Multi Feature Sound Classification using Deep Learning

[1]Vrushali K. Bongirwar, [2]Samir N. Ajani, [3]Archana V. Potnurwar

[1]Shri Ramdeobaba College of Engineering and Management, Nagpur, India.

[2]St. Vincent Pallotti College of Engineering and Technology, Nagpur, India.

[3]Priyadarshini College of Engineering, Nagpur, India.

Abstract.

Artificial Intelligence plays an important role in acoustics recognition. Importantly, being able to automatically and accurately identify environmental sounds opens up a broad range of applications. Deep learning techniques can assist in the recognition of sounds which we come across in our day-to-day life. Most of the previous work in environmental sound classification involves training a model on a single set of features. Convolutional neural network (ConvNet) is a class of deep feed-forward neural network which exploits the strong spatially local correlation in natural images. It achieves successful performance in visual analyzing area. This paper primarily focuses on two key aims: the first aim is to perform a multi-label classification system and the second aim is to develop Stacked Bidirectional Long Short-Term Memory (LSTM) with two hidden layers to categorize multiple UAVs sounds. There are three portions to perform environmental classification. Firstly, the input signal is converted into spectrogram image with time-frequency representation using short time Fourier transforms. Secondly, this spectrogram is used to extract features with local binary pattern of three different radius and neighborhood sizes. The three distinct features resulted from local binary pattern based on spectrogram are concatenated and used as one feature vector.

Keywords: *Deep learning, Long Short-Term Memory, Convolutional neural network.*

DOI Number: 10.14704/nq.2022.20.13.NQ88096 **Neuro Quantology 2022; 20(13):743-748**

1. Introduction:

Sound signs contain an abundance of data and are firmly connected with human existence. Accordingly, arranging the climate through the data gathered from the sound is the focal point of this paper. The order of metropolitan conditions in light of recordings and pictures requires an enormous number of video pictures and weighty computations. it causes specific constraints [1]. It is worth focusing on that the sound is an undeniable element to recognize in a metropolitan climate and has not disadvantages referenced previously. Besides, it flaunts different benefits, for example, advantageous information assortment, low organization cost, and simple recognizable proof of sound highlights. Accordingly, the sound is important in principle research and useful application for the grouping of metropolitan conditions. It has been broadly utilized in media recovery and metropolitan informatics.

Past examination of sound acknowledgment depended on format matching [2], include boundary matching [3] and secret Markov techniques [4]. The above strategies accomplish great execution in the early discourse acknowledgment frameworks. Be that as it may, they are not vigorous and can't deal with huge scope information. In 2006, Hinton utilized unaided pre-preparing to take care of the issue of slope vanishing [5], [6]. It makes profound brain networks teachable and extended the secret layers to 7 layers. Brain networks have "profundity" in the genuine feeling of the word. It advanced the improvement of profound learning and prompted the ascent of the third era of brain networks [7]. In 2012, Abdel-Hamid O et al. applied Convolutional Brain Organizations (CNN) to the sound acknowledgment interestingly. The trial results showed that the sound acknowledgment in view of CNN would be advised to execution with conventional techniques [8]. From that point onward, numerous researchers have likewise started to involve CNN for the ESC issues. Notwithstanding, machine recognizable proof has barely arrived at the level of the human.

2. Literature Survey:

Many exploration approaches [1], [6] like nourishment admission furthermore, discourse order is introduced surface elements in light of spectrogram. In [5], speaker distinguishing proof is recognized by utilizing example or surface coordinating. Surface elements utilized in time-recurrence portrayal same as the surface elements utilized in picture handling. Thus, numerous sorts of surface highlights can be applied in time-recurrence portrayal called spectrogram to arrange or perceive the sound [3]. Most applications [4], [8] utilize nearby parallel example (LBP) on account of its straightforward and quick properties. Neighborhood Double Example has interpretation invariant. This invariant is numerous valuable in spectrogram. A few applications [3], [14] present surface highlights getting on spectrogram characterize various melodic instruments by utilizing surface element extraction strategy based on time-recurrence coordinating. Jonathan Dennis [8] researched a system to group sound occasion with confounded conditions. Sound, right off the bat, is changed over into spectrogram utilizing discrete Fourier change. Then the dark scale spectrogram is quantized into various locales and focal minutes are utilized as the component vector of the sound. In [15], highlights are separated from 12-log gabor channel utilizing spectrogram. 12-log gabor channel with various scale and direction is utilized to remove highlights from spectrogram. The come about highlights are connected into one layered include vector. This one layered include vector is passed thought the MI models. The help vector machine is utilized for arrangement and 89.62% is reached.

