Research Article

*International Journal of Advanced Computer Research, Vol 10(46)*
*ISSN (Print): 2249-7277 ISSN (Online): 2277-7970*
*http://dx.doi.org/10.19101/IJACR.2019.940117*

# Latest trends in emotion recognition methods: case study on emotiw challenge

**Huma Naz[*] and Sachin Ahuja**
Chitkara University Institute of Engineering and Technology, Chitkara University, India

Received: 25-September-2019; Revised: 12-December-2019; Accepted: 15-December-2019

## Abstract

***Emotion recognition is becoming increasingly very active field in research. In recent past, this research field has emerged as a milestone in software engineering, website customization, education, and gaming. Moreover, Emotion recognition models are used by more and more intelligent system to improve the multimodal interaction. Therefore, this paper demonstrates the recent literature on the emotion recognition methods presented at Emotion Recognition in the Wild (EmotiW) challenge. EmotiW is a grand challenge organized every year in ACM international conference on multimodal interaction. There has been number of methods presented every year at EmotiW for emotion analysis which are incorporated in this paper on the basis of emotion categorization in different areas. This work depicts a broad methodical analysis of EmotiW challenge for sentiments analysis which can help researchers, IT professionals and academia to find worthy technique for emotion grouping in several areas. It would also provide aid to select the most suitable technique for emotion recognition on the basis of their applications.***

## Keywords

***Emotion recognition, Audio-video emotion recognition, Emotion recognition methods, EmotiW case study, Emotion analysis.***

## 1.Introduction

Emotion recognition has acknowledged as an emerging research area which has gain the attention of both researchers and IT industry [1]. In recent years, online communities are generating a great amount of data and sentimental analysis of that data is very valuable for customers as well as business owners. Thus, Emotion Recognition is an exercise of identifying human reaction by visual and verbal expressions. The methods of Sentiment analysis work in video, audio and wild emotion recognition. The idea of emotion recognition is motivated by a simple question- Can we build a device which could perform the human-machine interaction? If that would be possible then machine will be able to sense our emotion and it can help in the field of modern healthcare, industry 4.0 and smart homes [1]. As everyone is aware of that Emotion identification has gain significance nowadays. Other than that Emotion solution market is worth USD 12.37 billion in 2018 and estimated to reach USD 91.67 billion by 2024 [2]. Therefore, this paper presents the review of the different emotion recognition methods which were presented every year in EmotiW challenge.

EmotiW is a series of Emotion recognition challenge organized by ACM international conference on multimodal interaction (ICMI) every year in different countries. EmotiW is a grand challenge which provides a platform to the researchers where they can compete on the basis of their proposed ideas. The goal of this challenge is to acquire diverse methods of emotions from audio video database which can be examined on real world conditions and data [3]. Conventionally, Emotion analysis is achieved on laboratory data, despite of that data does not mimic the real-world conditions. Thus, it is worthy to explore the methods that works 'in the wild' emotion recognition.

EmotiW challenge seeks participation from the researchers those aim to validate their work on real world conditions and intend to submit their ideas on the basis of the theme in the challenge [4]. Along with the dataset, baseline of the theme is also provided by the organizers every year. Baseline is the accuracy achieved by the organizers on the provided dataset, so the task is to achieve the better outcomes with the similar theme and dataset. There are several methods for emotion identification presented every year which uses advance technologies like

[*]Author for correspondence

Convolution neural network, deep neural network, deep learning, supervised and unsupervised learning approaches from real world data with a better accuracy[5] .Emotiw is organized by ICMI which is a preeminent international conference for Multimodal research on interdisciplinary human to computer and human to human interaction. It is conference which brings together international leading scientist, researchers and academics for better perspective and understanding of human and machine interaction. Recently sixth EmotiW challenge held at ACM ICMI 2018, Colorado (USA), which consists of Audio, Video and Group based engagement in the Wild.

The recent presented methods for emotion recognition have been reviewed in this paper through audio-video and wild data. Approach of this paper considering the issues related to emotion classification, Audio and video characteristics. The EmotiW organizers provide an open platform for the participators to evaluate their recognition systems. The task of emotion recognition is a tough task because it includes speech as well as video.

This paper is ordered as follows: Section 2 draws basic history of challenge; section 3 briefly represents proposed scenarios to be applied on emotion classification. Moving towards in the paper section 4 shows comparative analysis and concluded result. At last, section 6 present conclusions.

## 1.1History of EmotiW challenge

The following section shows the history and basic idea of EmotiW challenges held from 2013 to 2019, Details contained name of the challenge, topics of the challenge which were related to emotion only. All other important details related to EmotiW are discussed here. *Figure 1* shows hierarchy of EmotiW challenge organized from 2013 till now, Seven EmotiW challenges has been organized till 2019 by organizers and planning for many more in upcoming years.

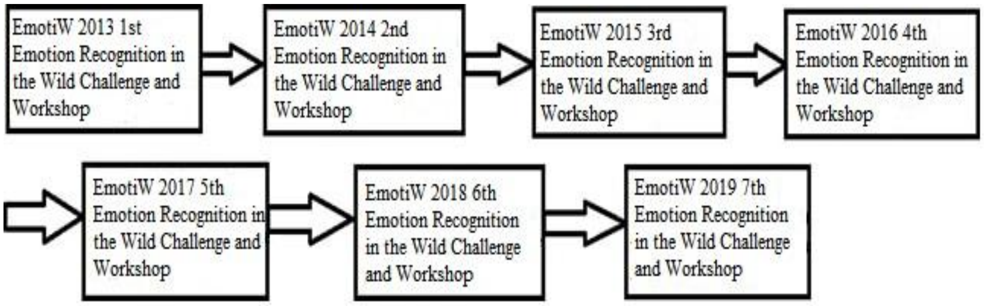
EmotiW 2013 1st Emotion Recognition in the Wild Challenge and Workshop
EmotiW 2014 2nd Emotion Recognition in the Wild Challenge and Workshop
EmotiW 2015 3rd Emotion Recognition in the Wild Challenge and Workshop
EmotiW 2016 4th Emotion Recognition in the Wild Challenge and Workshop
EmotiW 2017 5th Emotion Recognition in the Wild Challenge and Workshop
EmotiW 2018 6th Emotion Recognition in the Wild Challenge and Workshop
EmotiW 2019 7th Emotion Recognition in the Wild Challenge and Workshop

**Figure 1** Emotion classification challenges year wise

### 1.1.1EmotiW 2013

EmotiW 2013 is the first challenge on multimodal interaction organized by ACM ICMI at Sydney. It is a one-day event with theme of classifying the emotion of humans in the category of audio-video based emotion recognition (AVER) on real world conditions [6]. Along with the theme, organizers have also provided the acted facial expression in the Wild (AFEW) dataset to the contestants that was collected from movies showing close to real world conditions. According to the given baseline and dataset researchers presented their methods in Emotiw and few of them selected as winners conferring to their presented idea and accuracy. There were three papers selected as winners, 1st runner up and 2nd runner in EmotiW 2013.Therefore the presented ideas of winners are discussed here, which can help the scholars working in the area of multimodal interaction.

University of Montreal, Canada won the challenge with the accuracy of 41.03% which is favorably better than the baseline accuracy of 27.56% [3]. Kahou et al. [6] and her team presented an idea for the classification of emotion from short video clips. Their approach was to combine the four convolution network for different modalities like deep convolution network to capture the facial expressions within the video clips, deep belief network for analysis of audio information, deep auto encoder to classify the spatial characteristics for identifying the human action produced within the video and shallow network to focus on the extracted feature of the face in the given video. The best perform single model was convolution neural network which were trained using two large datasets i.e. Toronto Face Database and second self-made Google search images dataset. By combining the performance of best performed model the attained accuracy was achieved by the authors. *Table 1* describes the baseline accuracy and achieved accuracy by the winning teams of Emotiw 2013. *Figure 2* represents the comparison chart of accuracy achieved by different teams of EmotiW 2013 including winner, 1st runner up and 2$^{nd}$ runner up.

