors with 455 video samples. The actors were recorded to present one of seven emotions: anger, disgust, fear, happiness, neutral state, surprise, and sadness. The actors were recorded separately, and each emotion was executed five times.

### 5.2. Experimental setup

In order to verify the effectiveness of our proposed emotion recognition method, we conduct corresponding experiments on above datasets. We verified our unimodal emotion recognition method on the facial expression datasets (eNTERFACE05, CK+ and Aff-Wild2) and the body gesture emotion datasets (Emilya and BRED). On the other hand, we conducted bimodal fusion strategies exploration on multimodal emotion datasets ESVG for performance improvement.

***Facial expression recognition.*** The proposed facial expression-based model is deployed in the PyTorch framework. The initial parameters of SISTCM-ResNet module are set with the pre-trained ResNet on ImageNet. In the training stage, the Stochastic Gradient Descent optimizer is used with stochastic momentum of 0.9 and weight decay of 0.0001. In addition, 0.005 is set to the initial learning rate and reduce the rate to 50% of the before learning rate over 50 epochs.

***Body gesture emotion recognition.*** The proposed body gesture-based model is implemented in the PyTorch framework, and the Adam optimizer is used. In the training stage, the initial learning rate is set to 0.001, and after 50 epochs, the rate is set to 50% of the before learning rate.

***Bimodal Fusion.*** Score fusion is essentially predicting the final labels based on the prediction scores given by each modality to obtain

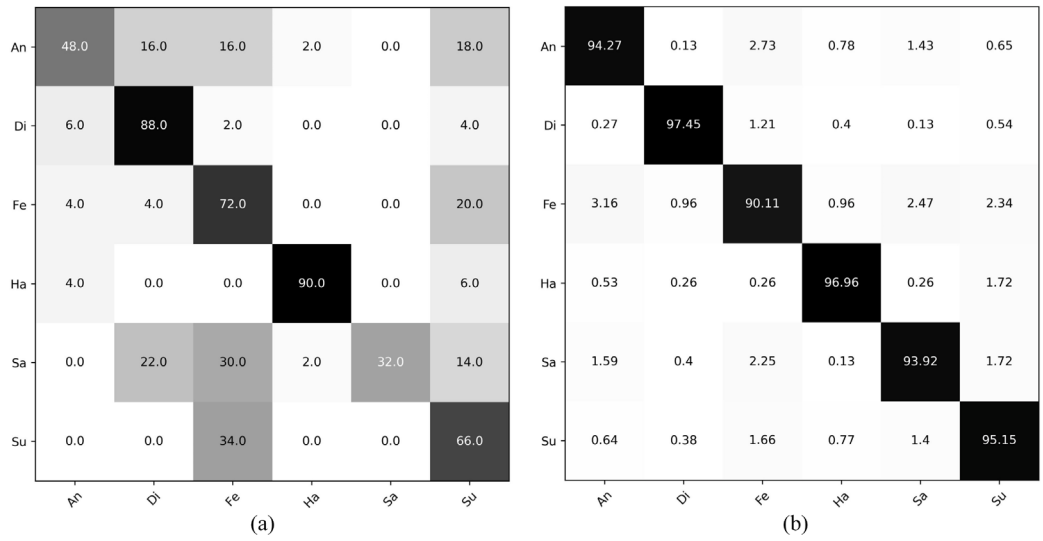



**Fig. 6.** Confusion matrix of our facial expression recognition model on eNTERFACE05 dataset under different cross-validation strategies. (a) is leave-one-subject-out, and (b) is K-fold.

better classification performance. We use two emotion scores obtained from facial expression and body gesture, and explore the combination of two emotion scores using the following three methods:

(1) Taking the average of the two modal emotion scores as the result: average($s1, s2$)

(2) Taking the maximum value of the two modal emotion scores as the result: max($s1, s2$)

(3) Taking the weighted sum of two modal emotion scores as the result: $\lambda s1 + \mu s2$ (The weights are determined by Random Search. First, the weights are sampled from [0,1], and the sum equals 1. Second, classification accuracy is used as the target function, and compute the accuracy using the weighted score $\lambda s1 + \mu s2$. Finally, repeat the above steps to find the set of weights with the best performance.)

where $s1, s2$ denote the facial expression and body gesture-based scores, respectively.

### 5.3. Experimental results

In this section, we present experimental results of unimodal and bimodal emotion recognition on public datasets. Aff-Wild2 dataset provides training/validation/testing partitions. As the testing set lacks labels, we only provide the experimental results on validation set. For other datasets, we follow related research (Filntisis et al., 2019; Fourati & Pelachaud, 2015; Kumawat et al., 2019; Sapiński et al., 2018; Zhalehpour, Onder, Akhtar, & Erdem, 2016; Zhi et al., 2022) and use K-fold cross-validation. In detail, on the ESVG dataset, we adopt the 5-fold cross-validation, and 10-fold cross-validation scheme is employed on the rest. Moreover, leave-one-subject-out cross-validation on the eNTERFACE05 dataset also provided. Tables 2–8 show the recognition accuracy on each dataset, and Figs. 6–8 show the confusion matrix on each dataset. All the recognition accuracy and confusion matrices are obtained based on the average recognition result.

#### 5.3.1. Unimodal emotion recognition performance for facial expression

The accuracy of our proposed method based on facial expression on the eNTERFACE05 dataset is shown in Table 2. In order to make the comparison fair, the models and algorithms based on visual modal are only considered. Firstly, Table 2 shows that our proposed method yields SOTA performance in both cross-validation strategies, especially in the leave-one-subject zero-shot challenging setting still improves by 4.35%, demonstrating that SISTCM module is powerful in feature learning. Secondly, in the leave-one-subject experiment, handcrafted features are used in the research of Zhalehpour et al. (2016), only spatial features in Avots, Sapiński, Bachmann, and Kamińska (2019), and deep spatio-temporal features are used in the research of others. We can find that deep spatio-temporal features are more conducive to facial expression recognition. Thirdly, in the K-folder experiment, even if the Tian et al. extracted landmarks information (Tian & She, 2022) or Shirian et al. utilized graph modeling dynamics (Shirian, Tripathi, & Guha, 2021), our proposed method significantly improves the 3.22% and 7.17% recognition accuracy, which further proves the superiority.

