A Compact Deep Learning Model for Robust Facial Expression Recognition

Chieh-Ming Kuo, Shang-Hong Lai
National Tsing Hua University
Hsinchu, Taiwan
lai@cs.nthu.edu.tw

Michel Sarkis
Qualcomm Technologies Inc.
San Diego, USA
msarkis@qti.qualcomm.com

Abstract

In this paper, we propose a compact frame-based facial expression recognition framework for facial expression recognition which achieves very competitive performance with respect to state-of-the-art methods while using much less parameters. The proposed framework is extended to a frame-to-sequence approach by exploiting temporal information with gated recurrent units. In addition, we develop an illumination augmentation scheme to alleviate the overfitting problem when training the deep networks with hybrid data sources. Finally, we demonstrate the performance improvement by using the proposed technique on some public datasets.

1. Introduction

Understanding human emotion from images has become increasingly important with the recent advances in deep learning and human computer interaction. Human emotion is expressed in multiple ways. Studies show that analysis of non-posed expression must rely on additional physiological signals, such as temperature dynamics and heart rate [44, 30, 38, 20, 37]. Unfortunately, these physical measurements are usually unavailable or infeasible to obtain in practice, which makes the research findings restricted to be used in laboratory environment.

Due to the ease of data acquisition, the video-based approach is most commonly used for expression recognition. Databases [27, 46] with quite restricted settings are usually used for performance benchmark for facial expression recognition. Traditional image-based methods for facial expression recognition employed hand-craft features, like LBP [29], BoW [35], HoG [5], or SIFT [26], and they have shown quite good results on several databases [27, 46, 28, 40, 11]. Furthermore, the sequence-based approach further modeled the temporal emotion variations with temporal hand-craft features extracted from videos [16, 24, 12, 19].

Recently, expression recognition in the wild [7, 6] has attracted considerable amounts of attention. This type of problem is challenging because the face images collected from internet are usually acquired under different illumination conditions and head poses. Researches, like EmotioNet [8], also showed that using downloaded images into the training set is quite useful to improve the generalization of model training. This inspires us to further investigate how the expression recognition task could benefit from model training from face image datasets acquired from unconstrained environments.

In this paper, we introduce a new convolutional neural networks (CNN) architecture for improving the performance and generalization with proper design of the deep networks. Experiments on standard databases also show that the proposed CNN model is appropriate for facial expression recognition with compact network parameters compared to the related deep learning based models. Moreover, we include several datasets of different types into the training dataset to improve the generalization of the learned CNN model. In addition, we develop an illumination augmentation scheme to improve the robustness of training the proposed CNN model. The main contributions in this paper can be summarized as follows:

- We propose a compact CNN model for facial expression recognition to compromise between recognition accuracy and model size.
- We evaluate our network model on two standard databases and show the proposed method is superior to the state-of-the-art methods.
- We collect three datasets of different scenarios which could be used to evaluate the cross-domain performance.
- We present leave-one-set-out experiments showing that the proposed illumination augmentation strategy alleviates the overfitting problem for model training with images from different sources.

2. Related Work

BDBN [25] showed that combination of feature extraction and selection with a unied boosted deep belief network achieved better performance. STM-ExpLet [24] used expressionlet-based spatio-temporal manifold to model the expression video clips. Exemplar-HMMs [33] combined HMMs and SVMs in a model-based similarity framework. The LOMo [34] combined different types of complimentary features, such as facial landmarks, LBP, SIFT, and geometry features, for expression recognition.

In recent years, deep learning has become very popular since CNN showed its unprecedented capability in many computer vision tasks. Various CNN models [22, 36, 13, 15] have been proposed for different image classification tasks [23, 21, 9]. However, these deep networks are not appropriate for small expression recognition databases.

The joint-fine tuning method [16] adopted data augmentation strategy with 7 different rotation angles to obtain 14 times more data. They train two different networks based on appearance and geometry features and combine the pretrained networks by joint-fine tuning. Researchers also showed that combining CNNs with recurrent neural networks (RNNs) performed excellently for expression recognition from video [17].

