Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator

Mundher Al-Shabi*, Wooi Ping Cheah, Tee Connie

FIST, Multimedia University, Melaka, Malaysia
*mundher.ahmed@hotmail.com

Abstract: Recognizing facial expression has remained a challenging task in computer vision. Deriving an effective facial expression recognition is an important step for successful human-computer interaction systems. This paper describes a novel approach towards facial expression recognition task. It is motivated by the success of Convolutional Neural Networks (CNN) on face recognition problems. Unlike other works, we focus on getting good accuracy results while requiring only a small sample data to train the model by merging the CNN and SIFT features. The proposed classification model is an aggregation of multiple deep convolutional neural networks and a hybrid CNN-SIFT classifiers. The goal of using SIFT features is to increase the performance on small data as SIFT does not require large training data to generate useful features. The model has been tested on FER-2013, SFEW 2.0 and CK+. The results showed how CNN-SIFT features improve the accuracy when added as a voting member in an ensemble classifier. It generates state-of-art results on FER-2013 and CK+ datasets, where it achieved 73.58% on FER-2013 and 99.35% on CK+.

1. Introduction

Automatic facial expression recognition is an interesting and challenging problem which has important applications in many areas such as human-computer interaction. It could help to build more intelligent robots with the ability to understand human emotions. Many other real-world applications such as call centre and interactive game development also benefit from such intelligence.

Ekman in early 1970s shown that there are six universal emotional expressions across all cultures. Those facial expressions include disgust, anger, happiness, sadness, surprise and fear [1]. The expressions could be identified by observing face signals. For example, a smile (raising of the mouth corners and tightening of the eyelids) is a signal of happiness.

Due to the importance of facial expression in designing Human–computer interaction systems, various feature extraction and machine learning algorithms have been developed for Facial Expression Recognition. Most of these methods are hand-crafted features extractions followed by a classifier such as [2] who's used Local binary pattern feature extractor with SVM classification, Haar[3], SIFT[4], Gabor filters with fisher linear discriminant[5], and Local phase quantization (LPQ) [6].

The recent success of convolutional neural networks (CNNs) in tasks such as image classification[7] has been extended to the problem of facial expression recognition[8]. Unlike traditional machine learning and computer vision approaches where features are defined by hand, CNN learns to extract the features directly from the training database using iterative algorithms like gradient descent. CNN is usually combined with feed-forward neural network classifier which makes the model end-to-end trainable on the dataset.

Like the ordinary neural network, CNN learns its weights using back-propagation algorithm. It has two main features local receptive fields and shared weights. In local receptive fields, each neuron connected to a local group of input space. The size of this group of pixels is equal to the filter size. The second feature, in CNN the same weights and bias used over all local receptive fields. These two features make CNN run faster and prone to over-fitting as the same weights applied to the entire image.

In most cases, CNN requires many training data to generalize very well. The availability of big datasets and the cheap computational power provided by the GPU increase the popularity of CNN. However, this is not the case in facial expression recognition where the datasets are limited. While Scale Invariant Feature Transform (SIFT)[9] and other hand-crafted methods provide less accurate results than CNN[10]–[12], they do not require extensive datasets to generalize. The limitation of the hand-crafted method is that their modeling capacities are limited by the fixed transformations (filters) that stay the same for a different source of data. In this paper, we propose a hybrid approach by combining SIFT and CNN to get the best of both worlds. We compare the individual CNN models, SIFT-CNN, and also aggregation over all the models. The methods are evaluated on three datasets, namely FER-2013, SFEW 2.0 and CK+. The contributions of this paper are two-fold: 1) we investigate the impact of combining SIFT with CNN feature to increase the performance on small data, and 2) designing a novel classifier for facial expression recognition by aggregating various CNN and SIFT models which achieved a state of art results on both FER-2013 and CK+ datasets.

2. Related Work

Automatic recognition of facial expressions has been an active research for a long time. Facial expression and emotion recognition with hand-crafted feature extractors were reported in[2]–[5].

