Going Deeper in Facial Expression Recognition using Deep Neural Networks

Ali Mollahosseini[1], David Chan[2], and Mohammad H. Mahoor[1,2]
[1] Department of Electrical and Computer Engineering
[2] Department of Computer Science
University of Denver, Denver, CO
ali.mollahosseini@du.edu, davidchan@cs.du.edu, and mmahoor@du.edu *†

Abstract

Automated Facial Expression Recognition (FER) has remained a challenging and interesting problem. Despite efforts made in developing various methods for FER, existing approaches traditionally lack generalizability when applied to unseen images or those that are captured in wild setting. Most of the existing approaches are based on engineered features (e.g. HOG, LBPH, and Gabor) where the classifier's hyperparameters are tuned to give best recognition accuracies across a single database, or a small collection of similar databases. Nevertheless, the results are not significant when they are applied to novel data. This paper proposes a deep neural network architecture to address the FER problem across multiple well-known standard face datasets. Specifically, our network consists of two convolutional layers each followed by max pooling and then four Inception layers. The network is a single component architecture that takes registered facial images as the input and classifies them into either of the six basic or the neutral expressions. We conducted comprehensive experiments on seven publically available facial expression databases, viz. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013. The results of proposed architecture are comparable to or better than the state-of-the-art methods and better than traditional convolutional neural networks and in both accuracy and training time.

1. Introduction

Current Human Machine Interaction (HMI) systems have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. Facial expression, which plays a vital role in social interaction, is one of the most important nonverbal channels through which HMI systems can recognize humans' internal emotions. Ekman *et al.* identified six facial expressions (viz. anger, disgust, fear, happiness, sadness, and surprise) as basic emotional expressions that are universal among human beings [11].

Due to the importance of facial expression in designing HMI and Human Robot Interaction (HRI) systems [33], numerous computer vision and machine learning algorithms have been proposed for automated Facial Expression Recognition (FER). Also, there exist many annotated face databases with either human actors portraying basic expressions [15, 35, 28, 29], or faces captured spontaneously in an uncontrolled setting [9, 29]. Automated FER approaches attempt to classify faces in a given single image or sequence of images as one of the six basic emotions. Although, traditional machine learning approaches such as support vector machines, and to a lesser extent, Bayesian classifiers, have been successful when classifying posed facial expressions in a controlled environment, recent studies have shown that these solutions do not have the flexibility to classify images captured in a spontaneous uncontrolled manner ("in the wild") or when applied databases for which they were not designed [30]. This poor generalizability of these methods is primarily due to the fact that many approaches are subject or database dependent and only capable of recognizing exaggerated or limited expressions similar to those in the training database. Many FER databases have tightly controlled illumination and pose conditions. In addition, obtaining accurate training data is particularly difficult, especially for emotions such as sadness or fear which are extremely difficult to accurately replicate and do not occur often real life.

Recently, due to an increase in the ready availability of computational power and increasingly large training databases to work with, the machine learning technique of neural networks has seen resurgence in popularity. Recent state of the art results have been obtained using neural networks in the fields of visual object recognition [20, 41], human pose estimation [45], face verification [43], and many

*This work is partially supported by the NSF grants IIS-1111568 and CNS-1427872.

†To be appear in IEEE Winter Conference on Applications of Computer Vision (WACV), 2016

more. Even in the FER field results so far have been promising [17]. Unlike traditional machine learning approaches where features are defined by hand, we often see improvement in visual processing tasks when using neural networks because of the network's ability to extract undefined features from the training database. It is often the case that neural networks that are trained on large amounts of data are able to extract features generalizing well to scenarios that the network has not been trained on. We explore this idea closely by training our proposed network architecture on a subset of the available training databases, and then performing cross-database experiments which allow us to accurately judge the network's performance in novel scenarios.

In the FER problem, however, unlike visual object databases such as imageNet [8], existing FER databases often have limited numbers of subjects, few sample images or videos per expression, or small variation between sets, making neural networks significantly more difficult to train. For example, the FER2013 database [1] (one of the largest recently released FER databases) contains 35,887 images of different subjects yet only 547 of the images portray disgust. Similarly, the CMU MultiPIE face database [15] contains around 750,000 images but it is comprised of only 337 different subjects, where 348,000 images portray only a "neutral" emotion and the remaining images do not portray anger, fear or sadness.

This paper presents a novel deep neural network architecture for the FER problem, and examines the network's ability to perform cross-database classification while training on databases that have limited scope, and are often specialized for a few expressions (e.g. MultiPIE and FERA). We conducted comprehensive experiments on seven well-known facial expression databases (viz. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013) and obtain results which are significantly better than, or comparable to, traditional convolutional neural networks or other state-of-the-art methods in both accuracy and learning time.