Recently, the peak-piloted method [47] successfully applied the GoogLeNet [39] for expression recognition by transferring the knowledge learned from large-scale face recognition database [43]. Their result also showed that the accuracy of the image-based approach is comparable to those of the sequence-based methods.

3. Proposed Framework

The overall pipeline of the proposed deep learning approach is depicted in Figure 1. Our framework is composed of two modules: Face pre-processing and CNN classification. To ensure our framework could be extended to different scenarios, we do not adopt any temporal normalization method [46] like [16].

3.1. Preprocessing

We first cropped the face region according to landmark points detected by IntraFace [42]. These landmarks could be used to extract the contour of eyebrows, eyes, nose and mouth. Large crop sizes could keep more information while small crop sizes could reduce noise come from background or head contour. In our implementation, the cropped image size L is determined by $L = \alpha \times max(d_v, d_h)$, where d_v is the distance between the uppermost landmark point and the lowermost landmark point, d_h is the horizontal distance between the leftmost landmark point and the rightmost landmark point, and α is a scalar used to control the size of face region. We set α to 1.05 for all experiments in this paper.

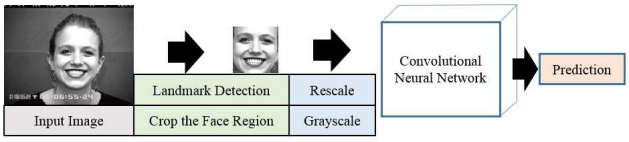

Figure 1. The proposed framework for image-based facial expression recognition. The CNN classifier takes a single gray-scale image as input, and output the corresponding expression category.

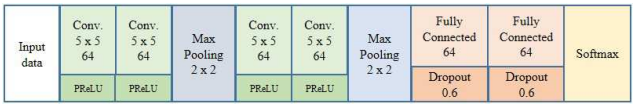

Figure 2. The architecture of our CNN model. Each convolution layer is equipped with a ReLU activation function. Dropout [14] is used after the fully-connected layer to prevent overfitting.

Once the crop size L is determined, we crop the face region center on the landmark point of nose and obtain modest face images for model training. The cropped images are resized to a fixed size 120×120, which is subsequently sent to the CNN classifier for expression recognition.

3.2. The CNN Model

The architecture of our CNN model is depicted in Fig. 2. Our model is composed of two convolutions and pooling blocks, followed by two fully-connected layers. We use ReLU [22] as activation function for each convolution layer. Dropout [14] is also applied after the fully-connected layers to preventing overfitting. Note that our model only uses the central 96×96 part of the resized face image as input. Details about model training will be described in the next subsection.

The proposed CNN structure could be considered as an improved version of DTAN in [16]. Their experiment already showed that this plain model could achieve good results for the expression recognition task. To further increase the discriminating power of our model, we stack two continuous convolution layers before the max pooling, like [36]. We also use bigger convolution filters which allow the neurons inside our model have larger receptive fields. After this modification, the receptive field of each neuron in the first fully-connected layer would become 36×36, which is about 14% of input 96×96 image, while the origin DTGAN is 16×16, which is only 6% of its input size 64×64.

Another important modification is that we substantially reduce the number of fully-connected neurons. We believe that expression in human face could be learned by a modest model size as long as we have suitable design of the receptive field. Later experiments given in this paper demonstrate that a suitable lightweight fully-connected network is not only compact in terms of model parameters but also accu-

rate for facial expression recognition.

