sensors

MDPI

*Article*

# Skeleton-Based Emotion Recognition Based on Two-Stream Self-Attention Enhanced Spatial-Temporal Graph Convolutional Network

**Jiaqi Shi [1,2,*], Chaoran Liu [2], Carlos Toshinori Ishi [2,3] and Hiroshi Ishiguro [1,2]**

[1] Graduate School of Engineering Science, Osaka University, Osaka 565-0871, Japan; ishiguro@irl.sys.es.osaka-u.ac.jp
[2] Advanced Telecommunications Research Institute International, Kyoto 619-0237, Japan; chaoran.liu@atr.jp (C.L.); carlos@atr.jp (C.T.I.)
[3] Interactive Robot Research Team, Robotics Project, RIKEN, Kyoto 351-0198, Japan
* Correspondence: shi.jiaqi@irl.sys.es.osaka-u.ac.jp

**Abstract:** Emotion recognition has drawn consistent attention from researchers recently. Although gesture modality plays an important role in expressing emotion, it is seldom considered in the field of emotion recognition. A key reason is the scarcity of labeled data containing 3D skeleton data. Some studies in action recognition have applied graph-based neural networks to explicitly model the spatial connection between joints. However, this method has not been considered in the field of gesture-based emotion recognition, so far. In this work, we applied a pose estimation based method to extract 3D skeleton coordinates for IEMOCAP database. We propose a self-attention enhanced spatial temporal graph convolutional network for skeleton-based emotion recognition, in which the spatial convolutional part models the skeletal structure of the body as a static graph, and the self-attention part dynamically constructs more connections between the joints and provides supplementary information. Our experiment demonstrates that the proposed model significantly outperforms other models and that the features of the extracted skeleton data improve the performance of multimodal emotion recognition.

**Keywords:** emotion recognition; gesture; skeleton; graph convolutional networks; self-attention



**Citation:** Shi, J.; Liu, C.; Ishi, C.T.; Ishiguro, H. Skeleton-Based Emotion Recognition Based on Two-Stream Self-Attention Enhanced Spatial-Temporal Graph Convolutional Network. *Sensors* **2021**, *21*, 205. http://doi.org/10.3390/s21010205

Received: 30 November 2020
Accepted: 27 December 2020
Published: 30 December 2020



## 1. Introduction

Multimodal emotion recognition has attracted a lot of attention due to its wide range of application scenarios. Many previous research efforts on emotion recognition process the information from different modalities and use multimodal clues to infer the emotional states. Body gestures are an integral part of nonverbal communication and deliver extremely important supplementary information when expressing emotions [1]. A robot or interactive system with the ability to recognize emotions from body movement can bring significant benefits to many applications, such as biometric security, healthcare, and gaming [2]. Body movement information has good robustness and can be a better alternative for emotion recognition from a distance [3]. However, considerably less work has been done on automatic emotion recognition using body movement.

The dynamic human skeleton is the most intuitional and natural method for depicting human actions, and it does so by structuring the body movement as a sequence of joint positions. The scarcity of labeled data containing 3D skeleton data is one of the significant reasons the body gesture modality has seldom received attention in emotion recognition tasks. Although several multimodal emotional databases include skeleton coordinates, e.g., the multimodal database created by Sapiński et al. (560 samples) [4] and the emoFBVP database (1380 samples) [5], the relatively small amount of samples and the lack of interpersonal dialogue and interaction limit the generalization and robustness of the models trained on them. By contrast, research using speech signals, textual transcriptions and

facial expressions mostly evaluate their models on large open-source multimodal emotional benchmark datasets, such as the interactive emotional dyadic motion capture database (IEMOCAP, over 10000 samples) [6]. However, these databases do not contain skeleton data representing the gesture modality, which makes them difficult to use in gesture emotion recognition.

Since deep learning has revolutionized many fields and had an outstanding performance, some recent research has utilized neural networks on gesture-based emotion recognition [7–9], which fed video frames or a sequence of joint coordinates into neural networks, e.g., convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to extract emotion-related features and make predictions. However, since the spatial connections and graphic structures between joints are seldom explicitly considered by these methods using image sequences and skeletons, the ability to understand the emotion expressed by the body movement is relatively limited.

In the field of action recognition, the Spatial Temporal Graph Convolutional Network (ST-GCN) [10] abstracts the skeletal structure of the human body as a spatial graph, in which the vertexes are joints and the edges represent natural bone connections, and then constructs a graph network using spatial-temporal convolutional blocks to extract features and make predictions. However, ST-GCN has some shortcomings: the topology of the graph is fixed and does not change with the information of each node, which means that the flow direction of information between nodes is predefined and limited to the natural connection between the joints of the body [11,12]. This is also unfavorable for emotion prediction tasks based on skeletal data and there is a lack of flexible connections that are necessary for emotional expression using body gestures. For example, when people express anger, the joints of the limbs, especially the two hands, tend to have relatively high gestural dynamics characteristics, e.g., movement speed and amount of movement [13]. A strong connection between these joints is likely necessary, but the fixed graph structure does not guarantee that the network can capture the appropriate dependency.

To solve these problems, we make the following contributions: (i) we extract 3D skeleton movement data from raw video based on pose estimation and the method can be used to expand existing databases to alleviate the lack of labeled data. (ii) We propose a self-attention enhanced spatial temporal graph convolutional network for skeleton-based emotion recognition, in which the self-attention part of the spatial graph convolutional layer dynamically constructs connections between the nodes of the entire graph and provides supplementary information for the spatial graph convolution. The model outperforms baselines by a significant margin. (iii) We fuse the representations extracted by the proposed model with the audio and text high-level features, to use audio, text, and skeleton modalities simultaneously. The performance significantly exceeds that of the bimodal model using only audio and text information, which shows the effectiveness of the extracted modality.

## 2. Related Work

### *2.1. Emotion Recognition*

Emotion not only plays a crucial role in interaction, decision making, and cognitive processes [14], but also further affects the accuracy of our memory of some events [15]. In human-human interaction, people naturally express and recognize their emotions through multiple modalities including speech, semantics, facial cues, and physical movement [16]. In recent years, the automatic detection of human emotional states has consistently drawn attention from researchers, due to its wide range of applications and growing demand in many different areas such as human-robot interaction, safety driving, website customization, and education [17,18]. For example, a driver emotion detection system can automatically infer the driver's emotional state and take corresponding measures to ensure road safety and human health [19]. Many previous research efforts on emotion recognition process the information from different modalities and use multimodal clues to infer the emotional states, and they show improvement of the overall performance by multimodal fusion [20]. Tzirakis et al. applied convolutional neural networks to extract features from

the speech and the facial expression and utilized long short-term memory networks to model the context and improve the performance [21]. Heusser et al. combined pretrained language models with speech emotion recognition and achieved a better accuracy [22]. However, these works rarely consider gesture as the source of information on recognizing emotions.

