sensors

MDPI

Article

Context-Aware Emotion Recognition in the Wild Using Spatio-Temporal and Temporal-Pyramid Models

Nhu-Tai Do [1], Soo-Hyung Kim [1,*], Hyung-Jeong Yang [1], Guee-Sang Lee [1] and Soonja Yeom [2]

[1] Department of Artificial Intelligence Convergence, Chonnam National University, 77 Yongbong-ro, Gwangju 500-757, Korea; donhutai@gmail.com (N.-T.D.); hjyang@jnu.ac.kr (H.-J.Y.); gslee@jnu.ac.kr (G.-S.L.)
[2] School of Technology, Environment and Design, University of Tasmania, Hobart, TAS 7001, Australia; Soonja.Yeom@utas.edu.au
* Correspondence: shkim@jnu.ac.kr

Abstract: Emotion recognition plays an important role in human–computer interactions. Recent studies have focused on video emotion recognition in the wild and have run into difficulties related to occlusion, illumination, complex behavior over time, and auditory cues. State-of-the-art methods use multiple modalities, such as frame-level, spatiotemporal, and audio approaches. However, such methods have difficulties in exploiting long-term dependencies in temporal information, capturing contextual information, and integrating multi-modal information. In this paper, we introduce a multi-modal flexible system for video-based emotion recognition in the wild. Our system tracks and votes on significant faces corresponding to persons of interest in a video to classify seven basic emotions. The key contribution of this study is that it proposes the use of face feature extraction with context-aware and statistical information for emotion recognition. We also build two model architectures to effectively exploit long-term dependencies in temporal information with a temporal-pyramid model and a spatiotemporal model with "Conv2D+LSTM+3DCNN+Classify" architecture. Finally, we propose the best selection ensemble to improve the accuracy of multi-modal fusion. The best selection ensemble selects the best combination from spatiotemporal and temporal-pyramid models to achieve the best accuracy for classifying the seven basic emotions. In our experiment, we take benchmark measurement on the AFEW dataset with high accuracy.

Keywords: video emotion recognition; spatiotemporal; temporal-pyramid; best selection ensemble; facial emotion recognition

Citation: Do, N.-T.; Kim, S.-H.; Yang, H.-J.; Lee, G.-S.; Yeom, S. Context-Aware Emotion Recognition in the Wild Using Spatio-Temporal and Temporal-Pyramid Models. *Sensors* **2021**, *21*, 2344. https://doi.org/10.3390/s21072344

Academic Editor: Stefano Berretti

Received: 10 March 2021
Accepted: 25 March 2021
Published: 27 March 2021

1. Introduction

Emotional cues provide universal signals that enable human beings to communicate during the course of daily activities and are a significant component of social interactions. For example, people will use facial expressions such as a big smile to signal their happiness to others when they feel joyful. People also receive emotional cues (facial expressions, body gestures, tone of voice, etc.) from their social partners and combine them with their experiences to perceive emotions and make suitable decisions. In addition, emotion recognition, especially facial emotion recognition, has long been crucial in the human–computer interaction (HCI) field, as it helps computers efficiently interact with humans. Recently, several scientific studies have been conducted on facial emotion recognition (FER) in an attempt to develop methods based on new technologies in the computer vision and pattern recognition fields. This type of research has a wide range of applications, such as advertising, health monitoring, smart video surveillance, and development of intelligent robotic interfaces [1].

Emotion recognition on the basis of behavioral expressions presents numerous challenges due to the complex and dynamic properties of human emotional expressions. Human emotions change over time, are inherently multi-modal in nature, and differ in terms of such factors as physiology and language [2]. In addition, use of facial cues, which are

considered the key aspect of emotional cues, still presents challenges owing to variations in such factors as head poses and lighting conditions [3]. Several factors, such as body expressions and tone of voice are also affected by noise in the environment and occlusion. In some cases, emotions cannot be interpreted without context [4]. In video-based emotion recognition, facial expression representation often includes three periods, onset, apex and offset [5,6], as shown in Figure 1. The lengths of the periods differ; the onset and offset periods tend to be shorter than the apex period. There are challenges regarding the unclear temporal border between periods, and spontaneous expressions lead to multiple apexes.

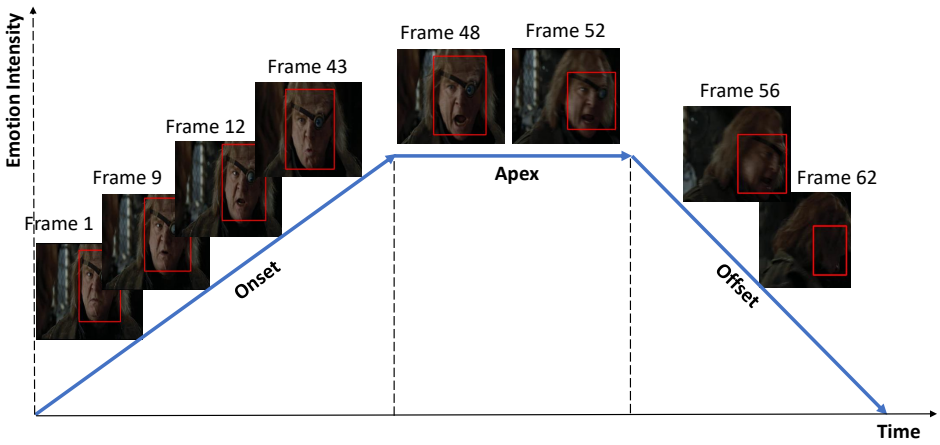

Figure 1. The three periods of facial expression representation are onset, apex, and offset. The duration of each varies, leading to unclear temporal borders. In addition, the appearance of spontaneous expressions leads to the presence of multiple apexes [6].

To address the above-mentioned challenges, both traditional and deep learning methods often focus on facial expressions that present changes in facial organs in response to emotional states, underlying intentions, and social interactions. Such methods attempt to determine facial regions of interest, represent changes in facial expressions, and divide emotions into six basic categories, namely, anger, disgust, fear, happiness, sadness, and surprise, as proposed by Ekman et al. [7].

In 2D image-based facial emotion recognition (2D FER), the main tasks focus on robust facial representation followed by classification. There are two approaches to feature representation, geometric- and appearance-based approaches. Geometric-based approaches represent facial expressions using geometric features of facial components (mouth, eyes, nose, etc.) in terms of shape, location, distance, and curvature [8–10]. Appearance-based approaches use local descriptors, image filters such as LBP [11], Gabor filters [12], PHOG [13], etc. to extract hand-crafted features for facial expression representation for traditional methods. In deep learning methods, feature representation is automatically extracted by convolutional neural networks (CNN) [14] that are trained on large-scale emotion recognition datasets such as RAF-DB [15] and AffectNet [16]. Geometric-based methods are often affected by noise and have difficulty showing small changes in facial details, while appearance-based methods are robust to noise and retain facial details. Deep learning models such as VGG-16 [17] and Resnet [18] demonstrate improved 2D FER performance [10,19].

In video-based emotion recognition, the main task focuses on efficiently exploiting spatiotemporal coherence to classify human emotion as well as integrating multiple modalities to improve overall performance. In the spatiotemporal approach, extensions of hand-crafted traditional features such as HOG, LBP and BoW are also proposed and applied using video-based emotion recognition methods such as 3D HOG [20], LBP-TOP [21], and Bag-Of-Word [22]. In addition, temporal models such as conditional random fields [23] and interval temporal Bayesian network [24] are used to exploit spatiotemporal relationships between different features. For deep learning-based methods, many works use CNNs for feature extraction followed by LSTM for exploiting spatiotemporal relations [25–27]. For the frame-level approach,

every frame in a video clip is subjected to facial feature extraction, concatenated together by a statistical operator (mim, mean, and std) using pre-determined time steps and finally classified by deep learning models or traditional classification methods such as SVM [28,29].

Recently, many works have focused on video-based emotion recognition to address challenges in emotion recognition using the deep learning approach. Zhu et al. [30] used a hybrid attention cascade network to classify emotion recognition with a hybrid attention module for the fusion features of facial expressions. Shi et al. [31] proposed a self-attention module integrated with the spatial-temporal graph convolutional network for skeleton-based emotion recognition. Anvarjon et al. [32] proposed deep frequency features for speech emotion recognition.

However, video-based emotion recognition also presents some challenges under in-the-wild conditions, such as problems involving head pose, lighting conditions, and the complexity in the facial expression representation due to spontaneous expression. Context is key in emotion recognition. For instance, in a dark environment or when the face of interest is tiny, it is possible to recognize emotions based off our experiences with related elements such as parts of the scene, body gestures, things, and other people in the scene. In addition, a hierarchical structure in the emotion feature representation is necessary to deal with unclear emotion temporal borders.

In this study, we propose an overall system with face tracking and voting to select the main face for emotion recognition using two models based on spatiotemporal and temporal-pyramid architecture to efficiently improve emotion recognition. For face tracking and voting, we use a tracking-and-detection template with robust appearance features as well as motion features to suggest faces and people. Then, through a voting scheme based on probabilities, occurrences, and sizes, we choose the face and person of interest in the video clip.

In video-based emotion recognition, we first deal with in-the-wild conditions by integrating contextual features, facial emotion probability, and facial emotion features to construct a robust set of facial emotion features. For unclear temporal border and spontaneous expression problems, we propose a temporal-pyramid architecture to integrate face-context features by time steps based on statistical information. The hierarchical structure of facial-context feature integration improves the emotion evaluation results of our system. Moreover, we also propose a spatiotemporal model using "Conv2D+LSTM+3DCNN+Classify" architecture to exploit spatiotemporal coherence among face-context emotion features in 3D and 2D+T strategies. Finally, we suggest the best ensemble method to choose the best combination among models. Our experiment was conducted on the AFEW dataset [33] which is the dataset of the EmotiW Challenge 2019 [34]. We achieved good performance on the validation set and test set.

The contributions of this paper are as follows: (1) We integrate facial emotion features with scene context features to improve performance. (2) We propose spatiotemporal models to exploit spatiotemporal coherence among face-context features using 3D and 2D+T temporal strategies. In addition, we build a temporal-pyramid model to exploit the hierarchical structure of overall face-context emotion features by statistical operator. (3) Our proposed system achieved good performance on a validation set taken from the AFEW dataset [33].

This paper is organized into seven sections. In Section 2, we briefly summarize related works. We describe our proposed idea in Section 3. We discuss the network architectures in Section 4 and the best selection ensemble method in Section 5. Our experiments are shown in Section 6. Finally, the conclusions are outlined in Section 7.

2. Related Works

2.1. Image-Based Facial Expression Recognition

Emotion recognition plays a fundamental role in human–computer interactions (HCIs). It is used to automatically recognize emotions for a wide range of applications, such as customer marketing, health monitoring, and emotionally intelligent robotic interfaces.

Emotion recognition remains a challenging task due to the complex and dynamic properties of emotions, their tendency to change over time, the fact that they are often mixed with other factors, and their inherently multi-modal nature in terms of behavior, physiology, and language.

To recognize emotion expression, the face is one of the most important visual cues. Facial expression recognition (FER) exploits the facial feature representation of static images [11] in the spatial domain. Traditional methods use handcrafted features such as local binary patterns (LBPs), speeded-up robust features (SURF), and scale-invariant feature transform (SIFT), to classify emotions. Recently, with the success of deep learning in computer vision tasks, FER problems raise the new challenge for classifying emotions under in-the-wild environments despite occlusions, illumination differences, etc. Many 2D FER image datasets such as AffectNet [16], RAF-DB [15], etc. have been published to promote technological development and fulfill the requirement for large-scale and real-world datasets.

2.2. Video-Based Emotion Recognition

From still images to video, emotion recognition presents many serious challenges; these involve, for example, behavioral complexities, environmental effects, and temporal changes in the video channel, as well as acoustic and language differences in the audio channel. To provide a baseline for video emotion recognition in the wild, the AFEW dataset [33] was built from many movies and TV shows. Emotions are classified into seven categories (anger, disgust, fear, happiness, neutrality, sadness, and surprise) under uncontrolled environments such as outdoor/indoor scenes, illumination changes, occlusions, and spontaneous expression. From 2013 to 2018, the emotion recognition research community made great strides through the EmotiW Challenge [34] on the basis of the AFEW dataset [33].

Because human emotions are almost always displayed on the face by movements of facial muscles, many studies have focused on facial representations in attempts to exploit the spatial and temporal information contained in a video. There are three main approaches to this problem: geometry, video-level, and frame-level approaches.

For the geometry approach, Liu et al. [26] computed 3D landmarks, normalized these landmarks and extracted features using Euclidean distances. They proposed the Landmark Euclidean Distance network. Kim et al. [27] proposed the CNN-LSTM network to classify emotions through sequential 2D landmark features.

For the spatiotemporal approach, Liu et al. [26] used the VGG Face network to extract facial features and then used these facial features to classify emotions. They showed an accuracy of 43.07% on the validation set. Lu et al. [25] proposed VGG-Face+BLSTM [35] for the spatiotemporal network using the VGG-Face network fine-tuned on facial expression images from video clips. This model showed an accuracy of 53.91%.

Finally, the main idea of the frame-level approach is to merge emotion features in every frame using an aggregation function (min, max, std, etc.). It addresses the invariance of the number of video frames. Bargal et al. [29] used facial emotion recognition networks to extract facial features and concatenated the results.For all frames, they used the statistical encoding module (STAT) to merge all frame-level features by min, max, variance, and average. They showed a high accuracy of 58.9% on the validation set. Knyazev et al. [28] later updated the STAT* module by scaling and normalization.

We realize that weak points exist in the above works that make use of the spatiotemporal, frame-level, and audio modalities. For instance, the spatiotemporal networks do not integrate 3DCNN [36] and BiLSTM [35] to find strong correlations between the spatial information in the data cube. Moreover, it would be better to use online fine-tuning in the video training process instead of offline feature extraction.

For the frame-level approach, STAT encoding does not utilize temporal information between the frame-level features. In addition, the frame-level features need to add more contextual information such as action information and scene information. The audio approach only uses one type of acoustic feature for emotion classification.

3. Proposed Idea

In this section, we define the problem that we wish to address and give a brief overview of our video emotion recognition system. Next, we explain our proposed method in detail, including the tracking and voting modules and method of face context feature extraction. The details of the model are discussed in the next section.