3.3. The Frame-to-Sequence Model

An image sequence in standard facial expression databases usually begins with a neutral expression and gradually proceeds to a peak expression. We could approximate this transformation process by a model $S(x)$, which takes a sequence of images $x_i^t, t = 1, ..., T$, as input and mapping each image sequence to its ground truth y_i as close as possible:

$$y_i \cong \widetilde{Y}_i = S(x_i^1, ..., x_i^T; \theta), \tag{1}$$

where T is the length of the image sequence and θ is a set of model parameters. Let p denote the probability of each expression produced by the sequence model, the sequence modeling problem could be formulated as maximizing the log-likelihood of a model given a training sequences, i.e.

$$\hat{\theta} = \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log p(\widetilde{Y}_i | x_i^1, ..., x_i^T; \theta) \tag{2}$$

Such a problem is difficult to be solved directly, thus we adopt a similar idea like [17] to use a pre-trained CNN as the feature extractor. The previous frame-based approach could be regarded as a mapping function $F(x)$, which maps each sample x_i^t to a probability distribution $\{p_1^t(j), j = 1, ..., m\}$ such that the index of the maximal probability $p_i^t(j)$, $\widetilde{y_i}$, is the same as its correct category y_i, i.e.

$$y_i \cong \widetilde{y_i} = \arg\max_{j} p_i^t(j) = F(x_i), \tag{3}$$

where $p_i^t = F(x_i) = [p_i^t(1), p_i^t(2), ..., p_i^t(m)]$.

Here, we use a sequence of probability distributions computed from the CNNs for the frame-based expression recognition instead of images as input for the expression recognition, which means

$$y_i \cong \widetilde{Y}_i = S(F(x_i^1), ..., F(x_i^T); \theta) \tag{4}$$

We can model $S(x)$ with a Gated Recurrent Neural Network [3]. Because we use the probability distribution for the frame-based classification as the feature representation, we expect that $S(x)$ could be well modeled by a shallow structure. The architecture of our frame-to-sequence model is composed of a single Gated Recurrent Units (GRU) layer with 128 hidden states and a softmax layer. The overall framework is shown in Figure 3.

3.4. Model Training

Small facial expression recognition datasets usually contain hundreds of image sequences, which may easily lead to the over-fitting problem during model training. For the model training, we adopt online augmentation strategy with both horizontally flipping and random shifting like [36]. In this paper, we set the maximal training iterations to 2000 epochs and report the best validation accuracy for training CNN model. For the frame-to-sequence model, we use ADAM [18] optimizer for the model training and run 10,000 iterations with the batch size set to 48 and at a fixed learning rate 0.01.

Table 1. Expression recognition accuracies of different methods on the CK+ database. The best result is marked in boldface.

Method	Input	Accuracy
BDBN [25]	Sequence	96.7
LOMo [34]	Sequence	92.0
Exemplar-HMMs [33]	Sequence	94.60
STM-ExpLet [24]	Sequence	94.19
DTGAN [16]	Sequence	97.25
Peak-Piloted [47]	Frame	**99.30**
Ours-frame	Frame	97.37
Ours-frame2seq	Sequence	98.47

4. Experiments on Standard Databases

The standard facial expression databases usually contain video sequences begins with a neutral expression and proceeds to a peak expression. For the frame-based approach, we use only the peak image for training and validation like [47]. We first evaluate the proposed framework on two well-known benchmark databases: the Extended Cohn-Kanade (CK+) database [27] and the Oulu-CASIA database [46].The CK+ database is composed of 327 labeled image sequences with seven emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. The Oulu-CASIA database contains 480 image sequences with one of six emotion labels: anger, disgust, fear, happiness, sadness, surprise. The resolution of CK+ and Oulu-CASIA database are 640×490 and 320×320, respectively. Details of these databases are shown in Table 4. For CNN model training, all weighted layers were initialized by xavier [10], the learning rate was fixed to 0.001 with the momentum set to 0.9. The weight decay method is also used for regularization with a factor of 0.001.

4.1. Frame-based Approach

To avoid subjects appearing in both the training and testing sets simultaneously, we divide the subjects into 10 subsets by their IDs in the ascending order, which is the same as the 10-fold cross validation protocol in [24]. This experiment protocol is used for all methods included in the experimental comparison in this paper.