On other hand, many work showed how convolution neural network could applied on facial expression recognition. In [13] the author analysed the features learned by the neural network and showed that the neural network can learn patterns from the face images that correspond to Facial Action Units (FAUs). He proposed to ignore the biases of the convolutional layers which gives him 98.3% on CK+ dataset. The winner of FER-2013 challenge[14] used a CNN layers followed by a linear one-vs-all SVM. Instead of minimization, the cross-entropy loss like vanilla CNN, he minimizes a margin-based loss with the standard hinge loss. His method achieved 71.2% on the private test. [15] Applied deeper neural network by constructing four Inception layers after two ordinary convolution layers. [16], [17] proposed a model based on transfer features from pre-trained deep CNN.

[18] Extracts fixed number of SIFT features from facial landmarks. Then, a feature matrix consisting of the extracted SIFT feature vectors is used as input to CNN. The matrix size is MxN where M is the number of SIFT features, and N is the size of each feature. The CNN architecture consists of six layers, two of them are projection layers, one 1D convolution layer, two fully connected layers and a softmax layer.

Another mixture of SIFT and deep convolution proposed in [19]. He used dense SIFT, LBP and CNN which extracted from AlexNet. Each of these features trained by linear SVM and Partial least squares regression. Finally combined the output from all classifiers using fusion network. While the proposed method shares with the [19] the advantages of using SIFT and CNN, the proposed method merge the SIFT and CNN during training and testing.

More recently, ensemble methods such as Bagging or Boosting have been used in facial expressions recognition. Several popular approaches such as [20] used CNN to analyse the video and deep belief net to capture audio information then aggregating top performing models into a single predictor. Moreover, in [21] combine multiple CNN models via learnable weights by minimizing the hinge loss. The winner of EmotiW2015 [8] Trained multiple CNN as committee members and combined their decisions via constructing a hierarchical architecture of the committee with exponentially-weighted decision fusion. He changed network architecture, input normalization, and random weight initialization to obtain varied decisions from deep CNNs.

3. Pre-Processing

We standardize the size of all image to 48x48 pixels. To make the model more robust to noise and slight transformations, data augmentation is employed. Each image is amplified ten times using different linear transformation as shown in Fig 1. These transformations are horizontal flip, rotation with a random angle between (-30, 30), skewing the center and zooming with cropping at four corners and of the image. Finally, all the images are normalized so as to have zero mean and unit variance.

Fig. 1. *Examples of the ten different images transformations*

4. Deep CNNs Architecture

An overview of the CNN network architecture is shown in Fig. 2. The network consists of six convolution layers, three Max-Pooling layers, followed by a two dense fully connected layers. Each time Max-Pooling is added, the number of the next convolution filters doubles. The number of convolution filters are 64, 128, and 256, respectively. The window size of those filters is 3x3. Max pooling layers with a stride of size 2x2 is placed after every two convolutional layers. Max-Pooling is used to summarize the filter area which is considered as a type of non-linear down-sampling. Max-Pooling is helpful in providing a form of translation invariance and reducing the computation for the deeper layers.

To retain the spatial size of the output volumes, zero-padding is added around the borders. The output of the convolution layers is flatted and fed to the dense layer. The dense layer consists of 2048 neurons linked as a fully connected layer.

Each of the Max-Pooling and dense layers is followed by a dropout layer to reduce the risk of network over-fitting by preventing co-adaptation of the feature extractor. Finally, a softmax layer with seven outputs is placed at the last stage of the network. To introduce non-linearity for CNN, we used Leaky Rectifier Linear Unit (Leaky ReLU)[22] as follows:

$$f(x) = \max(x, \frac{x}{20}) \tag{1}$$

Where the value 20 is selected by the FER-2013 validation set. The advantage of using Leaky ReLU over ordinary ReLU is to solve the dying ReLU problem. Instead of being zero when $x < 0$, a leaky ReLU will have a small negative slope. And its derivatives is not zero which make the network learn faster than ReLU.

