Acoustic Event Detection Using Fuzzy Integral Ensemble and Oriented Fuzzy Local Binary Pattern Encoded CNN

Achyut Mani Tripathi, Rashmi Dutta Baruah
Department of Computer Science & Engineering
Indian Institute of Technology Guwahati
Guwahati-781039, Assam, India
{t.achyut, r.duttabaruah}@iitg.ac.in

***Abstract*—In this paper, we propose a novel ensemble classifier using an Oriented Fuzzy Local Binary Pattern Encoded Convolutional Neural Network (CNN) for acoustic event detection (AED). The CNN has been widely used to perform acoustic event detection using a spectrogram image of the acoustic signals. The efficiency of the CNN depends on representation of the spectrogram images used during the training process. We propose the Oriented Fuzzy Local Binary Pattern (OFLBP) that extracts directional texture features from the spectrogram image by inspecting neighborhood pixels present at different angles from a central pixel. The proposed OFLBP technique is capable to deal with uncertainty present in the spectrogram image. The ensemble of the trained CNN is performed by a Fuzzy Integral method. The experiment and results show the proposed method outperforms to existing AED methods to classify the ESC-50 dataset.**

I. Introduction and Prior Work

An acoustic event detection (AED) mainly deals with identification of various events present in an audio stream. Detection of type of acoustic event plays a vital role in various applications such as multimedia [1], computer vision [2], robotics [3], [4] and activity recognition [5]. In a nutshell, the AED is classified into two categories first is identification of the single acoustic event and second is detection of occurrence of the multiple acoustic events together in the recorded acoustic signal [6]. The detection of the individual acoustic event is named as monophonic event detection [6] and identification of the multiple events is known as polyphonic event detection [6], [7], [8]. Our major aim in this paper is to detect the monophonic events. Initially the problem of AED is well explored using techniques that belong to develop automatic speech recognition (ASR) systems. The ASR system involves extraction of multiple features like MFCC [9], FFT [10], Wavelet transformation [11] and ZCR coefficients [12]. Various methods have been proposed to identify the acoustic events using the features mentioned above.

The recent advancement of the ASR [13] and increasing size of sound databases [14], [15] attracts researchers to apply deep learning methods for the AED. In [16], a deep convolutional neural network (DCNN) was proposed to identify the acoustic events. Zhang et al. [17] proposed a CNN based classifier that takes spectrogram images of the acoustic signals to identify the various acoustic events. In [18], Phan et al. proposed CNN classifier that incorporates 1-max pooling technique to identify the acoustic events. Lee et al. [19] proposed the ensemble of the CNN classifiers to detect the acoustic events in DCASE challenges dataset. Soo et al. [20] proposed a hybrid classifier that combines the LTSM and CNN to identify the acoustic events. The LSTM learns the sequential behavior of the audio streams and the CNN extracts features from the spectrogram images of the acoustic signals. The method yields better accuracy as compared to the conventional LSTM, DNN and CNN classifiers. Ozer et al. [21] proposed a noise robust acoustic event detection model that uses the CNN and spectrogram image features to detect the acoustic events. In [22], author investigated the performance of the attention based CNN to identify the acoustic events. In [23] Wang et al. employed RNN to identify the rare acoustic events in real life recordings. A novel 1-D convolutional recurrent neural network (CRNN) [24] was proposed by Wan et al. to detect the rare acoustic events in the acoustic streams.

Various techniques based on Local Binary Pattern (LBP) have been proposed to extract the acoustic features from the spectrogram images to detect the acoustic events [25], [26]. In [27], author combined temporal features with the LBP features to detect the acoustic events using the spectrogram images and the CNN. Majority of the methods mentioned above directly take the spectrogram images to train the deep learning models and failed to deal with the uncertainty present in the spectrogram image caused by noise in the acoustic signal. The fuzzy local binary pattern (FLBP) [28] is capable to deal with the uncertainty present in the images and well applied for the image classification tasks. In this paper the FLBP is used as a preprocessing tool to suppress the effect of the noise on the spectrogram images. The FLBP and LBP are capable to extract many key texture features from the spectrogram images but failed to extract directional texture features ie. line features which may play significant role during the pattern classification in the spectrogram images. In this paper, we introduced a directional version of the FLBP to extract more informative features from the spectrogram images.