**Table 2**
Facial expression recognition performance compared with previous works on eNTERFACE05 dataset under different cross-validation strategies. (Upper is leave-one-subject-out, and the below is K-fold).

| Method | Accuracy (%) |
|---|---|
| Zhalehpour et al. (2016) | 42.16 |
| Avots et al. (2019) | 48.31 |
| Miyoshi et al. (2021) | 49.26 |
| Zhang, Zhang, Huang, Gao, and Tian (2018) | 54.35 |
| Zhao, Liu, Huang, Lun, and Lam (2022) | 54.62 |
| Ma et al. (2019) | 58.19 |
| SISTCM-TLSTM (Ours) | **62.54** |
| Dong et al. (2023) | 81.04 |
| Chen et al. (2022) | 87.2 |
| Shirian et al. (2021) | 87.49 |
| Tang et al. (2022) | 88.11 |
| Zhi et al. (2022) | 89.25 |
| Tian and She (2022) | 91.44 |
| SISTCM-TLSTM (Ours) | **94.66** |

The confusion matrix of our facial expression recognition model on the eNTERFACE05 dataset is shown in Fig. 6. In the leave-one-subject experiment, the accuracy in the cases of anger and sadness is relatively poor less than 50%. It can be noticed that there is a high degree of confusion among them and disgust and fear, which may be because both belong to negative emotion and happen to look similar in particular facial expression. In the K-folder experiment, we can find that the accuracy of each emotion is improved because the subjects have appeared during training, further illustrating the challenge of leave-one.

The accuracy of our proposed method based on facial expression on the CK+ dataset is shown in Table 3. Zhi et al. (2022) extracted CNN-based deep features, Ravi, Yadhukrishna, et al. (2020) combined the Local Binary Patterns and CNN-based features, and Liu, Wang, and Feng (2021) proposed a hybrid feature representation in which four hand-crafted features are fusing while recognizing facial expression. Their

**Table 3**
Facial expression recognition performance compared with previous works on CK+dataset.

| Method | Accuracy (%) |
|---|---|
| Zhi et al. (2022) | 90.17 |
| Wu and Li (2023) | 96.36 |
| Liu et al. (2021) | 97.20 |
| Ravi et al. (2020) | 97.32 |
| Tian and She (2022) | 98.27 |
| Aouayeb et al. (2021) | **99.7** |
| SISTCM-TLSTM (Ours) | **97.79** |

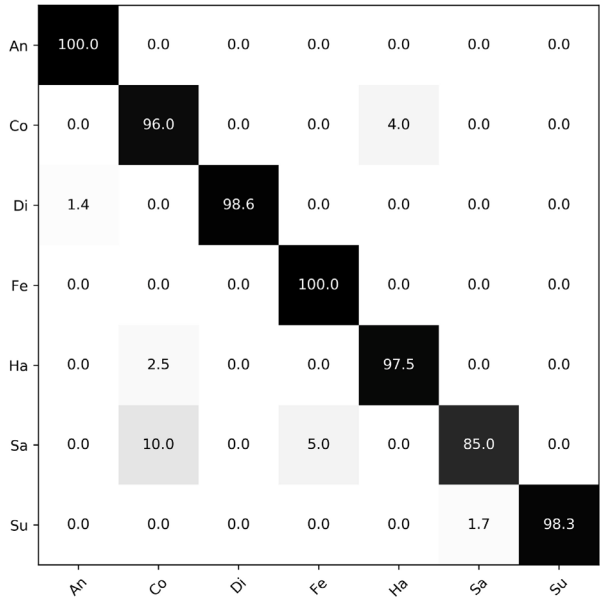



**Fig. 7.** Confusion matrix of our facial expression recognition model on CK+ dataset.

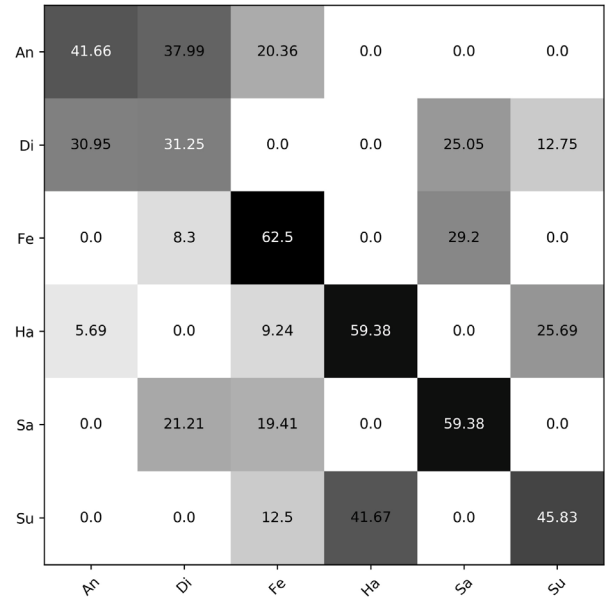



**Fig. 8.** Confusion matrix of our body gesture emotion recognition model on BRED dataset.

**Table 4**
Facial expression recognition performance compared with previous works on Aff-Wild2 ValidationSet.

| Method | F1 |
|---|---|
| Ma, Zhang, Qiu, and Ding (2023) | **44.60** |
| Liu et al. (2023) | 41.41 |
| Zhou, Lu, Xiong, and Wang (2023) | 41.38 |
| Zhang et al. (2022) | 39.4 |
| Kim, Kim, and Won (2022) | 35.71 |
| Ngan Phan, Nguyen, Huynh, and Kim (2022) | 35.87 |
| Baseline (Kollias, 2022) | 23.0 |
| SISTCM-TLSTM (Ours) | 36.70 |

performances are lower because both ignored the temporal relationship learning. Although Wu and Li (2023) created a stacked bidirectional LSTM (Bi-LSTM) model for temporal feature extraction, the cascaded features may be lost part of information compared to our directly extracted spatio-temporal features, which shows the effectiveness of our proposed model. However, Tian and She (2022) and Aouayeb, Hamidouche, Soladie, Kpalma, and Seguier (2021) gain better results than ours. Tian and She (2022) introduce facial landmarks containing extra geometry-appearance features that are beneficial to emotion recognition, but it increases the amount of calculation. Aouayeb et al. (2021) proposed ViT-based model indeed learns richer information than other CNN-based models, thus it also points us to a direction for later research.

