


Audio-Visual Emotion Recognition System Using Multi-Modal Features

Anand Handa, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India

Rashi Agarwal, UIET, CSJM University, Kanpur, India

 <https://orcid.org/0000-0002-5768-5894>

Narendra Kohli, Harcourt Butler Technical University, Kanpur, India

ABSTRACT

Due to the highly variant face geometry and appearances, facial expression recognition (FER) is still a challenging problem. CNN can characterize 2D signals. Therefore, for emotion recognition in a video, the authors propose a feature selection model in AlexNet architecture to extract and filter facial features automatically. Similarly, for emotion recognition in audio, the authors use a deep LSTM-RNN. Finally, they propose a probabilistic model for the fusion of audio and visual models using facial features and speech of a subject. The model combines all the extracted features and use them to train the linear SVM (support vector machine) classifiers. The proposed model outperforms the other existing models and achieves state-of-the-art performance for audio, visual, and fusion models. The model classifies the seven known facial expressions, namely anger, happy, surprise, fear, disgust, sad, and neutral, on the eNTERFACE'05 dataset with an overall accuracy of 76.61%.

KEYWORDS

Audio-Visual Emotion Recognition, Convolutional Neural Network (CNN), Long Short-Term Memory Recurrent Neural Network(LSTM-RNN), Low-Level Descriptors, Machine Learning, SVM (Support Vector Machine)

INTRODUCTION

Computer vision, in recent years, has witnessed outstanding and productive outcomes because of the tasks like face recognition, emotion recognition, and speech recognition. The reason is the adaptation of high-end techniques like machine learning. However, human expression recognition is still an onerous task. The first Emotion Recognition in Wild (EmotiW) (Dhall et al., 2013) challenge was held in the year 2013. Since then, the classification accuracy has increased to a great extent from a baseline figure of 38% but still, there is a scope of improvement. There are several reasons in the past for low accuracy percentage such as there is a lack of labeled video datasets, the nature of facial expressions is ambiguous, and the effectiveness of the methods of extracting facial expression is less. In the last few years, techniques like Deep Convolutional Neural Network (DCNN) (Schmidhuber, 2015) is proven to be outstanding in extracting features from an image. Also, Long Short Term Memory (LSTM) is proven to be the best in analyzing sequential data (Sak et al., 2014). Thus, by

DOI: 10.4018/IJCINI.20211001.0a34

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

applying all these recent and effective methods and combining them may increase the accuracy of classifying the human facial expressions more effectively. The main contributions of this paper can be summarized as follows:

- A separate feature selection model is introduced in AlexNet architecture which automatically filters the most prominent facial features. It helps in an overall improvement of the accuracy of the model.
- Separate models for audio and visual emotion recognition with better classification accuracy.
- A probabilistic audio-visual fusion model using SVM machine learning classifier which classifies the emotions with a better accuracy.

The rest of the paper is organized as follows: Section 2 discusses the related work. In section 3, the authors present the multi-modal emotion recognition framework, including the discussion of datasets, multi-modal features, and network architecture. In section 4, the authors present the experimental setup for the audio and visual emotion recognition. In section 5, the experimental results from the audio, video, and audio-visual fusion-based recognition models are discussed separately, and Section 6 concludes the paper.

RELATED WORK

A multi-modal approach for an emotion recognition system is more powerful and efficient than the bimodal and unimodal approaches because human emotions depend on both audio and visual information. In recent years, many studies came up, which are based on audio-visual recognition of human emotions and they also prove audio and visual fusion for emotion recognition to be advantageous. In this section, the authors discuss a few of them.

M. Mansoorizadeh et al. (Mansoorizadeh and Charkari, 2010) propose a fusion-based approach to emotion recognition. It uses both decision and feature level fusion. Features which are related to the same emotion has a higher chance of getting overlapped. The proposed framework combines features of the different modalities and generates a hybrid feature space. The experiments are performed on two different audio-visual emotion databases with a total number of 42 and 12 subjects. The proposed model accuracy is comparatively higher than the unimodal and bimodal face and speech-based individual systems.

An audio-visual recognition system based on the fusion of features is proposed by R. Gajsek et al. (Štruc et al., 2010). For the audio-based recognition model, the coefficients -- cepstral and prosodic are extracted, and for video-based recognition model, Gabor wavelets are considered as features. Lastly, to combine the outputs, a multi-class classifier is used.

International Journal of Cognitive Informatics and Natural Intelligence

In (Avots et al., 2019), authors present the analysis of an audio-visual model for emotion recognition. They use three different databases SAVEE, eNTERFACE'05, and RML for training the models and AFEW database is used as a testing set. MFCC coefficients are used to represent the emotional speech and SVM machine learning classifier is used for classification. The proposed multimodal emotion recognition is a decision-based fusion model. They perform the facial image classification using AlexNet. The reported accuracy for eNTERFACE'05 is 48.2%.

In another paper (Ouyang et al., 2017), authors present an audio-visual approach that uses transfer learning and a combination of models for the human emotion recognition system. The approach shows that transfer learning is an effective way of finding better image features. It also shows that the combination of recurrent neural network and geometric based video features are better and more effective for information extraction. In many scenarios like HCI, reinforcement learning is more effective. The proposed model achieves a final accuracy of 57.2% on the test dataset.