An ensemble method is a popular machine learning tech-

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

nique that combines different classifiers to boost the classification accuracy. The performance of the ensemble classifier is generally better than the performance of the individual classifier. The ensemble of classifiers requires combining the decision of each classifier through some method like mean or max voting, weighted or unweighted sum [29]. The ensemble techniques can be broadly classified based on the factors like ensemble size, diversity among the individual classifiers, and the combining methods [30]. The diversity among the individual classifiers used for the ensemble depends on the training methodology, nature of input variables, and the architecture of the classifier.

The traditional ensemble methods used for the AED assume no interaction between the classifiers used for the ensemble. But this assumption is false when the created ensemble models are applied to solve the problems related to real-world data such as the acoustic signals and video streams. The interaction between the classifiers can be positive or negative. In case of the positive interaction, all the classifiers strengthen each other and boost the overall accuracy. On the other side, the negative interaction degrade the performance of the ensemble classifier [31]. Fuzzy logic has shown its promising capabilities to solve various problems using approximate reasoning. Fuzzy integral (FI) [30] is a well used technique to combine the classifiers. The ensemble classifiers created by the FI have shown great performance when applied to solve the problems from diverse research domains. The fuzzy integral is approach is distinct from the other ensemble methods as it resolves all the pitfalls of traditional ensemble techniques as mentioned above. The fuzzy integral computes the importance of the individual classifiers to create a highly accurate classifier with less bias and variance [30]. All the reasons above motivated us to use the fuzzy integral method for ensemble of the CNN models.

To the best of our knowledge the FI-based ensemble model that combines the OFLBP encoded CNN has not been proposed for the AED. Our key contributions in this paper are as follows: First encoding of the spectrogram images using the OFLBP to train the CNN and second is the ensemble of the developed CNN models using the FI technique to improve the classification accuracy.

The organization of the paper is as follows: Section II provides a brief introduction of the LBP and FLBP . Section III presents the methodology used to develop the ensemble classifier using the FI technique and the OFLBP. Section IV provides details of the experiments and results and finally, future work and conclusion are presented in section V.

II. Preliminaries

This section explains the brief introduction of the LBP and FLBP.

A. *Local Binary Pattern (LBP)*

The local binary pattern (LBP) [32] is the popular feature extraction technique widely applied in the research domains such as signal processing and computer vision.

$$D_k = \begin{cases} 1 & if\ P_k >= P_C \\ 0 & if\ P_k < P_C \end{cases} \tag{1}$$

$$LBP = \sum_{k=1}^{7} D_k * 2^k \tag{2}$$

Computation of the LBP pattern is expressed by Eq.(1) and Eq.(2). Where P_k is a value of neighborhood pixel and P_C is a value of central pixel. D_k is a binary code computed using difference of the central and neighborhood pixel. The Eq.(2) is a binary weighted summation of the neighborhood pixels P_k of the central pixel P_C. Fig.(1) shows an example of computation of the LBP code from the image segment of a size (3*3).

The LBP has limitations, first it fails to capture the line features of the image in a final presentation of the image used for the image classification. Second the LBP is sensitive to noise and unable to capture small changes in neighborhood pixel values i.e. uncertainty present in the image.

B. *Fuzzy Local Binary Pattern (FLBP)*

To resolve the shortcoming of the LBP, Iakovidis et al. [28] proposed the FLBP. The FLBP is capable to deal with the uncertainty present in image. Instead of assigning binary code as 0 or 1 directly by comparing the central pixel against the neighborhood pixels using the Eq.(1), the FLBP computes degree of membership of the pixel that belong to two fuzzy sets named as small and greater fuzzy sets. Eq.(3) is used to calculate the membership of the k^{th} neighborhood pixel for the small fuzzy set μ_k^0. However, the membership of the greater fuzzy set μ_k^1 is computed using Eq.(4). Fig.(2) shows complete procedure to compute the FLBP with a threshold $(T = 5)$.

$$\begin{gathered} \mu_k^0 = 0, if\ P_k \geq P_C + T \\ \mu_k^0 = 1, if\ P_k \leq P_C - T \\ \mu_k^0 = \frac{T - P_k + P_C}{2 * T}, \\ if\ P_C + T > P_k > P_C - T \end{gathered} \tag{3}$$

$$\mu_k^1 = 1 - \mu_k^0 \tag{4}$$

Where P_k, P_C, T are the k^{th} neighborhood pixel, central pixel and threshold respectively.

III. Methodology

This section presents details of the OFLBP and ensemble of the CNN by the FI method.

