Efficient Deep Learning-Driven Systems for Real-Time Video Expression Recognition

by

James Ren Hou Lee

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following papers are used in this thesis. I was co-author with major contributions to the design, analysis, writing, and editing.

J. Lee and A. Wong, "TimeConvNets: A Deep Time Windowed Convolution Neural Network Design for Real-time Video Facial Expression Recognition," *2020 17th Conference on Computer and Robot Vision (CRV)*, 2020. This paper is incorporated in Chapter 3.

J. Lee, L. Wang, and A. Wong, "EmotionNet Nano: An Efficient Deep Convolutional Neural Network Design for Real-time Facial Expression Recognition," *arXiv e-prints, page arXiv:2006.15759*, 2020. This paper is incorporated in Chapter 4.

J. Lee and A. Wong, "AEGIS: A real-time multimodal augmented reality computer vision based system to assist facial expression recognition for individuals with autism spectrum disorder," *arXiv e-prints, page arXiv:2010.11884*, 2020. This paper is incorporated in Chapter 5.

Abstract

The ability to detect, recognize, and interpret facial expressions is an important skill for humans to have due to the abundance of social interactions one faces on a daily basis, but it is also something that most take for granted. Being the social animals that we are, expression understanding not only enables us to gauge current emotional states, but also allows for the recognition of conversational cues such as level of interest, speaking turns, and level of information understanding. For individuals with autism spectrum disorder, a core challenge that they face is an impaired ability to infer other people's emotions based on their facial expressions, which can cause problems when creating and sustaining meaningful, positive relationships, leading to troubles integrating into society and a higher prevalence of depression and loneliness. However, with significant recent advances in machine learning, one potential solution is to leverage assistive technology to aid these individuals to better recognize facial expressions. Such a technology requires reasonable accuracy in order to provide users with correct information, but also must follow a real-time constraint to be relevant and seamless in a social setting.

Due to the dynamic and transient nature of human facial expressions, a challenge during classification is the usage of temporal information to provide additional context to a scene. Many applications require the real-time aspect to be preserved, and thus temporal information must be leveraged in an efficient manner. Consequently, we explore the dynamic and transient nature of facial expressions through a novel deep time windowed convolutional neural network design called TimeConvNets, that is capable of encoding spatiotemporal information in an efficient manner. We compare against other methods capable of leveraging temporal information, and show that TimeConvNets can provide a real-time solution that is both accurate as well as architecturally and computationally less complex.

Even with the strong performances that the TimeConvNet architecture offers, additional architecture modifications tailored specifically for human facial expression classification can likely result in increased performance gains. Thus, we explore a human-machine collaborative design strategy for the purpose of further reducing and optimizing these facial expression classifiers. EmotionNet Nano was created and tailored specifically for the task of expression classification on edge devices, by leveraging human experience combined with the meticulousness and speed of machines. Experimental results on the CK+ facial expression benchmark dataset demonstrate that the proposed EmotionNet Nano networks achieved accuracy comparable to state-of-the-art, while requiring significantly fewer parameters, and are also capable of performing inference in real-time, making them suitable for deployment on a variety of platforms including mobile phones.

To train these models, a high quality expression dataset is required, specifically one that retains temporal information between consecutive image frames. We introduce FaceParty as a solution, which is a more difficult dataset created by the modified aggregation of six public video facial expression datasets, and provide details for replication. We hope that models trained using FaceParty can achieve increased generalization ability for faces in the wild due to the nature of the dataset.

Acknowledgements

First, I would like to thank my amazing supervisor, Professor Alexander Wong, for being the most knowledgeable, helpful, passionate, and caring mentor that anyone could ask for.

Next, I would like to thank Professors John Zelek and David Clausi for taking the time out of their busy schedules to read and review my thesis.

Third, I would like to thank my family, for always believing in me and supporting me behind the scenes.

Lastly, I would like to thank all of my friends and colleagues that I met before and during my Master's journey, for their never ending support and assistance, and also for making these past few years incredibly entertaining and educational.

Dedication

This thesis is dedicated towards my family, who somehow knew I would pursue higher education even before I did. It is also dedicated towards all those who can benefit from assistive living technology, and I hope that this thesis can help them in some way in the future.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The ability to detect, recognize, and understand different facial expressions is a crucial skill to have in everyday life, due to the abundance of social interactions one faces on a daily basis. This skill not only grants understanding of current emotional states, but also allows the user to recognize conversational cues such as level of interest, speaking turns, and level of information understanding [43]. Research has shown that a staggering 55% of the information behind a spoken message stems from facial cues, and only 7% is attributed to the words themselves [53]. Clearly, the visual component of the message is key when interpreting a message, even more important than the aural aspect, which only takes up 38% of the spoken message [53].

For the majority of individuals with autism spectrum disorder (ASD), a core challenge that they face is an impaired ability to infer other people's emotions based on their facial expressions. As such, the complexity of human societal interaction is further elevated, and can often lead to a higher prevalence of loneliness and depression, due to difficulties interacting with society. Motivated by this important social need as well as the significant recent advances in machine learning, one possible approach for assisting individuals with ASD to better recognize facial expressions is the design of computer vision-driven facial expression classification (FEC) systems for enabling improved emotion inference and improving social interactions. However, this does not come without its own set of challenges, which must be addressed before an accurate real-time system can be achieved.

1.1 Applications of Facial Expression Recognition

As stated previously, the main motivation behind this research is to assist individuals with ASD, in order to to help them interact with society without suffering from an impaired expression recognition ability. However, facial expression recognition (FER) systems possess utility beyond this application, and has already been used in many other fields ranging from driver state monitoring, where a camera system would be used to detect driver fatigue and mental state via their facial expressions [28], to marketing, where facial expressions could be used to determine approval and affinity towards a specific advertisement [3]. In addition, similar methods could be used for various other domains, such as depression detection, security checkpoints, ambient interfaces, and empathetic tutoring [2, 57].

In order for many of these applications to be used seamlessly in the real-world, the real-time aspect is crucial due to the nature of facial expressions and how fast they change over time. In addition, many of these fields require low-cost, small embedded devices, such that the average consumer can have access to them as well. For example, in the area of assistive technologies for improving quality of life, the majority of individuals using such technologies are unwilling to carry large, bulky, and expensive devices with them during their daily lives, as that would be a big hindrance that limits their ability to leverage the technologies in a seamless manner. As such, the assistive devices must leverage small, low-cost, embedded processors, yet provide low latency to enable real-time feedback to the user. In the in-car driver monitoring study done in [28], where a FEC system would record the driver and determine their current mental state, and warn them if their awareness level is deteriorating, the difference of a few milliseconds of processing is paramount for the safety of not only the user, but also other drivers on the road. A time differential of a few milliseconds on the highway can be the difference between a fatal car crash and a safe stop. In applications for fields such as marketing or security, real-time processing is important to provide salespeople or security guards immediate feedback such that an appropriate response can be made as soon as possible. For those relying on software assistance for social purposes, information is required at no delay in order to keep a conversation alive and not cause discomfort for both parties.

There are also many use cases for FER that do not have the real-time requirement, such as the review of an video in order to analyze a subject or subjects after the fact. If immediate feedback is not necessary, such as analyzing customer satisfaction via facial expressions, or investigating suspicious behaviour after an event, then a higher classification accuracy can be the main focus instead of a low complexity model. Naturally, these systems would likely be run in the background and on larger scale hardware due to their processing requirements, making them less portable and accessible for the average user, but the higher

quality results would compensate for this.

1.2 Challenges of Facial Expression Recognition

As mentioned in the previous section, in order for many FEC systems to function seamlessly in the real world, the real-time aspect is a key requirement that cannot be ignored. In fact, the sacrifice of a few percentage points in accuracy is sometimes necessary in order to obtain a faster processing speed, and this is a trade-off that must be explored on a case-by-case basis depending on the product domain. In the driver monitoring example, having the lowest possible inference time would be much safer for everyone involved, while a pseudo real-time system in a marketing scenario would still be acceptable due to the nature of the task.

For our specific purpose of helping individuals with ASD better recognize facial expressions during social interactions, there are a number of challenges that must be taken into account. First, there is a low latency requirement that must be included for fluent social interactions with others. It is required that the latency between the facial expression being made and the recognition and visualization of the emotion being conveyed be as low as possible, such that an individual with ASD can interpret the emotion as fast as possible to then properly engage the individual he or she is in conversation with. Low latency may not be a requirement in all fields involving FEC, such as when performing post-event video analysis, however, it is crucial for fields involving safety (i.e driver monitoring, security) and socialization (i.e live marketing events, in-person communication), due to the immediate feedback required. Seconds of inference delay for a security application could be disastrous, and in a social setting, feedback of someone's expression seconds after the fact is essentially worthless. Due to these reasons, researchers involved in real-time FEC tend to leverage lightweight networks or designs, use smaller inputs for classification, and avoid large amounts of data pre-processing [34].

However, a small trade-off for a higher accuracy can be made for our specific purpose of real-time FEC for individuals with ASD, as a seamless experience can still be obtained even with a lower number of "processed" frames, through the use of older predictions. As an example, if we assume that 30 frames per second can be acknowledged as real-time performance, we can actually process just one or two frames per second, and then use these as the predictions for the remaining frames. As a result, we can boost the classification accuracy of our networks through the use of additional information, at the cost of inference time, which leads us to our second challenge.

The next challenge that facial expression recognition faces is the inherent transient nature of these expressions. This interesting characteristic means that the expressions that we humans make consist of an onset, peak, and offset phase [24], meaning that temporal information becomes especially important in order to capture this smooth transition as well as the transient nuances of these expressions. Traditional 2D convolutional neural networks (CNN) have one major flaw in that they are primarily designed to capture spatial characteristics and not transient characteristics [18]. In order to encode the spatiotemporal relationships between consecutive frames, deep neural networks that leverage 3D convolutions (3DC) have been proposed, as well as architectures that leverage long short term memory (LSTM) units to learn temporal information while avoiding the vanishing or exploding gradient problems. However, such strategies can be very computationally complex to perform inference with, particularly since they fall within the dynamic video classification category of leveraging entire video sequences to predict expressions. This thesis proposes a solution to this complexity by using only a small portion of this information, in the form of "time windows," where we take a t-channel sub-sequence stack and use this as input, rather than the entire video sequence [33]. Thus, we can control the balance between the amount of temporal information leveraged, versus the amount of time taken for processing. The TimeConvNet [33] architecture will be discussed in detail in Chapter 3.

Even with the strong performances that the TimeConvNet architecture offers, it is likely that additional architecture modifications tailored specifically for human facial expression classification could result in increased performance gains. The design of network modules and architectures by hand is a difficult problem, one that is often time-consuming and tedious for humans to perform themselves. Allocating this task to machines instead has been a method explored in recent years via neural architecture search (NAS) strategies, but significant human effort is still required to design the search space in a way that reduces it to a feasible size, as well as defining a search strategy that can run within desired operational constraints and requirements in a reasonable amount of time. We address this challenge by exploring the use of a human-machine collaborative design strategy to create an efficient architecture catered specifically for the task of FEC. The resulting architecture, which we call EmotionNet Nano [26], is presented in Chapter 4, and compared against similar state-of-the-art models on a popular facial expression dataset.

In order for a model to be trained easily and accurately, large amounts of clean, organized, and error-free data are necessary, much of which is often difficult to obtain. For our purposes, an abundance of human face images labelled with specific expressions would be required in order to create an accurate facial expression predictor. Currently, the largest human face image database is EmotioNet [17], consisting of a million images collected from the internet labelled with 23 basic and compound expressions. Similar datasets to this ex-

ist, such as Multi-PIE [22], with 755,000 images classified into 7 classes, and AffectNet [44] comprising 450,000 images with 7 classes as well. However, these large datasets are unsuitable for our needs, due to their static image property. In order to train our networks for time-windowed learning, temporal data is necessary, which these datasets do not provide. Even if we were to train for single frame classification, another issue arises - the fact that these datasets do not share all of the same classes. As an example, Multi-PIE includes the "scream" and "squint" expressions, whereas most other datasets such as AffectNet do not. Once we limit the search of facial expression datasets to ones retaining temporal information, the amount of samples decreases dramatically. Whereas static single image datasets can contain hundreds of thousands of unique, labelled data points, video or image sequence datasets have only hundreds to a few thousand, limiting the scale at which these models can be trained. Thus, using any individual dataset would likely not be enough for a well trained model, and overfitting on the dataset is a real possibility that cannot be overlooked. To address this issue, we decided to combine many of these datasets and use them together in order to increase the generalization ability of our networks, due to the added amounts of unique subject faces the model would encounter. Using the seven class system described earlier as a baseline requirement for our database selections, we chose six public datasets - CK+ [39], BAUM-1 [64], eNTERFACE [42], AFEW [11], KDEF-dyn [6], and iSAFE [52]. Details on the first three datasets can be found in Chapter 3, as they were used in the first iteration of the amalgamated dataset, called BigFaceX. The following three datasets, AFEW, KDEF-dyn, and iSAFE, are discussed in Chapter 6, with details on how they were integrated into our proposed dataset, **FaceParty**.

Facial expression classification is a difficult problem by nature, one that even humans have trouble solving consistently. Ekman & Friesen [43] have found that the human "ceiling" in predicting facial expressions is 91.7 percent, but this is only for the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), and thus does not include the many other emotions we are capable of expressing, such as boredom, concentrated, intrigued, and so on. When classifying large static image datasets such as FER2013, even state-of-the-art models can only achieve roughly 75 percent accuracy [20], much lower than the performance attainable on other image recognition tasks. In an experiment performed by Goodfellow et al. [21], 1500 images of lab members performing facial expressions were collected and labelled. On this private dataset containing zero label noise, human accuracy was a mere 68±5%, meaning that around 1 in 3 predictions were incorrect! Experiments were also conducted on the FER2013 dataset in [21], where Goodfellow et al. found that human accuracy on FER2013 was around 65±5%, again showing how difficult accurate facial expression classification really is. Even though it is unlikely that models now and in the future will be able to reach perfect accuracy on expressions in the wild, human level

performance is completely acceptable due to the ambiguous nature of our facial expressions.

1.3 Thesis Contributions

There are three main contributions of this thesis:

1. TimeConvNets, which are novel time-windowed convolutional deep neural network designs, capable of leveraging temporal information in an efficient manner.
2. EmotionNet Nano, an exploration of a human-machine collaborative design strategy for a highly compact human facial expression classifier.
3. FaceParty, a custom, more difficult human facial expression dataset formed from the modified aggregation of six public datasets.

The goal of these three main contributions is to provide individuals with ASD additional tools that they can use to enhance their societal interactions with other people. These tools are also directed towards researchers in this field, so that they may use them to more easily create better FEC systems in the future.

In this thesis, a review of relevant background material is provided in Chapter 2. TimeConvNets are introduced and discussed in Chapter 3, showcasing their ability to leverage temporal information in a more efficient manner. In Chapter 4, a different strategy is employed, where we leverage a human-machine collaborative design strategy in order to create EmotionNet Nano. Following this, we propose AEGIS in Chapter 5, which is an assistive technology software created for the purpose of helping individuals with ASD better learn and understand facial expressions. Finally, a new dataset FaceParty is proposed in Chapter 6, building upon the one introduced in Chapter 3, in which six public datasets are used instead of just the three used in BigFaceX. Conclusions and parting thoughts are given in Chapter 7.