**Table 1** Accuracy achieved by the Winning teams

| Name of the team | Attained accuracy | Baseline accuracy |
|---|---|---|
| University of Montreal, Canada (UM) | 41.03%[6] | 27.56%[3] |
| University of California, San Diego (UCSD) | 37.08%[7] | 27.56%[3] |
| Institute of Computing Technology, Chinese Academy of Sciences, Beijing (ICT@CAS) | 35.85%[8] | 27.56%[3] |

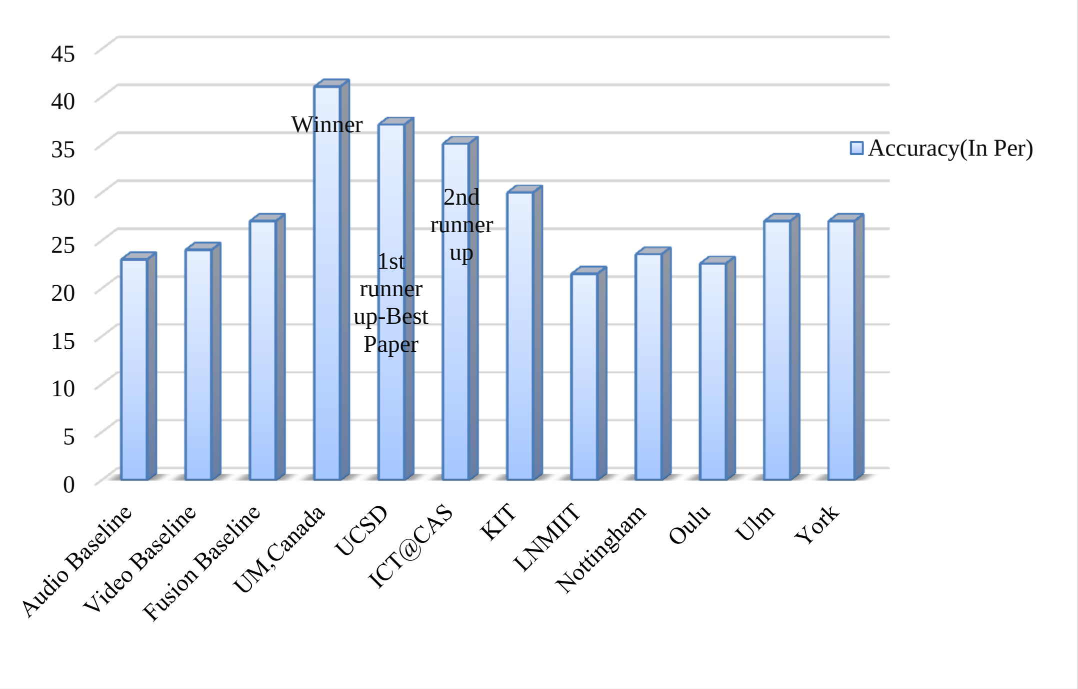

**Figure 2** Challenge result comparison of EmotiW 2013

### 1.1.2EmotiW 2014

After the successful completion of First EmotiW challenge organizers expand the series with second challenge which mimics real world conditions focused on affective sensing in unconstrained conditions. The challenge in the Wild 2014 held at ACMI ICMI in Istanbul [9]. Challengers asked researchers to propose diverse methods for AVER. The winning team of EmotiW 2014 was ICT, Beijing and their presented idea is discussed here.

ICT presented the method for automatic classification of human emotions from acted scenes in short video clips. They represented the video clips in three diverse types of image set model that were Gaussian distribution, linear subspace and covariance matrix. Kernel SVM, logistic regression and partial least squares were used for classification of these models. At final stage an optimal ensemble classifier is used from different audio video modalities. It has been conducted from different kernels for enhancing the performance and the final achieved accuracy is 50.4% with a considerable gain of 16.7 above the baseline accuracy of 33.7%. *Table 2* signifies the baseline accuracy and accuracy attained by the winning teams of Emotiw 2014.

*Figure 3* represents the chart of challenge result comparison which were achieved by different teams of EmotiW 2014 including winner, 1st runner up and 2nd runner up.

**Table 2** Accuracy achieved by the Winning teams

| Name of the team | Attained accuracy | Baseline accuracy |
|---|---|---|
| ICT@CAS | 50.4%[10] | 33.7%[9] |
| Beijing Normal University (BNU) | 47.17%[11] | 33.7%[9] |
| Hong Kong Polytechnic University (HKPU), Hong Kong | 45.21%[12] | 33.7%[9] |

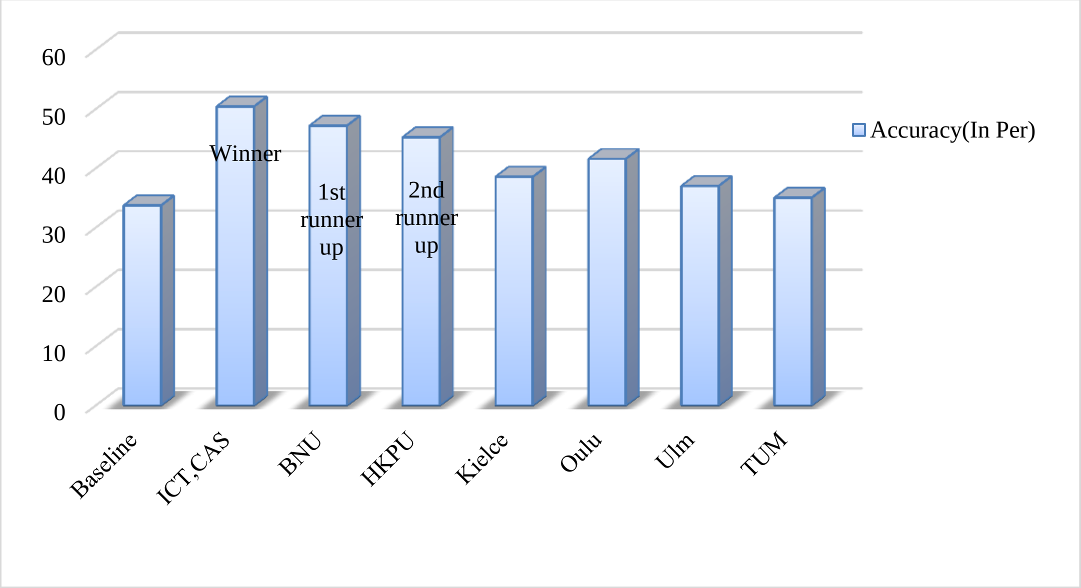

**Figure 3** Challenge result comparison of EmotiW 2014

### 1.1.3EmotiW 2015

The third Emotiw challenge divided into two sub challenges named as AVER and Image based static facial expression recognition (ISER). It was organized at ICMI settle in the year of 2015 and there were AFEW 5.0 and static facial expressions in the Wild (SFEW) 2.0 datasets which have been used in this challenge [13].

Dataset of the challenge contains differing natural movement of head pose, facial expressions, subjects from different ages and multiple subjects in a scene. The labelling of subjects has been done in the order of a name, character age, gender, pose and individual facial expressions. Therefore, competition has been done on two sub challenges and participants presented their ideas on both of the topics.

**AVER**

The task of the AVER challenge was to allocate a single label of emotion to the short clips from the six universal and Neutral emotion. The sub-challenge baseline was based on calculating LBPTOP descriptor and chi-square based SVM classifier. Val accuracy for the test case was 39.33% [13].

