symmetry

MDPI

Article

Deep Temporal–Spatial Aggregation for Video-Based Facial Expression Recognition

Xianzhang Pan [1,*], Wenping Guo [1], Xiaoying Guo [2], Wenshu Li [3], Junjie Xu [4] and Jinzhao Wu [5]

1 Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China; guowp@tzc.edu.cn
2 School of Software Engineering, Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China; guoxiaoying@sxu.edu.cn
3 College of information science and technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; Charlie@zstu.edu.cn
4 College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China; connyadmin@163.com
5 Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University for Nationalities, Nanning 530006, China; ziyaarnavut2011@hotmail.com
* Correspondence: pxz@tzc.edu.cn

Received: 28 October 2018; Accepted: 30 December 2018; Published: 5 January 2019

Abstract: The proposed method has 30 streams, i.e., 15 spatial streams and 15 temporal streams. Each spatial stream corresponds to each temporal stream. Therefore, this work correlates with the symmetry concept. It is a difficult task to classify video-based facial expression owing to the gap between the visual descriptors and the emotions. In order to bridge the gap, a new video descriptor for facial expression recognition is presented to aggregate spatial and temporal convolutional features across the entire extent of a video. The designed framework integrates a state-of-the-art 30 stream and has a trainable spatial–temporal feature aggregation layer. This framework is end-to-end trainable for video-based facial expression recognition. Thus, this framework can effectively avoid overfitting to the limited emotional video datasets, and the trainable strategy can learn to better represent an entire video. The different schemas for pooling spatial–temporal features are investigated, and the spatial and temporal streams are best aggregated by utilizing the proposed method. The extensive experiments on two public databases, BAUM-1s and eNTERFACE05, show that this framework has promising performance and outperforms the state-of-the-art strategies.

Keywords: facial expression recognition; convolutional neural networks; temporal-spatial features; optical flow; feature aggregation

1. Introduction

Facial expressions are non-verbal information and can complement our verbal information. Video-based facial expression recognition (VFER) aims to automatically classify human expression categories in video. A large number of researchers have become interested in VFER in the past decades. VFER is a challenging task because there is a large gap between visual features and emotions [1]. It has potential applications in healthcare, robotics, and driver safety [2–6]. The work of reference [7] defined the six facial expressions of anger, disgust, fear, happiness, sadness, and surprise in 1993.

There are usually two types of VFER, classified according to their feature representations. They are static image-based facial expression classification and video-based facial expression classification. The static image-based facial expression recognition aims to extract spatial information from a single image. It is the focus of previous works. Video-based facial expression recognition considers the temporal information between two adjacent video frames. It has been focused upon in recent years.

This requires the processing of a video sequence with spatial information and temporal information rather than a static image. VFER often has three stages: video preprocessing, visual feature extraction, and expression recognition. Video preprocessing aims to detect and crop facial images from video. Then, the visual features are extracted from video-frame facial images. Finally, a classifier is adopted to recognize expression.

Convolutional neural networks (CNNs) have improved image classification accuracy and outperform the hand-crafted feature method owing to the large-scale datasets [8]. However, current emotional datasets are small and have noise [9,10]. Furthermore, spatial signals and temporal signals are very important for modeling emotional video. Popular approaches for VFER employ a spatial signal and temporal signal to extract features from video, such as two different streams with a spatial stream and temporal stream and 3D convolutions [11–13]. They show rapid progress, but the 3D convolutions are difficult to scale to date for VFER, and the spatial–temporal network [14] only processes short-time video [12]. Thus, this is a key step to improving recognition performance by exploiting the comprehensive and effective features for VFER.

To obtain the above-mentioned goal, an end-to-end trainable video-level multimodal framework with convolutional neural networks (CNNs) is presented to extract and aggregate descriptors across whole portions of a static image and the temporal span of a video. The core of the descriptors is a spatial–temporal aggregation layer, called EmotionalVlan, inspired by the NetVLAD aggregation layer, which works well for object-level classification tasks in static images and video action recognition [15–17]. The aggregation is investigated at different layers of convolutional neural networks, and the penultimate layer is found to perform best in a facial expression recognition task. The different schemas are also investigated for fusing spatial and temporal information and the aggregating spatial and temporal information into single video-level descriptors.

2. Related Work

Among the above-mentioned three stages of VFER, facial feature extraction is the key role for VFER. A large number of the previous researchers have employed a variety of hand-crafted visual features for VFER [18–20]. These features are extracted from the key emotional video frames. For example, the method of [21] employs a set of words according to the multiscale dense Scale-invariant feature transform (SIFT) features to recognize facial expression. The sparse local Fisher discriminant analysis is utilized to extract visual features for facial expression recognition [22]. The method of Gabor filters is employed to extract visual features for emotion recognition [23]. Local phase quantization (LPQ) [24] is employed to extract visual features for VFER [25]. These hand-crafted facial features are low level, so they cannot effectively classify facial expression in videos.

To address the above problem, neural networks can provide us with the clues, especially the deep learning framework, which can improve the recognition accuracy of VFER because it achieves state-of-the-art level for many applications [26–30]. It can capture high-level abstractions by utilizing multiple nonlinear transformations. Typical learning methods, such as convolutional neural networks (CNNs), are effectively employed to classify facial expression [31,32]. For example, CNNs with three convolution and two subsampling layers are employed to extract facial features for VFER [33]. FaceNet2ExpNet is designed to classify facial expression, which adopts facial domain knowledge to improve recognition accuracy [34]. Recurrent neural networks (RNNs) and CNNs are integrated to classify facial expression [35]. The method of [36] integrates CNNs and long short-term memory (LSTM) to classify video-based facial expressions. A new neural network, called DeepSentiBank, is designed to generate visual features. Then, these features are fed into an architecture-frame-transformer for VFER [37]. Geometric–convolutional features are extracted for facial expression recognition; specifically, critical region geometric features are employed by the differential fusion network [38].

However, there is another characteristic of VFER: it consists of dynamic spatial and temporal variation parts, i.e., the mouth and eyes. The above-mentioned methods have difficulty generating the

powerful visual features hidden in video frames. For instance, a large number of previous researchers have only employed the spatial features for VFER and ignored the temporal features in video, which are complementary to recognize facial expression in video. The consecutive facial images are directly fed into the deep learning framework, but it is not possible to effectively utilize the temporal–spatial signals in this way, because the same emotion is shown through different variants of the face, and this is challenging due to the small training datasets. The recent deep spatial–temporal networks are designed to recognize facial expression in video and improve recognition performance [39], extracting deep spatial features and temporal features from the video. However, they largely ignore the long time structure of emotional video and only aim to operate on a few video frames (up to 10), and an emotional video usually has at least 25 video frames. They also provide us with clues. To address the above-mentioned shortcomings, a novel framework is designed using deep temporal–spatial networks to make full use of dynamic movement signals and static signals, adopting a multimodal deep CNN architecture and consisting of spatial CNNs and temporal CNNs, according to AlexNet, to extract high-level spatial features and temporal features from video, respectively. Then, an aggregation layer, called EmotionalVlan, is designed to aggregate convolutional descriptors across the spatial signals and temporal signals of the entire video. A global feature of the video is extracted. Finally, the global feature is fed into a classifier to recognize facial expression. Extensive experiments on two public databases, i.e., the BAUM-1 [9], and eNTERFACE05 [10], show the effectiveness of our method.