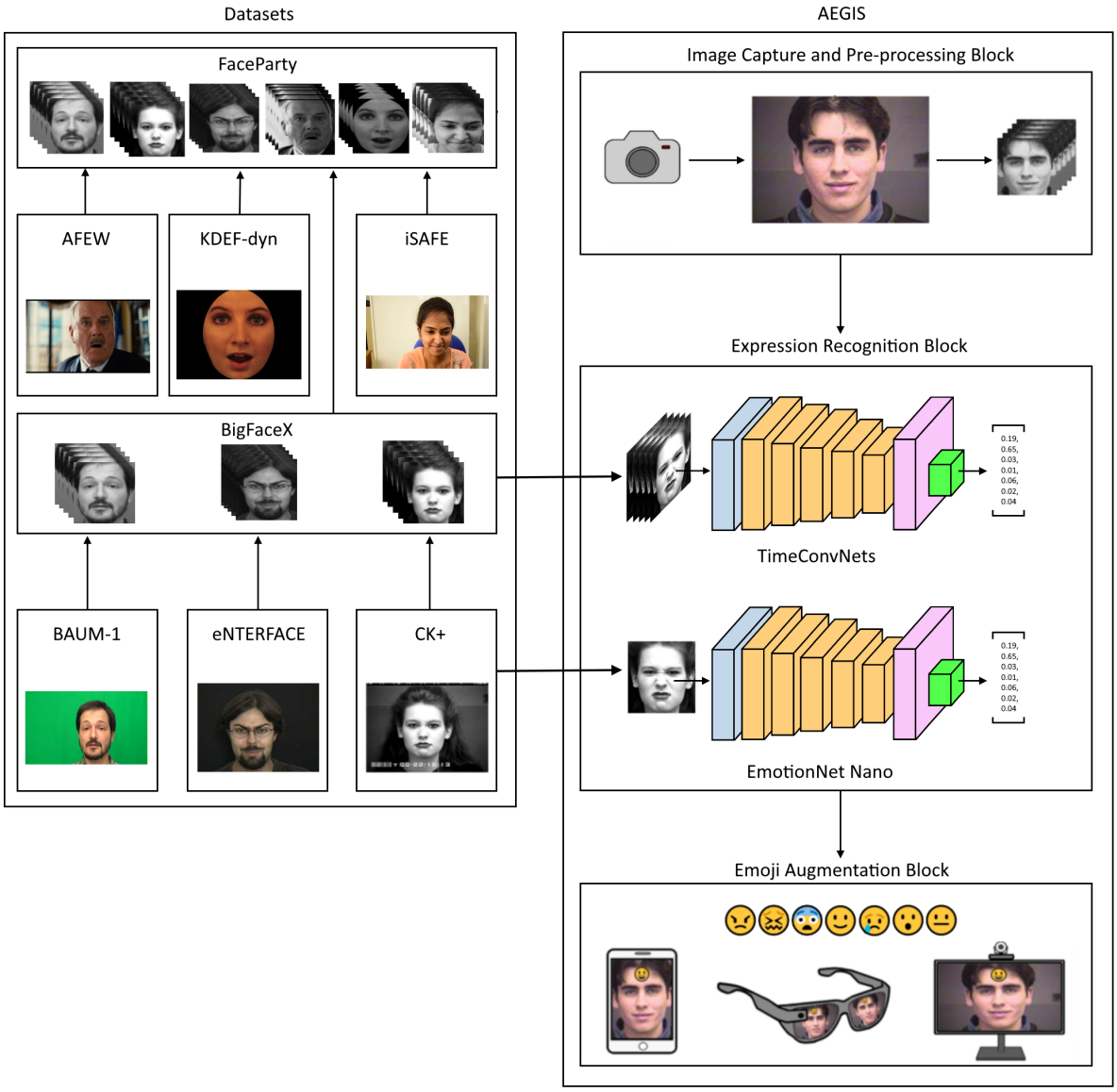

Figure 1.1: **Thesis Key Contribution Flowchart.** How the contributions of this thesis fit together. The AEGIS system, discussed in Chapter 5 leverages a TimeConvNet [33] (Chapter 3) or EmotionNet [26] (Chapter 4) model to perform expression recognition. These models can be trained using BigFaceX or FaceParty (Chapter 6), both of which are novel, more difficult datasets. The result is a real-time expression classifier that leverages temporal information to provide accurate feedback in an intuitive interface.

Chapter 2

Background

The field of facial expression classification has been an active area and has resulted in a wide variety of approaches over the years, to tackle challenging problems in fields such as security [5], safety [28], assistive living [38], marketing [3], depression detection, and empathetic tutoring [2, 57]. Traditionally, machine learning-driven strategies for FEC typically involve the coupling of a machine learning algorithm (e.g., support vector machines (SVM) [43], decision trees [45]) with a combination of different hand-engineered features and other computer vision techniques (e.g., local binary patterns, feature point tracking, and dense optical flow [2], histogram of oriented gradients [45]) in order to achieve high performance. More recently however, much of the focus and advances in FEC has evolved around deep learning, with a number of different deep neural network architectures being explored to tackle these problems.

The benefits of using deep learning instead of handcrafted features are two-fold. First, the generation of handcrafted features needs additional pre-processing steps and also sometimes requires a time consuming manual labelling step performed by human annotators, such as in the case of the Facial Action Coding System (FACS) [13]. Proposed in [13] by Ekman & Friesen, and now well known in facial expression recognition literature, the system resolves around the use of 44 unique action units that also vary in intensity based on a 5-point ordinal scale. As the FACS requires trained human observers to manually code all possible facial displays on video sequences frame by frame, significant human effort is required in order to create this detailed but very descriptive dataset. Alternate handcrafted feature techniques also exist, and are much less manual than the FACS, including Gabor wavelet features [36], local binary patterns [2], histogram of oriented gradients [45], hierarchical optical flow [9], and Essa & Pentland's [16] technique of extracting facial features using Principal Component Analysis (PCA) via Fast Fourier Transforms and local energy

computation. However, a key limitation to the use of handcrafted features is that they may not well capture the subtle spatiotemporal nuances in facial expressions, which can be critical for high facial emotion classification performance. Furthermore, a number of these handcrafted features can be time consuming to calculate, thus potentially reducing the speed in which facial emotion classification can be performed, especially on devices where memory and computation allocations are limited.

The second benefit of leveraging deep learning is that deep learning approaches enable the features to be learned directly from the wealth of image and video data available, rather than using handcrafted features that could be limiting in terms of expressiveness in characterizing spatiotemporal facial nuances. As a result, this can lead to diverse embedded features within a deep neural network that can provide a greater ability to capture such subtle yet critical facial nuances. Research leveraging deep learning has increased in popularity due to this, and a wide range of deep neural network architectures such as CNNs, RNNs [18], long short term memory (LSTM) networks [53], and DNNs [24] have been explored for the purpose of facial emotion classification. Furthermore, if the network has trouble learning accurate representations for facial expressions, additional mechanisms to further boost the performance of deep neural networks can be used, thus combining handcrafted features with deep learning techniques. Some examples include facial landmark injection [24], incorporating audio features in the network [18, 53], as well as ensembling with other machine learning strategies [2]. In [2] specifically, Bargal et al. trained three deep networks to create a concatenated feature vector that is classified using a support vector machine. Pan et al. [48] used a CNN-LSTM fusion network in order to utilize temporal information in video sequences. Sun et al. [53] performed a similar process but also included audio information, and used linear SVMs for classification. Hasani and Mahoor [24] modified an Inception [54] ResNet to use 3DC layers instead, followed by an LSTM unit.

One important detail to note from these studies is that the retention of temporal information can improve facial emotion classification performance by better capturing the dynamic nature of human expressions. However, a limitation with the aforementioned deep learning strategies for dynamic video classification is that they generally have high computational complexities, particularly given that they leverage entire video sequences for prediction, thus making them challenging to employ in real-time, low-latency scenarios such as assisting individuals with ASD during social interactions. In contrast, static single frame classifiers for FER also exist, and tend to be much simpler in terms of architectural and computational complexity, but come at the cost of a lower prediction accuracy.

2.1 Expression Classification Strategies

Throughout the field of expression classification, even with the wide range of methods used to solve the problem, most methods proposed in research literature tend to fall under one of two categories, either static single image classification, or dynamic video classification. The reasoning behind this binary separation is simply because of the balance between classification speed and accuracy. Classification performed using single frames would undoubtedly be faster than multi-frame input, but the increased information from the additional frames, if used properly, can result in accuracy increases.

2.1.1 Static single image classification

In this subset of FEC, a single image is used by the machine learning algorithm to predict a single emotion label (e.g., (anger, disgust, fear, happiness, sadness, surprise) [24, 43], or neutral). This category of strategies have faster inference speed but do not better capture spatiotemporal characteristics associated with facial expressions. Generally, classifiers belonging to this type tend to perform in real-time due to their lower computational complexity, and include systems using support vector machines [43] to classify displacements of manually defined facial landmarks, and local binary patterns [23] as feature vectors in combination with principal component analysis. Wang et al. [60] leverage Adaboost [51] to enhance the performance of a weak classifier comprised of a real-valued Haar feature based 2D Look-Up-Table, and demonstrate that high accuracy can be achieved on the JAFFE [41] dataset even with a low processing time of 0.11 milliseconds per image. Esau et al. [15] classifies facial expressions from image sequences in real-time using a fuzzy emotion model that outputs blended classification results with varying intensity for each emotion. While these low latency systems all run in real-time due to the nature of their design, they face particular difficulty when predicting emotions of faces eliciting non-peak expressions, as they do not capture the spatiotemporal characteristics of these faces. Thus, the ability to leverage both the high accuracy potential of spatiotemporal information as well as achieve low latency in facial emotion classification is highly desired.

An important pre-processing step is also required for real-time classification, which cannot be ignored as it can take just as much time as the classification itself. Face localization, determining if an image actually contains a face, or cropping an image to a facial bounding box can be done in many methods, including Haar Cascades [58] or face tracking using eye pupils [15], but generally require significant processing time that can cause a system to be unable to achieve real-time performance if not careful. Thus, many real-time systems tend

to use low resolution images for faster face localization, or computationally cheap feature extraction methods such as local binary patterns or histogram of oriented gradients. Static single frame real-time classifiers need only process one image at a time to perform inference which is relatively cheap in terms of processing, but when leveraging multiple frames for a single output prediction, the efficiency of pre-processing and extraction of key features becomes increasingly important.

2.1.2 Dynamic video classification

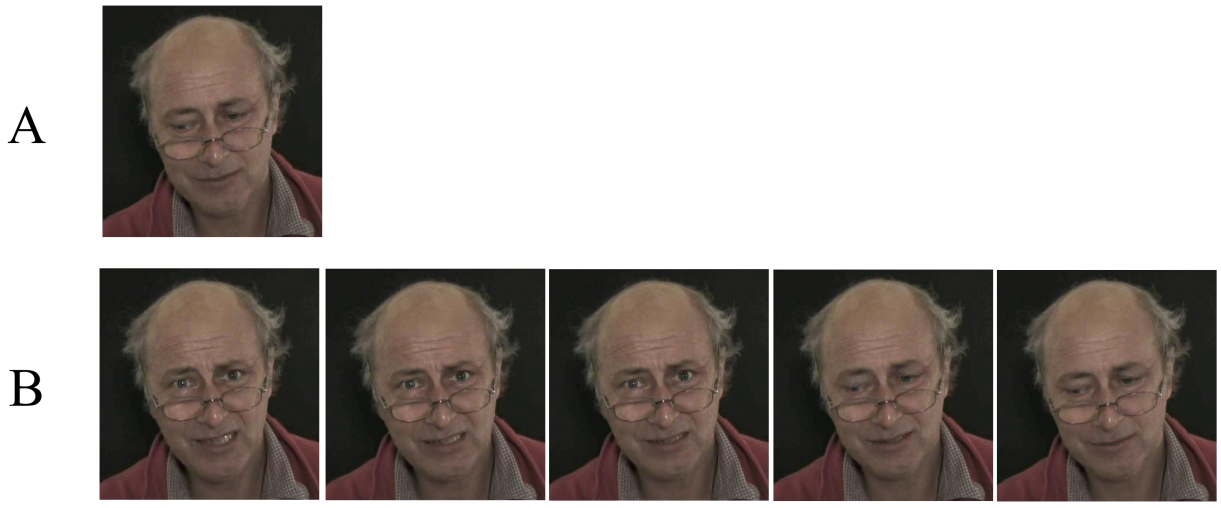

Figure 2.1: **The difficulties of classification using only a single frame.** Sequence A, consisting of only a single frame, is hard to classify due to an ambiguous expression (i.e. could be sadness, fear, or disgust). However, if previous temporal information is included, shown in Sequence B, we can see that the expression is most likely to be fear. Images taken from the eNTERFACE [42] dataset.

For dynamic video classification, an entire video sequence is typically used by the machine learning algorithm to predict a single emotion label. This category of strategies better captures the spatiotemporal characteristics embedded in the images, at the expense of computational complexity. Due to this, most classifiers leveraging entire video sequences cannot perform in real-time, due to the slow inference speeds associated with classifying videos consisting of hundreds to thousands of frames.

One of the challenges with analyzing human facial expressions is the fact that our expressions are of a transient and dynamic nature [24]. Not only do they consist of onset,

peak, and offset phases, the duration of each phase can range from milliseconds to seconds long, making some expressions incredibly difficult to capture even for humans ourselves. However, if each phase of the expression can be captured and analyzed as a whole, the deduction of an expression tends to be easier than if only a single frame was used. Figure 2.1 shows an example of such a problem. In Sequence A in the figure, we are shown only a single frame centered on a human face. To another human, we can be sure that the expression shown is not one of happiness, surprise, or anger, but it is much more ambiguous among other expressions such as sadness, fear, or disgust. Clearly, additional context would be helpful in this situation in order to determine the current expression, which is what is given in Sequence B. Previous frames are included, and with the additional context we can determine that the expression is most likely to be fear, rather than sadness or disgust.

In order to extract the temporal information embedded in video sequences, certain design changes are required. Traditional 2D CNNs have one major flaw in that they are primarily designed to capture spatial characteristics and not transient characteristics [18], meaning that they are not suited for the encoding of spatiotemporal relationships between consecutive frames. In order to leverage this additional information, authors have tried a variety of methods, such as 3D convolutions (3DC) [24], LSTM units to learn temporal information while avoiding the vanishing or exploding gradient problems [53], or by encoding the temporal difference information into the model inputs directly [37, 62]. However, such strategies can be very computationally complex to perform inference with, and sometimes also require additional information to be injected into the models, meaning that the additional pre-processing steps required can cause these networks to be infeasible for use in a real-time scenario. In the next chapter, we propose a simple method to address these issues, with a network architecture which we call TimeConvNets [33], that is able to learn temporal information while maintaining real-time performance.

