Audio-Visual Emotion Recognition using Deep Transfer Learning and Multiple Temporal Models

Xi Ouyang
Panasonic R&D Center
Singapore, Singapore
ouyang@hust.edu.cn

Shigenori Kawaai
Panasonic R&D Center
Singapore, Singapore
skawaai@gmail.com

Ester Gue Hua Goh
Panasonic R&D Center
Singapore, Singapore
1500080E@student.tp.edu.sg

Shengmei Shen
Panasonic R&D Center
Singapore, Singapore
shengmei.shen@sg.panasonic.com

Wan Ding
Central China Normal University
Wuhan, China
dingwan_cn@mails.ccnu.edu.cn

Huaiping Ming
I^{2}R/A*STAR
Singapore, Singapore
minghp@i2r.a-star.edu.sg

Dong-Yan Huang
I^{2}R/A*STAR
Singapore, Singapore
huang@i2r.a-star.edu.sg

ABSTRACT

This paper presents the techniques used in our contribution to Emotion Recognition in the Wild 2017 video based sub-challenge. The purpose of the sub-challenge is to classify the six basic emotions (angry, sad, happy, surprise, fear and disgust) and neutral. Our proposed solution utilizes three state-of-the-arts techniques to overcome the challenges for the wild emotion recognition. Deep network transfer learning is used for feature extraction. Spatial-temporal model fusion is to make full use of the complementary of different networks. Semi-auto reinforcement learning is for the optimization of fusion strategy based on dynamic outside feedbacks given by challenge organizers. The overall accuracy of the proposed approach on the challenge test dataset is 57.2%, which is better than the challenge baseline of 40.47% .

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**;

KEYWORDS

Emotion Recognition; Convolutional Neural Network; Long Short Term Memory network; 3D convolutional Network; Reinforcement Learning; Transfer Learning; Model Fusion

ACM Reference Format:
Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. 2017. Audio-Visual Emotion Recognition using Deep Transfer Learning and Multiple Temporal Models . In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3136755.3143012

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143012

1 INTRODUCTION

One of the main challenges for audio-visual emotion recognition in natural environment is the complexity of facial and vocal expressions along spatial-temporal dimensions. This paper tackles the problem by using two state-of-the-arts techniques. The first is transfer learning and the second is combination of temporal feature models. The proposed approach is composed of three modules. They are facial emotion recognition, audio emotion recognition and system fusion. For facial emotion recognition features transferred from deep face recognition networks are used to capture the spatial information. Three types of models are then used for temporal information extraction, based on the assumption that they are complementary. The first type is image-set classification model [8]. Image-set model treats images(video frames) order-less points in image feature space and studies the geometric structures of the points in set. The second type is LSTM-RNN and the third type is C3D [4]. The combination of Deep CNNs and LSTM-RNNs have been widely used [15]. However traditional types of CNN-RNN combinations are based on "stack", which presumes Markov properties between CNN and RNN nodes and may less flexible for complex visual tasks such as facial emotion recognition. In this paper, we utilize two more flexible network architectures to capture temporal features from facial videos for facial emotion recognition. The first network is ResNet-LSTM and the second is C3D Network. ResNet-LSTM allows nodes in lower CNN layers to directly contact with RNN and C3D network models the relations of spatial-temporal nodes as undirected graph. For audio emotion recognition the model is also transferred from pretrained speaker recognition LSTM. For system fusion we model the result submission process as a simulation of emotional human-computer interaction with emotions. The system predictions of testing data are the actions of computer and the returned results are the feedback/rewards (with emotions). We utilize

the reinforce learning strategy to study the system weights. Experimental results show that the combination of geometric structure, LSTM and C3D outperforms each single model for facial emotion recognition. The reinforcement learning fusion also improves the system performance. Experimental results show that the overall accuracy of the proposed approach on the challenge test dataset is 57.29%, which is better than the challenge baseline of 40.47% .

2 RELATED WORKS

2.1 Transfer Learning

Transfer learning has been proved effective for deep audio-visual emotion recognition based on small scale training data. The basic idea of transfer learning is the tasks share feature space and there are two open problems, which are the tasks selection and deep network architectures for transfer [2, 24]. Azizpour et.al. [2] and Yosinski et.al. [24] empirically studied the transferability of the convolutional neural network layer for different image classification tasks. For emotion recognition pretrained general image classification models and face recognition models are the most popular for transfer learning. Ng et.al. [16] and Ding et. al. [8] finetuned pretrained AlexNet for facial emotion recognition. Ding et al.[7] finetuned the network based on VGG-Face. Instead of weights transfer they transferred the distributions of outputs of the late VGG-Face layers. In other words they require the target network to act like source network without the constraints of same architecture. In this paper the transfer learning is weight transfer from pretrained face recognition models due to its effectiveness.