### *2.2. Gesture-Based Emotion Recognition*

There have been some studies about gesture-based emotion recognition over the past few years [23]. For example, Ahmed et al. [2,24] applied a feature selection algorithm to select features from the movement feature groups and proposed a genetic algorithm to recognize emotions from body movement. A big challenge of emotion recognition is to design suitable features to capture the emotional properties of body movement. Inspired by the great performance improvement brought by deep learning in many important tasks, some researchers also used deep learning in an end-to-end manner for gesture-based emotion recognition. Filntisis et al. [8] fed flattened 2D pose data into deep neural networks to get a representation and fused it with facial features to predict emotion, which achieved better results than the model using only facial data. Karumuri et al. [25] used convolutional neural networks for classification of emotions from (motion capture) data of dance movements. Deng et al. [26] proposed an attention-based bidirectional LSTM to explore the relationship between human behavior data and emotions.

Most of these studies used open-source multimodal datasets or created new datasets without a large amount of data. Since labeled emotional gesture data is scarcer than speech and facial data, these studies have been limited. Moreover, single-person action performance instead of natural dialogue-based interactive data has been used in these studies, which makes learning the natural expression of human emotions in interactive scenes difficult. Besides, most of the existing deep learning methods for gesture emotion recognition are based on RNNs or CNNs, which cannot represent the spatial connections of joints.

### *2.3. Graph Neural Networks*

Graph is a data structure composed of a series of nodes and edges. As non-Euclidean data, graph data have no regular spatial structure, and their complexity poses a great challenge to existing machine learning algorithms [27]. Recently, graph neural networks (GNNs) have been used in a wide variety of difficult tasks for previous machine learning algorithms, because of their ability to model the data generated from non-Euclidean domains and capture the internal dependence of the data [28], and they have achieved extensive success [29,30].

Encouraged by the success of CNNs in the computer vision domain, convolutional graph networks (GCNs) that generalize convolutions to graph data have attracted an increasing interest [27]. These approaches can be categorized as spectral approaches and spatial approaches [31]. Spectral approaches define graph convolutions by using the graph Fourier transform. Bruna et al. [32] first developed a spectral-based graph convolution, which is limited because of the huge calculation burden. The approaches proposed by Defferard et al. [33] improved efficiency and solved this problem. Spatial approaches directly use the topology of the graph and apply the convolution filter based on the neighborhood information of the graph. Monti et al. [34] proposed a spatial-domain model and generalized CNN architectures to non-Euclidean domains. Hamilton et al. [35] proposed a general inductive framework that samples and aggregates features from a node's local neighborhood to generate embeddings. This work follows the spatial approaches.

## 3. Skeletal Data Extraction

We extract dynamic body skeleton data from raw video of a large existing open-source emotional database to add a new modality representing gestures. To do this, we use a human pose estimation method to draw a temporal sequence of joint positions in the form

of 3D coordinates. Then some prepossessing steps are applied for the data before feeding it into the emotion recognition model.

### *3.1. Human Pose Estimation*

Human pose estimation is used to localize the human joints in images or videos, obtain 2D or 3D coordinates of the joints, and reconstruct the coordinate representations of the human body. With the increase of computing resources and the amount of data, the pose estimation method using learning has gradually matured and had considerable success. We first apply AlphaPose [36–38], an accurate open-source pose estimator, to detect the 2D position in the image frames of videos. The AlphaPose system pretrained on the COCO database [39] is used for 2D keypoint detection. Then we project the 2D detected results into 3D coordinates, using a pretrained fully convolutional model based on dilated temporal convolutions proposed by Pavllo et al. [40]. In this way, the dynamic skeleton representations of the human body in form of 3D coordinates are obtained.

### *3.2. Data Preprocessing*

In the 2D pose estimation results, there is high-frequency noise because of the image quality and the estimation error, which causes instability of the 3D coordinates. We add a low-pass filter after 2D joint position detection to filter out the noise and obtain clean skeleton data.

Since the lower halves of the actors' bodies are not visible in the video clips most of the time, the reliability of the estimation results of this part of the body is much lower than that of the upper body. Therefore, the lower body part is removed from the skeleton graph and only 10 joints of the upper body are considered in our research.

## 4. Proposed Networks

### *4.1. Graph Convolutional Network*

#### 4.1.1. Skeleton Graph Construction

The extracted body skeleton in each frame is represented in the form of 3D coordinate vectors of human joints. The dynamic motion data of each sample is a time sequence with different timesteps because of the variable lengths of the video clips. Following the work of [10], we construct an undirected spatial-temporal graph $G$ composed of vertexes $V$ and edges $E$, represented as $G = (V, E)$. The spatial-temporal graph is a representation of body movement along the spatial and the temporal dimensions and retains gestural static and dynamics characteristics, e.g., movement speed and amount of movement, which are helpful in recognizing emotions accurately [13,41]. Figure 1a shows the spatial-temporal graph of a skeleton sequence, where the vertexes denote the joints of the upper body, the blue lines represent the spatial edges based on the natural connection of human joints, and the green lines indicate the temporal edges that exist between two consecutive timesteps of the same joints. As illustrated in Figure 1b, the dots of different colors (in the dashed box) represent different subsets of the adjacency matrix of the graph based on the partitioning strategy. The root node and its neighboring nodes are divided into three subsets according to their distance from the body center (spinebase), i.e., the root node, the nodes closer than the root node, and the nodes farther than the root node.

(**a**) Graph architecture (**b**) Partitioning

**Figure 1.** Skeleton graph construction. (**a**) The spatial-temporal graph contains vertexes (blue dots), spatial edges (blue lines), and temporal edges (green lines). (**b**) The dots of different colors in the dashed box represent 3 subsets according to the spatial configuration partitioning strategy, that is, the root node (green), the nodes closer than the root node (yellow), and the nodes farther than the root node from the spinebase joint (blue).