2.2 Facial Expression Datasets

In order to build machine learning driven strategies for facial emotion classification, a key ingredient is the availability of a training dataset. One of the most widely-used datasets for static single image classification is FER2013 [21], which contains images of faces at the temporal peak of the expressions, extracted from online images via search engine queries. Many large static image datasets follow a similar process, and usually require an additional verification stage at the end where a human annotator manually checks every image for validity and correctness. Some examples include the AffectNet [44], Expression-in-the-Wild (ExpW) [65], and Real-world Affective Faces (RAF) [35] datasets, each ranging from

tens of thousands to hundreds of thousands of labelled face images. These datasets tend to consist of single frames each centered on one subject displaying a facial expression, but datasets containing multiple subjects in one frame also exist, such as the Happy People Images (HAPPEI) [10] database designed for group emotion recognition, or a variant of the ExpW dataset designed for interpersonal relation prediction.

Facial expression video datasets typically contain much fewer data samples due to the increased difficulty associated with collecting them. Most datasets involve the recording of real participants [39, 42, 49, 56, 64], which can be troublesome to coordinate and time consuming to collect. The Acted Facial Expressions in the Wild (AFEW) [11] dataset attempts to avoid this issue by obtaining the data from movies instead, but this raises other issues due to a lack of control over variables such as lighting, pose, background, and occlusions. In order to ease the neural network training process, researchers generally try to control these parameters as much as possible, and bring subjects into a laboratory controlled environment for data collection, thus keeping constant the background and lighting of the scene. Additional control over the type of stimuli can also be leveraged, in order to elicit either spontaneous or acted expressions from the participants. Naturally, the acted expressions tend to have a more extreme peak as the subjects try to make the expression more obvious, but this sometimes causes the result to be an inaccurate representation of the expression in the real-world. On the other hand, spontaneous reactions are much more natural, but can sometimes fail to elicit the desired emotional response from the subject, resulting in image frames that are ambiguous and hard to classify even for human annotators.

Widely-used video datasets for dynamic video classification strategies include the extended Cohn-Kanade (CK+) dataset [39], the MMI dataset [49], the AFEW dataset [11], and the Oulu-CASIA dataset [66]. Other less well known datasets include the BAUM-1 dataset [64], the eNTERFACE dataset [42], and the GEMEP-FERA [56] dataset. A more comprehensive comparison of these datasets is given in Table 2.1.

For the purposes of time windowed learning, static image datasets such as FER2013 or AffectNet were unsuitable for two reasons: first, each image was completely unrelated to the others, and second, the images in the dataset show only the peak of the expression. As such, only dynamic video datasets could be used in order to extract the temporal relationship between consecutive frames. One of the key challenges associated with facial expression classification is their dynamic nature, meaning that the retention of temporal information can cause higher accuracy to be achieved from the additional information gained. We explore this relationship in the next chapter, where we introduce TimeConvNets [33], leveraging our novel dataset BigFaceX, created by the modified aggregation of three public video datasets - the extended Cohn-Kanade (CK+) dataset [39], the BAUM-1 dataset [64],

Table 2.1: **A comparison of popular facial expression datasets.** This is in no means a comprehensive list of all public datasets available. The six basic expressions are Anger, Disgust, Fear, Happy, Sad, and Surprise.

Dataset	**Type**	**Subjects**	**Samples**	**Classes**
FER2013 [21]	Static	N/A	35,887	6 basic plus neutral
AffectNet [44]	Static	N/A	318,969	6 basic plus contempt and neutral
ExpW [65]	Static	N/A	91,793	6 basic plus neutral
RAF [35]	Static	N/A	29,672	6 basic plus neutral
JAFFE [41]	Static	10	213	6 basic plus neutral
CK+ [39]	Video	118	327	6 basic plus contempt
BAUM-1 [64]	Video	31	1,184	6 basic plus contempt, neutral, boredom, interest, bothered, concentrating, and thinking
eNTERFACE [42]	Video	42	1,287	6 basic
AFEW [11]	Video	N/A	1809	6 basic plus neutral
KDEF-dyn [6]	Video	40	240	6 basic
iSAFE [52]	Video	44	395	6 basic plus neutral
MMI [49]	Video	32	213	6 basic
Oulu-CASIA [66]	Video	80	2,880	6 basic

and the eNTERFACE dataset [42], and show that these TimeConvNets are able to learn temporal information while maintaining real-time performance.

Chapter 3

TimeConvNets

In this chapter, an intermediary design between static single frame classification and dynamic video classification is explored. The TimeConvNet architecture itself is shown in Section 3.2, where the backbone architectures used are also introduced. The experimental setup and experimental results are detailed in Sections 3.3 and 3.4, respectively.

3.1 Problem Formulation

A core challenge faced by the majority of individuals with ASD is an impaired ability to infer other people's emotions based on their facial expressions. With significant recent advances in machine learning, one potential approach to leveraging technology to assist such individuals to better recognize facial expressions and reduce the risk of possible loneliness and depression due to social isolation is the design of computer vision-driven facial expression recognition systems.

Most facial expression recognition systems either deal with single image inputs, or an entire video sequence at once. Even though high accuracy due to the additional context can be enjoyed by leveraging entire video sequences, the additional complexity and inference time required by such systems is sometimes not worth it, and in extreme cases unacceptable.

To address these needs, this chapter investigates a compromise between the fast inference speeds of single frame classification, and the high accuracy of including additional information in the inputs. The resulting novel deep time windowed convolutional neural network design (TimeConvNets), were designed for the purpose of real-time video facial

expression recognition while still maintaining high accuracy, through the use of spatiotemporal encodings of time windowed video frame sub-sequences.

To evaluate the proposed TimeConvNet design, we introduce a more difficult dataset called BigFaceX, composed of a modified aggregation of the extended Cohn-Kanade (CK+) [39], BAUM-1 [64], and the eNTERFACE [42] public datasets. Different variants of the proposed TimeConvNet design with different backbone network architectures were evaluated using BigFaceX alongside other network designs for capturing spatiotemporal information.

3.2 TimeConvNet Architecture

The proposed TimeConvNet, a deep time windowed convolutional neural network design, attempts to strike a balance between speed and accuracy by leveraging an efficient design for spatiotemporal encoding of time windowed video frame sub-sequences.

As discussed previously, temporal information is a key aspect in the classification of facial expressions due to their dynamic nature. Due to this property, dynamic video classification strategies use entire video sequences in order to maximize the amount of spatiotemporal information seen at once, at the cost of high computational complexity. By designing the proposed TimeConvNet to leverage time windowed sub-sequences, it allows us to reap the benefits of leveraging both spatial as well as transient facial cues, as dynamic video classification strategies do, while achieving significantly lower computational complexity when compared to such methods. The proposed TimeConvNet architecture explored in this study is shown in Figure 3.1. Using a streaming video sequence as input, a t-channel sub-sequence stack is constructed within a specific time window. This sub-sequence stack acts as the input to the TimeConvNet architecture, where a convolutional spatiotemporal encoding layer serves the purpose of capturing both spatial and transient characteristics within the time windowed sub-sequence stack through a set of learned convolutional filters within that layer. These learned convolutional filters, through training on time window sub-sequences of human expression video data, learns a diversity of spatiotemporal visual cues that well characterizes the different categories of emotion that we wish to predict. A key advantage of leveraging such a time windowed convolutional encoding layer early within the architectural design is that it allows for efficiently learning spatiotemporal embeddings without the need for computationally complex 3D convolutions, as well as without needing the entire video sequence.

In the next stage of the TimeConvNet architecture, the convolutional spatiotemporal encoding layer feeds into a backbone convolutional neural network architecture, where further hierarchical decomposition and encoding of the spatiotemporal representation from

the encoding layer at progressively higher levels of representational abstraction is performed for improved discrimination amongst the categories of facial expressions. Finally, a softmax layer is used to produce the final expression prediction. Given this design, the TimeConvNet is able to leverage the transient nuances of human facial expressions alongside spatial visual cues in a computationally efficient manner that facilitates for real-time scenarios.

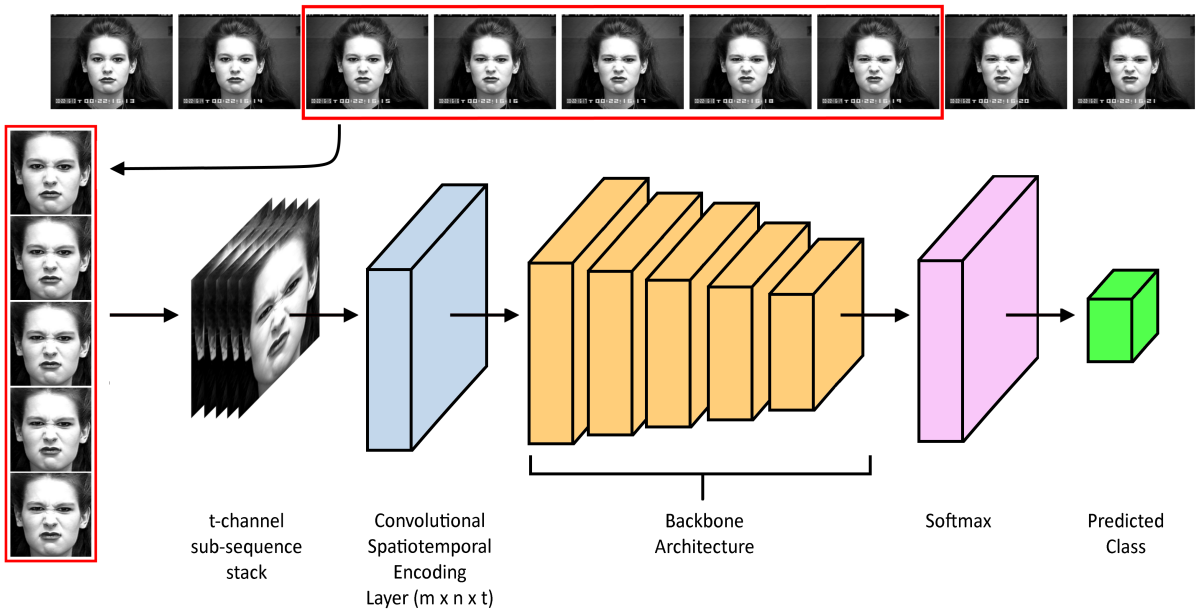

Figure 3.1: **Overview of the proposed TimeConvNet architecture.** Given a streaming video sequence, a t-channel sub-sequence stack is constructed within a particular time window. This sub-sequence stack is then passed into a convolutional spatiotemporal encoding layer, in which learned spatiotemporal filters capture the spatial as well as the transient characteristics exhibited in the facial expressions within the sub-sequence stack. A subsequent backbone convolutional neural network architecture is then leveraged to further decompose and encode the spatiotemporal representation from the convolutional spatiotemporal encoding layer at progressively higher levels of representational abstraction, followed by a softmax layer to produce the final facial emotion classification. By performing time windowed spatiotemporal encoding, TimeConvNets strike a better balance between speed and accuracy when compared to static single image classification strategies and dynamic video classification strategies.

3.2.1 Backbone Architectures

The latter stage of the TimeConvNet architecture involves the use of a backbone convolutional neural network architecture to perform more detailed decomposition of the inputs from the spatiotemporal encoding layer. We chose three different backbone architectures to evaluate in this study: i) mini-Xception [1], ii) ResNet20 [25], and iii) MobileNetV2 [50]. These three backbone architectures were chosen as they are widely used compact network architecture designs that provide a strong balance between efficiency and modeling performance.

The mini-Xception architecture [1] was originally introduced by Arriaga et al. in order to perform the tasks of face detection, gender classification, and emotion classification simultaneously in real-time using the FER-2013 [21] static face image dataset. Inspired by the Xception [8] architecture, the model leverages residual modules and depth-wise separable convolutions in order to achieve state-of-the-art level accuracy while reducing computational cost. As this mini-Xception model was already proven to perform well for the task of learning faces and their expressions, it was a natural choice to use in conjunction with time-windowed learning.

The use of residual connections [25] with deep learning has risen in popularity over the years due to their adaptability across almost all classification problems. These connections enable networks to learn faster and easier, with little additional cost to both architectural and computational complexity, and also provide a solution to the vanishing gradient problem due to their identity mapping options. Another benefit of these "shortcut connections" means that each consecutive layer should perform no worse than its previous layer, meaning that network architecture depth can be increased with less risk to the output model. As a result, residual network architecture designs have been shown to work well for the problem of FEC [24, 30, 67], resulting in our selection of the ResNet20 model as one of our backbone networks for TimeConvNets.

MobileNetV2 [50] introduces a neural network architecture specifically tailored towards mobile and resource constrained environments such as edge devices. It manages to retain comparable accuracy while reducing the number of operations significantly through the use of inverted residual connections with a linear bottleneck. As the main goal of our TimeConvNets is to use them on devices easily accessible to the average user, this mobile tailored model was also chosen to use as a backbone architecture.

3.3 Experimental Setup

3.3.1 Dataset

A critical factor in achieving strong facial emotion classification performance lies not just in the network architecture design, but also in quality of data in which the architecture is trained on. In order to properly evaluate our proposed TimeConvNet designs, a new dataset was needed that contained neither single image data points or entire video sequences. Naturally, famous facial expression datasets such as FER2013 [21] or AffectNet [44] could not be used, as these datasets are comprised only of single frame samples, making it infeasible to generate or retrieve any additional temporal information from them. Thus, our time windows could only be extracted from dynamic video datasets, such as the extended Cohn-Kanade (CK+) dataset [39].

Three public datasets were chosen - the extended Cohn-Kanade (CK+) dataset [39], the BAUM-1 dataset [64], and the eNTERFACE dataset [42]. These datasets were selected as they are relatively large, and consist of video data which were ideal for extracting video subsequences. In addition, they contained seven overlapping expression labels that we could use for our baseline classes, which are anger, disgust, fear, happiness, sadness, surprise, and neutral.

CK+

The CK+ dataset [39] is an improved version of the original Cohn-Kanade (CK) [29] database, with a 22 percent increase in video data and a 27 percent increase in the number of subjects. Participants ranged from 18 to 50 years of age, with a balanced mix of different genders and heritage. It contains 327 labelled emotion sequences across 123 subjects, with each sequence depicting the transition from the neutral face to the peak expression. One label is assigned to each video clip, which describes the peak expression. Each video was taken at a frame rate of 30 frames per second (FPS), with a resolution of either 640x490 or 640x480 pixels. Video clips that belong to the original CK database contained a timestamp at the bottom of each frame. As the CK+ dataset classes contain the additional emotion of contempt that is not common amongst other datasets, the samples pertaining to that emotion were not included in BigFaceX.

BAUM-1

Consisting of 1,184 video clips, with 31 different subjects performing a variety of spontaneous and elicited expressions, the BAUM-1 dataset [64] authors present it as the only non-posed database in literature at the time of publication that contains the six basic emotions in a non-English language. This dataset differs from the CK+ and eNTERFACE datasets in two major ways. First, the participants speak in Turkish rather than English, and second, it contains spontaneous reactions which are much closer to expressions in nature. Each video clip was recorded at 30 FPS, with a resolution of 720x576 pixels. 17 of the 31 subjects were female, and subjects ranged from 19-65 years of age, with a variety of hairstyles and facial hair. 3 of the subjects wore eyeglasses. This dataset also included several emotions not found in other datasets, such as the unsure, concentrating, and boredom emotions, and as such, the samples for those emotions are not used in BigFaceX.

eNTERFACE

The eNTERFACE'05 Audio-Visual Emotion Database [42] contained 34 male subjects and 8 female subjects from a mixture of countries. 13 of these individuals wore glasses, and 7 of them had a beard. The dataset consists of 1166 video clips, taken at a resolution of 720x576 pixels and 25 FPS. Due to the format of the data collection, all video sequences contain scripted responses, with acted expressions. One label is assigned to each video clip, which begins with a neutral expression and ends once the subject completes their scripted sentence.