Another fusion-based multi-modal recognition system is proposed by D. Datcu et al. (Datcu and Rothkrantz, 2011). It is a semantic fusion model which uses geometric features of a face either in the presence of speech or in the absence of speech. They consider only features related to eye and eyebrow to remove any effect of speech on the face.

Huang et al. (Huang et al., 2013) present an automatic multi-modal emotion recognition system. The main characteristic is that for each modality, it learns the decision parameters automatically. Each classifier acts as an expert and then equalized quantitatively and finally, the weighted sum of all the quantities that were equalized earlier, results in the final decision.

An approach based on semi-supervised clustering is proposed by Meghjani (Meghjani et al., 2009). Their approach divides the frames into two clusters. One is an emotion frame cluster, and another is a non-emotion frame cluster. The maximum number of continuous frames makes the distinction. Fusion is done by using the score and feature levels obtained from the above two mentioned modalities.

Paleari et al. (Paleari and Huet, 2008) propose a fusion-based recognition model and achieve an overall accuracy of 67%. In this paper, authors consider different scenarios for emotion recognition. It includes different modalities, different feature sets, the fusion algorithms, and also the different optimization methods. The optimization methods used are temporal averaging.

In (Wang et al., 2012), the authors discuss several kernel-based methods and propose an approach for the analysis of an existing modality relationship. They investigate the kernel-based methods for multi-modal information analysis. In this paper, the authors present a new approach to analyse the relationship between different modalities. The proposed method minimizes the Frobenius distance in the transformed domain to characterize the coupled patterns.

The face expression recognition basically involves three important steps. They are namely registration, feature extraction and classification (El Ayadi et al., 2011). In face registration, they locate the face in a particular image using a set of points known as landmarks. Face registration is carried out by using the faces detected in an image and then they are normalized to match a template image. Secondly, the features are categorized as one or more as a combination of pixel intensities, Local Binary Patterns (LBP), Histogram of Oriented Gradients (HoG) and Local Phase Quantization.

(X. Zhang & Ma, 2018) propose a multiple kernel type learning algorithm. However, it is the capability of the neural network to learn features that helps the network in classifying the input data with high accuracy. Lastly, the algorithm uses a machine learning algorithm to classify the facial expressions as one of the six basic emotions.

A deep neural network provides an architecture which is able to learn multiple levels of abstraction and multiple levels of representations. These models could help us in locating complex patterns in images and in various other sources. They also are able to perform exceptionally well in time restricted scenarios. There are various existing models that perform well and achieve good results. One of them is AlexNet (Ballester & de Araújo, 2016). It is one of the first networks that has introduced a method for solving the problem of overfitting and has used the dropout method. The limitation of all the traditional CNN models is a large amount of time and the number of operations. It takes nearly 100M operations for one iteration of full AlexNet.

Due to the availability of high computational power to work on the large database. Nowadays, the neural network architecture (Sánchez et al., 2017b) has become more popular. By using deep neural networks, one can obtain better results in various domains like object recognition, estimation of human poses, face verification, etc. In last few years, technique like DCNN is proven to be outstanding in extracting the features from an image. Long Short-Term Memory (LSTM) is proven to be the best in analyzing sequential data. Thus, by applying all these recent and effective methods and combining them may increase the accuracy of classifying the human facial expressions. It will also help in building effective multi-modal recognition systems (Melin & Sánchez, 2018) and (Sánchez et al., 2017a).

The neural network (Lopes et al., 2017) is unlike the earlier existing machine learning approaches because they are able to extract undefined features from the training database. In the proposed model, the authors train the neural network architecture on various openly available databases and then

Table 1. Summary of datasets used

Dataset	Facial Expressions	Sample Details
FER2013 (Carrier et al., 2013)	Neutral (Ne), Sad (Sa), Surprise (Su), Happy (Ha), Fear (Fe), Anger (An), Disgust (Di)	35887 static images of unknown subjects
eNTERFACE'05	Neutral (Ne), Sad (Sa), Surprise (Su), Happy (Ha), Fear (Fe), Anger (An), Disgust (Di)	Audio-visual clips of 44 subjects

they perform cross-validation on them. This helped the authors to accurately identify the network's performance.

In the proposed work, the authors present a multi-modal recognition technique, and they use different channels for expressions of a subject. One is an audio channel, and another is a video channel. The experiments are performed on the eNTERFACE'05 dataset (Martin et al., 2006). It is an audio-visual database of 44 subjects. Finally, the authors propose a fusion-based multi-modal recognition model that outperforms the other models, which are discussed in the literature, and the model achieves an overall accuracy of 76.61%.

PROPOSED FUSION MODEL

For the multi-modal recognition system, the authors first implement an emotion recognition model in a video using the Alexnet model. Then they implement an emotion recognition model in speech using the LSTM-RNN model. At last, the authors fuse both the models that result in the proposed fusion model for the audio-visual recognition system. An emotional state of a subject can be efficiently represented with the help of multi-modal features. Visual features such as MSDF (Sikka et al., 2012), and LBP-TOP and acoustic features such as MFCC, MFB, and PCM (Schuller et al., 2009) are considered for this multi-modal recognition system. The authors extract the low-level descriptors (LLD) (Schuller et al., 2007) by using the OpenSMILE toolkit (Eyben et al., 2010) and then apply the same set of transformations to these extracted LLD.

