The 15th International Conference on Auditory-Visual Speech Processing
10-11 August 2019, Melbourne, Australia

Learning Salient Features for Multimodal Emotion Recognition with Recurrent Neural Networks and Attention Based Fusion

Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, Clinton Fookes

Queensland University of Technology, Brisbane, Australia

dp.don@qut.edu.au, t.warnakulasuriya@qut.edu.au , s.denman@qut.edu.au,
s.sridharan@qut.edu.au, c.fookes@qut.edu.au

Abstract

Automatic emotion recognition is a challenging task since emotion is communicated through different modalities. Deep Convolution Neural Networks (DCNN) and transfer learning have shown success in automatic emotion recognition using different modalities. However significant improvement in accuracy is still required for practical applications. Existing methods are still not effective in modelling the temporal relationships within emotional expressions or in identifying the salient features from different modes and fusing them to improve accuracies. In this paper, we present an automatic emotion recognition system using audio and visual modalities. VGG19 models are used to capture frame level facial features followed by a Long Short Term Memory (LSTM) to capture their temporal distribution at a segment level. A separate VGG19 model captures auditory features from Mel Frequency Cepstral Coefficients (MFCC). The extracted auditory and visual features are fused together and a Deep Neural Network (DNN) with attention is used in classification using majority voting. Voice Activity Detection (VAD) on the audio stream improves performance by reducing the outliers in learning. The system is evaluated using Leave One Subject Out (LOSO) and K-fold cross-validation and our system outperforms state of the art methods on two challenging databases.

Index Terms: Attention based fusion, Deep learning, Facial expression recognition, Multi modal emotion recognition, Speech emotion recognition, Voice activity detection

1. Introduction

Emotion is a complex mental state that drives human thoughts and actions, and is a key aspect of human communication. Emotions are expressed through different modalities including speech, facial expressions and gestures. Humans use emotion information from such modalities in daily social interactions.

Automatic emotion recognition has become a major research area with an increasing focus on human-computer interaction applications. However, the task is still considered a challenging due to the complexity in generalizing expressed emotions. The auditory and visual clues of an expressed emotion are inherently ambiguous from one person to another although the verbal content may be the same. Hence, the challenging part of automating emotion recognition is the generalization. With deep learning architectures that extract much deeper features compared to traditional machine learning approaches, more robust features can be extracted to create a generalized representation of human emotions [1].

In earlier systems, emotions were recognized using unimodal information. In contrast, in a real world setting, humans express emotion as a collection of information through different sensory modalities and communication methods. Therefore, the emotions being exchanged are encapsulated in multi-modal sensory cues. Hence, for a complete and accurate estimate of emotion, the collective information from different modalities should be considered.

Transfer learning along with existing deep learning architectures have been extensively used for emotion feature extraction from different modalities [2] due to the unavailability of larger datasets to train from scratch for emotion recognition tasks. However, the temporal dependencies within each mode are ignored and in most of the existing methods, each feature at a given time step is considered as independent. Recently, 3D-CNN have been utilized to capture the features [3], but convolution over time does not give a reliable estimate intuition on temporal distribution of the features [4]. Recurrent Neural Networks (RNNs) are applied extensively to extract temporal information from samples, especially in speech recognition and machine translation [5] and have the potential to improve the emotion recognition performance.

Many different datasets have been used to carry out experiments on automatic emotion recognition, with popular datasets including RML [6] and eNTERFACE05 [7]. They are different from each other in elicitation method, modalities, language, number of subjects, samples and emotion categories. Considering these variations among datasets, the need for generalized models and effective mechanisms to capture emotional information from the multi-modal input streams is evident.

The main objective of this research is to implement a deep learning model that surpasses the state of the art for automated emotion recognition where we consider audio and visual streams for a video as modalities. In the proposed method, we utilize LSTMs for segment level temporal modelling of visual features, followed by an attention based fusion with auditory features. We use attention based fusion over simple concatenation for fusion since individual modalities can complement each other with salient information.