A. *Oriented Fuzzy Local Binary Pattern (OFLBP)*

The OFLBP considers directional neighborhood pixels while computation of the FLBP. Consideration of directional neighborhood pixels is helpful to extract more robust texture features of the image [33] using the FLBP. The OFLBP identify the directional neighborhood pixels using Eq.(5) and extracts FLBP using the procedure as mentioned in earlier

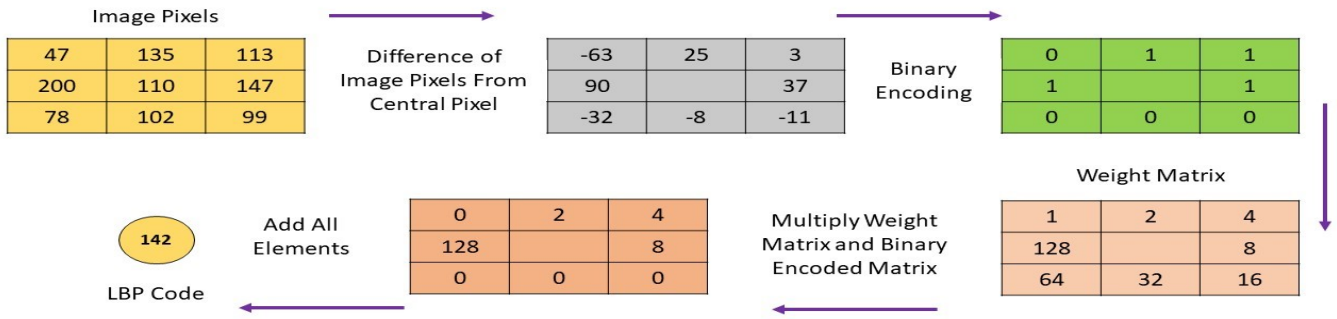

Fig. 1. Computation of LBP Code

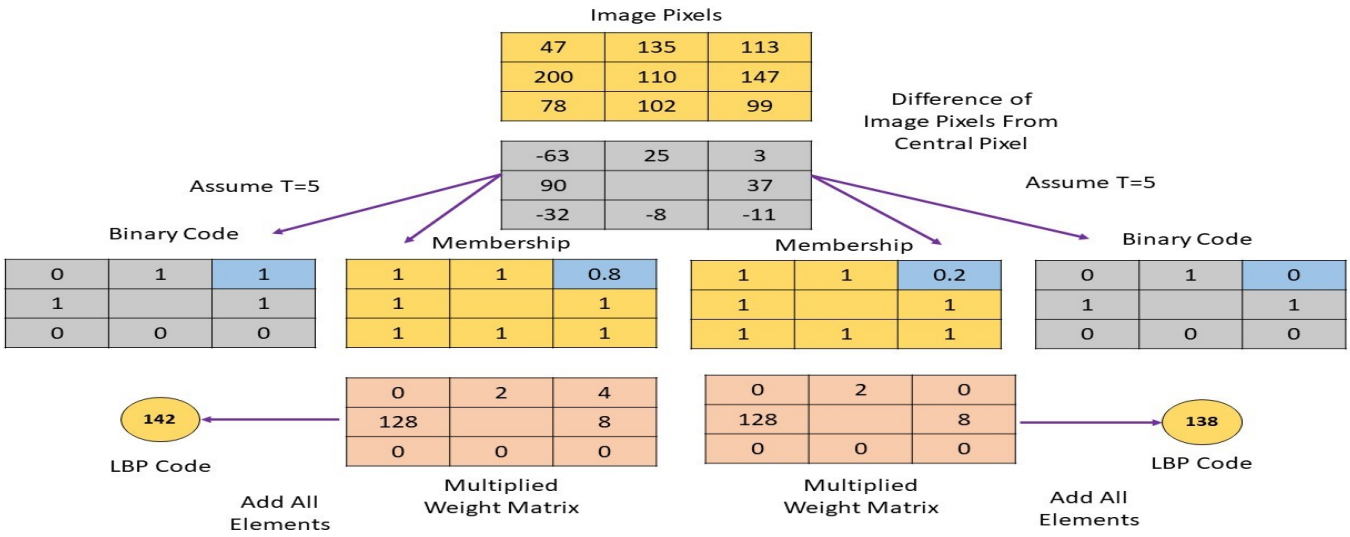

Fig. 2. Computation of FLBP Code

section. The OFLBP creates a pixel matrix of dimension $(9 * 9)$ and the central pixel is placed at location $(5 * 5)$. Eq.(5) and Eq.(6) are used to compute the coordinates of the neighborhood pixels that are present at a different angle (θ) form the central pixel [33]. Fig.(3) and Fig.(4) show the locations of the selected directional neighborhood pixels for the angles $\theta = [0^0, 30^0, 60^0]$ and $\theta = [90^0, 120^0, 150^0]$ respectively.

