

Article

Robust Human Face Emotion Classification Using Triplet-Loss-Based Deep CNN Features and SVM

Irfan Haider, Hyung-Jeong Yang , Guee-Sang Lee  and Soo-Hyung Kim *

Department of Artificial Intelligence Convergence, Chonnam National University,
Gwangju 500-757, Republic of Korea; irfan_haider99@hotmail.com (I.H.); hjyang@jnu.ac.kr (H.-J.Y.);
gslee@jnu.ac.kr (G.-S.L.)

* Correspondence: shkim@jnu.ac.kr

Abstract: Human facial emotion detection is one of the challenging tasks in computer vision. Owing to high inter-class variance, it is hard for machine learning models to predict facial emotions accurately. Moreover, a person with several facial emotions increases the diversity and complexity of classification problems. In this paper, we have proposed a novel and intelligent approach for the classification of human facial emotions. The proposed approach comprises customized ResNet18 by employing transfer learning with the integration of triplet loss function (TLF), followed by SVM classification model. Using deep features from a customized ResNet18 trained with triplet loss, the proposed pipeline consists of a face detector used to locate and refine the face bounding box and a classifier to identify the facial expression class of discovered faces. RetinaFace is used to extract the identified face areas from the source image, and a ResNet18 model is trained on cropped face images with triplet loss to retrieve those features. An SVM classifier is used to categorize the facial expression based on the acquired deep characteristics. In this paper, we have proposed a method that can achieve better performance than state-of-the-art (SoTA) methods on JAFFE and MMI datasets. The technique is based on the triplet loss function to generate deep input image features. The proposed method performed well on the JAFFE and MMI datasets with an accuracy of 98.44% and 99.02%, respectively, on seven emotions; meanwhile, the performance of the method needs to be fine-tuned for the FER2013 and AFFECTNET datasets.



Citation: Haider, I.; Yang, H.-J.; Lee, G.-S.; Kim, S.-H. Robust Human Face Emotion Classification Using Triplet-Loss-Based Deep CNN Features and SVM. *Sensors* **2023**, *23*, 4770. <https://doi.org/10.3390/s23104770>

Academic Editor: Wataru Sato

Received: 17 March 2023

Revised: 2 May 2023

Accepted: 7 May 2023

Published: 15 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: emotion classification; SVM; triplet loss; transfer learning; ResNet18

1. Introduction

Humans can communicate in a variety of ways, including through words, gestures and feelings. A comprehensive and accurate comprehension requires one's own sentiments and the hidden meanings it carried. Integration of these characteristics into machines that allow for a diverse and natural style of communication has become an attractive area of research in robotics, particularly in the realm of humanoid robots. Emotions are of several types, i.e., happy, sad, wonder, fear, guilt, confused and shocked. Mostly, six kinds of emotions are used. Emotions are detected for different purposes, i.e., healthcare apps, chat-bot apps and to make the system intelligent. EmotiChat [1] is a tool that is designed in this fashion. It is a chat program that can read your feelings. When the computer detects a positive emotion (a grin, for example), an emoticon is added to the conversation window. It may be useful in areas apart from human–computer interactions (HCI), such as surveillance or driver safety. Automatic systems can be made more responsive if they can gauge the driver's emotional state.

Recognizing a person's emotions based on their facial expressions is an integral part of interpersonal communication since they are an intuitive representation of a person's mental state and carry a wealth of emotional information. Zeng Guofan [2] was a well-known ancient Chinese thinker who developed techniques for emotion recognition from people's

faces. Today, researchers use various methods to detect facial expressions and to recognize emotions. Features can be extracted manually [3] or by an automatic detection method [1]. Action units (AU) are used to define facial expressions in the Facial Action Coding System (FACS). Raising one's inner brow is an example of an action unit. The emotions on someone's face are described by the simultaneous activation of many AUs [4]. To be able to make a claim regarding the degree of activation of the related emotions, it is helpful to be able to accurately recognize AUs. Facial landmarks can be hand-crafted, as in the work of Kotsia et al. [3]. It can be challenging to spot such features since their relative distances vary from person to person [5]. Besides emotions, AUs can also be used to identify facial textures. Filters can be used to identify the structural changes brought on by the expression of emotion on a person's face [5].

Face extraction from images is a crucial component of several approaches. The use of convolutional neural networks (CNNs) was validated for emotion recognition and feature extraction. Softmax [6] is often used to quantify the difference between the CNN's outputs and the supervised signal, and the triplet loss [7] and its variations were suggested to further impose intra-class compaction and inter-class distinctiveness of the learned features. We propose the quadruplet loss [8] as an enhancement of the triplet loss to further separate the classes. Adjusting the triplet loss's selection and margin specifications is a major focus for improving its functionality. During network optimization, the most difficult sample triplet is chosen to reduce intra-class distance and to increase inter-class distance. Nonetheless, the network is better able to learn from the most challenging data triplets due to these triplet losses. Using either intra-class distance or inter-class distance, the most difficult couples are selected as training samples by Song et al. [9]. convolutional neural network (CNN) models are very helpful in detecting emotions and feature extraction. CNN can achieve good performance in image recognition and attains higher recognition accuracy than conventional algorithms. CNN models also showed good results in detecting emotions by using EEG signals. In our study, we present primarily related material to the suggested method, as well as a few carefully chosen publications that provide an overall perspective on the various approaches.

However, current CNN models are not good enough. In our study, we customized ResNet18 to enhance the accuracy of emotion recognition. Our customized model performed well on two datasets, i.e., JAFFE and MMI. The main contribution of our model is to detect the emotions accurately and to customize ResNet18 to achieve better results than others.

2. Literature Review

Caleanu [10] and Bettadapura [11] provided an in-depth analysis of the field of facial expression recognition. Many detectors were used for recognizing emotion and for detecting faces from images or videos. Sarah et al. [12] used YOLO in their study for detecting faces from images. YOLO has many versions and is mostly used for the object detection task, and a face is also an object, which is it was used for emotion detection tasks. Liu et al. [13] used retina faces to detect faces from images, which he also used for aligning, analyzing and cropping the images into a size of 256×256 . In one study [14], faces were detected from video by using retina face detectors. Happy and Routray [5] distinguished between six fundamental emotions using the Extended Cohn–Kanade Dataset. Their method detected the facial area in an image and returned its coordinates. Facial features, such as the eyes or lips, are identified and designated from a specific area. Features are extracted from regions that show the largest contrast between two images. Firstly, reducing the features' dimensionality, we may then feed them into a support vector machine (SVM). A 10-fold cross-validation is used to test the efficacy of the procedure. The mean precision is 94.09%.

Byeon and Kwak [15] proposed using a video to detect emotional states. A 3D convolutional neural network (CNN) that takes input from sequences of five frames was constructed. An accuracy of 95% was achieved using a database of only 10 people. Song et al. [16] employed a deep convolutional neural network for face emotion recognition. A neural network

was built with a total of 65 thousand neurons for cascading five layers. An accuracy of 99.2 percent on the CKP set was accomplished with the help of convolutional, pooling, local filter, and one fully connected layer. Overfitting was prevented by employing the dropout technique.

The Extended Cohn–Kanade dataset was developed by Luecy et al. [17]. Annotations for both emotions and action units are included in this dataset. They also tested the datasets for classification accuracy using active appearance models (AAMs) in tandem with support vector machines (SVMs). They used AAM, which creates a mesh from the face, to pinpoint its location and to follow it across photos. They used this mesh to generate two feature vectors. To begin, the axes of rotation, translation, and scaling were applied to the vertices. Second, a grayscale image was derived from the mesh information and the input pictures. With the help of a cross-validation approach in which one participant is removed from the training phase, they were able to increase the accuracy to over 80%.

An emotion recognition system based on facial expressions was created by Anderson et al. [1]. There are three main parts to their system. The first is a tool developed for locating faces, called a face tracker (a ratio template derivative). The second part is a facial motion-tracking optical flow algorithm. The final piece is the recognition engine itself. It uses SVMs and multilayer perceptron (MLPs) as its foundational neural network models. EmotiChat takes this strategy and runs with it. Their rate of success in recognizing patterns is 81.82%.

