electronics

MDPI

Article

Facial Landmark-Based Emotion Recognition via Directed Graph Neural Network

Quang Tran Ngoc, Seunghyun Lee and Byung Cheol Song *

Department of Electronic Engineering, Inha University, Incheon 22212, Korea; quangtrandn93@gmail.com (Q.T.N.); lsh910703@gmail.com (S.L.)

* Correspondence: bcsong@inha.ac.kr; Tel.: +82-32-860-7413

Received: 8 April 2020; Accepted: 2 May 2020; Published: 6 May 2020

Abstract: Facial emotion recognition (FER) has been an active research topic in the past several years. One of difficulties in FER is the effective capture of geometrical and temporary information from landmarks. In this paper, we propose a graph convolution neural network that utilizes landmark features for FER, which we called a directed graph neural network (DGNN). Nodes in the graph structure were defined by landmarks, and edges in the directed graph were built by the Delaunay method. By using graph neural networks, we could capture emotional information through faces' inherent properties, like geometrical and temporary information. Also, in order to prevent the vanishing gradient problem, we further utilized a stable form of a temporal block in the graph framework. Our experimental results proved the effectiveness of the proposed method for datasets such as CK+ (96.02%), MMI (69.4%), and AFEW (32.64%). Also, a fusion network using image information as well as landmarks, is presented and investigated for the CK+ (98.47% performance) and AFEW (50.65% performance) datasets.

Keywords: facial emotion recognition; facial landmark; graph neural network

1. Introduction

Emotion recognition has been widely studied in various areas of computer vision as well as human–computer interactions (HCI). For a long time, various emotion recognition techniques which utilize different modalities such as video streams, audio signals, and bio-signals have been proposed. Kuo et al. [1] extracted appearance and geometry features from image sequences and combined them via a joint fine-tuning. Hossain et al. [2] developed a 2D convolutional neural network (CNN) architecture for an audio signal database. Zhang et al. [3] combined temporal information from EEG signals and spatial information from facial images for human emotion recognition.

Facial landmarks provide good information to analyze facial emotions. Yan et al. [4] defined facial landmarks as derivatives of action units (AUs) for describing facial muscle movements. Other studies [5,6] proposed several fusion mechanisms of landmarks and images. A work [7] introduced a fusion approach of landmarks and videos. Fabiano et al. [8] proposed a deformable synthesis model (DSM), which used 3D facial landmarks. These algorithms show that the landmark feature is effective; however, emotion recognition algorithms using landmark features have been rarely studied recently. This is due to the fact that appropriate tools for obtaining information from landmark features have not been properly selected rather than to the assumption that the information provided by landmark features is insufficient.

The above-mentioned approaches are too naïve to capture useful information from landmarks. They use typical deep neural networks such as multilayer perceptron (MLP) and CNN to analyze landmark features. However, conventional neural networks are not appropriate to capture geometrical information from facial landmarks. Furthermore, recurrent neural networks (RNNs) may cause a vanishing problem, and MLP cannot utilize temporal information. In general, CNN, RNN, and MLP

are good at extracting dense relations of features. However, since the relation between landmarks is so limited, these networks are not appropriate for extracting such information. Therefore, a third-party approach is required to analyze landmark features.

We gained insight into how to solve these problems from the similarity between landmarks and skeletons. Each skeleton point corresponds to a joint on a human body, and skeletons can be represented by a graph structure. Therefore, skeleton information is usually represented by graph convolution neural networks (GCNN). For instance, Zhang et al. [9] provided a spatio-temporal skeleton model which combines spatially neighboring edges and temporally neighboring edges and defined a new edge representation based on the skeleton model for action recognition task. Gao et al. [10] proposed a graph neural network (GNN) to extract intrinsic physical connections and spatial connectivity of skeleton data. Note that landmark data are semantically similar to skeleton data. The landmarks are connected to describe local patterns such as eyes, mouth, lips. Thus, landmark data can also be described by a graph structure. Consequently, a graph-based formulation can be an appropriate method to extract good emotion information from face images.

In this paper, we present a graph-based representation of facial landmarks via GNN and propose a facial emotion recognition (FER) algorithm using the graph-based representation. First, landmarks were mapped to vertices, and edges were built by the Delaunay method. Also, we generated a master node [11] for smooth information transfer between distant vertices. Next, vertices and edges fed into a DGNN. On the other hand, a straightforward temporal connection caused a vanishing gradient problem. So, we employed gated linear unit (GLU) to address the vanishing gradient problem. The conceptual view of the proposed method is shown in Figure 1. With various experiments using landmark features from the CK+ dataset, the proposed method was proven to provide higher performance than previous landmark-based FER methods. Furthermore, a score-level fusion of a landmark-based network and an image-based network achieved the outstanding accuracy of 98.7%, which is 1.45% higher than that reported in a previous study [5].