To obtain a better classification performance, multiple CNN models are used. Three models namely C1, C2 and C3 with different dropout probability have been built. The goal of varying the dropout probability is to increase the diversity among the models. Finally, a categorical cross-entropy method is used as the cost function and is optimized using Adam[23] which is an adaptive gradient-based optimization.

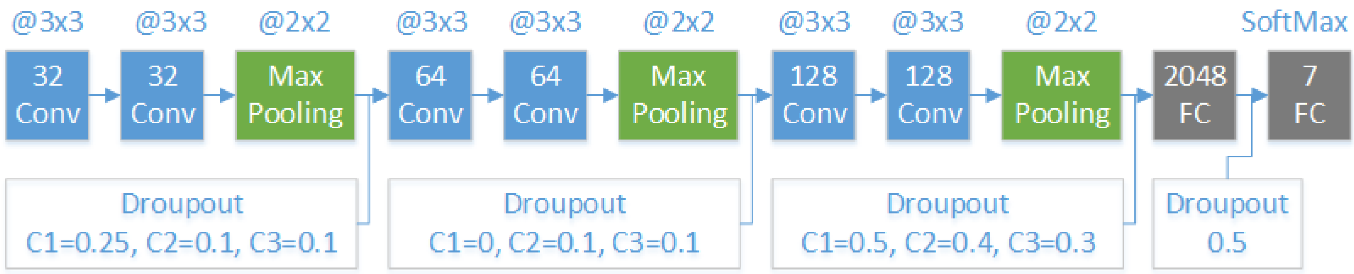

Fig. 2. *CNN model architecture*

5. SIFT AND BAG OF KEY-POINTS

For each image, Scale-invariant feature transform (SIFT)[9] is applied to extract the key-points from the facial image. After locating the key-points, direction and magnitude of the gradient are calculated using key-point neighbouring pixels. To identify the dominant directions, the gradient histogram is established as shown in Fig. 3. Finally, the SIFT descriptor is determined by partitioning the image into 4x4 squares. For each of the 16 squares, we have a vector of length 8. By merging all the vectors, we obtain a vector of size 128 for every key-point.

In order to use the key-point descriptors in classification, a vector of fixed-size is needed. For this purpose, K-means is used to group the descriptors into a set of clusters. Then a bag of key-points is formed via calculating the number of descriptors that are included in each cluster. The resulting feature vector has a size of K.

The K-vector passed through 4096 dense layer followed by a dropout. The weights of the feed forward layer is regularized by l2 norm with value 0.01. Finally merged with the C2 mode as shown in Fig. 3. Three distinctive models S1, S2, S3 has experimented. Each of those models has K size of 256, 512 and 1024 values correspondingly.

6. Aggregating All Models

In order to reduce variance and increase the accuracy we aggregate all the models as shown in Fig. 4. Where CNN-Only refers to the weighted average of C1, C2 and C3. And CNN-SIFT refers to the weighted average of S1, S2 and S3. Finally, we weighted average all the six models into the CNN-SIFT-Avg model.

7. Experimental Results

We tested our models on FER 2013, SFEW 2.0 and Extended Cohn-Kanade. The following sections describe the results of our models on these three datasets.

7.1. Experimental results for FER 2013

The FER-2013 was presented in the ICML 2013 Challenges in Representation Learning[22]. The dataset was retrieved using the Google image search API. Then OpenCV face recognition used to obtain bounding boxes around each face. Finally, the incorrectly labelled images rejected by a human.

The dataset contains 28709 training images, 3589 validation (public) and 3589 test (private) divided into seven types of expression Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral. Due to label noise, the human accuracy of this data is 68%.