3.1. Problem Definition

In this study, the input is a video clip $\mathbf{V} = \{\mathbf{S}, \mathbf{A}\}$ lasting 5 min or less consisting of a scene sequence $\mathbf{S}$ and audio stream $\mathbf{A}$. Certain cues play an important role in human emotion recognition, such as facial expression, body gestures, and tone of voice. In the scope of our work, we mainly focus on visual cues that are important to the perception of human feelings. Face tracking $\mathbf{F}$, along with the corresponding person tracking $\mathbf{P}$ comprising body and scene information, are the most important cues to solve this problem. Our objective is to effectively locate the significant face $\overline{\mathbf{F}}$ and corresponding person $\overline{\mathbf{P}}$ from the scene sequence $\mathbf{S}$. From there, we use the face and person image sequences $\mathbf{S}_{\overline{\mathbf{F}}}$ and $\mathbf{S}_{\overline{\mathbf{P}}}$ to classify the emotion $c_i \in \mathbf{C} = [0, 6]$ as one of seven basic emotions, namely, anger, disgust, fear, happiness, neutrality, sadness, and surprise.

Let $\mathbf{c}_{fj}^t \triangleq \left(x_{fj}^t, y_{fj}^t, w_{fj}^t, h_{fj}^t\right) \in \mathbb{R}^4$ and $\mathbf{c}_{pj}^t \triangleq \left(x_{pj}^t, y_{pj}^t, w_{pj}^t, h_{pj}^t\right) \in \mathbb{R}^4$ be the location of the jth face and person at time t in a scene sequence $\mathbf{S} \in \mathbb{R}^{W \times H \times T}$, respectively, and the tracking indices $g_j^t \in \mathbb{R}$ calculated using the tracking module shown in Figure 2, where $\left(x_{fj}^t, y_{fj}^t\right) / \left(x_{pj}^t, y_{pj}^t\right)$ is the face/person center, $\left(w_{fj}^t, h_{fj}^t\right) / \left(w_{pj}^t, h_{pj}^t\right)$ is the face/person size, $W \times H$ is the size of a scene, and T is the length of scene sequence $\mathbf{S}$. The scene sequence $\mathbf{S}$ then contains the face tracking $\mathbf{F}$ and person tracking $\mathbf{P}$ information, defined as follows:

$$\begin{aligned} \mathbf{F} &= \{\mathbf{f}_i\}^{i=1\ldots N_f} \text{ and } \mathbf{P} = \{\mathbf{p}_i\}^{i=1\ldots N_f}, \\ \mathbf{f}_i &= \left\{\mathbf{c}_{fj_k}^{t_k} | t_k < t_{k+1} \wedge g_{j_k}^{t_k} = i\right\}^{k=1\ldots M_i} \text{ and} \\ \mathbf{p}_i &= \left\{\mathbf{c}_{pj_k}^{t_k} | t_k < t_{k+1} \wedge g_{j_k}^{t_k} = i\right\}^{k=1\ldots M_i} \end{aligned} \tag{1}$$

where N_f is the number of tracked faces and persons and $\mathbf{f}_i / \mathbf{p}_i$ is the ith tracked face/person that contains the locations of the face/person in chronological order ($t_k < t_{k+1}$) and which has a length of M_i and the same tracking index ($g_{j_k}^{t_k} = i$).

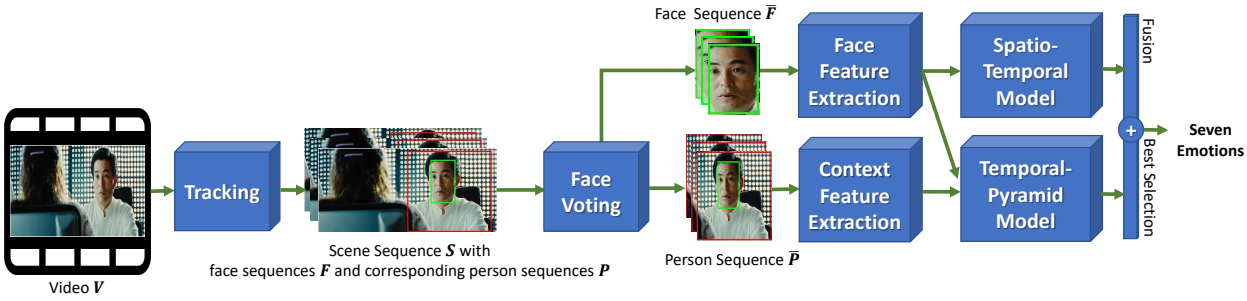

Figure 2. Overview of the proposed system for video emotion recognition in the wild.

We also denote $\mathbf{S}_{\mathbf{f}_i}$ and $\mathbf{S}_{\mathbf{p}_i}$ as, respectively, the image sequences of a tracked face and person, $\mathbf{f}_i$ and $\mathbf{p}_i$, extracted from the scene sequence $\mathbf{S}$. The emotional expression in the video $\mathbf{V}$ is mostly affected by the most significant face $\overline{\mathbf{F}}$, which appears more often and is larger than the other faces, and the corresponding person $\overline{\mathbf{P}}$, defined as follows:

$$\begin{aligned} \overline{\mathbf{F}} &= mode\{\mathbf{F}\} \\ \overline{\mathbf{P}} &= select_{g_{\overline{\mathbf{P}}} = g_{\overline{\mathbf{F}}}}\{\mathbf{P}\} \end{aligned} \tag{2}$$

where $g_{\overline{\mathbf{P}}}$ and $g_{\overline{\mathbf{F}}}$ are the tracking indices of $\overline{\mathbf{P}}$ and $\overline{\mathbf{F}}$, respectively.

The goal of our method is to classify the image sequences $\mathbf{S}_{\overline{\mathbf{F}}}$ and $\mathbf{S}_{\overline{\mathbf{P}}}$ of the dominant tracked face $\overline{\mathbf{F}}$ and corresponding person $\overline{\mathbf{P}}$ to classify what kind of emotions exist in the video $\mathbf{V}$. The classification result is denoted by a classification label $c \in [0, 6]$ corresponding to the seven basic emotions anger, disgust, fear, happiness, neutrality, sadness, and surprise.

3.2. Proposed System

An overview of our proposed system is shown in Figure 2. The system attempts to classify a video clip in the wild according to seven categorical emotions, namely anger, disgust, fear, happiness, neutrality, sadness, and surprise.

The key to this study is context-aware emotion recognition in video clips. The expression of the key face in a video clip signifies the emotion that the system will apply to that clip. The contextual features from the person region are used to improve the performance of the system when the key face is small and/or occluded. Our proposed model exploits the context-aware feature map to classify emotions into seven basic categories.

First, from an input video clip, our system effectively locates the most important tracked face $\overline{\mathbf{F}}$ and corresponding tracked person $\overline{\mathbf{P}}$ using the Tracking and FaceVoting module. These are considered the most significant characteristics to help our system classify emotional expression.

Second, the face context feature map is extracted from the significant face $\overline{\mathbf{F}}$ and person $\overline{\mathbf{P}}$ using the face feature extraction and context feature extraction models. The face feature extraction model is based on conventional models and uses pre-trained weights based on the AffectNet [16] and RAF-DB [15] datasets. The context feature extraction model is VGG16 [17], with pre-trained weights from ImageNet.

The context spatiotemporal LSTM-3DCNN model uses LSTM [37] or 3DCNN [36] to exploit the spatiotemporal correlation of the face context feature map and fine-tune the face feature extraction model. Its scheme is "FaceContext+LSTM+Conv3D+Classification" and it helps our system learn the feature map more deeply.

Moreover, we propose the context temporal-pyramid model based on the temporal-pyramid scheme instead of LSTM and 3DCNN. The face context feature map can be enhanced by the temporal-pyramid scheme as well as statistical operators (mean, max, and min). It exploits the long-term dependencies in all time-steps from the face context feature map. Our system applies categorical cross-entropy loss for training on the seven basic emotion classes for every video emotion model.

Finally, we fuse the classification features from all models to achieve the best accuracy in emotion classification. We propose the best selection ensemble and compare it to average fusion and join fine-tuning fusion [10]. The best selection ensemble finds the best combination of models by the heuristic principle when giving a first specific model. It attempts to find an unused model to help the current combination achieve the best accuracy with a smaller number of models to prevent over-fitting.

3.3. Face and Person Tracking

For the tracking module, we propose a tracking algorithm based on a tracking-by-detection scheme [38] and Hungarian matching method [39] to return the tracked faces $\mathbf{F}$ along with the corresponding tracked persons $\mathbf{P}$ from the scene sequence $\mathbf{S}$.

3.3.1. Tracking Database of Tracked Faces and Persons

It is assumed that there are tracked faces $\mathbf{F} = \{\mathbf{f}_i\}^{i=1...N_t}$ and corresponding tracked persons $\mathbf{P} = \{\mathbf{p}_i\}^{i=1...N_t}$ at the time t, where N_t is the number of tracked faces and persons, and $\mathbf{f}_i$ (or $\mathbf{p}_i$) is the location sequence of a tracked face (or person) as defined in Equation (1). Let $\mathbf{D} = \{\mathbf{d}_i\}^{i=1...N_t}$ be the tracking database containing appearance and motion observations.

Our algorithm uses the HSV color histogram and the face features to record appearance observations. The last face size and location of a tracked face record motion observations. Each element $\mathbf{d}_i = \left(\mathbf{d}_i^{hsv}, \mathbf{d}_i^{enc}, \mathbf{d}_i^{pos}, \mathbf{d}_i^{size}\right) \in \mathbf{D}$ is calculated as follows:

$$\begin{aligned} \mathbf{d}_i^{hsv} &= \left\{ \mathbb{H}_{hsv}\left(\mathbf{S}_{\mathbf{f}_i^j}\right) \right\}^{j=M_i-k+1\ldots M_i} \\ \mathbf{d}_i^{enc} &= \left\{ \mathbb{G}_{vggface2}\left(\mathbf{S}_{\mathbf{f}_i^j}\right) \right\}^{j=M_i-k+1\ldots M_i} \\ \mathbf{d}_i^{pos} &= (x,y)_{\mathbf{f}_i^{M_i}} \text{ and } \mathbf{d}_i^{size} = (w,h)_{\mathbf{f}_i^{M_i}} \end{aligned} \tag{3}$$

where $\mathbf{d}_i^{hsv}$ is the HSV color histograms of the last k-face images $\left\{\mathbf{S}_{\mathbf{f}_i^j}\right\}$ for the tracked face $\mathbf{f}_i$; M_i is the number of faces in $\mathbf{f}_i$; and the operator $\mathbb{H}(.)$ is used to generate 100 bin values of a 2D histogram using the H and S channels for color, and 20 bin values of a 1D histogram using the V channel for brightness, as mentioned in [40]. $\mathbf{d}_i^{enc}$ is the face encoding features of the last k-face images, which is extracted from the model $\mathbb{G}$ that uses pre-trained weights from VGGFace2 [41]. $\mathbf{d}_i^{pos}$ and $\mathbf{d}_i^{size}$ are respectively the last position and size of the tracked face $\mathbf{f}_i$.

3.3.2. Face and Person Candidates

For every scene $\mathbf{s}_t \in \mathbf{S}$, our algorithm uses Tiny Face Detector [42] to extract face candidates. This is is a robust detector that finds small faces with high efficiency. We also use SSD detection [43] trained on the VOC dataset to detect person candidates. For every face candidate, we find the person candidate that yields the smallest intersect over union (IoU) score. If this is not possible, the whole scene is used as the person region.

Let $\forall \mathbf{c}_{fj}^t = \left(x_{fj}^t, y_{fj}^t, w_{fj}^t, h_{fj}^t\right) \in \mathbf{C_F}$ and $\forall \mathbf{c}_{pj}^t = \left(x_{pj}^t, y_{pj}^t, w_{pj}^t, h_{p,j}^t\right) \in \mathbf{C_P}$ be, respectively, the face and person candidates in the scene $\mathbf{s}_t$, where $\left(x_{fj}^t, y_{fj}^t\right) / \left(x_{pj}^t, y_{pj}^t\right)$ is the face/person center, and $\left(w_{fj}^t, h_{fj}^t\right) / \left(w_{pj}^t, h_{pj}^t\right)$ is the face/person size. We need to extract appearance and motion observations of the face candidates. Let $\mathbf{O_F} = \{\mathbf{o}_j\}$ be the appearance and motion observations of face candidates $\mathbf{C_F}$; then, every element $\mathbf{o}_j = \left(\mathbf{o}_j^{hsv}, \mathbf{o}_j^{enc}, \mathbf{o}_j^{pos}, \mathbf{o}_j^{size}\right)$ is computed as follows:

$$\begin{aligned} \mathbf{o}_j^{hsv} &= \mathbb{H}_{hsv}\left(\mathbf{S}_{\mathbf{c}_{fj}^t}\right) \text{ and } \mathbf{o}_j^{enc} = \mathbb{G}_{vggface2}\left(\mathbf{S}_{\mathbf{c}_{fj}^t}\right) \\ \mathbf{o}_j^{pos} &= (x,y)_{\mathbf{c}_{fj}^t} \text{ and } \mathbf{o}_j^{size} = (w,h)_{\mathbf{c}_{fj}^t} \end{aligned} \tag{4}$$

where $\mathbf{S}_{\mathbf{c}_{fj}^t}$ is the corresponding image of $\mathbf{c}_{fj}^t$, the operator $\mathbb{H}(.)$ is used to extract the HSV color histogram, and the pre-trained VGGFace2 model $\mathbb{G}$ is used to compute the face-encoding features.

3.3.3. Face and Person Matching

Let $\mathbf{M}^v$ be the cost matrix of the observation $v \in \{hsv, enc, pos, size\}$ between the face candidates $\mathbf{C_F}$ and the tracked faces $\mathbf{F}$. We use a Euclidean distance operator $\mathbb{E}(.)$ to calculate every element $\mathbf{M}_{ij}^v \in \mathbf{M}^v$ as follows:

$$\mathbf{M}_{ij}^v = \begin{cases} \mathbb{E}\left(\overline{\mathbf{d}}_i^v, o_j^v\right) & if\ \mathbb{E}\left(\overline{\mathbf{d}}_i^v, o_j^v\right) \leq \mathbb{T}^v \\ \infty & otherwise \end{cases} \tag{5}$$

where the operator $\overline{\mathbf{d}}_i^v$ is the mean of $\mathbf{d}_i^v$, $\mathbb{T}^v$ is the valid threshold of observation v (determined experimentally), i is the face index in $\mathbf{F}$, and j is the candidate index in $\mathbf{C_F}$ and $\mathbf{C_P}$.