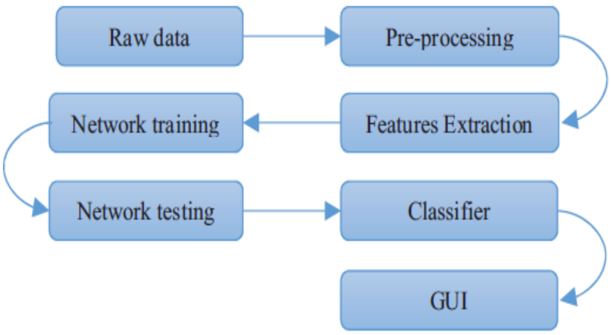
Takumi [16] created include extraction from spectrogram utilizing nearby twofold example. The came about designs are weighted by the nearby insights, mean and standard deviation and the L2-Hellinger standardization is utilized to improve the discriminative of elements. At last, support vector machine is utilized for arrangement.

Lately, profound convolution brain network has gained extraordinary headway in sound acknowledgment and different viewpoints [11], [12], [13]. Zhang et al. [14] utilized brief time frame Fourier change (STFT) and different techniques to change over birds sound into the range and utilized convolutional brain organization to group bird sounds. Not the same as utilizing a basic convolutional brain organization, Sankupellay et al. [15] utilized 50 layers lingering brain organization (Resnet50) to arrange the time range of bird sounds. Huang et al. [16] utilized thickly associated networks (Densenet) to separate time range includes and characterize them, which further developed the characterization impact. To additionally further develop the acknowledgment precision. Sheng et al. [17] utilized 1-layered CNN-LSTM, 2-layered vgg-style, and 3-layered densenet121 model as component extractors to remove progressed highlights, and afterward utilized a shallow classifier to perceive 43 sorts of bird sounds, accomplishing a decent exactness of 93.89%. The system [18] goes astray from the current methodologies by incorporating move learning. Utilizing like ResNet50, DenseNet201, InceptionV3, Xception, and EfficientNet can really separate and perceive the sound signs from various bird species with critical forecast precision. In the above profound learning model, the perplexing component ex-footing calculation is supplanted by different more profound and high-accuracy models with numerous boundaries, however this likewise deals with a similar issue. Countless boundaries will diminish the registering pace of the gadget, and complex model matches can't be applied to minimal expense computer chip. Running the models in minimal expense implanted devices is as yet ridiculous.

Moreover, albeit the vast majority of the examinations on bird sound acknowledgment have accomplished high acknowledgment exactness, how much informational index utilized in the exploration is little [17-36]. Most examinations are restricted to distinguishing a solitary bird animal varieties, and the quantity of bird species in the informational index utilized is simply 20 to 30 (in the accompanying, this paper will show a few similar information), so the proposed model doesn't have speculation capacity.

3. Proposed System

In this we have proposed a system to perceive bird species using Artificial Neural Network (ANN). ANN is applied to arrange and see the bird picture fetaures using Matlab programming. First thing, all fundamental data in term of power absurd thickness of bird is used to get data for each bird types. The accompanying cycle is to plan ANN to perceive sorts of birds. At long last, the graphical UI (GUI) of bird picture ID have been developed that vital the customer to dealt with set of picture as a commitment of bird to see bird species. The cycle in detail is portrayed under:



igure 1. Sound Processing using ANN

The rough data considered is the image of four birds types for the present circumstance . In pre-dealing with, power absurd thickness (PSD) was used to make different features of a bird present in picture. The faint scale picture was differentiated and special picture. From here on out, and the data will get ready using neural association. An Artificial Neural Network (ANN) classifier offers a transcendent system for picture preprocessing ,portrayal , assessment and affirmation. The multilayer perceptron (MLP) is used as classifier. The MLP takes a lot of picked features as data and has a substitute yield for each bird species to be seen. The researchers can use the system through Graphic User Interface (GUI). From GUI, customer can sort out what sort of bird will be requested. The Artificial Neural Network is an information methodology have been used to portray and recognize of bird type There have various procedures have been explored by various experts to do bird picture affirmation.