The overall accuracy of 10-fold cross validation on CK+ database is shown in Table 1. The accuracy of our frame-based approach outperforms most sequence-based methods and is second only to the peak-piloted method [47]. How-

Table 2. Expression recognition accuracies of different methods on the Oulu-CASIA database. The best result is marked in boldface.

Method	Input	Accuracy
HOG3D [19]	Sequence	70.63
LOMo [34]	Sequence	74.00
Exemplar-HMMs [33]	Sequence	75.62
STM-ExpLet [24]	Sequence	74.59
Atlases [12]	Sequence	75.52
DTGAN [16]	Sequence	81.46
Peak-Piloted [47]	Frame	84.59
Ours-frame	Frame	88.75
Ours-frame2seq	Sequence	**91.67**

ever, in [47], they pre-trained the CNN model with additional 500k images [43] that is over one thousand times more than the CK+ database size. The result shows that the proposed CNN model is very suitable for learning facial expression on a small database.

Expression recognition on the Oulu-CASIA database is more challenging because it contains more low-intensity expressions which are difficult to distinguish with insufficient image resolution. However, the Oulu-CASIA database is still a good benchmark with complete emotion samples for each subject. The results of our method outperformed all the other methods in the Oulu-CASIA database, as shown in Table 2.

The result shows that our framed-base CNN model has about four percent improvement compared to the state-of-the-art method [47]. Another strength of our framework is that we have only 8.62% absolute performance difference between these two databases while the state-of-the-art method [47] has 14.71% performance gap. The high recognition accuracy on the Oulu-CASIA database suggests that the proposed framework could maintain its discriminative power for more strict cases. Furthermore, our frame-based approach could be further extended to a sequence-based approach to boost recognition accuracy by exploiting temporal information.

4.2. Frame-to-Sequence Approach

To further exploit temporal information and improve the recognition accuracy, we develop a frame-to-sequence approach which uses multiple image frames as input and then produces a single prediction from the whole input sequence.

To avoid the problem due to sequence length deviations, we perform systematic uniform sampling in each of the original image sequence to normalize all training image sequences to a fixed length 9, which is also the shortest sequences length in Oulu-CASIA database.

For the model training, the sampled sequences were first augmented with mirror and random cropping as mentioned

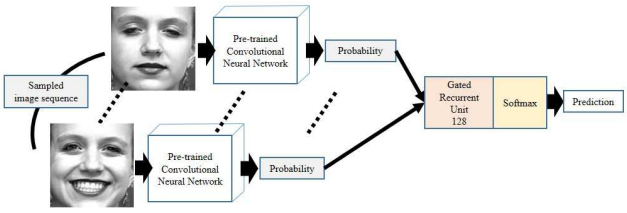

Figure 3. Framework of the proposed frame-to-sequence approach. The frame-to-sequence model takes features extracted by the pre-trained CNN model and uses their softmax outputs for classification.

in the previous section to become 100 times more. After that, we use our well-trained single frame CNN model to obtain the probability distribution for each frame, so every image sequence is represented as a 6-by-9 matrix for Oulu-CASIA dataset or 7-by-9 for CK+ dataset. That is, the model takes a feature vector of a single frame each time and then gives a classification decision after receiving the 9-th input. Note that all of the frame-to-sequence models use independent CNN models as the feature extractor which are trained with training data only. That is, the experiments performed for the sequence-based approach still follow the standard 10-fold cross-validation protocol.

As shown in Table 1 and 2, the proposed frame-to-sequence approach actually improves the performance for both CK+ and Oulu-CASIA databases. We can see that the improvement on Oulu-CASIA database is much more than that for the CK+ database because there are more weak-expression samples in Oulu-CASIA database which is hard to distinguish from the last frame only. Even though our method is still only second to to the peak-piloted method on the CK+ database, the performance gap between CK+ and Oulu-CASIA for our method is reduced to 6.8%. It indicates that our approach is a more generic solution for facial expression recognition in standard settings.