2. Related Work

Automatic Emotion Recognition (AER) is an extensively studied domain due to the challenges in identifying micro-level features associated with emotions. Both uni-modal and multi-modal methods have been studied, mainly using audio and visual features.

In AER with images, conventional methods typically detect the frontal face region and extract geometric and appearance features. In geometric feature extraction, relationship among different face components have been used as the features for training, specifically using facial landmarks and pose [8]. As an example, Ghimire et al. [9] used transformation of 53 facial feature points in each frame in a video relative to the first frame and used the relative Euclidean distance and angle as the

10.21437/AVSP.2019-5

feature vector [10] and Support Vector Machines (SVM) for the emotion classification. As an alternative to geometric feature based AER, attempts have been made considering the whole face region in appearance based feature extraction [11]. In [12], the authors considered Local Binary Patterns (LBP) as feature vectors while using Principal Component Analysis (PCA) for classification.

In addition to the RGB images, other types of image using infrared [13], depth information [14] and 3D images [15] have been used for AER. Infrared images have shown an increased accuracy compared to visible light images due to their invariability to illumination. In conventional methods for AER, Hidden Markov Models (HMM), SVM and PCA are extensively applied for classification. Mel Frequency Cepstral Coefficients (MFCC) and statistics of audio signals such as pitch and intensity are some of the audio features that have been used to obtain feature vector [16] for speech emotion recognition.

With the breakthrough of deep learning, conventional methods of emotion recognition have been outperformed by Deep Convolutional Neural Networks (DCNN) and Recurrent Neural Networks (RNN). In DCNNs, optimal features from the different modalities for the given task are determined by the DCNN itself. Deep learning architectures like ResNet [17], InceptionNet [18] and AlexNet [19] have been used in feature extraction [20]. To model the temporal relationship among facial features, LSTM and 3D-CNNs have been used [20]. With 3D-CNN, studies have been focused on multi-modal AER to increase the accuracy of the models by exploiting cues from multimodal information [21].

Ranganathan et al. [22] considered body gestures, physiological signals, face and voice as different modalities within a Convolutional Deep Belief Networks (CDBN) for emotion recognition. Kahou et al. [23] combined visual features extracted using CNN and audio features using deep belief nets together for emotion recognition. Chen et al. [24] have proposed a RNN based emotion recognition model using audio, visual and physiological features as modalities. Tzirakis et al. [25] applied a CNN to extract audio features, ResNet50 to extract visual features and a LSTM to remove outliers. In the more recent efforts Yan et al. [26] modelled human emotions by using CNNs to capture visual texture and audio clues and Bidirectional Recurrent Neural Networks to extract dynamic changes of facial textures and landmarks. Zheng et al. [27] have presented a recognition model combining eye movements electroencephalography (EEG) signals.

Even though significant advances have been made in exploiting multi-modal information through deep learning, existing methods are still not effective enough in modelling the temporal relationships of emotion expressions within a video, or in identifying the salient features from different modes in fusion. Our proposed system uses LSTMs to model the temporal relationships and attention to learn the significant features for fusion towards a generalized model.

3. Proposed Method

Our proposed DNN model consists of three main components; audio feature extractor, visual feature extractor and the fusion network as shown in Figure 2. Voice Activity Detection (VAD) is applied to each video sample during the pre-processing step to eliminate noise and remove uninformative features from audio and video streams respectively.

3.1. Voice Activity Detection (VAD)

Unlike in uni-modal emotion recognition, spurious auditory and visual frames may affect the recognition accuracy adversely [28]. In the RML dataset, more than 60% of the video clip contains irrelevant emotion and speechless regions that decrease the overall model accuracy. Figure 1 presents a typical audio signal from the RML dataset and the calculated corresponding Mel-Frequency Cepstral Coefficients (MFCC). It can be seen that a large section of the audio is inactive which may cause reduction in accuracy of the emotion recognition.