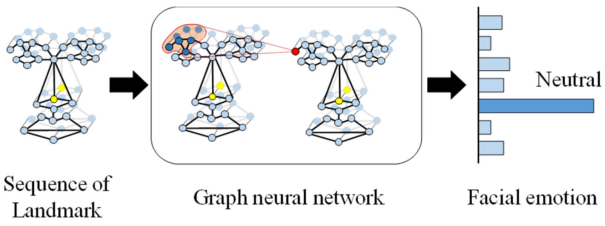

Figure 1. Conceptual figure of the proposed method. We constructed a sequence of landmark features to recognize facial emotions and adopted a graph neural network (GNN) to extract structural information.

The major contributions of the proposed method can be described in three aspects:

- We propose a new GNN structure with landmark features as input. The landmark features were used to construct a graph.
- A master node for better graph construction is presented.
- GLU is employed as a temporal block for boosting the performance of GNN.

2. Related Work

2.1. Landmark-Based Emotion Recognition

Various FER algorithms using landmark features have been developed. Rohr et al. [12] found that landmarks are prominent geometric features sparsely scattered over images and defined landmarks as a set of positions for analyzing emotion features. Ghimire et al. [13] defined specific handcrafted features based on landmarks and proposed an AdaBoost-based FER scheme which achieved a reasonable performance. Jung et al. [5] used a deep temporal geometry network (DTGN) framework to catch geometrical motion information of landmark points and predicted emotional labels via

joint fine-tuning. Hasani et al. [7] considered temporal changes in the network through landmark features. Previous studies showed that landmarks are important modalities for the recognition of facial emotion. For example, a study [5] demonstrated performance improvement by associating the landmark modality with other modalities.

However, compared to video-based FERs, landmark-based FERs have been rarely studied recently. This is because a proper deep learning model to extract useful information from landmarks has not yet been found. So far, neural networks developed to extract image features have been directly applied to landmark features without any analysis. For example, MLP is useful for analyzing dense relations between feature points. However, landmark points have sparse relations with each other. Therefore, a new algorithm is needed to analyze arbitrary relations between landmark points.

2.2. Graph Neural Network

Nowadays, GNN has been studied for various tasks. For example, Ma et al. [14] proposed a multi-dimensional graph convolutional GNN (mGCN) to capture node information of an entire graph. In addition, You et al. [15] proposed a graph convolutional policy network based on directed graph generation through reinforcement learning. The model discovered novel molecules with desired properties such as drug-likeness and synthetic accessibility. Furthermore, Li et al. [16] used a directed graph to capture both spatial and temporal dependencies among time series as well as address the traffic forecasting problem and long-term forecasting. Likewise, Li et al. [17] proposed a novel spatio-temporal graph routing (STGR) to use skeleton data for the action recognition task. For the STGR network, feature information was obtained by determining the connectivity among joints in subgroups of the spatial dimension. In addition, structural information was measured on the basis of the correlation degrees between temporal joint nodes.

Through a detailed analysis of GNN application, we found that GNN is very effective for sparse and arbitrary relational data, unlike CNN or MLP. We also found that data derived from the skeleton have very similar characteristics to landmark data. Each landmark point has a sparse relation with spatially adjacent points but has a continuous relation on the time axis. This is consistent with the characteristics of skeleton data. In STGR, it has been demonstrated that GNN effectively extracts such information. Therefore, we adopted GNN as a tool for obtaining landmark features.

3. Methods

This section describes the proposed directed graph neural network exploiting the landmark feature. The overall architecture of the proposed method is shown in Figure 2. First, we obtained landmarks from each frame via a conventional deep learning-based landmark extractor (see Section 3.1). Next, a graph structure was built based on landmark locations. In order to capture information among far distant nodes, we adopted a master node to graph the structure (see Section 3.2). Finally, we applied a directed graph neural network to analyze the graph structure (see Section 3.3). Also, GLU was added to prevent the vanishing gradient problem.

