Proceedings of 8th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2020 Month: June Volume: 8

Emotion Recognition from Video using Feature level Fusion

Binod Adhikari [a], Basanta Joshi [b]

[a,b] *Department of Electronics and Computer Engineering, IOE, TU, Kathmandu Nepal*
Corresponding Email: [a] 074msice006.binod@pcampus.edu.np, [b] basanta@ioe.edu.np

Abstract
Automatic emotion recognition is a challenging task as human uses more than one varieties of modalities to express their emotion. The applications of emotional intelligence can be found in many domains including multimedia retrieval and human-computer interaction. With arrival of deep neural network, there is great success in determining emotional states. Inspired by state-of-art on emotion recognition and its applications, we propose an emotion recognition system using audio and visual information. One of the challenging task in recognition of emotion is to extracting a robust feature, and deep neural network is one of game changer when there is an uncertainty about feature required by task in-hand, especially for computer-vision and speech-processing. For this purpose, utilization of convolutional neural network (CNN) at the front-end plays a key role because of its speciality on extracting robust task-based feature by parameter sharing and sparsity. we have utilized CNN at front layers for both auditory and visual modality. And for audio-modality LSTM is used to capture features related to change in time, as it is time-varying signal is applied. The system is trained completely in an end-to-end manner, with raw data for both audio and visual base models. End-to-end training and use of feature level fusion takes advantage of learning correlated features between different domain during combined training.The proposed model classify seven universal emotions, angry, sad, happy, fear, surprise, disgust and neutral. Traditional approaches based on auditory and visual handcrafted features for eNTERFACE05 is compared with this proposed learnt feature based emotion recognition system.

Keywords
Emotion, Deep Learning, CNN, LSTM

1. Introduction

Emotion Recognition System (ERS) has been a fundamental research problem in human-machine interaction for providing emotion intelligence to a machine. It allows machine to perceive an emotion expressed by human, and improves human-machine interaction by making it more natural. In natural process of expressing or perceiving emotion human uses different forms of information, and these can be describe as linguistics and para-linguistic. The use of linguistic, textual context, to express emotion varies with culture, geographical region. On the other hand, para-linguistic means use of facial-expression, tone and pitch of voice, body language and physiological signals.

The emotion recognition is challenging task as human emotions lack of temporal boundaries and ways of expressing emotion varies in human to human [1].Many studies have shown the favourable properties of deep neural network over the conventional machine learning as, it learns to represent intermediate features jointly with task and eliminate hand-engineered feature extraction part that makes learning process less depended on domain knowledge. Also, automatically learnt feature from raw input signal better suits the task in hand and leads to improvements. Nevertheless, majority of works on the field of emotion recognition is centred toward the use of hand-crafted feature specially for speech emotion recognition. When using raw waveform as input representation, phase invariant is critical as identical sounds may appear at distinct phase shifts. In 2015, Hoshen has proposed the CNN-DNN model for acoustic modeling using raw waveform [2]. They have tried to extract the frequency component by using segmented audio wave in smaller chunks of 275ms and processed it through deep CNN-DNN model to learn filter-banks. As audio waveform changes over time, the temporal feature should be

considered as an important feature, DNN with large hidden layer can addressed these issues.

Recently, Avots has proposed the multi-modal emotion recognition using audio and face by decision level fusion [3]. They have used the conventional machine learning method for speech emotion recognition by using Support Vector Machine (SVM) with different features such as Energy entropy tonal, power ratio, short-time energy spread, zero-crossing rate slope, spectral roll-off skewness and spectral centroid. For visual model they used transfer learning, AlexNet pre-trained model. And finally they have used a simple fusion technique of decision level to select the final emotion class based on maximum probability class from audio and visual mode.

In this study, the end-to-end deep learning approach for multi-modal emotion recognition from audio-visual information is presented to classify the discrete emotion labels. Instead of classical feature based approach, feature from raw audio signal is extracted using CNN-LSTM model and for visual information plain CNN model, inspired from VGG, is used.

2. Literature Review

Multi-modal emotion recognition is the concept of combining more than one modality, and this approach is highly depending on performance of its base models. In case of multi-modal emotion recognition using audio and visual information, audio and facial emotion recognition systems are uses as base models.

