Hindawi
Advances in Multimedia
Volume 2020, Article ID 8843413, 10 pages
https://doi.org/10.1155/2020/8843413

Research Article

Context-Aware Attention Network for Human Emotion Recognition in Video

Xiaodong Liu [1,2] **and Miao Wang** [1]

[1] *School of Computing Henan University of Engineering, Zhengzhou, China*
[2] *Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China*

Correspondence should be addressed to Xiaodong Liu; liuxiaodongxht@qq.com

Received 11 April 2020; Revised 18 October 2020; Accepted 25 October 2020; Published 12 November 2020

Academic Editor: Constantine Kotropoulos

Recognition of human emotion from facial expression is affected by distortions of pictorial quality and facial pose, which is often ignored by traditional video emotion recognition methods. On the other hand, context information can also provide different degrees of extra clues, which can further improve the recognition accuracy. In this paper, we first build a video dataset with seven categories of human emotion, named human emotion in the video (HEIV). With the HEIV dataset, we trained a context-aware attention network (CAAN) to recognize human emotion. The network consists of two subnetworks to process both face and context information. Features from facial expression and context clues are fused to represent the emotion of video frames, which will be then passed through an attention network and generate emotion scores. Then, the emotion features of all frames will be aggregated according to their emotional score. Experimental results show that our proposed method is effective on HEIV dataset.

1. Introduction

Estimating a person's emotional state is essential in our everyday life. This capacity is necessary to perceive and anticipate people's reactions [1]. Particularly, this emotion recognition challenge has a wide range of applications. For example, the emotional recognition platform can be used to recognize a potential suspicious person on intelligent security. Video recommendation services can match users' interest with video emotion, and the government sector can better understand people's response to hot events or new policies. Thus, human emotion recognition has attracted more and more attention as a new research field.

The human face contains rich emotional clues. Chu et al. [2] proposed a human emotion recognition method based on facial action coding system, which encodes facial expressions through a series of specific location movements of the face (action units). The action units can be identified by geometric features and appearance features extracted from face images [3]. Recently, along with the development of convolutional neural networks (CNNs), researchers attempt to further improve the performance of emotion recognition via CNNs [4]. Barrett et al. used CNNs to recognize action units and facial emotion. These studies were mainly focused on facial emotion recognition. However, the context information can also provide extra clues to recognize emotion. For example, persons are usually happy at a wedding and are usually sad at a funeral. When the context is incorporated, the recognition accuracy can be further improved. Previous researches have shown the importance of context in the perception of emotions [5]. In some scenarios, when we analyze a wider view rather than focus on the face of the person, we can more easily judge one's feeling. Kosti et al. [6] built an emotions-in-context database and showed the emotion recognition accuracy is improved when the person and the whole scene are jointly analyzed. Chen et al. [7] exploited context clues, including events, objects, and scenes for video emotion recognition, to improve performance. However, these methods treat the features of different frames equally and the difference of emotional information

contained in these frames is not considered. Although the research of context-aware video emotion recognition has made great progress, it still has two major challenges:

(1) Combination of face and context information. Face feature is associated with its context information. However, traditional video emotion recognition often computes the maximum or average value of images' feature of face and context separately and then fuses the features of these two modes, lack of organic fusion of face features, and context clues of the same image. Face feature and its context feature on the same image cannot be effectively integrated. As shown in Figure 1(a), when we try to estimate the emotion of the people in image sequences, context information is difficult to provide effective emotional features. For example, it is difficult to determine whether a person in the image sequences is teasing or being attacked by a dog through context information. However, when we combine face and context information in an image, it is easier to judge people's emotions as angry than using face information alone. Similarly detailed estimations can be made in Figure 1(b).

(2) Emotional differences in different images. Each frame in video contains a certain amount of emotional information, and these pieces of information can be complementary to each other. The most frequently used method is simply max/average pooling emotional features of all frames. However, different images of one video may contain different emotional information because of the difference of face size, pose, perspective, and context information. As an example, let us try to estimate the emotion of these people in Figure 2. In Figure 2(a), we can recognize that the emotion of the right image is joy with a greater probability. That is to say, the right image contains more emotional clues. Similar detailed estimations method can be made in the other images of Figure 2. Similarly, context information including surrounding environment and human body can also provide different emotional information. Therefore, how to solve the problem of emotional differences between different images is an important challenge for video emotional recognition.