$$if\ \theta\ \epsilon[0^0, 45^0]\ or\ \theta\ \epsilon[135^0, 180^0]$$

$$(P_k^x, P_k^y) = \begin{cases} P_k^x = P_C^x + \beta - k \\ P_k^y = P_C^y + \lfloor P_k^x * tan(\theta) \rfloor \end{cases} \quad (5)$$

$$if\ \theta\ \epsilon(45^0, 135^0)$$

$$(P_k^x, P_k^y) = \begin{cases} P_k^y = P_C^y + \beta - k \\ P_k^x = P_C^x + \lfloor P_k^y * (1/tan(\theta)) \rfloor \end{cases} \quad (6)$$

Where, P_k^x and P_k^y are the x and y coordinate positions of the k^{th} neighborhood pixel. P_C^x and P_C^y are the x and y coordinates of the central pixel and $\lfloor a \rfloor$ is a floor function. If P denotes number of neighborhood pixels required (In this paper P=7). Value of β is selected as 4 for the neighborhood pixels at the indexes 0 to P-4 and selected as 3 for the neighborhood pixels with the indexes P-3 to P .

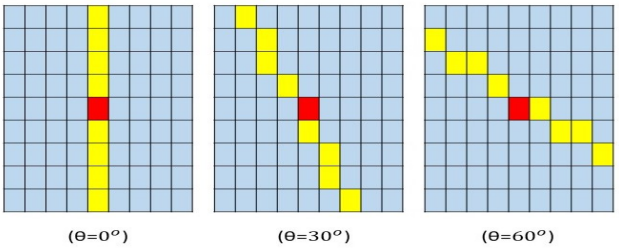

Fig. 3. Neighbors at $\theta = 0, 30, 60$

B. *Convolutional Neural Network (CNN)*

The CNN [34] is a popular deep learning technique that has been extensively applied to solve numerous computer vision problems. The CNN takes the image as input and learn the bias and weight parameters to perform the classification. The

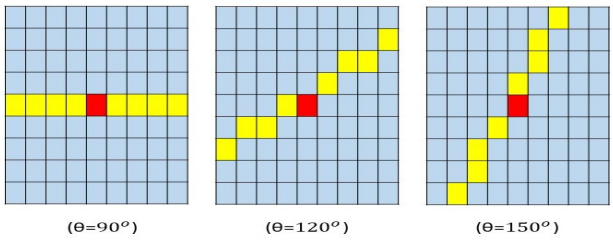

Fig. 4. Neighbors at $\theta = 90, 120, 150$

architecture of the CNN is inspired for the visual cortex part of a human brain. The architecture of the CNN contains an Input layer, Convolution layer, Pooling layer, Fully Connected layer, and an Output layer. Fig.(5) shows a basic architecture of the CNN.

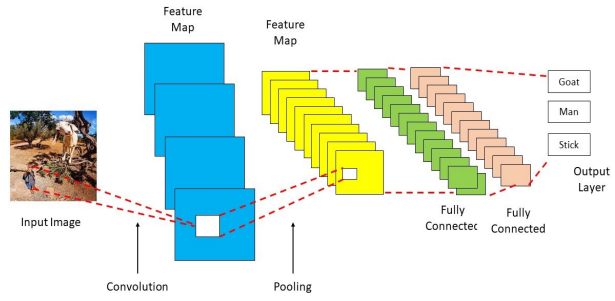

Fig. 5. Architecture of Convolutional Neural Network

C. *Fuzzy Integral Ensemble (FI)*

The major steps of the FI ensemble [30] are as follows:

1) Compute fuzzy densities $(g^1, .., g^Z)$ of all the classifiers using a training accuracy of the classifiers developed to classify the dataset with k classes as suggested in [30]. In our case we initialized the value of Z=6. Where Z is the number of developed classifiers.
2) Find the value of λ using Eq.(7) and the fuzzy densities g^Z obtained in the previous step.

$$\lambda + 1 = \prod_{r=1}^{Z} (1 + \lambda g^z) \tag{7}$$

Where λ is real root with a value greater than -1.