2.1 Facial emotion recognition (visual)

Facial expression is one of the most powerful, natural and universal signals for human being to convey their emotional states and intentions. In the field of computer vision and machine learning various facial expressions recognition (FER) systems have been explored to encode expressed information from facial representation. Traditional approaches usually consist of three steps: First a face image is detected from an input image, and the facial components (e.g., eyes, nose and mouth) or landmarks are identified from face region. Next, various spatial and temporal features are extracted from these facial components, generally known as hand-crafted feature. Finally, based on the extracted features, a classifier such as support vector machine (SVM), random forest is trained to classify emotion states [4]. The problems with traditional method is time consuming as it requires to map all facial land-marks and also need to have domain knowledge to select specific features (facial action units) which varies for person to person. In contrast to traditional approaches using handcrafted features, deep learning has emerged as a general approach to machine learning, yielding state-of-the-art results in many computer vision studies with the availability of big data. Deep-learning-based FER approaches highly reduce the dependence on face-physics-based models and other pre-processing techniques by enabling "end-to-end" learning to occur in the pipeline directly from the input images [5]. Among the several deep-learning models available, the convolutional neural network (CNN), a particular type of deep learning, is the most popular network model [6]. In CNN-based approaches, the input image is convolved through a filter collection in the convolution layers to produce a feature map. Each feature map is then combined to fully connected networks, and the face expression is recognized as belonging to a particular class-based on the output of the softmax algorithm.

2.2 Audio Emotion Recognition

Several prosodic and acoustic features have been used in the literature to teach machines how to detect emotions. Since emotional characteristics are more prominent in prosodic features, these features are widely used in the literature. For decades, spectrogram, generated by converting audio signal to frames by sliding window and applying Short-Time Fourier Transform (STFT) on such windowed frame. Another most popular feature mel frequency cepstral coefficients (MFCCs) have been used as the dominant acoustic feature representation for audio analysis tasks [7]. These are magnitude spectra projected to a reduced frequency bands, converted to logarithmic magnitudes and compressed with a discrete cosine transform (DCT). Both these features are time-frequency representation of audio and considering these features as 2D image for processing through CNN is very common approach for audio related community. However, these transformation is not homogeneous as horizontal and vertical axes in an image. Images are instantaneous snapshots of a target and often analyzed as a whole or in patches with little order constraints; however audio signals have to be studied sequentially in chronological order. These properties gave rise to audio-specific solutions.

Recently, many researcher tried to avoid relying only

on a designed filter bank based methods and data-driven statistical model learning. Use of raw waveform representation of the audio signals as inputs and learn data driven filters jointly with the rest of the network for the target tasks is gaining popularity. In Yedid Hoshen, et.al., have tried to use the lower layers of the model that follows the basic process of designed filter,log-mel spectrum, to discard the conventional approach of using designed filter [2]. The audio signal, represented as a sequence of either frames of raw audio or human engineered feature vectors can be analyzed by various deep learning models. Similar to other domains like image processing, for audio, multiple feed-forward, convolutional, and recurrent neural layers are usually stacked to increase the modeling capability. When using raw waveform as input representation, one of the difficulty is that the identical sounds may appear at distinct phase shifts. So, using a representation that is invariant to small phase shifts is critical. To achieve phase in-variance, researchers have usually used convolutional layers which pool in time and DNN with large number of hidden layers, which are able to capture the same filter shape at a variety of phases. This process requires very larger number of parameter to train, which leads model to over-fit under the condition of small data-set.

2.3 Multi-modal Emotion Recognition

The audio CNN takes the Mel-spectrogram segments of an audio signal and video CNN takes the face [8] . In the second phase, a DNN was trained that comprised of a number of fully-connected layers. The fusion of features extracted from two CNN was input to fully connected classifier. They show, among many fusion techniques, weighted mean wins other. To gain better performance, fusion methods that merge different modalities are essential. Fusion methods can be classified into feature, decision, and model level fusion. The most researchers chose late fusion. Vielzeuf discussed five fusion methods: majority vote, mean, mod-drop, score tree, and weighted [9]. The researcher concluded that the simple fusion technique (decision level) is better amongst other as base model already learns salient and discriminating features, and using complex fusion technique like feature level fusion will increase the complexity of model. This can be solved by minimizing the feature vector before meta-classifier to reduce parameters.