2. Background and Related Work

Algorithms for automated FER usually involve three main steps, viz. registration, feature extraction, and classification. In the face registration step, faces are first located in the image using some set of landmark points during "face localization" or "face detection". These detected faces are then geometrically normalized to match some template image in a process called "face registration". In the feature extraction step, a numerical feature vector is generated from the resulting registered image. These features can be *geometric features* such as facial landmarks [19], *appearance features* such as pixel intensities [32], Gabor filters [23], Local Binary Patterns (LBP) [37], Local Phase Quantization (LPQ) [52], and Histogram of Oriented Gradients (HoG) [29], or *motion features* such as optical flow [18], Motion History Images (MHI) [46], and volume LBP [51]. Current state-of-the-art methods, such as those used in Zhang *et al.* [49, 50] fuse multiple features using multiple kernel learning algorithms. However by using neural networks, we do not have to worry about the feature selection step - as neural networks have the capacity to learn features that statistically allow the network to make correct classifications of the input data. In the third step, of classification, the algorithm attempts to classify the given face as portraying one of the six basic emotions using machine learning techniques.

Ekman *et al.* [6, 12] distinguished two conceptual approaches to studying facial behavior: a "message-based" approach and a "sign-based" approach. Message-based approaches categorize facial behaviors as the the meaning of expressions, whereas sign-based approaches describe facial actions/configuration regardless of the action's meaning. The most well-known and widely used sign-based approach is the Facial Action Coding System (FACS) [12]. FACS describes human facial movements by their appearance on the face using standard facial substructures called Action Units (AUs). Each AU is based on one or a few facial muscles and AUs may occur individually or in combinations. Similarly, FER algorithms can be categorized into both message-based and sign-based approaches. In message-based approaches FER algorithms are trained on databases labeled with the six basic expressions [7], and more recently, embarrassment and contempt [27]. Unlike message-based algorithms, sign-based algorithms are trained to detect AUs in a given image or sequence of images [7]. These detected AUs are then converted to emotion-specified expressions using EMFACS [14] or similar systems [42]. In this paper, we develop a message-based neural network solution,

FER systems are traditionally evaluated in either a subject independent manner or a cross-database manner. In subject independent evaluation, the classifier is trained on a subset of images in a database (called the training set) and evaluated on faces in the same database that are not elements of the training set often using K-fold cross validation or leave-one-subject-out approaches. The cross-database method of evaluating facial expression systems requires training the classifier on all of the images in a single database and evaluating the classifier on a different database which the classifier has never seen images from. As single databases have similar settings (illumination, pose, resolution etc.), subject independent tasks are easier to solve than cross database tasks. Subject independent evaluation is not, however, unimportant. If a researcher can guarantee that the data will align well in pose, illumination and other factors with the training set, subject independent evaluation can give a reasonably good representation of the classification accuracy in an online system. Another technique,

subject dependent evaluation (person-specific), is also used in limited cases, e.g. FERA 2011 challenge [47]; often in these scenarios the recognition accuracy is more important than the generalization.

Recent approaches to visual object recognition tasks, and the FER problem have used increasingly "deep" neural networks (neural networks with large numbers of hidden layers). The term "deep neural network" refers to a relatively new set of techniques in neural network architecture design that were developed in order to improve the ability of neural networks to tackle big-data problems. With the large amount of available computing power continuing to grow, deep neural network architectures provide a learning architecture based in the development of "brain-like" structures which can learn multiple levels of representation and abstraction which allow algorithms for finding complex patterns in images, sound, and text.

It seems only logical to extend cutting-edge techniques in the field of "deep learning" to the FER problem. Deep networks have a remarkable ability to perform well in flexible learning tasks, such as the cross-database evaluation situation, where it is unlikely that hand-crafted features will easily generalize to a new scenario. By training neural networks, particularly deep neural networks, for feature recognition and extraction we can drastically reduce the amount of time that is necessary to implement a solution to the FER problem that, even when confronted with a novel data source, will be able to perform at high levels of accuracy. Similarly, we see deep neural networks performing well in the subject independent evaluation scenarios, as the algorithms can learn to recognize subtle features that even field experts can miss. These correlations provide the motivation for this paper, as the strengths of deep learning seem to align perfectly with the techniques required for solving difficult "in the wild" FER problems.

A subset of deep neural network architectures called "convolutional neural networks" (CNNs) have become the traditional approach for researchers studying vision and deep learning. In the 2014 ImageNet challenge for object recognition, the top three finishers all used a CNN approach, with the GoogLeNet architecture achieving a remarkable 6.66% error rate in classification [41, 36]. The GoogLeNet architecture uses a novel multi-scale approach by using multiple classifier structures, combined with multiple sources for back propagation. This architecture defeats a number of problems that occur when back-propagation decays before reaching beginning layers in the architecture. Additional layers that reduce dimension allow GoogLeNet to increase in both width and depth without significant penalties, and take an elegant step towards complicated network-in-network architectures described originally in Lin *et al.* [22]. In other word, the architecture is composed of multiple "Inception" layers, each of which acts like a micro-network in the larger network, allowing the architecture to make more complex decisions.