Emotion recognition by using features extraction has some limitations due to complex facial structures. To address this issue, a unique feature fusion dual-channel expression recognition algorithm was proposed in this paper using principles from machine learning and philosophy.

In particular, the issue of subtle variations in facial expression is not taken into consideration by the feature generated by a convolutional neural network (CNN). The proposed algorithm's initial fork takes as input a Gabor feature extracted from the ROI region. To make the most of the operational facial expression emotion region, it is important to first segment the area from of the main facial images before applying the Gabor transform to extract the area's emotion properties. Specific attention must be given to describing the surrounding area. The alternative method recommends a channel attention network that uses depth separable convolution to improve linear bottleneck structure, reduce network complexity, and safeguard against overfitting by way of the creation of a well-designed attention module that considers both depth of the convolution layer and spatial information. For the FER2013 dataset, it surpasses the competition by a wide margin due to its increased focus on feature extraction, which in turn leads to better emotion recognition accuracy.

Kotsia and Pitas [3] developed a method for emotion detection in which a candidate grid, a face mask with a small number of polygons, is superimposed on a person's face. The grid is superimposed on the image in a completely random location and then must be moved to the subject's face by hand. As the feeling progresses, a Kanade–Lucas–Tomasi tracker keeps tabs on the grid's position. For multiclass SVMs, the grid's geometric displacement information serves as the feature vector. Those feelings include anger, repugnance, fear, joy, sorrow, and surprise. They ran tests on the Cohn–Kanade dataset and found that their model has a 99.7 percent success rate. Based on local binary patterns (LBPs), Shan et al. [18] developed a system for emotion recognition. Facial area LBP features are extracted from the input face image. AdaBoost is used to locate the image regions that are most informative for classification. After testing numerous classification methods, we found that SVMs using Boosted-LBP features had the highest recognition accuracy (95.1% on the CKP dataset).

Using robust normalized cross-correlation (NCC), Zafar et al. [19] suggested an emotion recognition system in 2013. The “Correlation as a Rescaled Variance of the Difference between Standardized Scores” normalized coefficient calculator (NCC) was utilized here. Pixels that strongly or weakly influence the template matching are deemed outliers and are

ignored. This method was evaluated using both the AR FaceDB (85% recognition accuracy) as well as the Extended Cohn–Kanade databases (100% recognition accuracy).

2.1. Face Detection

The numerous real-world applications of face recognition have attracted a lot of research and development attention. In the days before the deep convolutional neural network (deep CNN), face detection depended primarily on characteristics that were created by hand [20]. Researchers presented a plethora of strong, manually created features, including HAAR, histogram of oriented gradients (HOG), LBP, scale-invariant feature transform (SIFT), and aggregate channel features (ACF). However, deep CNN has greatly outpaced the performance of these feature extractors. The last several years have seen an expansion of models, with deep CNN proving particularly effective at a wide variety of target identification applications. We characterize the target identification tasks such as a pair of tasks, one involving the classification of target candidate regions and the other using their regression. Object detection networks come in a wide variety; some examples include the RCNN family [20], SSD [21], YOLO series [22], FPN [23], MMDetection [24], EfficientDet, transformer (DETR), and Centernet.

Even though there are unique difficulties in detecting faces, such as multiscale, small face, low light, dense scene, etc., inter-class [25–27] is also an issue in image classification that misclassifies the emotions because of having the same person’s face in different classes, as can be seen in the example of having the same face in Figures 1 and 2. These problems are the same to those faced when searching for any object as shown in Figure 3. For this reason, broad object detection networks can be used as a stepping stone for training face detection networks. Scale, position, occlusion, expression, and cosmetics are all areas that might cause issues on their own. Researchers have come up with a number of approaches to these problems, including the Cascade CNN, multi-task cascaded convolutional networks (MTCNN), and the single-stage headless face detector (SSH).

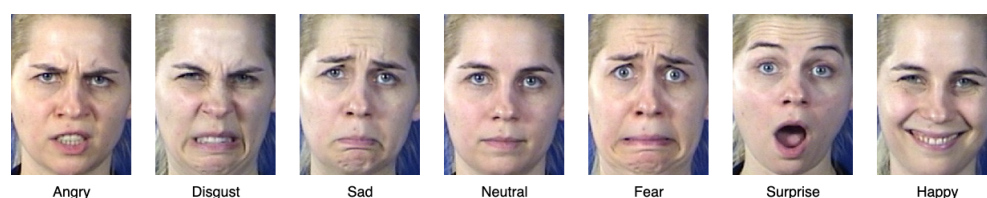
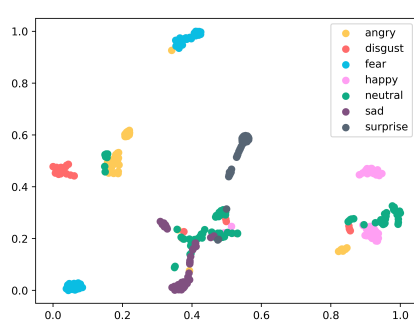


Figure 1. Inter-class sample.



(a) Scatterplot of t-SNE.



(b) t-SNE with image samples.

Figure 2. t-SNE visualization.

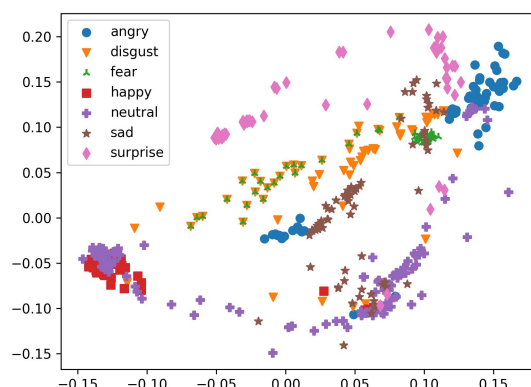


Figure 3. Scatterplot.

2.1.1. YOLO

The object detection community has contributed significantly to the evolution of the YOLO algorithm over the past five years, resulting in version 5. The creator of the YOLO algorithm proposed the first three versions, which are referred to as YOLOv1 [28], YOLOv2 [29], and YOLOv3 [24]. YOLOv3 is widely regarded as a major advancement in terms of performance and speed. The Softmax loss is replaced with the binary cross-entropy loss, and it identifies multiscale features (FPN) [23] and a superior backbone network (Darknet53). A separate research group unveiled YOLOv4 [30] in 2020's first few months. The group gave serious thought to a wide variety of YOLOv3 [24] variants, including the backbone along with the alleged bags of freebies and bags of specialties. A month later, a new research group unveiled YOLOv5 [31], which was substantially smaller in size, faster in speed [30], and fully implemented in Python (PyTorch). To this point, the field of object detection has greeted it with open arms.

2.1.2. Haar Cascade

To detect objects, the Haar classifier relies heavily on these Haar-like properties. These characteristics do not rely on pixel intensity but rather on the contrast shift between neighboring squares. Variations in brightness and darkness are identified based on the contrast differences between clusters of pixels. A Haar-like feature is formed by two or three neighboring groups that share a similar contrast variance. They use Haar-like features to detect images [32]. Raising or lowering the size of the pixel group that is being analyzed is a simple way to adjust the scale of Haar characteristics. This paves the way for characteristics to be utilized for detecting objects of varying sizes.

Haar classifier cascades must be trained before they can accurately detect human facial characteristics such as the mouth, eyes, and nose. Training classifiers requires the implementation of the mild AdaBoost algorithm and the Haar feature methods. Intel created Open Computer Vision Library, an open source library specifically designed to simplify the incorporation of programs linked to computer vision (OpenCV). Detection and training of Haar classifiers, such as implementation applications in HCI, robotics, authentication, computer vision, and other domains where visualization is essential, can make use of the OpenCV library. Ref. [32] is only one example. We need two separate photo collections to train the classifiers. Images and videos that do not feature the sought-after item form one set (a face, in this example). The term “negative images” is used to describe this group of pictures. The other group of pictures, the positive ones, show the object in use. Image name, top left pixel, height, and breadth of the object specify where the objects are located within the positive photos. Five thousand low-resolution photos (at least a megapixel) were utilized to train the face characteristics. Among these were pictures of paperclips and photographs of forests and mountains.