3) Create a decision profile matrix (DP) as shown in Eq.(8). Each entry in DP matrix shows the decision of Z^{th} classifier over the given input X for the class K.Here $d_{Z,k}(X)$ is decision of z^{th} classifier to classify the input X in class k.

$$DP(X) = \begin{bmatrix} d_{1,1}(X) & d_{1,2}(X) & -- & d_{1,k}(X) \\ d_{2,1}(X) & d_{2,2}(X) & -- & d_{2,k}(X) \\ d_{3,1}(X) & d_{3,2}(X) & -- & d_{3,k}(X) \\ -- & -- & -- & -- \\ d_{Z,1}(X) & d_{Z,1}(X) & -- & d_{Z,k}(X) \end{bmatrix} \tag{8}$$

4) Initialize the values of $g(1) = g^z, 1 \leq z \leq Z$ and recursively compute a value of $g(z)$ by solving Eq.(9).

$$g(z) = g^z + g(z-1) + \lambda g^z g(z-1), 2 \leq z \leq Z \tag{9}$$

5) Calculate a value of degree of support $\mu_L^k(X)$ as given in Eq.(10)

$$\mu_Z^k(X) = max\ (\ min\ (\mu_z^k(X), g^k(z))) \tag{10}$$

D. *Detection of Acoustic Events*

Methodology to identify the acoustic events using the FI based ensemble classifier is as follows:

1) Initially create the Z set of the OFLBP encoded spectrogram images for the given training signals.
2) Instead of using the OFLBP encoded spectrogram images directly to train the CNN we use a mapping technique proposed by Levi et al. [35]. The values computed by the OFLBP method is mapped to 3D metric space by approximation of the euclidean distance between the OFLBP codes. This mapping transforms the OFLBP encoded spectrogram image into a representation that can be used in the CNN.
3) Train the Z CNN models using the Z set of training data created in the previous step. Fig.(6) shows the procedure to train the CNN models.
4) Create the OFLBP encoded spectrogram image of the test signals and classify the image using the Z trained CNN classifiers.
5) Use the FI ensemble to identify the acoustic event present in the given test signal. Fig.(7) shows the procedure to detect the acoustic event in the given test signal.

IV. Experiments and Results

This section presents the experiments and results.

A. *Dataset Description*

To test the efficacy of the proposed method we selected a publicly available Environment Sound Classification (ESC-50) datasets [36]. The dataset contains 2000 recordings of the 50 different real life environmental sounds. Duration of each signal is 5 seconds. The signals of the ESC-50 dataset is further grouped into five major category as shown in Table I. Fig.(8)-Fig(11) show the spectrogram images of clock, engine sound, keyboard typing and water drop acoustic signal present in the ESC-50 dataset.The ESC-50 dataset provides prearranged files for the 5 fold cross validation thus the obtained results are directly compared with the baseline and state-of-the art methods. The spectrogram of the environmental sound is created using an in-built function of a MATLAB. The sampling rate of the signal is kept 32kHz and a frame size is selected as 30 ms with an overlapping window of 50%. The training of the CNN is performed with a learning rate of 0.01, 50 epochs and Stochastic Gradient Descent (SGD) solver.