#### 4.1.2. Self-Attention Enhanced Spatial Graph Convolutional Layer

As described in Section 1, in the work of Yan et al. [10], the calculation of graph convolution is based on the fixed topology of the graph, which may not be appropriate for emotion recognition tasks. To solve the problem, not only local neighborhood nodes that are predefined by the natural connection of the human body but also other nodes with high relevance of information need to be considered in the process of message passing. Therefore, we propose a self-attention enhanced spatial graph convolutional layer, in which the self-attention mechanism calculates the weighted sum of the values of all nodes to aggregate features from the entire graph and provide supplementary information for the spatial graph convolution module.

The structure of our self-attention enhanced spatial graph convolutional layer is illustrated in Figure 2. The layer is composed of three parts: the graph convolutional part, the self-attention part, and the gating mechanism.

Graph convolutional part: For the spatial graph convolutional part, the structure of the graph is predefined by the the adjacency matrix $\mathbf{A}$. According to the partitioning strategy, $\mathbf{A}$ is divided into $K$ subsets, i.e., $\mathbf{A} = \sum_k^K \mathbf{A}_k$, $\mathbf{A}_k \in \mathbb{R}^{V \times V}$. The input $\mathbf{f}_{in}$ is a tensor with the shape $(C_{in} \times T \times V)$. First, a convolution operation is applied for the input:

$$\mathbf{f} = \mathbf{W}_k \mathbf{f}_{in}, \tag{1}$$

where $\mathbf{W}_k \in \mathbb{R}^{(K \times C_{out}) \times C_{in} \times 1 \times 1}$ is a trainable matrix of the $1 \times 1$ convolution. Similarly, $\mathbf{f}$ is also divided into $K$ subsets, i.e., $\mathbf{f} = (\mathbf{f}_1, ..., \mathbf{f}_K)$, $\mathbf{f}_k \in \mathbb{R}^{C_{out} \times T \times V}$. The output of the spatial graph convolutional part is calculated as:

$$\mathbf{f}_g = \sum_k^K \mathbf{f}_k (\mathbf{A}_k \otimes \mathbf{M}_k), \tag{2}$$

where $\mathbf{M}_k \in \mathbb{R}^{V \times V}$ is a trainable weight matrix, and $\otimes$ denotes the element-wise product between matrices.

Self-attention part: The self-attention part applies a multi-head scaled dot product attention on the graph. In the field of Natural Language Processing, Vaswani et al. proposed multi-head self-attention, in which the queries and the corresponding keys are used to compute the weight and the weighted sum of the values is calculated as the output [42]. The self-attention model could be regarded as a spatial graph neural network with a fully-connected graph, in which the feature vectors of joints are nodes and every pair of nodes is connected. The model aggregates features from every node of the graph instead of neighborhood nodes when updating representations of nodes. First, the input is reshaped as a tensor with the shape $(T \times V \times C_{in})$, where $T$ is the temporal length, $V$ is the number of joints, and $C_{in}$ is the number of input channels.

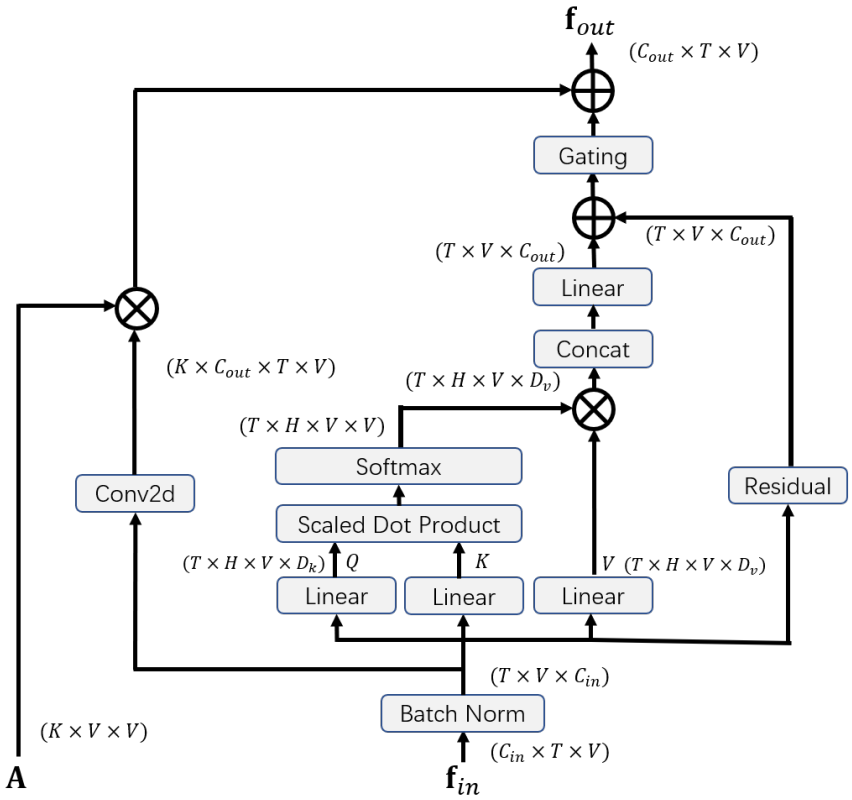


**Figure 2.** Illustration of self-Attention enhanced spatial graph convolutional layer.

For each timestep in one sample, we represent the feature vector of each joint as $a_i \in \mathbb{R}^{C_{in}}$, whose amount is V. The set of the vectors corresponding to the joint representations is $A = \{a_1, ..., a_V\}$. Then the similarity between the features of query $i$ and the features of key $j$ is calculated by the scaled dot product for each head:

$$e_{ij}^{head} = \frac{(a_i W_Q^{head})(a_j W_K^{head})^T}{\sqrt{d_k}}, \tag{3}$$

where $W_Q^{head} \in \mathbb{R}^{C_{in} \times d_k}$ and $W_K^{head} \in \mathbb{R}^{C_{in} \times d_k}$ are learnable matrices for one head, $d_k$ is the dimension of query. After this, the softmax function is applied to obtain the attention coefficient $\alpha$:

$$\alpha_{ij}^{head} = \frac{\exp e_{ij}^{head}}{\sum_{k=1}^{V} \exp e_{ik}^{head}}. \tag{4}$$