3.1. Landmark Features

We first describe a landmark extractor to recognize facial emotion. We used a deep learning-based method, following Dong et al. [18] to capture landmark features. The extracted landmark feature consisted of 68 parts on a human face, such as nose, eyes, eyebrow, and mouth. However, we found that the landmarks of the outer region had a negative effect on facial emotion recognition performance. Therefore, we used 51 landmarks as input features from each frame of a video, according to

$$\mathrm{LM} = \left\{ x_{t,p}, y_{t,p} \middle| 1 \le t \le T, 1 \le p \le P \right\} \tag{1}$$

where LM indicates a set of landmarks, $(x_{t,p}, y_{t,p})$ are the 2D coordinates of each landmark, and P and T denote the number of landmarks and the number of frames, respectively.

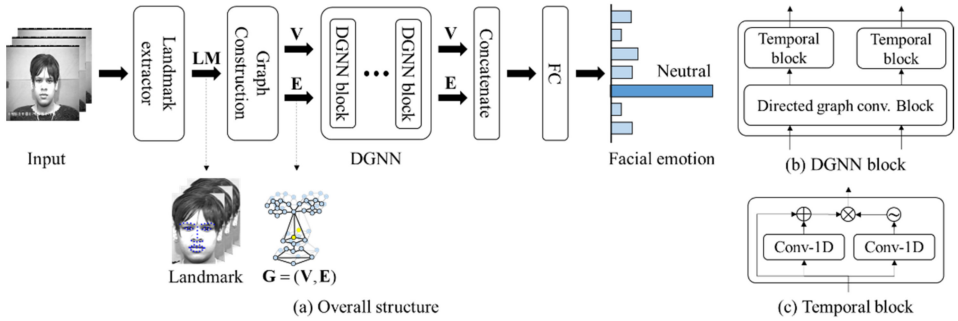

Figure 2. (**a**) Overview of the proposed method, (**b**) improved directed graph neural network (DGNN) block, and (**c**) gated linear unit (GLU). LM, **V**, **E**, **G**, and FC indicate landmark, vertex, edge, graph, and fully-connected layer, respectively.

3.2. Graph Construction

Each landmark portion presents a strong relationship with other specific landmark portions that are placed in a similar position or are connected by muscle. Therefore, the set of landmark locations can form a graph structure which can convey structure information. The proposed graph structure is shown in Figure 3.

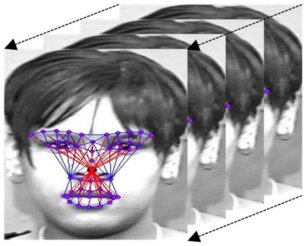

Figure 3. Depicting master nodes on a human face in a frame sequence. The nose node is considered a master node, then all nodes are connected to it. In this figure, the red edges are built on the basis of the master node, and the blue ones are built according to normal graph structure.

First, we defined a landmark location as a vertex. Each vertex is a 2D feature vector, which is by

$$\mathbf{V} = \left\{ \mathbf{v}_{t,p} \middle| 1 \le t \le T, 1 \le p \le P \right\} \tag{2}$$

$$\mathbf{v}_{i,p} = [x_{t,p}, y_{t,p}] \tag{3}$$

This approach is a simple way to define a vertex via location information. Its efficacy has been already proven in previous works [16,17], therefore we applied for our research.

Next, we constructed edges **E** using the Delaunay method [19]. The Delaunay method builds triangle meshes among all landmarks and [20] is useful to analyze facial emotions. However, a mesh structure only indicates whether edges are connected or not. Therefore, we multiplied the L_2 distance to represent edge strength, as follows:

$$\begin{aligned} \mathbf{E} &= \left\{ e_{t,i,j} \middle| 1 \le t \le T, 1 \le i, j \le P \right\} \\ &= \left\{ \left| \mathbf{v}_{t,i} - \mathbf{v}_{t,j} \right|_2 a_{t,i,j} \middle| 1 \le t \le T, 1 \le i, j \le P \right\} \end{aligned} \tag{4}$$

$$\begin{aligned} \mathbf{A} &= \mathrm{DM}(\left\{ x_{t,p} y_{t,p} \middle| 1 \le t \le T, 1 \le i, j \le P \right\}) \\ &= \left\{ a_{t,i,j} \middle| 1 \le t \le T, 1 \le i, j \le P \right\} \end{aligned} \tag{5}$$

where DM denotes the Delaunay method, and **A** denotes the adjacency matrix that contains binary values. Consequently, **V** and **E** are composed of 2D vectors and scalar values, respectively. In addition, these features compose a graph structure, which is defined as

$$\mathbf{G} = (\mathbf{V}, \mathbf{E}) \tag{6}$$

As a result, **G** contains geometric information of facial emotions. However, in our research, using a set of triangular meshes was not sufficient to extract emotional information. In a graph structure, geometric information is propagated through edges. However, a graph structure is built using landmarks. Therefore, the structure is larger and more complicated than other graphs used in computer vision tasks, such as for the skeleton. Due to this difficulty, in this structure, it is hard to propagate information from vertex to far distant vertex.