To overcome the above two challenges, inspired by the attention mechanism [8, 9], we propose a context-aware attention network (CAAN), which is robust to frames containing less emotional information and simultaneously uses the rich emotional clues provided by the other frames. Firstly, CAAN uses two subnetworks to extract both face and context features, respectively, and these two features on the same image are fused to represent the emotion of the image. Similar to literature [6], we take as input the entire image and extract global features for providing the necessary contextual support. Then, an attention network takes as input the image feature and generates emotional score of the image. Finally, the emotion features of all images in one video will be aggregated according to their emotional score, and the final emotion representation of the video is produced.

In addition, existing video emotion recognition datasets, such as video emotion dataset [10] and Ekman Emotion Dataset [11], mainly focus on the psychological feelings of viewers brought by video content and there are no humans in many videos, which cannot effectively evaluate the human emotion in the videos. Therefore, this paper builds a human emotion in video (HEIV) dataset, which is based on video emotion dataset [10, 11] and downloads some videos from the network. The HEIV dataset contains 1012 videos and the human emotions in the videos are annotated according to the emotion category defined by psychologists Ekman and Friesen, as well as the neutral emotional categories. Besides, some videos also are annotated by neutral. We will describe it in detail in Section 3. The performance of the CAAN network is evaluated on the HEIV dataset. It improves top-1 matching rates over the state of the art by 2.22%.

The main contributions of the paper are summarized as follows.

We constructed a HEIV dataset consisting of 1012 annotated videos, which mainly focuses on human emotion in the video rather than the psychological feelings of viewers brought by video content in existing video emotion recognition datasets. It is important for the design of good video emotion recognition model.

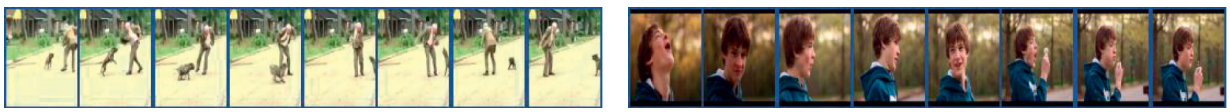
CAAN can automatically generate emotion scores for each frame and lead to better representation for the difference of emotional information in different video frames.

The effect of different weight function of attention mechanisms is evaluated which is helpful for the design of attention-based computational model.

The remainder of this paper is organized as follows. In the next section, we discuss related work on video-based emotion recognition. Section 3 describes the proposed dataset. Section 4 introduces the proposed CAAN. Section 5 gives experimental results. Section 6 concludes the paper and gives our future work.

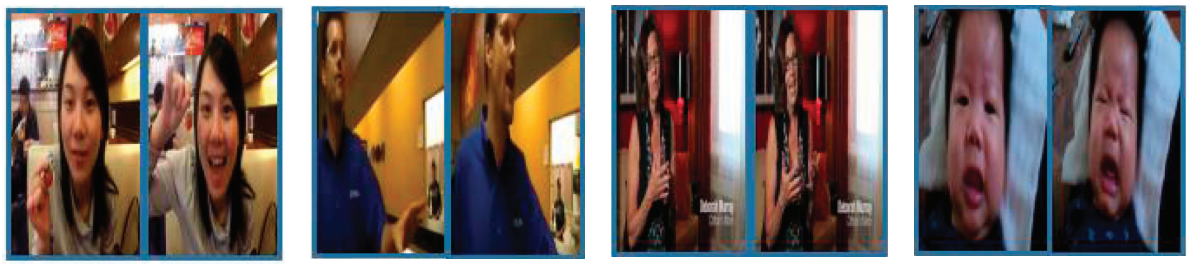
2. Related Work

2.1. Facial Emotion Recognition. Faces are the most commonly used stimuli to recognize the emotional states of people by researchers in computer vision. facial action coding system uses a set of specific localized movements of the face to encode the facial expression [2]. It deals with images in a nearly frontal pose [3]. However, facial images can be taken from multiple views or people may change their posture while being recorded. Some works that deal with multi-view emotion recognition have been proposed. Tariq et al. [12] learned a single classifier using data from multiple views. CSGPR [13] model performed the view normalization, where the features from different poses are combined. However, these approaches are not model relationships



(a) (b)

Figure 1: Illustration of information combination.



(a) (b) (c) (d)

Figure 2: Illustration of emotion difference.

among different views. Eleftheriadis et al. proposed a discriminative shared Gaussian process latent variable model for learning a discriminative shared manifold of facial expressions from multiple views [3]. Different from the existing multiple-view facial emotion recognition, this paper mainly solves the problem of different facial poses in emotion recognition in video. There are also a few emotion recognition works using other clues apart from the face. For example, Nicolaou et al. [14] considered the location of shoulders as additional information to the face features to recognize emotions.