Dataset

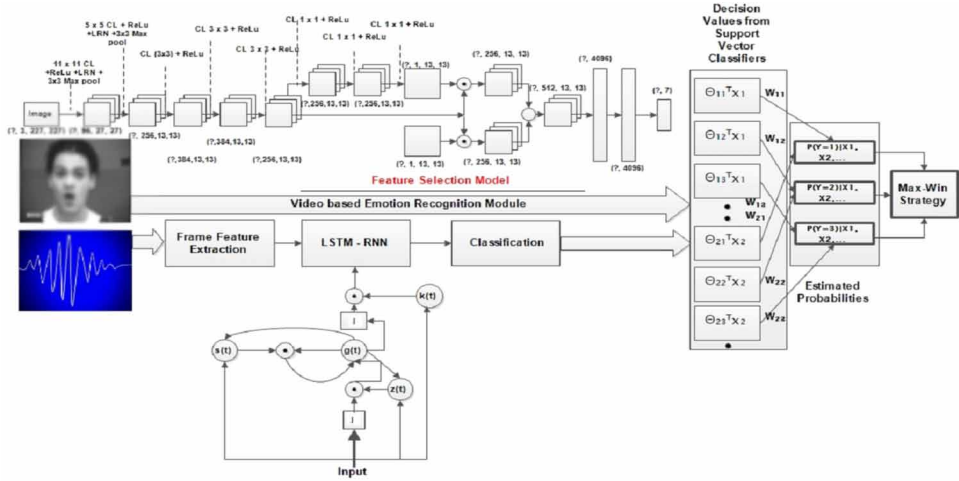
Table 1 shows the details of the datasets considered for training and testing the models.

Emotion Recognition System In A Video

The proposed network structure, as shown in Figure 1 is based on AlexNet architecture in which a CNN based feature selection mechanism is implemented in between the convolutional layer and the FC (fully connected) layer of the AlexNet model to locate the facial features. This is done to obtain details of the facial features more effectively, ignoring the negative effect of the background. The authors train the model using the images from the FER2013 dataset. In the first layer, to fit the model, the authors resize the images to 256×256 and later crop them to a size of 227×227 . Figure 2 shows a sample image frame from the dataset. The feature selection model is introduced because of the following two reasons:

1. The last convolutional layer of the first branch has only one feature map which helps in locating the prominent facial features.
2. It is to be noted that the weights of different facial features are different and hence every landmark on the face has a different level of importance. Therefore, the prominent landmarks can be highlighted, and the insignificant parts can even be eliminated or weakened. This is done to achieve better generalization efficiency.

Figure 1. Proposed Model with Audio, Video and Fusion based Recognition System. (CL- convolutional layer, ReLu- Rectified Linear Units, LRN- Local Response Normalization, I- activation functions)



The second branch of the feature selection model network uses a face mask as input data. It is used to get the facial landmarks from which the face masks are generated as described in Figure 1. The size of the layer in both the branches are same, i.e., 13×13 . The face mask helps in removing all the background features because they have a significant effect on FER model. Now, the weight map and face mask help in filtering out the facial features in the fifth convolutional layer. Figure 3 shows the feature selection model.

The feature selection model consists of two branches. One branch consists of three convolutional layers with 1×1 filter size followed by the selection layer. The second layer has an input layer that has a face mask, followed by a selection layer. After every iteration, the first branch is updated. The third layer output is known as a weight map that is used to estimate whether a feature in a given location is to be taken into consideration or should be left. So, the authors can now restrict the influence of features to a certain limit at unwanted locations of the face. Figure 3 shows the working of this

Figure 2. Facial landmarks on a sample video frame of eNTERFACE'05 dataset

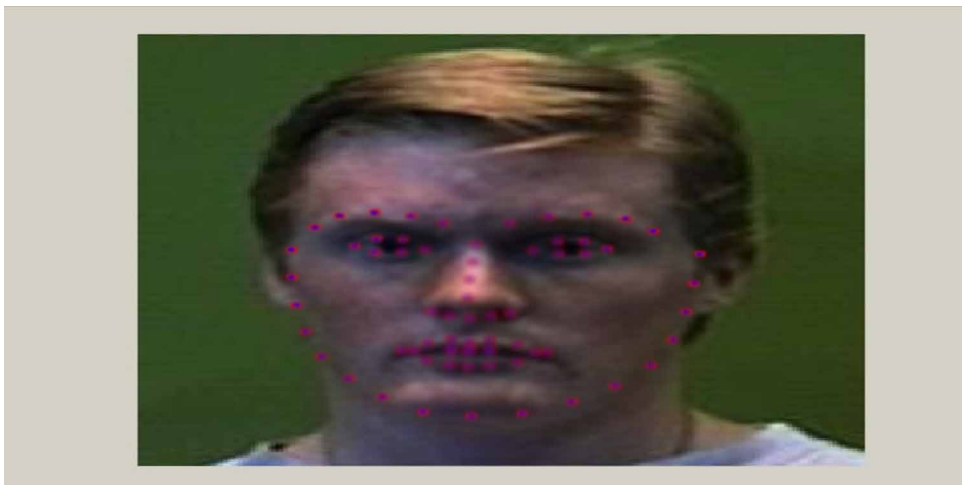
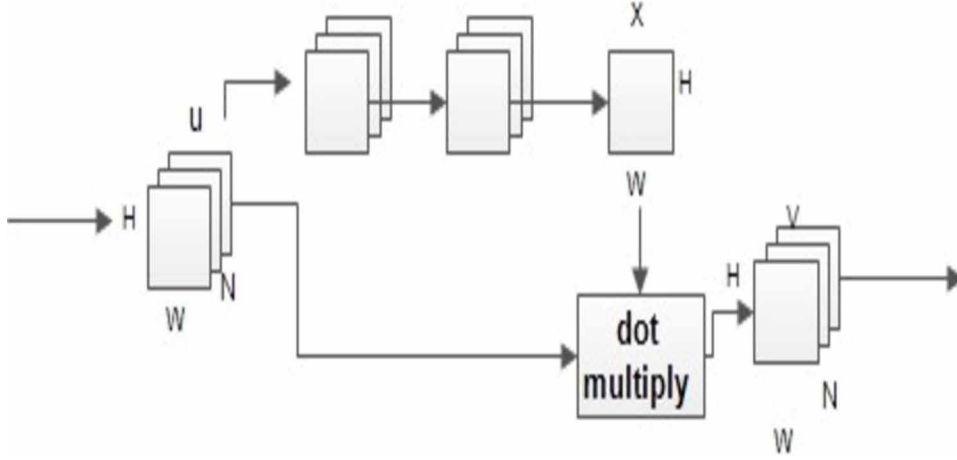