The weighted sum of values for each head is $\mathbf{s}^{head} = (\mathbf{s}_1^{head}, ..., \mathbf{s}_V^{head}) \in \mathbb{R}^{V \times d_v}$, where $d_v$ is the dimension of value. The calculation of $\mathbf{s}_i^{head}$ is formulated as:

$$\mathbf{s}_i^{head} = \sum_{j}^{V} \alpha_{ij}^{head} (a_j W_V^{head}), \tag{5}$$

where $W_V \in \mathbb{R}^{C_{in} \times d_v}$ is a learnable matrix. Then we calculate the self-attention result $\mathbf{f}_o$:

$$\mathbf{f}_o = Concat(\mathbf{s}^1, ..., \mathbf{s}^H) W_O, \tag{6}$$

where $H$ is the number of attention heads, and $W_O \in \mathbb{R}^{(H \times d_v) \times C_{out}}$ is a parameter matrix. To stabilize the training procedure, a residual is added for the output of the self-attention part:

$$\mathbf{f}_a = \mathbf{f}_o + Residual(\mathbf{f}_{in}), \tag{7}$$

$$Residual(\mathbf{f}_{in}) = \begin{cases} \mathbf{f}_{in}W_R & C_{in} \neq C_{out} \\ \mathbf{f}_{in} & C_{in} = C_{out} \end{cases}, \tag{8}$$

where $W_R \in \mathbb{R}^{C_{in} \times C_{out}}$ is a parameter matrix that transforms the number of input channels to the number of output channels.

Gating mechanism: Considering that the contributions of the two parts may be different for updating representations of nodes, we use a trainable coefficient $r$ to adjust the weight of the self-attention part, which is formulated as:

$$\mathbf{f}_{out} = \frac{\mathbf{f}_g + r \times \mathbf{f}_a}{2}, \tag{9}$$

where the trainable coefficient $r$ is initialized as 1.

#### 4.1.3. Self-Attention Enhanced Spatial Temporal Graph Convolutional Network

The basic block of the self-attention enhanced spatial temporal graph convolutional network (S-STGCN) is composed of a self-attention enhanced spatial graph convolutional layer, a temporal convolutional layer, and several functional layers (See Figure 3). The self-attention enhanced spatial graph convolutional layer is used to aggregate the information of the joints along the spatial dimension. The temporal convolutional layer of the network is the same as that of the ST-GCN, i.e., apply the convolution with receptive field $(K_t, 1)$ on the feature vectors of each node along the temporal dimension.

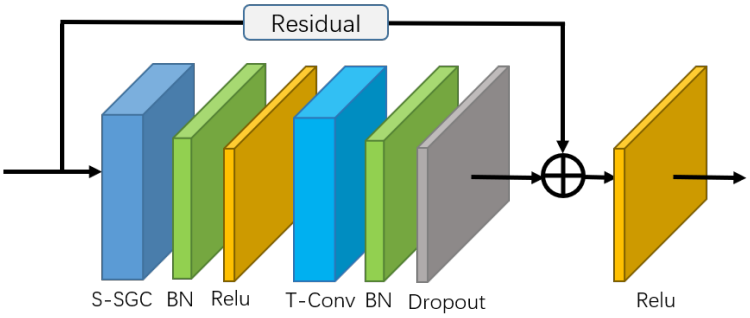


**Figure 3.** Illustration of basic graph convolutional block. The residual connection is applied to guarantee the stability of training. S-SGC represents the self-attention enhanced spatial graph convolutional layer. T-Conv represents the temporal convolutional layer.

As illustrated in Figure 4, there are 10 basic blocks in the model, with output channels of 32, 32, 32, 32, 64, 64, 64, 128, 128, and 128. After that, the output tensor is fed into a global average pooling layer to get an emotional feature vector for each sample. Finally, the vectors are passed into the output layer with a softmax function to obtain the prediction of the emotion classes.

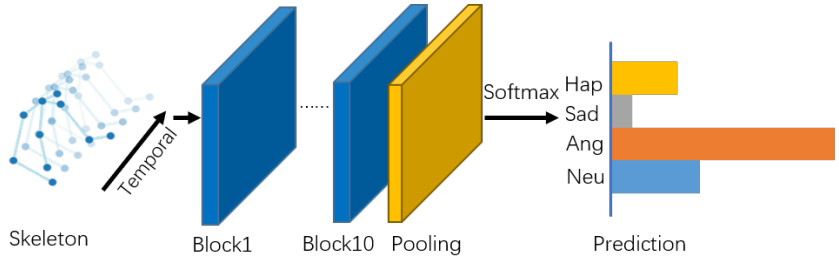


**Figure 4.** Overview of our proposed self-attention enhanced spatial temporal graph convolutional network.

#### 4.1.4. Two-Stream Architecture

In some previous works of action recognition [11,43], in addition to the joint positions, the second-order feature, i.e., the bone information representing the lengths and orientations of the human bones, has also been proved to be useful for skeleton-based action recognition tasks. Considering the bone information may also play an important role in the recognition of emotions in a similar way, we construct a two-stream network to simultaneously use the joint information and bone information.

Similar to the work of [11], we define the bone features as the vectors from the source nodes to the target nodes. Each bone is between two adjacent joints, in which the joint close to the center (spinebase) of the skeleton is defined as the source node, and the joint far away from the center is defined as the target node. For example, the bone vector between the source node $\mathbf{v}_1$ and the target node $\mathbf{v}_2$ is represented as $\mathbf{e}_{1,2} = \mathbf{v}_2 - \mathbf{v}_1$.

The architecture of the two-stream self-attention enhanced spatial temporal graph convolutional network (2s-S-STGCN) is illustrated in Figure 5. The joint data representing the positions of the joints and the bone data representing the lengths and directions of the bones are fed into two S-STGCNs and each stream is trained respectively. Then the output tensors of the two streams are fused to predict emotion labels.

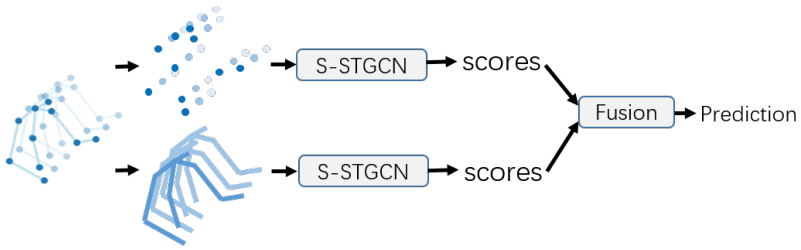


**Figure 5.** Two-stream network. The two self-attention enhanced spatial temporal graph convolutional networks (S-STGCNs) take the joint information and the bone information as input respectively. The fusion of the outputs of these models are used for final prediction.