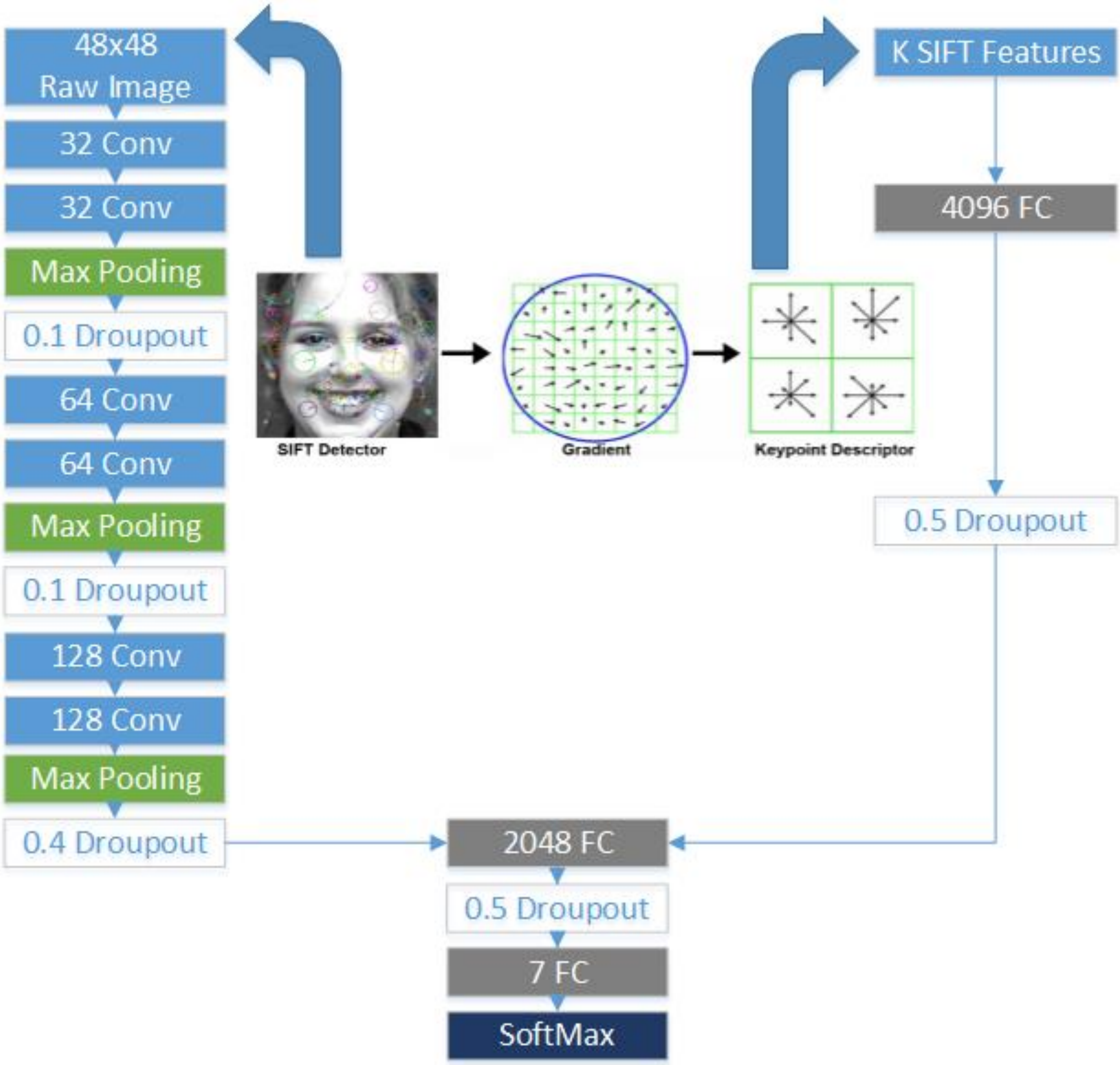

Fig. 3. *CNN-SIFT model architecture*

All the models are trained on 28709 examples. The public set is used as validation to tune the hyperparameters while the private set is used as test set. We initialized the weights as described in[24]. Each network trained for 300 epochs with a batch size of 128.