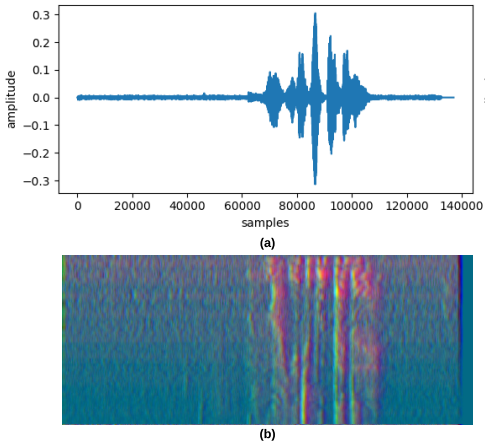

Figure 1: *(a) : Audio stream of a sample clip , (b) Generated MFCC features; VAD is used to remove inactive portions to improve the accuracy of emotion classification.*

Speech inactive audio-visual segments are trimmed at beginning and end of each clip, while still retaining the natural pauses in speech. The audio stream is segmented into time windows with an overlap and the speech activity of each window is detected based on the ratio between speech band energy and the total energy. The speech ratio of the time window t_{i} (SR_{i}) is calculated in the frequency domain as shown in Equation 1. A_{f}, f_{max}, f_{min}, f_{s}, f_{e} refer to amplitude of frequency f, maximum frequency of the signal, minimum frequency of the signal, frequency of start speech band and frequency of end speech band respectively. A time window is deemed voice active if SR_{i} is greater than SR_{th}.

$$SR_i = \frac{\sum_{f_s}^{f_e} A_f^2}{\sum_{f_{min}}^{f_{max}} A_f^2} \begin{cases} active\ \ when\ SR_i\ \geq\ SR_{th} \\ inactive\ when\ SR_i\ <\ SR_{th} \end{cases} \quad (1)$$

A region of interest from ($t_{\mathrm{start}} - \delta_1$) to ($t_{\mathrm{end}} + \delta_2$) is extracted from each video clip. δ_1 and δ_2 are used to capture subtle variations of facial expressions that can be observed before and after the speech activity region. Figure 3 illustrates transformation of facial expressions in a sample video from RML dataset. Apex refers to speech active region and onset and offset refer to pre and post-expressions. We extract speech active regions from each video in the proposed prepossessing step to feed into the deep neural network.

3.2. Audio Network

In our proposed method, we use MFCC channels as input to a VGG19 deep network pre-trained on ImageNet [30]. First, periodogram estimate of the power spectrum of the audio signal