### *4.2. Multimodal Emotion Recognition Network*

We construct a skeleton enhanced emotion recognition network (SERN), which integrates text and audio information with the features extracted by the self-attention enhanced spatial temporal graph convolutional network (See Figure 6). The multimodal dual recurrent encoder (MDRE) [44], composed of an audio recurrent encoder (ARE) and a text recurrent encoder (TRE), is used to extract high-level representations of audio and text, which takes as input prosody features, MFCC features and text transcripts. In the ARE, the extracted MFCC features $MFCC^i$ ($i$ represents the timestep) are fed into a gated recurrent unit (GRU) for each timestep, and the final hidden state $h_A^t$ of the GRU is concatenated with the prosodic feature vector to generate the representation vector of audio. In the TRE, the tokens of the text transcripts are passed through a word embedding layer, and the embedded tokens $(token^1, ..., token^t)$ are also fed into a GRU to obtain the representation vector of text. The text representation vectors extracted by TRE and the audio representation vectors extracted by ARE are concatenated and used for bimodal emotion recognition.

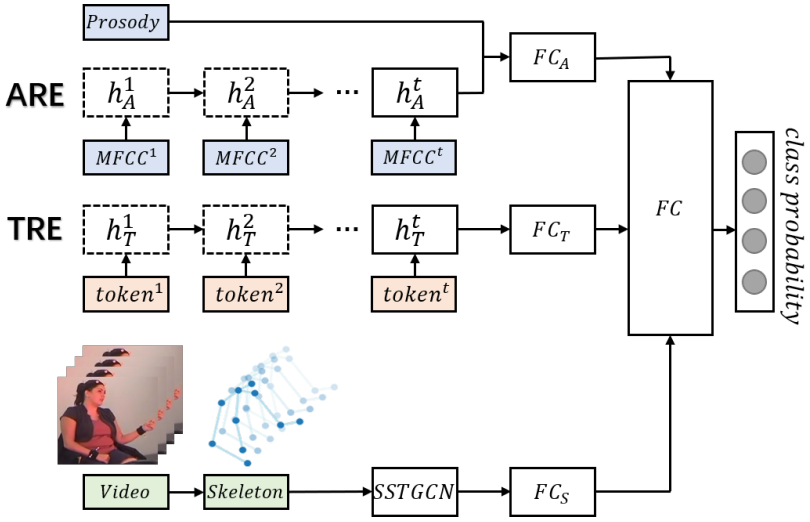


**Figure 6.** Skeleton enhanced emotion recognition network. Each *FC* represents a fully connected layer. The three fully connected layers on the left ($FC_A, FC_T, FC_S$) are used for adjusting the dimensions of the representation vectors of audio, text, and skeleton. The fully connected layer on the right is used to learn the weights of these features and make final prediction.

We propose a two-phase hierarchical network to simultaneously fuse the audio, text, and gesture information. In the first phase, the uni-modal features are fed into the ARE, the TRE, and the self-attention enhanced spatial temporal graph convolutional network, respectively, to obtain the audio, text, and gesture representations. In the second phase, these representations are passed through the fully connected layers and then concatenated to pass to the output layer to predict emotion.

## 5. Experiment

### *5.1. Dataset*

We usde the IEMOCAP database [6] in the experiment. The IEMOCAP database records audio and video data when two actors have a dialogue in hypothetical or scripted scenarios. Every utterance is annotated into 10 classes by three or four annotators, i.e., happy, sad, angry, surprised, afraid, disgusted, frustrated, excited, or other. However, for body movement, only motion capture data of the head and hands motions, instead of body skeleton data, is contained in the database, thus ignoring some parts that are also crucial for emotional expression, e.g., the spine, the shoulders, and the arms.

To keep consistent with previous work, we merged the samples with excitement labels into the happiness subset and used four emotion classes, i.e., neutral, happy, sad, and angry. The dataset used in the experiment contained a total of 5492 samples in those four classes (1606 happy, 1081 sad, 1102 angry, and 1703 neutral). Some examples of the dataset are shown in Figure 7.

(**a**) Happy (**b**) Sad (**c**) Angry (**d**) Neutral

**Figure 7.** Illustration of examples of four emotions from IEMOCAP database [6].

### *5.2. Feature Extraction and Experiment Setting*

Following the work of [44], we used the OpenSMILE toolkit [45] to extract MFCC features and prosodic features and to initialize the tokens with the pretrained 300-dimensional GloVe vector [46]. The samples were divided into a training set, a development set, and a test set, with a ratio of 8: 0.5: 1.5 in the training process. The cross-entropy loss was used as the loss function to back-propagation.

### *5.3. Results*

The performance of the proposed models and baseline models is shown in Table 1. We list the unweighted average recall (UAR) and the weighted average recall (WAR) of the unimodal and multimodal models utilizing audio signals, textual content, and body skeleton data. Considering the unbalance of the samples, the unweighted average recall was used to evaluate the model by treating each category equally. Since the value of the unweighted average recall was easily affected by the rare category, the weighted average recall was also applied to measure the overall prediction performance, which was numerically equal to the accuracy of the model.

**Table 1.** Comparison with baselines. We compare the performance of both the unimodal and multimodal models, for which A is the audio modality, T is the text transcription, and S is the skeleton data. SERN-2s is a skeleton enhanced emotion recognition network whose S-STGCN is replaced with 2s-S-STGCN.

| Model | Modality | UAR(%) | WAR(%) |
|---|---|---|---|
| ARE [44] | A | 59.7 | 57.1 |
| TRE [44] | T | 65.9 | 64.5 |
| ResNet18 [47] | S | 55.3 | 56.9 |
| ST-GCN [10] | S | 63.3 | 63.7 |
| S-STGCN | S | 68.4 | 68.4 |
| 2s-S-STGCN | S | **73.1** | **72.5** |
| MDRE[44] | A + T | 72.0 | 71.4 |
| SERN | A + T + S | 80.4 | 80.0 |
| SERN-2s | A + T + S | **82.2** | **82.3** |

For skeletal movement, ResNet18 [47] and ST-GCN [10] were selected as baseline models. ResNet is a convolutional network that adds a residual mechanism to the traditional convolutional neural network, which has been widely used in many tasks. In the experiment, we apply ResNet18 and ST-GCN to the skeleton data and compare their performance with our proposed model. Our S-STGCN model achieved a UAR of 68.4% and a WAR of 68.4%, significantly outperforming other networks in the emotion recognition task, which shows that the proposed model, especially the self-attention enhanced graph convolutional layer, could effectively extract emotion-related features from the dynamic skeleton sequence. It shows that the self-attention part obtained more flexible representations and provided effective supplementary information for the spatial graph convolution with the predefined graph. The two-stream architecture that integrated the joints and bones brought remarkable improvement, which indicates that this method could extract the emotional information from the skeleton modality more effectively.