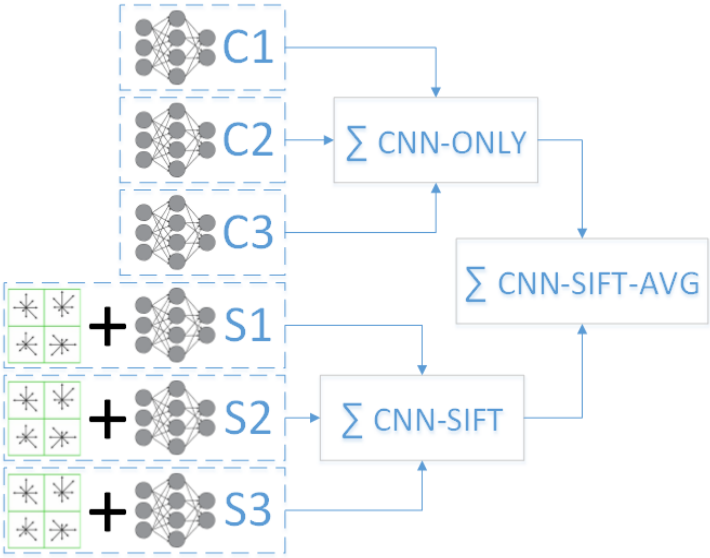

Fig. 4. *Aggregation of different models*

Fig. 5. Shows the accuracy of the models on the test data. As shown the ensemble models have significant improvement over individual models. The SIFT features work better as a voting member than as a single classifier. While CNN-SIFT surpasses CNN-Only, using both methods increases the performance extremely. Table 1 compares our models with other methods. It shows that CNN-SIFT-Avg model outperforming state of the art models [8], [14].

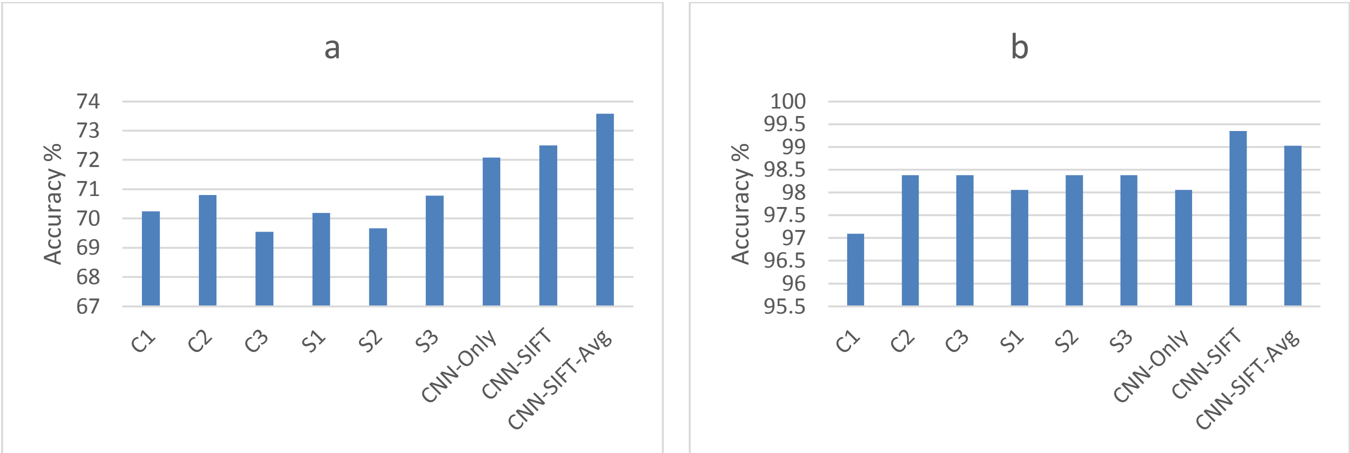

Fig. 5. *Classification accuracies of different models on a) FER-2013 b) CK+*

Table 1. *Classification accuracies of different models on FER*

Method	Accuracy %
CNN-Only	72.08
CNN-SIFT	72.5
CNN-SIFT-Avg	**73.58**
(Kim et al.) [8]	72.72
(Tang, 2013) [14]	71.2

Table 2. *Classification accuracies of different models on CK+*

Method	Accuracy %
CNN-Only	98.06
CNN-SIFT	**99.35**
CNN-SIFT-Avg	99.03
(Khorrami et al., 2015) [13]	98.3
(M. Liu, 2015) [12]	93.70

7.2 Experimental Results for CK+

The CK+ is a lab controlled dataset. Which consist of 327 images from 123 subjects. Each of which is assigned one of seven expression labels: anger, contempt, disgust, fear, happy, sad, and surprise. To make our experiments compatible with other woks [14], [15], [23] and FER-2013 dataset, the contempt example are deleted. So we trained our models on 309 images from the rest of the six expressions.