3. Proposed Method

In traditional machine learning one of the first step is to extract the feature sets from data. Different approach based on CNN as feature learner which is different from manual feature extraction from audio is used in this paper. CNN can be used by modeling it to respond as finite impulse response, which can represent audio in time-frequency decomposition and reduce noise influence. A key component of this model is use of raw data for both audio and visual models, by using 1D and 2D convolution respectively.

3.1 Audio Model

In contrast to traditional approach, where feature are first extracted and passed to machine learning algorithms, this audio model extracts feature itself jointly with classification task.

Input: first audio signals are segmented into 2 second each, and normalization is applied to have zero-mean and unit variance to account for outliers and different level of variation on pitch. Sampling is done at 16khz, which gives 32000 vector as input to audio model.

Convolution: Convolution over time (temporal convolution) is used to extract time-frequency representation of audio at high sampling rate. To avoid frequency variance smaller kernel size that represents 5ms is used, 32 filters are learn at this layer. Batch-Normalization (BN) layer normalizes the activation of the convolution layer at each batch, and improves the performance and stability of the deep networks. The transformation applied by BN maintains the mean activation close to 0 and standard deviation close to 1. The pooling layer can make the features robust against noise and distortion. Max-pooling, most commonly used non-linear functions, divides the inputs into a set of non-overlapping regions and outputs the maximum value of each sub-region and also down-sample the input signal.

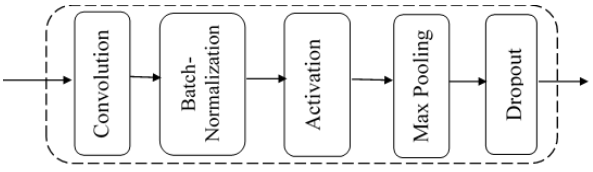

Figure 1: CNN block for audio model

In Figure 1, the CNN block is created to learn low level features from input audio. To learn features

hierarchically, 3 blocks are stacked to each-other. Combination of them learns the high level feature of audio signal by taking larger receptive field. As audio signal is time-varying signal, its current states depends over past states also. So, to capture these features, long-term dependencies from learnt features of CNN, LSTM layers is added on top. As LSTM is designed for capturing contextual features from sequential data, it perfectly fits for task such as in audio. The LSTM can remove or add information to the block state using four components: an input gate, an output gate, a forget gate and a cell with a self-recurrent connection. Cell, contains the states from previous sequence , which is used to predict future states from all previous states. For the classification task, output of LSTM is given to fully connected layer with softmax activation.

3.2 Visual model

In figure 2, the CNN block for visual model is created by stacking multiple convolution layers. As consecutive convolution layers allows subsequent convolution layer to learn higher-level features from features extracted in the previous layer. First steps in the traditional face recognition method is feature extraction utilizing hand-crafted representations such as Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), facial Landmarks (geometrical features). For this task, deep learning algorithms approach is used which takes pixel intensity and learns features itself during task. By using this method, model will learn appearance based features such as muscles shapes, textures and other. Inspired from VGG16, a plain CNN-DNN model is used for visual model in this study. Visual model is prepared as similar architecture of VGG16, however, only 3 blocks of CNN is used. Each block has stacked convolution layers. This stacking manner allows CNN to learn hierarchical decomposition of input from its previous layer. The input to the visual model is, pixel intensity of face-image which is extracted from frame of video file.