4.3. Implementation for Real-World Applications

In the previous section, we show that the propose method is superior to the previous methods on both accuracy and generalization. However, from the consideration of real-world application, an inevitable problem is the limitation of hardware storage and computation capability. The system developed with standard databases is difficult to be applied in practice since almost all face images are frontal faces in these datasets.

4.4. Parameter Efficiency

To overcome the storage of hardware limitation, we further reduce the number of convolution filters used in our CNN structure. This tiny version of the proposed CNN model only uses 16 filters in the first two convolution lay-

Table 3. Comparison of model sizes. Our original model only uses 50% fewer parameter and it gives the best result in average. The tiny model version of our method further reduces the model parameters to about 20% fewer, while keeping competitive performance.

CNN frameworks	AC. on CK+	AC. on OuluCASIA	number of parameters
DTGAN [16]	97.25	81.46	5950K
Peak-Piloted [47]	99.30	84.59	6798K
Ours-frame	97.37	88.75	2673K
Ours-frame2seq	98.47	91.67	2690K
Ours-Tiny	96.81	85.84	1229K

ers and 32 filters in the last two convolution layers, which is 25% and 50% of the number of filters in the original version. We repeat the same 10-fold validation as described in the previous section, the results are shown in Table 3. Even though this modification comes with a little accuracy drop, the tiny model only uses 50% fewer parameters compared to that of the original version, which is about 80% fewer than the state-of-the-art CNN methods [47, 16]. The small number of parameters makes the tiny model suited for portable devices or IoT applications with modest storage sizes. The average inference time of the tiny model is also reduced to 16ms from 21ms on a single NVIDIA GTX 970 GPU.

5. Experiments on Self-Collected Databases

Data collection for facial expression recognition is expensive and time-consuming. Research [8] indicates that using images downloaded from the Internet is helpful to model training for the expression recognition problem. To this end, we collect three additional datasets to improve the training of facial expression recognition, each representing specific data source. Moreover, to prevent subjective annotation, each dataset we collected is labeled with different approaches to ensure the annotation qualities. All datasets we collected were composed of six common expressions, like the Oulu-CASIA database [46], and an additional neutral expression, because the neutral face occurs most frequently. These datasets are hereinafter referred to as set A, set B and set C, respectively.

Set A was collected to represent applications in laptop scenario. In this dataset, we have 26 subjects and each of them was asked to sit on a chair in the lab and watch a series of videos which lead to different expressions on the faces of the subjects. The whole watching process was been recorded by three webcams from different near-frontal view angles simultaneously to enrich the head pose changes. Right after the subjects finished the task, we asked them to annotate the time intervals and the associated expressions by themselves. Then we clean these video clips

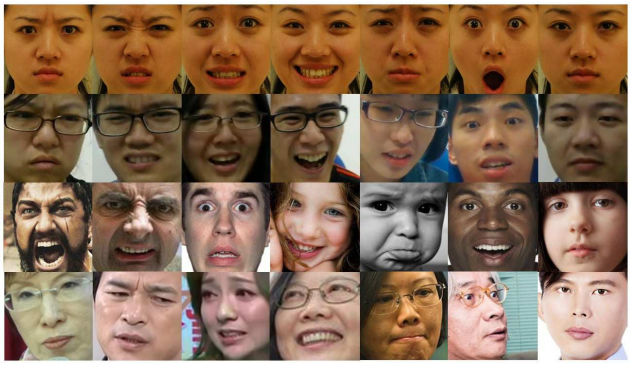

Figure 4. Examples of databases used for cross dataset evaluation. (Left to right) anger, disgust, fear, happiness, sadness, surprise and neutral, respectively. (Top to bottom) the TFEID database, set A, set B and set C, respectively. Variations in head poses and illumination conditions make the task of expression recognition more challenging.

to extract only the frames with peak expressions according to the subject's label. To make sure the data has a uniform distribution for all expressions, we sample 50 images for each category.