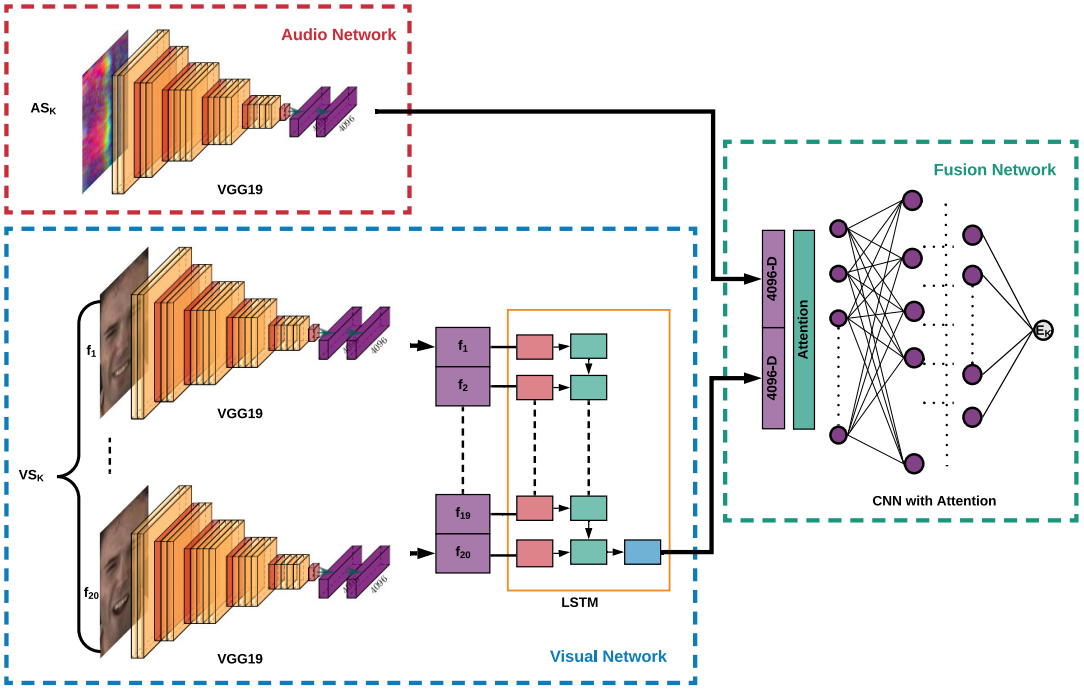

Figure 2: *Proposed architecture - Emotion prediction in the k^{th} video segment (E_k) from audio (AS_k) and visual (VS_k) streams. We use VGG19 [29] to extract audio and visual features. For audio we directly use the output 4096-D feature, for video we pass it temporally through LSTMs. We apply attention based fusion to effectively determine the salient components from the extracted feature vector for the emotion classification task*

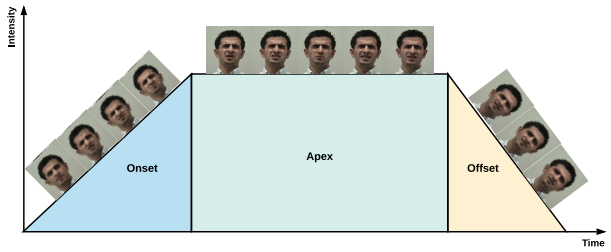

Figure 3: *Illustrates the evolution of active speech and the transition of facial expressions over time.The apex is the voice active region and high intensity emotion expressions are visible. In onset, neutral facial expressions evolves into intense facial expression. During offset, the high intensity of the apex region decreases towards a neutral expression.*

is calculated using a Discrete Fourier Transform (DFT) as in Equation 2 and 3 [31]. $S(n), i, P_i(k), h(n)$ refer to audio signal, frame number, power spectrum of frame i and Hamming window respectively.

$$SR_i(k) = \sum_{n=1}^{N} s_i(n)\, h(n)\, e^{\frac{-2jkn\pi}{N}} \tag{2}$$

$$P_i(k) = \frac{1}{N} \mid S_i(k) \mid^2 \tag{3}$$

A mel filter-bank is applied to the power spectrum and the logarithm of the summed energy of each filter is taken. Then MFCC values are calculated by taking the Discrete Cosine Transform (DCT) of log energies. Utterance level MFCC features (static) are then segmented into shorter frames with an overlap. Finally differential (delta) and acceleration (delta-delta) coefficients are calculated as in Equation 4. Finally, static, delta and delta-delta channels are rescaled to the range 0-255 to represent the three channels of RGB image.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{4}$$

We used 64 Mel-filter banks from 20 Hz to 8000 Hz to acquire log Mel-spectrogram with 10ms of overlap and 25ms Hamming window. A context window of 64 frames with a shift of 30 frames is used to obtain MFCC segments. The length of each corresponding audio segment is 655ms and the size of the output is 64*64*3. We use VGG19 to retrieve 4096-D audio features from fc7 layer which are fused with visual features from the corresponding video segment in a later step.

3.3. Visual Network

The visual network consists of two major components: one to capture facial expression in a given face and other to capture the temporal transformation of emotion expression within the segment. We align an audio segment with a video segment for meaningful feature level fusion. 20 video frames (655*30/1000 = 20) represent one audio segment when the frame rate is 30 frames per second (as in RML). First, we crop frontal face region which represents the emotion from each frame using a single-shot multi-box detectors [32] with ResNet backend and resize the region to 227*227*3. Adjustments are made in following occasions to ensure the correct number of frames from each segment of a video.

1. If a frontal face is not detected in a frame, then the frame is replaced by nearest frame with a detected frontal face.

2. If total frames per segment exceed 20, (n-20)/2 frames are removed from the beginning and end of the segment.
3. If total frames per segment are less than 20, then the first and last frame are duplicated.