BigFaceX Creation

Figure 3.2 gives a step-by-step explanation of how BigFaceX was created. First, for each video sequence in the CK+, BAUM-1, and eNTERFACE datasets, a window of frame size 5 was slid across the sequence with a stride of 1 and extracted. For each window, each image was cropped to the facial bounding box using Haar Cascades [58], in order to remove noise from various external factors such as clothing, background, or watermarks. The bounding boxes were extended by 10% on each side, as vanilla Haar Cascades can cause the sides of the face to be removed. Each image was then resized to a size of 48x48 pixels, normalized to the range of 0 to 1, and then merged together in the channel dimension, forming a final shape of 48x48x5. We labelled each sub-sequence with the original label of the video it was taken from.

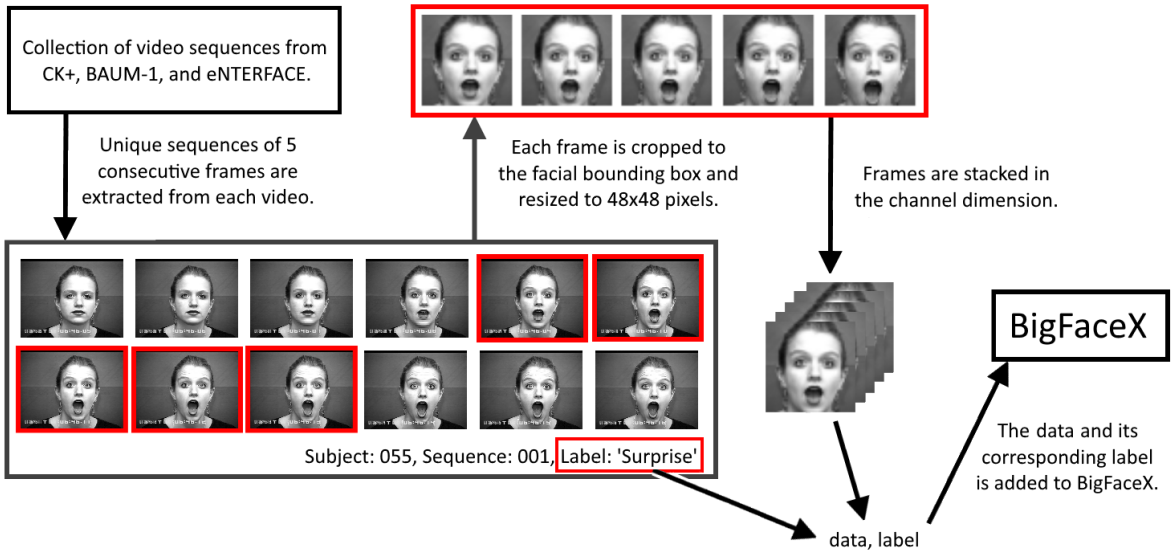

Figure 3.2: **The BigFaceX creation pipeline.** Unique sequences of consecutive frames are taken from each video in CK+, BAUM-1, and eNTERFACE. Each frame is cropped to the facial bounding box, and resized. Frames are stacked together and added to BigFaceX, with the same label as the original video clip.

Minor modifications were made when processing the BAUM-1 and eNTERFACE datasets, due to the fact that the initial neutral phase at the start of each video was much longer than in CK+. We chose to ignore the first 5 frames of each video in order to avoid creating large amounts of mislabelled data. Furthermore, for these two datasets, a stride of 2 was used instead of 1 due to the amount of frames in the video clip. This had the effect of increasing temporal width while maintaining the small window size. After all processing, the BigFaceX dataset contains 68,363 samples in total, and the class distribution is shown in Table 3.1. Example time windows from BigFaceX can be seen in Figure 3.3, where slight differences from frame to frame in each sequence can be observed. If classified individually, each frame could be ambiguous in their expression, but when used together, the expression is much more obvious (e.g. the first frame in row C of Figure 3.3 could easily be predicted as "anger" when used alone).

Table 3.1: **Distribution of the BigFaceX Dataset.** There are the most samples for the sad class, and least for the fear class.

Expression	Number of data points
Angry	8951
Disgust	8823
Fear	6069
Happy	12832
Sad	13870
Surprise	6197
Neutral	11621
Total	68363

3.3.2 Training Setup

In this study, a number of TimeConvNet architecture variants based on different backbone architecture were constructed. For all variants, 5-channel sub-sequence stacks were leveraged in order to reduce inference time while still providing a good characterization of transient facial behaviour. Each architecture variant was trained for 200 epochs using an initial learning rate of 1e−3, multiplied by 1e−1, 1e−2, 1e−3, and 0.5e−3 at epochs 81, 121, 161, and 181 respectively. Categorical cross-entropy loss was used with the Adam [31] optimizer. Data augmentation was applied to the inputs, including rotation, width and height shifts, zoom, and horizontal flips. For training data, we leveraged the proposed BigFaceX dataset, containing a total of 68,363 data points, and split it as follows: 70% for training, 10% for validation, and 20% for testing. All experiments were run using the Intel (R) Core (TM) i3-7100 3.90GHz x 4 CPU, and the GeForce RTX 2080 Ti GPU. We leveraged the Keras [7] library for this study.

3.3.3 Comparison Networks

To properly evaluate our TimeConvNet design, we also trained three other vanilla convolutional neural networks: i) a standard mini-Xception network architecture, ii) a 3D ResNet20-based deep convolutional neural network architecture, and iii) a (2+1)D ResNet20-based deep convolutional neural network architecture as suggested by Tran et al. [55].

The mini-Xception design was chosen to directly compare a model leveraging the TimeConvNet encoding layer with its vanilla version. While the TimeConvNet variant has an

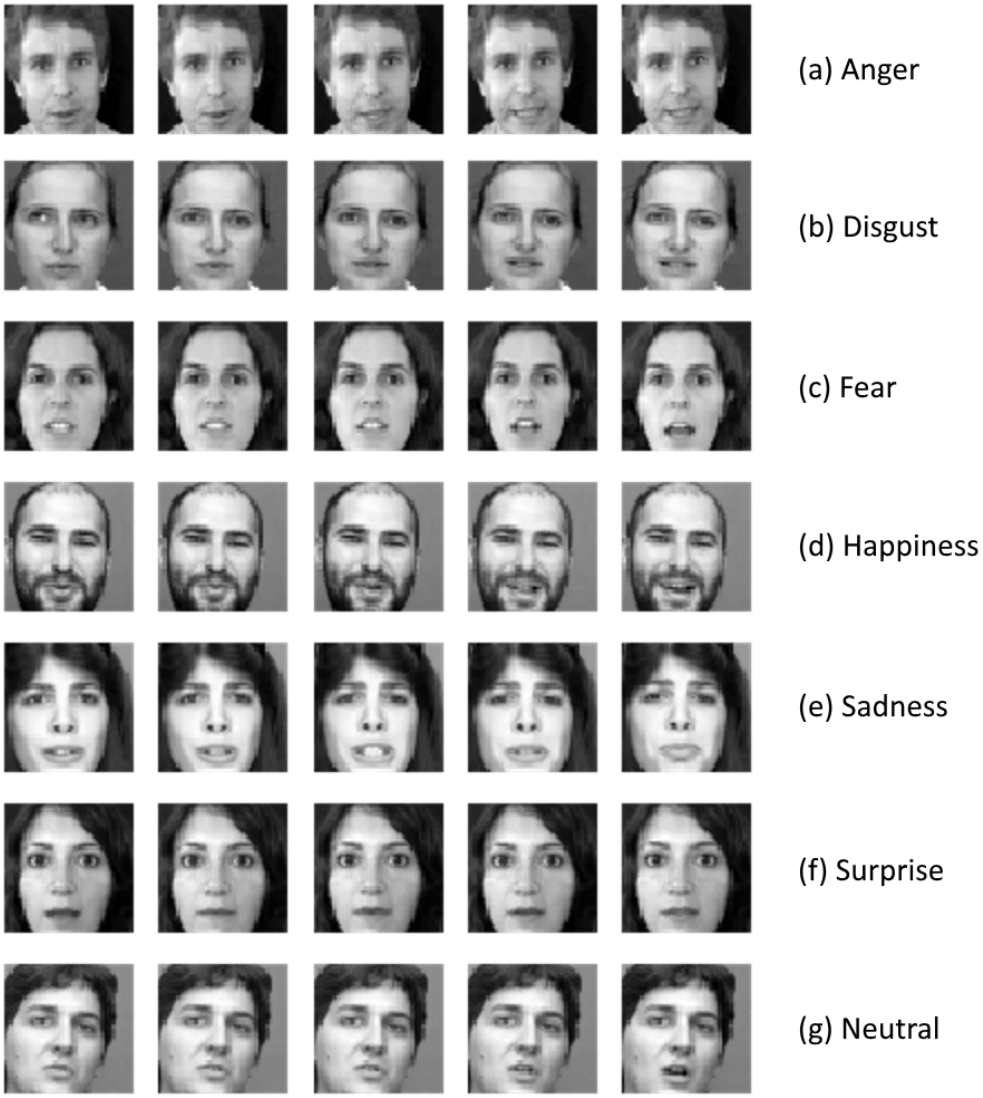

Figure 3.3: **Example facial expression data in BigFaceX.** Each row (left to right) represents a single 5-channel sub-sequence stack of frames. (a) Anger. (b) Disgust. (c) Fear. (d) Happiness. (e) Sadness. (f) Surprise. (g) Neutral.

input shape of 48x48x5, the vanilla version uses a 2D input of 48x48x1, meaning that we used only the last frame in each sample from BigFaceX for evaluation.

Next, to compare the TimeConvNet design against a straightforward 3D convolution,

we chose to use a 3D ResNet20-based neural network architecture, to observe if we could achieve 3D convolution level performance while avoiding the high complexities it is associated with. Additionally, this allows us to directly compare 3D convolution against time windowed convolution, as both backbone models leverage the ResNet20 architecture.

Lastly, we chose a design in between 2D and 3D convolution, known as (2+1)D convolution [55]. Tran et al. show that this architecture type is able to retain temporal relationships while lowering the number of parameters in the network when compared to 3D deep convolutional neural network architectures. It was included in order to analyze the balance between 2D and 3D convolution while also being compared to the TimeConvNet channel based inputs.

3.4 Experimental Results

To investigate the efficacy of the proposed TimeConvNet architecture design, we proposed and evaluated three different TimeConvNet variants using different backbone architectures: i) mini-Xception [1], ii) ResNet20 [25], and iii) MobileNetV2 [50], and compared them against the following network architecture designs: i) a standard mini-Xception network architecture, ii) a 3D ResNet20-based deep convolutional neural network architecture, and iii) a (2+1)D ResNet20-based deep convolutional neural network architecture as suggested by Tran et al. [55]. Note that since the standard mini-Xception network architecture leverages a single image, only the last frame from each sub-sequence in BigFaceX was used for evaluation instead.

The experimental results on the BigFaceX test set is shown in Table 3.2. The proposed TimeConvNet variant using the ResNet20 backbone achieved the highest top-1 accuracy (97.9%) while achieving significantly lower inference time (6.14ms) that is many folds lower than the 3D ConvNet and the (2+1)D ConvNet architectures. The fastest network architecture is the proposed TimeConvNet variant using the mini-Xception backbone, which was almost twice as fast as the ResNet20 variant (3.18ms) while still achieving higher accuracy than the the 3D ConvNet and the (2+1)D ConvNet architectures. Furthermore, the TimeConvNet variant with the mini-Xception backbone achieved approximately 10% higher accuracy than the standard mini-Xception network architecture, while having similar inference times. The TimeConvNet variant with the MobileNetV2 backbone provided the best balance between accuracy and inference time, but possesses significantly more parameters than other networks.

The validation accuracy training curves of each model are shown in Figure 3.4. The three TimeConvNet models have the highest accuracy, with the ResNet20 backbone model

Table 3.2: **TimeConvNet comparison verses vanilla models.** We compare in terms of accuracy, inference time, and parameter count. Accuracy was assessed on the BigFaceX test dataset. Inference times were averaged over 1000 runs. The best values for each column are shown in bold.

Network	Top 1 Accuracy	Inf. Time (ms)	Parameters
2D ConvNet (mini-Xception)	0.757	3.22	**58,423**
(2+1)D ConvNet (ResNet20)	0.848	52.70	523,357
3D ConvNet (ResNet20)	0.851	35.64	808,775
TimeConvNet (mini-Xception)	0.855	**3.18**	58,711
TimeConvNet (ResNet20)	**0.979**	6.14	274,535
TimeConvNet (MobileNetV2)	0.923	5.64	2,267,527

leading, followed by MobileNetV2 and then mini-Xception. The 3D ConvNet model and (2+1)D ConvNet models are next at roughly 75 percent, and finally the baseline 2D ConvNet shows the worst performance of them all. For the models that do not use mini-Xception as a backbone, there is a sudden jump in accuracy at epoch 81 which corresponds to the first learning rate reduction. We trained all models for 200 epochs each, but only show up to epoch 110 in Figure 3.4 as the networks do not show significant improvement past this point even after the learning rate changes in later epochs.

Confusion matrices are shown in Figure 3.5, where it can be seen that the ResNet20 backbone TimeConvNet has the least amount of error. A few interesting observations can be noted, especially related to the ambiguity of certain expressions and the overlap between facial features such as fear and sadness. Firstly, fear and surprise seem to be the classes with the highest prediction error, where many of the samples are incorrectly predicted to be the other class. If we consider what a human face looks like during these emotional states, it is understandable that there would be some level of ambiguity between the two. A similar scenario occurs between the expressions anger and disgust, anger and sadness, or sadness and neutral. Second, for most of the models, happiness is one of the easiest classes to predict successfully, likely due to the unique facial features associated with said expression. However, if other expressions were included in a future study such as "amused," it is hypothesized that the accuracy of the happiness class would decrease.

The results for the introduced TimeConvNets appear to be quite promising for the task of real-time video facial expression classification, which could potentially be very beneficial for assisting with ASD in real-time emotion recognition to improve social interactions.