Figure 3. Feature Selection Model



model. \mathbf{U} is an activation tensor that is generated from the last convolutional layer in the first part of this model. The number of feature channels are given by \mathbf{N} . The dimension of the tensor \mathbf{U} is $\mathbf{H} \times \mathbf{W}$. Finally, a matrix \mathbf{X} is generated. The selected features are obtained by multiplying the \mathbf{X} with initial activation function \mathbf{U} :

$$V_n = X.U_n \quad (1)$$

where U_n is the n^{th} feature channel of \mathbf{U} , and V_n is the selected feature. Gradient Normalization is done for \mathbf{X} by \mathbf{N} feature maps because \mathbf{N} features map equally affects all the maps in \mathbf{U} . To obtain the facial landmarks more precisely, the authors apply facial landmark detection algorithm (Baltrušaitis et al., 2016). This is done in the following way:

1. A face outline is created from the input face image.
2. Every pixel value which is inside the boundary is labeled as 1 and pixel outside is labeled as 0.
3. Finally, in order to fit the image which is output from last convolutional layer is resized to 13×13 .

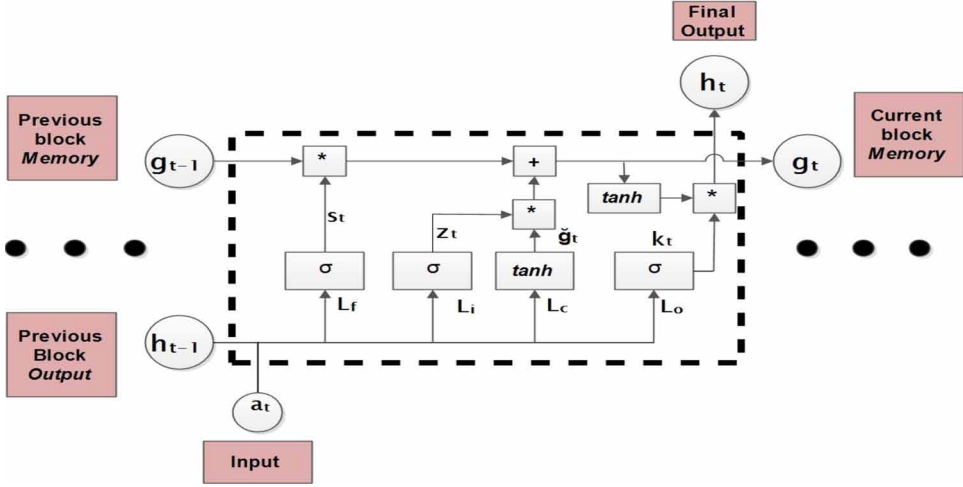
Speech Based Emotion Recognition System

For a speech-based emotion recognition system, the authors use an LSTM-RNN. The authors train the network and use memory block as the activation function. Figure 4 shows the single memory block and the sub- components. The mathematical equations for activation functions and memory blocks outputs are as follows:

$$z_t = \sigma(L_i a_t + M_i m_{t-1} + p_i) \quad (2)$$

$$s_t = s(L_f a_t + M_f m_{t-1} + p_f) \quad (3)$$

Figure 4. LSTM-RNN architecture for audio based emotion recognition (where z as Input gate, s as forget gate, k as output gate and g as memory cell)