More traditional CNN architectures have also achieved remarkable results. AlexNet [20] is an architecture that is based on the traditional CNN layered architecture - stacks of convolutions layers followed by max-pooling layers and rectified linear units (ReLUs), with a number of fully connected layers at the top of the layer stack. Their top=5 error rate of 15.3% on the ILSVRC-2012 competition revolutionized the way that we think about the effectiveness of CNNs. This network was also one of the first networks to introduce the "dropout" method for solving the over fitting problem (Suggested by Hinton *et al.* [38]) which proved key in developing large neural networks. One of the large challenges to overcome in the use of traditional CNN architectures is their depth and computational complexity. The full AlexNet network performs on the order of 100M operations for a single iteration, while SVM and shallow neural networks perform far fewer operations in order to create a suitable model. This makes traditional CNNs very hard to apply in time restrictive scenarios.

In [24] a new deep neural network architecture, called an "AU-Aware" architecture was proposed in order to investigate the FER problem. In an AU-Aware architecture, the bottom of the layer stack consists of convolution layers and max-pooling layers which are used to generate a complete representation of the face. Next in the layer stack, an "AU-aware receptive field layer" generates a complete representation over all possible spatial regions by convolving the dense-sampling facial patches with special filters in a greedy manner. Then, a multilayer Restricted Boltzmann Machine (RBM) is exploited to learn hierarchical features. Finally, the outputs of the network are concatenated as features which are used to train a linear SVM classifier for recognizing the six basic expressions. Results in [24] show that the features generated by this "AU-Aware" network are competitive with or superior to handcrafted features such as LBP, SIFT, HoG, and Gabor on the CK+, MMI and databases using a similar SVM. However, AU-aware layers do not necessarily detect FACS defined action units in faces.

In [17] multiple deep neural network architectures are combined to solve the FER problem in video analysis. These network architectures included: (1) an architecture similar to the AlexNet CNN run on individual frames of the video, (2) a deep belief network trained on audio information, (3) an autoencoder to model the spatiotemporal properties of human activity, and (4) a shallow network focused on the mouth. The CNN is trained on the private Toronto Face Database [40] and fine tuned on the AFEW database [9], yielded an accuracy of 35.58% when evaluated in a subject independent manner on AFEW. When combined with a single predictor, the five architectures produced an ac-

curacy of 41.03% on the test set, the highest accuracy in the EmotiW 2013 [9] challenge, where challenge winner 2014 [26] achieved 50.40% on test set using multiple kernel methods on Riemannian manifold.

A 3D CNN with deformable action parts constraints is introduced in [25] which can detect specific facial action parts under the structured spatial constraints, and obtain the discriminative part-based representation simultaneously. The results on two posed expression datasets, CK+, MMI, and a spontaneous dataset FERA achieve state-of-the-art video-based expression recognition accuracy.

3. Proposed Method

Often improving neural network architectures has relied on increasing the number of neurons or increasing the number of layers, allowing the network to learn more complex functions; however, increasing the depth and complexity of a topology leads to a number of problems such as increased over-fitting of training data, and increased computational needs. A natural solution to the problem of increasingly dense networks is to create deep sparse networks, which has both biological inspiration, and has firm theoretical foundations discussed in Arora *et al*. [3]. Unfortunately, current GPUs and CPUs do not have the capability to efficiently compute actions on sparse networks. The Inception layer presented in [36] attempts to rectify these concerns by providing an approximation of sparse networks to gain the theoretical benefits proposed by Arora *et al*., however retains the dense structure required for efficient computation.

Applying the Inception layer to applications of Deep Neural Network has had remarkable results, as implied by [39] and [41], and it seems only logical to extend state of the art techniques used in object recognition to the FER problem. In addition to merely providing theoretical gains from the sparsity, and thus, relative depth, of the network, the Inception layer also allows for improved recognition of local features, as smaller convolutions are applied locally, while larger convolutions approximate global features. The increased local performance seems to align logically with the way that humans process emotions as well. By looking at local features such as the eyes and mouth, humans can distinguish the majority of the emotions [4]. Similarly, children with autism often cannot distinguish emotion properly without being told to remember to look at the same local features [4]. By using the Inception layer structure and applying the network-in-network theory proposed by Lin *et al*. [22], we can expect significant gains on local feature performance, which seems to logically translate to improved FER results.