2.1.3. Retina Face

The cutting-edge RetinaFace facial detector for Python is based on deep learning and includes facial landmarks. When used in a crowded environment, its detection performance stands out as exceptional. To accomplish pixel-wise face localization, RetinaFace is a one-shot framework that employs three subtasks. Face detection, 3D reconstruction using a mesh decoder, and 2D alignment are the tasks at hand.

2.2. Loss Functions

2.2.1. Triplet Loss Function

For the objective of understanding the emotions expressed in a human voice, a neural embedded system that depends on triplet loss and residual learning was proposed by Kumar et al. [7]. The proposed approach uses the emotional tone of the speaker's words to learn the optimal embedding. The triplet loss function takes in an anchor image, a positive image, and a negative image as inputs. An already-labeled image serves as the "anchor", while an image with the same label as the anchor and an image with a different label serve as the "positive" and "negative", respectively. Discovering an embed of features for an image where the distance between the anchoring and the positive image is smaller than the distance seen between the anchor and the negative image is the goal of the triplet loss function. The triplet loss function can be written as:

$$L = \max(0, ||f(x_a) - f(x_p)||^2 - ||f(x_a) - f(x_n)||^2 + \text{margin})$$

where f is the feature embedding function, x_a is the anchor image, x_p is the positive image, and x_n is the negative image, $||\cdot||$ is the L_2 norm, and margin is a hyper-parameter that controls the margin between the positive and negative distances. Zhang et al. [8] provided identity loss in a supporting task to monitor the training process and to improve the durability of the elementary task without interrupting the entire training stage for facial expression recognition (FER) with deep metrics.

In order to bring images that have the same label closer to one another in feature space, the neural network is programmed to minimize the triplet loss function. After the network was trained, features embedding may be used to classify facial expressions by passing the image through it and by letting a straightforward classifier, such as a support vector machine classifier, make a prediction about the expression label. When using a triplet loss function, the distance among objects of various classes in the feature space is effectively increased [33]. To further improve the effectiveness of our method, we incorporated the triplet loss function.

2.2.2. Cross Entropy

Cross-entropy loss, also called log loss, is used to measure the efficiency of a classification process whose output is a probability value between 0 and 1. Cross-entropy loss increases as the difference between the expected probability and the actual label grows larger. The cross-entropy loss function and the loss function are both proposed in this method [34]. We will then evaluate the algorithm's performance based on the precision and recall of its detections. Simulation findings show that the proposed technique outperforms state-of-the-art classifiers such as CNN in terms of prediction accuracy.

2.3. Facial Emotion Classifiers

2.3.1. Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) employs a linear regression model for categorization and dimensionality reduction [35]. Features extracted from pattern categorization issues are its primary use. After its initial development by Fisher in 1936 for two classes, the linear discriminant was later generalized by C.R. Rao for many classes in 1948. By utilizing LDA to project data from a D -dimensional feature space onto a D' ($D > D'$) dimensional space, we may maximize variation across classes while decreasing variability within classes.

Logistic regression, one of the most popular linear classification models, does well for binary classification but poorly for multiple classification problems with clearly distinguishable classes. LDA can be used in data preprocessing in the same way that principal component analysis (PCA) can be used to minimize the number of characteristics and hence the computing expenses. Face recognition algorithms also make use of LDA. Similarly, LDA is used in face detection methods.

2.3.2. Support Vector Machine (SVM)

Invention of SVM can be credited to Vapnik et al. [36], with subsequent modifications by Corinna Cortes and Vapnik [10]. SVM is employed in data analysis and pattern recognition. Yag et al. mention [37] utilized support vector machines to categorize plant diseases. Pattern regression analysis and classifications are common applications for SVM, which is a supervised learning algorithm that infers a function or relationships from the given training data. In order to efficiently assign a new example point to one of two classes, an SVM algorithm requires a training set of examples from each class.

2.3.3. Softmax Regression

Assumptions for every class label are provided by Softmax classifiers, while the margin is provided by the hinge loss metric. As human beings, we can more easily grasp probabilities than range scores (such as in hinge loss and squared hinge loss). Neural network models that forecast a multinomial probability density function use the Softmax function as the input signal in the output layer. In other words, when determining class membership on more than two class labels, Softmax is utilized as the activation function.

3. Proposed Methodology

In this section, we first describe the overall proposed pipeline with a face detector to detect and tighten the face bounding box and a classifier to obtain the facial expression class of detected faces using deep features from a ResNet18 trained by using triplet loss. We use RetinaFace to detect and cut out the detected face regions from the original image for each input. Then, we train a ResNet18 model using cropped face images with triplet loss and use this model as a feature extractor. Afterward, deep features generated from the feature extractor are fed into an SVM classifier to produce the facial expression class of the face. A frame extractor was developed to process video input to extract the twenty representative images. The proposed pipeline is defined in Figure 4.

3.1. Frame Extractor

The proposed pipeline's encoder is designed for image input only. To eliminate this weakness and to make the pipeline work for image input and video, we constructed a frame extractor to extract the most representative frames of each video by considering the problem as anomaly detection on a time series.

Assume that we have a n – frame video input with resolution of $W \times H \times C$, which can be represented as $X = \{X_t\}^n$ for $X_t \in \mathbb{R}^{W \times H \times C}$. We aimed to construct a 1 – D time series $T = \{T_t\}^n$ for $T_t \in \mathbb{R}$ from this input and to detect the outliers in the series. First, we applied a Gaussian filter $G(x, y) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2+y^2}{2\sigma^2})$ on each frame $X'_t = G * X_t$ with $*$ as the convolution operator to remove the effect of noise between frames. Then, we applied a simple sum operator over X_t dimensions to map between T_t and X_t by $T_t = \sum X_t$. After obtaining the T series, we removed the first and last 20 values because they regularly represent a neutral facial expression. Then, the differencing method was applied to compute the difference between frames with a set of lag configs. With m – lag config, we obtained the new T' series with $T'_t = T_t - T_{t-m}$. In the last step, the outliers were detected using the mean absolute deviation method [38]. Across lag configs, we removed the duplicated frames and obtained the twenty unique and most representative frames. An example is shown in Figure 5.

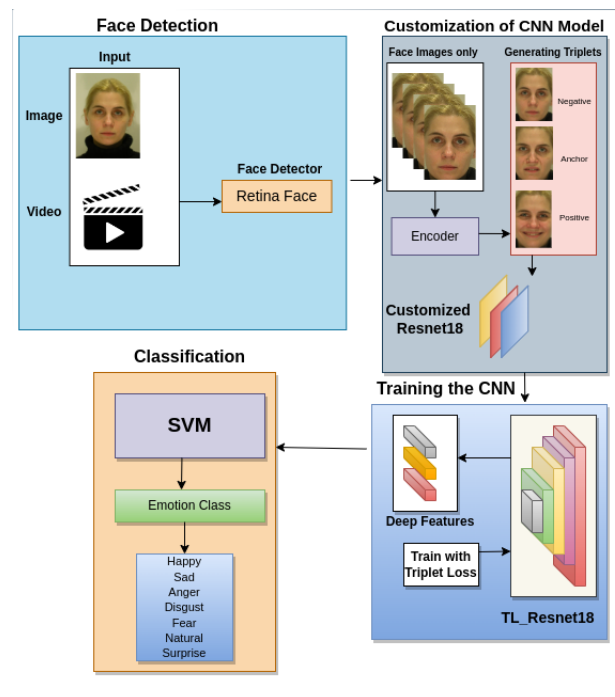


Figure 4. The proposed TLF-ResNet18 pipeline for facial expression recognition with a face detector to detect and tighten the face bounding box and an SVM classifier to obtain the facial expression class of detected faces using deep features from trained ResNet18.