The total cost matrix $\mathbf{M}$ is the weighted sum of $\forall \mathbf{M}^v$ with every element $\mathbf{M}_{ij}$ calculated as follows:

$$\mathbf{M}_{ij} = \sum_v w_v \frac{\mathbf{M}_{ij}^v}{\sum_{\notin \infty} \mathbf{M}^v} \tag{6}$$

where w_v is the weighted term of the observation v and $\sum_{\notin \infty}$ is the sum of elements other than ∞.

Our algorithm uses the Hungarian matching method [39] to find the optimal solution for which each tracking candidate $\mathbf{c}_{fj}^t$ (or $\mathbf{c}_{pj}^t$) is assigned to at most one tracking object $\mathbf{f}_i$ (or $\mathbf{p}_i$) and each tracking object $\mathbf{f}_i$ (or $\mathbf{p}_i$) is assigned to at most one tracking candidate $\mathbf{c}_{fj}^t$ (or $\mathbf{c}_{pj}^t$) as follows:

$$\sum_i \sum_j \mathbf{M}_{ij} \mathbf{X}_{ij} \rightarrow min \tag{7}$$

where $\mathbf{X}$ is a Boolean matrix with $\mathbf{X}_{ij} = 1$ if the tracking candidate $\mathbf{c}_{fj}^t$ (or $\mathbf{c}_{pj}^t$) is assigned to the tracking object $\mathbf{f}_i$ (or $\mathbf{p}_i$).

Then, we compute tracking indices $\left\{ g_j^t \right\}$ to assign the jth tracking candidate $\mathbf{c}_{fj}^t$ (or $\mathbf{c}_{pj}^t$) to the tracked objects $\mathbf{F}$ (or $\mathbf{P}$) as follows:

$$g_j^t = \begin{cases} i & if\ \mathbf{X}_{ij} = 1 \wedge \forall v, \mathbf{M}_{ij}^v \neq \infty \\ \infty & otherwise \end{cases} \tag{8}$$

3.3.4. Face and Person Update

For $g_j^t = i$, the tracking candidate $\mathbf{c}_{fj}^t$ (or $\mathbf{c}_{pj}^t$) is assigned to the tracked object $\mathbf{f}_i$ (or $\mathbf{p}_i$) as follows:

$$\begin{aligned} \mathbf{f}_i &= \mathbf{f}_i \oplus \mathbf{c}_j \\ \mathbf{d}_i^v &= \mathbf{d}_i^v \oplus \mathbf{o}_j^v\ ,\ v \in \{hsv, enc\} \\ \mathbf{d}_i^v &= \mathbf{o}_j^v\ ,\ v \in \{pos, size\} \end{aligned} \tag{9}$$

where the operator $\oplus$ is used to insert an element into the last position of an array.

Otherwise, for $g_j^t = \infty$, the candidate $\mathbf{c}_{fj}^t$ (or $\mathbf{c}_{pj}^t$) is a new tracking object to be inserted into the set of tracked objects $\mathbf{F}$ (or $\mathbf{P}$) as follows:

$$\begin{aligned} \mathbf{F} &= \mathbf{F} \oplus \left\{ \mathbf{c}_j \right\} \\ \mathbf{D} &= \mathbf{D} \oplus \left\{ \mathbf{o}_j^v \right\}, v \in \{hsv, enc, pos, size\} \end{aligned} \tag{10}$$

3.4. Face Voting

For the FaceVoting module, the system votes on the most significant face that has the largest influence on human emotional perception. Therefore, the inputs are the tracked faces $\mathbf{F}$ and tracked persons $\mathbf{P}$. The outputs are the most significant tracked face $\overline{\mathbf{F}}$ and the corresponding person $\overline{\mathbf{P}}$, which are used in the emotion classification.

The most important tracked face is the face that occurs more often and more clearly than the other tracked faces. It is assessed through frequency of occurrence, face size, and face probability. Given the tracked faces $\mathbf{F} = \{\mathbf{f}_i\}^{i=1...M_i}$ and persons $\mathbf{P} = \{\mathbf{p}_i\}^{i=1...M_i}$, the weighted terms of frequency of occurrence, face size, and face probability of each tracked face $\mathbf{f}_i$ and tracked person $\mathbf{p}_i$ are computed as follows:

$$
\begin{aligned}
w^i_{freq} &= \frac{M_i}{T} \\
w^i_{size} &= \frac{\sum_{j=1}^{M_i} w^i_j \times h^i_j}{M_i \times W \times H} \\
w^i_{prob} &= \frac{\sum_{j=1}^{M_i} p^i_j}{M_i}
\end{aligned}
\tag{11}
$$

where (W, H) and T are, respectively, the size and length of the scene sequence $\mathbf{S}$ and $\left(w^i_j, h^i_j\right)$ and p^i_j are, respectively, the size and detection probability of the jth face in the tracked face $\mathbf{f}_i$.

The weighted term of each tracked face $\mathbf{f}_i$ and tracked person $\mathbf{p}_i$ is calculated as follows:

$$
w^i = c_{freq} w^i_{freq} + c_{size} w^i_{size} + c_{prob} w^i_{prob} \tag{12}
$$

where $c_{x \in \{freq, size, prob\}}$ is a constant term that is used to adjust the priority of frequency of occurrence, face size, and face probability features in the face voting process.

The significant tracked faces $\overline{\mathbf{F}}$ and corresponding tracked persons $\overline{\mathbf{P}}$ have a weight that reaches a maximum value:

$$
\begin{aligned}
i_{max} &= \arg\max_{i \in [1 \ldots M_i]} w^i \\
\overline{\mathbf{F}} &= \mathbf{f}_{i_{max}} \\
\overline{\mathbf{P}} &= \mathbf{p}_{i_{max}}
\end{aligned}
\tag{13}
$$

From there, we extract the face images $\mathbf{S}_{\overline{\mathbf{F}}}$ and person images $\mathbf{S}_{\overline{\mathbf{P}}}$ based on tracked face $\overline{\mathbf{F}}$ and tracked person $\overline{\mathbf{P}}$, respectively.

3.5. Face and Context Feature Extraction

The Face and Context Feature Extraction module produces face and context features and probabilities for each of the seven emotions from the face and person regions using the face and context feature extraction models shown in Figure 3.

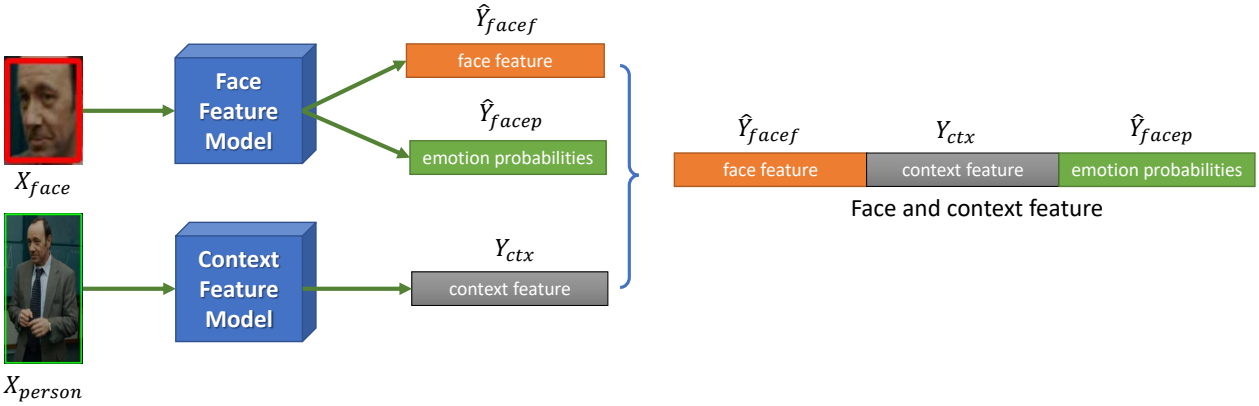

Figure 3. Face and Context Feature Extraction.

Let $\mathbb{M}_{face}$ be the face feature model which is built on conventional base networks such as Resnet [18], SEnet [44], Xception [45], Nasnet mobile [46], Densenet [47], Inception Resnet [48], VGG Face 1 [49], VGG Face 2 [41], and ImageNet [50]. The model receives

a face image X_{face} and returns prediction emotion probabilities $\hat{Y}_{facep} \in \mathbb{R}^7$ and feature vector $\hat{Y}_{facef} \in \mathbb{R}^K$ as follows:

$$\hat{Y}_{facep}, \hat{Y}_{facef} = \mathbb{M}_{face}\left(X_{face}\right) \tag{14}$$

where K is the feature size and $\hat{Y}_{facep}$ is the one-hot encoding vector used to determine the emotion label c by $c = \arg\max \hat{Y}_{facep}$.

In this study, we trained $\mathbb{M}_{face}$ on the AffectNet dataset [16] and fine-tuned it on the RAF-DB dataset [15] with category cross-entropy (CCE) loss as follows:

$$\mathcal{L}_{CCE} = -\sum_{c \in \mathbf{C}} Y_{facep} \log \hat{Y}_{facep} \tag{15}$$

where c is the emotion label in the set of seven basic emotions $\mathbf{C}$.

Similarly, let $\mathbb{M}_{ctx}$ be the context feature model, which extracts the context feature vector Y_{ctx} from the person image X_{person} as follows:

$$Y_{ctx} = \mathbb{M}_{ctx}\left(X_{person}\right) \tag{16}$$

where the context feature extraction model $\mathbb{M}_{ctx}$ is built on the VGG16 model [17] with weights pre-trained on ImageNet.

Formally, we want the face feature model $\mathbb{M}_{face}$ and the context model $\mathbb{M}_{ctx}$ to follow the following distribution:

$$p\left(c|X_{face}, X_{person}\right) = p\left(c|\hat{Y}_{facef}, \hat{Y}_{facep}, Y_{ctx}\right) \tag{17}$$

The context around a person's region is used to improve the performance of our model when the tracked face is very small or occluded. By extracting the feature vector with a model trained on ImageNet, we exploit the image diversity in ImageNet, and integrate this information into the face feature vector to identify correlations among the face and context characteristics and the emotion probability vector.

4. Network Architectures

4.1. Context Spatiotemporal LSTM-3DCNN Model

Overview. The context spatiotemporal LSTM-3DCNN model shown in Figure 4 incorporates the face, context feature blocks $\mathbb{M}_{face}$ and $\mathbb{M}_{ctx}$, the LSTM block $\mathbb{M}_{LSTM}$, the 3DCNN block $\mathbb{M}_{3dcnn}$, and the classification block $\mathbb{M}_{clas}$. Our proposed model uses the face and context feature blocks $\mathbb{M}_{face}$ and $\mathbb{M}_{ctx}$ to extract the face and context feature vectors. Use of the context feature vector helps to improve the accuracy of our model in difficult cases such as those with occluded face, small face, etc. Next, the LSTM block $\mathbb{M}_{LSTM}$ exploits the temporal correlation among the feature vectors and normalizes the information to a fixed-length spatiotemporal feature map where the first axis is the temporal dimension and the second and third axes are the spatial dimension. The 3DCNN block $\mathbb{M}_{3dcnn}$ learns spatiotemporal information from the spatiotemporal feature map to produce the high-level emotional features. From there, the classification block $\mathbb{M}_{clas}$ classifies the emotion as one of the seven basic categories.

The context feature vectors play an important role in performance improvement. It deals with the difficulties in emotion recognition when the faces are occluded and small. Moreover, it integrates contextual features with body posture, visual scene, social situations, etc. to explain human emotion instead of using only facial cues in emotion recognition.

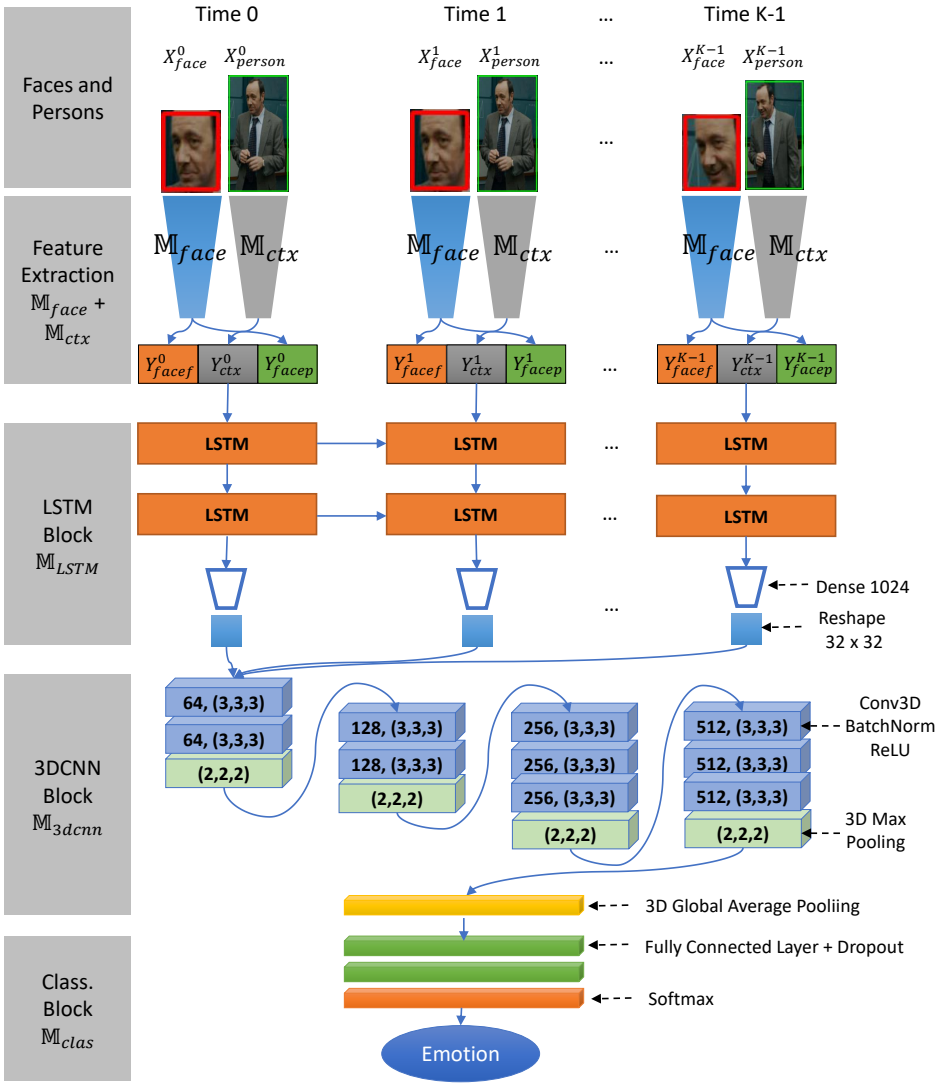

Figure 4. Context Spatio-Temporal LSTM-3DCNN Model.