Set B was collected from Google image search engine with keywords like anger face or neutral face, search results were then annotated by the keyword used. Compared to set A, the images downloaded from internet usually have much more head pose variations, occlusion and even watermarks that could make the system trained from the standard datasets fail easily.

The images in set C were collected from movies, dramas, news or TV shows, and these images were then labeled according to its story plot or scenario. Even though set C is also collected from the Internet like set B, they are quite different. Face images captured from movie usually contain strong illumination contrast and large head pose variation, which makes the samples in this dataset more complex than set B. We depict some examples of the additional databases in Figure 4 and the list of these databases is given in Table 4.

5.1. Leave-One-Set-Out Experiments

The generalization capability describes the viability of a framework in practice. To investigate the difficulty of CNN model generalization with unseen data type, we train our tiny-model using datasets with hybrid data type and then evaluate the generalization on the other data type, which we call leave-one-set-out experiment. We choose the first frame of anger category in the Oulu-CASIA database as its neutral sample so that it can be used in the following experiments. Another high-quality and well-posed database TFEID [2] is also included in this experiment.

In the leave-one-set-out experiments, we use our tiny-

Table 4. List of databases used in this paper. Note that we add neutral expression in the Oulu-CASIA by selecting the first frame in the anger sequences for all subjects. This approach is not suitable for the CK+ database because it may produce too many neutral samples due to non-uniform sequence distribution for all subjects.

Database	Ang.	Con.	Dis.	Fea.	Hap.	Sad.	Sur.	Neu.	Total
CK+	45	18	59	25	69	28	83	—	327
OuluCASIA	80	—	80	80	80	80	80	80	560
TFEID	34	68	40	40	40	39	36	39	268
set A	50	—	50	50	50	50	50	50	350
set B	50	—	50	50	50	50	50	50	350
set C	50	—	50	50	50	50	50	50	350
RAF	867	—	877	355	5957	2460	1619	3204	15339
GENKI	2162(smile)				1838(non-smile)				4000

Table 5. Leave-one-set-out validation accuracy with different illumination preprocessing. The notation w/o Oulu means this experiment uses Oulu-CASIA as validation data and the training uses the rest of datasets, and so on. Note that "No" means we did not perform any illumination normalization method. We abbreviated illumination normalization process as HE, LM and WS each represent histogram equalization, linear mapping and weighted summation method, respectively. The notation "Aug" means we use all normalization methods to augment the training set. The best 2 results in each experiment setting are marked in boldface.

Training	Validation	w/o Oulu	w/o TFEID	w/o set A	w/o set B	w/o set C	Average	AVE. improve
No	No	54.29	83.21	52.86	55.71	51.43	59.50	0
HE	HE	50.89	81.34	55.14	53.43	50.00	58.16	-1.34
LM	LM	57.32	85.07	54.00	57.71	**55.29**	61.88	2.38
WS	WS	55.36	83.95	**58.29**	55.71	52.57	61.18	1.68
Aug	No	56.43	**89.92**	54.29	**59.43**	52.29	62.47	2.97
Aug	HE	**58.93**	88.43	57.43	**58.57**	**54.86**	**63.64**	**4.41**
Aug	LM	**57.86**	86.94	**57.71**	56.29	54.57	**62.67**	**3.17**
Aug	WS	57.14	**89.55**	56.29	56.57	53.43	62.60	3.10

model for evaluation with the same hyper-parameter settings with the 10-fold validation described in the previous section. The experimental result is shown in the first row of Table 5. The high validation accuracy of the TFEID database suggests that the model training with hybrid data type could still learn a representation that generalizes well to the ideal case, like frontal face images with strong expressions. The low performance on the Oulu-CASIA database may be owing to the bias in the training data, which is mainly composed of Asian face samples. However, we could improve the recognition accuracy with appropriate augmentation strategy in the model training.