In order to solve this problem, we adopted a master node [11] to propagate information on a long path. By definition, the master node is connected to all the vertices of a graph structure. Therefore, information is effectively propagated through a shortcut connection. In our method, we set the center point of the nose as the master node.

In this way, our graph structure describes facial emotion information adequately. Our experimental results proved that this method achieves a high performance in emotion recognition task.

3.3. Directed Graph Neural Network

This section proposes a directed GNN (DGNN), which is shown in Figure 2b,c. A DGNN is composed of several directed graph convolution blocks that aggregate the information of vertices and edges. Each DGNN block consisted of temporal convolutional blocks to obtain temporal information from a video. In addition, in order to prevent the vanishing gradient problem, we applied GLU to the temporal convolutional blocks.

First, the directed graph convolutional block was adopted to extract geometric information. Two updating functions and two aggregating functions were used as vertices and edges. An updating function enhances vertex and edge information based on their relationship in a graph structure. An aggregating function sums up the attributes of vertices and edges for each iteration of the training phase. The updating operation of a vertex is followed by the edge update process.

Next, we used a temporal convolutional block. The proposed temporal convolutional block was composed of a 1D convolutional layer and a batch normalization layer for each vertex and edge. A temporal block not only extracts temporal information from a graph but also compresses it in a temporal dimension. However, with a naive convolutional layer, it is hard to obtain complex temporal features of a human face. In order to enhance a temporal block, we applied GLU. This gating mechanism allows the selection of the words or features that are important for predicting facial emotions. The effect of GLU is demonstrated in Section 4.

Thus, we propose a novel DGNN using a master node for the generalization of a graph structure. This method obtain more geometry information from each frame. Moreover, GLU contributes to solve the vanishing gradient problem.

4. Experimental Results

4.1. Pre-Processing Procedure

This section explains the pre-processing procedure for each dataset. We sampled 16 frames of a video sequence at the same intervals. If there were not enough frames, we applied the bilinear interpolation method to extract landmark locations. Next, we subtracted the landmark locations by the nose location and normalized them. We followed the data augmentation method of Jung et al. [5] to expand the dataset size. First, we added the Gaussian noise to the landmarks locations, as follows:

$$x\prime_i^{(t)} = x_i^{(t)} + N(0, \sigma^2) \tag{7}$$

where σ is the standard deviation of Gaussian noise. Next, we applied random rotation to each frame:

$$[\overline{x}_i^{(t)}, \overline{y}_i^{(t)}] = \mathbf{R}^{(t)}[x\prime_i^{(t)}, y\prime_i^{(t)}] \quad (8)$$

where $\mathbf{R}^{(t)}$ is a 2×2 rotation matrix in t^{th} frame. The range of the rotation angle was $[-\pi/10, \pi/10]$. Also, random flipping was adapted to the video sequences. We generated an original dataset, a randomly noised dataset (three times), a randomly rotated dataset (three times), and their flipped versions. As a result, we expanded the original datasets by 14 times.

4.2. Training Configuration

For all experiments, we used the same network architecture, which is described in Figure 4. In our experiment, the weights of all networks were initialized with He's initialization [21], and L_2 regularization was applied. The weight decay parameter was set to 5×10^{-4}. A stochastic gradient descent (SGD) [22] was used as the optimizer, and a Nesterov accelerated gradient [23] was applied. Momentum was set to 0.9. For the extended cohn-kanade (CK+) and MMI facial expression (MMI) datasets, the learning rate and the batch were set to 0.1 and 128, respectively. For the Acted Facial Expressions In The Wild (AFEW) dataset, the learning rate and the batch size were set to 0.01 and 256, respectively. The training phases proceeded for 1000, 700, and 700 epochs for CK+, MMI, and AFEW, respectively.

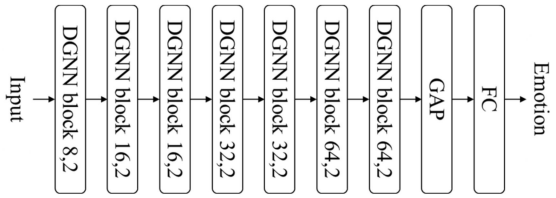

Figure 4. Network architecture used for the experiments. In the "DGNN block D, s", D and s denote depth and stride, respectively. GAP.