Implementation Details. Given the significant tracked face $\mathbf{S}_{\overline{\mathbf{F}}}$ and corresponding person $\mathbf{S}_{\overline{\mathbf{P}}}$ in the input image sequences, the module applies random temporal sampling to transform the input image sequences into sequences with a fixed length of K as follows:

$$\begin{aligned} \left\{X_{face}^{t}\right\}_{t=1}^{K} &= TemporalSampling\left(\mathbf{S}_{\overline{\mathbf{F}}}\right) \\ \left\{X_{person}^{t}\right\}_{t=1}^{K} &= TemporalSampling\left(\mathbf{S}_{\overline{\mathbf{P}}}\right) \end{aligned} \tag{18}$$

where K is the size of the sampling operator with a value of 32.

The network uses the face and context feature blocks $\mathbb{M}_{face}$ and $\mathbb{M}_{ctx}$ to transform every input face image X_{face}^{t} and person image X_{person}^{t} at time step $t = \overline{1,K}$ in the input sequences. The outputs return the face probability vector Y_{facep}^{t}, face feature vector Y_{facef}^{t}, and context feature vector Y_{ctx}^{t}:

$$\begin{aligned} Y_{facep}^{t}, Y_{facef}^{t} &= \mathbb{M}_{face}\left(X_{face}^{t}\right) \\ Y_{ctx}^{t} &= \mathbb{M}_{ctx}\left(X_{person}^{t}\right) \end{aligned} \tag{19}$$

Finally, they are combined to form the overall face context feature vector $Y^t_{face\_ctx}$ as follows:

$$\begin{aligned} Y^t_{face\_ctx} &= concat\left(Y^t_{facep}, Y^t_{facef}, Y^t_{ctx}\right) \\ &= concat\left(\mathbb{M}_{face}\left(X^t_{face}\right), \mathbb{M}_{ctx}\left(X^t_{person}\right)\right) \end{aligned} \tag{20}$$

where the *concat* operator is used to combine feature vectors.

We freeze the first layers of $\mathbb{M}_{face}$ with the exception of the end layers, which have roles in feature extraction and emotion classification. This helps the face feature model $\mathbb{M}_{face}$ not only transfer knowledge from the model pre-trained on large-scale image emotion recognition datasets [15,16] but also to be fine-tuned again at frame level on the video emotion dataset [33]. For $\mathbb{M}_{ctx}$, we freeze all layers and only extract the context feature that is learned from the model that is pre-trained on the large-scale ImageNet dataset [50].

To exploit the long-term dependencies, the LSTM block $\mathbb{M}_{LSTM}$ consists of stacked LSTM layers where each LSTM memory cell at layer i computes the hidden and state vectors h^t_i, c^t_i from the current face context feature $Y^t_{face\_ctx}$ (for layer 0) or the hidden vector h^t_{i-1} (for layer i>0), and the hidden and cell states after the previous LSTM memory cell h^{t-1}_i, c^{t-1}_i:

$$h^t_i, c^t_i = \begin{cases} LSTM\left(Y^t_{face\_ctx}, h^{t-1}_0, c^{t-1}_0\right), & i = 0 \\ LSTM\left(h^t_{i-1}, h^{t-1}_i, c^{t-1}_i\right), & 0 < i < L \end{cases} \tag{21}$$

where L is the number of LSTM layers in $\mathbb{M}_{LSTM}$. In this study, we chose $L = 2$ by experiment.

Next, we use the Dense and Reshape layers to normalize every hidden state vector h^t_{L-1} at the last LSTM layer to a specific length and produce the spatiotemporal feature map $Y_{lstm} \in \mathbb{R}^{K \times S \times S}$ of the face and context feature vectors $\left\{Y^t_{face\_ctx}\right\}$ as follows:

$$Y_{lstm} = \left\{Reshape_{S \times S}\left(Dense_L\left(h^t_{L-1}\right)\right)\right\} \tag{22}$$

where $L = S \times S$, and $(S \times S)$ are the fixed-length and (width, height) used to normalize and reshape the hidden state vector, respectively, and K is the number of time-steps.

To perform a deeper analysis of the spatiotemporal feature map Y_{lstm} in the temporal domain and ensure spatial coherence of the feature domain, the 3DCNN block $\mathbb{M}_{3dcnn}$ is used to produce the emotional high-level feature Y_{3dcnn} from Y_{lstm} as folows:

$$Y_{3dcnn} = \mathbb{M}_{3dcnn}(Y_{lstm}) \tag{23}$$

where $\mathbb{M}_{3dcnn}$ consists of four 3D convolutional blocks and a global average pooling layer. Every 3D convolutional block has 3D convolutional layers, followed by a batch normalization layer, and a rectified linear unit (ReLU), along with a 3D max pooling layer, at the end. The number of 3D convolutional layers and the kernel size of each one are, respectively: (2, 64), (2, 128), (3, 256), and (4, 512). All 3D convolutional layers use $3 \times 3 \times 3$ filters and a padding of 1. The 3D max pooling layers have a size of $2 \times 2 \times 2$.

Lastly, $\mathbb{M}_{clas}$ receives the emotion feature Y_{3dcnn} and classifies it into the seven basic emotions. $\mathbb{M}_{clas}$ comprises two fully-connected layers followed by ReLU layers and dropout layers. At the end of the block, a softmax layer is used to output the emotion probability vector $Y_{emotion}$ as follows:

$$Y_{emotion} = \mathbb{M}_{clas}(Y_{3dcnn}) \tag{24}$$

Finally, we use categorical cross-entropy loss for emotion classification as follows:

$$CCE(Y_{gt\_emotion}, Y_{emotion}) = -\sum_{i=1}^{C} Y_{i,gt\_emotion} \log Y_{i,emotion} \quad (25)$$

where $Y_{gt\_emotion}$ is the ground-truth; $Y_{emotion}$ is the prediction result of the model; and C is the number of emotion labels.

4.2. Context Temporal-Pyramid Model

Overview. The context temporal-pyramid model illustrated in Figure 5 comprises the face and context blocks $\mathbb{M}_{face}$ and $\mathbb{M}_{ctx}$, the temporal-pyramid block $\mathbb{M}_{stp}$, and the classification block $\mathbb{M}_{clas}$. The model has some similarities to the context spatiotemporal model in that it uses $\mathbb{M}_{face}$ and $\mathbb{M}_{ctx}$ for face context feature extraction and $\mathbb{M}_{clas}$ for emotion classification. However, the model exploits the face context features during all time steps in long-term temporal dependencies. The temporal-pyramid block $\mathbb{M}_{stp}$ provides all face context features from the feature extraction block to the statistical aggregation $\mathbb{M}_{stat}^{k=l_1,l_2,\ldots,l_P}$ models where P is the number of statistical aggregation models. Each $\mathbb{M}_{stat}^{k}$ builds the temporal pyramid features at level k. It will divide the time steps into 2^k feature subsequences and aggregate the face and context features using the mean operator and face probabilities by max, mean, and min operators. From there, all temporal pyramid features at all pyramid levels are combined into the context temporal pyramid feature to exploit the long-term dependencies of the face context features in all time-steps. Finally, emotion classification is done by $\mathbb{M}_{clas}$.

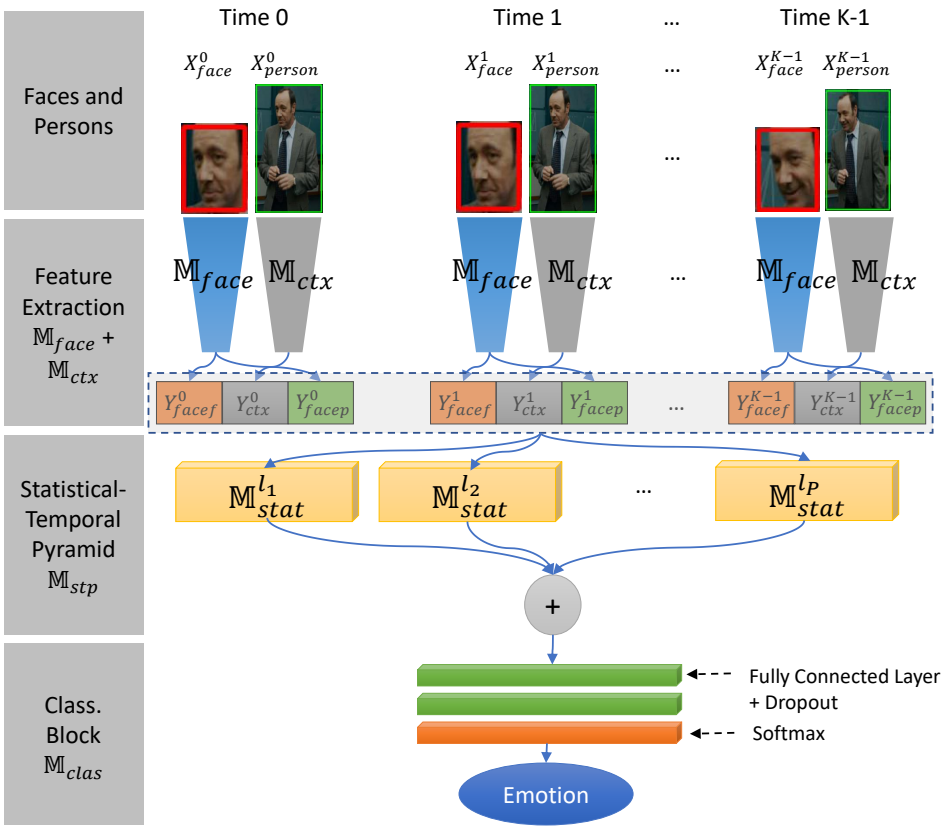

Figure 5. Context Temporal-Pyramid Model.

Implementation Details. The context temporal-pyramid model uses all faces and persons in $\mathbf{S}_{\overline{\mathbf{F}}}$ and $\mathbf{S}_{\overline{\mathbf{P}}}$ to extract face context features $\left\{Y_{face_ctx}^t = \left(Y_{facep}^t, Y_{facef}^t, Y_{ctx}^t\right)\right\}_{t=1\ldots M_{\overline{\mathbf{F}}}}$ using Equation (19), where $M_{\overline{\mathbf{F}}}$ is the number of elements in $\mathbf{S}_{\overline{\mathbf{F}}}$ and $\mathbf{S}_{\overline{\mathbf{P}}}$.

Then, the temporal-pyramid block $\mathbb{M}_{stp}$ exploits the long-term temporal dependencies during all time steps through a temporal pyramid scheme. It consists of statistical aggregation $\mathbb{M}_{stat}^{k=l_1,l_2,\ldots,l_P}$ models where every statistical aggregation $\mathbb{M}_{stat}^{k}$ transforms the face context features $\left\{Y_{fac\_ctx}^{t}\right\}$ into temporal pyramid features at level k, as shown in Figure 6.

The $\mathbb{M}_{stat}^{k}$ model divides all time steps into k time step sub-sequences $TS_j^k = \left[\frac{(j-1)n}{k}, \frac{j.n}{k}\right)_{j=1\ldots k}$ where $n = \mathbf{S}_{\overline{\mathbf{F}}}$ is the number of faces and persons in $\mathbf{S}_{\overline{\mathbf{F}}}$ and $\mathbf{S}_{\overline{\mathbf{P}}}$. The face context features $\left\{Y_{face\_ctx}^{i}\right\}_{i\in TS_j^k}$ in every time step sub-sequence j are transformed into the temporal pyramid feature V_j^k using the operator mean for face and context features and min, mean, and max for face probabilities, as follows:

$$\begin{aligned} V_{k,facef}^{j} &= mean_{i\in TS_j^k}\left\{Y_{facef}^{i}\right\} \\ V_{k,ctx}^{j} &= mean_{i\in TS_j^k}\left\{Y_{ctx}^{i}\right\} \\ V_{k,facep}^{j} &= concat\left(\min_{i\in TS_j^k}\left\{Y_{facep}^{i}\right\}, mean_{i\in TS_j^k}\left\{Y_{facep}^{i}\right\}, \max_{i\in TS_j^k}\left\{Y_{facep}^{i}\right\}\right) \end{aligned} \tag{26}$$

where the *mean*, *max*, and *min* operators are used to create an aggregate of the mean, max, and min from the vector values and the *concat* operator combines all values in a vector. The correlation between $\left(V_{k,facef}^{j}, V_{k,ctx}^{j}\right)$ and $V_{k,facep}^{j}$ is exploited at every time step sub-sequence j in pyramid level k, which helps our model learn the long-term temporal dependencies.