2.2 Deep Networks for Facial Emotion Recognition

Audio visual emotion recognition has attracted a lot of attentions in recent years. Various emotion recognition challenges such as the Facial Expression Recognition and Analysis Challenge (FERA) [21] and the Emotion Recognition in the Wild Challenge (EmotiW) [6] have become standard benchmarks. Based on benchmarks many deep learning based approaches have published and they achieved the state-of-the-arts performances. One trend of the deep learning based facial emotion recognition is the network architectures keep becoming deeper and more complex. as presented in Table 1. Gudi et.al. [11] trained a 3-layer convolutional neural network from scratch for static facials action recognition. Ng. et. al. [16] finetuned 5-layer AlexNet for emotion classification. Fan. et al. [10] 's approach is the combination of 16-layer CNN-RNN and 8-layer C3D network. Yao. et al. [23] 's approach is also based on 16-layer residual network instead of plain ones. Compared with plain networks residual networks, i.e., deep networks with identity mapping shortcut connections, converges much more effective during training when the network is very deep(e.g., 50 layers) [12]. Bargal. et.al. [3] combined three networks for frame feature extraction. They then model the videos based on the first and second orders of the frame feature sets.

3 VISUAL EMOTION RECOGNITION

This section introduces the end-to-end deep network module for facial emotion recognition. It is composed of two modules. The face detection module is a multi-task cascaded convolutional network [26] and the facial emotion recognition module is the fusion of three deep networks, i.e., The stacks of CNN and RNN (VGG-LSTM), ResNet-LSTM and C3D Network.

Table 1: Published facial expression CNNs in recent years .

Publish years	Depth	Type	Dataset
Gudi et.al. [11], 2015	3 conv-layers	Plain	FERA
Ng. et. al. [16], 2015	5 conv-layers	Plain	EmotiW
Fan. et al. [10], 2016	17 conv-layers	Plain	EmotiW
Yao. et al. [23], 2016	16 conv-layers	Residual	EmotiW
Bargal. et.al. [3], 2016	91 conv-layers	Residual	EmotiW

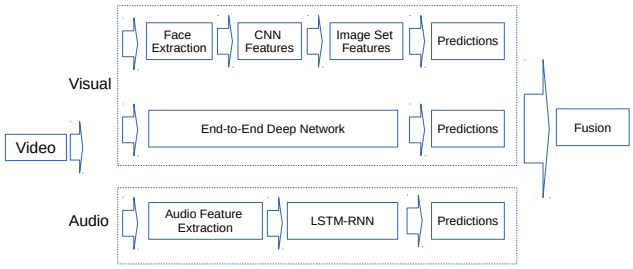

Figure 1: The overview of proposed approach.

3.1 Pre-processing

A Multi-Task Cascaded Convolutional Network (MTCNN) is used for face extraction and alignment for video frames. It treats the problem as cascade of sub-tasks and trains CNNs for each task. In more specific three tasks are defined and they are coarse background cutting, face candidate detection and fine bounding box localization & facial landmarks detection [26]. By presuming each video contains only one person's face, the candidate regions of video frames with highest confidence are taken as detected faces for the next step of feature extraction.

3.2 DCNN Feature Extraction

3.2.1 VGG-LSTM. The VGG-LSTM architecture is shown in Fig. 2. For one face video each frame goes through VGG-16 network [20] then the extracted CNN feature sequence passes to LSTM layers to predict the video emotion. The VGG-16 network is transferred from VGG-Face to compensate for the small scale of facial emotion samples for training.

3.2.2 Resnet-LSTM. The overall architecture of Resnet-LSTM [22] is shown in Fig. 3 and Fig. 4. In Resnet-LSTM the image features from multiple CNN layers are directly passed to LSTM network for recognition. In our approach the LSTM is composed of several smaller LSTMs as sub-modules, as presented in Fig. 4.

3.2.3 C3D. C3D network [4] uses 3-D convolutional kernals instead of traditional 2-D kernels to capture the spatial-temporal features for videos. The architecture of C3D is shown in Fig. 5.