Another benefit of the network-in-network method is that along with increased local performance, the global pooling performance is increased and therefore it is less prone to overfitting. This resistance to overfitting allows us to increase the depth of the network significantly without worrying about the small corpus of images that we are working with in the FER problem.

The work that we present in this paper is inspired by the techniques provided by the GoogLeNet and AlexNet architectures described in Sec. 2. Our network consists of two elements, first our network contains of two traditional CNN modules (a traditional CNN layer consists of a convolution layer by a max pooling layer). Both of these modules use rectified linear units (ReLU) which have an activation function described by:

$$f(x) = max(0,x)$$

where x is the input to the neuron [20]. Using the ReLU activation function allows us to avoid the vanishing gradient problem caused by some other activation functions (for more details see [20]). Following these modules, we apply the techniques of the network in network architecture and add two "Inception" style modules, which are made up of a 1×1, 3×3 and 5×5 convolution layers (Using ReLU) in parallel. These layers are then concatenated as output and we use two fully connected layers as the classifying layers (Also using ReLU). Figure 1 shows the architecture of the network used in this paper.

In this work we register facial images in each of the databases using research standard techniques. We used bidirectional warping of Active Appearance Model (AAM) [34] and a Supervised Descent Method (SDM) called IntraFace [48] to extract facial landmarks, however further work could consider improving the landmark recognition in order to extract more accurate faces. IntraFace uses SIFT features for feature mapping and trains a descent method by a linear regression on training set in order to extract 49 points. We use these points to register faces to an average face in an affine transformation. Finally, a fixed rectangle around the average face is considered as the face region. Figure 2 demonstrates samples of the face registration with this method. In our research, facial registration increased the accuracy of our FER algorithms by 4-10%, which suggests that registration (like normalization in traditional problems) is a significant portion of any FER algorithm.

Once the faces have been registered, the images are resized to 48×48 pixels for analysis. Even though many databases are composed of images with a much higher resolution testing suggested that decreasing this resolution does not greatly impact the accuracy, however vastly increases the speed of the network. To augment our data, we extract 5 crops of 40×40 from the four corners and the center of the image and utilize both of them and their horizontal flips for a total of 10 additional images.

In training the network, the learning rates are decreased in a polynomial fashion as: $base\_lr(1 - iter/max\_iter)^{0.5}$,

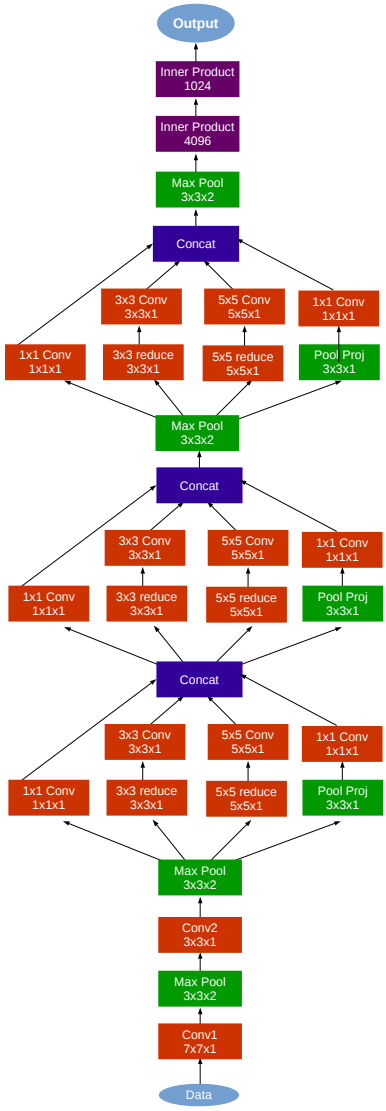

Figure 1. Network Architecture

Figure 2. Sample of the face registration. From left to right images are taken from MultiPIE, SFEW, MMI, CK+ and DISFA. First row shows the original images and the second row shows their registered images respectively.

where $base\_lr = 0.01$ is the base learning rate, *iter* is the current iteration and *max_iter* is the maximum allowed iterations. Testing suggested that other popular learning rate policies such as *fixed* learning rate, *step* where learning rate is multiplies by a gamma factor in each step, and exponential approach did not perform as well as the polynomial fashion. Using the polynomial learning rate, the test loss converged faster and allowed us to train the network for many iterations without the need for fine-tuning. We also trained the bias nodes twice as fast as the weights of the network, in order to increase the rate at which unnecessary nodes are removed from evaluation. This decreases the number of iterations that the network must run before the loss converges.

4. Experimental Results

4.1. Face Databases

We evaluate the proposed method on well-known publicly available facial expression databases: CMU Multi-PIE [15], MMI [35], Denver Intensity of Spontaneous Facial Actions (DISFA) [29], extended CK+ [27], GEMEP-FERA database [5], SFEW [10], and FER2013 [1]. In this section we briefly review the content of these databases.