2.2. Recognizing Emotion from Videos. There is some early work that recognized emotion through audio-visual features (e.g., [15–18]). Wang et al. [15] used audio-visual features to recognize emotion in 36 Hollywood movies. Irie et al. [16] extracted audio-visual features and combined them with a Hidden-Markov-like dynamic model. The audio-visual features fusion is evaluated by decision level fusion and feature level fusion [17]. However, they only use simple multimodal feature fusion without considering the potential relation of multimodal features, and the appearance features are low-level features. Singh et al. [19] proposed an improved technique for order preference by similarity to ideal solution (TOPSIS) method, which is based on the co-occurrence behavior of facial action coding system in the visual sequence to select key frames. Wang et al. [20] proposed two-level attention with two-stage multitask learning framework. Firstly, the features of corresponding region are extracted and enhanced automatically. Secondly, the relationship features of different layers are fully utilized by bidirectional RNN with self-attention. Wang et al. [21] defined a multimodal domain adaptive method to obtain the interaction between modes.

The performance of emotion recognition is evaluated by using different architectures CNN and different CNN feature layers in paper [11]. Nicolaou et al. [14] fused facial expression, shoulder posture, and audio clues for emotion recognition. Vielzeuf et al. [22] proposed a hierarchical approach, where scores and features are fused at different levels. It can retain the information of different levels, but the potential connection between multi-modal data in video has not yet been considered. Xue et al. [23] proposed a Bayesian nonparametric multimodal data modeling framework to learn emotions in videos, but it does not reflect the time evolution of emotional expression in videos. Kahou et al. [24] used CNN and RNN to model dynamic expression of videos, and the results show that the performance is better than the features fusion of frames. The temporal evolution of facial features is modeled through RNN in paper [25]. Zhang et al. [26] constructed kernel functions to convert CNN features into kernelized features. Xu et al. [27] conducted concept selection to investigate the relations between high-level concept features and emotions. This paper considers not only the emotional fusion of different facial features but also the difference of the amount of emotional information of video frames.

3. Human Emotion Dataset

We constructed a human emotion dataset based on video emotion dataset [10, 11] and videos downloaded from the web. Each video in the video emotion dataset is longer and contains multiple human clips in each video. It mainly focuses on the psychological feelings of viewers brought on by video content. We clipped human clips from videos of video emotion dataset and annotate the emotions of humans in the video. We also downloaded short video clips from YouTube. The database contains a total number of 1012 videos, and it uses a training set of 607 videos and a testing set of 405 videos. Figure 3 shows example frames of each emotion category from the HEIV dataset.

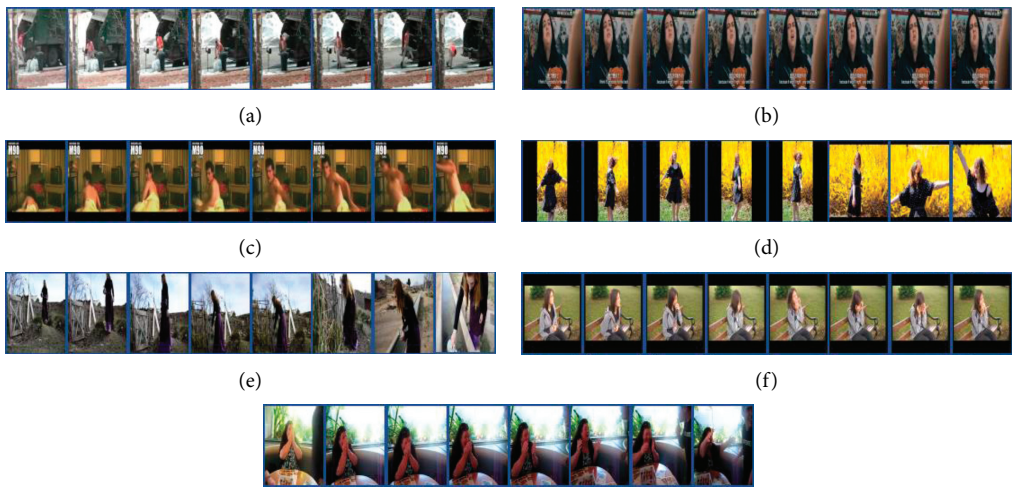


Figure 3: Example frames of each emotion category from the HEIV dataset. (a) Anger. (b) Disgust. (c) Fear. (d) Joy. (e) Neutral. (f) Sadness. (g) Surprise.