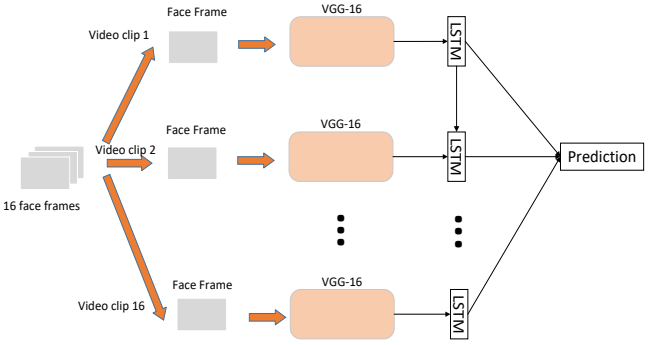

Figure 2: The overall structure of VGG-LSTM model.

4 AUDIO EMOTION RECOGNITION

Deep neural network (DNN) and transfer learning are applied for the emotion recognition of audio signals. The amount of data provided by Emotiw2017 is quite limited for deep learning method. Under the condition of scarce training data, one solution is to borrow information from related domains where large amount of labelled data is available. In this work, we borrow information for audio emotion recognition task from gender classification task with a large scale (1000 hours) corpus of read English speech named LibriSpeech [17].

4.1 Pre-processing

First of all, the audio signals extracted from the videos are converted to 16k Hz sampled, 16 bit quantized mono signals. The speech signals of LibriSpeech are also converted to the same data format. Then the sample value of all the signals are normalized to the same range as the volume varies for different audio signals. Audio features use the extended version of GeMAPS (eGeMAPS) provided by AVEC 2015 [18]. The audio features are normalize to zero mean and unit variance as the DNN input.

4.2 Transfer Learning

The basic idea of transfer learning is to share common knowledge learned from one machine learning task (source) with another related machine learning task (target). This technique of borrowing information from related domain has been proved to be effective for many deep learning tasks with limited data [9, 13, 19]. In our case, the source task is gender classification and the target task is emotion classification. The features including energy, fundamental frequency and spectral are essential for both gender classification and emotion classification. The proposed framework of transfer learning for audio emotion recognistion is shown in Fig. 6.

The DNN of source task is trained with LibriSpeech which contains 1201 female speakers and 1283 male speakers. About 85% of the data is applied as training set and another 15% is applied as the validation set. As shown in the left part of Fig. 6, the neural network with two feed forward hidden layers and each layer 256 nodes is applied for the gender classification task.

5 REINFORCEMENT LEARNING FOR AUDIO AND VISUAL FUSION

Reinforcement learning is important for emotion recognition for natural human-computer interaction. However, one challenge is the lack of public available data. We consider the results submission process as simulation of emotional HCI for preliminary research. During reinforcement learning the computer actions are prediction results submission and the human feedbacks are the returned distributions of results. The goal is to minimize the diversity between distributions by tuning weights of sub-systems for fusion. Fig. 7 presents an overview of the reinforcement learning framework.

The final fusion is based on three sub-systems. They are imageset modeling [8] and visual and audio sub-systems introduced in Section 3 and 4. For score level fusion three weighted vector $w_i = [\lambda_1, \lambda_2, ..., \lambda_c], i = 1, 2, 3$ are introduced, where c denotes the number of emotion classes for recognition. The final score S is calculated as

$$S = \sum_{i=1}^{3} w_i S_i \tag{1}$$

where S_1, S_2, S_3 denotes the audio and visual based scores respectively. For reinforcement learning the score function L_t at Step t is defined as

$$L_t = W(A_{t-1})D - C \tag{2}$$

where A_{t-1} denotes the feedback at Step $t-1$, $D = [d_1, ...d_c]$ denotes the differences of predicted/true distributions on testing data and W denotes the weights of D. C is static, e.g., the overall accuracy on validation data.

6 EXPERIMENTAL RESULTS

6.1 Emotiw 2017 Database

Video based emotion recognition challenge is one of the subchallenges of Emotion Recognition in the Wild Challenge (EmotiW2017) [1]. It contains audio-video short clips labeled using a semi-automatic approach defined in [5] and the task is to assign a single emotion label to the video clip from seven universal emotion states (Anger, Disgust, Fear, Happiness, Neutral, Sad and Surprise). This challenge is a continuation from EmotiW2013-16. Compared to last year a major change is extra 60 video clips extracted from TV series (The Big Bang Theories) was introduced to the test set to determine the generalizability of methods trained on more normal emotion expressions. In this sub-challenge, the dataset contains 1799 video clips: 773 for training, 373 for validation and 653 for testing. The training and validation data of last and this year are same.