We used the well-known tenfold cross-validation method for the CK+ and MMI datasets. Nine subsets were used for training, and the remaining one was utilized for validation. For the AFEW dataset, the dataset comprises training and validation data, and we did not apply the tenfold cross-validation.

4.3. Experimental Results for the CK+ Dataset

In this section, we show experimental results for the CK+ dataset. The CK+ dataset is mainly used for emotion recognition research. It is composed of 327 video sequences presenting seven emotional labels, i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise. Also, all videos contain frontal face images without disturbing elements such as strong illumination and different faces. Due to these characteristics, our landmark extractor could extract precise landmarks' locations and provide excellent information to the DGNN. The experimental results are shown in Table 1.

The experimental result showed that our DGNN provided a performance of 96.02, outperforming not only a comparative landmark-based method but also most of the video- and image-based methods. This result shows that the landmark features can be considered an effective modality to analyze facial emotional information. However, the confusion matrix in Figure 5 shows that DGNN often failed to discriminate contempt from sadness. This is because landmarks such as eyebrow and lip shape related to contempt and sad emotions are similar. In order to describe this clearly, Figure 6 shows some prediction results. As mentioned above, the proposed method provides very high confidence scores for the evaluation of 'surprise' and 'happiness', which are easy to be recognized by landmark features. On the other hand, the proposed method confuses the emotion 'sadness' with 'contempt.' Such emotions are hard to discriminate by landmark features. Also, in order to verify that our algorithm

can be fused with other modality-based algorithms, we show the fused result of DGNN and C3D-GRU. DGNN provided a 1.22% higher performance than C3D-GRU, which provided a state-of-the-art (SOTA) performance. In order to analyze this result more clearly, Figure 7 illustrates the confusion matrices of fused DGNN–C3D-GRU algorithms. The confusion matrix showed that the fused algorithm of DGNN and C3D-GRU worked well for most emotional classes but tended to be more biased to contempt than that of DGNN. Since the fused algorithm confused sadness and contempt, its performance for sadness was lower than that of DGNN. However, the overall performance of the fused algorithm was much higher than those of the other methods. Thus, we can say that landmark feature use may be effective when applying multi-modal algorithms.

Table 1. Experimental results for the CK+ dataset. The proposed methods are highlighted in bold.

Method	Data Type	Accuracy
3DCNN-DAP [24]	Video	92.4
3DCNN [24]	Video	85.9
STRNN [3]	Video	95.44
STM-ExpLet [25]	Video	94.2
DTAN [5]	Video	91.4
F-Bases [26]	Video	92.6
C3D-GRU [27]	Video	97.25
DNN [28]	Static image	90.91
IPA2LT(LTNet) [29]	Static image	91.67
IACNN [30]	Static image	95.37
DTGN [5]	Landmark	92.35
DGNN	**Landmark**	**96.02**
DGNN + C3D-GRU	**Landmark + video**	**98.47**

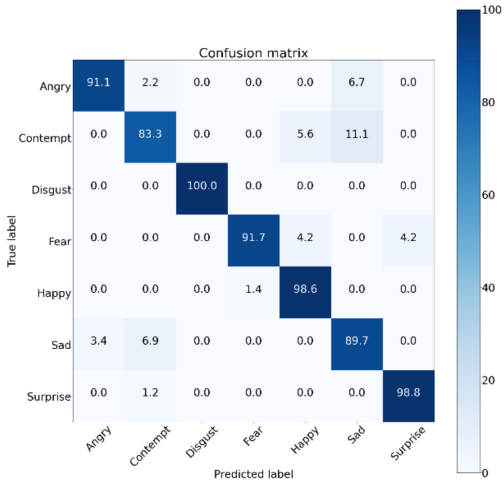

Figure 5. Confusion matrix of our DGNN method for the CK+ dataset.

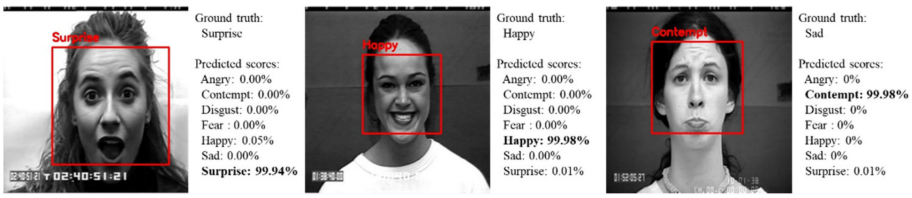

Figure 6. Prediction results of the proposed method. Red boxes indicate faces detected by the landmark extractor.