5.2. Illumination Normalization

The image data collected from the Internet usually comes with diverse illumination conditions that may hinder the model training. Illumination normalization is widely used in various computer vision tasks. We take histogram equalization [31] and linear mapping, which maps the minimum and maximum pixel values to an interval $[0, 1]$ by a linear transformation, into the comparison. However, directly applying histogram equalization may overemphasize local contrast as shown in Figure 5 and linear mapping did not work well when the image already have large global contrast. Therefore, we propose a weighted summation approach to take advantage of both normalization methods:

$$I_{ws}(x, y) = (1 - \lambda) \times I_{he}(x, y) + \lambda \times I_{lm}(x, y),$$

where λ is a weight factor which decides how much the pixel of the combined image I_{ws} takes reference from the histogram equalized image I_{he} and the linearly mapped image I_{lm}, we set λ to 0.5 in our implementation. Some results are shown in the rightmost column in Figure 5. To do a fair comparison between these normalization methods, we apply each of normalization method on both training and validation data and evaluate it with the leave-one-set-out protocol.

The experimental results are reported in the upper four rows in Table 5. Even though linear mapping brings the largest performance boost among all these methods, the weighted summation approach achieves the highest improvement, about 6.5%, when using set B as validation.

Figure 5. Examples of different illumination normalization methods. (Left to right) original image, histogram equalization, linear mapping and weighted summation, respectively. (Top to bottom) the TFEID database, set B and set C, respectively. Histogram equalization sometimes overemphasize local contrast or watermarks as shown in the upper two rows of middle left column. Linear mapping could not enhance the contrast of image with large global contrast like the sample in the third row of middle right column.

5.3. Illumination Augmentation

To take advantage of all the aforementioned normalization methods, an intuitive idea is to use all of them as training data to form the illumination augmentation approach. For the validation, we still use only one of the above normalization methods to keep the whole framework stay concise during the classification process. As shown in the lower four rows in Table 5, no matter what kind of normalization method the validation set is applied, using the proposed illumination augmentation strategy actually improves the overall accuracy.

More interestingly, if we only apply one of normalization strategy, the histogram equalization method was the only one which degrades the average prediction accuracy. Nevertheless, When we adopt the proposed illumination augmentation strategy and only apply histogram equalization on the validation data, we will gain the largest performance boost instead. One possible explanation to this phenomenon could be that mixing the illumination normalization methods encourages the CNN model to learned representations with better robustness against illumination variations during the training process.

6. Experiments on "In The Wild" Databases

In this section, we further test the proposed framework on two public in the wild datasets, the MPLab-GENKI [41] dataset and the newest Real-world Affective Faces (RAF) Database [32]. Because the RAF database has serious data imbalance problem as shown in Table 4, we duplicate the number of image in class with fewer samples, so that the model would see those rare samples more often during the training.

For the GENKI dataset, our model achieved 95.33% average accuracy and 0.34% standard deviation by using the 4-fold cross-validation protocol used in the previous works, as shown in Table 7. Even though the GENKI dataset only has two categories: smile and non-smile, it is still a challenging dataset with very low resolution face parts. The smallest face image detected only has resolution about 20-by-20 pixels. We found that there is no accuracy boost from the proposed illumination augmentation strategy because most images in the GENKI dataset are captured under well illuminated conditions and the accuracy on this dataset is already close to saturation. However, the result achieved by our method is still the state-of-the-art and it is obtained without any further tuning.

For the RAF database, our model achieves 65.52% accuracy under their evaluation metric, the performance could be further improved to 67.55% with the proposed illumination augmentation strategy. Our result is competitive as both VGG and AlexNet models achieved 58.22% and 55.60% accuracies, respectively. However, our model is much smaller in size. It is about only 5% of the size of those two previous models. Even though our result is inferior to the best accuracy 74.20% reported with DLP-CNN [32], but our model used 87.45% less parameters. In addition, all the other models were used only for feature extraction and an additional multiclass SVM [1] was needed as a classifier while our CNN model provides an end-to-end expression recognition system.