Our contributions are as follows. Firstly, an effective entire video representation through a trainable spatial–temporal aggregation layer with state-of-the-art CNNs is proposed, which can avoid the limited emotional video datasets, and the trainable strategy can learn to better represent the entire video. Secondly, the trained schema, non-trained schema, and the schema of inserting the EmotionalVlan layer in different places are investigated. As a result, trained schema and inserting the EmotionalVlan layer after fc7 are utilized to achieve better recognition performance than other schema for VFER. Thirdly, the proposed framework performs better than state-of-the-art technology for public datasets.

3. The Proposed Method

There are spatial and temporal signals in emotional video. The spatial signal, called spatial information, consists of video-frame facial appearance, and the temporal signal, called optical flow information or temporal information, consists of the movement of the facial video frame. Temporal information is also considered to be facial critical area movement; for instance, shrinking the eyebrows and eyes indicates disgust. Our framework is based on spatial information and temporal information. A trainable end-to-end model for facial expression recognition is designed, which can learn comprehensive and effective features for videos.

To achieve this goal, the structure of our method is shown in Figure 1. This proposed method contains 30 individual CNN streams, i.e., the temporal CNN network and the spatial CNN network. The temporal CNN network aims to generate high-level temporal features from the optical flow signals. The spatial CNN network aims to generate spatial features from facial images of video. Then, an aggregation layer, called EmotionalVlan, is designed to aggregate temporal features and spatial features. The output of EmotionalVlan represents the entire video features, and it is fed into Softmax layer for VFER. The best parameters of this framework are obtained through a training stage.

The temporal CNN network is designed to make full use of the temporal signal. Its input is the optical flow displacement fields between two video frames. Facial expression recognition will be made easier in this way; it is not necessary to estimate facial motion implicitly using our framework. The temporal CNN network is designed according to this idea.

The spatial CNN network is designed to process static video frames. The static appearance of the face is a useful clue, because the facial appearance and facial expressions have a strong correlation. Our spatial CNN network is designed according to this idea. A recent successful CNN model is utilized here, such as AlexNet.

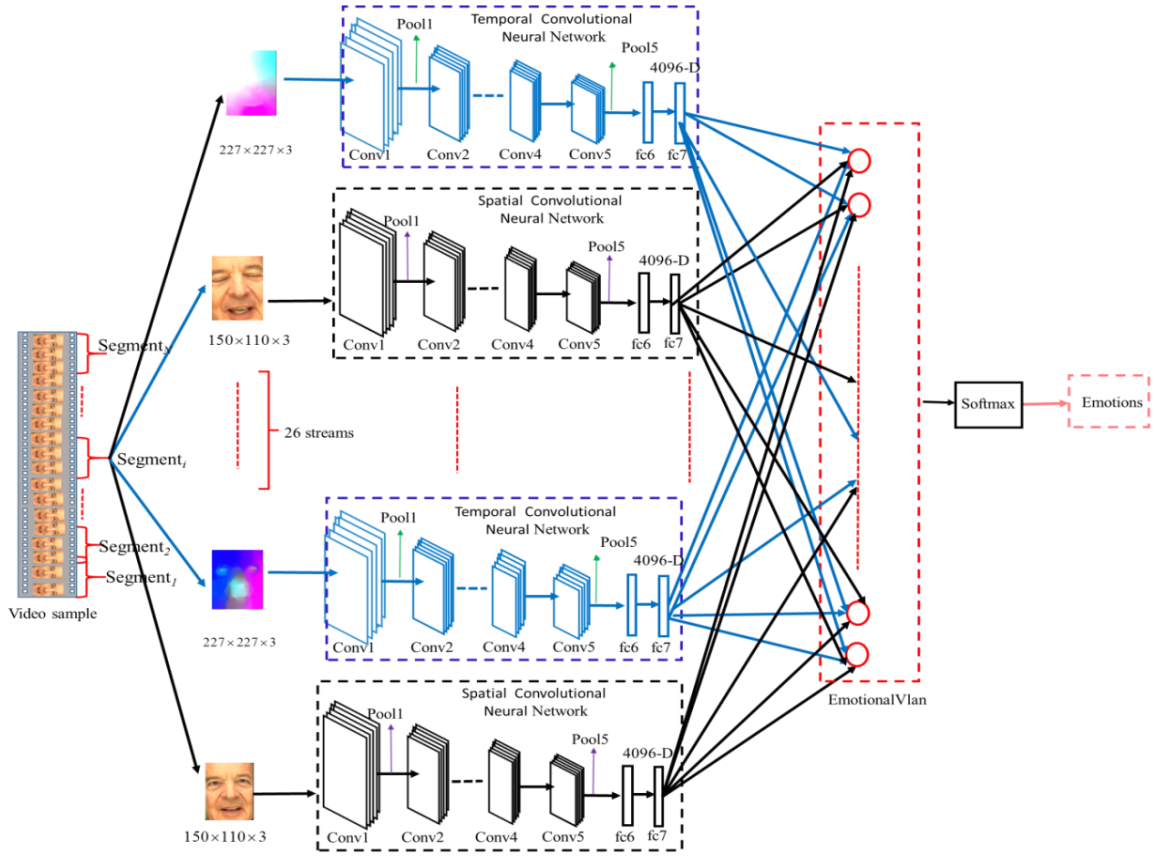

Figure 1. The structure of our method.

In order to overcome limited emotional video data, the temporal CNN network and spatial CNN network have the same network structure based on AlexNet architecture. There are five convolution layers, three max-pooling layers, and four fully connected (fc) layers (fc6, fc7, fc8, fc9) in our CNNs. There are 4096 units in the former two fc (fc6, fc7) layers, respectively. The max-pooling layers are employed to avoid overfitting. The spatial and temporal expression feature vectors are extracted by fc7 and fed into the EmotionalVlan layer to generate the entire video feature, respectively. Finally, the entire video feature is fed into Softmax to classify facial expression. Our CNN input is $227 \times 227 \times 3$ RGB images. Firstly, the extant parameters of AlexNet are copied to initialize each stream of our framework; then, our framework is fine-tuned to the target video segments [40]. Figure 2 shows different iterations for validation recognition accuracy of BAUM-ls in stage of spatial CNNs fine-tuning. Figure 3 shows different iterations for validation recognition accuracy of BAUM-ls in stage of temporal CNNs fine-tuning. Figure 4 shows different iterations for validation recognition accuracy of eNTERFACE05 in stage of spatial CNNs fine-tuning. Figure 5 shows different iterations for validation recognition accuracy of eNTERFACE05 in stage of temporal CNNs fine-tuning. Figure 6 shows different iterations for recognition accuracy of BAUM-ls in stage of EmotionalVlan layer training. Figure 7 shows different iterations for recognition accuracy of eNTERFACE05 in stage of EmotionalVlan layer training. From these Figures, we can see that recognition accuracy cannot improve after 5000 steps. The algorithm of this work is described in Algorithm 1, and the flow of the proposed method is shown in Figure 8. The detailed steps of the proposed method are described as follows.