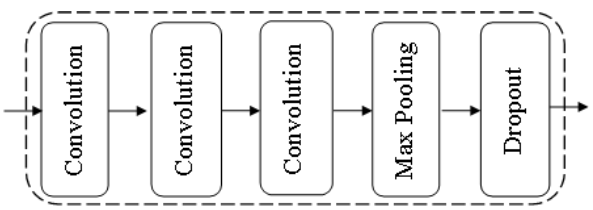

Figure 2: CNN block for Visual model

3.3 Multi-modal Training

The block-diagram for audio-visual based multi-modal emotion recognition system using feature level fusion is shown in figure 3. As in figure 3, learnt features from audio and visual model is fused and connected to meta-classifier of dense layer to create multi-modal emotion recognition system. The process fusion is simply linear, concatenation or element wise addition. To train final multi-modal model, at first, classifier network of both audio and video network were discarded. And then, Feature learning layers (output of LSTM for audio, and output of average pooling layer for visual) are concatenated. On top of concatenated layer, classifier network (dense layers) is added, which is followed by another dense-layer with soft-max activation. Before training of multi-modal, weights of two uni-modal (audio and visual) are initialize by previously trained uni-modal, and weights of classifier (afterward concatenated layer) are initialized randomly. To minimize the complexity and fine-tuning the parameters of classifier network, some initial layers of uni-modal are set non-trainable. Finally, the whole network is trained end-to-end. The soft-max activation layer gives predicted probabilities for all considered emotion classes (7 classes in this paper). Predication of emotion class for a given input (audio and visual), class with maximum predicted probability is taken as final output of the model.

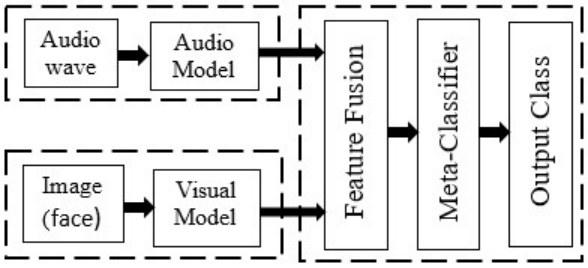

Figure 3: Block diagram of multi-modal using feature fusion

4. DATASET

All dataset used in this paper are publicly avilable for academic purpose. The used dataset are eNTERFACE'05 [10], FER2013 [11] and emoDB [12].

4.1 eNTERFACEO5

This dataset contains the 1166 video file. Each video clip is labeled with one emotion class from set of six

different emotions (happiness, sadness, anger, surprise, disgust, fear) and neutral. Different 44 participants, including from 14 nationalities were used to capture video clips.The video was processed using a 720x576 AVI format. The frame rate of 25fps was used. The audio was taken at sampling rate of 48000 Hz.

4.2 FER2013

FER2013 is publicly available dataset.This data-set consists of 35887 labeled images of face. This dataset contain static gray-scaled images, each having size of 48*48. The dataset is divided as; 28709 for training, 3589 for evaluation and the rest for testing. This dataset contains 7 different emotion; happiness, sadness, anger, surprise, disgust, fear and neutral.

4.3 emoDB

This dataset is prepared for speech emotion recognition. It contains 500 audio files taken at sampling rate of 48 Khz. Data was created by using 10 different actors. This dataset contains 7 different emotion classes; happy, angry, anxious, fearful, bored and disgusted as well as neutral. This dataset is prepared for speech emotion recognition.

5. Experimental Results and Discussion

At first, audio and visual network were trained individually, learning rates were choose 0.001. To prevent over-fitting, dropout (rate 0.25) regularization is applied. The values for learning rate initially set at 0.01, as choosing higher learning rate reduce training duration. However, higher rate gives overshooting problems. To solve this issue, reduce learning at plateau is applied to automatically decrease learning rate during training. During training from this approach best result is achieved at learning rate 0.001 and selected for further training. Whereas, dropout rate for preliminary experiments were set to zero to investigate model performance without constraint.After finalizing model architectures and parameters to overcome over-fitting problem, when using dropout rate zero, different dropout-rate is tested and least over-fitting is witnessed on dropout rate at 0.25. Adam optimizer is used for whole training, adam is popular as it provides faster convergence with good generalization capability. Categorical cross-entropy as loss function is used as it is suitable for multi-class classification task. The complete process of model building, training and testing are done on keras with tensor-flow as back-end. And GPU (Tesla K80) of google colab is used for training. The multi-modal is trained for 100 epochs which took approximately 6hours.