Figure 5. The twenty unique and most representative frames extracted from an angry expression video of MMI database.

3.2. Face Detector

We only need the human face image regions to know which facial emotion is expressed. Images with complex backgrounds and small faces can affect the performance of the facial expression recognition pipeline. We decided to plug in a face detector to detect only face regions in each image and to remove unrelated regions that can harm our pipeline performance.

We surveyed many existing face detectors, such as RetinaFace [39] and YoloFace [40], to obtain the best face detector for our pipeline. All face detectors were compared using the WiderFace dataset [41]. We finally chose RetinaFace because it has the best performance based on the WiderFace dataset.

With the RetinaFace model, we cropped face regions on each candidate image, obtained only the largest detected face, and kept it for the latter pipeline component input, as shown in Figure 4. The face detector pipeline can be seen in Figure 6.

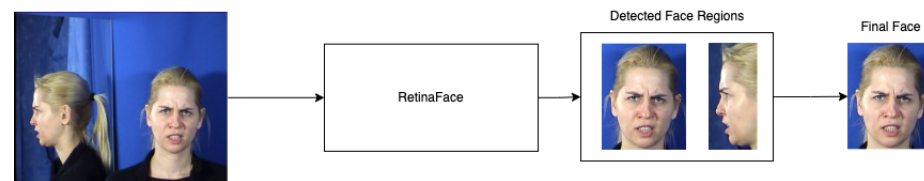


Figure 6. The face detector pipeline with multiple faces in the input image and output as the largest face only.

3.3. Deep Feature Extractor

After receiving a proper dataset with cropped faces, we used a convolutional neural network [42] to extract the deep features of each image. A CNN model can contain many types of layers, such as convolution, activation, and pooling layers. Each layer extracts higher features than the previous layer [39]. In order to obtain the best result, we applied the advanced CNN architect proposed in [43]. With ResNet [43], we can stack more convolution layers to the model without worrying about the vanishing gradient problem, and the model can achieve higher performance.

After the last pooling layer of the CNN model, we added a fully connected layer that outputs a 128 – D embedding vector or deep features for each input. Then, this embedding vector is normalized to a unit sphere as in [44] before being fed into any loss functions or classifiers. Following the FaceNet paper [44], we used the triplet loss function to optimize the extracted embeddings directly. We adopted the online triplet mining technique introduced in [45]. This technique enables us to generate triplets based on data batches in each learning iteration. The whole training diagram is shown in Figure 7.

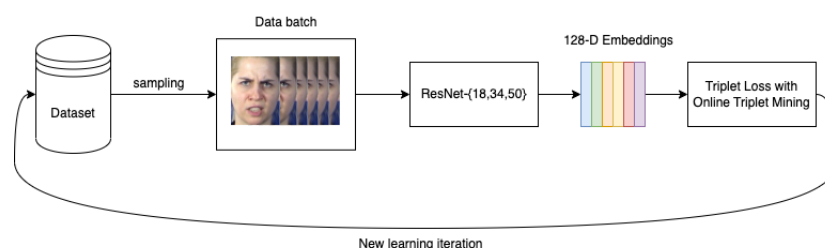


Figure 7. Training diagram with online triplet mining technique from [9].

3.4. Facial Expression Classification

We also used many different classifiers (SVM, LDA, Softmax Regression) to obtain the facial expression class. At the end of each training iteration, the produced embeddings with labels of their original images were used to fit classifiers. Once trained, these classifiers were evaluated on the validation and test set. In this work, we used scikit-learn’s implementation of three classifiers.

4. Results and Evaluation

4.1. Datasets

JAFFE [46]: This database was introduced by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba [46] and has a total of 213 images. In this datasets, the number of subjects is 10, and the images are grayscale in the size of 256×256 . The photos were annotated, with average semantic scores for each facial emotion, by 60 annotators after each subject performed seven facial expressions (six basic and one neutral).

FER2013 [47]: This dataset contains images and categories describing the emotion of the person in it. The total number of grayscale images in the size of 48×48 in the dataset is 35,953 images in seven different facial expression classes. The dataset has three subsets for training, validation, and testing. While analyzing the dataset, we found that the dataset has a significant data imbalance problem due to the “disgust” class only having a few samples compared to the others. FER2013 distribution is depicted in Figure 8.

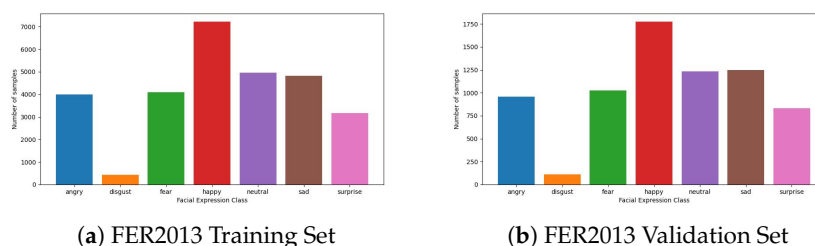


Figure 8. Dataset distribution of the training and validation set of FER2013.

AFFECTNET [48]: This dataset includes samples of various sizes and excellent images in RGB or grayscale-based color. There are eight distinct classes (surprise, angry, sad, contempt, disgust, fear, neutral, and happy). The validation and training sets are divided as of the FER2013. Similar to the FER2013 dataset, AFFECTNET also suffers from the data imbalance problem, but this is due to the dominance of the number of the “happy” class. AFFECTNET distribution is depicted in Figure 9.

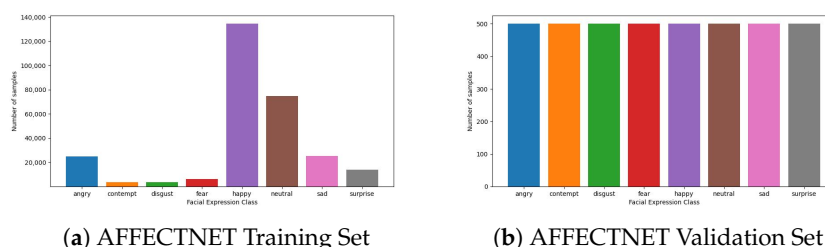


Figure 9. Dataset distribution of training and validation set of AFFECTNET.

MMI [49]: This is a database containing 236 videos of people showing emotions with annotations. Twenty frames from each video that best capture its content were automatically extracted by the frame extractor discussed in Section 3. A total of 4756 photos of frontal views of the human face was retrieved.

4.2. Experiment Design

In order to assess the effectiveness of the proposed pipeline, we evaluated its performance on the JAFFE, FER2013, AFFECTNET, and MMI datasets. For better classification, we obtained optimized deep features at the “avg_pool” layer of the ResNet18 model and fed it into SVM for facial emotion prediction. The proposed approach was validated on the four standard datasets, JAFFE, FER2013, Affectnet and MMI. The developed TLF-ResNet18 model outperformed on the JAFFE and MMI datasets and offered comparable results on the FER2013 dataset. The proposed model outperformed the state-of-the-art approaches, which show its effectiveness. The developed pipeline is flexible and can be extended to address more facial expressions. In all cases, we measured the benefit of the proposed method compared to other previous results. As seen below, our pipeline outperforms many current SoTA results.

All experiments were set up using ResNet18 as the feature extractor to generate the unified embeddings from original human face images. We applied basic data augmentations to improve generalization performance, such as horizontal flip, vertical flip, random rotation, and random crops. All images were normalized to grayscale-based images. We found that 48×48 is the optimal image size when considering both speed and accuracy for facial expression recognition; increasing the image size does not increase the accuracy much. We also implemented a weighted data sampling method that will assign a smaller weight to all samples in a class with a larger total number of samples. In each learning iteration, the data loader prioritized creating batches of data on samples that have a larger weight. This method greatly improved the imbalance datasets such as FER2013 and Affectnet.