6.2 Visual Network

6.2.1 Network Architectures. For experiments we resampled videos to a fixed length of sixteen frames to balance the tradeoff between computing resource and model accuracy. For VGG-LSTM the size of input face images was 224×224, the CNN features are the outputs of the fc6 layer and the LSTM is one hidden-layer 128 embedding nodes. A 0.7 ratio dropout layer was also added on the top of hidden LSTM layer. For Resnet-LSTM the input size is 112×96, the CNN features are the outputs of pool1b, pool3 and fc5 layers.

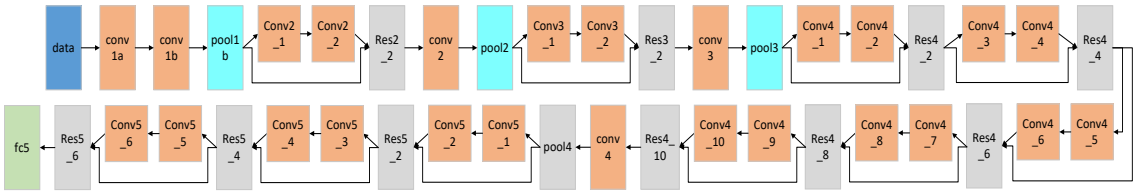

Figure 3: The structure of face residual network.

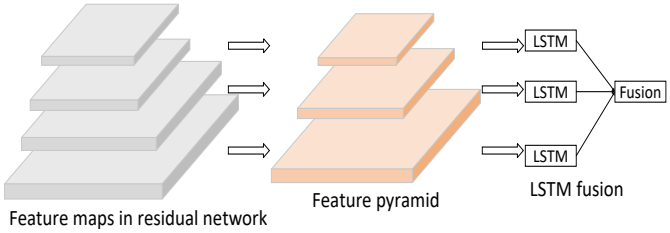

Figure 4: Feature pyramid.

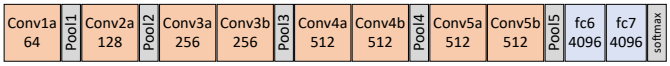

Figure 5: The structure of C3D network.

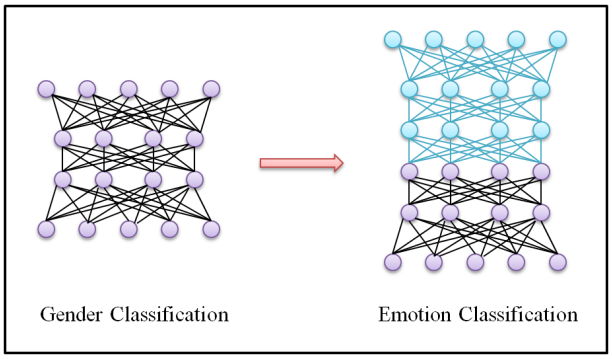

Figure 6: The proposed transfer learning framework.

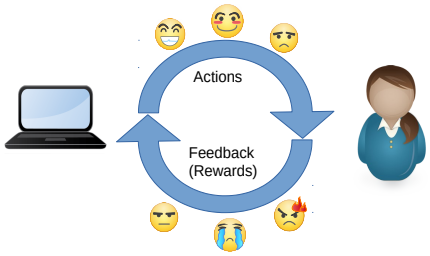

Figure 7: Reinforcement learning for emotion recognition.

Three one hidden layer LSTMs with 128 nodes took the outputs of CNN layers respectively and the final output is the element wise sum of three LSTMs. For C3D network, we added dropout layers after fc6 and fc7 layers with 0.5 dropout ratio to avoid overfitting.

6.2.2 Training. Pre-trained models were used for training three networks. We used VGG-face as pre-trained model for VGG-LSTM, caffe-face [22] for Resnet-LSTM and C3D model pre-trained on Sport-1M [14]. The learning rate starts with 0.001, decreases to its 1/10 every 1000 iterations and training stops at 5000 iterations. Two data augmentation techniques were also used for training. They are horizontal flipping and random cropping.