The temporal pyramid feature V_j^k is a combination of $V_{k,facef}^{j}$, $V_{k,ctx}^{j}$, and $V_{k,facep}^{j}$ as follows:

$$V_k^j = concat\left(V_{k,facef}^{j}, V_{k,ctx}^{j}, V_{k,facep}^{j}\right) \tag{27}$$

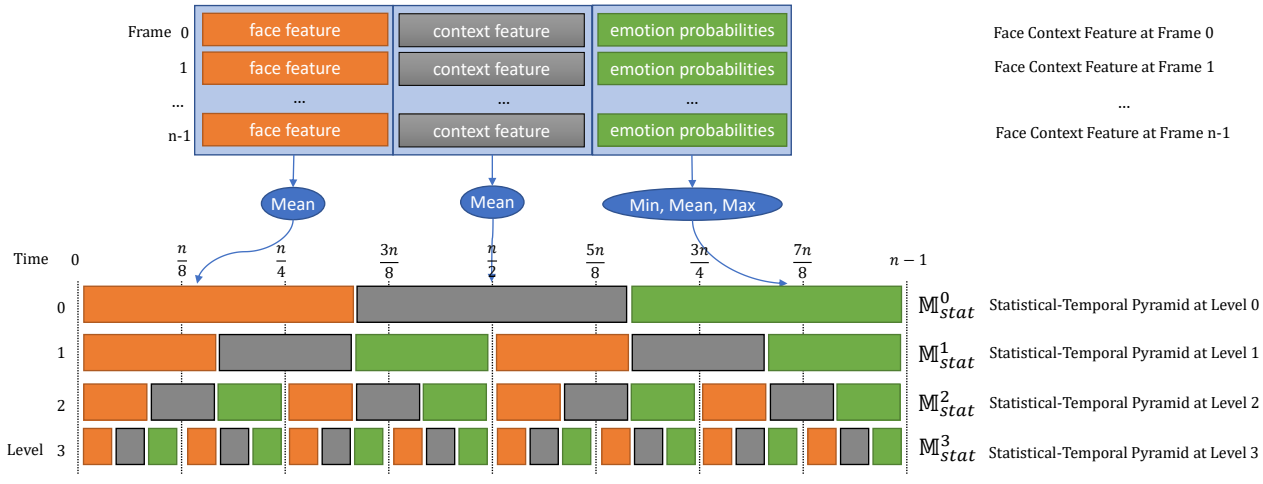

Figure 6. Context Temporal-Pyramid Features.

Finally, the temporal-pyramid block $\mathbf{M}_{stp}$ incorporates $\mathbb{M}_{stat}^{k}$ models in pyramid levels $k = \{l_1, l_2, \ldots, l_P\}, l_i \in [0\ldots 3]$ to produce the context temporal-pyramid feature Y_{stp} as follows:

$$Y_{stp} = concat_{k\in\{l_1,l_2,\ldots,l_P\}}\mathbb{M}_{stat}^{k}\left(\left\{\left(Y_{facep}^{t}, Y_{facef}^{t}, Y_{ctx}^{t}\right)\right\}_{t=1\ldots M_{\overline{\mathbf{F}}}}\right) \tag{28}$$

From there, we use the classification block $\mathbf{M}_{clas}$, the architecture of which is similar to that of $\mathbf{M}_{clas}$ in the context spatiotemporal model for emotion classification to produce emotion probabilities, as shown in Equation (24). We also apply categorical cross-entropy loss to train the model, as shown in Equation (25).

5. Best Selection Ensemble

The main idea of an ensemble method is to identify the best combination of the given models to solve the same tasks. The main advantage of ensemble methods is that they effectively use the large margin classifiers to reduce variance error and bias error [51].

We propose a **best selection ensemble** method to combine multi-modality information to address the bias error problem. Our method applies the heuristic principle to find the best combination of the given models at every selection step. We search all model combinations with the given first model and keep the shortest combination to prevent over-fitting.

First, it is assumed that the outputs of the $\{M_k\}_{k=1...K}$ models are prediction emotion probability vectors $\{\hat{Y}_k\}_{k=1...K}$ defined as follows:

$$\hat{Y}_k = \{\hat{y}_{k,i}\}_{i\in[1,N_E]}, \sum_{i\in[1,N_E]} \hat{y}_{k,i} = 1 \tag{29}$$

where K and $N_E = 7$ are the number of models and emotion labels, respectively. The average fusion $\mathbb{F}_{avg}$ of $\{M_k\}_{k=1...K}$ is calculated as follows:

$$\mathbb{F}_{avg}(\{M_k\}) = \left\{ \frac{\sum_{k=1}^{K} \hat{y}_{k,i}}{K} \right\}_{i\in[1,N_E]} \tag{30}$$

The multi-modal score is calculated based on the accuracy metric between the fusion result and the ground truth, as follows:

$$\mathbf{Score}_{acc}(\{M_k\}) = acc\big(\mathbb{F}_{avg}(\{M_k\}), Y_{gt}\big) \tag{31}$$

where the acc operator is used to calculate the accuracy of the prediction compared to ground truth.

Without loss of generality, we assume that $\{M_k\}_{k=1...K}$ is sorted in descending accuracy where $\mathbf{Score}_{acc}(M_i) > \mathbf{Score}_{acc}(M_j)$ if $i < j$.

Let $Select$ be the model-combination set. Initially, $Select$ is empty. We sequentially choose the first model M_{s_1} from left to right in $\{M_k\}_{k=1...K}$ and attempt to find the optimal list of model selections corresponding to the given model M_{s_1}.

Let $Open = \{M_k\}_{k=1...K} \backslash \{M_{s_1}\}$ be the open list of models that can be selected for processing. $Close = \{M_{s_1}\}$ is then the closed list of the selected models.

At step l, it is assumed that $Open = \{M_k\}_{k=1...K} \backslash \left\{M_{s_{j=1...l}}\right\}$ and $Close = \left\{M_{s_{j=1...l}}\right\}$. We select the first model M_v from left to right in $Open$ such that the following is satisfied:

$$\begin{aligned} &\mathbb{F}_{avg}\Big(Closed \bigcup \{M_v\}\Big) > \mathbb{F}_{avg}(Closed) \\ &\Big|Closed \bigcup \{M_v\}\Big| <= \mathbb{T} \end{aligned} \tag{32}$$

where $\mathbb{T} = 5$ is the threshold of the number of models in $Closed$ (determined experimentally).

If a model M_v cannot be found, we stop at this step and update the $Select$ list as follows:

$$Select = Select \bigcup \{Closed\} \tag{33}$$

We then repeat the process to select the first model in the next position. Finally, we choose the model combination in *Select* with the highest accuracy and smallest number of models.

6. Experiments and Discussion

6.1. Datasets

6.1.1. Image-Based Emotion Recognition in the Wild

In this work, we chose suitable datasets for training of the face feature extraction model. The datasets must deal with the in-the-wild environments where there are many unconstrained conditions, such as occlusion, poses, illumination, etc. AffectNet [16] and RAF-DB [15] are by far the largest datasets satisfying the above criteria. The images in the datasets are collected from the Internet based on emotion-related keywords. Emotion labels are annotated by experts to guarantee reliability.

AffectNet [16] contains two data groups, manual and automatic groups, with more than 1,000,000 images that are labeled with 10 emotion categories as well as dimensional emotion (valence and arousal). We used only images in the manual group belonging to seven basic emotion categories (anger, disgust, fear, happiness, neutrality, sadness, and surprise). Thus, we used 283,901 images for training and 3500 images for validation. The data distributions of in the training and validation sets are shown in Figure 7.

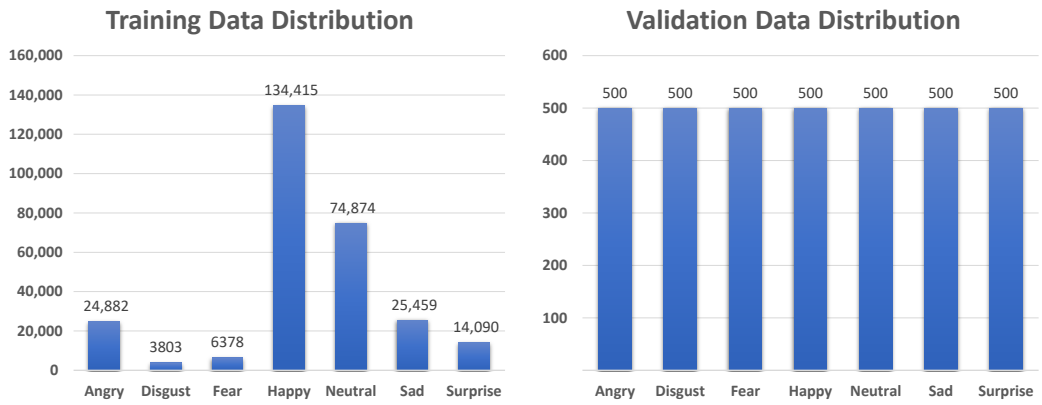

Figure 7. Data distribution for training and validation on AffectNet dataset [16].

The RAF-DB dataset [15] consists of about 30,000 facial images in the basic and compound emotion groups which were taken under the in-the-wild conditions with illumination changes, uncontrolled poses, and occlusion. In this study, we chose 12,271 images for training and 3068 images for validation, all of which were from the basic emotion group. The data distributions of the training and validation sets are shown in Figure 8.

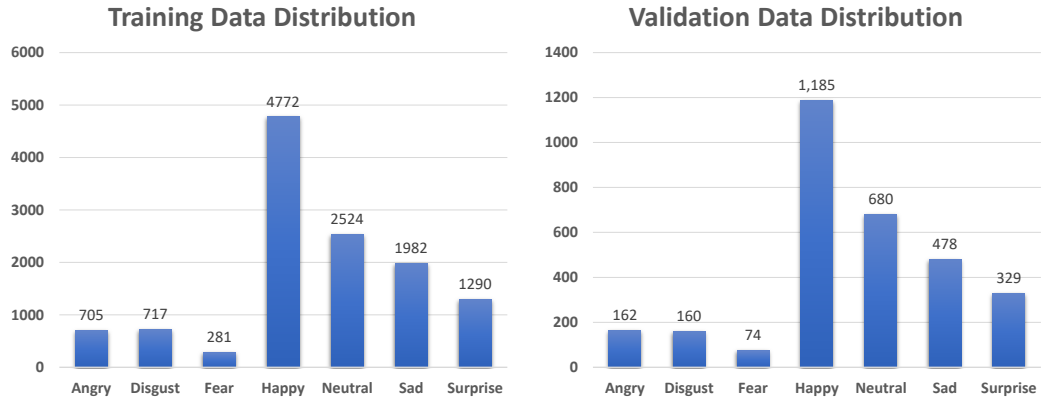

Figure 8. Data distribution for training and validation on RAF-DB dataset [15].

6.1.2. Video-Based Emotion Recognition in the wild

For facial emotion recognition in video clips, we used the AFEW dataset [33] to evaluate our study. The video clips in the dataset are collected from movies and TV shows under uncontrolled environments in terms of occlusion, illumination, and head poses. Each video clip was chosen based on its label, which contains emotion-related keywords corresponding to the emotion illustrated by the main subject. Use of this dataset helped us to address the problem of temporal facial expressions in the wild.

From the AFEW dataset, we used 773 video clips for training and 383 video clips for validation with labels corresponding to the seven basic emotion categories (anger, disgust, fear, happiness, neutrality, sadness, and surprise). The distribution of this dataset is shown in Figure 9.

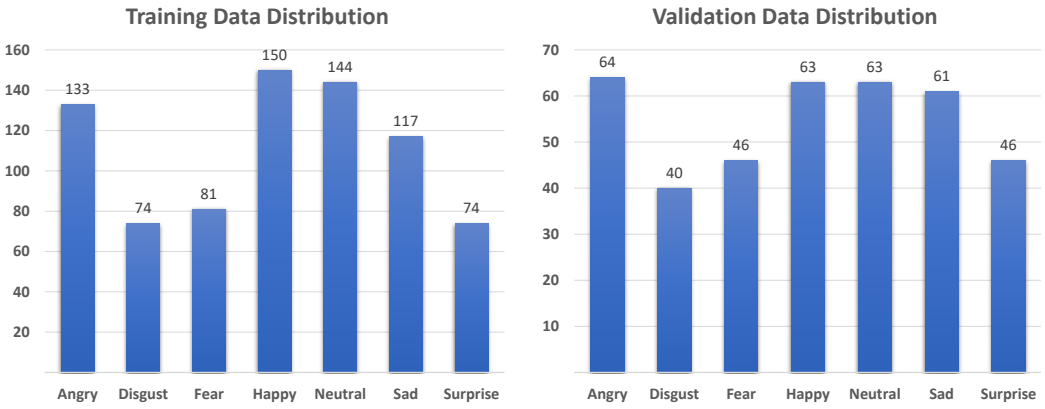

Figure 9. Data distribution for training and validation on AFEW dataset [33].

Table 1 shows the datasets used for in image and video emotion recognition in this study:

Table 1. Image and video emotion recognition datasets.

Emotion	AffectNet [16]		RAF-DB [15]		AFEW [33]	
	Training	Validation	Training	Validation	Training	Validation
Angry	24,882	500	705	162	133	64
Disgust	3803	500	717	160	74	40
Fear	6378	500	281	74	81	46
Happy	134,415	500	4772	1185	150	63
Neutral	74,874	500	2524	680	144	63
Sad	25,459	500	1982	478	117	61
Surprise	14,090	500	1290	329	74	46
Total	283,901	3500	12,271	3068	773	383

6.2. Environmental Setup, Evaluation Metrics, and Experimental Setup

Environment. We used Python 3.7 with Tensorflow 2.1 and Keras to develop our program. Our experiments were conducted on a Desktop PC with Intel Core I7 8700, 64 GB RAM and two Nvidia GeForce GTX 1080 Ti graphic cards with 11GB memory.

Evaluation Metrics. We used accuracy (*Acc.*) and F_1 score as the quantitative measurements in this study. We also used the average $Mean_{Acc.}$ and standard deviation $Std_{Acc.}$ of the accuracy values on the main diagonal of the normalized confusion matrix M_{norm} to evaluate the performance results, as in [15]. These metrics are calculated as follows:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + TN + FN} \\ F_1 &= 2\frac{Precision.Recall}{Precision + Recall} \\ Mean_{Acc.} &= \frac{\sum_{i=1}^{n} g_{i,i}}{n} \\ Std_{Acc.} &= \sqrt{\frac{\sum_{i=1}^{n} (g_{i,i} - Mean_{Acc.})^2}{n}} \end{aligned} \tag{34}$$

where $g_{i,i} \in diag(M_{norm})$ is the ith diagonal value of the normalized confusion matrix M_{norm}, n is the size of M_{norm}, and TP, TN, FP, and FN, respectively, are true positive, true negative, false positive, and false negative. The precision is the ratio of correctly predicted positive samples to all predicted positive samples. The recall is the ratio of correctly positive prediction to all true samples. They are calculated as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned} \tag{35}$$

The accuracy metric measures the ratio of correctly predicted samples to all samples; it ranges from 0 (worst) to 1 (best). It allows us to assess the performance of our model given that the data distribution is almost symmetric.