The confusion matrix of our facial expression recognition model on the CK+ dataset is shown in Fig. 7. An overall view of the matrixes reveals that the recognition performance of sadness is not ideal. The main reason may be that the data amount of sadness emotion is small, and the emotion clues are not learned well. The accuracy of other emotions is good, and the accuracy of anger and fear reaches 100%. In addition, the recognition accuracy of each emotion is higher than the eNTERFACE05 dataset, which indicates that spontaneous emotions are more difficult to recognize than acted emotions.

In addition, we conducted experiments on the Aff-Wild2 dataset, and the results are presented in Table 4. Firstly, it can be observed that the recognition performance is not satisfactory compared to the two previous datasets. This is attributed to the various diversity among subjects in the wild dataset, posing greater challenges for recognition. Secondly, our approach achieved moderate recognition accuracy, falling short of the results achieved by ABAW5 challenge teams, such as Netease Fuxi. On the one hand, this might be due to the advantage of recent network architectures, while on the other hand, it could be attributed to the information supplementation from the audio modality. Furthermore, compared to some prior works that only utilize facial data, our model still demonstrates a certain degree of superiority, indicating potential for further optimization.

### 5.3.2. Unimodal emotion recognition performance for body gesture

We establish a corresponding body gesture emotion classification model for each action category on the Emilya dataset. Table 5 shows the classification performance for each action. Firstly, our method obtains the best recognition results in all actions, especially in MB, KD, and Lf actions, which are higher than 90%. Secondly, the experimental results show that the accuracy of emotion recognition in the BS and SD actions is lower than 80%. In sitting actions, the actors have limited activity space and expressiveness, so there is no significant difference. Finally, compared with (Fourati & Pelachaud, 2015) and Crenn, Meyer, Konik, Khan, and Bouakaz (2020), the performance of our proposed change representation of body gesture without timeline shows a lower level. However, the proposed change representation of body gesture with timeline achieves the better results in all actions except for SW and WH, which demonstrated the importance of temporal relationship in gesture representation. After fusing these two methods, the overall average accuracy is increased by 3.88% compared with the state-of-the-art (Crenn et al., 2020), which proves that our method has better performance and applicability.

In order to further analyze the recognition performance of each emotion, Table 6 shows the accuracy comparison of each emotion in all

**Table 5**
Body gesture emotion recognition performance compared with previous works on Emilya dataset.

| Methods | Accuracy (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Actions | BS | KD | Lf | MB | SD | SW | Th | WH | Average |
| Fourati and Pelachaud (2015) | 67.9 | 82.4 | 78.7 | 83.1 | 68.5 | 84.8 | 79.4 | 84.2 | 78.63 |
| Crenn et al. (2020) | – | – | – | – | – | – | – | – | 82.20 |
| Without timeline | 52.00 | 68.10 | 70.92 | 73.50 | 54.80 | 67.55 | 64.85 | 67.86 | 64.95 |
| With timeline | 72.40 | 90.30 | 89.49 | 88.80 | 73.80 | 79.18 | 86.80 | 80.31 | 82.64 |
| Fusion (Ours) | **75.58** | **92.08** | **92.02** | **91.38** | **76.60** | **85.71** | **89.05** | **86.22** | **86.08** |

**Table 6**
The recognition accuracy of each emotion in all actions (On Emilya dataset). Notably, the bold parts mean that the performance of our method is better than Fourati and Pelachaud (2015) proposed.

| Methods | Emo | Accuracy (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Actions | – | BS | KD | Lf | MB | SD | SW | Th | WH | Average |
| Fourati and Pelachaud (2015) | Ag | 83.5 | 92.3 | 84.8 | 91.8 | 81.3 | 80.1 | 97.4 | 88.6 | 87.5 |
| | Ax | 53.0 | 71.6 | 66.1 | 71.8 | 55.1 | 82.6 | 71.3 | 80.1 | 68.9 |
| | Jy | 54.0 | 73.7 | 78.6 | 77.6 | 57.1 | 84.4 | 74.6 | 77.6 | 72.2 |
| | Nt | 73.1 | 75.4 | 75.0 | 84.8 | 69.9 | 88.3 | 64.6 | 90.7 | 77.7 |
| | PF | 56.6 | 83.6 | 80.8 | 85.7 | 59.8 | 82.7 | 80.2 | 77.1 | 75.8 |
| | Pr | 77.6 | 86.4 | 83.2 | 81.7 | 82.8 | 89.8 | 78.5 | 81.6 | 82.7 |
| | Sd | 74.9 | 84.8 | 81.0 | 90.4 | 75.5 | 88.9 | 80.3 | 93.6 | 83.7 |
| | Sh | 70.8 | 90.2 | 79.0 | 82.3 | 65.8 | 82.9 | 84.6 | 87.0 | 80.3 |
| Fusion (Ours) | Ag | 81.15 | 92.29 | **94.42** | **93.54** | **82.50** | **87.50** | 93.75 | 87.92 | **89.13** |
| | Ax | **72.69** | **90.96** | **88.27** | **88.04** | **73.27** | 78.85 | **84.38** | **83.46** | **82.49** |
| | Jy | **73.46** | **86.73** | **90.38** | **88.96** | **76.92** | **87.69** | **83.08** | **85.77** | **84.12** |
| | Nt | **77.22** | **95.28** | **95.63** | **96.11** | **74.17** | **93.06** | **90.28** | **91.11** | **89.11** |
| | PF | **69.42** | **94.29** | **90.77** | **91.79** | **75.96** | **83.08** | **88.46** | **80.58** | **84.29** |
| | Pr | **81.92** | **90.19** | **95.77** | **92.69** | 73.08 | **94.42** | **89.04** | **90.96** | **88.51** |
| | Sd | 71.35 | **92.71** | **90.21** | **94.23** | **79.81** | 83.54 | **89.17** | 85.83 | **85.86** |
| | Sh | **77.88** | **94.82** | **91.92** | **87.31** | **76.35** | 79.81 | **94.62** | 85.77 | **86.06** |