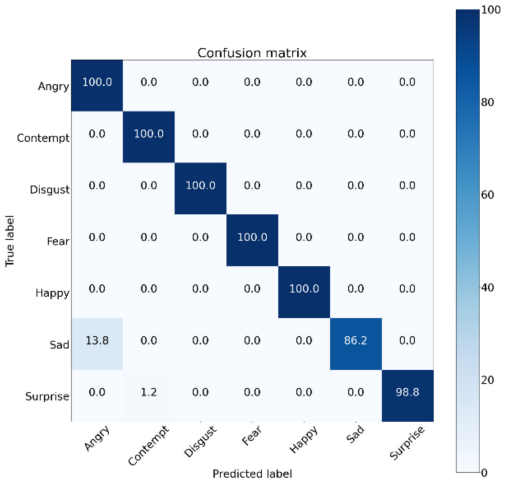

Figure 7. Confusion matrix of fused DGNN + C3D-GRU algorithms for the CK+ dataset.

Next, we evaluated the computational cost of the proposed algorithm. Table 2 shows that the landmark extractor required most of the computation time. Thus, in order to achieve real-time processing, a faster landmark extractor is required.

Table 2. Processing times of the proposed algorithm. L.E., landmark extractor.

L.E.	DGNN	Total
17.35 ms	4.85 ms	22.20 ms

4.4. Experimental Results for the MMI Dataset

In order to verify the efficacy of the landmark feature, we provide additional experimental results for the MMI dataset. The MMI dataset is composed of 205 image series showing frontal faces. The database displays six kinds of emotion, i.e., anger, disgust, fear, happiness, sadness, and surprise. The MMI dataset contains video sequences showing a frontal face with relatively active motion. The experimental results re shown in Table 3.

Table 3. Experimental results for the MMI dataset. The proposed method is highlighted in bold.

Method	Data Type	Accuracy
3DCNN-DAP [24]	video	63.4
3DCNN [24]	video	53.2
DTAN [5]	video	62.45
AURF [31]	static image	69.88
IPA2LT(LTNet) [29]	static image	65.61
IACNN [30]	static image	69.48
DTGN [5]	Landmark	59.02
DGNN	**Landmark**	**69.64**

Our method provided a performance of 69.64%, outperforming the conventional landmark-based method. Also, the proposed method is comparable in performance with the SOTA image-based and video-based methods. As a result, we can conclude that landmark feature is a very effective modality to recognize facial emotions in various environments.

4.5. Experimental Result for the AFEW Dataset

This section provides the experimental results for the AFEW dataset, a challenging case for landmark-based algorithms. The AFEW dataset includes a huge amount of movie clips. There are

seven labels, i.e., anger, disgust, fear, happiness, sadness, surprise, and neutral expression. Unlike other datasets, AFEW contains videos showing actively moving faces in various environments. In addition, some of them include images displaying a head from the side or the back, from which it is impossible to extract landmarks. This means that only limited frames provide meaningful landmark features, and it is hard to extract its information. However, our proposed method could extract useful information even from the noisy landmarks, providing better performance with respect to other modality-based algorithms. Therefore, we fused the proposed algorithm with that of C3D-GRU [27], which is one of the state-of-the-art FER algorithms used with videos, and verified the performance of the fused algorithm. The experimental results are shown in Table 4.

Table 4. Experimental results for the AFEW dataset. The proposed methods are highlighted in bold.

Method	Data Type	Accuracy
SSE-HoloNet [32]	Video	46.48
VGG-LSTM [33]	Video	48.60
C3D-LSTM [33]	Video	43.20
C3D-GRU [27]	Video	49.87
DGNN	**Landmark**	**32.64**
DGNN-C3D-GRU	**Landmark-video**	**50.65**

DGNN provided a 0.78% higher performance than C3D-GRU, which is the SOTA method for the AFEW dataset. In addition, comparing the confusion matrices of Figures 8 and 9, we found that DGNN improved the performance for some classes of emotion. This result indicates that landmarks could still provide proper information on the emotions in the videos.