F_1 score can be used to more precisely evaluate the model in the case of an uneven class distribution, as it takes both FP and FN into account. F_1 score is a weighted average of precision and recall and ranges from 0 (worst) to 1 (best). In this study, due to the multi-class classification problem, we report the F_1 score as the weighted average F_1 score of each emotion label with weighting based on the number of labels.

Moreover, we also used $Mean_{Acc.}$ and $Std_{Acc.}$ to consider emotion evaluation under in-the-wild conditions with an imbalanced class distribution. This can be done in place of the accuracy metric, which is sensitive to bias under an uneven class distribution.

Experimental Setup. In this study, we conducted four experiments corresponding to: (1) the face and context feature extraction models; (2) the context spatiotemporal models; (3) the context temporal-pyramid model; and (4) the ensemble methods. Finally, we compared our results to related works on the AFEW dataset for video emotion recognition.

6.3. Experiments on Face and Context Feature Extraction Models

Overview. We used six conventional architectures to build a face feature extraction model to integrate into the facial emotion recognition models for video clips shown in Table 2. They consisted of Resnet 50 [18], Senet 50 [44], Densenet 201 [47], Nasnet mobile [46], Xception [45], and Inception Resnet [48]. Besides training from scratch, weights pre-trained on VGG-Face 2 [41], VGG-Face 1 [49], and ImageNet [50] were also used for transfer learning to leverage the knowledge from these huge facial and visual object datasets. For the context feature extraction model, we used the VGG16 model [17] with weights pre-trained on ImageNet [50] to extract the context feature around the person region.

Training Details. We first trained the models on the AffectNet dataset. We then fine-tuned the models on the RAF-DB dataset. Because the training and testing distributions differed, we applied a sampling technique to ensure that every emotion label in every batch had the same number of elements. Every image was resized to 224 × 224 and data augmentation was applied with random rotation, flip, center crop, and transition. The batch size was 8. The optimizer was Adam [52] with a learning rate of 0.001 and plateau reduction when

training on the Affect-Net dataset. For fine-tuning on RAF-DB, we used SGD [53] with a learning rate within the range of 0.0004 to 0.0001 using the cosine annealing schedule.

Results and Discussion. Table 2 shows the performance measurements of the face feature extraction models on the validation sets of the AffectNet and RAF-DB datasets.

As shown in Table 2, the performance results on AffectNet could be separated into three distinct groups, which are, in descending order: Group 1 (Inception Resnet, ResNet 50, and Senet 50), Group 2 (Densenet 201 and Nasnet mobile), and Group 3 (Xception). Group 1 had three metrics greater than 61% with the highest accuracy value of 62.51%, F_1 score of 62.41% and $Mean_{Acc.}$ of 62.51% for the Inception Resnet model.

Table 2. Performance of face feature extraction models on the AffectNet and RAF-DB validation sets.

No	Model	Pre-Train Weight	Affectnet [16]			RAF-DB [15]		
			Acc.	F_1	$Mean_{Acc} \pm Std$	*Acc.*	F_1	$Mean_{Acc} \pm Std$
1	ResNet 50 [18]	VGGFace2 [41]	61.57%	61.46%	61.57% ± 10.79%	**87.22** %	**87.38**%	**82.45**% ± **09.20**%
2	Senet 50 [44]	VGGFace 1 [49]	61.51%	61.50%	61.51% ± 10.40%	83.64%	83.81%	76.96% ± 11.12%
3	Nasnet mobile [46]	ImageNet [50]	59.20%	58.88%	59.20% ± 13.95%	80.74%	81.01%	74.05% ± 12.44%
4	Densenet 201 [47]	ImageNet [50]	59.31%	58.91%	59.31% ± 14.12%	83.08%	83.23%	76.94% ± 11.31%
5	Inception Resnet [48]	Scratch	62.51%	62.41%	62.51% ± 09.63%	81.23%	81.79%	77.08% ± 08.10%
6	Xception [45]	Scratch	56.26%	56.38%	56.26% ± 11.18%	80.90%	81.03%	74.71% ± 14.28%

After fine-tuning on the RAF-DB dataset using the weights from pre-training on the AffectNet dataset, the ResNet 50 model achieved the best performance, with the accuracy of 87.22%, F_1 score of 87.38%, and $Mean_{Acc.}$ of 82.45%. $Mean_{Acc.}$ was 82.44% greater than that of the DLP-CNN baseline in the RAF-DB dataset (74.20%) [15]. Therefore, we chose to use this model as the face feature extraction model for video emotion recognition.

Figure 10 shows the confusion matrix of the ResNet 50 model on the validation sets of the AffectNet and RAF-DB datasets. For the results of the ResNet 50 model on AffectNet, the happiness emotion label achieved the highest accuracy of 85%, while the remaining emotion labels showed similar accuracies, ranging from 53.6% to 63%. After fine-tuning in the RAF-DB dataset, the accuracy of the images labeled neutrality, sadness, surprise, and anger were significantly enhanced from 83.9% to 88.3%, nearly reaching the accuracy of 91.8% for the happiness label. The disgust and fear categories showed the lowest accuracy. In addition, the values of $Mean_{Acc.} \pm Std$ on AffectNet and RAF-DB were 61.57% ± 10.78%, and 82.44% ± 9.20%, respectively.

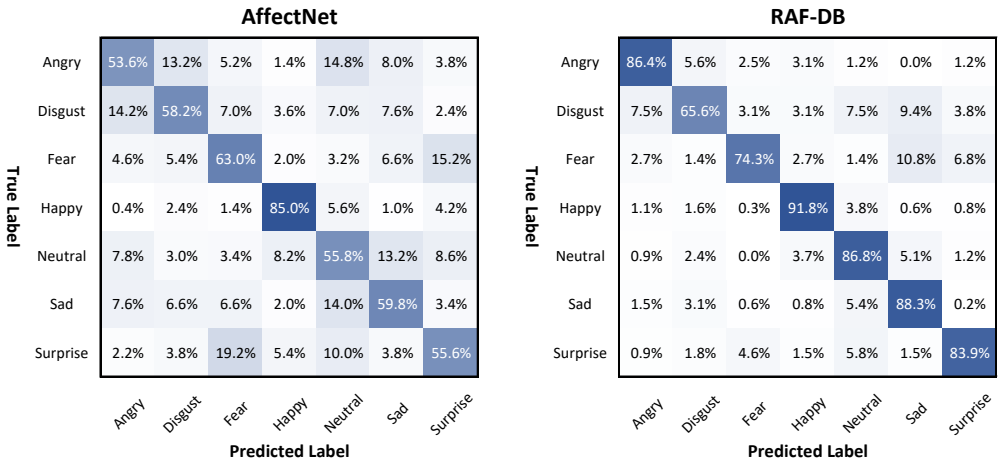

Figure 10. Confusion matrix of ResNet 50 model on the AffectNet and RAF-DB validation sets.

6.4. Experiments on Spatiotemporal Models

Overview. The spatiotemporal models consist of four blocks, namely feature extraction block, LSTM block, 3DCNN block, and classification block that receives input from the face and person sequences. In this experiment, we built three different models from the spatiotemporal approach, as shown in Table 3.

Model 1, "Spatiotemporal Model + Fix-Feature," used only the face sequence with the ResNet 50 face feature extraction model. The ResNet 50 model used weights that were pre-trained on the AffectNet and RAF-DB datasets, as discussed above. Moreover, all layers of the ResNet 50 model were frozen. Thus, the face feature extraction model was not fine-tuned during video-based emotion recognition training. Model 2, "Spatiotemporal Model + NonFix-Feature," was different from the first model in that only three blocks of the ResNet 50 model were frozen, and the feature block of the ResNet 50 model was fine-tuned. Model 3, "Spatiotemporal Model + NonFix-Feature + Context," expanded the context feature of Model 2 using input from both face and person sequences and used the pre-trained weights from the VGG16 model on ImageNet for context feature extraction.

Training Details. We trained our models on the AFEW dataset. During video batch sampling, every emotion label appeared with the same frequency to overcome the uneven class distribution and differences in distribution between the training and validation sets. We randomly extracted 32 frames per video clip in the training phase. For the validation phase, we averaged five predictions per clip by randomly extracting 32 frames. For data augmentation, we transformed the whole face and person sequence by resizing to 224 × 224, applying random horizontal flip, spatial rotation ±15°, and scaling ±20%. Training was done using SGD optimizer with early stopping at 40 epochs, an initial learning rate of 0.0004, and a reduction in the learning rate on the plateau.

Result Discussion. Table 3 illustrates the performance results of the spatiotemporal models on the validation set of the AFEW dataset.

Table 3. Performance results of the spatiotemporal models on the AFEW validation set.

No	Method	Context	Feature	*Acc.*	F_1	$Mean_{Acc.} \pm Std$
1	Spatiotemporal Model + Fix-Feature		Fix	51.70%	46.17%	46.51% ± 34.38%
2	Spatiotemporal Model + Nonfix-Feature		Nonfix	52.22%	48.26%	47.33% ± 31.73%
3	Spatiotemporal Model + Nonfix-Feature + Context	✓	Nonfix	**54.05**%	**50.78**%	**48.98**% ± **32.28**%

Model 1, with fixed face features due to frozen face feature extraction, obtained an accuracy of 51.70%, F_1 score of 54.17%, and $Mean_{Acc.}$ of 46.51%. Through fine-tuning on the feature block of the ResNet 50 model, Model 2showed an enhancement of accuracy by 0.52%, F_1 score by 2.09%, and $Mean_{Acc.}$ by 0.82%. Due to use of the context with the person region, Model 3 showed significant increases of 1.82%, 2.52%, and 1.65% for the accuracy, F_1 score, and $Mean_{Acc.}$, respectively. Model 3 also showed the highest accuracy of 54.05%, F_1 score of 50.78%, and $Mean_{Acc.} \pm Std$ of 48.98% ± 32.28% among all the spatiotemporal models.

Figure 11 shows the confusion matrix among the three models using the spatiotemporal approach. By fine-tuning the feature block of the face feature extraction model, Model 2 obtained an accuracy of 73% in the neutrality emotion label, compared to 58.7% for Model 1. Furthermore, Model 3, which took context into account, showed an enhancement of the accuracy of the sadness and surprise emotion labels, with accuracies of 62.3%, and 32.6%, respectively. These figures represent increases of 13.1% and 17.2% for the two emotion labels compared to the second approach. Moreover, Model 3 showed $Mean_{Acc.} \pm Std$ of 48.98% ± 32.28%, which is greater than the 47.33% ± 31.73% of Model 2, and 46.51% ± 34.38% of Model 1.

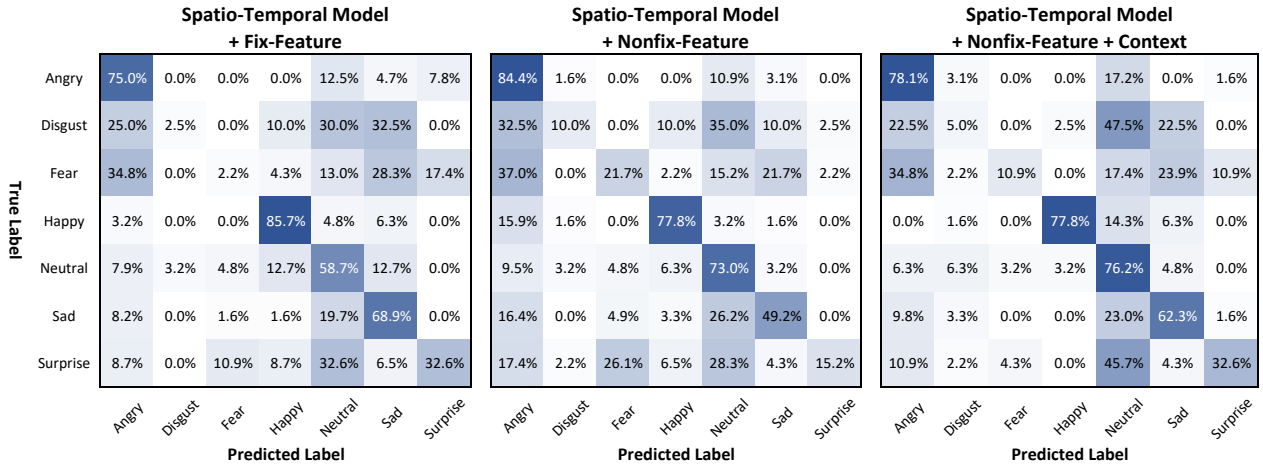

Figure 11. Confusion matrix of Models 1–3 (spatiotemporal approach) on the AFEW validation set.

6.5. Experiments on Temporal-Pyramid Models

Overview. For the temporal-pyramid model, we performed an ablation study on the context and scale factors, as shown in Table 4. For the context factor, Models 4–6 without context used only the ResNet 50 face feature extraction model, while Models 7–9 with context combined the face and context features from face and person sequences. When shown a face frame, a model without context produced one vector with a length of 2048 for the face feature and 21 probability outputs corresponding to the seven emotion labels and three statistical operators (min, mean, and max). The context feature vector from the VGG 16 model using pre-trained weights form ImageNet had a length of 2048. Therefore, the models without/with context had lengths of 2069/4117 per frame.

For the level factor, we conducted experiments on three groups of levels, {3}, {4}, and {0,1,2,3}. At level k, all processing frames are divided into 2^k sub-sequences and all sub-sequences are combined in the same interval by the mean operator in the face and context features and three operators (min, mean, and max) in the emotion probability outputs. For example, for level group {0, 1, 2, 3}, we divided all face and context frames in a video clip into 1, 2, 4, and 8 sub-sequences at Levels 0, 1, 2, and 3, respectively. In total, 15 sub-sequences were used to capture the emotion based on statistical information from whole frames or small chunks of frames with various lengths. Therefore, the length of the temporal-pyramid features without and with context is 15 * 2069 = 31,035 and 15 * 4117 = 61,755, respectively.