3.1. Video Annotation. The HEIV dataset was manually annotated by 10 annotators: 5 males and 5 females. Neutral and six emotion categories, including "anger," "disgust," "fear," "joy," "sadness," and "surprise," defined by psychologists Ekman and Friesen [28] are considered. In order to ensure the quality of the annotations, some videos clips with emotion labels coming from existing video emotion recognition dataset are exercised by annotators. After learning and practicing, annotators are asked to annotate our HEIV dataset. When we show a video with a person marked, we ask the annotators to select one of the emotion categories that suit that video. Each annotator independently annotates emotions, and emotion catalogue of a video marked by the most annotators is selected as the emotion label of the video. Furthermore, the gender (male/female) and the age range (child, teenager, adult) of persons in the video are also annotated.

3.2. Database Statistics. Of the 1012 annotated videos, 64% are males and 36% are females. Their ages are distributed as follows: 10% children, 11% teenagers, and 79% adults. Table 1 shows the number of videos for each of the categories.

Table 1: The number of videos per emotion category in HEIV dataset.

Category	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Number	103	105	121	207	125	158	197

4. Context-Aware Attention Network

In our work, we focus on improving the accuracy of emotion recognition. The primary challenge in human emotion recognition is the difference of facial scale, poses, perspective, and different degrees of contextual information. We aim to tackle this by context-aware attention network (CAAN), where facial and context emotion features are fused and the emotional scores of the fusion feature are generated by attention network. The fusion features of all images and their emotion scores are aggregated to make a human emotion prediction in video.

Our proposed framework is shown in Figure 4. The architecture consists of three main modules: two emotion feature extractors and an attention fusion module. The face feature extraction module takes as input the region of the human face and extracts its facial emotion features. The context extraction module takes as input the entire frame, which extracts global features for providing the necessary contextual information. Finally, the third module is an attention fusion network which takes as input the fusion features of face and context information. It is composed of two branches. The first branch is a tiny CNN network which takes as input the fusion feature and generates emotion features of frames. The other branch is also a tiny CNN network and is used to generate an emotion score for each frame. Then, the emotion features of frames and their emotion scores will be aggregated, and the final emotion representation of the human in the video will be produced.

4.1. Emotion Features Extraction. The face is the main part of a human to express emotion. Previous research on emotion recognition mainly focused on facial expression. However, the context plays an important role in emotion recognition, and when context information is incorporated, recognition accuracy can be further improved. To jointly analyze the human face and context features to recognize rich information about human emotion in the video, facial and contextual emotion features are extracted separately and then are fused. Meanwhile, the importance of fusion feature is judged by the attention mechanism. This section describes facial and contextual emotion features extraction.

In the emotion feature extraction stage, face features extraction module and context features extraction module

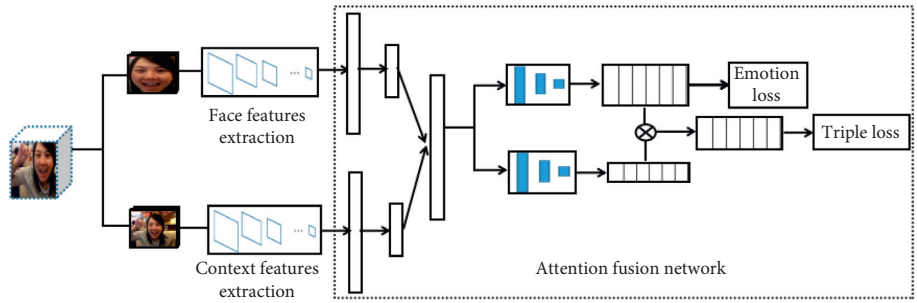


Figure 4: CAAN structure.

are used to extract face and context information, respectively. Given a video $V = (v_1, v_2, \ldots, v_K)$ with emotion labels where K is the total number of frames of the video V and v_i is the i-th video frame, human faces are first extracted from video frames by faster-rcnn [29] trained on WIDER face dataset [30]. Then, the faces detected in videos are resized to 224×224. Let n be the number of frames with faces in the video. The human faces in video V can be denoted as $F = (f_1, f_2, \ldots, f_n)$, where f_i is the human face extracted from v_i. VGG-Face model trained on the VGG-Face dataset [31] as initialization is used to extract facial emotional features. It is trained on the emotion recognition dataset and obtains the facial emotion feature extractor. In this paper, HEIV dataset is used to train face feature extractor and context feature extractor. HEIV is mainly used for human emotion recognition in video, and most images in video contain human face. VGG-Face uses face images as training samples and is supervised by emotional category to which the image belongs. The trained VGG-Face can extract the emotional features of different images. Therefore, given an image sequence $F = (f_1, f_2, \ldots, f_n)$, the fc6 feature is extracted as the facial emotion feature of each image through the forward propagation operation of face feature extractor. Let $X = \{x_i | i = 1, 2, \ldots, n\}$ denote the fc6 layer features of F, where x_i is the fc6 layer feature of f_i. In order to fuse face and context information on the same image, only image sequences containing faces $S = (I_1, I_2, \ldots, I_n)$ are selected in this paper. VGG network [32] is selected as context feature extractor. It is pre-trained on ImageNet dataset [33] and then trained on HEIV emotion recognition dataset. The whole image containing face is taken as training sample, and the emotional category of the video is taken as supervisory signal. The VGG trained can extract the contextual emotional features about the scenes, environments, and backgrounds of different images. Therefore, given an image sequence $S = (I_1, I_2, \ldots, I_n)$, the fc6 feature is extracted as the context emotion feature of each image through the forward propagation operation of context feature extractor. Let $C = (c_1, c_2, \ldots, c_n)$ denote the fc6 layer features of V, where c_i is the context information of v_i. The obtained facial features and context features are fed to the attention fusion network for effective fusion, so as to further improve the accuracy of video emotion recognition.