We used Adam [50] as the optimizer to optimize the loss value with a learning rate of 0.001, which is reduced by a factor of 0.9 if there is no improvement in validation accuracy in the last five epochs. We ran the training for 100 epochs and evaluated the learned embeddings at the end of each learning iteration by fitting the SVM, LDA, and Softmax regression classifiers. All experimental results can be seen in Figure 10. We present and discuss in detail each case in the next subsections.

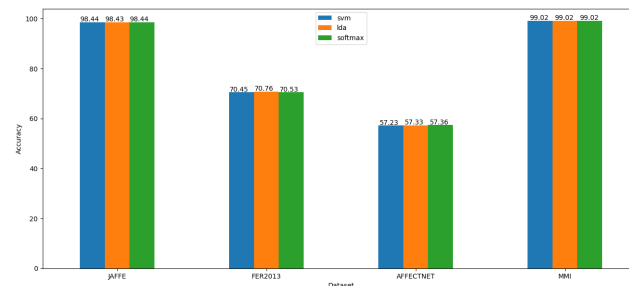


Figure 10. The best accuracy of each classifier when trained and evaluated on embeddings learned on each dataset.

4.3. Small and Balanced Datasets: JAFFE, and MMI

JAFFE and MMI are small datasets for facial expression recognition and do not have a separate evaluation dataset. We split these datasets randomly with a ratio 7/3 of training and validation sets to evaluate the unified embeddings on these datasets. Specifically, the JAFFE dataset that has a total of 214 samples in seven emotion classes is split to have 149 samples in the training set and 65 samples in the validation set; finally, the MMI dataset is split to have 3329 samples in the training set and 1427 in the validation set.

The training curve of these three datasets can be seen in Figures 11 and 12. The results of classifiers have no difference and demonstrate SoTA performances. They also show that our method is robust and has achieved the ultimate performance that could not be better. The best accuracy of the proposed method and previous works on the JAFFE and MMI dataset are shown in Tables 1 and 2.

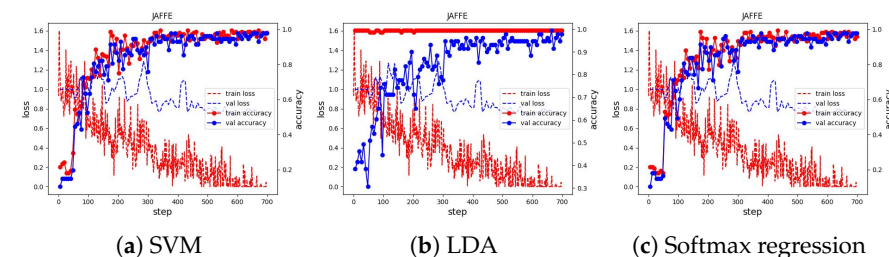


Figure 11. Training curve of the three classifiers on the JAFFE dataset.

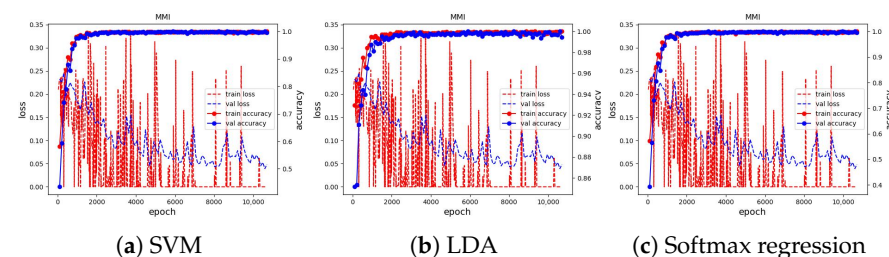


Figure 12. Training curve of three classifiers on MMI dataset.

Table 1. Performance comparison between our proposed pipeline with previous works on the JAFFE dataset.

	JAFFE
Proposed method	98.44%
Reference [51]	92.8%
Reference [52]	73.28%
Reference [53]	94.83%
Reference [54]	96.30%
Reference [55]	97.46%

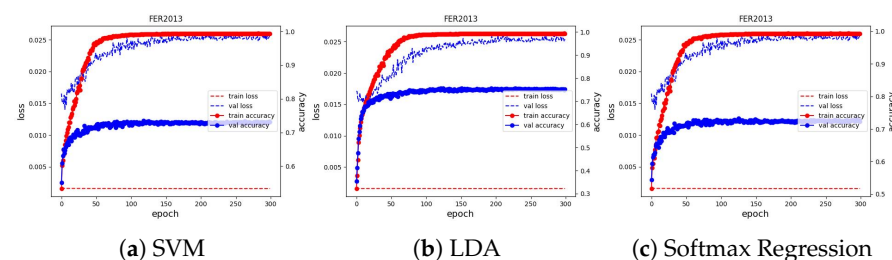
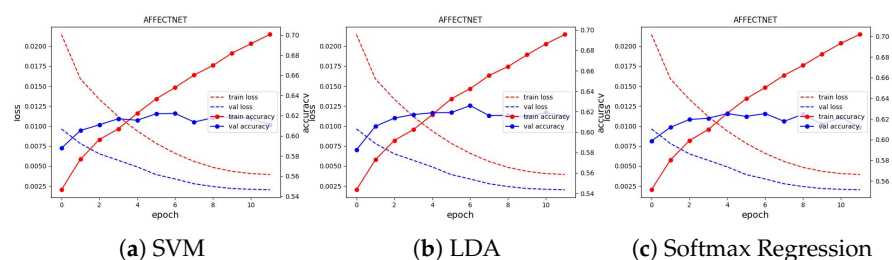
Table 2. Performance comparison between our proposed pipeline with previous works on the MMI database.

	MMI
Proposed method	99.02%
Reference [56]	98.63%
Reference [57]	82.74%
Reference [58]	83.41%
Reference [59]	83.56%

4.4. Imbalance Dataset: FER2013 and AFFECTNET

In contrast, the FER2013 and AFFECTNET datasets are much larger than previous ones, with a total of 26,899 and 1,749,905 samples, respectively. In addition, these two datasets have a problem of imbalanced data samples between classes. Figures 8 and 9 show the sample distribution of the FER2013 and AFFECTNET datasets that reveal a serious imbalance problem in both datasets.

We trained the feature extractor ResNet18 with 300 epochs on FER2013 and 12 epochs on AFFECTNET; other training configurations are kept the same as the experiments on the JAFFE and MMI datasets. The training curve is depicted in Figures 13 and 14. All classifiers are not able to obtain good results on these datasets, with approximately 74.64% and 62.78% in accuracy, and they are comparable with the current SoTA results, with 75% on FER2013 and 3.03% on AFFECTNET. For better analysis, we provide confusion matrices and ROC curves of the best model on each dataset in Figures 15 and 16, respectively.

**Figure 13.** Training curve of three classifiers on the FER2013 dataset.**Figure 14.** Training curve of three classifiers on the AFFECTNET dataset.

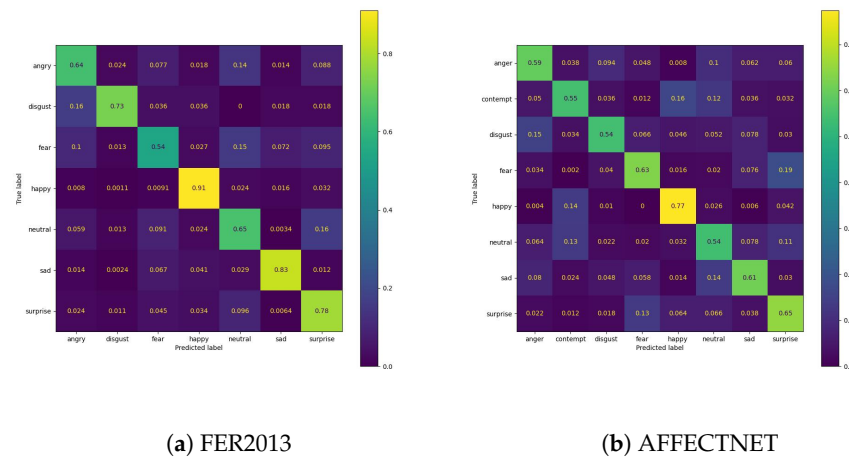


Figure 15. Confusion matrix of best models on the validation set of each dataset.