6.3 Speech Network

We train the speech network with the Microsoft Cognitive Toolkit (CNTK) [25]. The feed forward network with tanh function as the activation function is applied for the hidden layers of both source and target task:

$$f(\mathbf{X}) = \tanh(\mathbf{X}^T w + b), \tag{3}$$

where $\mathbf{X}$ is the input, w is the weight and b is the bias. The $softmax$ function is applied for the output layer to derive a probability representation for each class, and the predicted probability for the j^{th} class is

$$\mathbf{P}(y = j|\mathbf{X}) = softmax(\mathbf{X}^T w_j + b_j). \tag{4}$$

The cross entropy with $softmax$ layer is used as the criterion node and the classification error is used as evaluation node in the training process.

The minibatch size is 64 and the momentum per minibatch is 0.9 for both source task and target task. The learning rates per minibatch are 0.004 and 0.002 for source task and target task respectively. The learning is reduced by decrease factor of 0.618 if the improvement is less than 0, and it is increased by factor of 1.382 if the improvement is larger than 0.5. The maximum number of epochs is 40 for both source and target task.

The confusion matrix of seven classes on the validation set is shown in Table 2. The classification accuracy for Angry and Neutral is relatively good, but the accuracy for Disgust and Surprise is poor. The average accuracy is 35.2%.

6.4 Fusion

For each step the weights W defined in Eq. 2 was manually set based on the feedback at previous step. Then the optimum weights of sub-systems were searched by brutal-force. Since there are $3x7 = 21$ independent weights to determine, we constrained the searching space for each weight in [0,0.5,1]. The final weights were determined based on the observation that some emotions (such as disgust) are difficult to recognize if only based on face and audio information. So

Table 2: Confusion matrix of the audio emotion recognition system on the Validation set.

	Ang	Dis	Fea	Hap	Neu	Sad	Sur
Ang	**40**	0	4	5	6	8	2
Dis	14	**1**	2	5	11	7	0
Fea	5	0	**18**	6	6	10	1
Hap	12	1	4	**20**	14	12	0
Neu	5	0	1	18	**34**	5	0
Sad	4	2	7	11	17	**20**	0
Sur	4	0	5	17	13	5	**2**

we assigned low weights to disgust scores. Based on reinforcement learning strategy the best weight we found is

Table 3: The best weights of sub-systems for fusion.

	An	Di	Fe	Ha	Ne	Sa	Su
ImageSet	0	0	0	0.5	0.5	0	0
Visual LSTM	0.5	0	0	1	0.5	1	0.5
Audio LSTM	0.5	0	1	0	0	0	0

6.5 Evaluation Results

6.5.1 Visual Emotion Recognition. Table 4 are the results of visual modules on validation set. The results shows that traditional CNN-LSTM architectrue achieves better performances than more free architectures such as Pyramid CNN-LSTM and C3D. One possible reason to this observation is the lack of data makes training of more free (more complex) networks more difficult. However different network architecutres still present high complementary, as the accurcy of combination achieves 53% outperforms the best single model by 6%. Besides the RNN based temporal features and the Geometry Structure features also present complementary. as the overall accuracy increase by 1% after combination.

Table 4: The validation results of visual deep models.

Deep models	Accuracy
VGG-LSTM	47.4%
Resnet-LSTM	46.7%
C3D	35.2%
End to End Module	53.1%
CNN+ImageSet Model	50.1%
Whole Visual Module	54.2%

Compared with last year's results shown in Table 6, our system (Table 5) gains improvements on Happy (74->89) and Sad (42->55) but decreases on Angry (77->67), Fear (42->32) and Surprise (14->0). The reason may be the proposed LSTM based system is better at capturing the long lasting temporal context features but not good at the peak features. The same system was tested as well using the Chinese Movie dataset of multimodal emotion recognition challenge

Table 5: Best results on testing data. Overall accuracy is 57.2%

	An	Di	Fe	Ha	Ne	Sa	Su
An	67.35	0.00	3.06	11.22	10.20	6.12	2.04
Di	15.0	0.00	0.00	17.50	25.00	40.0	2.50
Fe	27.14	0.00	32.86	4.29	7.14	21.43	7.14
Ha	4.17	0.00	0.00	89.58	1.39	4.86	0.00
Ne	9.84	0.00	1.55	14.00	58.03	16.06	0.51
Sa	15.00	0.00	3.75	17.50	8.75	55.00	0.00
Su	25.00	0.00	10.71	14.29	28.57	21.43	0.00