3.1 Softmax Regression on Binary Attributes

The Softmax backslide model relied upon matched data of 312 bird attributes (wing tone, charge shape, tail plan, etc...) that was actually assembled in the Caltech-UCSB study and thusly didn't use PC vision. Or then again perhaps, this model was required to fill in as a benchmark for our gathering execution. The data was run for unequivocally organized (ex. Prairie and Pine Warblers in disconnected classes with 200 classes hard and fast) and moreover on 'thoroughly gathered data' (ex. all Warblers fall into one class with 71 classes full scale).

3.2 Multiclass SVM on HOG+RGB features

The part vectors used in this model are the histogram of organized points (HOG) connected with RGB histogram regards. Pigs are incorporate descriptors which had wide accomplishment in picture acknowledgment (Dalal and Triggs 2005) due to its ability to perceive blueprints and invariance to numerical and photometric changes (except for thing heading). Since bird tone is key in distinctive its class, RGB histograms of the photos are used to help in unmistakable evidence.

3.3 CNNs and Transfer Learning

proposed a novel visual after computation subject to the depictions from a discriminatively arranged Convolutional Neural Network (CNN). Our count pretrains a CNN using a huge plan of chronicles with following groundtruths to get a nonexclusive target depiction. Our association is made out of shared layers and various pieces of room unequivocal layers, where zones contrast with solitary planning progressions and each branch is at risk for equal request to recognize the target in each region.

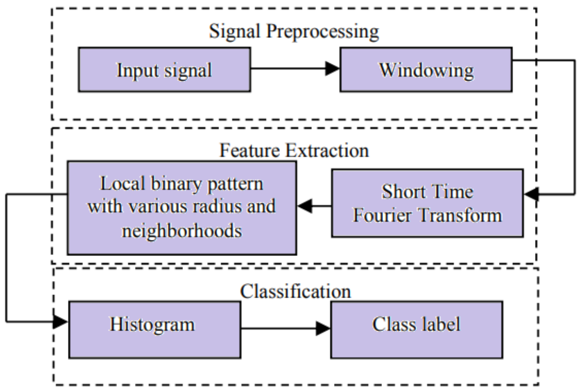


Figure 2. Transfer Learning Model

4. Results

Our proposed system approach has been tested over 15 species of flying wild creatures considered from Caltech image dataset. We have done training of our system using T-1 image dataset that represents actually around total number of images of that Caltech image class. All prediction we had obtained is true for every unknown image present in dataset. Unknown image which is not available in the dataset is removed out from the database for checking the true prediction for every unknown image. From the experiment we can observe that the TP(U) and TP(D) has been always more than 0.5%.In most of the cases both the TD(U) and TP(U) values are mostly more than 08%.These proves that how much accurate is our proposed algorithm is.

Similarly ,the wrong prediction(false prediction) values is range from 0.02 to 0.2 and yes not more than that, which proves that the accuracy level of our proposed algorithm which is shown in table(1).

When we look at the training accuracy of the system it generally starts from 28% and drops at 100% for the rest of the image dataset in the class.

As shown in figure(1) ,the validation accuracy percentage and the cross entropy percentage are inversely proportional to each other.

Step	Train accuracy (%)	Cross entropy	Validation accuracy (%)
0	29.00	2.432535	17.00
100	87.00	1.200826	89.00
500	96.00	0.447151	89.00
1000	99.00	0.235821	96.00
1500	100.00	0.17149	96.00
2000	100.00	0.116124	96.00
2500	100.00	0.096606	95.00
2800	100.00	0.068634	96.00
3999	100.00	0.060213	96.00