Training Details. In the training phase, we created temporal-pyramid features at level groups {3}, {4}, and {0,1,2,3} with and without context using the face feature model and context feature model with pre-trained weights in the Resnet 50 model from AffectNet and RAF-DB, and pre-trained weights for the VGG-16 model from ImageNet. For every level group, we used data augmentation to process 10 instances in every video clip. Data augmentation was applied to all frames with the same transformations: resizing to 224 × 224, random horizontal flip, scaling, and rotation. When sampling to get a minibatch, we randomly chose eight video clips with one of ten instances in data augmentation for every video clip, where the results satisfied the balance between emotion labels in a minibatch. We used the same training configuration as used in the training phase of the spatiotemporal models with the SGD optimizer, an initial learning rate 0.0004, and learning rate reduction on the plateau.

Results and Discussion. Table 4 depicts the experimental results of the temporal-pyramid models with adjustment of context and level factors.

Table 4. Performance results of the temporal-pyramid models on the AFEW validation set.

No	Method	Context	Level	*Acc.*	F_1	$Mean_{Acc.} \pm Std$
4	Temporal-Pyramid Model + Level 3		3	55.87%	52.76%	51.21% ± 29.87%
5	Temporal-Pyramid Model + Level 4		4	55.87%	52.51%	51.23% ± 30.12%
6	Temporal-Pyramid Model + Level 0,1,2,3		0,1,2,3	**55.87**%	**54.06**%	**51.85**% ± **25.98**%
7	Temporal-Pyramid Model + Level 3 + Context	✓	3	56.14%	54.61%	52.35% ± 25.53%
8	Temporal-Pyramid Model + Level 4 + Context	✓	4	56.40%	53.99%	52.18% ± 27.47%
9	Temporal-Pyramid Model + Level 0,1,2,3 + Context	✓	0,1,2,3	**56.66**%	**56.50**%	**54.25**% ± **16.63**%

For the level factor, Models 4–6, respectively, were set to level groups {3}, {4}, and {0,1,2,3} without context. The performance results of the three models were the same, with an accuracy of 55.87%. However, Model 6, with many level factors, gave better results in terms of F_1 score and $Mean_{Acc.}$ (54.06% and 51.85%, respectively, compared to 52.76% and 51.21% and 52.51%, and 51.23% for Models 4 and 5, respectively). Similarly, Model 9, using many level factors, also showed an F_1 score and $Mean_{Acc.}$ of 56.50% and 54.25%, which were superior to the results of Models 7 and 8. Therefore, the level factor affected the F_1 score and $Mean_{Acc.}$.

For the context factor, Models 7–9. respectively, increased accuracy, F_1 score, and $Mean_{Acc.}$ by 0.26%, 1.86%, and 1.15%; 0.52%, 1.49%, and 0.94%; and 0.78%, 2.44%, and 2.41% over the corresponding values of Models 4–6. In the same level group, the context factors helped Models 7–9 provide better results than Models 4–6, respectively. Moreover, Model 9, with many level factors, showed a significant increase in F_1 score and $Mean_{Acc.}$, as it had the highest values of 56.50% and 54.25%, respectively.

Figure 12 shows the confusion matrices of Models 6, 9, and 8. For the same level group {0,1,2,3}, Model 9, with context, showed an enhancement in the accuracy of the difficult emotion labels, disgust, fear, and surprise, by 30.0%, 43.5%, and 43.5%, respectively, compared to 20.0%, 32.6%, and 23.9% for Model 6. The $Mean_{Acc.} \pm Std$ of Model 9 was 54.25% ± 16.63%, which is greater than the 51.84% ± 25.98% of Model 6 (without context) and 52.18% ± 27.47% of Model 8 (with only one level {4}).

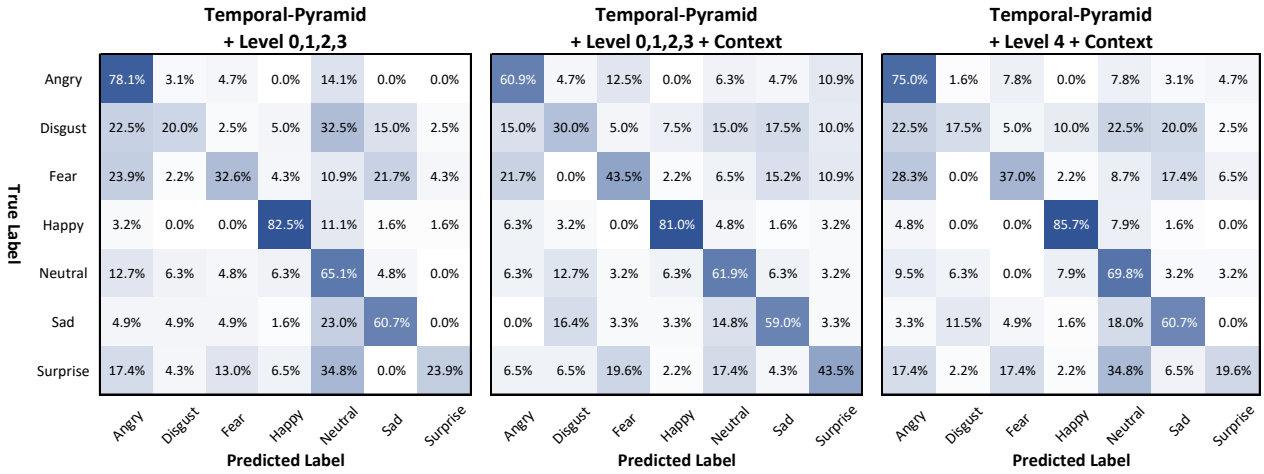

Figure 12. Confusion matrices of Models 6, 9, and 8 (from left to right), which use the temporal-pyramid approach on the AFEW validation set.

6.6. Experiments on Best Selection Ensemble

Overview. We conducted ensemble experiments through three approaches to exploit the complementary nature and redundancy among the models, as shown in Table 5. We first used the average fusion method, which combines the seven emotion probability outputs of all models with an average operator. The second approach was the multi-modal joint late-fusion method [10]. In this approach, we divided all models into two groups,

spatiotemporal (Models 1–3) and temporal-pyramid (Models 4–9) groups. This method used the average operator to merge all probability outputs of the emotion models in the same group, called the probability-merged layer, followed by a dense layer, and a softmax layer for classification into the seven emotion categories. The role of each group's outputs guarantees the accuracy of each branch. In addition, the model had a joint branch to merge the probability-merged layers of the two groups with a concatenation operator to give the emotion outputs.

The last approach was the best selection ensemble method. It chooses one of the models as the first element and then repeats the process by adding one of the remaining models using the average operator on the probability outputs with the previous models to help current combination increase. The process ends when there are no additional unused models to help increase the accuracy of the model combination or all models are selected.

Results and Discussion. The results of our experiments on the average fusion, multi-model joint late fusion, and best selection ensembles are shown in Table 5.

Table 5. Validation results of the ensemble experiments on the AFEW validation set.

No	Method	*Acc.*	F_1	$Mean_{Acc.} \pm Std$
10	Average Fusion	57.70%	55.00%	53.18% ± 29.62%
11	Multi-modal Joint Late-Fusion [10]	58.49%	57.40%	54.92% ± 23.50%
12	Best Selection Ensemble	**59.79%**	**58.48**%	**56.24**% ± **23.26**%

The best selection method showed the highest accuracy and F_1 score of 59.79% and 58.48%, respectively, representing significant increase in accuracy and F_1 score of 2.09% and 3.48% and 1.3% and 1.08% compared to the average fusion method and multi-modal joint late-fusion method, respectively. The combination models in the best selection method that gave the best scores were Models 3, 6, 7, and 9.

The confusion matrix in the best selection method shown in Figure 13 gave the highest $Mean_{Acc}$ 56.24% with the smallest $Std_{Acc.}$ of 23.26% compared to the average fusion method and multi-modal join late-fusion method. Moreover, this method showed an improvement in performance for the more difficult emotion labels: disgust, 25.0%; fear, 39.1%; and sadness, 37.0%.

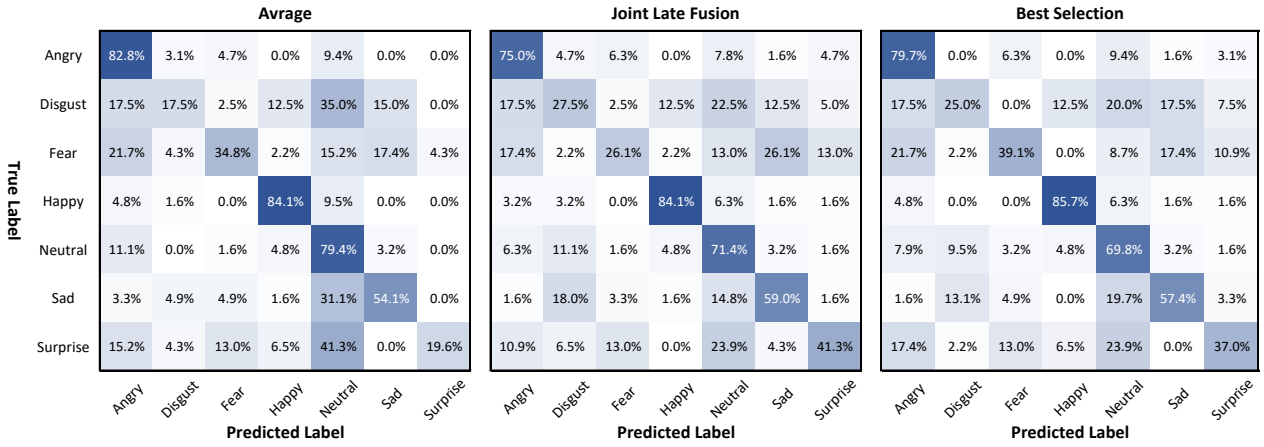

Figure 13. Confusion matrices of the average fusion method, multi-modal join late-fusion method and best selection ensemble method on the AFEW validation set.

6.7. Discussion and Comparison with Related Works

Discussion. Figure 14 presents the results of the three experiments on the AFEW validation set. First, the context factor played an important role in enhancing the performance of spatiotemporal Model 3 compared to Models 1 and 2 using the same approach, as well as temporal-pyramid Models 7–9 compared to the corresponding Models 4–6. This finding

confirms that context is key to interpretation facial expression to access the emotional state of a person [54], especially, in cases in which the facial region is small and blurry.

Second, use of multi-level factors {0,1,2,3} in temporal-pyramid models provided more robust features than were seen in the models using only a single level ({3} and {4}). For instance, Model 6 gave better results than Models 4 and 5. Similarly, the performance of Model 9 was better than that of Models 7 and 8. This shows that division of time periods in facial expression representation in a hierarchical structure creates robust features to capture human emotions under in-the-wild conditions, such as unclear temporal border and multiple apexes from spontaneous expressions.

Finally, when integrating multiple-modalities, the best selection ensemble method achieved better results than average fusion method, and multi-modal joint late-fusion method.

The main advantage of our ensemble method is that it allows the identification of the best combination of a large number of models through a multi-modal approach as well as derivation of instances from many training times. We were able to expand the average operator through use of other operators, such as skew, min, max, and median, as well as by combining many operators. In this study, the average and median operator were more useful than the others.

Comparison with related works. The accuracy measurements of our proposed methods and related methods on the AFEW validation set are shown in Table 6.

Table 6. Performance comparison with related studies on AFEW validation set.

Authors	Method	Approach	Modality	Year	Accuracy
Fan et al. [55]	CNN+LSTM	Spatiotemporal (2D+T)	Visual	2016	45.43%
	C3D	Spatiotemporal (3D)	Visual		39.69%
Yan et al. [56]	Trajectory Features + SVM	Geometry	Geometry	2016	37.37%
	CNN Features + Bi-directional RNN	Spatiotemporal (2D+T)	Visual		44.46%
	Fusion	Fusion	Visual+Geometry		49.22%
Vielzeuf et al. [57]	VGG-LSTM	Spatiotemporal (2D+T)	Visual	2017	48.60%
	LSTM C3D	Spatiotemporal (3D)	Visual		43.20%
	ModDrop Fusion	Fusion	Visual		52.20%
Hu et al. [58]	Face Features + Supervised Scoring Ensemble	Frame-Level	Visual	2017	44.67%
Knyazev et al. [28]	Face Features + STAT (min,std,mean) + SVM	Frame-Level	Visual	2017	53.00%
	Weighted Average Score	Fusion	Visual		55.10%
Kaya et al. [59]	CNN-FUN Features + Kernel ELMPLS	Spatiotemporal (3D)	Visual	2017	51.60%
Lu et al. [25]	VGG-Face + BLSTM	Spatiotemporal (2D+T)	Visual	2018	53.91%
	C3D	Spatiotemporal (3D)	Visual		39.36%
	Weighted Average Fusion	Fusion	Visual		56.05%
Liu et al. [26]	VGG16 FER2013 + LSTM	Spatiotemporal (2D+T)	Visual	2018	46.21%
	Face Features + STAT (min,std,mean) + SVM	Frame-Level	Visual		51.44%
	Landmark Euclidean Distance	Geometry	Geometry		39.95%
	Weighted Average Fusion	Fusion	Visual+Geometry		56.13%
Vielzeuf et al. [60]	Max Score Selection + Temporal Pooling	Frame-Level	Visual	2018	52.20%
Fan et al. [61]	Deeply-Supervised CNN (DSN)	Frame-Level	Visual	2018	48.04%
	Weighted Average Fusion	Fusion	Visual		57.43%
Duong et al. [62]	CNN Features + LSTM	Spatiotemporal (2D+T)	Visual	2019	49.30%
Li et al [63]	VGG-Face Features + Bi LSTM	Spatiotemporal (2D+T)	Visual	2019	53.91%
Meng et al. [64]	Frame Attention Networks (FAN)	Frame-Level	Visual	2019	51.18%
Lee et al. [65]	CAER-Net	Spatiotemporal (2D+T)	Visual	2019	51.68%
Kumar et al. [66]	Noisy Student Training + Multi-level attention	Frame-Level	Visual	2020	55.17%
Our method	**Spatiotemporal model**	Spatiotemporal (2D+T)	Visual		**54.05**%
	Temporal-pyramid model	Frame-Level	Visual		**56.66**%
	Best Selection Ensemble	Fusion	Visual		**59.79**%

Our spatiotemporal method outperforms other recently reported methods using the same approach, by around 0.14% compared with Li et al. [63]. Recently, Kumar et al. [66] used multi-level attention with an unsupervised approach by iterative training between student and teacher models. Their method showed a highest accuracy of 55.17%, which is lower than that of our temporal-pyramid method, 56.66%. To compare the fusion and ensemble methods, we searched for related studies that used multiple-modalities using visual and geometric information of facial expressions. Our ensemble method achieved the highest accuracy of 59.79%, which is better than that shown in related studies, where the highest reported accuracy was 57.43% by Fan et al. [61].