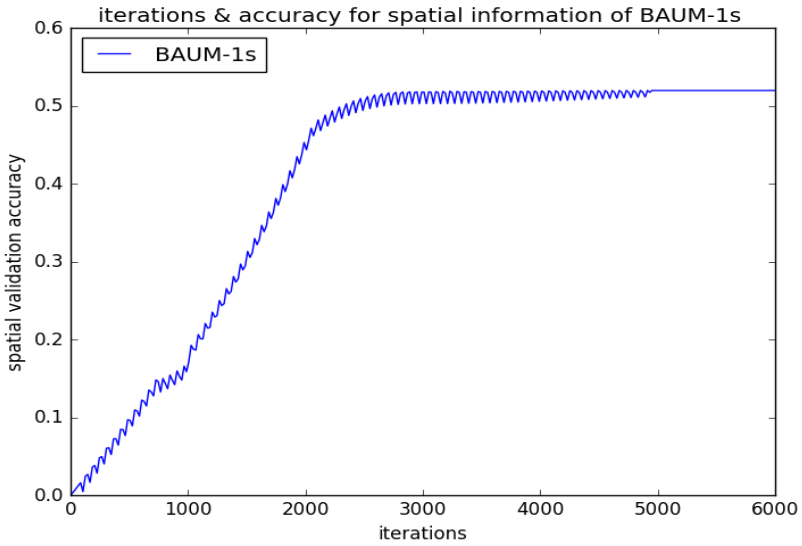

Figure 2. Different iterations for validation recognition accuracy of BAUM-ls in stage of spatial CNNs fine-tuning.

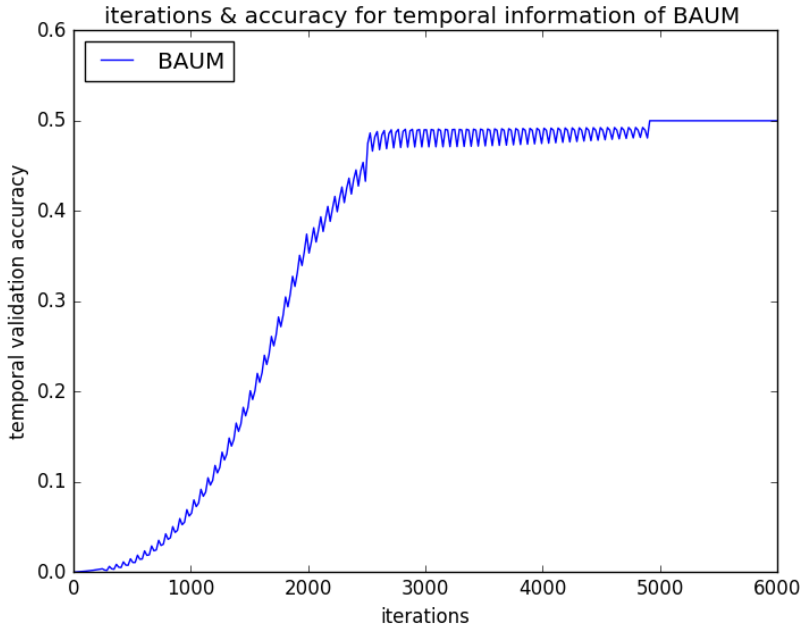

Figure 3. Different iterations for validation recognition accuracy of BAUM-ls in stage of temporal CNNs fine-tuning.

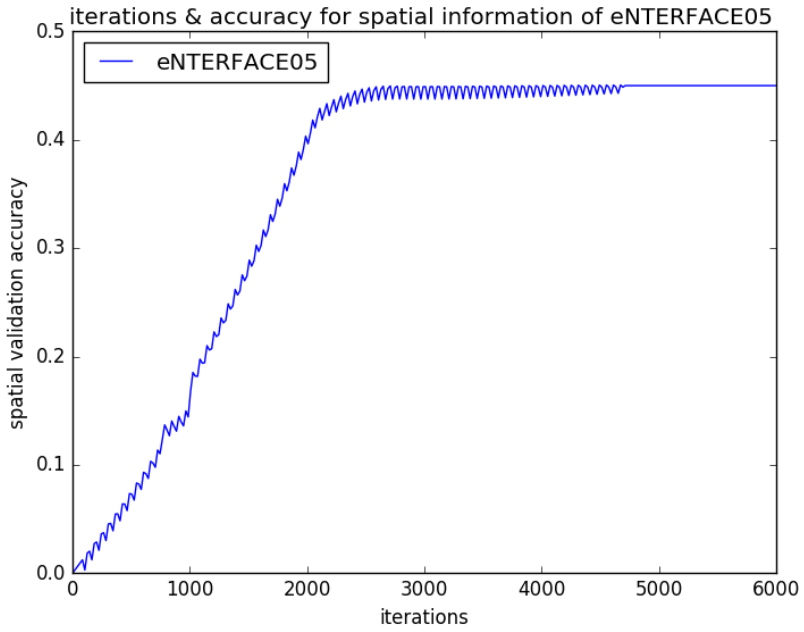

Figure 4. Different iterations for validation recognition accuracy of eNTERFACE05 in stage of spatial CNNs fine-tuning.

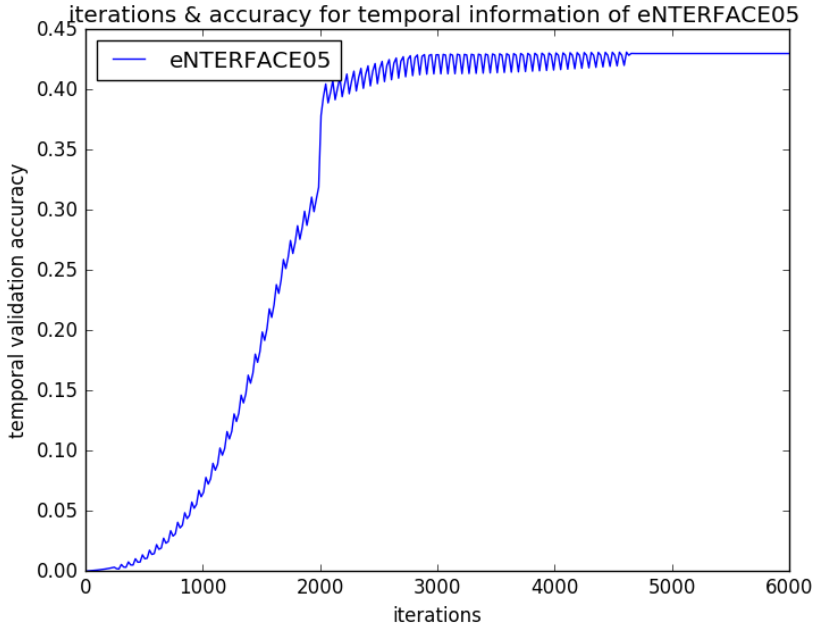

Figure 5. Different iterations for validation recognition accuracy of eNTERFACE05 in stage of temporal CNNs fine-tuning.