**ISER**

Facial and face part expressions were measured using Mix pictorial form structure for static Facial based emotion recognition and then SVM is trained using feature ensemble on vector. After training the model Baseline accuracy achieved on test case was 39.13%. We discuss here the outcome obtained by the winning team of EmotiW 2015 with the Baseline comparison. *Table 3* denotes the baseline accuracy and accuracy attained by the winning teams of Emotiw 2015 in ASER.

*Figure 4* denotes the comparison chart of challenge result achieved by winner, 1st runner up and 2nd runner up of EmotiW 2015(ASER).

*Table 4* denotes the baseline accuracy and accuracy attained by the winning teams of Emotiw 2015 in the subarea of ISER.

*Figure 5* denotes the chart of challenge result comparison attained by winner, 1st runner up and 2nd runner up of EmotiW 2015(ISER).

**Table 3** Accuracy achieved by the Winning teams of ASER

| Name of the team | Attained accuracy | Baseline accuracy |
|---|---|---|
| Intel Labs China (ILC), Beijing | 53.8%[14] | 39.33%[13] |
| Bogazici University (BI) Istanbul, Turkey | 53.6%[15] | 39.33%[13] |
| Ecole Polytechnique de Montréal, Canada | 49.9%[16] | 39.33%[13] |

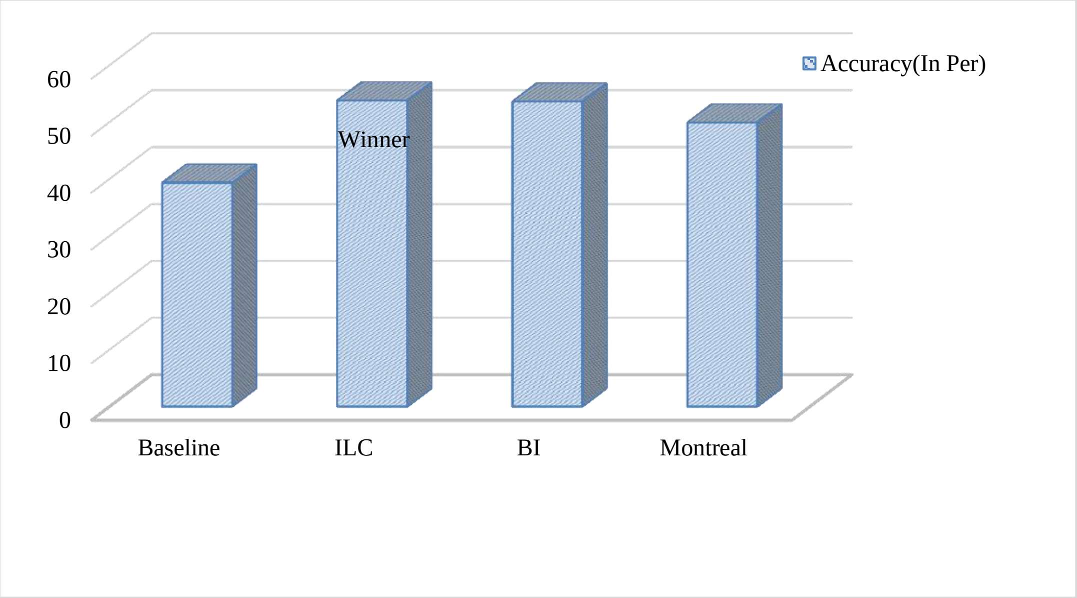

**Figure 4** Challenge result comparison of EmotiW 2015

**Table 4** Accuracy achieved by the Winning teams of ISER

| Name of the team | Attained accuracy | Baseline accuracy |
|---|---|---|
| Korea Advanced Institute of Science and Technology (KAIST) Daejeon, | 61.6%[17] | 39.13%[13] |
| Carnegie Mellon University (CMU), Forbes Ave Pittsburgh | 61.2%[18] | 39.13%[13] |
| Advanced Digital Sciences Center (ADSC) University of Illinois at Urbana-Champaign, Singapore | 53.8%[19] | 39.13%[13] |

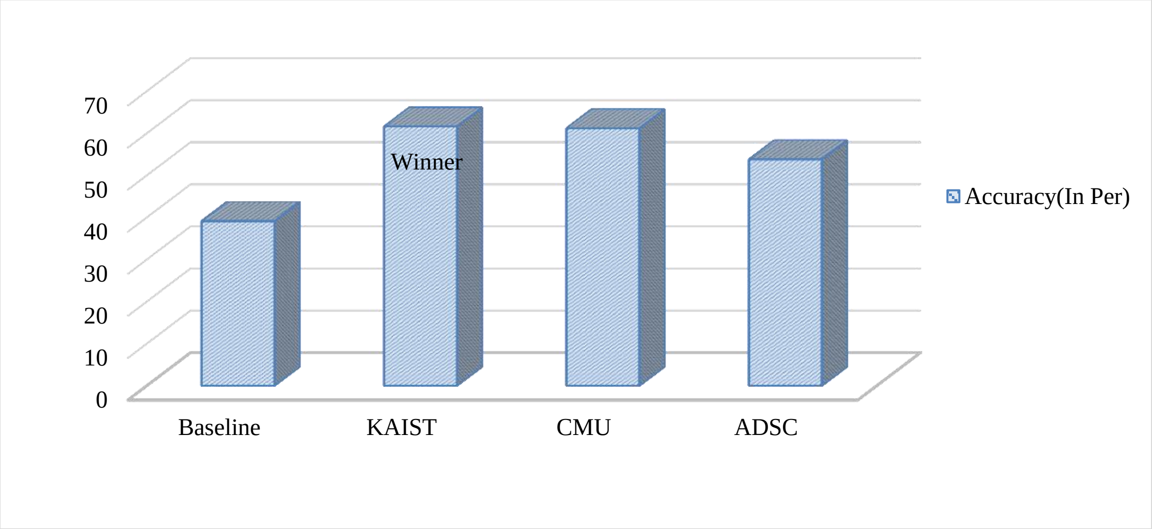

**Figure 5** Challenge result comparison of EmotiW 2015

### 1.1.4EmotiW 2016

**Video based emotion recognition sub-challenge (VBER)**

This series of challenge in continuation from 2013 to 2015 EmotiW. It is fourth challenge which was held at ACM ICMI in Tokyo and focuses on affective sensing in unconstrained conditions[20]. The major change from the earlier challenge the composition of TV reality data is compounded with the movie data in test data. Motivation here is that the TV data is rather more unstructured than Movie data, thus organizers challenge to present the methods which can classify the test data with one emotion label. Baseline accuracy is achieved as 40.27% and dataset used in this challenge was AFEW 6.0[21].

**Group level emotion recognition sub-challenge (GLER)**

GLER was the second sub challenge in EmotiW 2015, it contains the image of people in group. The task here is given is to predict the happiness intensity of a group on the scale 5.The dataset used in the sub challenge is Happie [22] dataset which collects the group images with the bottom up, top down and centrist features. Top down attributes here refer to the external factors like (scenes, neighbour), bottom up consist of the features like facial expression, face gestures etc. and centrist is the combination of both the approaches. Winner decided on the basis of the least root mean square error (RMSE) which is a proposed metric of the challenge. The RMSE for Test set was 1.30. *Table 5* signifies the baseline and winners achieved accuracy of Emotiw 2015 in the subarea of VBER.

*Figure 6* represents the chart of challenge result comparison attained by winner, 1st runner up and 2nd runner up of EmotiW 2015(VBER).

*Table 6* signifies the baseline accuracy in the form of RMSE and minimal RMSE attained by Emotiw 2016 winners in the subarea of GLER.