OpenCV Cascade Classifier has been chosen to detect faces landmarks in images and use these landmarks to crop the faces. The model is pre-trained on FER-2013 training set first. Then fine-tuned the parameter on CK+ dataset. We used all the 309 images for training and testing using 10-fold cross-validation. All the networks trained only for 20 epochs to prevent overfitting due to the limited size of the training set.

Again the CNN-SIFT proves its performance gain against CNN-Only as shown in Fig. 5. However, this time, the accuracy of the CNN-SIFT exceed the CNN-SIFT-Avg. As the number of examples goes small which is the case in CK+, the performance of CNN-SIFT increase compared to CNN-Only and CNN-SIFT-Avg. That is due to that SIFT features do not require big data. The cross validation result of all the models is shown in Table 2. All the fold tested on 31 examples except the last one tested on only 30.

7.3 Experimental Results for SFEW 2.0

Static Facial Expressions in the Wild (SFEW 2.0)[25], [26] is created by selecting static frames from Acted Facial Expressions in the Wild (AFEW). The dataset covered varied head poses, different face resolutions, large age range and close to real-world illumination. The dataset consists of 891 training examples and 431 for validation. Due to the non-public availability of test set, we tested our models on the validation set. We used the pre-aligned faces that provided by SFEW creator which misses some faces The problems in the alignment of faces affected our results, particularly those of SIFT features. As SIFT relies on edges to locate the key-points and some of these images all black or haven't any edges at all. The result is shown in Fig. 6, the S models separately perform badly. On another hand, it performs as well as CNN-ONLY when aggregated as CNN-SIFT.

Fig. 7, 8 and 9 show the confusion matrix of CNN-SIFT-Avg model and CNN_SIFT model on both datasets. In fig. 7, the disgust expression has the heights accuracy. Followed by happy and surprise. The sad has the lowest accuracy among all expressions in FER-2013. Things slightly different for CK+ dataset where most of the expressions score 100% except angry and sad.

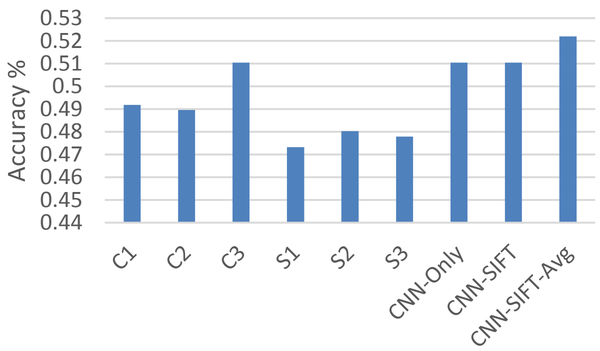

Fig. 6. *Classification accuracies of different models on SFEW*

	Ang.	*Disg.*	*Fear*	*Hap.*	*Sad*	*Sur.*	*Neu.*
Ang.	*67.74*	*3.01*	*12.26*	*2.37*	*8.6*	*1.29*	*4.73*
Disg.	*7.5*	*92.5*	*0*	*0*	*0*	*0*	*0*
Fear	*7.18*	*0.26*	*71.54*	*1.28*	*11.28*	*4.36*	*4.1*
Hap.	*1.57*	*0.11*	*1.46*	*89.47*	*2.58*	*2.13*	*2.69*
Sad	*11.94*	*0*	*14.65*	*2.55*	*58.28*	*2.07*	*10.51*
Sur.	*1.2*	*0.24*	*8.37*	*3.35*	*0.96*	*84.45*	*1.44*
Neu.	*6.75*	*0.13*	*6.89*	*4.5*	*15.5*	*1.06*	*65.17*