As shown in Table 6, training with illumination augmentation achieves higher accuracy on both disgust and fear categories compared to direct training, and they are usually more difficult to classify. The improvement on average accuracy and the reduction of standard deviation across performance of different emotions also indicated that the proposed illumination augmentation strategy could help the CNN model to learn a more general feature representation as we mentioned in the previous section.

7. Conclusion

In this paper, we proposed a new CNN architecture for facial expression recognition which outperforms the state-of-the-art methods. The frame-to-sequence approach successfully exploits temporal information and it improves the accuracies on the public benchmarking databases. The proposed framework was demonstrated to provide better generalization while keeping high parameter efficiency, which is a very important issue for applications on portable devices. Experimental results showed that the proposed system provides the state-of-the-art accuracies for facial expression recognition on several datasets. We also demon-

Table 6. Expression recognition accuracies and the CNN model sizes of different methods on the RAF database.

Method	Ang.	Dis.	Fea.	Hap.	Sad.	Sur.	Neu.	Ave.	Std.	Model size	Ratio
VGG + mSVM[32]	68.52	27.50	35.13	85.32	64.85	66.32	59.88	58.22	18.63	54458K	20.3
AlexNet+mSVM[32]	58.64	21.87	39.19	86.16	60.88	62.31	60.15	55.60	18.68	43501K	16.2
DLP-CNN+mSVM[32]	71.60	52.15	**62.16**	**92.83**	**80.13**	81.16	**80.29**	**74.20**	**12.56**	19655K	7.35
Ours-frame	**82.07**	44.59	41.25	81.01	44.14	**90.12**	75.44	65.52	19.64	**2673K**	**1**
Ours-frame*	74.47	**67.57**	46.88	82.28	57.95	84.57	59.12	67.55	12.78	**2673K**	**1**

Table 7. Expression recognition accuracies of different methods on the GENKI database.

Method	Accuracy(%)
Pair-wide Distance Vector [4]	93.42 ± 1.46
CNN-2Loss [45]	94.60 ± 0.29
Ours-frame	$\mathbf{95.33 \pm 0.34}$
Ours-frame*	95.15 ± 0.44

strate that the proposed illumination augmentation strategy is very effective through experiments on the public "in the wild" databases.

References

[1] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[2] L.-F. Chen and Y.-S. Yen. Taiwanese facial expression image database . brain mapping laboratory, institute of brain science, national yang-ming university, taipei, taiwan., 2007.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[4] D. Cui, G.-B. Huang, and T. Liu. Smile detection using pair-wise distance vector and extreme learning machine. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 2298–2305. IEEE, 2016.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] A. Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[7] 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.

[8] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. 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.

[9] L. Fei-Fei. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, 2010.

[10] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. 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.

[12] Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *Computer Vision–ECCV 2012*, pages 631–644. Springer, 2012.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[14] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[15] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[16] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[17] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang. How deep neural networks can improve emotion recognition on video data. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 619–623. IEEE, 2016.

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[20] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[21] A. Krizhevsky, V. Nair, and G. Hinton. The cifar-10 dataset, 2014.

[22] 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.

[23] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, 1998.

[24] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[25] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[26] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[28] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

[29] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[30] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[31] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[32] W. D. Shan Li and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017.

[33] K. Sikka, A. Dhall, and M. Bartlett. Exemplar hidden markov models for classification of facial expressions in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–25, 2015.

[34] K. Sikka, G. Sharma, and M. Bartlett. Lomo: Latent ordinal model for facial analysis in videos. *arXiv preprint arXiv:1604.01500*, 2016.

[35] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*, pages 250–259. Springer, 2012.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[38] G. Stemmler. Methodological considerations in the psychophysiological study of emotion. *Handbook of affective sciences*, pages 225–255, 2003.

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[40] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.

[41] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):2106–2111, 2009.

[42] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[43] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[45] K. Zhang, Y. Huang, H. Wu, and L. Wang. Facial smile detection based on deep learning features. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 534–538. IEEE, 2015.

[46] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[47] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. In *European Conference on Computer Vision*, pages 425–442. Springer, 2016.