**Table 5** Accuracy achieved by the Winning teams of VBER

| Name of the team | Attained accuracy | Baseline accuracy |
|---|---|---|
| iQIYI Co. Ltd, Beijing, China | 59.02%[23] | 40.47%[20] |
| Boston University Dept. Computer Science Boston, MA, USA | 56.66%[24] | 40.47%[20] |
| ILC, Beijing, China | 57.84%[25] | 40.47%[20] |
| Southeast University (SU), Nanjing, China | 55.14%[26] | 40.47%[20] |

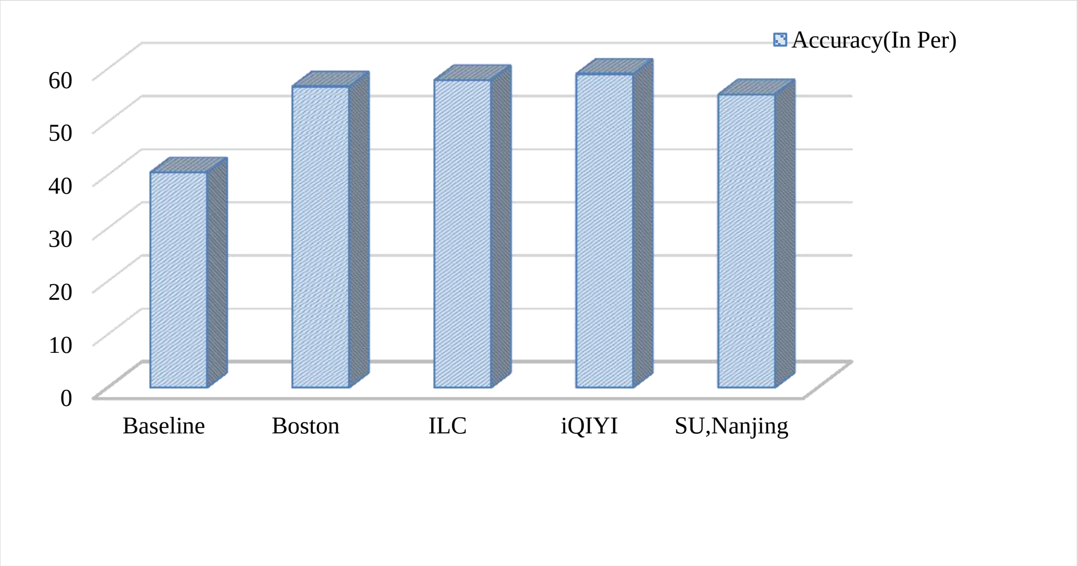

**Figure 6** Challenge result comparison of EmotiW 2016

**Table 6** Accuracy achieved by the Winning teams of GLER

| Name of the team | Rmse | Baseline (Rmse) |
|---|---|---|
| National University of Singapore, Singapore | 0.822[27] | 1.3[22] |
| University of Illinois at Urbana-Champaign, Singapore | 0.831[28] | 1.3[22] |
| BNU, Beijing, China | 0.836[29] | 1.3[22] |

### 1.1.5EmotiW 2017

The fifth EmotiW grand challenge held at ACM ICMI in Glasgow. Challenge was covering two main topics which were-

**Audio-video based emotion recognition (Movie + reality TV + Sitcom)**

This challenge is in continuation from 2013 to 2016 and contains short video clips. The exercise was to accredit a single emotion label from Positive, Neutral or Negative to the video clip from the six emotions and Neutral. The sub-challenge baseline accuracy was 38.81% [30].

**Group level emotion recognition challenge (positive/neutral/negative)**

Second main topic of the challenge was Group based emotion recognition, which taken the images from group effect dataset [31] to classify the group's perceived emotion into a single label. In today’s era of social networking people click selfie and upload these pictures over the internet. These group pictures can be of positive environment like convocations, marriages, party or negative events like funeral or neutral one. *Table 7, 8* signifies the baseline accuracy and accuracy attained of Emotiw 2017 winners in the subarea of VLER & GLER.

**Table 7** Accuracy achieved by the Winning teams of VLER

| Name of the team | Accuracy achieved | Baseline(accuracy) |
|---|---|---|
| NTechLab Moscow, Russia | 60.03%[32] | 38.81%[30] |
| ILC, Beijing, China | 60.34%[33] | 38.81%[30] |
| Orange Labs Cesson-Sevigne, France | 58.8%[34] | 38.81%[30] |

**Table 8** Accuracy achieved by the Winning teams of GLER

| Name of the team | Accuracy achieved | Baseline(accuracy) |
|---|---|---|
| Chinese Academy of Sciences, PR China | 83.09%[35] | 53.62%[31] |
| University of Delaware Newark<br>DE, USA | 80.61%[36] | 53.62%[31] |
| BNU, China | 79.78%[37] | 53.62%[31] |

### 1.1.6EmotiW 2018

It is 6th EmotiW challenge which was held in august 2018 at ICMI in Colorado. The challenge is divided in subparts which are Engagement in the Wild, Audio-video Emotion Recognition and Group-based Emotion Recognition. There were total 12 papers [38]–[48] accepted in this challenge out of the total number of papers submitted. *Figure 7* shows chart of winners result accuracy (in percentage) with respect to year, every year one or two team declared as winner by considering their methods and result accuracy.

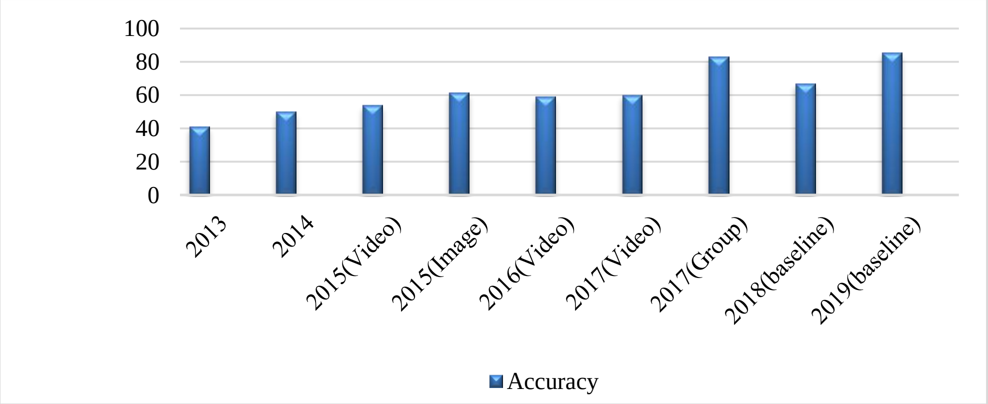

**Figure 7** EmotiW winners result accuracy (in per) with respect to year

### 1.1.7EmotiW 2019

EmotiW 2019 is seventh workshop held at ICMI china which mainly focused on transferring current research from lab based and small scaled environment to large scale and real-world corpus. There were broadly their topics which were covered in the challenge (AVER, Group-level Cohesion sub-challenge and Engagement prediction in the Wild (EW)). This workshop invited researchers to submit their methodologies on any of these topics. The main task of the challenge is to classify the engagement of the subject in the video and to predict the cohesion in the group images. Mean square error (MSE) on the training dataset is 0.10[49].

## 2.Literature review

This section describes about the diverse emotion recognition methods proposed by researchers on AVER and emotion recognition in wild. In this regard, review work of previous years presented methods in EmotiW challenge has been covered in this section. There are numbers of methods available through which emotion can be discriminated by applying diverse mechanism, which are explained here.

### 2.1Audio-video methods

AVER methods consider verbal and visual expression and researchers have used dissimilar procedures like neural network, data mining learning methods and ensemble of different model for precise result. Audio-video together gives better accuracy that's why both of the expressions included in recognition methods.