4.2. Attention Fusion Network. For a video clip v, we now have two high-level semantic features (X, C). These two features characterize the human of video from different perspectives and there are also differences in the amount of emotional information contained in different video frames. In order to fuse the features of face image sequence and context image sequence into a unified feature representation, the face feature sequence and context feature sequence can be fused separately, and then the face fusion feature and context fusion feature are fused. However, this will isolate the face feature from its context in the same image. Faces are closely related to the context of the same image, and their emotional features can complement each other to reflect a person's emotions in the image more comprehensively. In addition, traditional average pooling or maximum pooling feature fusion methods are difficult to effectively mine the complementarity between different image features and cannot reflect the emotional differences of different images. Therefore, this paper proposes an attention fusion network to effectively fuse image feature sequences of face and context. It can quantify the emotion difference of different video frames and fuse the features of face image sequence and context image sequence according to their importance and derive a unified feature representation.

More precisely, inspired by [7], we first transform face and context features of all images to a high-level space (1024 neurons for face and context features) and then face and context feature of each image are fused. Therefore, their distinct properties can be preserved and discriminative ability will be increased. Since these features are extracted from different video frames and have different discrimination, we then use an attention fusion mechanism to fuse these features, which can be able to be robust to frames with poor emotion information and simultaneously use the rich emotion information provided by the other video frames. Our basic idea is that each emotion feature can have an emotion score in aggregation, and the emotion features are aggregated according to their emotion scores. For that, emotion features are passed through two branches and then aggregated together. The first branch named fusion feature generation part extracts higher-level fusion emotion feature, and the other branch named emotion score generation part predicts an emotion score for each fusion feature. Features of video frames are then aggregated according to emotion scores.

Since the face and context features are extracted from the same frames demonstrating the same emotion in different forms, fusion feature generation part first fuses them using a fusion layer with 2048 neurons for absorbing all the information to obtain a shared representation, which can be expressed by the following:

$$r_i = \Phi(x_i, c_i), \tag{1}$$

where r_i is the fusion feature of the i-th video frame and $\Phi(,)$ is a fusion function.

The fusion feature r_i will be fed to two branch networks. The first branch named fusion generates subnetworks for generating higher-level fusion features, and it can be expressed by a fully connected layer.

$$g_i = W_1^f \times r_i + b_1^f. \tag{2}$$

The obtained higher-level fusion features are passed through a fully connected layer and generate emotion prediction vector. This branch is supervised by softmax-loss, which optimizes the probability of each image feature.

The other branch is the emotion score generation subnetwork, which is used to generate an emotion score for fusion feature of each image. We rely on an attention mechanism to obtain an emotion score. Its responsibility is first to analyze the amount of emotional information contained in video frames and then generate an emotion score which is used to bestow the feature with as much emotion information as possible. We use higher-level fusion features g_i to represent fusion features of each image, and its corresponding emotion score can be calculated using a fully connected layer that has only one cell, which is signed as 1CF,

$$s_i = W_1^s \times g_i + b_1^s, \tag{3}$$

where W_1^s and b_1^s are parameters to be learned for the emotion score generation part. Similarly, emotion score can also be generated by two or three successive fully connected layers, which is signed as 2CF and 3CF separately. In the experiments in Section 5, we will compare the effects of these different weighting functions.

The emotion feature representation of video V can be obtained by aggregating the fusion features g_i and emotion scores s_i of all images. It can be expressed as follows:

$$R_V = \frac{\sum_{i=1}^{n}(g_i \times s_i)}{\sum_{i=1}^{n} s_i}, \tag{4}$$

where R_V is the emotion feature of the video V. It is supervised by triplet loss [34], which minimizes variances of intra-class samples and discrimination of the emotion representation of the video.