actions. Firstly, from the overall observation, each emotion recognition accuracy of our method is better than Fourati and Pelachaud (2015). The neutral recognition accuracy is 89.11%, and it is higher by 11.41%, which illustrates the performance improvement. Secondly, the recognition accuracy of anxiety is lowest in all actions, especially in BS and SD actions, its worst classification result is less than 80%. The possible reason is that most actors tend to move their body when expressing anxiety, but lower body motion clues cannot be captured well under seating actions. Finally, our method has improved the recognition performance of all emotions in MB and Lf action, indicating that these two actions are the easiest to represent with our method.

The accuracy of our proposed method based on body gesture on the BRED dataset is shown in Table 7. Firstly, an overall view of Table 7 reveals that the proposed two gesture representation methods have achieved better recognition performance compared with Filntisis et al. (2019) and Atanassov, Pilev, Tomova, and Kuzmanova (2021) and baseline. It illustrates that our proposed body gesture-based method has better effectiveness and superiority in the emotion recognition task. Secondly, according to the experimental results, the representation of body gesture with timeline has a better recognition performance than the representation without timeline. Finally, the fusion recognition result of two representation methods is better than the single, which demonstrate that the proposed two methods have complementarity in gesture representation. The confusion matrix of our body gesture emotion recognition model on BRED dataset is shown in Fig. 8. It can be observed that the recognition accuracy of each emotion on BRED is limited compared to the Emilya dataset. There may be two reasons for the limited recognition accuracy: first, the data amount is less and unbalanced, thus the effective features are not learned well; second, children have little dependence on body gesture to express emotions, so there are no enough emotion clues to use.

### 5.3.3. Bimodal emotion recognition performance

In order to further verify the effectiveness of the unimodal method and the performance improvement brought by the visual bimodal fusion, we conduct related experiments on the ESVG dataset. The experimental results are shown in Table 8, Table 9, and Fig. 9.

**Table 7**
Body gesture emotion recognition performance comparisons with previous works on BRED dataset.

| Method | Accuracy (%) |
|---|---|
| Baseline | 35.00 |
| Atanassov et al. (2021) | 32.00 |
| Filntisis et al. (2019) | 34.00 |
| Without timeline | 43.31 |
| With timeline | 47.63 |
| Fusion (Ours) | **49.88** |

Firstly, from Table 8, the accuracy of emotion recognition using the facial expression is higher than using the body gesture. There is no doubt that facial expression is the most intuitive visual modality that conveys rich emotional cues, and facial expression recognition achieves high performance on various emotion datasets. Secondly, both the facial expression recognition method and the body gesture emotion recognition method proposed in this paper largely improve the recognition accuracy, especially the recognition performance based on body gesture is improved by 19.35%, which illustrates that our proposed two methods have more superiority in emotion recognition. Thirdly, the proposed body gesture-based emotion recognition method provides comparable recognition performance to the proposed and other facial expression-based recognition. Finally, comparing the experimental results in Tables 8 and 9, we find that taking the maximum value of the two modal emotion scores as the fusion result (max score fusion) hardly improves the recognition performance. Because the recognition performance of the two visual modalities is quite different, taking the maximum value as fusion result did not exploit their advantages. The average score fusion and weighted score fusion improve the recognition result compared to the unimodal, and the weighted score fusion improves the recognition accuracy to 98.42%. It leverages the complementarities of the two modalities, indicating the advantages of fusing two visual modality information.

We further analyze the recognition accuracy for each emotion to intuitively comprehend the performance. Fig. 9 shows the experimental

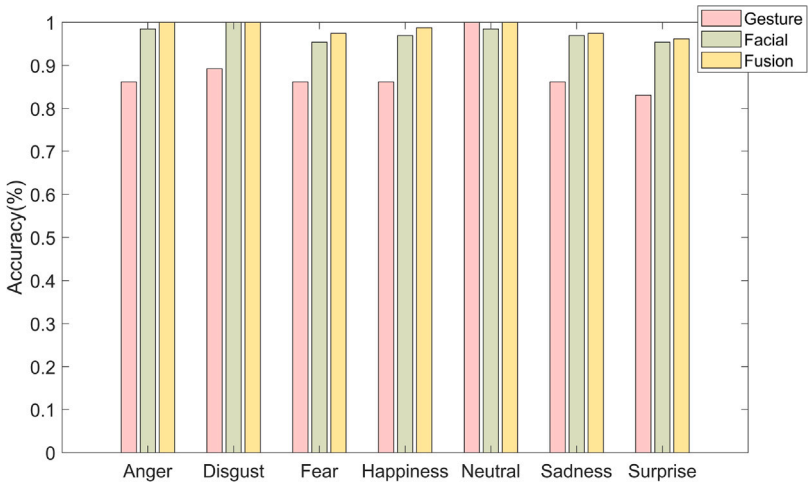



**Fig. 9.** Recognition accuracy for each emotion category on ESVG dataset.

**Table 8**
Unimodal emotion recognition performance comparison with other methods on ESVG dataset.

| Modality | Method | Accuracy (%) |
|---|---|---|
| Facial expression | Baseline (Sapiński et al., 2018) | 94.67 |
| | Wei et al. (2021) | 87.69 |
| | SISTCM-TLSTM (Ours) | **97.67** |
| Body gesture | Baseline (Sapiński et al., 2018) | 62.83 |
| | Sapiński, Kamińska, Pelikant, and Anbarjafari (2019) | 69.00 |
| | ACCM (ours) | **88.35** |

**Table 9**
Bimodal emotion recognition performance comparison at different score fusion strategy on ESVG dataset.

| Score fusion (Facial + Gesture) | Accuracy (%) |
|---|---|
| Average | 98.19 |
| Max | 97.80 |
| Weighted sum | **98.42** |

comparison results. Firstly, after using the weighted score fusion on the facial expression and body gesture, the recognition accuracy of each emotion has been improved, which further verifies the complementarity of two visual modality information. Secondly, the accuracy of neutral, angry, and disgust have achieved 100% after fusion, showing the best recognition performance compared with the other emotions. Finally, the recognition accuracy of neutral based on body gesture is higher than based on facial expression. The reason may be that the body gesture basically does not change significantly when expressing this emotion, thus the gesture clues are very consistent and easier to find features.