As the research area of emotion classification is on the boom, therefore Poria et al. [50] proposed a data set called Multimodal Emotion Lines Dataset (MELD), which was created for enhancing and extending the emotion lines because there is no any previously introduced dataset available which has more than one speaker for audio recognition. MELD contains more than 3000 utterance from 1433 dialogues which were taken from Friend TV series, it is comparatively better because it contains the multiparty conversation and the utterance in MELD is twice in comparison with the available datasets SEMAINE and IEMOCAP. It not only contains dialogues in textual form but also their corresponding audio and video. Therefore, multimodal multiparty conversational emotion recognition dataset called MELD is proposed to help researchers to get solid baseline results of their work.

Knyazev et al. [32] projected an ensemble model for emotion classification to get entry in emotion recognition challenge EmotiW 2017.Authors presented a hybrid model, which considers audio and spatial features from videos .CNN are used for extracting spatial features that are pretrained on large face recognition datasets. This model has been proposed because emotion recognition has number of applications in research, academia, industry and it is also very important of Artificial Intelligence. This work uses Ensemble models and also uses SVM technique, Video features are extracted using CNNs and Audio features are collected the visual models with an additional modality. This work shows that leveraging large dataset can show better results.

Wankhade and Kukade [51] suggested a method which can change the emotional state through altering speech signal .It takes words and letters by different emotional situations. Emotion classification based on audio is a challenging task which author handled by presented experimental study on emotion recognition through speech. The emotions categories for this experimental study are neutral, anger, joy and sadness.

Kahou et al. [6] proposed method of emotion recognition in audio and video by Faces and Convolution Network and used SVM techniques for result. SVM is discriminative classifier which gives more accuracy separated by hyper plane; it is a supervised learning algorithm which is used for classification and regression. SVM gives better accuracy and prediction quality in comparison with other classification algorithm. The accuracy differs with different datasets but it also has a drawback which is smaller prediction error[25] and Vapnik developed the algorithm of SVM for better classification [52].Authors have proposed a method in which convolution neural network and audio features are combined together are through SVM it generates Emotions and Activity recognition, bag of mouth are also applied and combinable creates output and used large dataset to train the model for emotion recognition.

Chen et al. [53] suggested a method of emotion recognition for the categorization of YouTube videos into six emotion categories (i.e., happiness, anger, disgust, fear, sadness, and surprise). Because numbers of online videos are increasing tremendously and YouTube videos watching are very customary action among online users. Millions of data is available on video sharing websites, So it is very difficult to find required data from these websites, That's why this approach improvise the searching of content and the effectiveness of video recommendations. This approach used supervised and unsupervised learning methodology for emotion classification and for better results an ensemble model was used for combining results. As proposed this approach uses supervised and unsupervised learning methods initially and for improving result ensemble model was subsequently used according to the emotion category.

## 2.2Wild methods review

Wild methods include indoor outdoor scenes, different changing lighting conditions, external noise, pose changing issues and motion blur, as well as uncontrolled scenarios. Hence Emotion recognition "in the wild" is a difficult and demanding task and researchers have been used different scenarios for recognizing emotions to controlling an open condition.

Dhall et al.[13] Proposed method which describes base line method, challenge protocol and data, b for third EmotiW challenge which was held in 2015. It consists of static image based facial expression and an audio-video based emotion recognition. Traditionally lab controlled environment data is used for emotion recognition in a traditional way, but emotion discrimination in wild is a challenging task because of the diversity of head scenes and background noise, so the researches have shown the model to identify the state line, postures in wild, but efficiency is still a major and open problem in the wild.

Tan et al. [35] suggested a technique for audio video based emotion identification to get ingress in EmotiW challenge 'in the wild' 2018.Challengers assigned participants a single emotion label from the six emotions (Anger, Disgust, Fear, Happiness, Sad and Surprise) to the video clip. To fulfil the criteria multimodal emotion recognition system takes into account audio and video and also text information. Temporal and non-temporal classifiers are used to obtain the best classification results of uni-modal emotions. This model extracted features from diverse modalities and train it through passing them into BS-Fusion. They have participated in EmotiW 2018 challenge and get a result: 60.34% on the testing dataset that is promising for participation.

Ng et al. [19] proposed an idea of adding deep supervision in emotion recognition, It is recent and great idea to deal with the problem of emotion grouping, because the EmotiW challenge has been successfully organized by six times, emotion recognition in audio-video and 'in the wild' has always been one of the main task in EmotiW challenge. The Proposed method uses single emotion label from the category of six emotions and neutral emotion.

Afshar and Ali Salah[54]proposed method for face and landmark detection, they introduce a hybrid method by combination of two different methods for making the system more robust and reliable. Supervised Descent Method (SDM) had been used by researchers because this method is fast and can detect facial expression precisely and then alignment is performed by using generalized Procrustes analysis on detected faces for removing translation.

Hossain and Muhammad[55] proposed an emotion recognition technique processing through deep learning approach with emotional Big Data. In the proposed system, emotion grouping is done through deep learning along with Big Data which comprises of speech and video. Speech signal and video signals were processed separately. For speech signal, an image was produced through processing speech signal into the frequency domain. By following this process, a Mel spectrogram was acquired which then fed to a CNN. For video signals, Video segment is divided into representative frames and which is fed to the CNN. Then output of speech signal and video signal are fused using two consecutive Extreme Learning Machine (ELM). Then final output was given to an SVM for emotion classification. There were two audio-visual emotional databases were used for the evaluation of this proposed system, one of which is Big Data. The results of experiments assure the efficiency of the proposed system.

# 3.Discussion and analysis

This section elaborates the block diagram of the work done in this paper and short description of the terms used throughout the paper for better understanding. Furthermore, this section describes the table of methods used in Emotion recognition with their merits demerits and achieved accuracy. *Figure 8* represents the step by step process of the study and analysis done in this case study of EmotiW challenge.

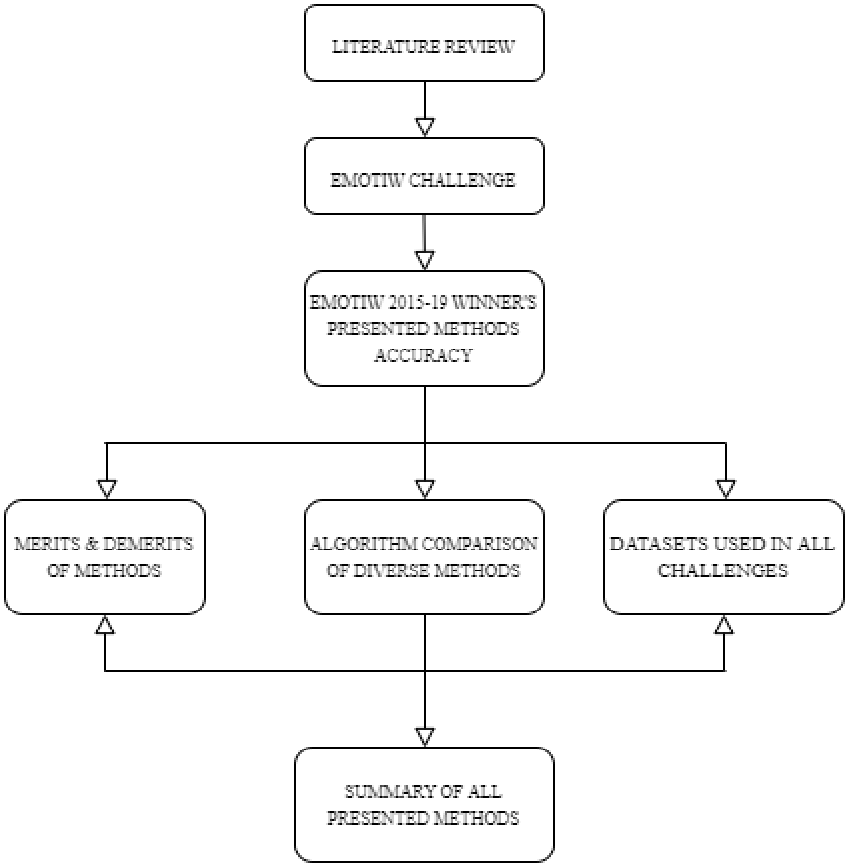
LITERATURE REVIEW
EMOTIW CHALLENGE
EMOTIW 2015-19 WINNER'S PRESENTED METHODS ACCURACY
MERITS & DEMERITS OF METHODS
ALGORITHM COMPARISON OF DIVERSE METHODS
DATASETS USED IN ALL CHALLENGES
SUMMARY OF ALL PRESENTED METHODS

**Figure 8** Flow chart of the study and analysis

### 3.1Convolutional neural network (CNN)

CNN is a deep neural network class which is mostly used to analyze the visual images in deep learning [56]. The images as input, assign the weight or bias to different objects in the images, so that it can be differentiated with each other. CNN need lesser preprocessing of input in comparison with other available methods, it has capability to self-train the data.