5. Experiments

5.1. Effect of Weighting Function. In this subsection, we analyze the effect of different weighting function of emotion score generation part on the emotion recognition performance. First of all, we extract the fc6 layer feature of the face and context features by face feature extraction part and context feature contraction part. The fc6 features of face and context information are first passed through a fully connected layer with 1024 neurons and then fused. Fusion features are fed to two branches: one is used to generate higher-level fusion features, and the other is used to generate emotion scores. These two branches will be aggregated to generate the final emotion representation of the video. We consider three different weight functions of attention network, 1CF, 2CF, and 3CF, as described in Section 4.2.

We also give the evaluation results by the attention network which takes as inputs face and context information separately. For the face or context information, the network is divided into two branches beginning with pool5 layer features. The first branch is used to extract facial or context features through pre-training vgg face or vgg16 model, and the other branch takes as input the middle features of face or context information and generates emotion score for each face or context feature. Then, the facial or context emotion features and their emotion scores will be aggregated, and the final emotion representation of the face or context will be produced. Similar to the attention fusion network, emotion score can be calculated by one or two or three convolution layers and a fully connected layer that has only one cell, which is also signed as 1CF and 2CF and 3CF separately. Table 2 shows the accuracy of emotion recognition using different weight functions in attention networks on HEIV datasets.

As shown in Table 2, we observe that the recognition accuracy is different with different weighting functions in emotion score generation part, which means that attention mechanism can play an effective role in this situation. We also observe that 3CF is slightly better than 2FC and 1FC for fusion emotion features and context information, but 1CF is slightly better than 2FC and 3FC for the face features. From these three emotion feature accuracies, we can infer that deeper attention networks can get better emotion score, but when the attention network exceeds a certain degree, we cannot get better emotion score. We rely on the 3CF weighting function for fusion features and context features emotion score generation part and 1CF weighting function for facial emotion score generation part as the default in all subsequent experiments.

5.2. Effect of Attention Mechanism and Feature Fusion. In this subsection, we evaluate the performance of attention mechanism and feature fusion. In order to validate the effectiveness of our attention mechanism and feature fusion, we implement the following three average aggregate baseline approaches:

Face Average Aggregate (FAA). The fc6 layer features of all faces are extracted by VGG-Face. These features are aggregated by average pool and then passed through two successive fully connected layers and are supervised by softmax-loss.

Context Average Aggregate (CAA). The fc6 layer features of all context images are extracted by VGG16. These features

Table 2: Accuracy of emotion recognition on HEIV dataset.

Layers	Fusion features accuracy (%)	Context features accuracy (%)	Facial features accuracy (%)
1FC	50.37	42.72	48.89
2FC	51.11	43.46	44.44
3FC	51.85	43.95	47.90

are aggregated by the average pool and then passed through two successive fully connected layers and are supervised by softmax-loss.

Fusion Feature Average Aggregate (FFAA). The fc6 layer features of all faces and context images are extracted by VGG-Face and vgg16 separately. These two features are first passed through a fully connected layer with 1024 neurons and then fused. The fusion feature is passed through two successive fully connected layers and is supervised by softmax-loss.

Table 3 shows the accuracy of emotion recognition using attention mechanism and the above three average aggregation methods: FAA, CAA, and FFAA. As shown in Table 3, on HEIV dataset, attention mechanism increases top-1 emotion recognition accuracy by 5.43%, 2.22%, and 4.94%, respectively, compared with FAA, CAA, and FFAA. We also notice that feature fusion increases top-1 emotion recognition accuracy by 3.45% and 5.18%, respectively, compared with face features and context features on average aggregation and feature fusion increases top-1 emotion recognition accuracy by 2.96% and 7.9%, respectively, compared with face features and context features on attention mechanism. Based on these experiments, we can infer that attention fusion network outperforms the average aggregate method on HEIV dataset, and feature fusion outperforms single face or context feature on HEIV dataset.

5.3. Visualization of CAAN. In order to visualize the CAAN, some image sequences in the test set and their corresponding emotional scores are shown in Figure 5. As shown in Figure 5, the emotional scores of different images are different because of the difference of their facial posture and context information. Some images contain abundant emotion clues on human face and the context information, such as the 3rd image in Figure 5(b) and the 5th image in Figure 5(f); thus, CAAN gives these images higher emotional scores. On the contrary, some images contain little emotion clues on human face and the context information, such as the 1st image in Figure 5(a) and the 7th image in Figure 5(e), and CAAN gives these images lower emotional scores.

5.4. Comparison with State of the Art. We also compare state-of-the-art performance in recent literature. To validate the effectiveness of our CAAN method, we compare with the following state-of-the-art approaches on HEIV dataset.