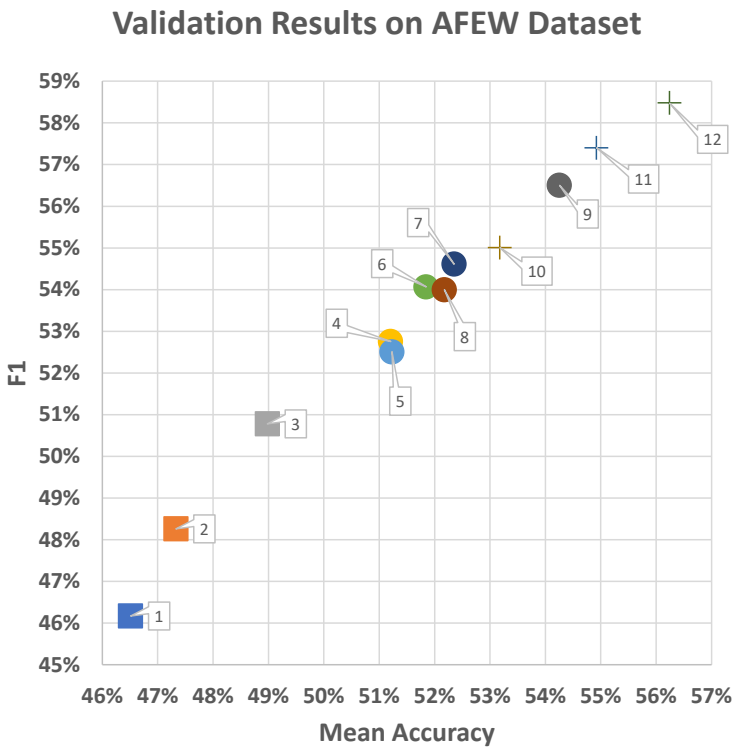

Figure 14. Results of our proposed models on the AFEW validation set. The rectangle data points represent spatiotemporal approaches, with context (Model 3) and without context (Models 1 and 2). The circular data points represent models based on temporal-pyramid approaches, which consist of two groups: without context (Models 4–6) and with context (Models 7–9) with level groups of {3}, {4}, {0,1,2,3}, respectively, in each group. Finally, the "plus sign" data points represent the average fusion method (Model 10), multi-modal joint late-fusion method (Model 11), and best selection method (Model 12).

7. Conclusions

In this study, we built an emotion recognition system to track the main face and recognize its facial expression in a video clip. We propose a face-person tracking and voting module to help our system detect the main face and person in a video clip for emotion recognition. Our tracking algorithm is based on a tracking-by-detection scheme with robust appearance observations to suggest facial and human regions, while the voting module uses relevant information about frequency of occurrences, size, and face detection probability to determine a main human and face sequence. In the next step, our emotion recognition models detects facial expressions through two main approaches, the spatiotemporal approach and the temporal-pyramid approach. Finally, the best selection ensemble method selects the best combination of models from among many training models to predict facial expression in a video clip. Compared to previous results on the AFEW dataset, our work shows improvement in every domain.

In the spatiotemporal models, we use 2D CNN facial and context blocks followed by an LSTM block and 3D CNN block to exploit the spatiotemporal coherence of facial and context features and facial emotion probabilities. The context factor is a significant key that increases the the performance of our model from 52.22% to 54.05%. Moreover, we achieved an accuracy that is better than that reported by related studies on the AFEW validation set.

For the temporal-pyramid models, we apply data augmentation on facial and context regions and extracted facial and person features and face emotion probabilities from every frame of the video clip. Using temporal-pyramid strategies, we created robust hierarchical features to feed into a simple neural network for classification of facial expression. Our method exploits the high correlation of features in the temporal domain. Due to the improvements mentioned above, we achieved an accuracy of 56.66% on the validation set, which is better than the accuracies of related studies using a single model with the same approach.

Finally, we propose a best selection ensemble to select a suitable combination of models from a large number of model instances during training with tuning of hyper-parameters, adjustment of levels, and configuration of context factor. Our ensemble method achieved an accuracy of 59.79%, which is better than that of the average fusion and multi-modal joint late-fusion method as well as related studies on the AFEW validation set.

In the further works, we will apply a multi-level attention mechanism to highlight the spatiotemporal correlations between emotion features over time. In addition, we use a graph convolution network to express movement of facial action units, which helps our system to better classify human expression.

Author Contributions: Conceptualization, N.-T.D. and S.-H.K.; Funding acquisition, S.-H.K., G.-S.L. and H.-J.Y.; Investigation, N.-T.D.; Methodology, N.-T.D.; Project administration, S.-H.K., G.-S.L., H.-J.Y. and S.Y.; Supervision, S.-H.K; Validation, N.-T.D.; Writing—original draft, N.-T.D.; and Writing—review and editing, N.-T.D. and S.-H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3A03000947).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Corneanu, C.A.; Simon, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [CrossRef]
2. Bänziger, T.; Grandjean, D.; Scherer, K.R. Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion* **2009**, *9*, 691–704. [CrossRef]
3. Martinez, B.; Valstar, M.F. Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition. In *Advances in Face Detection and Facial Image Analysis*; Springer International Publishing: Cham, Switzerland, 2016; pp. 63–100. [CrossRef]
4. Wieser, M.J.; Brosch, T. Faces in Context: A Review and Systematization of Contextual Influences on Affective Face Processing. *Front. Psychol.* **2012**, *3*. [CrossRef]
5. Koelstra, S.; Pantic, M.; Patras, I. A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1940–1954. [CrossRef] [PubMed]
6. Bernin, A.; Müller, L.; Ghose, S.; Grecos, C.; Wang, Q.; Jettke, R.; von Luck, K.; Vogt, F. Automatic Classification and Shift Detection of Facial Expressions in Event-Aware Smart Environments. In Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greece, 26–29 June 2018; pp. 194–201. [CrossRef]
7. Ekman, P.; Friesen, W.V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.

8. Kotsia, I.; Pitas, I. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Trans. Image Process.* **2007**, *16*, 172–187. [CrossRef] [PubMed]
9. Pantic, M.; Patras, I. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2006**, *36*, 433–449. [CrossRef]
10. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991. [CrossRef]
11. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
12. Liu, W.; Song, C.; Wang, Y.; Jia, L. Facial expression recognition based on Gabor features and sparse representation. In Proceedings of the 2012 12th International Conference on Control, Automation, Robotics and Vision, ICARCV 2012, Guangzhou, China, 5–7 December 2012; pp. 1402–1406. [CrossRef]
13. Dhall, A.; Asthana, A.; Goecke, R.; Gedeon, T. Emotion recognition using PHOG and LPQ features. In Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, Santa Barbara, CA, USA, 21–25 March 2011; pp. 878–883. [CrossRef]
14. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
15. Li, S.; Deng, W.; Du, J. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2584–2593. [CrossRef]
16. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
19. Ding, H.; Zhou, S.K.; Chellappa, R. FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017. [CrossRef]
20. Klaeser, A.; Marszalek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of the British Machine Vision Conference 2008, Leeds, UK, 1–4 September 2008; pp. 99.1–99.10. [CrossRef]
21. Zhao, G.; Pietikäinen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**. [CrossRef] [PubMed]
22. Sikka, K.; Wu, T.; Susskind, J.; Bartlett, M. Exploring bag of words architectures in the facial expression domain. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012. [CrossRef]
23. Jain, S.; Changbo Hu.; Aggarwal, J.K. Facial expression recognition with temporal modeling of shapes. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1642–1649. [CrossRef]
24. Wang, Z.; Wang, S.; Ji, Q. Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3422–3429. [CrossRef]
25. Lu, C.; Zheng, W.; Li, C.; Tang, C.; Liu, S.; Yan, S.; Zong, Y. Multiple Spatio-temporal Feature Learning for Video-based Emotion Recognition in the Wild. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; Volume III, pp. 646–652. [CrossRef]
26. Liu, C.; Tang, T.; Lv, K.; Wang, M. Multi-Feature Based Emotion Recognition for Video Clips. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 630–634. [CrossRef]
27. Kim, D.H.; Lee, M.K.; Choi, D.Y.; Song, B.C. Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction—ICMI 2017, Glasgow, UK, 13–17 November 2017; ACM Press: New York, New York, USA, 2017; pp. 529–535. [CrossRef]
28. Knyazev, B.; Shvetsov, R.; Efremova, N.; Kuharenko, A. Leveraging Large Face Recognition Data for Emotion Classification. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 692–696. [CrossRef]
29. Bargal, S.A.; Barsoum, E.; Ferrer, C.C.; Zhang, C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction—ICMI 2016, Tokyo Japan, 12–16 November 2016; ACM Press: New York, NY, USA, 2016; pp. 433–436. [CrossRef]
30. Zhu, X.; Ye, S.; Zhao, L.; Dai, Z. Hybrid attention cascade network for facial expression recognition. *Sensors* **2021**, *21*, 2003. [CrossRef]

31. Shi, J.; Liu, C.; Ishi, C.T.; Ishiguro, H. Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network. *Sensors* **2021**, *21*, 205. [CrossRef] [PubMed]
32. Anvarjon, T.; Mustaqeem; Kwon, S. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors* **2020**, *20*, 5212. [CrossRef] [PubMed]
33. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]
34. Dhall, A.; Roland Goecke, S.G.; Gedeon, T. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In Proceedings of the ACM International Conference on Mutimodal Interaction, Suzhou, China, 14–18 October 2019.
35. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* **2005**, *18*, 602–610. [CrossRef]
36. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; Volume 2015, pp. 4489–4497. [CrossRef]
37. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
38. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1409–1422. [CrossRef]
39. Kuhn, H.W. The Hungarian method for the assignment problem. In *50 Years of Integer Programming 1958–2008: From the Early Years to the State-of-the-Art*; Springer: Berlin/Heidelberg, Germany, 2010. [CrossRef]
40. Pérez, P.; Hue, C.; Vermaak, J.; Gangnet, M. Color-Based Probabilistic Tracking. In Proceedings of the European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; pp. 661–675. [CrossRef]
41. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG), Xi'an, China, 15–19 May 2018; pp. 67–74. [CrossRef]
42. Hu, P.; Ramanan, D. Finding tiny faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 951–959. [CrossRef]
43. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Leibe, B., Matas, J., Welling, M., Sebe, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37. [CrossRef]
44. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, 2011–2023. [CrossRef]
45. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
46. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning Transferable Architectures for Scalable Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8697–8710. [CrossRef]
47. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
48. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017, San Francisco, CA, USA, 4–9 February 2017.
49. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015; pp. 41.1–41.12. [CrossRef]
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**. doi:10.1016/j.protcy.2014.09.007. [CrossRef]
51. Rokach, L. Ensemble Methods for Classifiers. In *Data Mining and Knowledge Discovery Handbook*; Springer: New York, NY, USA, 2005; pp. 957–980. [CrossRef]
52. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
53. Schaul, T.; Zhang, S.; LeCun, Y. No more pesky learning rates. In Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013.
54. Barrett, L.F.; Mesquita, B.; Gendron, M. Context in emotion perception. *Curr. Dir. Psychol. Sci.* **2011**. [CrossRef]
55. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 445–450. [CrossRef]
56. Yan, J.; Zheng, W.; Cui, Z.; Tang, C.; Zhang, T.; Zong, Y.; Sun, N. Multi-clue fusion for emotion recognition in the wild. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; ACM: New York, NY, USA, 2016; pp. 458–463. [CrossRef]

57. Vielzeuf, V.; Pateux, S.; Jurie, F. Temporal multimodal fusion for video emotion classification in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 569–576. [CrossRef]
58. Hu, P.; Cai, D.; Wang, S.; Yao, A.; Chen, Y. Learning supervised scoring ensemble for emotion recognition in the wild. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 553–560. [CrossRef]
59. Kaya, H.; Gürpınar, F.; Salah, A.A. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image Vis. Comput.* **2017**, *65*, 66–75. [CrossRef]
60. Vielzeuf, V.; Kervadec, C.; Pateux, S.; Lechervy, A.; Jurie, F. An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 589–593. [CrossRef]
61. Fan, Y.; Lam, J.C.K.; Li, V.O.K. Video-based Emotion Recognition Using Deeply-Supervised Neural Networks. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 584–588. [CrossRef]
62. Nguyen, D.H.; Kim, S.; Lee, G.S.; Yang, H.J.; Na, I.S.; Kim, S.H. Facial Expression Recognition Using a Temporal Ensemble of Multi-level Convolutional Neural Networks. *IEEE Trans. Affect. Comput.* **2019**, *33*, 1. [CrossRef]
63. Li, S.; Zheng, W.; Zong, Y.; Lu, C.; Tang, C.; Jiang, X.; Liu, J.; Xia, W. Bi-modality Fusion for Emotion Recognition in the Wild. In Proceedings of the 2019 International Conference on Multimodal Interaction, Jiangsu, China, 14–18 October 2019; ACM: New York, NY, USA, 2019; pp. 589–594. [CrossRef]
64. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame Attention Networks for Facial Expression Recognition in Videos. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3866–3870. [CrossRef]
65. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-aware emotion recognition networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 10142–10151. [CrossRef]
66. Kumar, V.; Rao, S.; Yu, L. Noisy Student Training Using Body Language Dataset Improves Facial Expression Recognition. In *Computer Vision—ECCV 2020 Workshops*; Bartoli, A., Fusiello, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 756–773. [CrossRef]