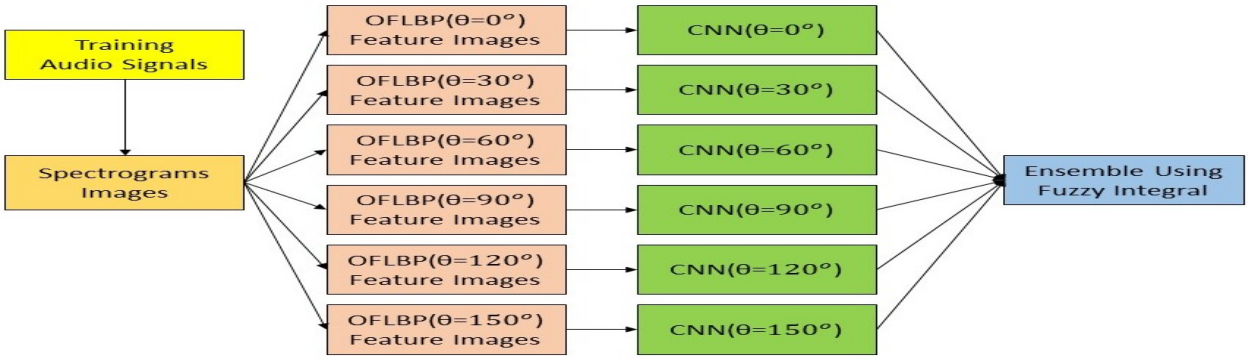

Fig. 6. Training Using CNN Classifiers

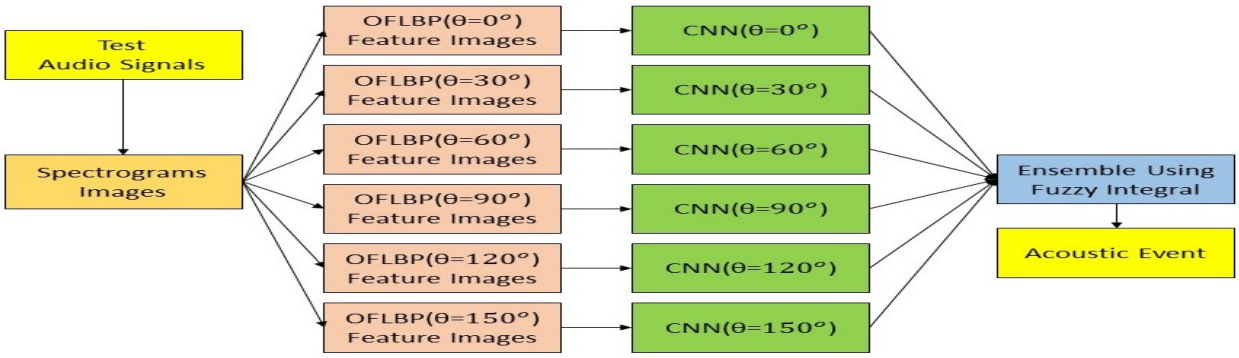

Fig. 7. Acoustic Event Detection using CNN Classifier

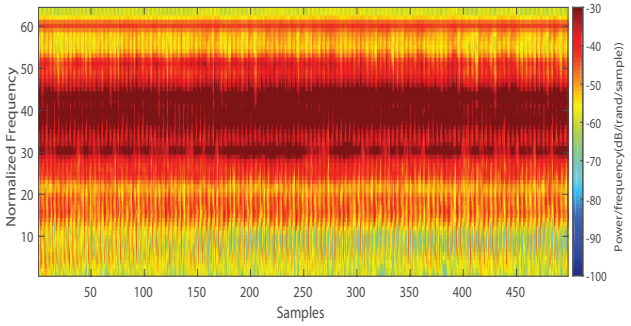

Fig. 8. Spectrogram of Clock Sound

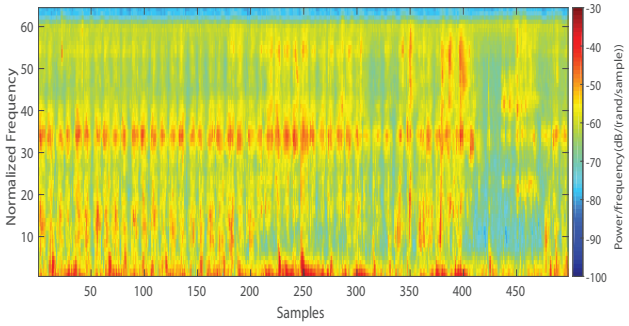

Fig. 9. Spectrogram of Engine Sound

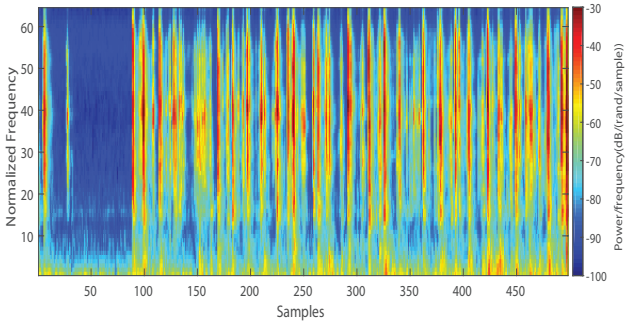

Fig. 10. Spectrogram of Keyboard Sound

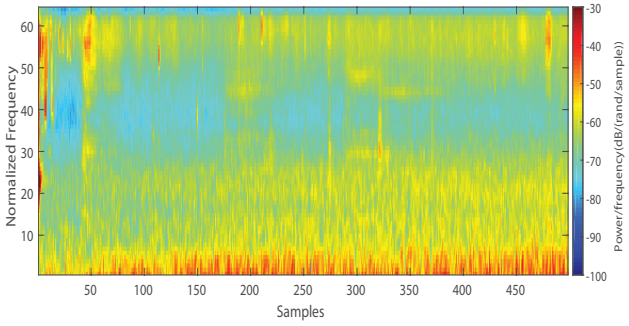

Fig. 11. Spectrogram of Water Drop Sound

TABLE I
MAJOR CATEGORIES OF ESC-50 DATASET

S.No.	Category
1	Animals
2	Natural soundscapes & water sounds
3	Human, non-speech sounds
4	Interior/domestic sounds
5	Exterior/urban noises