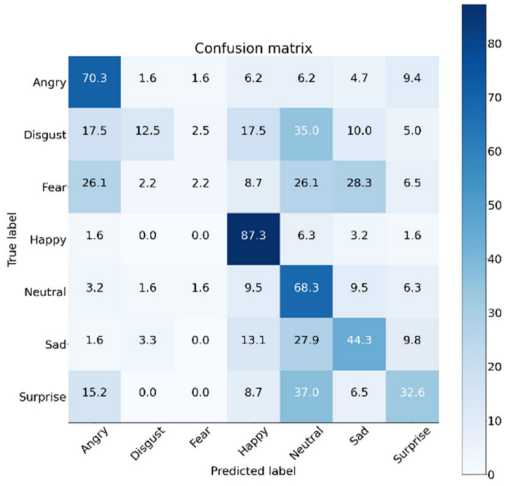

Figure 8. Confusion matrix of C3D-GRU for the AFEW dataset.

4.6. Ablation Study

We present the results of an ablation study to verify the ability of each module, i.e., master node and GLU. We used the CK+ dataset for this experiment. The experimental results are shown in Table 5. We set the same hyper-parameters as in Section 4.3, except for DGNN's modules.

Table 5. Experimental results of the ablation study. MS, master node.

Method	Result
DGNN (baseline)	95.02
DGNN + MS	95.72
DGNN + GLU	95.70
DGNN + MS + GLU	96.02

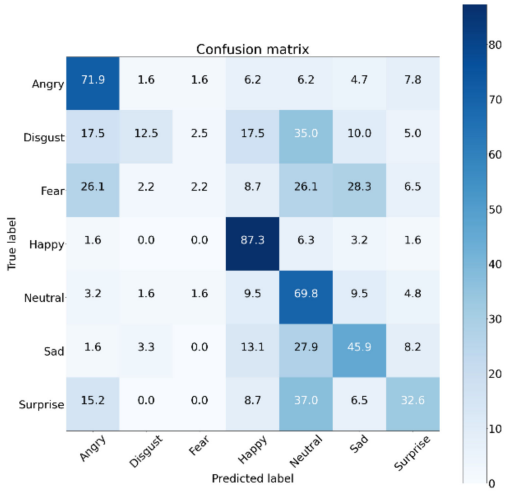

Figure 9. Confusion matrix of DGNNC3D-GRU for the AFEW dataset.

The baseline DGNN, which adopts graph convolutional layers, provided a performance of 95.02% comparable to that of conventional methods. However, the master node and GLU enhanced DGNN's geometry and temporal information, respectively. Therefore, the overall performance was reasonably increased. Also, when they were applied at the same time, DGNN provided an accuracy of 96.02%, which was 1% higher than the baseline accuracy of DGNN. Therefore, every module of the proposed method contributed to improving performance.

5. Limitations

The proposed method is based on landmark features and thus presents two limitations. First, the performance of the proposed method depends on the accuracy of the landmarks, and it can be impossible to extract facial landmarks from certain images. Second, noise can affect the performance of the proposed scheme. Marginal noise could be overcome by the robustness of the proposed algorithm. However, as noise increased, the overall performance decreased. Therefore, we adopted a state-of-the-art landmark extractor. Nonetheless, even the landmark extractor struggled to extract accurate landmark information in a varied environment such as the AFEW dataset.

Also, the overall latency of the algorithm depends on that of the landmark extractor. The input data of the GNN, i.e., the landmark features, have lower dimensionality than images or videos. So, the cost of GNN is much lower than that of CNN. However, in order to get accurate landmark features, we should adopt a computationally heavy landmark extractor than DGNN. Actually, in the case of a video sequence, latency is 22.20 ms, which is not a negligible cost for a multi-modal algorithm. Most of the computing time was required by the landmark extractor.

6. Conclusions

We believe that landmarks play an important role in an FER system. However, despite their importance, they have been barely studied. This paper presents a novel directed graph neural network for FER based on landmark features, showing that it achieved comparable performance to those of SOTA image- or video-based algorithms. Also, we demonstrate that, even in situations where the landmark features were difficult to solve, the proposed method was effective when fused with a conventional video-based method. As a result, the proposed algorithm provided SOTA performance for the CK+ and AFEW datasets, which corresponded to 98.47% and 50.65%, respectively. Our future plan is to construct a multi-modal FER system, possibly using various modalities such as facial features, audio signals, and body movements. In a multi-modal FER system, the landmarks will play an important role.

Author Contributions: The work described in this article is the collaborative effort of all authors. All authors contributed to data processing and designed the algorithm. Q.T.N., S.L. and B.C.S. made contributions to data measurement and analysis. All authors participated in the writing of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Industrial Technology Innovation Program funded by the Ministry of Trade, Industry and Energy (MI, Korea) [10073154, Development of human-friendly human–robot interaction technologies using human internal emotional states].