Then we retrieve visual features of a video segment sequentially passing each frame in the segment through a VGG19 model, pre-trained on ImageNet. We extract feature representations of last fully connected layer (fc7) and stack them together to form a feature shape (20*4096-D). Then the stacked features are passed through a LSTM module to capture the temporal relationships among stacked features where LSTM model represents the input using a 4096-D feature vector. the 4096-D feature vector from the LSTM module is then fused with audio features in the next step.

3.4. Feature Fusion

Feature level fusion is used to fuse the audio and video modes. Given a video segment, the deep learning model learns feature vectors of 4096-D for both auditory and visual streams. Feature vectors are concatenated together to obtain fused vector of size 8192-D. A neural network with attention is used to identify significant features with reduced dimensionality. We calculate attention weights and the weighted combined vector as shown in Equation 5, 6 and 7 [33]. c_t, h_t, β_t, $\hat{h}$ and q refer to merged feature vector, neural network output, attention score, context vector which is randomly initialized and jointly trained with other components of the network, and output respectively.

$$h_t = tanh\,(\,W_h\,c_t\,+\,b_h\,) \tag{5}$$

$$\beta_t = \frac{exp\,(\,|h_t|^\top\,\hat{h}\,)}{\sum_t\,exp\,(\,|h_t|^\top\,\hat{h}\,)} \tag{6}$$

$$q = \sum_t\,\beta_t\,\times c_t \tag{7}$$

Finally, the segment is classified using softmax activation in the final dense layer of the DNN. Video level emotion is obtained by majority voting over all segments.

4. Experimental Setup

Experiments are conducted on two audio-visual datasets, RML and eNTERFACE05, where both contain six basic emotions; anger, disgust, fear, sadness, happiness and surprise. RML contains 500 acted video samples from eight subjects. "eNTERFACE05" contains 1166 induced video samples from 42 subjects (34 men and 8 women). RML includes video clips from six different languages while eNTERFACE05 is only in English. During the preprocessing phase, some video samples with rapid face movements (where frontal face cannot be detected in many successive frames) are removed from dataset. We utilize Leave One Subject Out (LOSO) cross validation and k-fold cross validation since both the methods are used in literature.

5. Results and Discussion

To verify the performance of our proposed AER system, we report results on two public datasets, eNTeRFACE05 and RML. We train our models in two phases: feature extraction and fusion network. We train our audio and visual feature extraction networks first and then the resultant features are passed through an attention based neural network for AER. We have selected baseline models that have used multimodal fusion.

We considered the model proposed by Zhang et al. [34] as our baseline model for LOSO cross-validation. They utilized a structure similar to AlexNet to extract features and used a simple neural network for multimodal fusion. Our model outperformed their accuracy as shown in Table 1, as we employ LSTMs to model the temporal feature distribution and attention to capture salient features from each feature set. Figures 4 and 5 show the confusion matrices of test results of the proposed model when tested on RML and eNTERFACE datasets, respectively.

Table 1: *Recognition accuracy for RML & eNTERFACE05 datasets with LOSO cross-validation compared with state of the art methods.*

Dataset	Refs.	Accuracy
RML	$Sarvestani\ et\ al.$ [35]	72.03%
	$Zhang\ et\ al.$ [34]	74.32%
	Ours	**76.13%**
$eNTERFACE$	$Sarvestani\ et\ al.$ [35]	70.11%
	$Bejani\ et\ al.$ [36]	77.02%
	Ours	**78.49%**

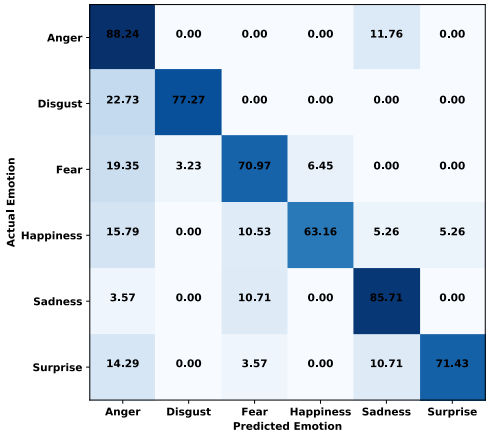

Figure 4: *Confusion matrix of the proposed approach on the RML dataset for a LOSO evaluation.*