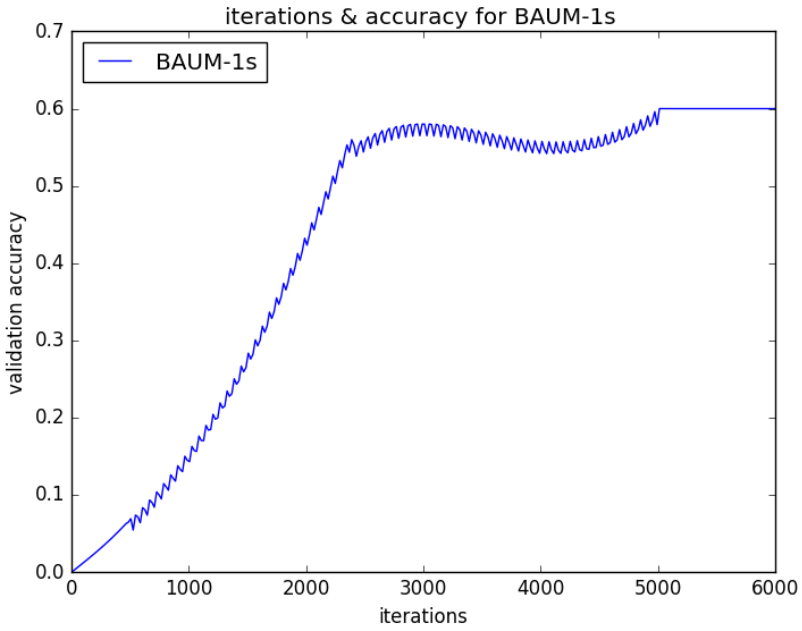

Figure 6. Different iterations for recognition accuracy of BAUM-ls in stage of EmotionalVlan layer training.

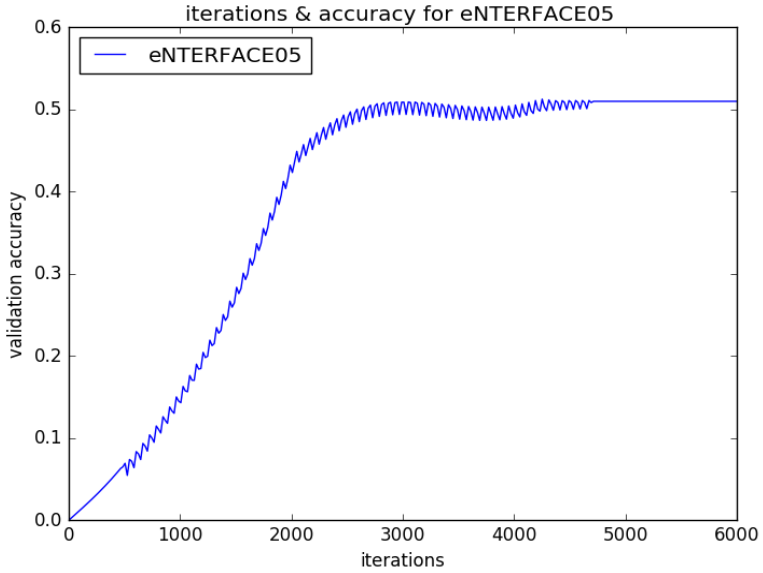

Figure 7. Different iterations for recognition accuracy of eNTERFACE05 in stage of EmotionalVlan layer training.

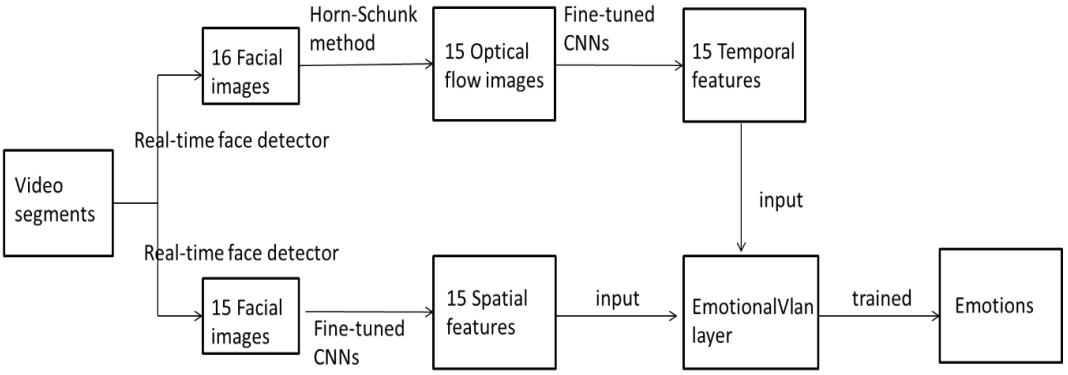

Figure 8. The flow of the proposed method.

Step 1: There are 16 facial images cropped from every video segment.
Step 2: There are 15 optical flow images generated from 16 facial images.
Step 3: 15 static facial images before and 15 optical flow images are fed into the 15 spatial CNNs and 15 temporal CNNs for fine-tuning, respectively.
Step 4: 15 spatial features are extracted by 15 fine-tuned spatial CNNs, and 15 temporal features are extracted by 15 fine-tuned temporal CNNs.
Step 5: The 15 spatial features and 15 temporal features are fed into EmotionalVlan layer for training and recognition of facial expressions.

Algorithm 1. The algorithm of this work

Input: 15 static video-frames and 15 optical flow images
Output: facial expressions
1. Copy AlexNet parameters to initialize our framework parameters: mini-batch is 40, and learning rate is 0.001
2. While epoch number<=5000{
//the forward propagation
For each neuron k in the hidden layer, that is, before the EmotionalVlan layer{
F(k) = max(0,x) //the output of neuron k, x represents the input of neuron k,
}
For each neuron k in EmotionalVlan{
$V[k] = \sum_{i=1}^{T} e^{a_k}(x_{s,i} + x_{t,i})$ //the output of neuron k, $x_{s,i}$ and $x_{t,i}$ represents the output of fc7 layer in spatial streams and temporal streams separately.
}
// back propagation error
For each neuron k in the output layer{
$E_k = O_k(1 - O_k)(T_k - O_k)$ // compute error, T_k represents the true emotional label, O_k represents the prediction emotional label
}
For every hidden neuron {
Update the weight
}
}

Firstly, the feature aggregation layer is described in the following. Then, the two steps of our method are given as follows: (1) video preprocessing and (2) network training.