### 3.2Deep neural network (DNN)

Deep Neural Network can be called as neural network with a certain convolution or it is a class of multilayer feed-forward perceptron model which simplifies the properties of artificial neural networks using stochastic gradient descent back-propagation. DNN is a simply a neural network with more than two layers and consist of some mathematical modeling for complex data processing. DNN contain multiple layers emulating nodes and neurons, directed in uni-direction (one-way connection). Each node is connected to the next node in a single way connection[57] and trains a local copy of data using global model parameters. Further, it uses multiple threads to process the global model parameters and apply the averaging for contribution access across the whole network.

### 3.3Support vector machine (SVM)

SVM is a non-parametric supervised algorithm which is applied to classify the data. In recent years classification algorithm has gain attention of the researchers due to its classification accuracy on geospatial data. It can be treated as regression algorithm and first proposed by the Vapnik and Learner (1963) and then N. Barakat et. al[58] updates it with the theory of statistical learning. It can be used for regression problems and SVM use hyper plane to classify the data to its related class.

### 3.4Extreme learning machines (ELM)

ELM is a layered feed forward neural network which can be used for both classification and regression. The need of ELM arises due to the slow learning speed of neural networks than required and it could be a reason of its slow gradient based learning algorithm or number of iterations in each learning step. ELM only have one abstraction layer that's why it is not feasible for deep learnings but faster than the existing neural network. ELM is tend to provide the best classification with very fast learning speed and optimum results[59].

### 3.5Supervised descent method (SDM)

SDM is a method of weighted gradient learning which can be used for averaging the conflicts between the diverse gradient methods. It is a local supervised method and used for minimization the Non-linear type of least square methods. Moreover SDM is more robust, fast and fast supervised method for non-linear optimization of gradient methods[60]. *Table 9* shows merits and demerits of the methods used by different research in their EmotiW proposed techniques.

*Table 10* represents the comparison of the diverse techniques in terms of the accuracy which were presented in challenge.

**Table 9** Advantages and disadvantages of the some of the methods used in this area

| S.No | Reference | Method(S) | Advantages | Disadvantages |
|---|---|---|---|---|
| 1 | [6] | Deep Convolutional Network #1 | Using of large scale google images avoided the over fitting | Cost increased due to use of large-scale database |
| 2 | [11] | Hierarchal Fusion Classifier | Diverse features were trained which have discriminative ability for emotion identification. | Feature level fusion could be explored for better results |
| 3 | [12] | Multiple Kernel Learning with SVM | Novel feature descriptor has been proposed for facial expression application named as Histogram of Oriented Gradients from three Orthogonal Planes (HOG-TOP) | High Computation cost for larger datasets |
| 4 | [10] | Optimal fusion of logistic regression kernel SVM and partial least squares | Multiple Riemannian manifold with diverse ensemble classifier are examined for prominent result | Difficult categories of emotion identification can cost the downgrading in result |
| 5 | [7] | Multiple Kernel leaning with paralinguistic audio feature using SVM | Used a novel combination of Ramanan's deformable parts model and supervised descent Method for face identification tracking | Computation cost is high for feature extraction and grid search |
| 6 | [15] | Least square based classifier with multi-level weighted fusion | Multi-level fusion with single alignment is more accurate | Multi-level fusion with different audio modalities can cost downgrading in outcome. |

**Table 10** Methods result comparison

| S.no. | Reference | Method (S) | Results |
|---|---|---|---|
| 1 | [6] | Deep Convolutional Network #1 | 41.03% |
| 2 | [11] | Hierarchal Fusion Classifier | 47.17% |
| 3 | [12] | Multiple Kernel Learning with SVM | 45.21% |
| 4 | [10] | Optimal fusion of logistic regression kernel SVM and partial least squares | 50.4% |
| 5 | [7] | Multiple Kernel leaning with paralinguistic audio feature using SVM | 37.08% |
| 6 | [15] | Least square based classifier with multi-level weighted fusion | 53.6% |
| 7 | [16] | Hybrid CNN-Recurrent Neural Network (RNN) | 49.9% |
| 8 | [17] | Multiple deep CNNs | 61.6% |
| 9 | [18] | deep CNNs with multiple improved framework | 61.2% |
| 10 | [19] | Transfer learning approach for deep CNN | 53.8% |
| 11 | [23] | Enhanced HoLo CNN | 50.2% |
| 12 | [24] | Hybrid RNN and 3d CNN | 56.66% |
| 13 | [26] | Multi-cue fusion emotion recognition framework | 55.14% |
| 14 | [27] | CNN and RNN | 0.822(RMSE) |

| S.no. | Reference | Method (S) | Results |
|---|---|---|---|
| 15 | [28] | Partial Least Squares regression | 0.831(RMSE) |
| 16 | [29] | CNN-Long-Short Term Memory Model | 0.836(RMSE) |

## 3.6Results & comparative analysis

*Table 11* represents the comparative analysis of methods/technique/tools, advantages and disadvantages which were presented at EmotiW challenge from 2013-2018.

**Table 11** Comparison of various Emotion recognition methods of EmotiW

| Authors | Year | Method/ approach/ summary | Tools/ techniques/algorithm | Merits | Demerits | Method |
|---|---|---|---|---|---|---|
| Kahou et al. [6] | 2013 | deep auto encoder,<br>deep belief network,<br>deep convolutional neural network | Support vector machine, CNN, aggregate models | It overcomes the problems of over fitting.<br>Uses both SVM and MLP aggregator.<br>Low complexity<br>Result is of highly constrained. | Result can be improvised as SVM does not as much accuracy comparatively. | Audio-Video |
| Dhall et al. [13] | 2015 | AFEW 5.0<br>SFEW 2.0 | Deep learning, multiple kernel learning | It contains both audio video and Static image emotion recognition/.<br>It performs emotion recognition on wild data in comparison of lab-controlled data. | AFEW a dataset which is being used for training, validation and testing and it is available as open source dataset.<br>testing will be done on real data in wild, for that it may not provide that accurate result | Wild |
| Afshar and Ali Salah [54] | 2015 | Supervised descent method,<br>Supervised learning | Generalized Procrustes analysis, Discriminative Response Map Fitting, supervised descent method, | It combines two methods as for making system reliable and robust.<br>SDM method which is used in this approach is very fast for execution.<br>It processing speed for finding face and their corresponding Landmark is too fast. | DRMF Methods implementation speed is slow but it works well in the wild. | Wild |