Fig. 7. *The confusion matrix of CNN-SIFT-Avg on FER*

	Ang.	Disg.	Fear	Hap.	Sad	Sur.
Ang.	95.65	0	0	0	2.17	2.17
Disg.	0	100	0	0	0	0
Fear	0	0	100	0	0	0
Hap.	0	0	0	100	0	0
Sad	3.57	0	0	0	96.43	0
Sur.	0	0	0	0	0	100

Fig. 8. *The confusion matrix of CNN-SIFT-Avg on CK+*

	Ang.	*Disg.*	*Fear*	*Hap.*	*Sad*	*Sur.*
Ang.	*97.78*	*0*	*0*	*0*	*3.57*	*0*
Disg.	*0*	*100*	*0*	*0*	*0*	*0*
Fear	*0*	*0*	*100*	*0*	*0*	*0*
Hap.	*0*	*0*	*0*	*100*	*0*	*0*
Sad	*2.22*	*0*	*0*	*0*	*96.43*	*0*
Sur.	*0*	*0*	*0*	*0*	*0*	*100*

Fig. 9. *The confusion matrix of CNN-SIFT on CK+*

8. Conclusion

In this paper, a hybrid Convolutional Neural Network and Scale Invariant Feature Transform aggregator approach is proposed to recognize facial expression. We have shown how the SIFT features and convolution neural network could work together. This hybrid approach gets the strength of both methods. While the combination of CNN and SIFT does not provide much improvement, the SIFT features work well as a group member in the aggregator model. The improvement is significantly noticed when the data are small as the case in CK+. Our experiments demonstrate a clear advantage of aggregating SIFT and CNN models by achieving outstanding results on both FER-2013 and CK+ datasets.

9. References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[2] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[3] J. Whitehill and C. W. Omlin, "Haar features for FACS AU recognition," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, p. 5 pp.-pp.101.

[4] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A Set of Selected SIFT Features for 3D Facial Expression Recognition," in *2010 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 4125–4128.

[5] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[6] Z. Wang and Z. Ying, "Facial Expression Recognition Based on Local Phase Quantization and Sparse Representation," in *2012 Eighth International Conference on Natural Computation (ICNC)*, 2012, pp. 222–225.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *ArXiv151203385 Cs*, Dec. 2015.

[8] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jan. 2016.

[9] D. G. Lowe, "Object recognition from local scale-invariant features," in *The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999*, 1999, vol. 2, pp. 1150–1157 vol.2.

[10] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2015, pp. 1–6.

[11] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," *ArXiv14066909 Cs*, Jun. 2014.

[12] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired Deep Networks for Facial Expression Feature Learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015.

[13] P. Khorrami, T. L. Paine, and T. S. Huang, "Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?," *ArXiv151002969 Cs*, Oct. 2015.

[14] Y. Tang, "Deep Learning using Linear Support Vector Machines," *ArXiv13060239 Cs Stat*, Jun. 2013.

[15] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[16] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2015, pp. 443–449.

[17] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *2015 11th International Conference on Natural Computation (ICNC)*, 2015, pp. 702–708.

[18] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Trans. Multimed.*, pp. 1–1, 2016.

[19] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *J. Electron. Imaging*, vol. 25, no. 6, pp. 061407–061407, 2016.

[20] S. E. Kahou *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*, New York, NY, USA, 2013, pp. 543–550.

[21] Z. Yu and C. Zhang, "Image Based Static Facial Expression Recognition with Multiple Deep Network Learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2015, pp. 435–442.

[22] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[23] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *ArXiv14126980 Cs*, Dec. 2014.
[24] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” *ArXiv150201852 Cs*, Feb. 2015.
[25] A. Dhall, S. Member, S. Lucey, and T. Gedeon, *Collecting Large, Richly Annotated Facial-Expression Databases from Movies*. .
[26] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, “Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol,” in *Proceedings of the 16th International Conference on Multimodal Interaction*, New York, NY, USA, 2014, pp. 461–466.