3.1. Temporal–Spatial Aggregation Layer

The $x_{s,i} \in R^D$ represents a D-dimensional spatial descriptor extracted from the i-th index of video-frame; the $x_{t,i} \in R^D$ represents a D-dimensional temporal descriptor extracted from the i-th index of a video frame and the $(i + 1)$-th index of a video frame. It would be preferable to aggregate $x_{t,i}$ and $x_{s,i}$ over the entire video to form the video descriptor and preserve the information content of the video. In order to achieve the above goal, each video descriptor is then assigned to one of the EmotionalVlan cells. EmotionalVlan has 4096 cells.

$$V[k] = \sum_{i=1}^{T} e^{a_k}(x_{s,i} + x_{t,i}) \tag{1}$$

where a_k is a tunable parameter of the k-th cell. It should be noted that the $(x_{s,i} + x_{t,i})$ represents the temporal–spatial information aggregation of the i-th video-frame. The summing operator indicates aggregation over the entire spatial signal and temporal signal of video. The output is a vector V, where $V[k]$ indicates the aggregated descriptor of the k-th cell and $V[k] \in R^{4096\times15}$ is a single descriptor of the video.

Intuitively, $(x_{s,i} + x_{t,i})$ records the appearance of the video frame and the movement of the i-th video frame. Then, they are aggregated by computing the sum of all the video frames inside every cell. Specifically, in order to better recognize facial expression, the end-to-end schema is employed to learn all parameters. From Formula (1), it can be seen that the spatial–temporal aggregation is differentiable, so it can conduct back-propagating error gradients in our framework.

In theory, we can place the spatial–temporal aggregation layer shown above at any level in our network to pool the spatial and temporal features. Therefore, this work considers pooling the output of fully connected layers (fc6 and fc7 are considered). The 4096-dimensional spatial features and temporal features from every video frame are pooled. As described in Section 4.2, the fc7 layer is the best place to obtain the best performance by pooling features.

3.2. Video Preprocessing

There are different durations in every video. Every video is split into 16 frames of segments as inputs of our framework, because we investigated the integer L in a range of [2,20] and found that L = 16 obtains the best performance for facial expression tasks. Figure 9 describes the comparisons of different L for recognition accuracy. The training data is enlarged by utilizing the above method. If a segment has more than 16 frames, the first and last $(L - 16)/2$ frames are dropped. If a segment contains fewer than 16 frames, the first and last $(16 - L)/2$ frames are repeated. A segment of L = 16 produces 15 temporal images, because two video frames produce one temporal image, which indicates the corresponding position change between two video frames. The temporal information is considered as a set of motion vectors d_t, which is computed from video frames t and $t + 1$, and a temporal image I_t contains $d_t x$ and $d_t y$. The $d_t x$ and $d_t y$ respectively indicate the horizontal and vertical corresponding position change between the two video frames t and $t + 1$ They are computed through the Horn–Schunck method [41]. Considering the input size of our CNNs, the $d_t z$ is computed, indicating the third channel of input temporal image, as shown below. It is noted that the 15 static video frames before and 15 temporal images are fed into the proposed framework.

$$d_t z = \sqrt{{d_t}^2 x + {d_t}^2 y} \tag{2}$$

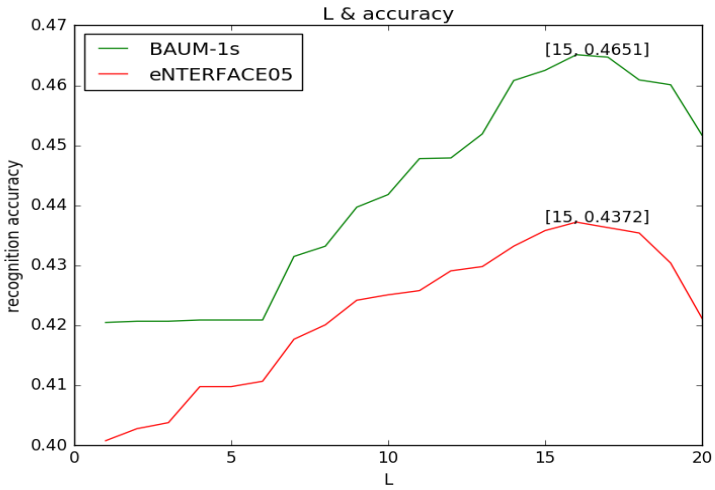

Figure 9. Comparisons of different L for recognition accuracy.

For the preprocessing of the spatial stream, a robust real-time face detector [42] is employed to crop a facial image from each video frame. In a facial image, we can observe that its width is approximately twice as long as the distance between two eyes and the height is approximately three times longer than the distance between two eyes. We crop the resized image of 150 × 110 × 3 from a facial image based on the normalized distance between two eyes, as in [43].

3.3. Network Training

We employ the cross-entropy loss to train our framework, which is the next layer of our EmotionalVlan described in Figure 1. The output probabilities are obtained using a Softmax. For every stream, we replace the fc8 layer of AlexNet with our EmotionalVlan. Then, a new fc layer based on the number of emotional categories is designed. Finally, in order to minimize Formula (3) in the training stage, we utilize the algorithm of standard back propagation to obtain the parameters of our framework as shown below.

In Figure 1, every stream of CNN has five convolution layers, three max-pooling layers, and four fully connected (fc) layers (fc6, fc7, fc8, fc9). The fc6 and fc7 layers contain 4096 units, the fc8 is our EmotionalVlan. The last fc9 is a label vector based on emotional categories. It should be noted that, in the process of training, we separately resize the optical image and static image with a size of $227 \times 227 \times 3$.

Given data $X = \{(b_i, y_i)\}_{i=1,2,\cdots N}$, i indicates the index of the video frame, b_i represents the output of the EmotionalVlan layer, and y_i represents the segment label according to the entire video label. For our entire network B, the back propagation method is utilized to solve the following problem:

$$\min_{W^B, \lambda^B} \sum_{i=1}^{N} H(softmax(W^B \cdot v^B(b_i; \lambda^B)), y_i), \tag{3}$$

where W^B represents the weights of the Softmax layer, and $v^B(b_i; \lambda^B)$ indicates the 122880-D ($R^{4096 \times 30}$) output of the EmotionalVlan layer in network B with parameters λ^B. The log-loss of Softmax is obtained by

$$H(B, y) = -\sum_{j=1}^{k} y_i \log({y_j}^B), \tag{4}$$

where ${y_j}^B$ is the j-th output of the Softmax layer, and k represents the total number of emotions.

4. Experiment Studies and Result Analysis

To evaluate the recognition accuracy of our framework, extensive experiments on two public datasets of emotional videos were conducted: BAUM-1 [9] and eNTERFACE05 [10]. The implementation details are as follows. For the training of CNNs, the mini-batch is 40, and the learning rate is 0.001. The epoch number is 5000. The CNNs are implemented based on MatConvNet. An NVIDIA GPU with 25 GB memory is utilized to train our framework. The subject-independent cross-validation experiments are employed to test the performance of our framework [44]. The two datasets contain more than 10 subjects; we conduct leave-one-subject-group-out (LOSGO) with five subject groups for the experiments. The average accuracy is utilized to evaluate our framework.