TABLE II
DIFFERENT MODELS DEVELOPED FOR AED

Model	Description
M1	CNN model trained with OFLBP encoded images with $\theta = 0^0$
M2	CNN model trained with OFLBP encoded images with $\theta = 30^0$
M3	CNN model trained with OFLBP encoded images with $\theta = 60^0$
M4	CNN model trained with OFLBP encoded images with $\theta = 90^0$
M5	CNN model trained with OFLBP encoded images with $\theta = 120^0$
M6	CNN model trained with OFLBP encoded images with $\theta = 150^0$
M7	FI ensemble of M1, M2, M3, M4, M5, M6

B. *Model Description*

Total seven models were developed to examine the performance of the proposed method. Table II shows the details of all the seven models. Table III shows details of the architecture of the CNN model used to classify the environmental sounds. Later the six different CNN models M1, M2, M3, M4, M5 and M6 are developed using the set of OFLBP images created with the six different values of angle $(\theta) = [0^0, 30^0, 60^0, 90^0, 120^0, 150^0]$. The models M1 to M6 are ensemble using the FI technique to create the model M7. In our case we selected the value of parameter T as 2 while the computation of OFLBP features from the spectrogram images.

C. *Results*

Table IV shows the accuracy achieved by the seven models to detect the environmental sounds present in the SEC-50 dataset. The model M1 $\theta = 0^0$ shows the lowest accuracy of 84.53%. The ensemble model M7 shows the highest accuracy of 90.03%. The models M2, M3, M4, M5 and M6 show the accuracy of 86.61%, 85.91%, 84.69%, 88.27% and 86.89% respectively. We also compared the results of models M1-M6 using traditional ensemble techniques. The ensemble of models M1-M6 using max vote rule attains the accuracy of 84.15%, however the accuracy of the same models using mean vote rule yields the accuracy of 83.58%.

The baseline accuracy of KNN, SVM and random forest is 32.20%, 39.60% and 44.30% respectively.

TABLE III
DETAILS OF TRAINING OF CNN

Layer	Filter Size	Number of Filters	Activation Function	Max -Pooling Filter	Dropout
Convolutional Layer	(11 , 11)	96	relu	(2 , 2)	
Convolutional Layer	(5 , 5)	256	relu	(2 , 2)	
Convolutional Layer	(3, 3)	384	relu	-	
Convolutional Layer	(3, 3)	384	relu	-	
Convolutional Layer	(3 , 3)	256	relu	(2 , 2)	
Fully Connected Layer (4096)	-	-	relu	-	0.5
Fully Connected Layer (4096)	-	-	relu	-	0.5
Sofmax Layer 5 Classes	-	-	-	-	-

The fuzzy integral-based ensemble model (M7) yields highest accuracy among all the developed models and state of the art methods to classify the ESC-50 dataset as shown in the Table IV. It is clear from the Table IV that the OFLBP encoding of the spectrogram image efficiently deals with the uncertainty present in the spectrogram image and provides more useful features to train the CNN. The ensemble of the CNN using fuzzy integral technique enhances the accuracy of the classification of acoustic events.

TABLE IV
ACCURACY OF DEVELOPED MODELS AND STATE OF THE ART METHODS

S.No.	Model	Accuracy(%)
1	M1	84.53
2	M2	86.61
3	M3	85.91
4	M4	84.69
5	M5	88.27
6	M6	86.89
7	**M7**	**90.03**
8	Hardik et al [37]	86.50
9	Yuji et al. [38]	84.90
10	Kumar et al. [39]	83.50
11	Baseline KNN [36]	32.20
12	Baseline SVM [36]	39.60
13	Baseline Random Forest [36]	44.30
14	Ensemble Using Max Vote Rule	84.15
15	Ensemble Using Mean Vote Rule	83.58

V. CONCLUSION

In this paper we proposed the FI-based ensemble classifier that combines the decision of multiple OFLBP encoded CNNs to detect the desired acoustic event. The OFLBP efficiently extracts useful directional texture features from the directional neighbor pixels and also capable to deal with the uncertainty present in the spectrogram image of the acoustic signal. The experiments and results indicate the proposed method shows