CMU MultiPIE: CMU MultiPIE face database [15] contains around 750,000 images of 337 people under multiple viewpoints, and different illumination conditions. There are four recording sessions in which subjects were instructed to display different facial expressions (i.e. Angry, Disgust, Happy, Neutral, Surprise, Squint, and Scream). We selected only the five frontal viewpoints (-45°to +45 °), giving us a total of around 200,000 images.

MMI: The MMI [35] database includes more than 20 subjects of both genders (44% female), ranging in age from 19 to 62, having either a European, Asian, or South American ethnicity. An image sequence is captured that has neutral faces at the beginning and the end for each session and subjects were instructed to display 79 series of facial expressions, six of which are prototypic emotions. We extracted static frames from each sequence, where it resulted in 11,500 images.

CK+: The Extended Cohn-Kanade database (CK+) [27] includes 593 video sequences recorded from 123 subjects ranging from 18 to 30 years old. Subjects displayed different expressions starting from the neutral for all sequences, and some sequences are labeled with basic expressions. We selected only the final frame of each sequence with peak expression in our experiment, which results in 309 images.

DISFA: Denver Intensity of Spontaneous Facial Actions (DISFA) database [29] is one of a few naturalistic databases that have been FACS coded with AU intensity values. This database consists of 27 subjects, each recorded while watching a four minutes video clip by two cameras. Twelve AUs

Table 1. Network Configuration

Layer type	Patch Size / Stride	Output	1 x 1	3 x 3	3 x 3 reduce	5 x 5	5 x 5 reduce	Proj Pooling	# Operations
Convolution - 1	7×7 / 2	$24 \times 24 \times 64$							5.7M
Max pool - 1	3×3 / 2	$12 \times 12 \times 64$							5.7M
Convolution - 2	3×3 / 1	$12 \times 12 \times 192$							1.4M
Max Pool - 2	3×3 / 2	$6 \times 6 \times 192$							1.4M
Inception - 3a			64	128	96	32	16	32	2.6M
Inception - 3b			128	192	128	96	32	64	4.5M
Max Pool - 4	3×3 / 2	$3 \times 3 \times 480$							0.6M
Inception - 4a			192	208	96	48	16	64	1.3M
Avg Pooling - 6		$1 \times 1 \times 1024$							25.6K
Fully Connected - 7		$1 \times 1 \times 4096$							0.2M
Fully Connected - 8		$1 \times 1 \times 1024$							51K

are coded between 0-5, where 0 denotes the absence of the AU, while 5 represents maximum intensities. As DISFA is not emotion-specified coded, we used EMFACS system [14] to convert AU FACS codes to expressions, which resulted in around 89,000 images in which the majority have neutral expressions.

FERA: The GEMEP-FERA database [5] is a subset of the GEMEP corpus used as database for the FERA 2011 challenge [47]. It consists of recordings of 10 actors displaying a range of expressions. There are seven subjects in the training data, and six subjects in the test set. The training set contains 155 image sequences and the testing contains 134 image sequences. There are in total five emotion categories in the database: Anger, Fear, Happiness, Relief and Sadness. We extract static frames from the sequences with six basic expressions, which resulted to in around 7,000 images.

SFEW: The Static Facial Expressions in the Wild (SFEW) database [10] is created by selecting static frames from Acted Facial Expressions in the Wild (AFEW) [9]. The SFEW database covers unconstrained facial expressions, different head poses, age range, and occlusions and close to real world illuminations. There are a total of 95 subjects in the database. In total there are 663 well-labeled usable images.

FER2013: The Facial Expression Recognition 2013 (FER-2013) database was introduced in the ICML 2013 Challenges in Representation Learning [1]. The database was created using the Google image search API and faces have been automatically registered. Faces are labeled as any of the six basic expressions as well as the neutral. The resulting database contains 35,887 images most of them in wild settings.

Table 2 shows the number of images for six basic expressions and neutral faces in each database.

Table 2. Number of images per each expression in databases

	AN	DI	FE	HA	NE	SA	SU
MultiPie	0	22696	0	47338	114305	0	19817
MMI	1959	1517	1313	2785	0	2169	1746
CK+	45	59	25	69	0	28	83
DISFA	436	5326	4073	28404	48582	1024	1365
FERA	1681	0	1467	1882	0	2115	0
SFEW	104	81	90	112	98	92	86
FER2013	4953	547	5121	8989	6198	6077	4002

* AN, DI, FE, HA, Ne, SA, SU stand for Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprised respectively.