4.1. Datasets

The BAUM-1s database [9] has 1222 video samples from 31 people. They exhibited joy, anger, sadness, disgust, fear, surprise, boredom, and contempt. In order to obtain spontaneous facial expressions, the emotions are inspired by watching films. The video frame size in a video is $720 \times 576 \times 3$. Our work aims to classify the emotions of anger, disgust, fear, joy, sadness, and surprise, which are described from 521 samples of video. Some BAUM-1s database samples of cropped facial images are shown in Figure 10.

Figure 10. Some cropped facial images on the BAUM-1s database.

TThe eNTERFACE05 database contains 1290 video samples from 43 people [10]. They exhibited anger, disgust, fear, joy, sadness, and surprise. Every video sample has an average duration of approximately four seconds. The video frame size is 720 × 576 × 3. Some eNTERFACE05 database samples of cropped facial images are shown in Figure 11.

Figure 11. Some cropped facial images on the eNTERFACE05 database.

Since a video is split into many segments as inputs to the CNNs, the training dataset is increased. In this work, 7000 segments are produced from the BAUM-1s database, and 19350 segments are produced from the eNTERFACE05 database.

4.2. Experimental Results and Analysis

To test the effectiveness of our framework, it is compared with non-trained schema. Specifically, the parameters before EmotionalVlan are obtained. As described in Table 1, the non-trained schema is that the last 2-layers are removed and the EmotionalVlan layer is utilized to pool the descriptor of fc7, and then the output of the EmotionalVlan layer is fed into a logistic regression [45–47] to classify emotions. Our trained schema is that the EmotionalVlan layer is trained together with the preceding layers. It can be seen that the trained schema performs better than the non-trained schema. The trained schema separately improves the recognition accuracy from 46.02% to 46.51% on BAUM-1s and from 42.97% to 43.72% on eNTERFACE05 compared with non-trained schema. The reason is that trained schema can learn more comprehensive and effective features from video.

Table 1. Recognition performance of trained and non-trained schema.

Method	BAUM-1s	eNTERFACE05
trained schema	46.51%	43.72%
non-trained schema	46.02%	42.97%

We also test the different places in our framework where the EmotionalVlan layer can be inserted. In particular, we compare placing EmotionalVlan after the two fully-connected layers (fc6, fc7). In every case, the EmotionalVlan layers are trained together with the preceding layers. In the case of fc6, fc7 will be dropped, and in case of fc7, fc6 will be dropped. The results of Table 2 show that inserting the layer after fc7 obtains the best performance. Inserting after fc7 separately improves the recognition accuracy from 45.01% to 46.51% on BAUM-1s and from 42.11% to 43.72% on eNTERFACE05 compared with inserting the layer after fc6. The reason is that the features of fc7 are more semantic and varied, and the EmotionalVlan layer can capture complex distributions of feature space.

Table 2. Recognition performance of inserting the EmotionalVlan layer after different places.

Method	BAUM-1s	eNTERFACE05
fc7	46.51%	43.72%
fc6	45.01%	42.11%

We also evaluate the recognition accuracy of the learned temporal and spatial CNN features, respectively. Table 3 gives the recognition accuracy of different streams on the BAUM-1s database and eNTERFACE05 database. It is noted that spatial stream recognition performance indicates that only 15 spatial streams are utilized to classify emotions, and temporal stream recognition performance indicates that only 15 temporal streams are utilized to classify emotions, and spatial–temporal streams indicate that both 15 spatial streams and 15 temporal streams are utilized to classify emotions, and they all utilize trained schema. From Table 3, it can be seen that temporal–spatial streams are useful for VFER and improve recognition performance, because temporal–spatial streams utilize both the spatial information and temporal information in video. Spatial streams only utilize the spatial information of video rather than temporal information, and temporal streams only utilize the temporal information of video rather than spatial information.

Table 3. Recognition performance of different streams.

Stream	BAUM-1s	eNTERFACE05
Spatial–temporal streams	46.51%	43.72%
Spatial streams	42.68%	41.03%
Temporal streams	41.09%	39.75%

Finally, this framework is compared with recent popular deep network methods, such as HOG 3D [48] and 3DCNN [49]. From the results in Tables 3 and 4, it can be seen that 3DCNN and HOG 3D have better recognition accuracy than single spatial streams and single temporal streams, because they utilize both spatial information and temporal information rather than single information of videos. This proposed work can achieve better recognition accuracy than 3DCNN and HOG 3D, indicating that the designed framework can generate more comprehensive and effective features for VFER.

Table 4. Recognition performance of different deep network methods.

Method	BAUM-1s	eNTERFACE05
Ours	46.51%	43.72%
HOG 3D	43.02%	41.23%
3DCNN	42.89%	41.05%

To describe recognition performance per facial expression, Figures 12 and 13 describe the confusion matrices of the best recognition performance achieved by utilizing the proposed method on the BAUM-1s database and eNTERFACE05 database, respectively. From Figures 12 and 13, it can be seen that "fear" is identified badly with an accuracy of about 50%, which indicates that it is

more difficult to classify than other emotions. Because "fear" is similar to other emotions, such as sadness, our framework cannot effectively recognize "fear" owing to the limited and low-quality dataset. Our framework depends on large-scale and quality datasets.

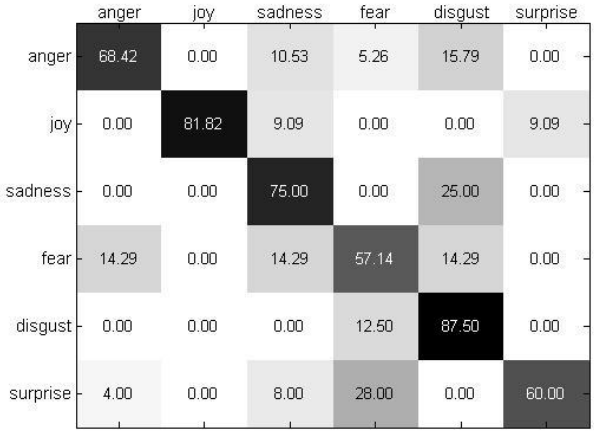

Figure 12. The confusion matrix of the best recognition performance on the BAUM-1s database.

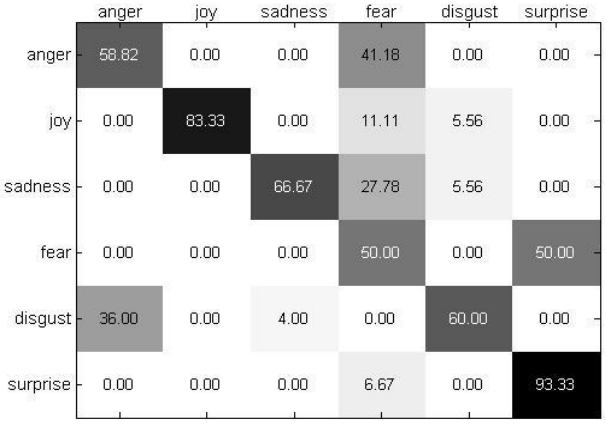

Figure 13. The confusion matrix of the best recognition performance on the eNTERFACE05 database.