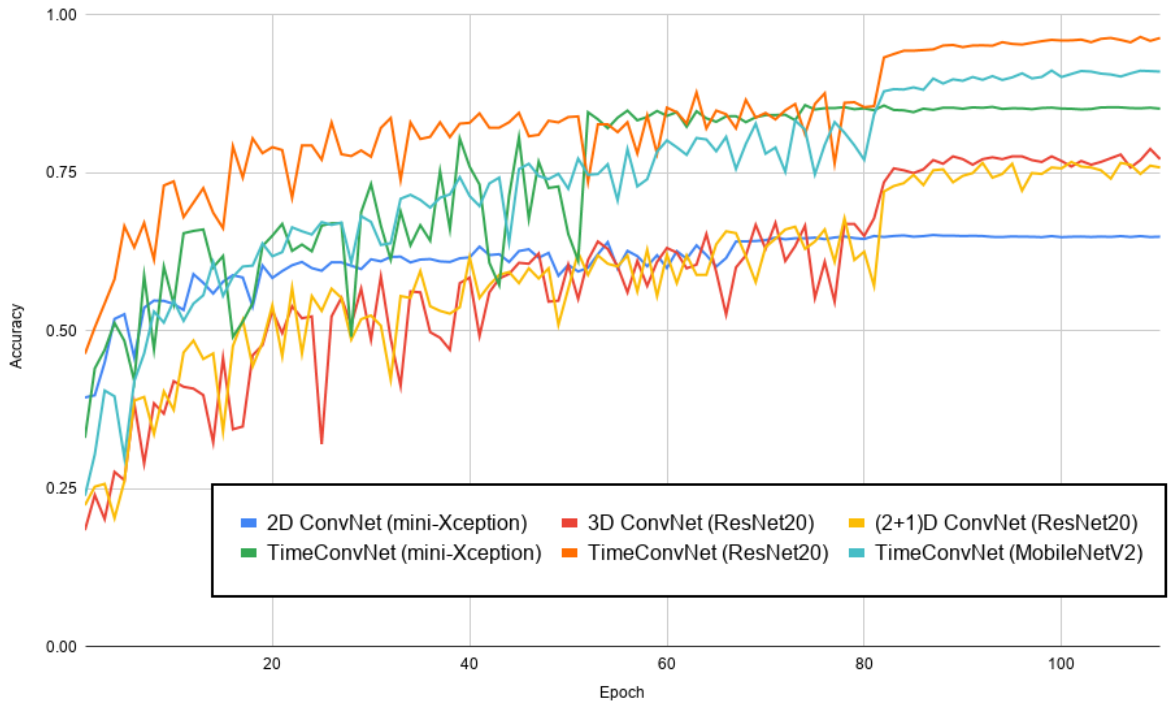

Figure 3.4: **Validation accuracy of the six models trained using BigFaceX.** Graph only shows up to epoch 110 as models do not improve further. Models without mini-Xception as a backbone show a performance jump when the learning rate is changed at epoch 81. ResNet20 has the highest final accuracy, with mini-Xception having the lowest. Best viewed in colour.

Given the promising results of TimeConvNets and the BigFaceX dataset, there are a number of considerations that must initially be taken if leveraged within a real-time emotion recognition system.

First, it is important to consider that in addition to the time required to run inference using the network, there is also significant processing overhead associated with frame processing as well as face detection. Based on our empirical evaluation, given that the TimeConvNets have inference times of ∼3-6 ms, plus additional processing overhead of around ∼35 ms, we can achieve a facial expression classification pipeline that takes around 40 ms per video frame, which translates to approximately 25 frames per second (FPS).

Second, it is important to note that BigFaceX, like most facial expression datasets,

2D ConvNet (mini-Xception)

Label \ Prediction	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	69.43%	6.97%	6.45%	1.68%	11.17%	2.42%	1.88%
Disgust	3.25%	86.48%	2.44%	2.78%	2.89%	0.44%	1.72%
Fear	6.44%	6.84%	64.71%	1.17%	12.79%	4.07%	3.99%
Happy	2.60%	3.80%	1.55%	85.43%	1.61%	1.69%	3.33%
Sad	4.77%	5.13%	4.19%	0.81%	77.74%	2.11%	5.27%
Surprise	6.49%	2.45%	12.75%	3.31%	8.63%	64.08%	2.29%
Neutral	2.47%	2.93%	0.94%	6.90%	7.97%	2.62%	76.17%

TimeConvNet (mini-Xception)

Label \ Prediction	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	81.91%	3.92%	5.74%	0.93%	5.02%	1.45%	1.03%
Disgust	1.28%	92.35%	1.30%	1.09%	2.29%	0.24%	1.45%
Fear	3.03%	4.51%	83.21%	0.79%	4.53%	2.11%	1.81%
Happy	1.27%	1.95%	0.98%	90.85%	1.33%	1.48%	2.14%
Sad	2.13%	1.78%	3.66%	1.17%	85.31%	1.98%	3.97%
Surprise	2.36%	1.57%	7.29%	2.07%	4.28%	81.14%	1.31%
Neutral	1.94%	0.79%	1.70%	2.71%	3.03%	1.26%	88.56%

3D ConvNet (ResNet20)

Label \ Prediction	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	75.53%	4.23%	3.63%	3.30%	8.31%	2.76%	2.23%
Disgust	2.20%	86.41%	2.21%	2.06%	4.28%	1.39%	1.44%
Fear	4.61%	4.02%	74.28%	2.95%	7.78%	3.71%	2.65%
Happy	1.28%	1.01%	0.53%	92.29%	0.88%	1.29%	2.73%
Sad	2.37%	1.72%	2.38%	0.98%	88.12%	1.57%	2.86%
Surprise	4.53%	1.15%	5.10%	6.28%	5.87%	74.18%	2.89%
Neutral	0.59%	0.39%	0.67%	3.73%	4.72%	0.36%	89.54%

TimeConvNet (ResNet20)

Label \ Prediction	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	97.90%	0.50%	0.50%	0.28%	0.58%	0.17%	0.07%
Disgust	0.27%	99.14%	0.24%	0.08%	0.14%	0.07%	0.07%
Fear	0.31%	0.43%	98.20%	0.13%	0.63%	0.25%	0.05%
Happy	0.23%	0.16%	0.21%	98.83%	0.16%	0.30%	0.10%
Sad	0.31%	0.08%	0.22%	0.09%	98.75%	0.35%	0.19%
Surprise	0.16%	0.11%	0.61%	0.23%	0.48%	97.90%	0.50%
Neutral	0.03%	0.08%	0.08%	0.28%	0.69%	0.06%	98.78%

(2+1)D ConvNet (ResNet20)

Label \ Prediction	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	81.16%	3.23%	2.42%	2.69%	6.00%	2.42%	2.07%
Disgust	4.61%	87.15%	1.81%	2.15%	2.56%	0.79%	0.92%
Fear	6.84%	5.49%	70.97%	1.96%	6.44%	5.11%	3.20%
Happy	2.42%	0.94%	0.69%	89.67%	1.04%	2.56%	2.67%
Sad	3.23%	1.82%	2.43%	1.02%	85.36%	2.05%	4.07%
Surprise	5.24%	0.77%	3.79%	3.73%	5.39%	78.28%	2.79%
Neutral	1.11%	0.89%	0.40%	2.41%	3.56%	0.48%	91.15%

TimeConvNet (MobileNetV2)

Label \ Prediction	Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral
Anger	92.56%	1.46%	1.97%	0.39%	2.65%	0.49%	0.48%
Disgust	0.87%	96.02%	0.90%	0.36%	0.86%	0.28%	0.70%
Fear	2.34%	0.99%	92.22%	0.44%	1.70%	1.07%	1.24%
Happy	1.28%	0.81%	0.56%	94.46%	0.83%	0.95%	1.11%
Sad	0.89%	0.86%	1.78%	0.40%	93.21%	1.43%	1.43%
Surprise	1.45%	0.53%	3.81%	0.79%	3.29%	89.48%	0.65%
Neutral	1.24%	0.22%	0.59%	1.25%	1.52%	0.07%	95.11%

Figure 3.5: **Confusion matrices of the models on the BigFaceX dataset.** A visualization of the distribution of the model predictions. Darker colours are better along the diagonal, but indicate a higher rate of incorrect classification and overlap between classes if seen elsewhere.

exhibit class imbalances, with noticeably more samples for more frequently occurring classes such as happy or sad versus less frequently encountered expressions such as fear or surprise. This can be seen in Table 3.1, and can potentially lead to biases in the predictions made by the networks trained on BigFaceX. One area of improvement in the future is a larger effort to re-balance BigFaceX.

Third, while the BigFaceX dataset attempts to mimic real-world scenarios much better than datasets such as FER2013, as it includes non-peak expressions as well as spontaneous reactions, it is still unable to account for all possibilities, such as variation in mouth movements during speech using different languages, or varying amounts of facial occlusion. We did attempt to account for this in our design choices, such as by including the BAUM-1 dataset, as it includes non English speakers which should help the models generalize to

other languages, but it is likely that a larger dataset containing languages from around the world is required. Certain amounts of facial occlusion is accounted for as datasets such as eNTERFACE include subjects with facial hair or eyeglasses, but more unique features such as piercings or tattoos may still cause inaccurate detection and prediction. Future work should include more variation in subjects, occlusion types, facial markings, and brightness levels, in order to better represent a real-life scenario that a user would likely experience.

3.5 Summary

In this chapter, a novel design for an efficient spatiotemporal encoding method for facial expression classification is investigated. As an intermediary design between the speeds of single frame classification and the accuracy of video classification, we hypothesized that a customizable balance between the two could be achieved.

Based on the experimental results, it can be seen that the models using the proposed TimeConvNet designs are not only more accurate, but also take much less time for inference when compared to the (2+1)D and 3D convolution-based networks. Based on a quantitative analysis, the model with the most optimal trade-off between accuracy, inference time, and model size is the ResNet20 TimeConvNet network, which is able to achieve accurate real-time performance while being relatively lightweight.

Even with the compact and lightweight performance boost that the TimeConvNet design offers, it is unlikely that it is the most efficient network design possible. Human based design principles, while backed with logic and precedent, are unable to perform the fine tuning of network modules efficiently and in a reasonable time frame. Thus, neural architecture search (NAS) strategies, or machine based neural network design algorithms are required in order to come up with new and improved tweaks to any existing models. In the next chapter, a human-machine collaborative design strategy tailored specifically for the task of human facial expression classification is leveraged, in order to find the optimal balance between accuracy and architectural and computational complexity.

Chapter 4

EmotionNet Nano

In this chapter, an human-machine collaborative design strategy tailored specifically for the task of human facial expression classification is investigated. Section 4.2 provides a detailed description of the two phase design strategy used for the creation of EmotionNet Nano and its variants. Experimental setup is discussed in Section 4.3, where the dataset used and training protocols can be found, and Section 4.4 contains the results of the experiments.

4.1 Problem Formulation

Even though the performance of deep learning-based FEC systems continue to rise, widespread deployment of such systems is limited, with one of the biggest hurdles being the high architectural and computational complexities of the deep neural networks that drive such systems. This hurdle is particularly limiting for real-time embedded scenarios, where low latency operation is required on the low-cost embedded devices. For example, in the area of assistive technologies for improving quality of life, the majority of individuals using such technologies are unwilling to carry large, bulky, and expensive devices with them during their daily lives, as that would be a big hindrance that limits their ability to leverage the technologies in a seamless manner. As such, the assistive devices must leverage small, low-cost, embedded processors, yet provide low latency to enable real-time feedback to the user.

The previous chapter discussed the use of TimeConvNets as a way to perform more efficient facial expression classification while maintaining a low architectural and computational complexity. While human network architects are able to create compact and efficient

networks designs by hand, these methods are typically time consuming and often lead to lower performance as a result. Interestingly, the addition of machines to form a human-machine collaborative design strategy has shown recent success in the design of highly compact deep CNNs [61], by coupling principled network design prototypes with machine driven design exploration, constrained by human-specified design requirements.

In this chapter, we explore the efficacy of leveraging a human-machine collaborative design strategy that leverages human experience and ingenuity with the raw speed and meticulousness of machine driven design exploration, in order to find the optimal balance between accuracy and architectural and computational complexity for the specific task of human facial expression classification. The resulting deep neural network architecture, which we call **EmotionNet Nano**, is specifically tailored for real-time embedded facial expression recognition and created via a two phase design strategy. We present two variants of EmotionNet Nano, each with a different trade-off between accuracy and complexity, and evaluate both variants on the CK+ [39] benchmark dataset against state-of-the-art facial expression classification networks.

4.2 Human Machine Collaborative Design Strategy

EmotionNet Nano was created using a two phase design strategy, the first of which leveraged residual architecture design principles to capture the complex nuances of facial expressions. Next, machine-driven design exploration was employed to generate the final tailor-made architecture design that achieves high architectural and computational efficiency while maintaining a high performance.

4.2.1 Principled Network Design Prototyping

In the first design stage, an initial network design prototype, φ, was designed using human-driven design principles in order to guide the subsequent machine-driven exploration design stage. In this study, the initial network design prototype of EmotionNet Nano leveraged residual architecture design principles [25], as it was previously demonstrated to achieve strong performance on a variety of recognition tasks. More specifically, the presence of residual connections within a deep neural network architecture have been shown to provide a good solution to both the vanishing gradient and curse of dimensionality problems. Residual connections also enable networks to learn faster and easier, with little additional cost to architectural or computational complexity. Additionally, as the network architecture depth increases, each consecutive layer should perform no worse than its previous layer

due to the identity mapping option. As a result, residual network architecture designs have been shown to work well for the problem of FEC [24, 30, 67].

In this study, the final aspects of the initial network design prototype, φ, consists of an average pooling operation followed by a fully connected softmax activation layer to produce the final expression classification results. The final macroarchitecture and microarchitecture designs of the individual modules and convolutional layers of the proposed EmotionNet Nano were left to the machine-driven design exploration stage to design in an automatic manner. To ensure a compact and efficient real-time model catered towards embedded devices, this second stage was guided by human-specified design requirements and constraints targeting embedded devices possessing limited computational and memory capabilities.

4.2.2 Machine Driven Design Exploration

Following the initial human-driven network design prototyping stage, a machine-driven design exploration stage was employed to determine the macroarchitecture and microarchitecture designs at the individual module level to produce the final EmotionNet Nano. In order to determine the optimal network architecture based on a set of human defined constraints, generative synthesis [61] was leveraged for the purpose of machine-driven design exploration. Defined in Equation 4.1, we can formulate generative synthesis as a constrained optimization problem, where the goal is to find a generator $\mathcal{G}$ that, given a set of seeds $\mathcal{S}$, can generate networks $\{\mathcal{N}_s | s \in \mathcal{S}\}$ that maximize a universal performance function $\mathcal{U}$ while also satisfying constraints defined in an indicator function $1_r(\cdot)$,

$$\mathcal{G} = \max_{\mathcal{G}} \mathcal{U}(\mathcal{G}(s)) \text{ subject to } 1_r(\mathcal{G}(s)) = 1, \forall s \in \mathcal{S} \tag{4.1}$$

As such, given a human-defined indicator function $1_r(\cdot)$ and an initial network design prototype φ, generative synthesis is guided towards learning generative machines that generate networks within the human-specified constraints.