International Journal of Cognitive Informatics and Natural Intelligence

$$g_t^c = \tanh(L_c a_t + M_c m_{t-1} + p_c) \quad (4)$$

$$g_t = z_t * g_t^c + s_t * g_{t-1} \quad (5)$$

$$k_t = \sigma(L_o a_t + M_o m_{t-1} + N_o g_t + p_o) \quad (6)$$

$$h_t = k_t * \tanh(g_t) \quad (7)$$

Classifier

SVM is a binary classifier which helps in classifying the data points either in one category or in the rest categories at any particular time. The authors use linear SVM (Hearst et al., 1998) as their basic classifier as they are having all linearly separable features. SVM solves an unconstrained optimization problem (Fan et al., 2008) represented as follows:

$$\min \frac{1}{2} \theta^T \theta + c \sum_i \xi(\theta; \theta_i, y_i) \quad (8)$$

where c is the penalty parameter and the loss function is $\xi(\theta; \theta_i, y_i) = \max(1 - y_i \theta^T x_i, 0)$. SVM uses the following mathematical formulation for testing data for predicting it as either +ve or -ve.

$$f(x) = \begin{cases} +ve, & \text{for } \theta^T x > 0 \\ -ve, & \text{for } \theta^T x \leq 0 \end{cases} \quad (9)$$

$\theta^T x$ is the decision value of the above classifier. It is fed to the next process, which is the fusion process. One-vs.-Rest strategy is followed because SVM is a binary classifier which helps in classifying the data points. It either classifies to one or to the rest ones of the categories.

Fusion Model

In this section, the authors describe a fusion network to integrate the output of each SVM classifier. A set of $m \times n$ decision values are generated by the classifiers which are represented as $a_{(j,k)} = \theta_{jk}^T X_j$ where $j = 1, \dots, m$ and $k = 1, \dots, n$. Fusion network uses these values as input.

Consider, $h_w(c)$ as an input function for any input c . Probability $P(q=k/c)$ is the probability of a y -labeled class and it takes n different possible values. Finally, a dimensional vector of size n representing n different probabilities is derived. The methodology used for the final prediction is “*max-win methodology*”. The most likely labels are chosen over the other class labels. The hypothesis function $h_w(c)$ is as:

$$h_w(c^{(i)}) = \begin{bmatrix} P(q^{(i)} = 1|c^{(i)}; W) \\ P(q^{(i)} = 2|c^{(i)}; W) \\ \vdots \\ P(q^{(i)} = n|c^{(i)}; W) \end{bmatrix} = \frac{1}{\sum_{k=1}^n e^{\sum_{j=1}^m W_{jk}^T c_{jk}^{(i)}}} \begin{bmatrix} e^{\sum_{j=1}^m W_{j1}^T c_{j1}^{(i)}} \\ e^{\sum_{j=1}^m W_{j2}^T c_{j2}^{(i)}} \\ \vdots \\ e^{\sum_{j=1}^m W_{jn}^T c_{jn}^{(i)}} \end{bmatrix} \quad (10)$$

In the proposed model, the authors use $J_{(w)}$ as a loss function for optimization and to get the optimized values for W , gradient descent method is used. It updates W to $W - \tilde{W}_{(w)}$ after every iteration. The mathematical equation for calculating $J_{(w)}$ is given as:

$$J(W) = -\frac{1}{L} \left[\sum_{i=1}^L \sum_{k=1}^m 1\{y^{(i)} = k\} \log \frac{e^{\sum_{j=1}^m W_{jk}^T c_{jk}^{(i)}}}{\sum_{k=1}^n e^{\sum_{j=1}^m W_{jk}^T c_{jk}^{(i)}}} \right] + \frac{\sum_{j=1}^m \sum_{k=1}^n W_{jk}^2}{2} \quad (11)$$

$$\nabla W_k = -\frac{1}{L} \sum_{i=1}^L \left[x^{(i)} \left(1\{y^{(i)} = k\} - P(q^{(i)} = k|c^{(i)}; W) \right) \right] + \sum_{j=1}^m W_{jk} \quad (12)$$

λ is defined as the parameter for L-2 regularization and the indicator function is $1\{\cdot\}$. The indicator function $1\{\cdot\}$ is defined on a set say Y that indicates membership of an element in a subset A of Y . Here, $1\{\cdot\}$ represents an indicator function on a set of values $\{y^{(i)} = k\}$ which means that the value of the function is 1 when $y^{(i)} = k$ otherwise 0. In the fusion network, same value of W_j is used which is shared between the other feature's decision values ($W_{j1}=W_{j2}=W_{j3}=\dots=W_{jn}$).

EXPERIMENTAL SETUP

Configuration Details For Emotion Recognition System In A Video

Caffe toolkit is applied to fine-tune AlexNet Deep CNN model. The model is pre-trained, and the authors approximately use 28,000 images for tuning and scaling the FER2013 training data. It is done for scaling the FER2013 $48 \times 48 \times 1$ image data to $256 \times 256 \times 3$ to meet the input requirement of the CNN model. The initial learning rate is set to 0.001 and the weight decay is set to 0.001 because hyper-parameters are applied to the model. The batch size is initialized as 128, and the stochastic gradient descent is considered as 0.9. It is to be noted that the FC layer is completely retrained to make the learning faster by changing the multiplier for learning rate from 0.001 to a value of 0.004. Also, after every 10 epochs, the new learning rate is *one-tenth* of the previous epoch.

Table 2. Bidirectional LSTM configuration

Model	Size of Input	Layer 1	Layer 2	Layer 3	Layer 4	Size of Output
M1	88	128	128	-	-	7
M2	88	192	192	-	-	7
M3	88	256	256	-	-	7
M4	88	128	256	128	-	7
M5	88	128	256	256	128	7

Configuration Details For Emotion Recognition System In An Audio

Table 2 shows the Bidirectional LSTM-RNN architecture. The authors evaluate five different Bidirectional LSTM-RNN architectures. For training Long Short Term Memory networks, CURRENT Toolkit (Weninger et al., 2015) is used. Out of the five architectures from M1 to M5, the authors select the best performing model as the final model.