Table 1. Accuracy Measure

S. No.	BS	TI	TP(*D*)	TP(*U*)	FP(*D*)	FP(*U*)
1	B1	36	0.86	0.58	B3- 0.11	B3- 0.28
					B2- 0.02	B9- 0.03
					B6- 0.01	B110.02
					B5- 0.01	B5- 0.02
2	B2	30	0.96	0.96	B1- 0.01	B1- 0.02
3	B3	30	0.71	0.5	B2- 0.21	B2- 0.25
					B1- 0.04	B1- 0.21
					B7- 0.01	B5- 0.01
					B6- 0.01	B7- 0.01
4	B4	35	0.99	0.92	B1- 0.01	B9- 0.03
					B1- 0.01	B110.02
					B9- 0.01	B100.01
5	B5	31	0.91	0.95	B1- 0.01	B5- 0.01
					B7- 0.05	B7- 0.02
					B8- 0.01	B11- 0.01
					B6- 0.01	B8- 0.01
6	B6	30	0.67	0.89	B4- 0.13	B5- 0.05
					B7- 0.07	B11- 0.02
					B5- 0.05	B7- 0.02
					B8- 0.02	B9- 0.01
7	B7	33	0.87	0.79	B6- 0.08	B10- 0.09
					B5- 0.02	B7- 0.065
					B10- 0.01	B11- 0.02
					B11- 0.01	B9- 0.01
8	B8	32	0.63	0.95	B11- 0.15	B5- 0.01
					B5- 0.07	B11- 0.01
					B10- 0.05	B10- 0.01
					B9- 0.03	B6- 0.005
9	B9	30	0.97	0.87	B11- 0.02	B11- 0.12
					B10- 0.01	B10- 0.01
					B8- 0.001	B5- 0.001
					B12- 0.001	B6- 0.001
10	B10	30	0.65	0.97	B9- 0.287	B11- 0.01
					B6- 0.026	B5- 0.010
					B11- 0.018	B9- 0.004
					B4- 0.005	B12- 0.01
11	B11	30	0.81	0.96	B6- 0.089	B9- 0.036
					B10- 0.042	B10- 0.01
					B9- 0.028	B5- 0.001
					B8- 0.01	B12- 0.01

Table 2. Sound Feature Prediction

5. Conclusion:

This research paper has a wide exploration of various unique techniques for identification of sound and its classification with different types Deep learning algorithms. Image processing based various techniques available in previous work has been evaluated and refined for the use of our approach. Co-related features of sound samples can be extracted from the image set of sound has been also identified in this paper. The various physical appearance of sound and deep learning approaches can be combine together to obtained a better result and evaluated the data again to check whether we can do further improvement to achieve more accuracy. Redundant information has been removed in our processing system to obtained enhance and better result.

References:

[1] C. Herley and M. Vetterli, "Wavelets and recursive filter banks," in *IEEE Transactions on Signal Processing*, vol. 41, no. 8, pp. 2536-2556, Aug. 1993.

[2] S. O. Maruf, M. U. Azhar, S. G. Khawaja and M. U. Akram, "Crackle separation and classification from normal Respiratory sounds using Gaussian Mixture Model," *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, Peradeniya, Sri Lanka, 2015, pp. 267-271.

[3] Ajani, S.N., Amdani, S.Y. (2021). Agent-Based Path Prediction Strategy (ABPP) for Navigation Over Dynamic Environment. In: Muthukumar, P., Sarkar, D.K., De, D., De, C.K. (eds) Innovations in Sustainable Energy and Technology. Advances in Sustainability Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-16-1119-3_16

[4] M. Lozano, J. A. Fiz and R. Jané, "Automatic Differentiation of Normal and Continuous Adventitious Respiratory Sounds Using Ensemble Empirical Mode Decomposition and Instantaneous Frequency," in *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 486-497, March 2016.

[5] Y. P. Kahya, E. Ç. Güler, B. Sankur and T. Engin, "Detection and clustering analysis of crackles in respiratory sounds," *1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Paris, France, 1992, pp. 2527-2528.

[6] S. N. Ajani and S. Y. Amdani, "Probabilistic path planning using current obstacle position in static environment," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170727.