An important factor in leveraging generative synthesis for machine-driven design exploration is to define the operational constraints and requirements based on the desired task and scenario in a quantitative manner via the indicator function $1_r(\cdot)$. In this study, in order to learn a compact yet highly efficient facial expression classification network architecture, the indicator function $1_r(\cdot)$ was set up such that: i) accuracy $\geq 92\%$ on CK+ [39], and ii) network architecture complexity $\leq$ 1M parameters. These constraint values were chosen to explore how compact a network architecture for facial expression classification

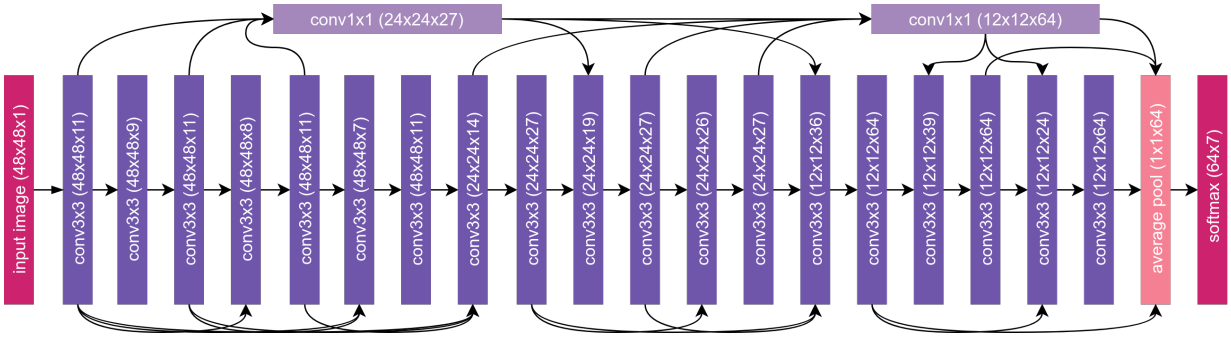

Figure 4.1: **The EmotionNet Nano Architecture.** The network architecture exhibits high macroarchitecture and microarchitecture heterogeneity, customized towards capturing deep facial features. Furthermore, the network architecture exhibits selective long-range connectivity throughout the network architecture. The number of channels per layer are based on EmotionNet Nano-B.

can be while still maintaining sufficient classification accuracy for use in real-time embedded scenarios. As such, we use the accuracy of Feng & Ren [19] as the reference baseline for determining the accuracy constraint in the indicator function.

4.2.3 Final Architecture

The final network architecture of the proposed EmotionNet Nano is shown in Figure 4.1, generated after both design phases. A number of notable characteristics of the proposed EmotionNet Nano network architecture design are worth discussing as they give insights into architectural mechanisms that strike a strong balance between complexity and accuracy.

First, the macroarchitecture and microarchitecture heterogeneity of the network allows it to achieve high efficiency even with a low number of parameters. Unlike hand-crafted architecture designs, the macroarchitecture and microarchitecture designs within the EmotionNet Nano network architecture as generated via machine-driven design exploration differ greatly from layer to layer. For instance, there are a mix of convolution layers with varying shapes and different number of channels per layer depending on the needs of the network. As shown in Figure 4.1, there are a greater number of channels needed as the sizes of feature maps decrease.

The benefit of high microarchitecture and macroarchitecture heterogeneity in the EmotionNet Nano network architecture is that it enables different parts of the network archi-

tecture to be tailored to achieve a very strong balance between architectural and computational complexity while maintaining model expressiveness in capturing necessary features. The architectural diversity in EmotionNet Nano demonstrates the advantage of leveraging a human-collaborative design strategy as it would be difficult for a human designer, or other design exploration methods to customize a network architecture to the same level of architectural granularity.

Secondly, EmotionNet Nano exhibits selective long range connectivity throughout the network architecture. The use of long range connectivity in a very selective manner enables a strong balance between model expressiveness and ease of training, and computational complexity. Most interesting and notable is the presence of two densely connected 1×1 convolution layers that take in outputs from multiple 3×3 convolution layers as input, with its output connected farther down at later layers. Such a 1×1 convolution layer design provides dimensionality reduction while retaining salient features of the channels through channel mixing, thus further improving architectural and computational efficiency while maintaining strong model expressiveness.

4.3 Experimental Setup

4.3.1 Dataset

To evaluate the efficacy of the proposed EmotionNet Nano, we examine the network complexity, computational cost and classification accuracy against other facial expression classification networks on the CK+ [39] dataset, which is the most extensively used laboratory-controlled FEC benchmark dataset [34, 49].

The Extended Cohn-Kanade (CK+) [39] dataset contains 593 video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age with a variety of genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak expression, recorded at 30 frames per second (FPS) with a resolution of either 640x490 or 640x480 pixels. Out of these videos, 327 are labelled with one of seven expression classes, anger, contempt, disgust, fear, happiness, sadness, and surprise. The CK+ database is widely regarded as the most extensively used laboratory-controlled FEC database available, and is used in the majority of facial expression classification methods [34, 49]. Figure 4.2 shows that the CK+ dataset has good diversity for each expression type, which is important from an evaluation perspective. However, as the CK+ dataset does not provide specific training, validation, and test set splits, a mixture of splitting techniques

can be observed in literature. For experimental consistency, we adopt the most common dataset creation strategy where the last three frames of each sequence is extracted and labeled with the video label [34]. In this study, we performed subject-independent 10-fold cross validation on the resulting 981 facial expression images.

Figure 4.2: **Diversity of expressions in the CK+ dataset.** Example faces for each expression type in CK+ is shown. Contempt not included as relevant subjects did not give publication consent.

4.3.2 Training Setup

EmotionNet Nano was trained for 200 epochs using an initial learning rate of 1e−3, multiplied by 1e−1, 1e−2, 1e−3, and 0.5e−3 at epochs 81, 121, 161, and 181 respectively. Categorical cross-entropy loss was used with the Adam [31] optimizer. Data augmentation was applied to the inputs, including rotation, width and height shifts, zoom, and horizontal flips. Following this initial training, we leveraged a machine-driven exploration stage to fine tune the network specifically for the task of FEC. Training was performed using a GeForce RTX 2080 Ti GPU. The Keras [7] library was leveraged for this study.

4.4 Experimental Results

4.4.1 State-of-the-art Performance Comparison

Two variants of EmotionNet Nano were created to examine the different trade-offs between architectural and computational complexity and accuracy. In order to demonstrate the efficacy of the proposed models in a quantitative manner, we compare the performance of both variants against state-of-the-art facial expression classification networks introduced in literature, shown in Table 4.1. It can be observed that both EmotionNet Nano-A and Nano-B networks achieve strong classification accuracy, with EmotionNet Nano-A in particular achieving comparable accuracy with the highest-performing state-of-the-art networks that are more than a magnitude larger. While EmotionNet Nano-B has lower accuracy than the highest-performing networks, it is still able to achieve comparable accuracy as [19] while being three orders of magnitude smaller. A more detailed discussion of the performance comparison will be provided in the next section; overall, it can be observed that both EmotionNet Nano variants provide the greatest balance between accuracy and complexity, making it well-suited for embedded scenarios.

4.4.2 Speed and Energy Efficiency

We also perform a speed and energy efficiency analysis, shown in Table 4.2, to demonstrate the efficacy of EmotionNet Nano in real-time embedded scenarios. Here, an ARM v8.2 64-Bit RISC embedded processor was used for evaluation. Referring to Table 4.2, both EmotionNet Nano variants are able to perform inference at >25 FPS and >70 FPS on the tested embedded processor at 15W and 30W respectively, which more than fulfills a real-time system constraint. In terms of energy efficiency, both EmotionNet Nano

Table 4.1: **EmotionNet Nano networks compared against state-of-the-art on the CK+ dataset.** We report 10-fold cross-validation average accuracy on the CK+ dataset with 7 classes (anger, contempt, disgust, fear, happiness, sadness, and surprise).

Method	Params (M)	Accuracy (%)
Ouellet [47]	58	94.4
Feng & Ren [19]	332	92.3
Wang & Gong [59]	5.4	97.2
Otberdout et al. [46]	11	98.4
EmotionNet Nano-A	0.232	97.6
EmotionNet Nano-B	0.136	92.7

variants demonstrated high power efficiency, with the Nano-B variant running at 5.29 images/sec/watt on the embedded processor.

Table 4.2: **EmotionNet Nano Speed and Energy Efficiency.** All metrics are computed on an ARM v8.2 64-Bit RISC embedded processor at different power levels.

	15W		30W	
Model	**FPS**	$[\frac{\textbf{images/s}}{\textbf{watt}}]$	**FPS**	$[\frac{\textbf{images/s}}{\textbf{watt}}]$
EmotionNet Nano-A	25.8	1.72	70.1	2.34
EmotionNet Nano-B	**32.8**	**2.19**	**72.9**	**2.43**

4.4.3 Limitations

The distribution of expressions in CK+ is unequal, which results in an unbalanced dataset both for training and testing. The effects of this are prevalent when classifying the contempt or fear expressions, both of which are underrepresented in CK+ (e.g. there are only 18 examples of contempt, whereas there are 83 examples of surprise). Due to the nature of human facial expressions, similarities between expressions do exist, but the networks are generally able to learn the high-level distinguishing features that separate one expression from another. However, incorrect classifications can still occur, as shown in Figure 4.3, where a "disgust" expression is falsely predicted to be "anger."

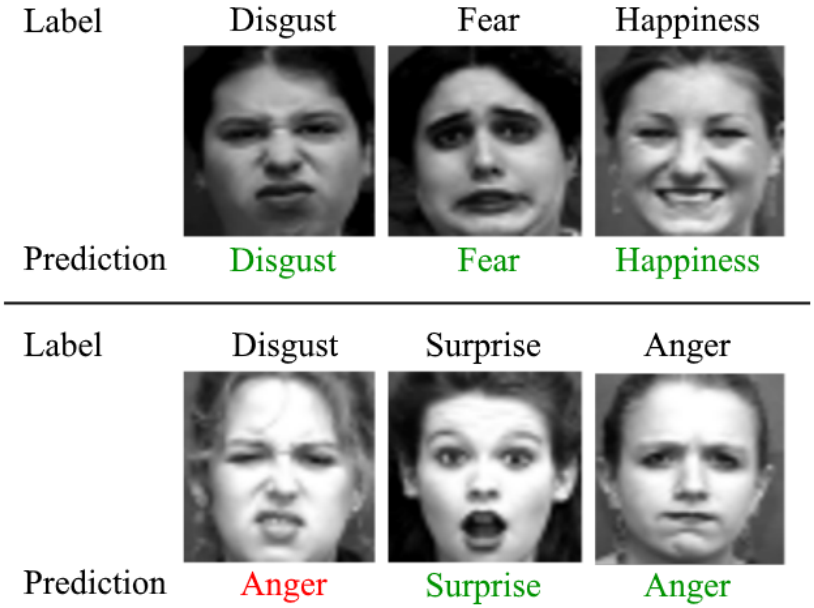

Figure 4.3: **Example expression predictions of faces in the CK+ dataset using EmotionNet Nano-A.** Five of the faces are classified correctly, indicated in green, with an example of a misclassified expression (disgust), shown in red.

4.5 Summary

In this chapter, we introduced EmotionNet Nano, a highly efficient deep convolutional neural network design tailored for facial expression classification in real-time embedded scenarios by leveraging a human-machine collaborative design strategy. By leveraging a combination of human-driven design principles and machine-driven design exploration, the EmotionNet Nano architecture design possesses several interesting characteristics (e.g., architecture heterogeneity and selective long-range connectivity) that makes it tailored for real-time embedded usage. Two variants of the proposed EmotionNet Nano network architecture design were presented, both of which achieve a strong balance between architecture complexity and accuracy while illustrating performance trade-offs at that scale. Using the CK+ dataset, we show that the proposed EmotionNet Nano can achieve comparable accuracy to state-of-the-art facial expression classification networks (at 97.6%) while possessing a significantly more efficient architecture design (possessing just 232K parameters). Furthermore, we demonstrated that EmotionNet Nano can achieve real-time inference speed on an embedded processor at different power levels, thus further illustrating its suitability

for real-time embedded scenarios.

The combination of EmotionNet Nano with the TimeConvNet architecture design is also possible, and is in fact quite desirable for the task of expression classification. Facial expressions are highly dynamic and transient in nature [24], meaning that information about the previous expression is valuable when predicting the current expression. Therefore, the retention of temporal information can lead to increased performance, at the expense of computational complexity. Investigating this trade-off between computational complexity and improved performance when leveraging temporal information would be worthwhile. In the next chapter, we introduce a real-world system that leverages these compact networks in order to assist individuals with ASD with the task of expression recognition. The Augmented-reality Expression Guided Interpretation System, or AEGIS for short, is able to run in real-time on edge devices such as mobile phones, and provides intuitive feedback to the user via augmented reality images on the screen, providing them with instant and seamless feedback which they can use to enhance their social interactions.

Chapter 5

AEGIS

Individuals living with ASD can have an increased difficulty in interpreting and understanding facial expressions and their corresponding emotions [32, 63], which can cause problems when creating and sustaining meaningful, positive relationships, leading to troubles integrating into society and a higher prevalence of depression and loneliness. However, studies have shown that after the use of assistive technology (AT) software, participants with ASD showed improvement on facial emotion recognition for emotions shown in the software, as well as emotions not included in the software [32]. This finding leads us to believe that the existence of a real-time expression classification system could help alleviate some of these issues faced by peoples with ASD, and thus we propose AEGIS (Augmented-reality Expression Guided Interpretation System), designed in order to assist these individuals learn to better identify expressions and improve their social experiences. AEGIS is a multimodal augmented reality (AR) assistive technology system deployable on a wide range of user devices including tablets, smartphones, video conference systems, and smartglasses, showcasing its extreme flexibility in a variety of use cases, to allow integration into daily life with ease. AEGIS leverages the use of computer vision and deep convolutional neural networks in order to achieve accurate yet real-time performance, granting the user a seamless, intuitive experience that can assist them in their societal interactions. More specifically, it leverages a TimeConvNet [33] network design from Chapter 3 as the initial prototype, and can easily support an EmotionNet Nano design mentioned in Chapter 4 as well.

5.1 Approach

AEGIS uses a device with a screen and a video streaming camera, which could range from a handheld device such as a mobile phone or tablet, to wearable technology such as smartglasses. The AEGIS software would be installed on the device, and would run in real-time. The user would face the camera towards the desired area, which would be the faces of the people they are planning on talking to, while watching the screen of the device themselves. As shown in Figure 5.1, AEGIS takes in a streaming video camera source as input and processes each real-world frame before automatically passing them to our novel deep convolutional time windowed neural network. Within the model, sets of learned convolutional filters leverage both spatial and temporal information in order to provide an accurate expression prediction. The original real-world camera frame is then augmented with the corresponding emoji, and finally shown to the user on the device screen. This entire process takes milliseconds to complete, allowing for a seamless and immersive experience where the user sees augmented real-world frames with emojis overlaid on top of each person in real-time.