EXPERIMENTAL RESULTS

The authors measure the performance for all the seven facial expressions. Figure 5, 6, and 7 shows the confusion matrices for various recognition models based on video and audio features separately and also the confusion matrix for fusion-based multi-modal recognition system. The performance of fusion-based multi-modal recognition system is 76.61%.

In the proposed model, the authors divide the original dataset into two parts. One is a training set, and another is a validation set. During the training phase, the authors monitor the validation accuracy. A state with the highest validation set accuracy is then saved. A trained model is then estimated using test sets for the final accuracy evaluations. Figure 8 shows us the accuracy of the validation and testing of dataset.

In this work, the authors also validate the experimental results using ROC (Receiver Operating Characteristic) and PR (Precision-Recall) curves. The ROC curves use False Acceptance Rate (FAR) and Validation Rate (VR) whereas

Figure 5. Confusion matrix for speech-based emotion recognition model

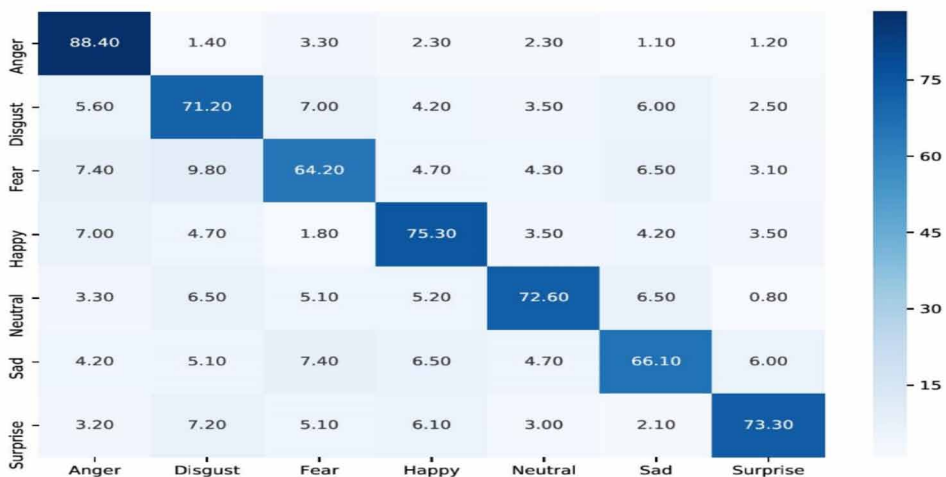


Figure 6. Confusion matrix for video-based emotion recognition model



International Journal of Cognitive Informatics and Natural Intelligence

the PR curve is often used in information retrieval. Figure 9 shows the ROC curve with area under the curve as 0.7664. The blue curve represents the micro-average ROC curve, the red curve represents the macro-average ROC curve, and dotted black line which represents the Area Under the Curve (AUC) = 0.5 (in case of completely random classifier).

Comparison With Existing Models

The authors compare the proposed method with the other methods described in the literature, which are shown in Table 3. The approach outperforms the other existing approaches in the category of video-based, audio-based, and fusion-based recognition model with an accuracy of 62.17%, 73.01%, and 76.61% respectively.

Figure 7. Confusion matrix for fusion model

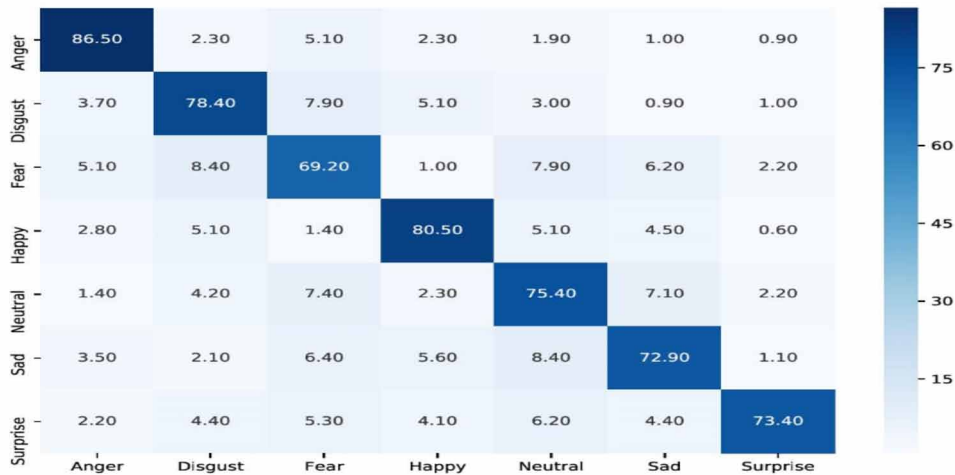


Figure 8. Validation accuracy (blue curve) and training accuracy (orange curve)