Table 6: Our EmotiW2016 Best results. Overall accuracy is 53.9%

	An	Di	Fe	Ha	Ne	Sa	Su
An	77.11	01.20	04.82	03.61	04.82	04.82	03.61
Di	22.22	02.78	05.56	25.00	13.89	27.78	02.78
Fe	24.24	01.52	42.42	03.03	13.64	03.03	12.12
Ha	13.33	00.00	02.22	74.07	04.44	05.93	00.00
Ne	14.94	01.72	00.57	08.05	53.45	18.39	02.87
Sa	22.54	00.00	02.82	19.72	08.45	42.25	04.23
Su	25.00	03.57	17.86	07.14	17.86	14.29	14.29

(MEC) 2017. It achieves top 2 in Audiovisual emotion recognition sub-challenge of MEC 2017. Our emotion recognition system is language independent technology. Of course, the performance can be better if we conduct fine-tuning for different languages (e.g., English, Chinese).

7 CONCLUSION

We proposed an audio-visual based approach that applies transfer learning and combination of models for emotion recognition in the wild and achieved a final accuracy of 57.2% on the test set, which is better than the baseline of 40.47%. The facial emotion recognition subsystem demonstrates that transfer learning is an effective way to find good image features; the combination of geometric and RNN based video features is more effective than single model for temporal information extraction; Reinforcement learning is promising for emotion recognition in HCI scenarios while one of the challenges is how to define the reward function.

Our future work will be conducted in two directions. The first is to find more effective structures of deep models for spatial-temporal feature extraction. The second is to conduct a deeper study on HCI based reinforcement learning emotion recognition in real scenarios.

REFERENCES

[1] Dhall Abhinav, Goecke Roland, Ghosh Shreya, Joshi Jyoti, Hoey Jesse, and Gedeon Tom. 2017. From Individual to Group-level Emotion Recognition: EmotiW 5.0. In *ACM ICMI 2017*.
[2] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. 2015. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 36–45.
[3] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. 2016. Emotion recognition in the wild from videos using images. In *ACM International Conference on Multimodal Interaction*. 433–436.
[4] R. Fergus L. Torresani D. Tran, L. Bourdev and M. Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. *ICCV* (2015).

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19, 3 (2012), 34–41.

[6] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 423–426.

[7] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. 2017. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 118–126.

[8] Wan Ding, Mingyu Xu, Dongyan Huang, Weisi Lin, Minghui Dong, Xinguo Yu, and Haizhou Li. 2016. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. (2016), 506–513.

[9] Zhengming Ding, Nasser M Nasrabadi, and Yun Fu. 2016. Task-driven deep transfer learning for image classification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2414–2418.

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

[11] Amogh Gudi, H. Emrah Tasli, Tim M. Den Uyl, and Andreas Maroulis. 2015. Deep learning based FACS Action Unit occurrence and intensity estimation. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. 1–5.

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

[13] Ying Huang, Mingqing Hu, Xianguo Yu, Tao Wang, and Chen Yang. 2016. Transfer Learning of Deep Neural Network for Speech Emotion Recognition. In *Chinese Conference on Pattern Recognition*. Springer, 721–729.

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*.

[15] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie K Dagli, and Thomas S Huang. 2016. How Deep Neural Networks Can Improve Emotion Recognition on Video Data. (2016), 619–623.

[16] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 443–449.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. (2015), 5206–5210.

[18] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. (2015), 3–8.

[19] Yoshihide Sawada and Kazuki Kozuka. 2015. Transfer learning method using multi-prediction deep Boltzmann machines for a small scale dataset. (2015), 110–113.

[20] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR* (2015).

[21] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. 2015. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 6. IEEE, 1–8.

[22] Z. Li Y. Wen, K. Zhang and Y. Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. *ECCV* (2016), 499–515.

[23] Anbang Yao, Dongqi Cai, Ping Hu, Shandong Wang, Liang Sha, and Yurong Chen. 2016. HoloNet: towards robust emotion recognition in the wild. In *The ACM International Conference on Multimodal Interfaces*. 472–478.

[24] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

[25] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al. 2014. An introduction to computational networks and the computational network toolkit. *Microsoft Technical Report MSR-TR-2014–112* (2014).

[26] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.