4.2. Results

We evaluated the accuracy of the proposed deep neural network architecture in two different experiments; viz. subject-independent and cross-database evaluation. In the subject-independent experiment, databases are split into training, validation, and test sets in a strict subject independent manner. We used the K-fold cross validation technique with K=5 to evaluate the results. In FERA and SFEW, the training and test sets are defined in the database release, and the results are evaluated on the database defined test set without performing K-fold cross validation. Since there are different samples per emotion per subject in some databases, the training, validation and test sets have slightly different sample sizes in each fold. On average we used 175K samples for training, 56K samples for validation, and 64K samples for test. The proposed architecture was trained for 200 epochs (i.e. 150K iterations on mini-batches of size 250 samples). Table 3 gives the average accuracy when classifying the images into the six basic expressions and the neutral expression. The average confusion matrix for subject-independent experiments can be seen in Table 4.

Here, we also report the top-2 expression classes. As Table 3 depicts, the accuracy of the top-2 classification is 15% higher than the top-1 accuracy in most cases, especially in the wild datasets (i.e. FERA, SFEW, FER2013). We believe

that by assigning a single expression to a image can be ambiguous when there is transition between expressions or the given expression is not at its peak, and therefore the top-2 expression can result in a better classification performance when evaluating image sequences.

Table 3. Average Accuracy (%) for subject-independent

	Top-1	Top-2	State-of-the-arts
MultiPIE	94.7±0.8	98.7±0.3	70.6 [21], 90.6 [13]
MMI	77.6±2.9	86.8±6.2	63.4 [25], 74.7 [24], 79.8 [30], 86.9 [37]
DISFA	55.0±6.8	69.8±8.6	-
FERA	76.7±3.6	90.5±4.6	56.1 [25], 75.0 [2], 55.6 [47]
SFEW	47.7±1.7	62.1±1.2	26.1 [24], 24.7 [13]
CK+	93.2±1.4	97.8±1.3	84.1 [30], 84.4 [21], 88.5 [42], 92.0 [24] 92.4 [25], 93.6 [49]
FER2013	66.4±0.6	81.7±0.3	69.3[44]

Table 4. Average (%) confusion matrix for subject-independent

		predicted						
		AN	**DI**	**FE**	**HA**	**NE**	**SA**	**SU**
Actual	**AN**	**55.0**	7.0	12.8	3.5	7.6	8.5	5.3
	DI	1.0	**80.3**	1.8	5.8	8.5	2.2	0.1
	FE	7.4	4.3	**47.0**	8.1	18.7	8.6	5.5
	HA	0.7	3.2	2.4	**86.6**	5.5	0.2	1.0
	NE	2.3	6.3	7.8	5.5	**75.0**	1.3	1.4
	SA	6.0	11.3	8.9	2.7	13.7	**56.1**	0.9
	SU	0.8	0.1	2.8	3.5	2.5	0.6	**89.3**

The proposed architecture was implemented using the Caffe toolbox [16] on a Tesla K40 GPU. It takes roughly 20 hours to train 175K samples for 200 epochs. Figure 3 shows the training loss and classification accuracy of the top-1 and top-2 classification labels on the validation set of the subject-independent experiment over 150,000 iterations (about 150 epochs). As the figure illustrates, the proposed architecture converges after about 50 epochs.

In the cross-database experiment, one database is used for evaluation and the rest of databases are used to train the network. Because every database has a unique fingerprint (lighting, pose, emotions, etc.) the cross database task is much more difficult to extract features from (both for traditional SVM approaches, and for neural networks). The proposed architecture was trained for 100 epochs in each experiment. Table 5 gives the average cross-database accuracy when classifying the six basic expressions as well as the neutral expression.

The experiment presented in [30] is a cross-database experiment performed by training the model on one of the CK+, MMI or FEEDTUM databases and testing the model on the others. The reported result in Table 5 is the average results for the CK+ and MMI databases.

Different classifiers on several databases are presented in [37] where the results is still one of the state-of-the-art methods for person-independent evaluation on the MMI database (See Table 3). The reported result in Table 5 is the best result using different SVM kernels trained on the CK+ database and evaluated the model on the MMI database.