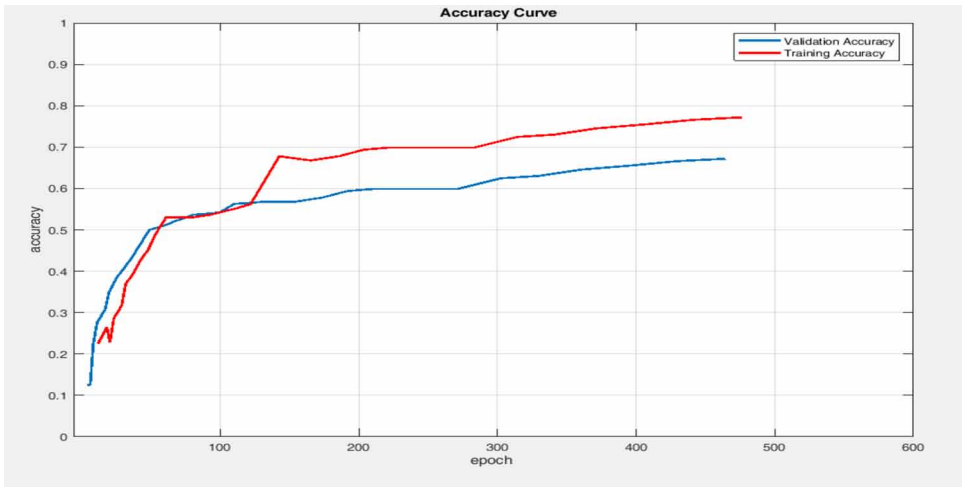


Figure 9. ROC curve representation for eNTERFACE'05 dataset

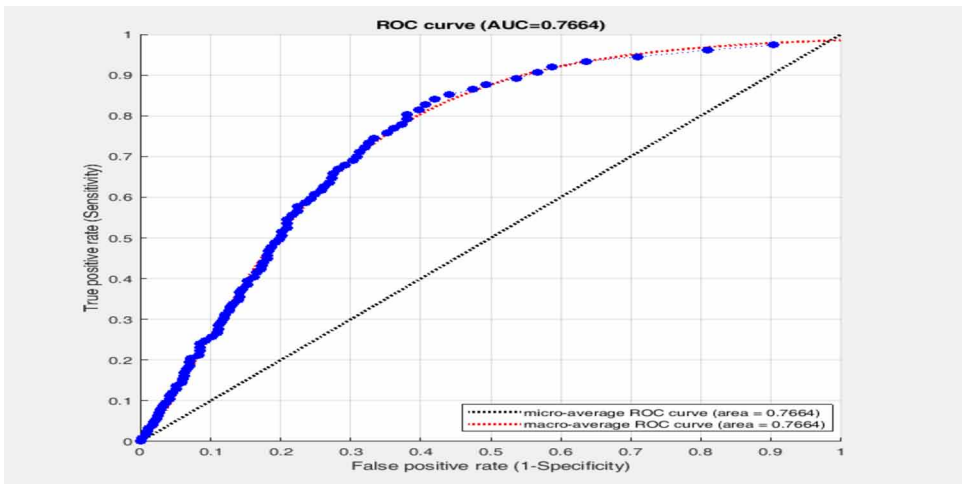


Table 3. Comparison of the proposed approach with other models

Emotion recognition model in audio		Emotion recognition model in video		Fusion-based emotion recognition model	
Authors	Accuracy (%)	Authors	Accuracy (%)	Authors	Accuracy (%)
M. Mansoorizadeh et al.	33	M. Mansoorizadeh et al.	37	D. Datcu et al.	56.3
Y. Wang et al.	38	D. Datcu et al.	37.7	Ouyang et al.	57.2
D.Datcu et al.	55.9	Huang et al.	52.3	Huang et al.	61.1
Huang et al.	56.4	R. Gajsek et al.	54.7	M. Paleari et al.	67
R. Gajsek et al.	62.9	Y. Wang et al.	58	M. Mansoorizadeh et al.	71
-	-	-	-	R. Gajsek et al.	71.3
-	-	-	-	Y. Wang et al.	76
Proposed approach	73.01	Proposed approach	62.17	Proposed approach	76.6

CONCLUSION

A fusion-based multi-modal emotion recognition system is proposed in this paper. The proposed model is an automatic audio and video-based emotion recognition system that classifies facial expressions among seven facial expressions classes, which are anger, happy, neutral, sad, disgust, fear, and surprise. The paper demonstrates a probabilistic audio-visual fusion model using SVM machine learning classifier which classifies the emotions into the classes mentioned above with better accuracy as compared to the previous works in the literature. In this paper, the authors explain the architecture of the models for audio, video, and the fusion model using CNN and LSTM. In the proposed model, the authors use AlexNet model in which the authors integrate a feature selection model to recognize the facial expressions with an accuracy of 62.17% in a video dataset. The authors also implement an audio-based emotion recognition model using Bidirectional LSTM-RNN, which recognizes the emotions with an accuracy of 73.01%. Lastly, the authors fuse both the models and achieve the highest accuracy of 76.61% on eINTERFACE'05 dataset.

The future work will have following directions – the authors will design more lightweight models to improve the accuracy and performance of the models. Also, the Human-Computer Interfaces models should be like -- they can understand humans' feedback, and acknowledge them appropriately. They must be capable of improving the efficiency and must allow engagement of prevailing interfaces. The authors will try to integrate the feature-level fusion into the

fusion network and investigate more types of fusion network models to improve the recognition performance further.