In Figures 12 and 13, the numbers in different background colors are used to represent the recognized probability. The numbers in dark background colors represent the correctly recognized probability, and the numbers in other background colors represent the incorrectly recognized probability.

4.3. Comparisons with State-Of-The-Art Methods

Now, the current work is compared with previous work on the BAUM-1s database and eNTERFACE05 database. It is noted that previous research employed the same subject-independent test runs as ours. Table 5 shows a comparison of the state-of-the-art methods. The results of Table 5 indicate that our method obviously outperforms the state-of-the-art strategies. This shows the advantages of our framework. Specifically, in comparison with the previous best work, our method separately improves accuracy from 45.04% to 46.51% for the BAUM-1s database and from 42.16% to 43.72% on the eNTERFACE05 database. The methods in references [9,50] employ hand-crafted methods and shallow learning neural networks to extract low-level features. This also demonstrates that our high-level features perform better than hand-crafted features and indicates the validity of the deep learning method. AlexNet has the shortcoming of having too many parameters and consuming expensive computing time; however, our framework based on AlexNet has more parameters than AlexNet, and it is very difficult to adjust the parameters according to training datasets.

Table 5. Comparisons of the state-of-the-art methods.

Datasets	Refs	Accuracy
BAUM-1s	Zhalehpour et al., [9]	45.04%
	Ours	46.51%
eNTERFACE05	Zhalehpour et al., [9]	42.16%
	Mansoorizadeh et al., [50]	37.00%
	Bejani et al., [51]	39.27%
	Ours	43.72%

5. Comparison and Discussion

The results proved that our framework can effectively address the problem of VFER. We designed two types of streams to separately process the spatial signals and temporal signals of video. The spatial streams are effective and compact at characterizing the static appearance of video, and the temporal streams can effectively capture the dynamic changes of facial movements. The temporal streams can complement the spatial streams to recognize emotions. In order to avoid overfitting to a limited dataset, every stream utilizes state-of-the-art networks. It is challenging to pool the spatial streams and temporal streams; thus, we investigate the trained schema, non-trained schema, and the schema of inserting the EmotionalVlan layer at different places. As a result, utilizing the trained schema and inserting the EmotionalVlan layer after fc7 achieves the best recognition performance compared to other schema, as shown in Tables 1 and 2. Our framework can generate more comprehensive and effective features for VFER than the state-of-the-art method owing to the EmotionalVlan layer, which is a trainable aggregation layer and pools the temporal features and spatial features separately extracted by fc7.

6. Conclusions

This paper proposes a novel method of VFER by aggregating spatial–temporal features in order to classify video-based facial expression owing to the gap between the visual descriptors and emotions. This method can aggregate spatial and temporal convolutional features across the entire extent of a video. This proposed framework with an aggregating layer is end-to-end trainable and outperforms state-of-the-art methods. It can effectively avoid overfitting to the limited emotional video datasets, and the trainable strategy can learn to better represent an entire video. The experimental results show that the proposed method can separately improve accuracy from 45.04% to 46.51% on the BAUM-1s database and from 42.16% to 43.72% on the eNTERFACE05 database. Therefore, it is proved that the proposed method is helpful for future facial expression recognition in long videos. It is noted that the designed framework has a large number of parameters and requires expensive computing time. In future, it is necessary to further study how to reduce the parameters of our framework in order to accelerate the computing time.

Author Contributions: X.P. was responsible for the methodology, software, validation, and writing. W.G. was responsible for data collection. X.G. and W.L. were responsible for parts of the writing of the paper. J.X. was responsible for part of the writing the paper, and J.W. was responsible for parts of the code.

Funding: This work was supported by Zhejiang Provincial Public Welfare Project of China (Grant No. LGF19F020009), National Natural Science Foundation of China (Grant No.31771224, 61603228), National Key Research and Development Program of China (Grant No. 2018YFB1004901), Zhejiang Provincial Natural Science Foundation of China (Grant No. Y17C090031), the Science and Technology Program of Guangxi (Grant No.AB17129012), the Science and Technology Major Project of Guangxi (Grant No.AA17204096), Special Fund for Bagui Scholars of Guangxi, Shandong Social Science Planning Research Special, High Education Science and Technology Planning Program of Shandong Provincial Education Department (Grant No.18CHLJ18, J18KA340).

Acknowledgments: The authors would like to thank the anonymous reviewers and editors for their constructive comments and suggestions, which helped to enhance the presentation of this paper.

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

References

1. Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 3030–3043. [CrossRef]
2. Williams, A.C. Facial expression of pain: An evolutionary account. *Behav. Brain Sci.* **2002**, *25*, 455–488. [CrossRef]
3. Zhao, H.; Sun, M.; Deng, W.; Yang, X. A New Feature Extraction Method Based on EEMD and Multi-Scale Fuzzy Entropy for Motor Bearing. *Entropy* **2017**, *19*, 14. [CrossRef]
4. Jabon, M.; Bailenson, J.; Pontikakis, E.; Takayama, L.; Nass, C. Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Comput.* **2011**, *10*, 84–95. [CrossRef]
5. Deng, W.; Zhang, S.J.; Zhao, H.M.; Yang, X.H. A novel fault diagnosis method based on integrating empirical wavelet transform and fuzzy entropy for motor bearing. *IEEE Access* **2018**, *6*, 35042–35056. [CrossRef]
6. Leo, M.; Carcagnì, P.; Distante, C.; Spagnolo, P.; Mazzeo, P.; Rosato, A.; Petrocchi, S.; Pellegrino, C.; Levante, A.; De Lumè, F.; et al. Computational Assessment of Facial Expression Production in ASD Children. *Sensors* **2018**, *18*, 3993. [CrossRef] [PubMed]
7. Ekman, P. Facial expression and emotion. *Am. Psychol.* **1993**, *48*, 384–392. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
9. Zhalehpour, S.; Onder, O.; Akhtar, Z.; Erdem, C.E. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Trans. Affect. Comput.* **2017**, *8*, 300–313. [CrossRef]
10. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE' 05 Audio-Visual Emotion Database. In Proceedings of the International Conference on Data Engineering Workshops, Atlanta, GA, USA, 3–7 April 2006.
11. Ren, Z.; Skjetne, R.; Gao, Z. A Crane Overload Protection Controller for Blade Lifting Operation Based on Model Predictive Control. *Energies* **2019**, *12*, 50. [CrossRef]
12. Huibin, L.I.; Sun, J.; Zongben, X.U.; Chen, L. Multimodal 2D+3D Facial Expression Recognition with Deep Fusion Convolutional Neural Network. *IEEE Trans. Multimed.* **2017**, *19*, 2816–2831.
13. 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.
14. Zhang, K.; Huang, Y.; Du, Y.; Wang, L. Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks. *IEEE Trans. Image Process.* **2017**, *26*, 4193–4203. [CrossRef] [PubMed]
15. Zhao, S.; Liu, Y.; Han, Y.; Hong, R.; Hu, Q.; Tian, Q. Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1839–1849. [CrossRef]
16. Zhao, H.; Yao, R.; Xu, L.; Yuan, Y.; Li, G.; Deng, W. Study on a Novel Fault Damage Degree Identification Method Using High-Order Differential Mathematical Morphology Gradient Spectrum Entropy. *Entropy* **2018**, *20*, 682. [CrossRef]
17. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5297–5307.
18. Borza, D.; Danescu, R.; Itu, R.; Darabant, A. High-Speed Video System for Micro-Expression Detection and Recognition. *Sensors* **2017**, *17*, 2913. [CrossRef] [PubMed]
19. Liu, Y.; Li, Y.; Ma, X.; Song, R. Facial Expression Recognition with Fusion Features Extracted from Salient Facial Areas. *Sensors* **2017**, *17*, 712. [CrossRef] [PubMed]
20. Xie, W.; Shen, L.; Yang, M.; Lai, Z. Active AU Based Patch Weighting for Facial Expression Recognition. *Sensors* **2017**, *17*, 275. [CrossRef] [PubMed]
21. Sikka, K.; Wu, T.; Susskind, J.; Bartlett, M. Exploring bag of words architectures in the facial expression domain. In Proceedings of the International Conference on Computer Vision, Xiamen, China, 16–18 December 2012; pp. 250–259.
22. Wang, Z.; Ruan, Q.; An, G. Facial expression recognition using sparse local Fisher discriminant analysis. *Neurocomputing* **2016**, *174*, 756–766. [CrossRef]

23. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with Gabor wavelets. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998.
24. Nanni, L.; Brahnam, S.; Lumini, A. Local phase quantization descriptor for improving shape retrieval/classification. *Pattern Recognit. Lett.* **2012**, *33*, 2254–2260. [CrossRef]
25. Kayaoglu, M.; Erdem, C.E. Affect Recognition using Key Frame Selection based on Minimum Sparse Reconstruction. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 519–524.
26. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
27. Haryanto, I.; Ariyanto, M.; Caesarendra, W.; Dewoto, H.K. Development of Speech Control for Robotic Hand Using Neural Network and Stream Processing Method. *Internetworking Indones. J.* **2017**, *9*, 59–64.
28. Caesarendra, W.; Wijaya, T.; Tjahjowidodo, T.; Pappachan, B.K.; Wee, A.; Roslan, M.I. Adaptive Neuro-Fuzzy Inference System for Deburring Stage Classification and Prediction for Indirect Quality Monitoring. *Appl. Soft Comput.* **2018**, *72*, 565–578. [CrossRef]
29. Gajewski, J.; Vališ, D. The determination of combustion engine condition and reliability using oil analysis by MLP and RBF neural networks. *Tribol. Int.* **2017**, *115*, 557–572. [CrossRef]
30. Wilk-Kolodziejczyk, D.; Regulski, K.; Gumienny, G.; Kacprzyk, B.; Kluska-Nawarecka, S.; Jaskowiec, K. Data mining tools in identifying the components of the microstructure of compacted graphite iron based on the content of alloying elements. *Int. J. Adv. Manuf. Technol.* **2018**, *95*, 3127–3139. [CrossRef]
31. Kim, B.K.; Lee, H.; Roh, J.; Lee, S.Y. Hierarchical Committee of Deep CNNs with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition. In Proceedings of the ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 427–434.
32. Deng, W.; Yao, R.; Zhao, H.M.; Yang, X.H.; Li, G.Y. A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. *Soft Comput.* **2017**, 1–18. [CrossRef]
33. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going Deeper in Facial Expression Recognition using Deep Neural Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016.
34. Ding, H.; Zhou, S.K.; Chellappa, R.; Ding, H.; Zhou, S.K.; Chellappa, R. FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017.
35. Kahou, S.E.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent Neural Networks for Emotion Recognition in Video. In Proceedings of the ACM International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 467–474.
36. Rodriguez, P.; Cucurull, G.; Gonzalez, J.; Gonfaus, J.M.; Nasrollahi, K.; Moeslund, T.B.; Roca, F.X. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Trans. Syst. Man Cybern.* **2017**. [CrossRef] [PubMed]
37. Gao, J.; Fu, Y.; Jiang, Y.G.; Xue, X. Frame-Transformer Emotion Classification Network. In Proceedings of the ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 6–9 June 2017; pp. 78–83.
38. Tang, Y.; Zhang, X.M.; Wang, H. Geometric-Convolutional Feature Fusion Based on Learning Propagation for Facial Expression Recognition. *IEEE Access* **2018**, *6*, 42532–42540. [CrossRef]
39. Kim, D.H.; Baddar, W.; Jang, J.; Yong, M.R. Multi-Objective based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition. *IEEE Trans. Affect. Comput.* **2017**. [CrossRef]
40. Ballester, P.L.; Araujo, R.M. On the performance of GoogLeNet and AlexNet applied to sketches. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 1124–1128.
41. Bruhn, A.; Weickert, J.; Schnörr, C. Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods. *Int. J. Comput. Vis.* **2005**, *61*, 211–231. [CrossRef]
42. Viola, P.; Jones, M. Robust Real-time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
43. Zhang, S.; Zhao, X.; Chuang, Y.; Guo, W.; Chen, Y. Learning Discriminative Dictionary for Facial Expression Recognition. *IETE Tech. Rev.* **2017**, *33*, 1–7. [CrossRef]

44. Müller, C. The INTERSPEECH 2010 Paralinguistic Challenge. In Proceedings of the Interspeech, Chiba, Japan, 26–30 September 2010; pp. 2794–2797.
45. Deng, W.; Zhao, H.; Yang, X.; Xiong, J.; Sun, M.; Li, B. Study on an improved adaptive PSO algorithm for solving multi-objective gate assignment. *Appl. Soft Comput.* **2017**, *59*, 288–302. [CrossRef]
46. Deng, W.; Zhao, H.; Zou, L.; Li, G.; Yang, X.; Wu, D. A novel collaborative optimization algorithm in solving complex optimization problems. *Soft Comput.* **2017**, *21*, 1–12. [CrossRef]
47. Krishnapuram, B.; Carin, L.; Figueiredo, M.A.T.; Hartemink, A.J. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 957–968. [CrossRef] [PubMed]
48. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor based on 3D Gradients (HOG3D). In Proceedings of the BMVC 2008—19th British Machine Vision Conference, Leeds, UK, 1–4 September 2008.
49. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
50. Mansoorizadeh, M.; Charkari, N.M. Multimodal information fusion application to human emotion recognition from face and speech. *Multimed. Tools Appl.* **2010**, *49*, 277–297. [CrossRef]
51. Bejani, M.; Gharavian, D.; Charkari, N.M. Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. *Neural Comput. Appl.* **2014**, *24*, 399–412. [CrossRef]