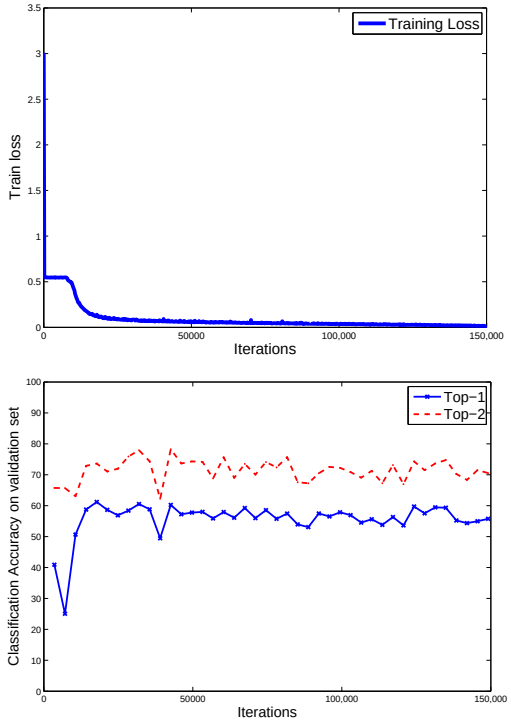

Figure 3. Training loss and classification accuracy on validation set

Table 5. Average Accuracy (%) on cross database

	top-1	top-2	[30]	[37]	[31]	[49]
MultiPIE	45.7	63.2	-	-	-	-
MMI	55.6	68.3	51.4	50.8	36.8	66.9
DISFA	37.7	53.2	-	-	-	-
FERA	39.4	58.7	-	-	-	-
SFEW	39.8	55.3	-	-	-	-
CK+	64.2	83.1	47.1	-	56.0	61.2
FER2013	34.0	51.7	-	-	-	-

A supervised kernel mean matching is presented in [31] which attempts to match the distribution of the training data in a class-to-class manner. Extensive experiments were performed using four classifiers (SVM, Nearest Mean Classifier, Weighted Template Matching, and K-nearest neighbors). The reported result in Table 5 are the best results of the four classifier when training the model on the MMI and Jaffe databases and evaluating on the CK+ database as well as when the model was trained on the CK+ database and evaluated on the MMI database.

In [49] multiple features are fused via a Multiple Ker-

nel Learning algorithm and the cross-database experiment is trained on CK+, evaluated on MMI and vice versa. Comparing the result of our proposed approach with these state-of-the-art methods, it can be concluded that our network can generalized well for FER problem. Unfortunately, there is not any study on cross-database evaluation of more challenging datasets such as FERA, SFEW and FER2013. We believe that this work can be a baseline for cross-database of these challenging datasets.

In [13], a Shared Gaussian Processes for multiview and viewinvariant classification is proposed. The reported result is very promising on MultiPIE database which covers multiple views, however on wild setting of SFEW it is not as efficient as MultiPIE. A new sparse representation is employed in [21], aiming to reduce the intra-class variation and by generating an intra-class variation image of each expression by using training images.

[47] is the FERA 2011 challenge baseline and [2] is the result of UC Riverside team (winner of the challenge). [42] detects AUs and uses their composition rules to recognize expressions by means of a dictionary-based approach, which is one of the state-of-the-art "sign-based" approaches. [44] is the winner of the ICML 2013 Challenges on FER2013 database that employed a convolutional neural network similar to AlexNet [20] but with linear one-vs-all linear SVM top layer instead of a Softmax function.

Table 6. Subject-independent comparison with AlexNet results (% accuracy)

	Proposed Architecture	AlexNet
MultiPie	94.7	94.8
MMI	77.9	56.0
DISFA	55.0	56.1
FERA	76.7	77.4
SFEW	47.7	48.6
CK+	93.2	92.2
FER2013	66.4	61.1

As a benchmark to our proposed solution, we trained a full AlexNet from scratch (as opposed to fine tuning an already trained network) using the same protocol as used to train our own network. As shown in Table 6, our proposed architecture has better performance on MMI & FER2013 and comparable performance on the rest of the databases. The value of the proposed solution over the AlexNet architecture is its training time - Our version of AlexNet performed more than 100M operations, whereas the proposed network performs about 25M operations.

5. Discussion

As shown in Tables 3 and 5, the results in the subject-independent tests were either comparable to or better than the current state of the art. It should be mentioned that we have compared our results with the best methods on each database separately, where the hyper parameters of the presented models are fine-tuned for that specific problem. We perform significantly better than the state of the art on MultiPIE and SFEW (no known state of the art has been reported for the DISFA database). The only exceptions to the improved performance are with the MMI and FERA databases. There are a number of explanations for this phenomenon.

One of the likely reasons for the performance discrepancies on the subject-independent databases is due to the way that the networks are trained in our experiments. Because we use data from all of the studied databases to train the deep architecture, the input data contains image that do not conform to the database setting such as pose and lighting. It is very difficult to avoid this issue as it is hard or impossible to train such a complex network architecture on so little data without causing significant overfitting. Another reason for the decreased performance is the focus on cross-database performance. By training slightly less complicated architectures, or even using traditional methods such as support vector machines, or engineered features, it would likely be possible to improve the performance of the network on subject-independent tasks. In this research however, we present a comprehensive solution that can generalize well to the FER "in the wild" problem.