"Anger" has the best recognition accuracy for both the datasets while happiness performs worst. However, recognition accuracy remains lower in LOSO methods since the subjects in the testing samples are withheld from training data. Due to the small sized datasets, deep learning model is unable to generalize which causes a decrease accuracy on the test set. It can be observed that majority of the misclassified emotions have been identified as "anger". Due to the synthetic nature of datasets, video level emotion may not be represented in each frame. In particular, disgust and fear may have similar mouth motions to anger which make them hard to distinguish.

Next, we compare our model with state of the art methods using a K-fold cross-validation. We use the model proposed by Seng et al. [37] as our baseline model. They have used Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) for visual feature extraction with PCA for dimensionality reduction. They have combined prosodic features with spectral features for audio emotion classification with a rule based ap-

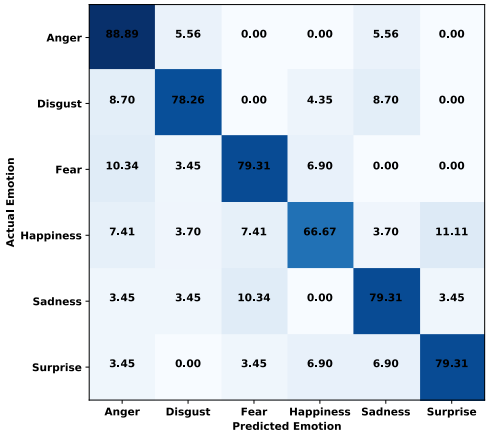

Figure 5: *Confusion matrix of the proposed approach on the eNTERFACE05 dataset for a LOSO evaluation.*

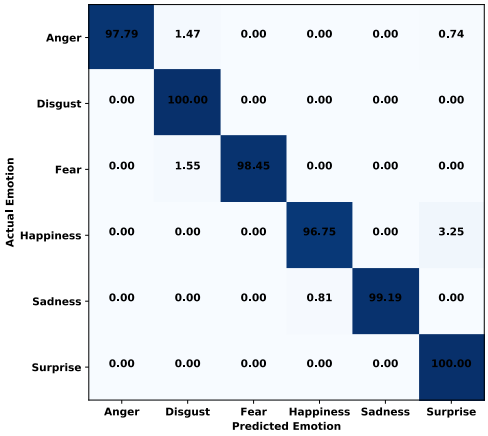

Figure 6: *Confusion matrix of the proposed approach on the RML dataset for a K-fold cross validation evaluation.*

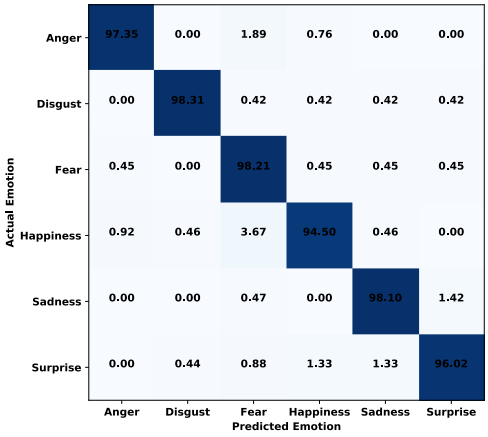

Figure 7: *Confusion matrix of the proposed approach on the RML dataset for a K-fold cross validation evaluation.*

proach followed by multimodal fusion. Wang et al. [6] used kernel cross-modal factor analysis to represent the patterns between features of different modalities. They use kernel canonical correlation analysis to maximize the correlation and kernel matrix fusion for emotion classification.

Our proposed model outperformed the accuracy of the above method by nearly 5%. Figure 6 & 7 presents the confusion matrices for both the datasets in K-fold cross validation. Our deep networks for audio-visual feature extraction have outperformed their conventional method of feature extraction. Temporal modelling and attention based fusion are the other key factors which have led to the increased accuracy of our method.