Table 1: Parameters and environments for training

Model	Learning rate	Dropout	Number of epochs	GPU	Training duration
Avots[3]	1e-4	0.2	10	NVIDIA Titan XP Pascal	NA
This paper	0.01	0.25	100	NVIDIA Tesla K80	approx. 6hrs

5.1 Visual Modality

Avots have used transfer learning approach for visual modality. They fine-tuned the AlexNet pre-trained model by using 30*30 gray-scale image. They have used more than one frames from a single video , frame level prediction, and used the majority voting after AlexNet to convert prediction to video level. They applied the frame selection method by comparing frame similarity between consecutive frames. They took the average of sum of absolute pixel different of past 10 frames. They selected the key frame, if pixel intensity of next frame is less than average value *1.5. Whereas, in this paper we have used simple key-frame selection method by randomly selecting frames after every 10 frames. We have used the bigger resolution of image (96*96) then in base-line paper. It increased computational cost, but helps to extracting better features. As features of facial-emotion-recognition can better represent by shape of facial muscles and smaller section such as eyes, mouth, furrows. Taking bigger size of image improves recognition with computational cost.

Data augmentation was used for visual model training. This includes rotation, zoom, flip and in addition, color augmentation is used by introducing random brightness and saturation to the image. Input to visual model in this paper is raw pixel intensity. The visual model is first trained on FER2013 dataset and the learnt weights are later used as pre-trained weights for multi-modal. FER2013 is popular dataset for facial emotion recognition as it have expressive images with huge varieties of participant. Also, use of pre-trained weights learnt from dataset of same domain helped to learn salient and discriminate features.

5.2 Audio Modality

The audio model used in paper [3]is prepared by classical machine learning approach using SVM. Two

different set of features was used, 21 statistical feature set and 13 MFCC (spectral). Audio from each video files was first segmented into 400ms and then both statistical and MFCC feature sets were extracted. These features are stacked together and used to train SVM. Finally, prediction for each video was made by using majority voting from output of SVMs for each audio segments.

In this paper , we have used raw audio signal as input to audio model instated of using hand-crafted features. Also the longer sequence of audio is used 2000ms instead of 400ms. Longer sequences increases computational cost. However,it is effective to extract long-term dependencies. At first audio signals are extracted from video and segmented to have equal length of 2seconds. Then down-sampled and normalized. Audio model used in this paper is complete deep learning and training was done in end-to-end manner. Input shape to audio model is (32000,1). Pre-training was done in dataset of same domain (emoDB). The learnt weights are then used in multi-modal training.

5.3 Multi-modal

In paper [3] the multi-modal was created by using majority voting of each frame level prediction from audio and visual model and, final predication was done by summing and normalizing the output of individual model and taking maximum probability as final prediction. The author in paper [3] have used the decision level fusion, and mentioned that decision fusion didn't improved the recognition rate. They have used majority voting as decision fusion. To get the final predication of one emotion class is done as; each audio and visual models gives 6 predicted probabilities of specific class. And highest value amongst these 12 probabilities is considered as final prediction. They conclude during this process individual models didn't worked as complement to each-other.

Whereas in this paper, multi-modal is created by feature level fusion. The use of feature-level fusion helps to learn correlated features between different modalities. However, the process of feature fusion adds additional computation. As meta-classifier used in this fusion needs to train again with individually trained base models (audio and visual). The following process is applied to create feature level fusion: first classifier end of each base model is discarded and their features is concatenated by element-wise stacking. This concatenated layer create feature vector of 512, by combining 256 features from visual and 256 features from audio model. Concatenated layer is further stacked to DNN with two FC (256 units) layer and softmax activation is used to made final prediction. The alignments between audio signal and visual information for training of multi-modal system is done by using one audio sample and one key-frame that represent video. These audio sample and one key-frame gives representation of complete video, and for alignment between two domain, these two samples are used together as input to the multi-modal system. The model is then re-trained on eNTERFACE05 dataset and evaluated on 20 percentage of total data using confusion matrix and classification report.

5.4 Comparison and analysis with state-of-art

Result seen after using multi-modal, improvement can notice from figure 4, 5 and 6. In figure 6, we can see that miss-classification rates are lower than in individual audio and visual model. Comparison with state-of-art shows, use of raw data instead of manually extracted features and feature-level fusion instated of decision fusion with majority voting improved recognition rate for this task. However both of these changes are computationally expensive and requires more amounts of data than that in base-line paper. To solve the less data issue initially training of individual base models on dataset of similar domain and using transfer learning to train on dataset of interest (eNTERFACE05) is found effective but also time and resource consuming. Also the use of larger images size and raw-audio signal help to learn better feature, however, these increases input size to respective models. So, it requires more computation resource than in hand-crafted feature base approach.