[7] Samir N Ajani Piyush K. Ingole , Apeksha V. Sakhare “Modality of Multi-Attribute Decision Making for Network Selection in Heterogeneous Wireless Networks”, Ambient Science - National Cave Research and Protection Organization, India,2022, Vol.9, Issue.2, ISSN- 2348 5191. DOI:10.21276/ambi.2022.09.2.ta02

[8] S. Emrani, T. Gentimis and H. Krim, "Persistent Homology of Delay Embeddings and its Application to Wheeze Detection," in *IEEE Signal Processing Letters*, vol. 21, no. 4, pp. 459-463, April 2014.

[9] Sen, M. Saraclar and Y. P. Kahya, "A Comparison of SVM and GMM-Based Classifier Configurations for Diagnostic Classification of Pulmonary Sounds," in *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 7, pp. 1768-1776, July 2015.

[10] Ajani, S., Amdani, S.Y. (2022). Obstacle Collision Prediction Model for Path Planning Using Obstacle Trajectory Clustering. In: Sharma, S., Peng, SL., Agrawal, J., Shukla, R.K., Le, DN. (eds) Data, Engineering and Applications. Lecture Notes in Electrical Engineering, vol 907. Springer, Singapore. https://doi.org/10.1007/978-981-19-4687-5_8

[11] R. G. Stockwell, L. Mansinha and R. P. Lowe, "Localization of the complex spectrum: the S transform," in *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 998-1001, April 1996.

[12] A. Moukadem, A. Dieterlen and C. Brandt, "Shannon Entropy based on the S-Transform Spectrogram applied on the classification of heart sounds," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 704-708.

[13] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[14] Ajani, S.N., Bhanarkar, P. (2022). Design a Mechanism for Opinion Mining. In: Sharma, S., Peng, SL., Agrawal, J., Shukla, R.K., Le, DN. (eds) Data, Engineering and Applications. Lecture Notes in Electrical Engineering, vol 907. Springer, Singapore. https://doi.org/10.1007/978-981-19-4687-5_35

[15] S. Jeon, J. -W. Shin, Y. -J. Lee, W. -H. Kim, Y. Kwon and H. -Y. Yang, "Empirical study of drone sound detection in real-life environment with deep neural networks," *2017 25th European Signal*

Processing Conference (EUSIPCO), Kos, Greece, 2017, pp. 1858-1862.

[16] S. Al-Emadi, A. Al-Ali, A. Mohammad and A. Al-Ali, "Audio Based Drone Detection and Identification using Deep Learning," *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, Tangier, Morocco, 2019, pp. 459-464.

[17] M. Z. Anwar, Z. Kaleem and A. Jamalipour, "Machine Learning Inspired Sound-Based Amateur Drone Detection for Public Safety Applications," in *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2526-2534, March 2019.

[18] H. Chen *et al.*, "AdderNet: Do We Really Need Multiplications in Deep Learning?," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 1465-1474.

[19] G. -Y. Wang, Y. -M. Zhang, M. -L. Sun, X. Wang and Y. Zhang, "Speech signal feature parameters extraction algorithm based on PCNN for isolated word recognition," *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, Shanghai, China, 2016, pp. 679-682.

[20] S. Mao, D. Tao, G. Zhang, P. C. Ching and T. Lee, "Revisiting Hidden Markov Models for Speech Emotion Recognition," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 6715-6719.

[21] Graves, A. -r. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 6645-6649.

[22] B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," in *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[23] K. J. Piczak, "Environmental sound classification with convolutional neural networks," *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, USA, 2015, pp. 1-6.

[24] Ajani, S. and Amdani, S.Y., 2020. Dynamic Path Planning Approaches based on Artificial Intelligence and Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *11*(3), pp.2084-2098.

[25] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 2721-2725.

[26] Zhu *et al.*, "Learning Environmental Sounds with Multi-scale Convolutional Neural Network," *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 2018, pp. 1-8.

[27] F. Yu, V. Koltun and T. Funkhouser, "Dilated Residual Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 636-644.

[28] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," in *IEEE Access*, vol. 7, pp. 125868-125881, 2019.

[29] Q. Liang, S. Xiang, Y. Wang, W. Sun and D. Zhang, "RNTR-Net: A Robust Natural Text Recognition Network," in *IEEE Access*, vol. 8, pp. 7719-7730, 2020.