Table 2: *Accuracy obtained for RML & eNTERFACE05 datasets with K-fold cross-validation compared with state of the art methods*

Dataset	Refs.	Accuracy
RML	$Wang\ et\ al.$ [38]	82.22%
	$Seng\ et\ al.$ [37]	90.83%
	Ours	**98.70%**
$eNTERFACE$	$Wang\ et\ al.$ [38]	72.47%
	$Seng\ et\ al.$ [37]	86.67%
	Ours	**97.75%**

6. Conclusion

In this paper, we present a model for multi-modal emotion recognition, in which we capture the evolution of the emotion and use attention to fuse salient feature from each mode. We used transfer learning on a VGG19 model pretrained on ImageNet to retrieve acoustic features of an audio segment and visual features of a video frame, and we modeled the temporal flow of visual features of a video using LSTMs. The extracted spatial-temporal features from visual and audio models are fused using an attention-based neural network for segment level classification followed by majority voting for video level classification.

We have achieved accuracies of 76.1% and 78.5% for a LOSO cross-validation and 98.7% and 97.8% for a K-fold cross-validation for RML and eNTERFACE05 respectively. For both databases and evaluations our method outperformed the state of the art as shown in Table 1 and 2.

We have used two-stage learning to train the AER in which we trained audio-visual network and fusion network in the two phases. LOSO cross-validation shows less accuracy compared to K-fold cross-validation due to the limited generalization of the AER model as a result of small datasets. We are planning to investigate methods of obtaining more generalized emotion representations by deep learning models using small datasets in the future.

7. Acknowledgements

This research was supported by an Australia Research Council (ARC) Discovery grant DP140100793.

8. References

[1] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3687–3691.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[3] A. T. Lopes, E. De Aguiar, and T. Oliveira-Santos, "A facial

expression recognition system using convolutional networks," in *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2015, pp. 273–280.

[4] S. Arif, J. Wang, T. Ul Hassan, and Z. Fei, "3D-CNN-based fused feature maps with LSTM applied to action recognition," *Future Internet*, vol. 11, no. 2, p. 42, 2019.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[6] Y. Wang and L. Guan, "Recognizing human emotion from audiovisual information," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2. IEEE, 2005, pp. ii–1125.

[7] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[8] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system-A case study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 132–137.

[9] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and Support Vector Machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[10] 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*, 2016, pp. 5562–5570.

[11] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1386–1398, 2015.

[12] S. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *2012 4th International conference on intelligent human computer interaction (IHCI)*. IEEE, 2012, pp. 1–5.

[13] W. Wei, Q. Jia, and G. Chen, "Real-time facial expression recognition for affective computing based on Kinect," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2016, pp. 161–165.

[14] M. Szwoch and P. Pieniazek, "Facial emotion recognition using depth data," in *2015 8th International Conference on Human System Interaction (HSI)*. IEEE, 2015, pp. 271–277.

[15] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," 2009.

[16] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] 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*, 2015, pp. 1–9.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] B. H. Mohammad Mahoor *et al.*, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 30–40.

[21] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition," *Computer Vision and Image Understanding*, vol. 174, pp. 33–42, 2018.

[22] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[23] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[24] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using Recurrent Neural Networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 49–56.

[25] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[26] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.

[27] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE transactions on cybernetics*, no. 99, pp. 1–13, 2018.

[28] P. Harar, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2017, pp. 137–140.

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

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[31] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[32] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot Multibox Detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[33] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.

[34] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 281–284.

[35] R. R. Sarvestani and R. Boostani, "FF-SKPCCA: Kernel probabilistic canonical correlation analysis," *Applied Intelligence*, vol. 46, no. 2, pp. 438–454, 2017.

[36] M. Bejani, D. Gharavian, and N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Computing and Applications*, vol. 24, no. 2, pp. 399–412, 2014.

[37] K. P. Seng, L.-M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 3–13, 2018.

[38] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia - TMM*, vol. 14, pp. 597–607, 06 2012.