Acknowledgments: The authors would like to thank the support of Inha University.

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

References

1. Kuo, C.M.; Lai, S.H.; Sarkis, M. A compact deep learning model for robust facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2121–2129.
2. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]
3. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* **2018**, *49*, 839–847. [CrossRef] [PubMed]
4. Yan, J.; Zheng, W.; Cui, Z.; Tang, C.; Zhang, T.; Zong, Y. Multi-cue fusion for emotion recognition in the wild. *Neurocomputing* **2018**, *309*, 27–35. [CrossRef]
5. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression –recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
6. Kollias, D.; Zafeiriou, S. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *arXiv* **2019**, arXiv:1910.01417.
7. Hasani, B.; Mahoor, M.H. Facial expression recognition using enhanced deep 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 30–40.
8. Fabiano, D.; Canavan, S. Deformable synthesis model for emotion recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019.
9. Zhang, X.; Xu, C.; Tian, X.; Tao, D. Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**. [CrossRef] [PubMed]
10. Gao, X.; Hu, W.; Tang, J.; Liu, J.; Guo, Z. Optimized skeleton-based action recognition via sparsified graph regression. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 601–610.
11. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, Australia, 6–11 August 2017; pp. 1263–1272.
12. Rohr, K.; Stiehl, H.S.; Sprengel, R.; Buzug, T.M.; Weese, J.; Kuhn, M.H. Landmark-based elastic registration using approximating thin-plate splines. *IEEE Trans. Med. Imaging* **2001**, *20*, 526–534. [CrossRef] [PubMed]
13. Ghimire, D.; Lee, J. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors* **2013**, *13*, 7714–7734. [CrossRef] [PubMed]
14. Ma, Y.; Wang, S.; Aggarwal, C.C.; Yin, D.; Tang, J. Multi-dimensional graph convolutional networks. In Proceedings of the 2019 SIAM International Conference on Data Mining, Calgary, AB, Canada, 2–4 May 2019; pp. 657–665.
15. You, J.; Liu, B.; Ying, Z.; Pande, V.; Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 3–8 December 2018; pp. 6410–6421.
16. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv* **2017**, arXiv:1707.01926.

17. Li, B.; Li, X.; Zhang, Z.; Wu, F. Spatio-temporal graph routing for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Napa, CA, USA, 16–17 July 2019; Volume 33, pp. 8561–8568.
18. Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Style aggregated network for facial landmark detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 379–388.
19. Delaunay, B. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* **1934**, *7*, 1–2.
20. Golzadeh, H.; Faria, D.R.; Manso, L.J.; Ekárt, A.; Buckingham, C.D. Emotion Recognition using Spatiotemporal Features from Facial Expression Landmarks. In Proceedings of the 2018 International Conference on Intelligent Systems (IS), Funchal-Madeira, Portugal, 25–27 September 2018; pp. 789–794.
21. 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, 26 June–1 July 2016; pp. 770–778.
22. Kiefer, J.; Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **1952**, *23*, 462–466. [CrossRef]
23. Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence O (1/kˆ 2). *Doklady AN USSR* **1983**, *269*, 543–547.
24. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 143–157.
25. Liu, M.; Shan, S.; Wang, R.; Chen, X. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1749–1756.
26. Sariyanidi, E.; Gunes, H.; Cavallaro, A. Learning bases of activity for facial expression recognition. *IEEE Trans. Image Process.* **2017**, *26*, 1965–1978. [CrossRef] [PubMed]
27. Lee, M.K.; Choi, D.Y.; Kim, D.H.; Song, B.C. Visual Scene-aware Hybrid Neural Network Architecture for Video-based Facial Expression Recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–8.
28. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter conference on applications of computer vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
29. Zeng, J.; Shan, S.; Chen, X. Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 222–237.
30. Meng, Z.; Liu, P.; Cai, J.; Han, S.; Tong, Y. Identity-aware convolutional neural network for facial expression recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 558–565.
31. Liu, M.; Li, S.; Shan, S.; Chen, X. Au-aware deep networks for facial expression recognition. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6.
32. Hu, P.; Cai, D.; Wang, S.; Yao, A.; Chen, Y. Learning supervised scoring ensemble for emotion recognition in the wild. In Proceedings of the 19th ACM international conference on multimodal interaction, Glasgow, UK, 13–17 November 2017; pp. 553–560.
33. Vielzeuf, V.; Pateux, S.; Jurie, F. Temporal multimodal fusion for video emotion classification in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow UK, 13–17 November 2017; pp. 569–576.