#### 5.3.4. Cross-database evaluation

To verify the generalization of our proposed FER method, we conduct the cross-database evaluation on ESVG dataset. We train the model on the different datasets respectively (Enterface05 & CK & Finetune represents that we train on the Enterface05 & CK+ datasets and finetune the last layers on ESVG dataset), then evaluate on ESVG. The number of emotion classes differs, so only six overlapping basic emotions are adopted. An overall view of Table 10 shows that the experimental performance on the cross-dataset is degraded, indicating that the model lacks good generalization ability. We consider that identity information may have a negative impact on recognition. We will try to decouple the identity features and focus on general emotional features in the future. What is more, we also found that the recognition performance shows a growth trend as the amount of training data increases, indicating the effectiveness of the model supported by sufficient data, and shows pre-training the model on large-scale datasets is necessary for small dataset learning.

**Table 10**
Comparison of Cross-Database Evaluation on ESVG dataset.

| Train dataset | Accuracy (%) |
|---|---|
| Enterface05 | 50.52 |
| CK+ | 45.90 |
| Enterface05 & CK+ | 62.82 |
| Enterface05 & CK+& Finetune | 68.59 |

### 5.4. Ablation studies

In this section, to investigate the effectiveness of each module in our proposed method, we carry out ablation experiments.

We evaluate facial expression recognition method SISTCM-TLSTM by removing SISTCM and TLSTM module individually. Further, we use the pre-trained EfficientNet[3] as feature extractor and train a classifier as baseline. Table 11 shows the performance comparison results. Firstly, our proposed methods gained better performance than the baseline, and it shows the importance of spatio-temporal relations. Secondly, compared to ResNet+TLSTM that only extracts spatial features, SISTCM-TLSTM extracts the spatio-temporal features and obtains better emotion recognition performance, which shows that the mining of temporal relationships is very important. Thirdly, we can observe that the recognition performance of the two-stream LSTM is better than only using a single LSTM which further learns the global temporal relationship of local spatio-temporal features, indicating that considering the changing of the clip-level emotion is beneficial to the judgment of the final emotional state. Finally, we noticed that removing TLSTM cause a sharp performance drop on eNTERFACE05. Therefore, we could infer that the emotional expression has more obvious progressive relationship with time in the wild dataset.

From the aspects of gesture representation method and recognition model, we conduct corresponding experiments to verify the effectiveness of proposed method based on body gesture. Table 12 shows the performance comparison results. Firstly, the representation of body gesture with timeline has a better recognition performance than without timeline. Secondly, the fusion result of two representation methods is

[3] https://github.com/HSE-asavchenko/face-emotion-recognition.

**Table 11**
Ablation experiments of facial expression on the four datasets.

| | CK+ | eNTERFACE05 | Aff-Wild2 | ESVG |
|---|---|---|---|---|
| SISTCM-TLSTM (Proposed) | **97.79** | **62.54** | **36.70** | **97.67** |
| ResNet + TLSTM | 95.81 (↓1.98) | 59.13 (↓3.41) | 32.24 (↓4.46) | 97.05 (↓0.62) |
| SISTCM + LSTM | 94.64 (↓3.15) | 53.45 (↓9.09) | 31.25 (↓5.45) | 92.75 (↓4.92) |
| Baseline | 93.33 (↓4.46) | 53.24 (↓9.3) | 29.31 (↓7.39) | 92.34 (↓5.33) |

**Table 12**
Ablation experiments of body gesture emotion on the three datasets.

| | Emilya | BRED | ESVG |
|---|---|---|---|
| Proposed | **86.08** | **49.88** | **88.35** |
| Proposed-UpperBody | 72.50 (↓13.58) | 40.24 (↓9.64) | 75.60 (↓12.75) |
| ACCM(With timeline) | 82.64 (↓3.44) | 47.63 (↓2.25) | 86.24 (↓2.11) |
| CNN (With timeline) | 75.79 (↓10.29) | 42.02 (↓7.86) | 83.58 (↓4.77) |
| ACCM(Without timeline) | 64.95 (↓21.13) | 43.31 (↓6.57) | 62.56 (↓25.79) |
| CNN (Without timeline) | 56.78 (↓29.30) | 38.93 (↓10.95) | 59.46 (↓28.89) |

better than the single, which demonstrates that the proposed two methods have complementarity in gesture representation. Thirdly, the emotion recognition performance of the proposed ACCM is better than traditional CNN. It illustrates that adding the attention layer and channel-wise convolution layer facilitates the performance of body gesture-based emotion recognition. Moreover, considering some datasets might only contain the upper-body parts, we also conduct the experiments with upper-body joints data. Although the performance has a slight drop because of the fewer information, it is still works for the upper body.

## 6. Conclusion

In this paper, we take advantages of the two visual modal characteristics of facial expression and body gesture, and propose suitable video emotion recognition methods separately. For facial expression sequences, we propose SISTCM to extract local spatio-temporal features and learn clip-level emotional states, and further build a two-stream LSTM to further learn global temporal cues based on these features for final emotion recognition. Two-stream LSTM can recognize the emotion more accurately based on the fusion of feature and emotion two paths. Additionally, for the body gesture sequence, we propose a novel method to represent the changes of body gesture with joints information, and built ACCM to recognize emotion. The representation method uses body joints to simplify the gesture information, and the complexity of training can be reduced. ACCM can maximize the advantages of key channel features and preserve the independent properties of each channel. Extensive experimental results show the superiority of the proposed unimodal emotion recognition methods over other methods. Moreover, the fusion of facial expression-based method and body gesture-based method effectively improves the accuracy of emotion recognition.