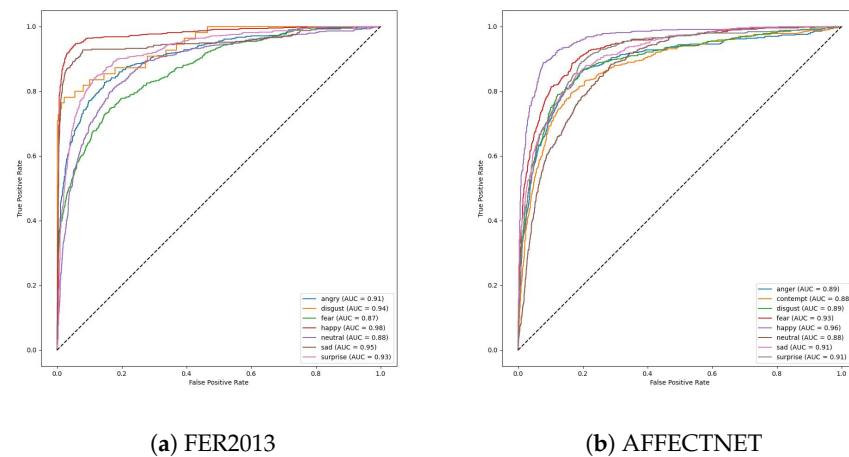


Figure 16. ROC curve of best models on the validation set of each dataset.

In FER2013, the “happy” and “sad” emotion classes have the best performance with 91% and 83% in accuracy, respectively. They also have the largest AUC compared to others in ROC figures. In contrast, the “fear” class has the worst performance, and samples in these classes are usually misclassified to each other. The “disgust” class has very few samples in the FER2013 dataset but achieves an acceptable performance with 73% accuracy.

In AFFECTNET, the “happy” emotion class has a dominant number of samples in it and has achieved the best performance in the AFFECTNET dataset with 77% in accuracy. The worst performance is in the “neutral” class, with only 54% in accuracy, and its samples are usually misclassified with the “contempt” class.

We also provide the best accuracy of the proposed method and comparison of previous works in Tables 3 and 4.

Table 3. Performance comparison between our proposed pipeline and previous works on the FER2013 dataset.

	FER2013	Extra Training Data
Proposed method	74.64%	✗
Reference [60]	75.97%	✓
Reference [61]	64.46%	✗
Reference [62]	69.57%	✗
Reference [63]	70.04%	✗
Reference [64]	74.59%	✗
Reference [6]	72.03%	✗
Reference [52]	73.28%	✗
Reference [65]	74.14%	✗
Reference [66]	72.16%	✗
Reference [67]	75.42%	✓
Reference [47]	76.82%	✓

Table 4. Performance comparison between our proposed pipeline and previous works on the AFFECTNET dataset.

	AFFECTNET 8 Emotions	AFFECTNET 7 Emotions	Extra Training Data
Proposed method	62.78%		✗
Reference [68]		66.46%	✗
Reference [69]		66.37%	✓
Reference [70]	65.20%		✗
Reference [6]	63.36%		✗
Reference [71]	63.03%	66.29%	✗
Reference [72]	63.00%		✗
Reference [73]	62.09%	65.69%	✗
Reference [74]	61.60%	65.40%	✓
Reference [75]	61.32%	65.74%	✓
Reference [76]	53.93%		✗
Reference [77]	59.30%		✗

5. Ablation Study

In this section, we investigate the effect of different training methods on model performance. Tables 5 and 6 show that the triplet loss and weighted data sampling methods have significant impact on the model performance when training on imbalanced datasets such as FER2013 and AFFECTNET.

Table 5. Performance comparison between different training methods on the FER2013 dataset.

	FER2013
Training with Linear Kernel SVM	74.64%
Training with with RBF kernel SVM	74.52%
Training with with Polynomial kernel SVM	74.50%
Training with with Sigmoid kernel	74.49%
Training with LDA	74.62%
Training with Softmax Regression	74.53%
Training with Cross-Entropy Loss	71.40%
Training without Weighted Data Sampling method	64.56%

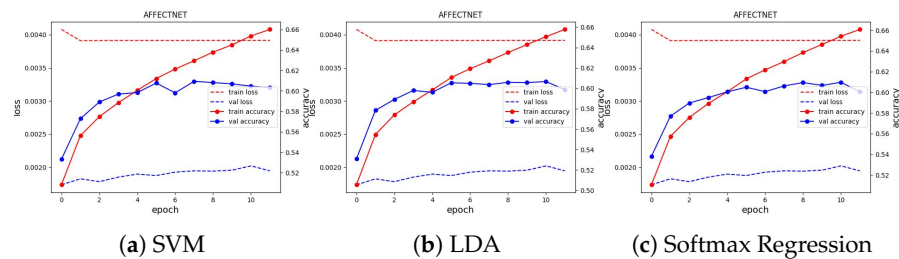
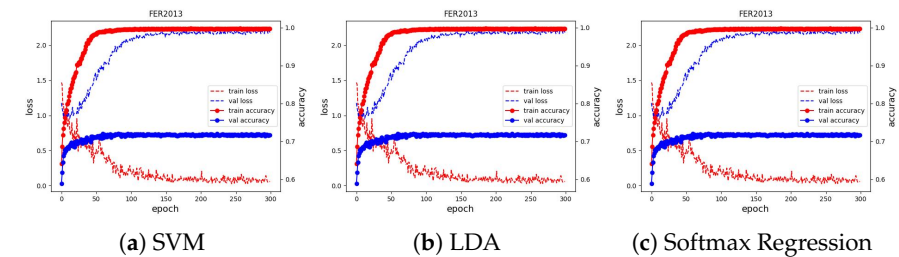
Table 6. Performance comparison between different training methods on the AFFECTNET dataset.

AFFECTNET	
Training with Linear Kernel SVM	62.78%
Training with LDA	62.76%
Training with Softmax Regression	62.73%
Training with Cross-Entropy Loss	60.70%
Training without Weighted Data Sampling method	51.32%

5.1. Using Cross-Entropy Loss

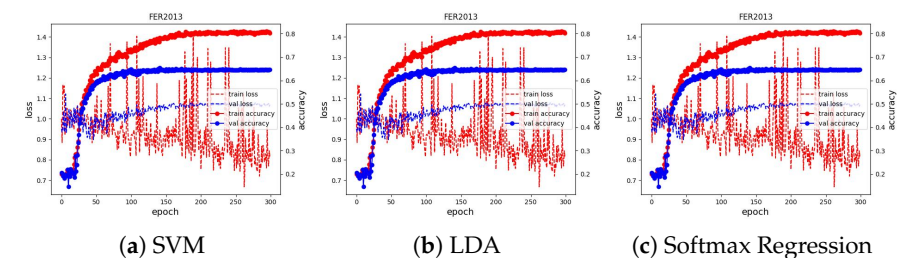
We conducted some experiments by using other loss functions such as cross-entropy loss instead of triplet loss to justify the effectiveness of our proposed method.

Figures 17 and 18 show that cross-entropy loss generates worse results on the FER2013 and AFFECTNET datasets than the proposed method, with only 74.1% and 60.70%, respectively. This indicates that the proposed method can produce deep features with stronger discriminative power.

**Figure 17.** Training curve of three classifiers using Cross Entropy loss on the AFFECTNET dataset.**Figure 18.** Training curve of three classifiers using Cross Entropy loss on the FER2013 dataset.

5.2. Effect of Weighted Data Sampling Method

As described in Section 4.2, we implemented and applied the weighted data sampling method to deal with the data imbalance problem. We found that the data imbalance problem in FER2013 and AFFECTNET has a strong negative effect on model performance. In this subsection, we provide experimental results without applying weighted data sampling in Figures 19 and 20.