We also compared the trimodal SERN with the MDRE [44] that was performed as the basic model of the audio and text modalities. Similarly, we applied the UAR and the WAR to the models as performance metrics. The model using text, audio, and skeleton information exceeded the MDRE that does not use skeleton information by a UAR of 8.4% and a WAR of 8.6%. The results show that the new extracted body skeleton modality contained some emotional information that was not contained in the audio and text modalities, and the use of the skeleton data enhanced the emotion recognition of the model. Moreover, we further compared a multimodal model using 2s-S-STGCN to extract skeleton features with

the multimodal model using single-stream S-STGCN. The result showed that the use of the two-stream method also brought the improvement of the performance in multimodal emotion recognition.

## 6. Ablation Study and Discussion

### *6.1. Effect of the Preprocessing*

In the data preprocessing (See Section 3.2), we applied the low-pass filter for the estimated position data of each joint to reduce the impact of the high-frequency noise. The performance of the preprocessed and unpreprocessed data is listed in Table 2. The result shows that the noise in the estimation result was harmful for skeleton-based emotion recognition and that the preprocessing brings considerable improvement.

**Table 2.** Comparison between preprocessed and unpreprocessed data. The unpreprocessed noisy data and the preprocessed data are fed into S-STGCN respectively to compare their performance.

| Data | UAR(%) | WAR(%) |
|---|---|---|
| Unpreprocessed | 66.5 | 66.4 |
| Preprocessed | 68.4 | 68.4 |

### *6.2. Gating Mechanism*

In Section 4.1.2, we introduced the gating mechanism to adjust the weight of the graph convolutional part and the self-attention part. In order to confirm the effect of the gating mechanism, the coefficient $r$ was set to constant 1 in the ablation experiment. Thus Equation (9) was transformed into:

$$\mathbf{f}_{out} = \frac{\mathbf{f}_g + \mathbf{f}_a}{2}. \tag{10}$$

The result is shown in Table 3. The model using the gating mechanism obtained better prediction results, which suggests that the gating mechanism could flexibly adjust the importance of the self-attention part and was beneficial to the improvement of performance.

**Table 3.** Comparison for the use of the gating mechanism.

| Model | UAR(%) | WAR(%) |
|---|---|---|
| S-STGCN w/o G | 67.7 | 67.3 |
| S-STGCN | 68.4 | 68.4 |

We also visualized the coefficient $r$ of each layer in Figure 8. The weights of the self-attention parts in the last several layers were higher than those in the other layers, and the weights in the last two layers were more than 1, which indicates that the features extracted by the self-attention part in the top layers were more informative than that of the bottom layers and that in last two layers the self-attention parts even were more important than the spatial graph convolutional part for the feature extraction.

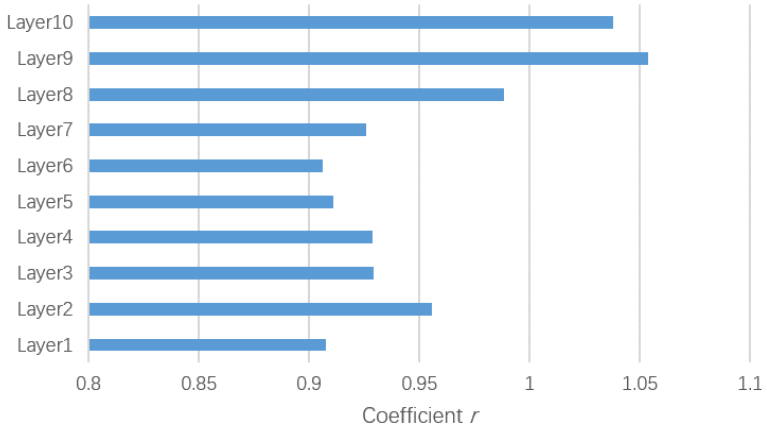


**Figure 8.** Visualization of the coefficient $r$ of each layer.

### *6.3. Effect of the Bone Information*

As introduced in Section 4.1.4, we fused the joint and the bone information to make the final prediction. Here we compared two fusion strategies: (i) The softmax scores of the two models were directly added as the final output. (ii) The scores were concatenated and then fed into a fully connected layer to obtain the result.

The experimental result is shown in Table 4. For single-input models, the performance of the joint model was slightly lower than that of the bone model. For the two-stream models with different fusion strategies, the trained model with the fully connected layer was better than the model using the summation strategy. Both of the two-stream models outperformed the single-input model, which shows the effectiveness of the two-stream architecture and the complementarity between the joint and bone information.

**Table 4.** Performances of different inputs and fusion strategies. J represents the joint information. B represents the bone information. 2s-S-STGCN-a is the two-stream model using the summation fusion strategy. 2s-S-STGCN-c is the two-stream model using the fully connected layer.

| Model | Input | UAR(%) | WAR(%) |
|---|---|---|---|
| S-STGCN | J | 68.4 | 68.4 |
| S-STGCN | B | 68.8 | 68.9 |
| 2s-S-STGCN-a | J&B | 71.9 | 72.3 |
| 2s-S-STGCN-c | J&B | **73.1** | **72.5** |

From the confusion matrix in Figure 9, it can be seen that the joint and bone modalities both performed best in the prediction of sad emotions. For joint information, the anger class was not well recognized, while for bone information, the samples of the happy class were often incorrectly recognized. For both of the single-stream models, other classes were frequently confused with the neutral class.

Overall, both of the two-stream models obtained better performance for most of the classes. The two-stream model using the summation of the scores also frequently incorrectly classified the samples of angry, happy, and sad classes as the neutral class. The other two-stream model that fed concatenation of the scores into the fully connected layer alleviated this problem to a certain extent and performed better on most of the categories.