## CRediT authorship contribution statement

**Jie Wei:** Conceptualization, Validation, Writing – original draft. **Guanyu Hu:** Visualization, Validation, Writing – review & editing. **Xinyu Yang:** Conceptualization, Writing – review & editing, Supervision. **Anh Tuan Luu:** Formal analysis, Writing – review & editing, Supervision. **Yizhuo Dong:** Methodology, Investigation, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


Abdullah, S. M. S., & Abdulazeez, A. M. (2021). Facial expression recognition based on deep learning convolution neural network: A review. *Journal of Soft Computing and Data Mining*, *2*(1), 53–65. http://dx.doi.org/10.30880/jscdm.2021.02.01.006.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274, https://psycnet.apa.org/doi/10.1037/0033-2909.111.2.256.

Aouayeb, M., Hamidouche, W., Soladie, C., Kpalma, K., & Seguier, R. (2021). Learning vision transformer with squeeze and excitation for facial expression recognition. http://dx.doi.org/10.48550/arXiv.2107.03107, arXiv preprint arXiv:2107.03107.

Atanassov, A. V., Pilev, D. I., Tomova, F. N., & Kuzmanova, V. D. (2021). Hybrid system for emotion recognition based on facial expressions and body gesture recognition. In *2021 International conference automatics and informatics* (pp. 135–140). IEEE, http://dx.doi.org/10.1109/ICAI52893.2021.9639829.

Avola, D., Cinque, L., Fagioli, A., Foresti, G. L., & Massaroni, C. (2020). Deep temporal analysis for non-acted body affect recognition. *IEEE Transactions on Affective Computing*, *13*(3), 1366–1377. http://dx.doi.org/10.1109/TAFFC.2020.3003816.

Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2019). Audiovisual emotion recognition in wild. *Machine Vision and Applications*, *30*(5), 975–985. http://dx.doi.org/10.1007/s00138-018-0960-9.

Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, *59*(1–2), 213–225. http://dx.doi.org/10.1016/S1071-5819(03)00050-8.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE conference on computer vision and pattern recognition* (pp. 7291–7299). IEEE, http://dx.doi.org/10.48550/arXiv.1611.08050.

Chen, J., Chen, Z., Chi, Z., & Fu, H. (2018). Facial expression recognition in video with multiple feature fusion. *IEEE Transactions on Affective Computing*, *9*(1), 38–50. http://dx.doi.org/10.1109/TAFFC.2016.2593719.

Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., & Hirota, K. (2022). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, *70*(1), 1016–1024. http://dx.doi.org/10.1109/TIE.2022.3150097.

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, 1–18. http://dx.doi.org/10.1007/s00521-021-06012-8.

Crenn, A., Meyer, A., Konik, H., Khan, R. A., & Bouakaz, S. (2020). Generic body expression recognition based on synthesis of realistic neutral motion. *IEEE Access*, *8*, 207758–207767. http://dx.doi.org/10.1109/ACCESS.2020.3038473.

Deng, J. J., Leung, C. H. C., Mengoni, P., & Li, Y. (2018). Emotion recognition from human behaviors using attention model. In *2018 IEEE first international conference on artificial intelligence and knowledge engineering* (pp. 249–253). IEEE, http://dx.doi.org/10.1109/AIKE.2018.00056.

Dong, D., Ji, R., & Mei, Y. (2023). Dual-sequence LSTM multimodal emotion recognition based on attention mechanism. In *Intelligent robotics: Third China annual conference, CCF CIRAC 2022, Xi'an, China, December 16–18, 2022, Proceedings* (pp. 145–157). Springer, http://dx.doi.org/10.1007/978-981-99-0301-6_12.

Farzaneh, A. H., & Qi, X. (2021). Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2402–2411). IEEE.

Filntisis, P. P., Efthymiou, N., Koutras, P., Potamianos, G., & Maragos, P. (2019). Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. *IEEE Robotics and Automation Letters*, *4*(4), 4011–4018. http://dx.doi.org/10.1109/LRA.2019.2930434.

Fourati, N., & Pelachaud, C. (2014). Emilya: Emotional body expression in daily actions database. In *The 9th international conference on language resources and evaluation* (pp. 3486–3493). ELRA.

Fourati, N., & Pelachaud, C. (2015). Multi-level classification of emotional body expression. *1*, In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/FG.2015.7163145.

Fu, X., Xue, C., Li, X., Zhang, Y., & Cai, T. (2020). A review of body gesture based affective computing. *Journal of Computer-Aided Design & Computer Graphics*, *32*(7), 1052–1061.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition* (pp. 7132–7141). IEEE.

Huang, K., Li, J., Cheng, S., Yu, J., Tian, W., Zhao, L., et al. (2020). An efficient algorithm of facial expression recognition by TSG-RNN network. In *International conference on multimedia modeling* (pp. 161–174). Springer, http://dx.doi.org/10.1007/978-3-030-37734-2_14.

Kim, J.-H., Kim, B.-G., Roy, P. P., & Jeong, D.-M. (2019). Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access*, *7*, 41273–41285. http://dx.doi.org/10.1109/ACCESS.2019.2907327.

Kim, J.-H., Kim, N., & Won, C. S. (2022). Facial expression recognition with swin transformer. http://dx.doi.org/10.48550/arXiv.2203.13472, arXiv preprint arXiv:2203.13472.

Kollias, D. (2022). Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2328–2336). http://dx.doi.org/10.48550/arXiv.2202.10659.

Kollias, D., & Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. http://dx.doi.org/10.48550/arXiv.1811.07770, arXiv preprint arXiv:1811.07770.

Kollias, D., & Zafeiriou, S. (2020). Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the OMG in-the-wild dataset. *IEEE Transactions on Affective Computing*, *12*(3), 595–606. http://dx.doi.org/10.1109/TAFFC.2020.3014171.