REFERENCES

- Avots, E., Sapin'ski, T., Bachmann, M., & Kamin'ska, D. (2019). Audiovisual emotion recognition in wild. *Machine Vision and Applications*, 30(5), 975–985. doi:10.1007/s00138-018-0960-9
- Ballester, P., & de Araújo, R. M. (2016). On the Performance of GoogLeNet and AlexNet Applied to Sketches. In AAAI (pp. 1124–1128). Academic Press.
- Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–10.
- Carrier, P.-L., Courville, A., Goodfellow, I. J., Mirza, M., & Bengio, Y. (2013). *FER-2013 face database*. Université de Montréal.
- Datcu, D., & Rothkrantz, L. J. (2011). Emotion recognition using bimodal data fusion. *Proceedings of the 12th International Conference on Computer Systems and Technologies*, 122–128.
- Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2013). Emotion recognition in the wild challenge 2013. *Proceedings of the 15th ACM on International conference on multimodal interaction*, 509–516. doi:10.1145/2522848.2531739
- El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587.
- Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9(Aug), 1871–1874.
- Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems & their Applications*, 13(4), 18–28.
- Huang, K.-C., Lin, H.-Y. S., Chan, J.-C., & Kuo, Y.-H. (2013). Learning collaborative decision-making parameters for multimodal emotion recognition. *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, 1–6.
- Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61, 610–628.
- Mansoorizadeh, M., & Charkari, N. M. (2010). Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, 49(2), 277–297. doi:10.1007/s11042-009-0344-2
- Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The enterface'05 audio-visual emotion database. *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, 8–8.
- Meghjani, M., Ferrie, F., & Dudek, G. (2009). Bimodal information analysis for emotion recognition. *Applications of Computer Vision (WACV), 2009 Workshop on*, 1–6.
- Melin, P., & Sánchez, D. (2018). Multi-objective optimization for modular granular neural networks applied to pattern recognition. *Information Sciences*, 460, 594–610.
- Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H., & Huang, D.-Y. (2017). Audio-visual emotion recognition using deep transfer learning and multiple temporal models. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 577–582. doi:10.1145/3136755.3143012
- Paleari, M., & Huet, B. (2008). Toward emotion indexing of multimedia excerpts. *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, 425–432. doi:10.1109/CBMI.2008.4564978
- Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Fifteenth annual conference of the international speech communication association*.
- Sánchez, D., Melin, P., & Castillo, O. (2017a). A grey wolf optimizer for modular granular neural networks for human recognition. *Computational Intelligence and Neuroscience*.

Sánchez, D., Melin, P., & Castillo, O. (2017b). Optimization of modular granular neural networks using a firefly algorithm for human recognition. *Engineering Applications of Artificial Intelligence*, 64, 172–186.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. doi:10.1016/j.neunet.2014.09.003 PMID:25462637

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., & Kessous, L. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. *Eighth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. *Tenth Annual Conference of the International Speech Communication Association*.

Sikka, K., Wu, T., Susskind, J., & Bartlett, M. (2012). Exploring bag of words architectures in the facial expression domain. *European Conference on Computer Vision*, 250–259.

Štruc, V., & Mihelc, F. (2010). Multi-modal emotion recognition using canonical correlations and acoustic features. *Pattern Recognition (ICPR), 2010 20th International Conference on*, 4133–4136.

Wang, Y., Guan, L., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14(3), 597–607. doi:10.1109/TMM.2012.2189550

Weninger, F., Bergmann, J., & Schuller, B. (2015). Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *Journal of Machine Learning Research*, 16(1), 547–551.

Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the ACM International Conference on Multimodal Interaction* (pp. 435–442). ACM.

Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., & Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273, 643–649.

Anand Handa is a research scholar in computer science and engineering department at Dr. APJ Abdul Kalam Technical University, Lucknow, UP, India. His areas of interest include image processing, computer vision, malware analysis, and machine learning. He has completed his Masters in Computer Science and Engineering from Rajiv Gandhi Proudhyogiki Vishwavidyalaya (RGPV), Bhopal, India. He pursued his bachelor's in computer science and engineering from the University Institute of Engineering and Technology, CSJM University, Kanpur. He has published various International and National papers at multiple conferences and in peer-reviewed journals.

Rashi Agarwal is the dean of Entrepreneurship and incubation center CSJM University Kanpur and head in the Department of Information Technology at the University Institute of Engineering and Technology, CSJM University, Kanpur, UP, India. She has done her Ph.D. in image processing from Dr. APJ Abdul Kalam Technical University, Lucknow, in 2006. She pursued her B.Tech in computer science and engineering from Harcourt Butler Technical Institute, Kanpur. Her area of interest includes image processing, data mining, and machine learning. She also authored and co-authored various books and multiple International and national conference and journal papers.

Narendra Kohli is a professor in the Department of Computer Science and Engineering at Harcourt Butler Technical University, Kanpur, UP, India. He has done his Ph.D. in image processing from the Indian Institute of Technology, Kanpur, and masters from Kiev University Kiev, Ukraine (Erstwhile USSR). His area of interest includes medical imaging, computer vision, and image processing. He is a senior member of various reputed organizations and agency including IEEE, IETE, etc. He has published many international and national research papers in peer-reviewed conferences and journals. He is also a reviewer and editor of various international journals.