| Authors | Year | Method/ approach/ summary | Tools/ techniques/algorithm | Merits | Demerits | Method |
|---|---|---|---|---|---|---|
| Chen et al.[53] | 2017 | Machine learning, Emotion dictionary approach and ensemble model | Supervised machine learning, Unsupervised machine learning | Provides better results by the combination of two learning technique i.e. Supervised & unsupervised. | Accuracy rate of unsupervised classification Using the emotion dictionary approach were Unsatisfactory. | Audio-video |
| Knyazev et al. [32] | 2017 | Deep convolution neural networks, proprietary state of the art, face recognition networks | Dlib face detector, VGG-Face[15],<br>FR-Net-A,<br>FR-Net-B,<br>FR-Net-C | Ensemble of Models has been applied that Results in better accuracy of emotion classification than the Winners of 2016 EmotiW challenge.<br>Audio features also complement the result in terms of better accuracy | Leveraging large data is trained for achieving better results which can be time consuming and for small data it may not provide that better results. | Audio-Video |
| Hu et al. [33] | 2017 | Deep CNNs, Feature extraction, Ensemble model | Supervised learning | As this method is using deep CNNs, so it exhibits the Property of deep supervision.<br>It provides the supervision to diverse layers jointly rather providing only to the top layers | It is only applicable for audio video but not for text information | Wild |
| Yan et al. [26] | 2018 | Deep neural networks,Beam Search Fusion | Transfer learning, Fusion Methods | It extracts more discriminative Features.<br>Provides emotion classification for audio, video and text information (multimodal classification) | Video description has not been used since<br>Emotion is only related to video content, so it can provide<br>Better Results. | Wild |

| Authors | Year | Method/ approach/ summary | Tools/ techniques/algorithm | Merits | Demerits | Method |
|---|---|---|---|---|---|---|
| Hossain and Muhammad [55] | 2019 | CNNs and extreme learning machines | Deep learning approach, Big Emotional data, Mel spectrogram, Support Vector machines | Provide hybrid results of CNNs and the ELMs. Used Big Emotional data. Mel spectrogram are used for audio and video feature | Space complexity, Big data takes huge space and training for model | Wild |

### 3.6.1Results

Despite the long history and popularity of EmotiW challenge, there is no comprehensive review is existing which can provide the insights about the challenge. Therefore, on the basis of existing research works, a comparative analysis is provided in *Table 9-11* which incorporates a comparison of various proposed methods for emotion classification for the achievement of high accuracy rate in Algorithms, Applications, and Techniques.

- The focus is to provide the insight of the challenge, details of the participant, their presented methods and higher achieved accuracy. It will be very helpful in finding appropriate procedure according to the type of data.
- It includes the EmotiW challenge and participants presented methods with increasing accuracy year by year using modern machine learning methods. It can be beneficial for the research family of Emotiw to find the suitable technique.
- This review table lists all the Emotion classification methods along with the associated techniques which seems to be very beneficial for future researchers to select the best suitable tools. Key aspects are compared between the proposed methods and it has been observed that the fusion of diverse algorithm provide more accurate results.

## 4.Research challenges

1. More effective fusion category of difficulty situation can be explored for better outcomes[10].
2. There are more nuanced emotion available which are difficult to identify correct label and it can cost to the performance of the model[19].
3. Diverse good objective function can be evaluated for training deep CNN in order to get more accuracy and diversified outcomes[17].
4. Alternative modalities of audio can be explored for training classifier to exploit better ways which have not been elaborated yet [15].
5. The future work in this area can be done with more complex top down scene related features, it could improve the performance of the model.
6. Some of the methods are evaluated on the provided datasets only, the accuracy can vary with more complex and real-life scenario.
7. Offline mitigation can be explored for further improvement in the accuracy or more accurate emotion prediction.

## 5.Conclusion

Emotion recognition from speech and videos helps to understand the gestures and emotions of a person more accurately and assure naturalness in the performance of existing speech and visual systems. In this paper, we have reviewed and presented an ample amount of significant work which has been done in the recent past years related to methods used by researchers for emotion recognition with considerable accuracy. Since EmotiW 1st challenge 2013 accuracy is increasing in every Challenge by using the latest methods of machine learning as represented in *Table 9-11*. This paper presents those methods with a technique, tool year wise. These days the latest machine learning methods like deep learning, CNN, DNN, neural networks are in trend for better and accurate results. Our work shows the latest methods used for improvisation of accuracy which can help young researchers for finding appropriate methods for their emotion grouping work. Moreover, it shows a comprehensive review of EmotiW emotion recognition methods and develops a categorization scheme to analyze the existing techniques and their best use in diverse Areas.

## Acknowledgment

None.

## Conflicts of interest

The authors have no conflicts of interest to declare.

## References

[1] Zhao M, Adib F, Katabi D. Emotion recognition using wireless signals. In proceedings of the 22nd annual international conference on mobile computing and networking 2016 (pp. 95-108). ACM.

[2] https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market. Accessed 20 August 2019.

[3] Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T. Emotion recognition in the wild challenge 2013. In proceedings of the ACM on international conference on multimodal interaction 2013 (pp. 509-16). ACM.

[4] Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, et al. Depression, mood, and emotion recognition workshop and challenge. In proceedings of the international workshop on audio/visual emotion challenge 2016 (pp. 3-10). ACM.

[5] Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology. 2018; 21(1):93-120.

[6] Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülçehre Ç, Memisevic R, et al. Combining modality specific deep neural networks for emotion recognition in video. In proceedings of the ACM on international conference on multimodal interaction 2013 (pp. 543-50).

[7] Sikka K, Dykstra K, Sathyanarayana S, Littlewort G, Bartlett M. Multiple kernel learning for emotion recognition in the wild. In proceedings of the ACM on international conference on multimodal interaction 2013 (pp. 517-24). ACM.

[8] Liu M, Wang R, Huang Z, Shan S, Chen X. Partial least squares regression on grassmannian manifold for emotion recognition. In proceedings of the ACM on international conference on multimodal interaction 2013 (pp. 525-30). ACM.

[9] Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T. Emotion recognition in the wild challenge 2014: baseline, data and protocol. In proceedings of the international conference on multimodal interaction 2014 (pp. 461-6). ACM.

[10] Liu M, Wang R, Li S, Shan S, Huang Z, Chen X. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In proceedings of the international conference on multimodal interaction 2014 (pp. 494-501). ACM.

[11] Sun B, Li L, Zuo T, Chen Y, Zhou G, Wu X. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In proceedings of the international conference on multimodal interaction 2014 (pp. 481-6). ACM.

[12] Chen J, Chen Z, Chi Z, Fu H. Emotion recognition in the wild with feature fusion and multiple kernel learning. In proceedings of the international conference on multimodal interaction 2014 (pp. 508-13). ACM.

[13] Dhall A, Ramana Murthy OV, Goecke R, Joshi J, Gedeon T. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In proceedings of international conference on multimodal interaction 2015 (pp. 423-6). ACM.

[14] Yao A, Shao J, Ma N, Chen Y. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In proceedings of the ACM on international conference on multimodal interaction 2015 (pp. 451-8). ACM.

[15] Kaya H, Gürpinar F, Afshar S, Salah AA. Contrasting and combining least squares based learners for emotion recognition in the wild. In proceedings of the ACM on international conference on multimodal interaction 2015 (pp. 459-66). ACM.

[16] Ebrahimi Kahou S, Michalski V, Konda K, Memisevic R, Pal C. Recurrent neural networks for emotion recognition in video. In proceedings of the ACM on international conference on multimodal interaction 2015 (pp. 467-74). ACM.

[17] Kim BK, Lee H, Roh J, Lee SY. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In proceedings of the international conference on multimodal interaction 2015 (pp. 427-34). ACM.

[18] Yu Z, Zhang C. Image based static facial expression recognition with multiple deep network learning. In proceedings of the international conference on multimodal interaction 2015 (pp. 435-42). ACM.

[19] Ng HW, Nguyen VD, Vonikakis V, Winkler S. Deep learning for emotion recognition on small datasets using transfer learning. In proceedings of the international conference on multimodal interaction 2015 (pp. 443-9). ACM.

[20] Dhall A, Goecke R, Joshi J, Hoey J, Gedeon T. Video and group-level emotion recognition challenges. In proceedings of the international conference on multimodal interaction 2016 (pp. 427-32). ACM.