5.2 Design Options

5.2.1 Neural Network Design

As we are aiming for a real-time system, we must choose a good balance between performance accuracy and inference speed. We use a novel deep time windowed convolutional neural network design discussed in Chapter 3 which we call a TimeConvNet. Given a streaming video sequence, if there is a face detected in the frame, we process said frame by cropping it to the facial bounding box and then resizing to a size of 48x48 pixels. We then add these processed frames into a first-in-first-out queue of length t. By doing so, we create a dynamic t-channel video sub-sequence, which we then stack together in the channel dimension (creating an input of shape 48x48xt), and provide this stack as the input to our TimeConvNet system, where a set of learned convolutional filters within the convolutional spatiotemporal encoding layer captures both the spatial and temporal attributes of the time window. This time windowed method allows us to capture the transient nuances of dynamic facial expressions without the use of computationally complex 3D convolutions or needing an entire video sequence. The spatiotemporal encoding layer is fed into a backbone convolutional neural network architecture, where progressively higher levels of abstraction are performed for improved discrimination amongst the facial expression categories. Fi-

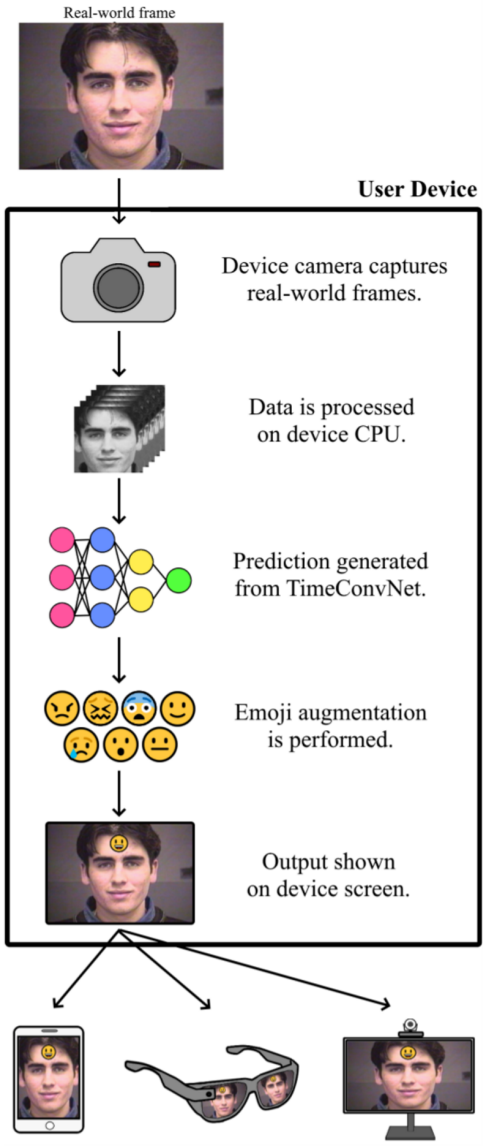

Figure 5.1: **System overview of AEGIS.** Given a streaming video camera as input, each real-world frame is processed on the device CPU, then passed to our novel TimeConvNet. Emoji augmentation is performed based on the network prediction, and the output is shown on the user device. AEGIS is deployable on a variety of devices, including smartphones, smartglasses, and video conference systems.

nally, a softmax layer determines the prediction, which is one of seven classes (anger, fear, disgust, happiness, sadness, surprise, and neutral).

Based on the results given in Chapter 3, Table 3.2, and due to the low latency requirement for AEGIS, as users would want expression feedback in real-time for a seamless experience, we ultimately decided on the ResNet20 backbone architecture as it provided the best balance compared to the other models. We found that by using the ResNet20 TimeConvNet in our system, we could achieve a run time frame rate of roughly 25 FPS, even including all image processing, inference, and augmentation steps, which can allow for smooth social interactions.

5.2.2 Visualization Choices

Naturally, the way we present the information to the user is a key factor in how well they understand the given message. In order to provide the user with an immersive experience, the expression information must be presented in a way such that the user can understand the information unambiguously and without thought. The work presented in [63] reports that 50 percent of individuals diagnosed with ASD are highly visual-oriented and possess strong visual-spatial abilities, and additional studies have also found that these individuals respond much better to visual based stimuli when compared to auditory based stimuli. Inspired by these findings, and motivated by the speed and clarity requirements, we designed our visualization choices to best suit these visual-based needs.

The first option we thought of was using a coloured outline, which would essentially show each person in the frame with a glowing outline of a certain colour, based on their predicted emotion. However, we realized that it might be difficult to associate certain colours to certain emotions, as it may vary for each individual. Also, there might not exist a colour that most accurately describes an emotion - for example, anger might be thought of as red, but what about emotions such as surprise or happiness? Thus, we decided on using emojis as our visualization choice, where we would show an emoji floating over each person's head. The reasoning behind this choice was because of their growing popularity among people, both young and old, and also due to their expressiveness and ability to immediately convey certain tones. This came with its own set of challenges, but we managed to find emojis that corresponded well to each of the seven emotions we planned on showing. An example of what the augmented reality scenes would look like are shown in Figure 5.2. On the left side, the "surprise" expression is misclassified as "fear," but on the right, the system correctly recognizes the "sadness" expression.

For those individuals who prefer a text-based system, a text overlay with a predic-

Figure 5.2: **Assistive technology for Autistic Spectrum Disorder.** Example of how EmotionNet Nano can be leveraged to assist individuals with Autistic Spectrum Disorder to better infer emotional states from facial expressions during social interactions in the form of augmented reality.

tion confidence percentage is also a possible design. In literature, some real-time FEC approaches tend to show their prediction results with labelled bounding boxes around the individual faces [1, 60], but for visual-oriented users, this method is suboptimal, as an emoji is able to convey a message tone much faster than a word can, and are generally much more intuitive. Thus, for initial prototypes of the system, we chose to use the emoji-based visual-oriented approach.

5.2.3 Hardware Choices

The requirements of the device used in AEGIS was one that was simple to use, easy to set up, and also non-obstructive towards the everyday life of the user. Camera input was needed, and a screen was required to display the output. These simple requirements allow AEGIS to be deployable on a variety of devices, including tablets and the increasingly popular smartphone. The system is also usable on video conference systems and smart-glasses, to account for the fact that certain interactions may be negatively impacted due to people feeling uncomfortable having a camera pointed at them during conversation, and also to allow for a wider variety of use cases including in business social interactions. For now, initial prototypes will be made to run on smartphones, but deployment to these other

platforms is possible in the future due to the extreme flexibility of the system.

5.3 Implications

Once deployed, AEGIS will be easily accessible and usable by anyone owning a smartphone device. The system comes at virtually no cost, and as long as the user owns the appropriate hardware device, they will be able to access and use the application. With the use of our multimodal system, these individuals with ASD will no longer be required to guess the emotional states of other people, and can easily interpret their expressions via the non-intrusive yet simple to understand emojis. Further research will need to be done to determine if users with ASD actually benefit from using our system, but studies such as [32] which employ similar assistive technologies have shown that improved emotion recognition skills can be achieved after the use of their system. In [32] specifically, when compared against pretest performances, participant scores were on average 19% higher on the task where emotions were inferred from facial expression video clips, and on average 9% higher on the task where emotions were inferred from voice audio clips. These promising results lead us to believe that AEGIS can achieve comparable improvements due to the similarities between the assistive technologies.

However, one problem is that users may be unwilling to use it in a social setting at first, as they will need to have their phone cameras on and pointed towards anyone they are speaking to. Also, social interactions may be negatively influenced due to the other party being unable or unwilling to accommodate the system. Thus, we propose that the system be initially used for training purposes, in an environment where all parties are comfortable and familiar with the system and its use, such as at home. Through use of the system, those living with ASD may be able to learn and interpret social cues firstly for people close to them, and then be able to generalize to a wider range of interactions in society. Once AEGIS is deployed on video conference systems and smartglasses, individuals will also be able to use our system in more unfamiliar settings, allowing them to gain confidence in societal interaction and thus improve their social lives.

Privacy concerns are often associated with camera based devices, and AEGIS is no exception. We can alleviate some of these concerns by ensuring that no personal information or images are stored on the devices after each use. At any given moment, the most frames stored on the device should only be the amount of frames needed for inference, which in this case is 5 frames, due to the 5-frame image stack. In addition, all processing is done locally on the user's personal device, meaning that a network connection is not necessary for AEGIS to function. This also means that the user can use AEGIS with peace of mind,

knowing that the image frames captured on their device are not being sent over the internet to any third parties.

5.4 Summary

This chapter presented the development of AEGIS, a novel multimodal augmented reality assistive technology system designed to help individuals with ASD with the detection and interpretation of facial expressions in social settings. The first iteration of the system is planned to run on a smartphone device, and overlay expression information via emojis on top of each real-world camera frame, in real-time. We propose that the system be initially used in a home setting for training purposes, and then allow the user to generalize to experiences in society. Future work involves deploying the system to other devices such as tablets and smartglasses, and validating the benefits of using our system.

AEGIS can be used with any real-time neural network that leverages time-windowed convolution due to its modular design, allowing for easy modifications of the network itself. In order to improve the accuracy of the models used, a high quality temporal dataset is required for training. The BigFaceX dataset mentioned in Chapter 3 is a good starting point due to the increased generalization ability of models trained using it as it increases the number of unique subjects and images the model encounters during training. We take this one step further and introduce FaceParty in the next chapter, where we add in three new datasets to BigFaceX (AFEW [11], KDEF-dyn [6], and iSAFE [52]), hypothesizing that this amalgamation can further improve model accuracy on expressions in the wild.

Chapter 6

FaceParty

In this chapter, the methodology behind how the FaceParty dataset was created is revealed. FaceParty builds upon the BigFaceX dataset introduced in Chapter 3, by introducing three additional public facial expression databases and a new processing method. These three datasets, AFEW [11], KDEF-dyn [6], and iSAFE [52], are discussed, and the processing methods are disclosed for ease of reproduction.

6.1 Datasets

In addition to the CK+, BAUM-1, and eNTERFACE datasets from BigFaceX, three additional datasets are included. They were selected due to their database sizes, types of data included, and also because they contained the seven expression classes used previously: anger, disgust, fear, happiness, sadness, surprise, and neutral.

6.1.1 AFEW

The Acted Facial Expressions in the Wild (AFEW) database differs from most laboratory controlled datasets as it contains only images and scenes that were extracted from movies, rather than from a controlled setting. Much like other facial expression datasets, it includes a wide range of subjects, each with varying ages, genders, ethnicity, and occluding objects. However, the video clips in AFEW are much closer to a real-life situation, as the actors in the movies are generally acting out a scenario in a real world setting rather than in a laboratory. Additionally, the subjects in the video clips are not controlled, meaning that

Figure 6.1: **An example image sequence from the AFEW dataset.** In this scene, the man is labelled with the "surprise" expression. Note that frames have been skipped in between each image in the sequence.

their head poses and positions vary from frame to frame, thus increasing the difficulty of classification by a large margin. Furthermore, the illumination levels in each scene differ greatly, depending on the context of the scene they were taken from, and includes indoor, night-time, outdoor, bright, and dark illuminations.

AFEW contains 957 video sequences, containing a total of 1259 expressions, as some clips contain multiple subjects. Each sequence was saved in the AVI digital format, and has a run-time length ranging from 300 to 5400 milliseconds.

6.1.2 KDEF-dyn

The original static Karolinska Directed Emotional Faces (KDEF) database [40] is a popular face database that has been widely used in the behavioural and neurophysiological research fields [6]. This dynamic version of the dataset, KDEF-dyn [6], was built by applying morphing animations to KDEF photographs, thus transforming the initial neutral face expression image into a peak facial expression. The animation mimics real-life expressions and the natural speed of an expression event, and was easily adjustable in terms of duration, speed, and intensity. Even though the resulting video clips are not taken from real-life subjects acting out the emotion, studies have indicated that natural expressions tend to unfold in a uniform and ballistic manner, meaning that these dynamically generated expressions can capture the same properties as performed by a human subject [6]. The inclusion of KDEF-dyn in the FaceParty dataset can allow for additional dynamically generated images to be included in the future, and can also help neural network models during training due to the controlled nature of these samples.

The KDEF-dyn dataset only contains 20 male and 20 female subjects, with each displaying six basic expressions, happiness, sadness, anger, fear, disgust, and surprise. As the videos started from the neutral expression, the first few frames of each clip was used as a neutral expression sample. In total, 240 dynamic video clips were constructed, each

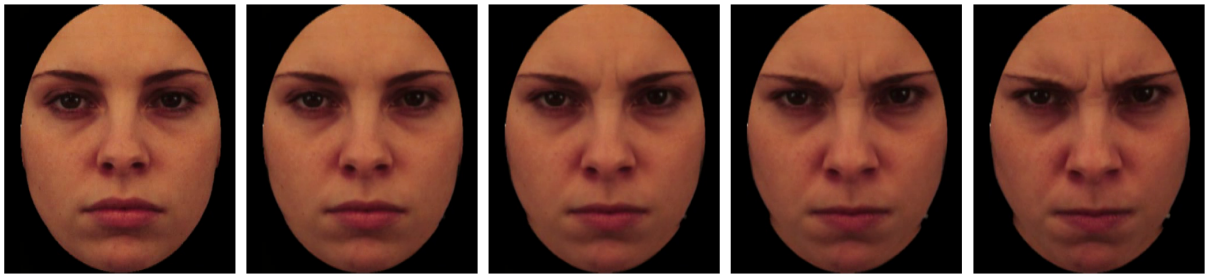

Figure 6.2: **An example image sequence from the KDEF-dyn dataset.** In this clip, the subject starts from the "neutral" expression, and has their expression morphed into the "anger" expression. Note that frames have been skipped in between each image in the sequence.

with a length of 1033 milliseconds. Each video sequence started with an original KDEF photograph (frame 0), and ended with the peak expression at frame 30, which is also an image from the original KDEF dataset.

6.1.3 iSAFE

The Indian Semi-Acted Facial Expression database [52], or iSAFE, is a human facial expression recognition database designed specifically for the indian demographic. As most FER databases contain subjects primarily of Caucasian descent, these traditionally available public datasets tend to perform poorly when introduced to a subject with a different skin color or ethnically different facial features. The inclusion of iSAFE into FaceParty ensures that models trained using FaceParty will also be able to learn expressions from a wider range of subjects.

The dataset contains 395 clips of 44 different subjects between the ages of 17 to 22. Each video clip was manually segmented from a video recording of the subject viewing a specific expression stimuli, and each clip is labelled by an annotator as well as the subject themselves. Handcrafted features such as histogram of oriented gradients are also included in the dataset.

Figure 6.3: **An example image sequence from the iSAFE dataset.** In this video, the subject can be observed acting out the "disgust" expression. Note that frames have been skipped in between each image in the sequence.

6.2 FaceParty Creation

FaceParty was created in a similar method to BigFaceX, discussed in Chapter 3. However, certain changes were made in order to increase generalization ability and ease of training. The original BigFaceX dataset created as many time windows as possible from each video in order to try and maximize the number of data points, but a side effect of this was a large amount of windows in which the expression was not entirely present. To solve this, FaceParty instead limits the number of time windows created per video to just 3.

In order to pinpoint the time windows with the highest manifestation of the target emotion, we first trained a static frame facial expression classifier using the AffectNet [44], ExpW [65], and FER2013 [21] static image datasets. Each of these datasets contains thousands of static face images labelled with an expression, most importantly including the classes that we used in BigFaceX. All three of these datasets use images collected from the internet, where emotion related keywords were combined with various words related to gender, age, ethnicity, or occupations as a search query. Images returned from the online search engine were then collected, processed using a face recognition software, and then cropped to a facial bounding box. Lastly, manual verification performed by human annotators was done to ensure the validity of the dataset. AffectNet contains a total of 450,000 labelled expressions (of which 313,834 are of our desired classes), ExpW contains 91,793, and FER2013 consists of 35,887 images. An additional pre-processing step was necessary in order to synchronize the class labels, as different datasets used different numberings for each expression.