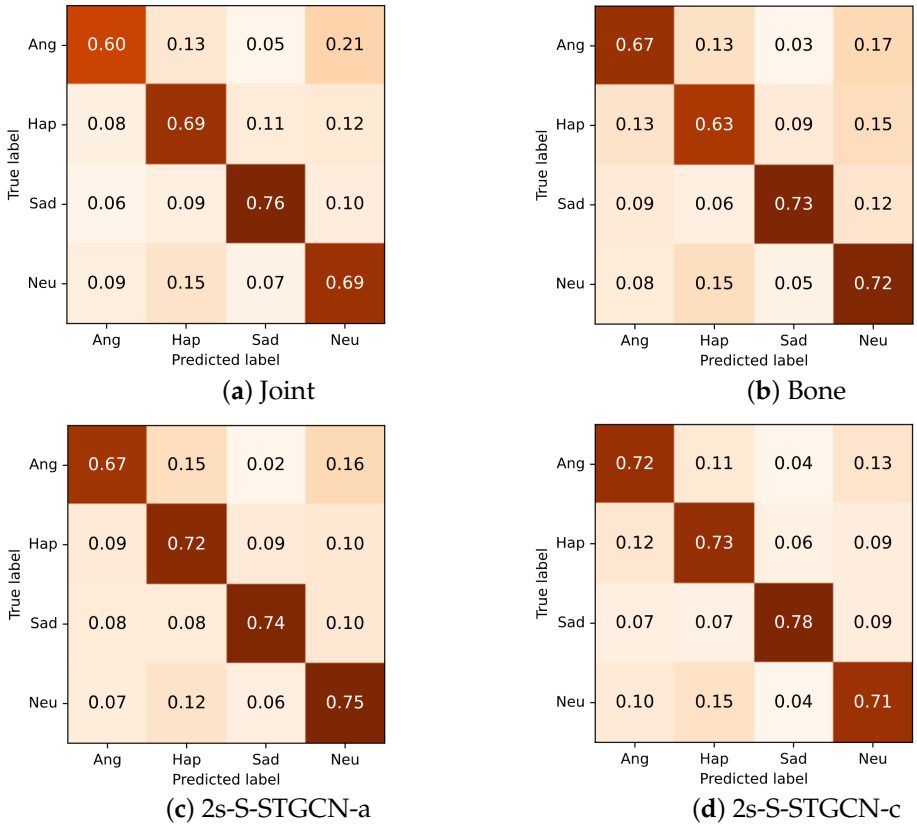


**Figure 9.** Confusion matrices for single-stream and two-stream models.

### *6.4. Multimodal Analysis*

Different modalities showed different characteristics in the ability of emotion prediction. Figure 10 shows the confusion matrices of the unimodal and multimodal models. For skeleton modality (Figure 10a), the model showed the best performance in predicting the sad class. We speculate that this may be related to the posture of the spine and head. When people expressed their sadness, they tended to bend over and lower the head, which was obviously different from happy, angry, and neutral. As shown in (Figure 10b), for the audio model, anger and happiness were often incorrectly recognized as each other. For text modality (Figure 10c), although anger and happiness were correctly distinguished, angry, happy, and sad classes were frequently misclassified as the neutral class. The multimodal network that integrated the information from the three modalities (Figure 10d) achieved high and balanced performance for every class by leveraging the strengths of unimodal models and compensating for their weaknesses. From the error analysis, it can be seen that each modality had its unique strengths.

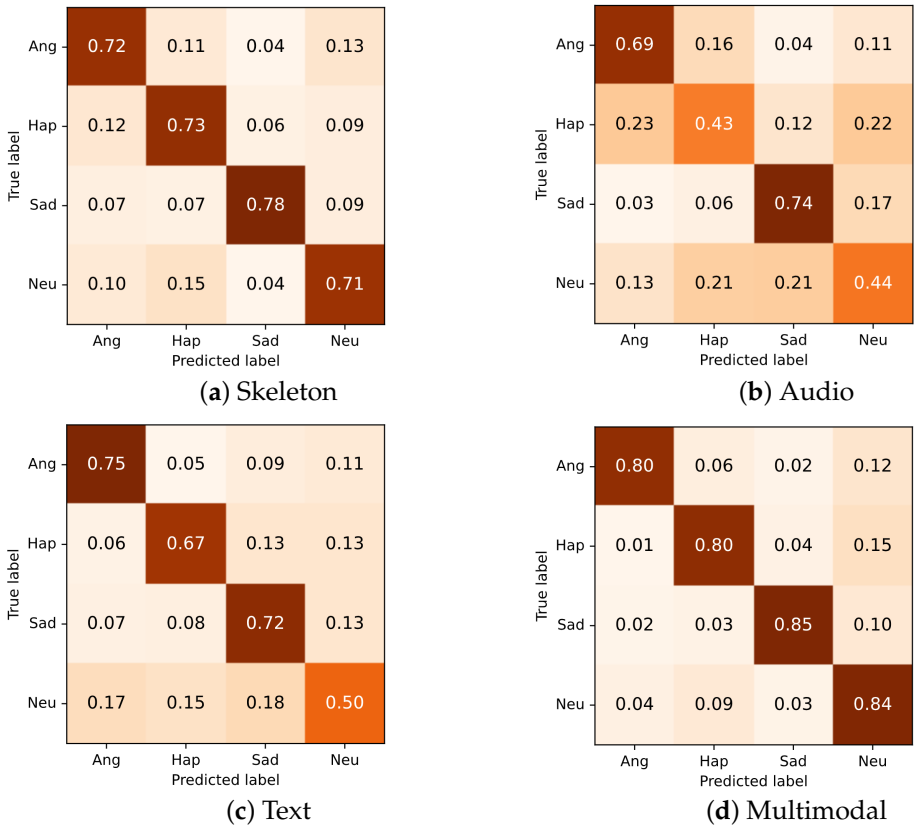


**Figure 10.** Confusion matrix for each modality. For skeleton modality, the two-stream self-attention enhanced spatial temporal graph convolutional network is used.

## 7. Conclusions

In this paper, we extracted body skeleton data using a pose estimation based approach from the videos of the IEMOCAP database. We proposed a novel two-stream self-attention enhanced spatial temporal graph convolutional network for emotion recognition based on the skeleton data. It models the body skeleton as a graph and constructs both static and flexible connections between the joints to update the representations. Through our experiment, we demonstrated that the performance of the proposed model is superior to that of the ST-GCN on emotion recognition tasks and that the two-stream architecture further improves the performance. Besides, we integrated information from audio signals, text transcripts, and skeletal movement, showing that our method outperforms the bimodal model, which indicates that the extracted skeleton data can provide important supplementary information for emotion detection. Since the multimodal fusion strategy that directly concatenates different modalities may not be optimal, we will focus on exploring more multimodal fusion strategies. We also plan to generate emotional gestures using the extracted skeleton data in our future work.