6. Conclusion

This work presents a new deep neural network architecture for automated facial expression recognition. The proposed network consists of two convolutional layers each followed by max pooling and then four Inception layers. The Inception layers increase the depth and width of the network while keeping the computational budget constant. The proposed approach is a single component architecture that takes registered facial images as the input and classifies them into either of the six basic expressions or the neutral.

We evaluated our proposed architecture in both subject-independent and cross-database manners on seven well-known publicly available databases. Our results confirm the superiority of our network compared to several state-of-the-art methods in which engineered features and classifier parameters are usually tuned on a very few databases. Our network is first which applies the Inception layer architecture to the FER problem across multiple databases. The clear advantage of the proposed method over conventional CNN methods (i.e. shallower or thinner networks) is gaining increased classification accuracy on both the subject independent and cross-database evaluation scenarios while reducing the number of operations required to train the network.

7. Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] Challenges in representation learning: Facial expression recognition challenge: http://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge. 2, 5, 6

[2] Facial expression recognition and analysis challenge 2011 (results): http://sspnet.eu/fera2011/. 7, 8

[3] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013. 4

[4] E. Bal, E. Harden, D. Lamb, A. Van Hecke, J. Denver, and S. Porges. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of Autism and Developmental Disorders*, 40(3):358–370, 2010. 4

[5] T. Bänziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, pages 271–294, 2010. 5, 6

[6] J. F. Cohn, Z. Ambadar, and P. Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, pages 203–221, 2007. 2

[7] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual analysis of humans*, pages 377–409. Springer, 2011. 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[9] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013. 1, 3, 6

[10] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 2011. 5, 6

[11] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 1

[12] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 2

[13] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multi-view and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 24(1):189–204, 2015. 7, 8

[14] W. V. Friesen and P. Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2:36, 1983. 2, 6

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 1, 2, 5

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 7

[17] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013. 2, 3

[18] M. Kenji. Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 74(10):3474–3483, 1991. 2

[19] H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, volume 4, pages 3732–3737. IEEE, 1997. 2

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 3, 4, 8

[21] S. H. Lee, K. Plataniotis, and Y. M. Ro. Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing*, page 1, 2014. 7, 8

[22] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 3, 4

[23] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, 11(4):467–476, 2002. 2

[24] M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013. 3, 7

[25] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Computer Vision–ACCV 2014*, pages 143–157. Springer, 2014. 4, 7

[26] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014. 4

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010. 2, 5

[28] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face*

and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pages 200–205. IEEE, 1998. 1

[29] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013. 1, 2, 5

[30] C. Mayer, M. Eggers, and B. Radig. Cross-database evaluation for facial expression recognition. *Pattern recognition and image analysis*, 24(1):124–132, 2014. 1, 7

[31] Y.-Q. Miao, R. Araujo, and M. S. Kamel. Cross-domain facial expression recognition using supervised kernel mean matching. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 326–332. IEEE, 2012. 7

[32] M. Mohammadi, E. Fatemizadeh, and M. Mahoor. Pca-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, 25(5):1082 – 1092, 2014. 2

[33] A. Mollahosseini, G. Graitzer, E. Borts, S. Conyers, R. M. Voyles, R. Cole, and M. H. Mahoor. Expressionbot: An emotive lifelike robotic face for face-to-face communication. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 1098–1103. IEEE, 2014. 1

[34] A. Mollahosseini and M. H. Mahoor. Bidirectional warping of active appearance model. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 875–880. IEEE, 2013. 4

[35] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005. 1, 5

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014. 3, 4

[37] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 2, 7

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3

[39] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015. 4

[40] J. M. Susskind, A. K. Anderson, and G. E. Hinton. The toronto face database. *Technical report, UTML TR 2010-001, University of Toronto*, 2010. 3

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 1, 3, 4

[42] S. Taheri, Q. Qiang, and R. Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 23(8):3590, 2014. 2, 7, 8

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 1

[44] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013. 7, 8

[45] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014. 1

[46] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 635–640. IEEE, 2004. 2

[47] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926. IEEE, 2011. 2, 6, 7, 8

[48] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 4

[49] X. Zhang, M. H. Mahoor, and S. M. Mavadati. Facial expression recognition using {l} _ {p}-norm mkl multiclass-svm. *Machine Vision and Applications*, pages 1–17, 2015. 2, 7

[50] X. Zhang, A. Mollahosseini, B. Kargar, H. Amir, E. Boucher, R. M. Voyles, R. Nielsen, and M. Mahoor. ebear: An expressive bear-like robot. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 969–974. IEEE, 2014. 2

[51] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007. 2

[52] W. Zhen and Y. Zilu. Facial expression recognition based on local phase quantization and sparse representation. In *Natural Computation (ICNC), 2012 Eighth International Conference on*, pages 222–225. IEEE, 2012. 2