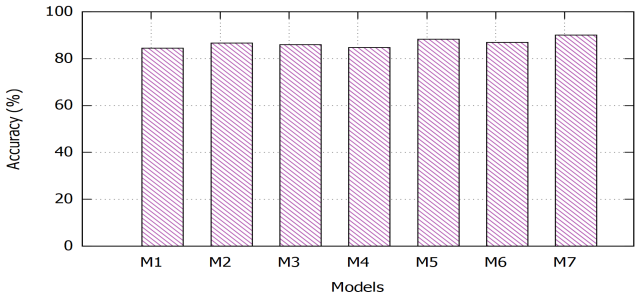

Fig. 12. Accuracy Of Models M1-M7

better accuracy as compared to conventional classifiers used to perform the acoustic event detection.

In future we would like to extend our work to detect the acoustic events in overlapping acoustic signals.

REFERENCES

[1] P. K. Atrey, N. C. Maddage, and M. S. Kankanhalli, "Audio based event detection for multimedia surveillance," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5. IEEE, 2006, pp. V–V.

[2] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos." in *INTERSPEECH*, 2013, pp. 2929–2933.

[3] R. Gomez, K. Inoue, K. Nakamura, T. Mizumoto, and K. Nakadai, "Speech-based human-robot interaction robust to acoustic reflections in real environment," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1367–1373.

[4] S. Aziz, M. Awais, T. Akram, U. Khan, M. Alhussein, and K. Aurangzeb, "Automatic scene recognition through acoustic classification for behavioral robotics," *Electronics*, vol. 8, no. 5, p. 483, 2019.

[5] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[6] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[7] A. M. Tripathi and R. D. Baruah, "Acoustic event classification using Cauchy non-negative matrix factorization and fuzzy rule-based classifier," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2017, pp. 1–6.

[8] ——, "Incremental Cauchy non-negative matrix factorization and fuzzy rule-based classifier for acoustic source separation," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019, pp. 1–6.

[9] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International Conference on Computer Graphics, Simulation and Modeling*, 2012, pp. 135–138.

[10] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Speech emotion recognition using eigen-fft in clean and noisy environments," in *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2007, pp. 689–694.

[11] A. Bhowmick and M. Chandra, "Speech enhancement using voiced speech probability based wavelet decomposition," *Computers & Electrical Engineering*, vol. 62, pp. 706–718, 2017.

[12] M. R. Gamit and K. Dhameliya, "Isolated words recognition using mfcc, lpc and neural network," *International journal of Research in Engineering and technology*, vol. 4, no. 6, pp. 146–9, 2015.

[13] H. Geoffrey, D. Li, Y. Dong, E. D. George, and A.-r. Mohamed, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[14] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[15] H. Fan, J. Zhou, and C. Fuegen, "Facebook acoustic events dataset," in *ICASSP*, 2018.

[16] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[17] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 559–563.

[18] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *arXiv preprint arXiv:1604.06338*, 2016.

[19] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[20] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of lstm and cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[21] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing*, vol. 272, pp. 505–512, 2018.

[22] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based cldnns for short-duration acoustic scene classification." in *INTERSPEECH*, 2017, pp. 469–473.

[23] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2742–2746.

[24] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 80–84.

[25] D. Battaglino, L. Lepauloux, L. Pilati, and N. Evans, "Acoustic context recognition using local binary pattern codebooks," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[26] T. Kobayashi and J. Ye, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3052–3056.

[27] W. Yang, S. Krishnan, W. Yang, and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 6, pp. 1315–1321, 2017.

[28] D. K. Iakovidis, E. G. Keramidas, and D. Maroulis, "Fuzzy local binary patterns for ultrasound texture characterization," in *International conference image analysis and recognition*. Springer, 2008, pp. 750–759.

[29] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[30] D. Ralescu and G. Adams, "The fuzzy integral," *Journal of Mathematical Analysis and Applications*, vol. 75, no. 2, pp. 562–570, 1980.

[31] J. Zhai, H. Xu, and Y. Li, "Fusion of extreme learning machine with fuzzy integral," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 21, no. supp02, pp. 23–34, 2013.

[32] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European conference on computer vision*. Springer, 2004, pp. 469–481.

[33] M. Kass and A. Witkin, "Analyzing oriented patterns," *Computer vision, graphics, and image processing*, vol. 37, no. 3, pp. 362–385, 1987.

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

[35] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 503–510.

[36] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on*

Multimedia. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[37] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification." in *INTERSPEECH*, 2017, pp. 3107–3111.

[38] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[39] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 326–330.