5.4.1. Attention-Based Network. QAN [8] and attention clusters [9] are two attention-based networks. QAN is a quality network which takes as input images of video on HEIV dataset, and attention clusters are a multimodal attention network which takes as input fc6 layer features of face and context on HEIV dataset.

Table 3: Performance evaluation of attention and feature fusion.

Methods	Average aggregate accuracy (%)	Attention accuracy (%)
Face	43.46	48.89
Context	41.73	43.95
Fusion	46.91	51.85

5.4.2. Feature Fusion Network. Recent literature [6, 7, 24, 29] used multimodal feature fusion network. It implemented two modes of face and context on HEIV dataset.

Table 4 gives top-1 accuracy (%) of different methods on HEIV. As shown in Table 4, our context-aware attention fusion network achieves 2.22% performance gain on HEIV dataset. We also noticed that the performance of QAN only taking as input video frames is lower than fusion feature. By the attention mechanism, the performance of attention clusters [9] taking as input two modes of face and context is higher than feature fusion without attention mechanism. Note that our CAAN attains superior performance for two reasons: firstly, attention mechanism is robust to frames containing less emotional information and simultaneously uses the rich emotional clues provided by the other frames. Secondly, our feature fusion not only jointly exploits the face features and context information but also preserves their distinct properties. Based on these experiments, CAAN outperforms state-of-the-art results on HEIV datasets. The improvement of CAAN network proves CAAN's ability to deal with videos with different emotion information.

5.5. Confusion Matrix. To analyze the recognition accuracy of different emotion categories, we gave the confusion matrix of the recognition accuracy using CAAN on HEIV which is shown in Table 5. The vertical is true label, and the horizontal is the recognition accuracy of each emotion category. We observed that the surprise, fear, and disgust are well recognized, and the anger, neutral, and joy have a greater number of false positives. We inferred that it is because anger and joy add more emphasis on psychological activities, and their behavior expression is relatively low. We also noticed that 30.59% of joy is recognized as a surprise. We inferred that it is because some humans have both feelings of joy and surprise, and it is hard to determine which emotion dominates. We also observed that happiness is not recognized as disgust and neutral is not recognized as sadness. It is because the expressions of these two emotion categories are quite different.

Figure 5: Samples with their emotion scores predicted by CAAN.

Table 4: Top-1 accuracy (%) compared with state-of-the-art methods on HEIV

Method	Result (%)
Quality-aware network [8]	43.95
Fan et al. [25]	45.68
Vielzeuf et al. [22]	45.93
Chen et al. [7]	46.17
Kosti et al. [6]	46.42
Attention clusters [9]	49.63
Ours	51.85

Table 5: Confusion matrix.

	Predicted label						
True label	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	38.10	9.52	14.29	4.76	16.67	4.76	11.90
Disgust	2.33	55.81	11.63	4.65	2.33	16.28	6.98
Fear	2.13	4.26	57.45	2.13	8.51	19.15	6.38
Joy	1.18	0	8.24	47.06	9.41	3.53	30.59
Neutral	12.24	8.16	6.12	18.37	42.86	0	12.24
Sadness	1.64	11.48	14.75	8.20	1.64	54.10	8.20
Surprise	1.28	1.28	8.97	8.97	8.97	7.69	62.82

5.6. Result on Ekman-6 and VideoEmotion-8. In this section, we conduct experiments on Ekman-6 [35] and VideoEmotion-8 [10] datasets to further evaluate the effectiveness of our method.

Ekman-6 dataset contains 1637 videos, and it uses a training set of 819 videos and a testing set of 818 videos. It was manually annotated by 10 annotators according to Ekman's theory [28] on six basic human emotion categories, with a minimum of 221 videos per category.

VideoEmotion-8 dataset contains 1101 videos collected from YouTube and Flickr. The average duration of videos is 107 seconds. The experiments were conducted 10 times according to train/test splits provided by [10].

Table 6 gives top-1 accuracy (%) of different methods on Ekman-6 and VideoEmotion-8 datasets. As shown in Table 6, our context-aware attention fusion network achieves 1.83% and 1.68 performance gain on Ekman-6 and VideoEmotion-8 dataset, respectively. The results show that our methods achieve the state-of-the-art results on both Ekman-6 and VideoEmotion-8 datasets.

Table 6: Top-1 accuracy (%) compared with state-of-the-art methods on Ekman-6 and VideoEmotion-8.