Kumawat, S., Verma, M., & Raman, S. (2019). LBVCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences. In *IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. pp. 1–10). IEEE.

Lamba, P. S., & Virmani, D. (2021). CNN-LSTM-based facial expression recognition. In *Proceedings of 3rd international conference on computing informatics and networks* (pp. 379–389). Springer, http://dx.doi.org/10.1007/978-981-15-9712-1_32.

Li, M., Chen, L., Wu, M., Pedrycz, W., & Hirota, K. (2021). Multimodal information-based broad and deep learning model for emotion understanding. In *2021 40th Chinese control conference* (pp. 7410–7414). IEEE, http://dx.doi.org/10.23919/CCC52363.2021.9549897.

Li, H., & Xu, H. (2020). Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems*, *204*, Article 106172. http://dx.doi.org/10.1016/j.knosys.2020.106172.

Liang, D., Liang, H., Yu, Z., & Zhang, Y. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. *The Visual Computer*, *36*(3), 499–508. http://dx.doi.org/10.1007/s00371-019-01636-3.

Liu, J., Wang, H., & Feng, Y. (2021). An end-to-end deep model with discriminative facial features for facial expression recognition. *IEEE Access*, *9*, 12158–12166. http://dx.doi.org/10.1109/ACCESS.2021.3051403.

Liu, C., Zhang, X., Liu, X., Zhang, T., Meng, L., Liu, Y., et al. (2023). Facial expression recognition based on multi-modal features for videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5871–5878).

Lo, L., Xie, H.-X., Shuai, H.-H., & Cheng, W.-H. (2020). MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks. In *2020 IEEE conference on multimedia information processing and retrieval* (pp. 79–84). IEEE, http://dx.doi.org/10.1109/MIPR49039.2020.00023.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops* (pp. 94–101). IEEE, http://dx.doi.org/10.1109/CVPRW.2010.5543262.

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, *46*, 184–192. http://dx.doi.org/10.1016/j.inffus.2018.06.003.

Ma, B., Zhang, W., Qiu, F., & Ding, Y. (2023). A unified approach to facial affect analysis: The MAE-face visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5923–5932).

Maret, Y., Oberson, D., & Gavrilova, M. (2018). Identifying an emotional state from body movements using genetic-based algorithms. In *Artificial intelligence and soft computing: 17th international conference, ICAISC 2018, Zakopane, Poland, June 3-7, 2018, proceedings, part I 17* (pp. 474–485). Springer, http://dx.doi.org/10.1007/978-3-319-91253-0_44.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *22nd International conference on data engineering workshops* (p. 8). IEEE, http://dx.doi.org/10.1109/ICDEW.2006.145.

Miyoshi, R., Nagata, N., & Hashimoto, M. (2021). Enhanced convolutional LSTM with spatial and temporal skip connections and temporal gates for facial expression recognition from video. *Neural Computing and Applications*, *33*(13), 7381–7392. http://dx.doi.org/10.1007/s00521-020-05557-4.

Ngan Phan, K., Nguyen, H.-H., Huynh, V.-T., & Kim, S.-H. (2022). Expression classification using concatenation of deep neural network for the 3rd ABAW3 competition. http://dx.doi.org/10.48550/arXiv.2203.12899, arXiv e-prints, arXiv–2203.

Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, pp. 1–20. http://dx.doi.org/10.1109/TAFFC.2018.2874986.

Park, S.-J., Kim, B.-G., & Chilamkurti, N. (2021). A robust facial expression recognition algorithm based on multi-rate feature fusion scheme. *Sensors*, *21*(21), 6954. http://dx.doi.org/10.3390/s21216954.

Pease, B., & Pease, A. (2008). *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam.

Piana, S., Staglianò, A., Odone, F., & Camurri, A. (2016). Adaptive body gesture representation for automatic emotion recognition. *ACM Transactions on Interactive Intelligent Systems*, *6*(1), 1–31. http://dx.doi.org/10.1145/2818740.

Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K. C., Dimitropoulos, K., et al. (2016). Multimodal affective state recognition in serious games applications. In *2016 IEEE international conference on imaging systems and techniques* (pp. 435–439). IEEE, http://dx.doi.org/10.1109/IST.2016.7738265.

Ravi, R., Yadhukrishna, S., et al. (2020). A face expression recognition using CNN & LBP. In *2020 fourth international conference on computing methodologies and communication* (pp. 684–689). IEEE, http://dx.doi.org/10.1109/ICCMC48092.2020.ICCMC-000127.

Razzaq, M. A., Bang, J., Kang, S. S., & Lee, S. (2020). Unskem: unobtrusive skeletal-based emotion recognition for user experience. In *2020 International conference on information networking* (pp. 92–96). IEEE, http://dx.doi.org/10.1109/ICOIN48656.2020.9016601.

Revina, I. M., & Emmanuel, W. S. (2021). A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, *33*(6), 619–628. http://dx.doi.org/10.1016/j.jksuci.2018.09.002.

Saha, S., Datta, S., Konar, A., & Janarthanan, R. (2014). A study on emotion recognition from body gestures using Kinect sensor. In *2014 International conference on communication and signal processing* (pp. 056–060). IEEE, http://dx.doi.org/10.1109/ICCSP.2014.6949798.

Sapiński, T., Kamińska, D., Pelikant, A., & Anbarjafari, G. (2019). Emotion recognition from skeletal movements. *Entropy*, *21*(7), 646. http://dx.doi.org/10.3390/e21070646.

Sapiński, T., Kamińska, D., Pelikant, A., Ozcinar, C., Avots, E., & Anbarjafari, G. (2018). Multimodal database of emotional speech, video and gestures. In *International conference on pattern recognition* (pp. 153–163). Springer, http://dx.doi.org/10.1007/978-3-030-05792-3_15.

Shan, C., Gong, S., & McOwan, P. W. (2007). Beyond facial expressions: learning human emotion from body gestures. In *British machine vision conference* (pp. pp. 1–10). Citeseer.