**Figure 19.** Training curve of three classifiers on the FER2013 dataset without applying a weighted data sampling method.

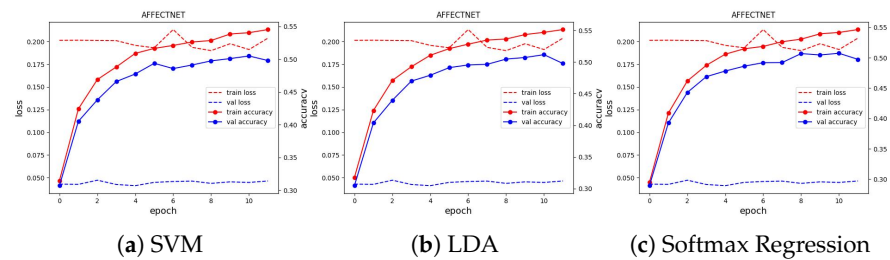


Figure 20. Training curve of three classifiers on the AFFECTNET dataset without applying a weighted data sampling method.

6. Discussion

Our proposed method uses triplet loss to enable the model to generate deep features for each image input. With this loss function, we may bring together, in latent space, the deep features of every pair of samples from the same class and perform the opposite for every pair of samples from different classes. As a result, our model is better able to discriminate across classes, leading to improved accuracy even on datasets with high inter-class similarity. The triplet loss makes the deep features of each sample pair of a same class closer in latent space and vice versa for each data pair of two different classes. The method also achieves SoTA performance without the need for face registration, extensive data augmentation, or extra training data or features. Our FER method is conceptually easier to understand than earlier methods because it does not require face registration and is not impacted by registration errors. We anticipate that using supplemental training data and thorough FER-specific data augmentation would considerably enhance the performance.

Our approach already outperforms earlier efforts on FER and achieves the best results on the JAFFE and MMI datasets. Our approach is most effective for those datasets that have inter-class similarity. To obtain better results, however, the performances on FER2013 and AFFECTNET need to be severely fine-tuned, as they are still far from SoTA performance. This paper proposes a robust pipeline and studies the FER performance on many FER datasets with different characteristics (from small and balanced datasets, such as JAFFE and MMI, to larger and imbalanced datasets such as FER2013 and AFFECTNET). However, due to dataset bias, the results from this and other FER datasets are only representative of real-world FER performance. This restriction applies to FER research in general as well as to this particular study.

We further speculate that the decreased performance compared to state-of-the-art approaches is due to the absence of fine-tuning on the FER2013 and AFFECTNET datasets. Additional training data and comprehensive, FER-specific data augmentation, as proposed in the research, are expected to greatly improve performance on these datasets.

7. Conclusions

In this paper, we proposed a method to solve inter-class problems in emotion recognition by using images and videos. Our pipeline uses RetinaFace as a face detector to detect and crop human face regions in images in order to remove unnecessary information from the input. We also implemented a frame extractor that is based on the MAD outlier detection method in time series to extract the most representative frames in videos. Then, each human face region was fed into a feature extractor ResNet18 model to produce deep features, and the deep features were fed into an SVM classifier to obtain the facial expression class. Performance on other classifiers, such as LDA and Softmax regression, are also reported in this paper. We achieved state-of-the-art accuracies on two datasets named JAFFE and MMI, with comparable results on the FER2013 and AFFECTNET datasets. Multimodal approaches that combine image- and video-based methodologies with other modalities such as voice, text, or physiological inputs should be examined to improve the accuracy of emotion recognition in future research. These methods may improve the overall performance of emotion detection systems by leveraging multiple sources of information, making them more resilient to different environmental conditions.

Author Contributions: Conceptualization, I.H.; methodology, I.H.; writing—original draft preparation, I.H.; writing—review and editing, I.H., S.-H.K. and G.-S.L.; supervision, S.-H.K., G.-S.L. and H.-J.Y.; resources, S.-H.K., G.-S.L.; project administration, S.-H.K., G.-S.L. and H.-J.Y.; funding acquisition, S.-H.K., G.-S.L. and H.-J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2018R1D1A3B05049058, NRF-2021R1I1A3A04036408).

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

- Anderson, K.; McOwan, P.W. A real-time automated system for the recognition of human facial expressions. *IEEE Trans. Syst. Man Cybern.* **2006**, *36*, 96–105. [\[CrossRef\]](#) [\[PubMed\]](#)
- Song, Z. Facial expression emotion recognition model integrating philosophy and machine learning theory. *Front. Psychol.* **2021**, *12*, 759485. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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. [\[CrossRef\]](#)
- Chinnappa, G.; Rajagopal, M.K. Residual attention network for deep face recognition using micro-expression image analysis. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *13*, 117. [\[CrossRef\]](#)
- Happy, S.L.; Routray, A. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **2015**, *6*, 1–12. [\[CrossRef\]](#)
- Fard, A.; Mahoor, M. Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild. *IEEE Access* **2022**, *10*, 26756–26768. [\[CrossRef\]](#)
- Puneet, K.; Jain, S.; Raman, B.; Roy, P.P.; Iwamura, M. End-to-end triplet loss based emotion embedding system for speech emotion recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021.
- Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 403–412.
- Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.
- Caleanu, C.-D. Face expression recognition: A brief overview of the last decade. In Proceedings of the 2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 23–25 May 2013.
- Bettadapura, V. Face Expression Recognition and Analysis: The State of the Art. *arXiv* **2012**, arXiv:1203.6722.
- Sarath, S. Human emotions recognition from thermal images using Yolo algorithm. In Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSPP), Chennai, India, 28–30 July 2020.
- Liu, H.; Zeng, J.; Shan, S.; Chen, X. Emotion recognition for in-the-wild videos. *arXiv* **2020**, arXiv:2002.05447.
- Liu, C.; Jiang, W.; Wang, M.; Tang, T. Group level audio-video emotion recognition using hybrid networks. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 25–29 October 2020.
- Byeon, Y.-H.; Kwak, K.-C. Facial Expression Recognition Using 3D Convolutional Neural Network. *Int. J. Adv. Comput. Sci. Appl.* **2014**, *5*, 107–112. [\[CrossRef\]](#)
- Song, I.; Kim, H.-J.; Jeon, P.B. Deep learning for real-time robust facial expression recognition on a smartphone. In Proceedings of the 2014 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 10–13 January 2014.
- Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, San Francisco, CA, USA, 13–18 June 2010.
- Shan, C.; Gong, S.; McOwan, P. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [\[CrossRef\]](#)
- Zafer, A.; Nawaz, R.; Iqbal, J. Face recognition with expression variation via robust NCC. In Proceedings of the 2013 IEEE 9th International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, 9–10 December 2013.
- Qi, D.; Tan, W.; Yao, Q.; Liu, J. YOLO5Face: Why Reinventing a Face Detector. *arXiv* **2022**, arXiv:2105.12931.
- Wei, L.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, Ch.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14. October 2016, Proceedings, Part I 14; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
- Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.

23. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2017**, arXiv:1612.03144.
24. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
25. Hassaballah, M.; Saddam Bekhet, Amal AM Rashed, and Gang Zhang. Facial features detection and localization. In *Recent Advances in Computer Vision: Theories and Applications*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1, pp. 33–59.
26. Sun, J.; Yang, W.; Gao, R.; Xue, J.-H.; Liao, Q. Inter-class angular margin loss for face recognition. *Signal Process. Image Commun.* **2020**, *80*, 115636. [\[CrossRef\]](#)
27. Shi, X.; Dou, Q.; Xue, C.; Qin, J.; Chen, H.; Heng, P.A. An active learning approach for reducing annotation cost in skin lesion analysis. In *Machine Learning in Medical Imaging, Proceedings of the 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 13 October 2019, Proceedings 10*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 628–636.
28. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection Cvfoundation. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
29. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2017**, arXiv:1612.08242.
30. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
31. Ultralytics. YOLOv5. February 2021. Available online: <https://github.com/ultralytics/yolov5> (accessed on 15 November 2022).
32. Khan, M.; Chakraborty, S.; Astya, R.; Khepra, S. Face Detection and Recognition Using OpenCV. In Proceedings of the 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 18–19 October 2019.
33. Venkataramanan, A.; Laviale, M.; Figus, C.; Usseglio-Polatera, P.; Pradalier, C. Tackling inter-class similarity and intra-class variance for microscopic image-based classification. In Proceedings of the Computer Vision Systems: 13th International Conference, ICVS 2021, Virtual Event, 22–24 September 2021; Proceedings 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 93–103.
34. Lalitha, S.D.; Thyagarajan, K.K. Micro-facial expression recognition based on deep-rooted learning algorithm. *arXiv* **2020**, arXiv:2009.05778.
35. Hassan, M.; Ali, S.; Alquhayz, H.; Safdar, K. Developing intelligent medical image modality classification system using deep transfer learning and LDA. *Sci. Rep.* **2020**, *10*, 12868. [\[CrossRef\]](#)
36. ImageNet. Available online: <http://www.image-net.org/challenges/LSVRC/2014> (accessed on 1 November 2022).
37. Yağ, İ.; Altan, A. Artificial Intelligence-Based Robust Hybrid Algorithm Design and Implementation for Real-Time Detection of Plant Diseases in Agricultural Environments. *Biology* **2022**, *11*, 1732. [\[CrossRef\]](#)
38. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **2013**, *49*, 764–766. [\[CrossRef\]](#)
39. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-stage Dense Face Localisation in the Wild. *arXiv* **2019**, arXiv:1905.00641.
40. Chen, W.; Huang, H.; Peng, S.; Zhou, C.; Zhang, C. YOLO-face: A real-time face detector. *Vis. Comput.* **2020**, *37*, 805–813. [\[CrossRef\]](#)
41. Yang, S.; Luo, P.; Loy, C.; Tang, X. WIDER FACE: A Face Detection Benchmark. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
42. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference On Computer Vision And Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
44. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference On Computer Vision And Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
45. Zeiler, M.; Fergus, R. Visualizing and Understanding Convolutional Networks. *arXiv* **2013**, arXiv:1311.2901.
46. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with Gabor wavelets. In Proceedings of the Third IEEE International Conference On Automatic Face And Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
47. Goodfellow, I.; Erhan, D.; Carrier, P.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.; et al. Challenges in representation learning: A report on three machine learning contests. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2013**, *64*, 59–63. [\[CrossRef\]](#)
48. Mollahosseini, A.; Hasani, B.; Mahoor, M. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [\[CrossRef\]](#)
49. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference On Multimedia And Expo, Amsterdam, The Netherlands, 6 July 2005; p. 5.
50. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Minaee, S.; Abdolrashidi, A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* **2019**, *21*, 3046. [\[CrossRef\]](#)

52. Khairuddin, Y.; Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *arXiv* **2021**, arXiv:2105.03588.
53. Aouayeb, M.; Hamidouche, W.; Soladié, C.; Kpalma, K.; Séguier, R. Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition. *arXiv* **2021**, arXiv:2107.03107.
54. Boughida, A.; Kouahla, M.; Lafifi, Y. A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. *Evol. Syst.* **2022**, *13*, 331–345. [\[CrossRef\]](#)
55. Shaik, N.; Cherukuri, T. Visual attention based composite dense neural network for facial expression recognition. *J. Ambient. Intell. Humaniz. Comput.* **2022**, 1–14. [\[CrossRef\]](#)
56. Burkert, P.; Trier, F.; Afzal, M.; Dengel, A.; Liwicki, M. DeXpression: Deep Convolutional Neural Network for Expression Recognition. *arXiv* **2015**, arXiv:1509.05371.
57. Chen, Y.; Wang, J.; Chen, S.; Shi, Z.; Cai, J. Facial Motion Prior Networks for Facial Expression Recognition. In Proceedings of the 2019 IEEE Visual Communications And Image Processing (VCIP), Sydney, NSW, Australia, 1–4 December 2019; pp. 1–4.
58. Sadeghi, H.; Raie, A. HistNet: Histogram-based convolutional neural network with Chi-squared deep metric learning for facial expression recognition. *Inf. Sci.* **2022**, *608*, 472–488. [\[CrossRef\]](#)
59. Han, B.; Hu, M.; Wang, X.; Ren, F. A Triple-Structure Network Model Based upon MobileNet V1 and Multi-Loss Function for Facial Expression Recognition. *Symmetry* **2022**, *14*, 2055. [\[CrossRef\]](#)
60. Vignesh, S.; Savithadevi, M.; Sridevi, M.; Sridhar, R. A novel facial emotion recognition model using segmentation VGG-19 architecture. *Int. J. Inf. Technol.* **2023**, 1–11. [\[CrossRef\]](#)
61. Abdulsattar, N.; Hussain, M. Facial Expression Recognition using Transfer Learning and Fine-tuning Strategies: A Comparative Study. 2022 International Conference On Computer Science And Software Engineering (CSASE), Duhok, Iraq, 15–17 March 2022; pp. 101–106.
62. Bodapati, J.; Srilakshmi, U.; Veeranjanyulu, N. FERNet: A deep CNN architecture for facial expression recognition in the wild. *J. Inst. Eng. Ser. B* **2022**, *103*, 439–448. [\[CrossRef\]](#)
63. Oguine, O.; Oguine, K.; Bisallah, H.; Ofuani, D. Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction. *arXiv* **2022**, arXiv:2206.09509.
64. Qi, Y.; Zhou, C.; Chen, Y. NA-Resnet: Neighbor block and optimized attention module for global-local feature extraction in facial expression recognition. *Multimed. Tools Appl.* **2023**, *82*, 16375–16393. [\[CrossRef\]](#)
65. Pham, L.; Vu, T.; Tran, T. Facial Expression Recognition Using Residual Masking Network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4513–4519.
66. Vulpe-Grigoras, I.A.; Grigore, O. Convolutional Neural Network Hyperparameters optimization for Facial Emotion Recognition. In Proceedings of the 2021 12th International Symposium On Advanced Topics In Electrical Engineering (ATEE), Bucharest, Romania, 25–27 March 2021; pp. 1–5.
67. Georgescu, M.; Ionescu, R.; Popescu, M. Local Learning With Deep and Handcrafted Features for Facial Expression Recognition. *IEEE Access* **2018**, *7*, 64827–64836. [\[CrossRef\]](#)
68. Antoniadis, P.; Filntisis, P.; Maragos, P. Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition. *arXiv* **2021**, arXiv:2106.03487.
69. Ryumina, E.; Dresvyanskiy, D.; Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing* **2022**, *514*, 435–450. [\[CrossRef\]](#)
70. Farzaneh, A.; Qi, X. Facial Expression Recognition in the Wild via Deep Attentive Center Loss. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2402–2411.
71. Savchenko, A.; Savchenko, L.; Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2132–2143. [\[CrossRef\]](#)
72. Kollias, D.; Zafeiriou, S. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *arXiv* **2019**, arXiv:1910.04855.
73. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition. *arXiv* **2021**, arXiv:2109.07270.
74. Schoneveld, L.; Othmani, A.; Abdelkawy, H. Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition. *arXiv* **2021**, arXiv:2103.09154.
75. Savchenko, A. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In Proceedings of the 2021 IEEE 19th International Symposium On Intelligent Systems And Informatics (SISY), Subotica, Serbia, 16–18 September 2021; pp. 119–124.
76. Zhou, H.; Meng, D.; Zhang, Y.; Peng, X.; Du, J.; Wang, K.; Qiao, Y. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In Proceedings of the 2019 International Conference On Multimodal Interaction, Suzhou, China, 14–18 October 2019.
77. Siqueira, H.; Magg, S.; Wermter, S. Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks. *arXiv* **2020**, arXiv:2001.06338.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.