Method	Ekman	VideoEmotion-8
Emotion in context [7]	51.8	50.6
Xu et al. [11]	50.4	46.7
Kernelized feature [26]	54.4	49.7
Concept selection [27]	54.40	50.82
Ours	56.23	52.5

6. Conclusion and Future Work

In this paper, we first built a video dataset with 7 categories of human emotion, named human emotion in the video (HEIV). With the HEIV dataset, we trained a context-aware attention network (CAAN) to recognize human emotion. CAAN consists of three modules. Two emotion feature extraction modules are used to extract face and context features, respectively. Attention fusion network fuses these two features and generates an emotion score for each fusion feature. Then, the fused emotion features will be aggregated according to their emotion score, and the final emotion representation of the video is produced. The performance of the CAAN network is evaluated and it can achieve excellent results on the HEIV dataset.

Although our approach obtains a promising performance in video emotion recognition, however, because of the diversity of human emotion expression, human emotion can be expressed through multiple body parts. In future work, we will further combine human part semantics for better recognition performance.

Data Availability

HEIV can be obtained by contacting liuxiaodongxht@qq.com.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the scientific and technological research project of Henan Provincial Science and Technology Department (182102310919) and the Foundation of He'nan Educational Committee (21A520006).

References

[1] K. Byoung, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, pp. 401–420, 2018.

[2] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 529–545, 2017.

[3] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 189–204, 2015.

[4] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[5] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[6] R. Kosti, J. M. Alvarez, A. Recasens et al., "Emotion recognition in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1960–1968, Honolulu, HI, USA, July 2017.

[7] C. Chen, Z. Wu, and Y. G. Jiang, "Emotion in context: deep semantic feature fusion for video emotion recognition," *ACM on Multimedia Conference*, vol. 16, pp. 127–131, 2016.

[8] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4694–4703, Honolulu, HI, USA, July 2017.

[9] X. Long, C. Gan, G. D. Melo et al., "Attention clusters: purely attention based local feature integration for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7834–7843, Honolulu, HI, USA, July 2017.

[10] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pp. 73–79, Québec, Canada, July 2014.

[11] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 15–22, New York, NY, USA, June 2016.

[12] U. Tariq, J. Yang, and T. S. Huang, "Multi-view facial expression recognition analysis with generic sparse coding feature," in *Proceedings of the Computer Vision - ECCV 2012. Workshops and Demonstrations European Conference on Computer Vision*, pp. 578–588, Firenze, Italy, October2012.

[13] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1357–1369, 2013.

[14] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[15] H. L. Wang and L. F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, 2006.

[16] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523–535, 2010.

[17] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2014.

[18] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," *Multimedia Tools and Applications*, vol. 61, no. 1, pp. 21–49, 2012.

[19] L. Singh, S. Singh, and N. Aggarwal, "Improved TOPSIS method for peak frame selection in audio-video human emotion recognition," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 6277–6308, 2019.

[20] X. Wang, M. Peng, L. Pan, H. Min, J. Chunhua, and R. Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 62, no. 7, pp. 217–225, 2019.

[21] Y. Wang, J. Wu, and H. Keiichiro, "Multi-attention fusion network for video-based emotion recognition," in *Proceedings of the 2019 International Conference on Multimodal Interaction*, pp. 595–601, Suzhou, China, October 2019.

[22] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 569–576, New York, NY, USA, November 2017.

[23] J. Xue, Z. Luo, K. Eguchi, T. Takiguchi, and T. Omoto, "A Bayesian nonparametric multimodal data modeling framework for video emotion recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 601–606, Hong Kong, China, July 2017.

[24] S. E. Kahou, V. Michalski, K. Konda et al., "Recurrent neural networks for emotion recognition in video," in *Proceedings of the ACM International Conference on Multimodal Interaction*, pp. 467–474, Seattle, WA, USA, November 2015.

[25] L. Fan and K. Yunjie, "Spatiotemporal Networks for Video Emotion Recognition," 2017, http://arxiv.org/abs/1704.00570.

[26] H. Zhang and M. Xu, "Recognition of emotions in user-generated videos with kernelized features," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2824–2835, 2018.

[27] B. Xu, Y. Zheng, H. Ye et al., "Video motion recognition with concept selection," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 406–411, Shanghai, China, July 2019.

[28] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[29] S. Ren, K. He, R. Girshick et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 91–99, Montreal, Canada, December 2015.

[30] S. Yang, P. Luo, C. L. Chen, and X. Tang, "Wider face: a face detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5525–5533, Las Vegas, NV, USA, June 2016.

[31] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, pp. 1–12, Swansea, UK, September 2015.

[32] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, http://arxiv.org/abs/1409.1556.

[33] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2014.

[34] F. Schroff, D. Kalenichenko, and P. James, "Facenet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, Boston, MA, USA, June 2015.

[35] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 255–270, 2018.