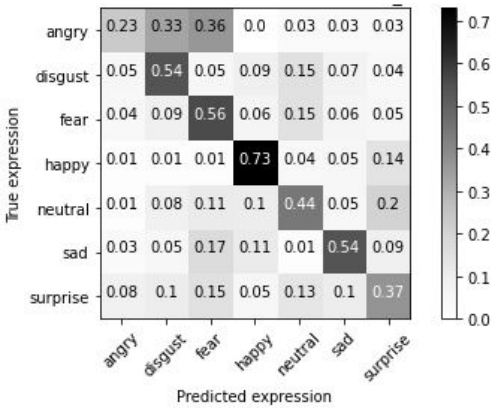

Figure 4: Confusion matrix of audio model

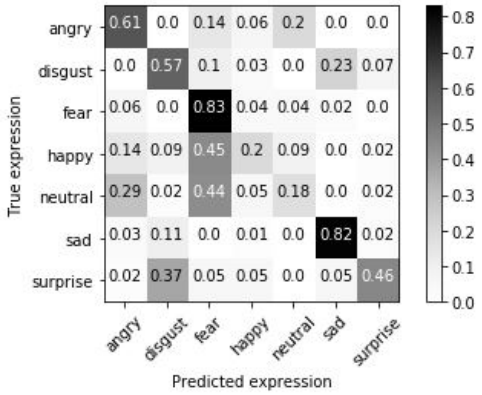

Figure 5: Confusion matrix of visual model

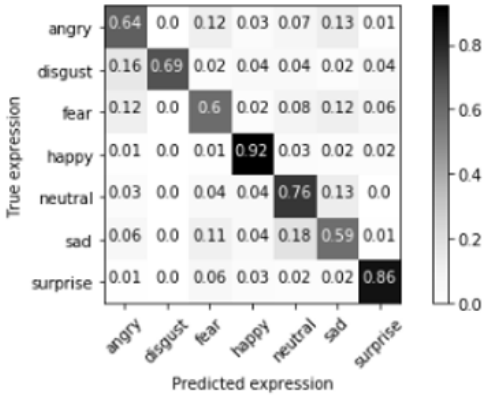

Figure 6: Confusion matrix of multi-model

Table 2: Result comparison with state-of-art

Model	Dataset	Features	Fusion Technique	Accuracy
Avots[3]	eNTER-FACE05	Extracted Manually	Decision level	44.5%
This paper	eNTER-FACE05	Raw data	Feature level	71.5%

6. Conclusion

Emotion recognition is a complex task by nature. As human uses more than one modality to express their emotion, incorporating more than one modality to recognition is not only more natural but also improves the performance of model as one modality acts as complementary to another. In this paper Deep feature fusion based method is used to improve the accuracy of the Emotion recognition to 71%. For multi-modal emotion recognition, feature-level fusion is better suitable for using different modalities as complementary information. Also, the use of raw data as input to audio and visual for feature extraction has potential to improve recognition rate. However, these process requires more computational resources, data.

7. Future work

Further comparative analysis can be made by incorporating different dataset. More clear insights can be made by testing different detests on proposed multi-modal with different fusion technique. Also, on availability of larger dataset, recognition rate can increase without over-fitting by adding more layers (CNN blocks) on individual base models.

Acknowledgments

The authors convey their special thanks of gratitude to all faculties of Department of Electronics and Computer Engineering of IOE, Pulchowk Campus for providing an opportunities and supporting with resources for conducting this work.

References

[1] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological psychiatry*, 44(12):1248–1263, 1998.

[2] Yedid Hoshen, Ron J Weiss, and Kevin W Wilson. Speech acoustic modeling from raw multichannel waveforms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4624–4628. IEEE, 2015.

[3] Egils Avots, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. Audiovisual emotion recognition in wild. *Machine Vision and Applications*, 30(5):975–985, 2019.

[4] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017.

[5] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.

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

[7] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*, 2014.

[8] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013.

[9] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576, 2017.

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

[11] Pierre-Luc Carrier, Aaron Courville, Ian J Goodfellow, Medhi Mirza, and Yoshua Bengio. Fer-2013 face database. *Universit de Montral*, 2013.

[12] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.