Shen, Z., Cheng, J., Hu, X., & Dong, Q. (2019). Emotion recognition based on multi-view body gestures. In *2019 Ieee international conference on image processing* (pp. 3317–3321). IEEE, http://dx.doi.org/10.1109/ICIP.2019.8803460.

Shirian, A., Tripathi, S., & Guha, T. (2021). Dynamic emotion modeling with learnable graphs and graph inception network. *IEEE Transactions on Multimedia*, *24*, 780–790. http://dx.doi.org/10.1109/TMM.2021.3059169.

Shukla, A., Gullapuram, S. S., Katti, H., Kankanhalli, M., Winkler, S., & Subramanian, R. (2020). Recognition of advertisement emotions with application to computational advertising. *IEEE Transactions on Affective Computing*, 1–13. http://dx.doi.org/10.1109/TAFFC.2020.2964549.

Siegman, A. W., & Feldstein, S. (2014). *Nonverbal behavior and communication*. Psychology Press.

Sun, B., Cao, S., He, J., & Yu, L. (2018). Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. *Neural Networks*, *105*, 36–51. http://dx.doi.org/10.1016/j.neunet.2017.11.021.

Sun, W., Zhao, H., & Jin, Z. (2019). A facial expression recognition method based on ensemble of 3D convolutional neural networks. *Neural Computing and Applications*, *31*(7), 2795–2812. http://dx.doi.org/10.1007/s00521-017-3230-2.

Tang, G., Xie, Y., Li, K., Liang, R., & Zhao, L. (2022). Multimodal emotion recognition from facial expression and speech based on feature fusion. *Multimedia Tools and Applications*, 1–15. http://dx.doi.org/10.1007/s11042-022-14185-0.

Tian, J., & She, Y. (2022). A visual–audio-based emotion recognition system integrating dimensional analysis. *IEEE Transactions on Computational Social Systems*, http://dx.doi.org/10.1109/TCSS.2022.3200060.

Val-Calvo, M., Álvarez-Sánchez, J. R., Ferrández-Vicente, J. M., & Fernández, E. (2020). Affective robot story-telling human-robot interaction: exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access*, *8*, 134051–134066. http://dx.doi.org/10.1109/ACCESS.2020.3007109.

Wang, X., Hou, D., Hu, M., & Ren, F. (2017). Dual-modality emotion recognition based on composite spatio-temporal features. *Journal of Image and Graphics*, *22*(01), 39–48.

Wang, Y., Ma, H., Xing, X., & Pan, Z. (2020). Eulerian motion based 3dCNN architecture for facial micro-expression recognition. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part I* (pp. 266–277). Springer, http://dx.doi.org/10.1007/978-3-030-37731-1_22.

Wang, R., & Shi, Z. (2021). Personalized online education learning strategies based on transfer learning emotion classification model. *Security and Communication Networks*, *2021*, 1–11. http://dx.doi.org/10.1155/2021/5441631.

Wei, J., Yang, X., & Dong, Y. (2021). Time-dependent body gesture representation for video emotion recognition. In *International conference on multimedia modeling* (pp. 403–416). Springer, http://dx.doi.org/10.1007/978-3-030-67832-6_33.

Wu, Y., & Li, J. (2023). Multi-modal emotion identification fusing facial expression and EEG. *Multimedia Tools and Applications*, *82*(7), 10901–10919. http://dx.doi.org/10.1007/s11042-022-13711-4.

Wu, J., Zhang, Y., Sun, S., Li, Q., & Zhao, X. (2022). Generalized zero-shot emotion recognition from body gestures. *Applied Intelligence*, 1–19. http://dx.doi.org/10.1007/s10489-021-02927-w.

Yan, J., Zheng, W., Xin, M., & Yan, J. (2014). Integrating facial expression and body gesture in videos for emotion recognition. *IEICE Transactions on Information and Systems*, *97*(3), 610–613. http://dx.doi.org/10.1587/transinf.E97.D.610.

Zepf, S., Hernandez, J., Schmitt, A., Minker, W., & Picard, R. W. (2020). Driver emotion recognition for intelligent vehicles: a survey. *ACM Computing Surveys*, *53*(3), 1–30. http://dx.doi.org/10.1145/3388790.

Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2016). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, *8*(3), 300–313. http://dx.doi.org/10.1109/TAFFC.2016.2553038.

Zhang, W., Qiu, F., Wang, S., Zeng, H., Zhang, Z., An, R., et al. (2022). Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops* (pp. 2428–2437). http://dx.doi.org/10.48550/arXiv.2203.12367.

Zhang, Y., & Zhang, L. (2015). Semi-feature level fusion for bimodal affect regression based on facial and bodily expressions. In *2015 international conference on autonomous agents and multiagent systems* (pp. 1557–1565). IFAAMAS.

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2018). Learning affective features with a hybrid deep model for audio–Visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030–3043. http://dx.doi.org/10.1109/TCSVT.2017.2719043.

Zhao, R., Liu, T., Huang, Z., Lun, D. P., & Lam, K.-M. (2022). Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition. *IEEE Transactions on Affective Computing*, http://dx.doi.org/10.1109/TAFFC.2022.3181736.

Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 915–928. http://dx.doi.org/10.1109/TPAMI.2007.1110.

Zhi, J., Song, T., Yu, K., Yuan, F., Wang, H., Hu, G., et al. (2022). Multi-attention module for dynamic facial emotion recognition. *Information*, *13*(5), 207. http://dx.doi.org/10.3390/info13050207.

Zhou, W., Lu, J., Xiong, Z., & Wang, W. (2023). Leveraging TCN and transformer for effective visual-audio fusion in continuous emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops* (pp. 5755–5762). http://dx.doi.org/10.48550/arXiv.2303.08356.

Zhu, Q., Mao, Q., Jia, H., Noi, O. E. N., & Tu, J. (2022). Convolutional relation network for facial expression recognition in the wild with few-shot learning. *Expert Systems with Applications*, *189*, Article 116046. http://dx.doi.org/10.1016/j.eswa.2021.116046.