A ResNet20 model was used for this static image classifier, trained over 200 epochs using a batch size of 512 with the Adam [31] optimizer. Images were resized to a shape of 48x48x1. Due to the large class imbalances in each of the datasets, we under-sampled each dataset in order to maintain the same amount of data samples for each class. Image

augmentation was performed, including up to 10 degree rotations, width and height shifts of up to 10 percent, horizontal flips, brightness shifts of up to 10 percent, and zoom transforms of up to 10 percent.

For the datasets CK+ and KDEF-dyn, the location of the peak expression was at the end of each video, and so the process for these two datasets was just to extract the three time windows at the end of the video sequence. iSAFE follows a similar description, however, the expressions were much less controlled and could end in a non-peak pose, and thus we used the same method for iSAFE as we did for AFEW, BAUM-1, and eNTERFACE. For these datasets, we loaded every frame in the video sequence, cropped each to a facial bounding box, and resized to 48x48 pixels in size. Next, we had our static image classifier predict the expression for every frame that contained a face, and sorted the resulting softmax predictions by the true label of the video. The top three frames that displayed the expression were selected, and time windows of size 5 were constructed by taking the previous 4 frames and using the "peak" frame as the end frame. In the case that the expression was at the start of the video prior to frame 5, frames 1 to 5 were used instead. Finally, each time window was manually verified by a human annotator, and invalid windows including frames that were mostly occluded or did not contain the entire face were deleted. The final class distribution of the FaceParty dataset can be seen in Table 6.1, with happiness having the most samples and neutral having the least.

The method used for the facial bounding box cropping step differs from how it was performed in BigFaceX. Rather than increasing the Haar Cascades [58] generated bounding boxes by 10 percent to retain most of the face, we shrank the bounding boxes in order to increase the amount of relevant information in each image frame. Bounding boxes were shrunk by 10 percent on each side on the x-axis, and shrunk by 20 percent from the top of the image.

6.3 Summary

In this chapter, the FaceParty dataset was introduced, created from a modified aggregation of six public facial expression recognition datasets, CK+, BAUM-1, eNTERFACE, AFEW, KDEF-dyn, and iSAFE. Each of these datasets were chosen due to their shared class labels, and each serves a specific purpose in allowing greater generalization ability for any models trained using it. Future work includes adding in more datasets, modifying the images themselves with various occlusion types such as head-wear, facial hair, or shadows, and a more careful class balancing of the samples in FaceParty.

Table 6.1: **Class Distribution of the FaceParty Dataset.** There are the most samples for the happy class, and the least samples for the neutral class.

Expression	**Number of data points**
Angry	824
Disgust	1007
Fear	979
Happy	1209
Sad	939
Surprise	1035
Neutral	597
Total	6590

Chapter 7

Conclusion

In this chapter, a brief summary of the thesis and the key contributions are described. Controversy surrounding the field of facial expression classification is mentioned, followed by some parting thoughts and future work.

7.1 Summary of Thesis and Contributions

In this thesis, three main contributions were disclosed. First, a novel deep time windowed convolutional neural network design was proposed, which are capable of leveraging temporal information in an efficient and compact manner. Based on comparisons against other model architectures leveraging other variants of temporal convolution, it can be seen that the proposed TimeConvNet designs are both more accurate as well as faster in terms of inference time. Out of all the models proposed, the ResNet20 TimeConvNet variant demonstrates the best balance between accuracy, inference time, and parameters, and is able to achieve real-time level performance while maintaining a high accuracy.

In the following chapter, a human-machine collaborative design strategy tailored specifically for the task of human facial expression classification is explored. EmotionNet Nano, a highly compact human facial expression classifier is discussed, and tested against other state-of-the-art models using the CK+ dynamic video dataset. Results indicate that the proposed EmotionNet Nano variants can achieve comparable accuracy with state-of-the-art, while also possessing a significantly more efficient architecture design. In addition, we demonstrated that EmotionNet Nano can also achieve real-time performance on an embedded processor at varying power levels, thus further illustrating its suitability for real-time embedded scenarios, such as for individuals with ASD to use.

The Augmented-reality Expression Guided Interpretation System (AEGIS) is then proposed in Chapter 5. While not a main contribution, it demonstrates the applicability of the TimeConvNet design in an embedded scenario, also validating the use cases of the EmotionNet Nano networks. AEGIS overlays expression information via emojis on top of real-world image frames, providing users with instantaneous feedback of a target's facial expression, thus assisting them with the detection and interpretation of these expressions in social settings.

The third main contribution of this thesis is the FaceParty dataset. Built from six public dynamic facial expression databases, it is a custom, more difficult dataset that we hope future researchers can leverage to train models with greater generalization abilities. Details to create the FaceParty dataset are disclosed in Chapter 6.

7.2 Controversy

With the recent advances of facial recognition software over the past decade, privacy concerns over the storage of personal information and images has been on the rise. A malicious third party could track another person without their consent by matching images of that person's face against a security database, thus knowing their location as long as a camera is nearby. Devices using a malicious application could store images of their user's faces and send them back to a third party, granting them access to the user's private documents if they had bio-metric face recognition based passwords set up.

Of course, there is an argument that public and private security systems can receive a massive upgrade with the use of facial recognition software. Tracking each individual who enters a building, knowing where suspicious targets are at any given moment, or even looking up someone's identity are valuable resources for any security organization in the world. In the field of assistive technology, facial recognition software also provides huge benefits, by not only improving user experiences through custom profiles, but also enhancing other aspects of life by analyzing their user's emotional state.

The main argument against the widespread adoption of facial recognition is privacy concerns, and attempts to alleviate some of these issues have been proposed. Erkin et al. [14] suggests a system where the face image provider and the facial template database can perform recognition and provide a result efficiently while using a custom algorithm that prevents each party from learning the other parameters. In this scenario, the image provider would be unable to learn more than the basic parameters of the database, while the database would be unable to learn the input image altogether, thus providing security from both ends.

7.3 Future Work

The TimeConvNet system design explored in Chapter 3 showcases a solid balance between speed and accuracy by leveraging a time window size of 5 frames, but exploring various other sizes is worthwhile, to investigate if additional temporal information can enhance the classification accuracy further. Combining the TimeConvNet design with the human-machine collaborative design strategy is also worth investigating, to see if additional performance gains can be achieved.

A thorough testing of multiple models using the FaceParty dataset will be conducted, in order to validate and confirm the datasets usability in the field. Efforts to increase the size of the database will also be performed, as the number of samples is still relatively small in comparison to other available datasets. Alternate image augmentation techniques can also be explored in conjunction with FaceParty, in order to increase the amount of available training data for models to learn from. Furthermore, an additional effort to balance the dataset is necessary due to the large class imbalances present in the current iteration. Models trained using FaceParty should experience an enhanced generalization ability due to the variety of subjects and datasets used, but confirming this and making sure no inherent biases exist in the dataset is important if FaceParty is to be used extensively in the future.

The validation of AEGIS and testing the system in the wild is also worthwhile, in order to verify the effects of the system for individuals with ASD. Initial experiments can be performed in a similar fashion to [32], moving on to tests in a home environment, and finally to social interactions in the wild. Extending AEGIS onto additional platforms such as smartglasses and web cameras is a natural progression when aiming for a seamless and unobtrusive design, and future iterations should have this functionality included. Allowing additional customization for each user is also viable, as some participants may prefer text based cues or colour based feedback rather than emoji augmentation. Although it may take extended use of AEGIS in order for any major effects to be noticeable, we believe that with consistent use early on, users will gain an enhanced ability to detect, recognize, and interpret facial expressions, thus improving their relationships and interactions with society.

7.4 Parting Thoughts

Facial expression recognition is a powerful ability to have, one that most of us take for granted in our everyday lives. We socialize and communicate, all while not noticing that

we pick up thousands of emotional cues on a daily basis. Our natural ability to discern a targets emotional state from their subtle facial transitions and immediately understand and adjust the flow of the conversation is something that has allowed humans to be the social animals that we are. When this ability is impaired, it severely restricts the types of social interactions one can have, and can damage interpersonal relationships and cause loneliness and depression.

This thesis aims to provide a first step to assist these individuals, and all fields that rely on fast, compact, and accurate facial expression classifiers. With the use of some of the contributions listed in this thesis, we hope that future researchers will be able them as a starting point for more reliable systems that can better the world.

References

[1] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017.

[2] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436. ACM, 2016.

[3] AM Barreto. Application of facial expression studies on the field of marketing. *Emot. Expr. brain face, no. June*, pages 163–189, 2017.

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[5] Mrs Ayesha Butalia, Maya Ingle, and Parag Kulkarni. Facial expression recognition for security. *International Journal of Modern Engineering Research (IJMER)*, 2(4):1449–1453, 2012.

[6] Manuel G Calvo, Andrés Fernández-Martín, Aida Gutiérrez-García, and Daniel Lundqvist. Selective eye fixations on diagnostic face regions of dynamic emotional expressions: Kdef-dyn database. *Scientific reports*, 8(1):1–10, 2018.

[7] François Chollet et al. Keras. https://keras.io, 2015.

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[9] Jeffrey F Cohn, Adena J Zlochower, James J Lien, and Takeo Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 396–401. IEEE, 1998.

[10] Abhinav Dhall, Roland Goecke, and Tom Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[11] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted facial expressions in the wild database. 10 2011.

[12] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. pages 2106–2112, 11 2011.

[13] Paul Ekman and Wallace V. Friesen. Facial action coding system: Manual. 1978.

[14] Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Lagendijk, and Tomas Toft. Privacy-preserving face recognition. In *International symposium on privacy enhancing technologies symposium*, pages 235–253. Springer, 2009.

[15] Natascha Esau, Evgenija Wetzel, Lisa Kleinjohann, and Bernd Kleinjohann. Real-time facial expression recognition using a fuzzy emotion model. In *2007 IEEE international fuzzy systems conference*, pages 1–6. IEEE, 2007.

[16] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997.

[17] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[18] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM, 2016.

[19] Duo Feng and Fuji Ren. Dynamic facial expression recognition based on two-stream-cnn with lbp-top. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 355–359. IEEE, 2018.

[20] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019.

[21] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[22] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[23] SL Happy, Anjith George, and Aurobinda Routray. A real time facial expression classification system using local binary patterns. In *2012 4th International conference on intelligent human computer interaction (IHCI)*, pages 1–5. IEEE, 2012.

[24] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3D convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–40, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] James Ren Hou Lee, Linda Wang, and Alexander Wong. EmotionNet Nano: An Efficient Deep Convolutional Neural Network Design for Real-time Facial Expression Recognition. *arXiv e-prints*, page arXiv:2006.15759, June 2020.

[27] James Ren Hou Lee and Alexander Wong. AEGIS: A real-time multimodal augmented reality computer vision based system to assist facial expression recognition for individuals with autism spectrum disorder. *arXiv e-prints*, page arXiv:2010.11884, October 2020.

[28] Mira Jeong and Byoung Chul Ko. Driver's facial expression recognition in real-time for safe driving. *Sensors*, 18(12):4270, 2018.

[29] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE, 2000.

[30] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 19–27, 2015.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Paul G Lacava, Ofer Golan, Simon Baron-Cohen, and Brenda Smith Myles. Using assistive technology to teach emotion recognition to students with asperger syndrome: A pilot study. *Remedial and Special Education*, 28(3):174–181, 2007.

[33] James Ren Hou Lee and Alexander Wong. Timeconvnets: A deep time windowed convolution neural network design for real-time video facial expression recognition. In *2020 17th Conference on Computer and Robot Vision (CRV)*, pages 9–16. IEEE, 2020.

[34] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[35] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[36] Gwen Littlewort, Ian Fasel, M Stewart Bartlett, and Javier R Movellan. Fully automatic coding of basic expressions from video. *University of California, San Diego, San Diego, CA*, 92093, 2002.

[37] Wei Liu, Zhifeng Li, and Xiaoou Tang. Spatio-temporal embedding for statistical face recognition from video. In *European Conference on Computer Vision*, pages 374–388. Springer, 2006.

[38] Elena Lozano-Monasor, María T López, Francisco Vigo-Bustos, and Antonio Fernández-Caballero. Facial expression recognition in ageing adults: from lab to ambient assisted living. *Journal of Ambient Intelligence and Humanized Computing*, 8(4):567–578, 2017.

[39] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[40] Daniel Lundqvist, Anders Flykt, and Arne Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91(630):2–2, 1998.

[41] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[42] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[43] Philipp Michel and Rana El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM, 2003.

[44] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[45] Carlos Orrite, Andrés Gañán, and Grégory Rogez. Hog-based decision tree for facial expression classification. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 176–183. Springer, 2009.

[46] Naima Otberdout, Anis Kacem, Mohamed Daoudi, Lahoucine Ballihi, and Stefano Berretti. Automatic analysis of facial expressions based on deep covariance trajectories. *IEEE transactions on neural networks and learning systems*, 2019.

[47] Sébastien Ouellet. Real-time emotion recognition for gaming using deep convolutional network features. *arXiv preprint arXiv:1408.3750*, 2014.

[48] Xianzhang Pan, Guoliang Ying, Guodong Chen, Hongming Li, and Wenshu Li. A deep spatial and temporal aggregation framework for video-based facial expression recognition. *IEEE Access*, 7:48807–48815, 2019.

[49] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005.

[50] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[51] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[52] Shivendra Singh and Shajulin Benedict. Indian semi-acted facial expression (isafe) dataset for human emotions recognition. In *International Symposium on Signal Processing and Intelligent Recognition Systems*, pages 150–162. Springer, 2019.

[53] Bo Sun, Qinglan Wei, Liandong Li, Qihua Xu, Jun He, and Lejun Yu. LSTM for dynamic emotion and group emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 451–457. ACM, 2016.

[54] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[55] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[56] Michel F Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. The first facial expression recognition and analysis challenge. In *Face and Gesture 2011*, pages 921–926. IEEE, 2011.

[57] Michel F Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43, 2011.

[58] Paul Viola, Michael Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1(511-518):3, 2001.

[59] Guan Wang and Jun Gong. Facial expression recognition based on improved lenet-5 cnn. In *2019 Chinese Control And Decision Conference (CCDC)*, pages 5655–5660. IEEE, 2019.

[60] Yubo Wang, Haizhou Ai, Bo Wu, and Chang Huang. Real time facial expression recognition with adaboost. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 926–929. IEEE, 2004.

[61] Alexander Wong, Mohammad Javad Shafiee, Brendan Chwyl, and Francis Li. Ferminets: Learning generative machines to generate efficient neural networks via generative synthesis. *arXiv preprint arXiv:1809.05989*, 2018.

[62] Osamu Yamaguchi, Kazuhiro Fukui, and K-i Maeda. Face recognition using temporal image sequence. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323. IEEE, 1998.

[63] Helena Song Sook Yee. Mobile technology for children with autism spectrum disorder: Major trends and issues. In *2012 IEEE Symposium on E-Learning, E-Management and E-Services*, pages 1–5. IEEE, 2012.

[64] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313, 2016.

[65] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[66] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[67] Yitao Zhou, Fuji Ren, Shun Nishide, and Xin Kang. Facial sentiment classification based on resnet-18 model. In *2019 International Conference on Electronic Engineering and Informatics (EEI)*, pages 463–466. IEEE, 2019.