[21] Dhall A, Goecke R, Lucey S, Gedeon T. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia. 2012; 19(3):34-41.

[22] Dhall A, Goecke R, Gedeon T. Automatic group happiness intensity analysis. IEEE Transactions on Affective Computing. 2015; 6(1):13-26.

[23] Yao A, Cai D, Hu P, Wang S, Sha L, Chen Y. Holonet: towards robust emotion recognition in the wild. In proceedings of the international conference on multimodal interaction 2016 (pp. 472-8). ACM.

[24] Fan Y, Lu X, Li D, Liu Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In proceedings of the international conference on multimodal interaction 2016 (pp. 445-50). ACM.

[25] Bargal SA, Barsoum E, Ferrer CC, Zhang C. Emotion recognition in the wild from videos using images. In proceedings of the international conference on multimodal interaction 2016 (pp. 433-6). ACM.

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

[27] Li J, Roy S, Feng J, Sim T. Happiness level prediction with sequential inputs via multiple regressions. In proceedings of the international conference on multimodal interaction 2016 (pp. 487-93). ACM.

[28] Vonikakis V, Yazici Y, Nguyen VD, Winkler S. Group happiness assessment using geometric features and dataset balancing. In proceedings of the international conference on multimodal interaction 2016 (pp. 479-86). ACM.

[29] Sun B, Wei Q, Li L, Xu Q, He J, Yu L. LSTM for dynamic emotion and group emotion recognition in the wild. In proceedings of the international conference on multimodal interaction 2016 (pp. 451-7). ACM.

[30] Dhall A, Goecke R, Ghosh S, Joshi J, Hoey J, Gedeon T. From individual to group-level emotion recognition: Emotiw 5.0. In proceedings of the international conference on multimodal interaction 2017 (pp. 524-8). ACM.

[31] Dhall A, Joshi J, Sikka K, Goecke R, Sebe N. The more the merrier: analysing the affect of a group of people in images. In international conference and workshops on automatic face and gesture recognition (FG) 2015 (pp. 1-8). IEEE.

[32] Knyazev B, Shvetsov R, Efremova N, Kuharenko A. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598. 2017.

[33] Hu P, Cai D, Wang S, Yao A, Chen Y. Learning supervised scoring ensemble for emotion recognition in the wild. In proceedings of the international conference on multimodal interaction 2017 (pp. 553-60). ACM.

[34] Vielzeuf V, Pateux S, Jurie F. Temporal multimodal fusion for video emotion classification in the wild. In proceedings of the international conference on multimodal interaction 2017 (pp. 569-76). ACM.

[35] Tan L, Zhang K, Wang K, Zeng X, Peng X, Qiao Y. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In proceedings of the international conference on multimodal interaction 2017 (pp. 549-52). ACM.

[36] Guo X, Polanía LF, Barner KE. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In proceedings of the international conference on multimodal interaction 2017 (pp. 603-8). ACM.

[37] Wei Q, Zhao Y, Xu Q, Li L, He J, Yu L, et al. A new deep-learning framework for group emotion recognition. In proceedings of the international conference on multimodal interaction 2017 (pp. 587-92). ACM.

[38] Yang J, Wang K, Peng X, Qiao Y. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In proceedings of the on international conference on multimodal interaction 2018 (pp. 594-8). ACM.

[39] Niu X, Han H, Zeng J, Sun X, Shan S, Huang Y, et al. Automatic engagement prediction with GAP feature. In proceedings of the on international conference on multimodal interaction 2018 (pp. 599-603). ACM.

[40] Vielzeuf V, Kervadec C, Pateux S, Lechervy A, Jurie F. An occam's razor view on learning audiovisual emotion recognition with small training sets. In proceedings of the international conference on multimodal interaction 2018 (pp. 589-93). ACM.

[41] Thomas C, Nair N, Jayagopi DB. Predicting engagement intensity in the wild using temporal convolutional network. In proceedings of the international conference on multimodal interaction 2018 (pp. 604-10). ACM.

[42] Chang C, Zhang C, Chen L, Liu Y. An ensemble model using face and body tracking for engagement detection. In proceedings of the international conference on multimodal interaction 2018 (pp. 616-22). ACM.

[43] Guo X, Zhu B, Polanía LF, Boncelet C, Barner KE. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In proceedings of the international conference on multimodal interaction 2018 (pp. 635-9). ACM.

[44] Wang K, Zeng X, Yang J, Meng D, Zhang K, Peng X, et al. Cascade attention networks for group emotion recognition with face, body and image cues. In proceedings of the international conference on multimodal interaction 2018 (pp. 640-5). ACM.

[45] Khan AS, Li Z, Cai J, Meng Z, O'Reilly J, Tong Y. Group-level emotion recognition using deep models with a four-stream hybrid network. In proceedings of the international conference on multimodal interaction 2018 (pp. 623-9). ACM.

[46] Gupta A, Agrawal D, Chauhan H, Dolz J, Pedersoli M. An attention model for group-level emotion recognition. In proceedings of the international conference on multimodal interaction 2018 (pp. 611-5). ACM.

[47] Liu C, Tang T, Lv K, Wang M. Multi-feature based emotion recognition for video clips. In proceedings of the on international conference on multimodal interaction 2018 (pp. 630-4). ACM.

[48] Fan Y, Lam JC, Li VO. Video-based emotion recognition using deeply-supervised neural networks. In proceedings of the international conference on multimodal interaction 2018 (pp. 584-8). ACM.

[49] Ghosh S, Dhall A, Sebe N, Gedeon T. Predicting group cohesiveness in images. In international joint conference on neural networks 2019 (pp. 1-8). IEEE.

[50] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. Meld: a multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508. 2018.

[51] Wankhade VA, Kukade RV. Categorization and analysis of emotion from speech signals. Themed Section: Engineering and Technology. 2018; 4(7):395-8.

[52] Zhang LM. Genetic deep neural networks using different activation functions for financial data mining. In international conference on big data 2015 (pp. 2849-51). IEEE.

[53] Chen YL, Chang CL, Yeh CS. Emotion classification of youtube videos. Decision Support Systems. 2017; 101:40-50.

[54] Afshar S, Ali Salah A. Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding. In proceedings of the conference on computer vision and pattern recognition workshops 2016 (pp. 1517-25).

[55] Hossain MS, Muhammad G. Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion. 2019; 49:69-78.

[56] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. 2014.

[57] Huang KY, Wu CH, Hong QB, Su MH, Chen YH. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In international conference on acoustics, speech and signal processing 2019 (pp. 5866-70). IEEE.

[58] Barakat N, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Transactions on Information Technology in Biomedicine. 2010; 14(4):1114-20.

[59] Huang GB, Zhu QY, Siew CK. Extreme learning machine: a new learning scheme of feedforward neural networks. Neural Networks. 2004; 2:985-90.

[60] Xiong X, De la Torre F. Global supervised descent method. In proceedings of the conference on computer vision and pattern recognition 2015 (pp. 2664-73).

**Huma Naz** is currently pursuing her Masters in Engineering at Chitkara University in the Department of Computer Science and Engineering. Her research interests include Data Mining, Machine Learning and Developing Prediction Models for various applications.

Email: huma.naz@chitkara.edu.in

**Dr. Sachin Ahuja** is serving in Chitkara University as Director Research. He holds a PhD in Data mining. His research interests include educational data mining. Specifically, under data mining he is interested in predictions, designing of survey questionnaires, measuring and comparing the academic performance of students and comparison of traditional teaching with flipped and blended learning models. Apart from data mining, his teaching interests include Big Data, Relational Database and Procedural Languages. In addition to this, he is heading the Office of Patent facilitation where he has facilitated inventors from Chitkara University in filing patents by guiding them in licensing and consultancy.

Email: Sachin.ahuja@chitkara.edu.in