**Author Contributions:** Conceptualization, J.S., C.L. and C.T.I.; methodology, J.S. and C.L.; software, J.S.; validation, J.S. and C.L.; formal analysis, J.S., C.L. and C.T.I.; supervision, C.T.I. and H.I.; project administration, H.I.; funding acquisition, H.I. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partly supported by the Grant-in-Aid for Scientific Research on Innovative Areas JP20H05576, and the JST ERATO Grant Number JPMJER1401.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Noroozi, F.; Kaminska, D.; Corneanu, C.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* **2018**. [CrossRef]
2. Ahmed, F.; Bari, A.H.; Gavrilova, M.L. Emotion Recognition From Body Movement. *IEEE Access* **2019**, *8*, 11761–11781. [CrossRef]
3. Wallbott, H.G. Bodily expression of emotion. *Eur. J. Soc. Psychol.* **1998**, *28*, 879–896. [CrossRef]
4. Sapiński, T.; Kamińska, D.; Pelikant, A.; Ozcinar, C.; Avots, E.; Anbarjafari, G. Multimodal database of emotional speech, video and gestures. In Proceedings of the International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018; pp. 153–163.
5. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Multimodal emotion recognition using deep learning architectures. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
6. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
7. Sapiński, T.; Kamińska, D.; Pelikant, A.; Anbarjafari, G. Emotion recognition from skeletal movements. *Entropy* **2019**, *21*, 646. [CrossRef] [PubMed]
8. Filntisis, P.P.; Efthymiou, N.; Koutras, P.; Potamianos, G.; Maragos, P. Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction. *IEEE Rob. Autom Lett.* **2019**, *4*, 4011–4018. [CrossRef]
9. Ly, S.T.; Lee, G.S.; Kim, S.H.; Yang, H.J. Gesture-Based Emotion Recognition by 3D-CNN and LSTM with Keyframes Selection. *Int. J. Contents* **2019**, *15*, 59–64.
10. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv* **2018**, arXiv:1801.07455.
11. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12026–12035.
12. Cai, Y.; Huang, L.; Wang, Y.; Cham, T.J.; Cai, J.; Yuan, J.; Liu, J.; Yang, X.; Zhu, Y.; Shen, X.; et al. Learning Progressive Joint Propagation for Human Motion Prediction. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 226–242.
13. Dael, N.; Goudbeek, M.; Scherer, K.R. Perceived gesture dynamics in nonverbal expression of emotion. *Perception* **2013**, *42*, 642–657. [CrossRef]
14. Schwarz, N. Emotion, cognition, and decision making. *Cogn. Emot.* **2000**, *14*, 433–440. [CrossRef]
15. Kensinger, E.A. Negative emotion enhances memory accuracy: Behavioral and neuroimaging evidence. *Curr. Directions Psychological Sci.* **2007**, *16*, 213–218. [CrossRef]
16. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
17. Kołakowska, A.; Landowska, A.; Szwoch, M.; Szwoch, W.; Wrobel, M.R. Emotion recognition and its applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*; Springer: Cham, Switzerland, 2014; pp. 51–62.
18. Franzoni, V.; Milani, A.; Nardi, D.; Vallverdú, J. Emotional machines: The next revolution. *Web Intell.* **2019**, *17*, 1–7. [CrossRef]
19. Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; Picard, R.W. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–30. [CrossRef]
20. Yoon, S.; Dey, S.; Lee, H.; Jung, K. Attentive modality hopping mechanism for speech emotion recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3362–3366.
21. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
22. Heusser, V.; Freymuth, N.; Constantin, S.; Waibel, A. Bimodal Speech Emotion Recognition Using Pre-Trained Language Models. *arXiv* **2019**, arXiv:1912.02610.
23. Kaza, K.; Psaltis, A.; Stefanidis, K.; Apostolakis, K.C.; Thermos, S.; Dimitropoulos, K.; Daras, P. Body motion analysis for emotion recognition in serious games. In Proceedings of the International Conference on Universal Access in Human-Computer Interaction, Toronto, ON, Canada, 17–22 July 2016; pp. 33–42.
24. Ahmed, F.; Gavrilova, M.L. Two-layer feature selection algorithm for recognizing human emotions from 3d motion analysis. In Proceedings of the Computer Graphics International Conference, Calgary, AB, Canada, 17–20 June 2019; pp. 53–67.
25. Karumuri, S.; Niewiadomski, R.; Volpe, G.; Camurri, A. From Motions to Emotions: Classification of Affect from Dance Movements using Deep Learning. In Proceedings of the the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–6.
26. Deng, J.J.; Leung, C.H.C.; Mengoni, P.; Li, Y. Emotion recognition from human behaviors using attention model. In Proceedings of the 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 26–28 September 2018; pp. 249–253.
27. Zhou, J.; Cui, G.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *arXiv* **2018**, arXiv:1812.08434.

28. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**. [CrossRef]
29. Bastings, J.; Titov, I.; Aziz, W.; Marcheggiani, D.; Sima'an, K. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv* **2017**, arXiv:1704.04675.
30. Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W.L.; Leskovec, J. Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018; pp. 974–983.
31. Hu, F.; Zhu, Y.; Wu, S.; Wang, L.; Tan, T. Hierarchical graph convolutional networks for semi-supervised node classification. *arXiv* **2019**, arXiv:1902.06667.
32. Bruna, J.; Zaremba, W.; Szlam, A.; LeCun, Y. Spectral networks and locally connected networks on graphs. *arXiv* **2013**, arXiv:1312.6203.
33. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3844–3852.
34. Monti, F.; Boscaini, D.; Masci, J.; Rodola, E.; Svoboda, J.; Bronstein, M.M. Geometric deep learning on graphs and manifolds using mixture model cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5115–5124.
35. Hamilton, W.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the Advances in neural information processing systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1024–1034.
36. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
37. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.S.; Lu, C. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. *arXiv* **2018**, arXiv:1812.00324.
38. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient Online Pose Tracking. *arXiv* **2018**, arXiv:1802.00977.
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European conference on computer vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
40. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7753–7762.
41. Coulson, M. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *J. Nonverbal Behav.* **2004**, *28*, 117–139. [CrossRef]
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in neural information processing systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
43. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *arXiv* **2019**, arXiv:1912.06971.
44. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118.
45. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
46. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014, pp. 1532–1543.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.