Automated Voice-to-Image Generation Using Generative Adversarial Networks in Machine Learning

Lakshmi Prasanna *Yeluri*[1*], *G.*Ramesh[1], Vijayalata *Y*[3] , *Khaja* Shareef[4], Shailesh *Chamola*[2] and Mallikarjuna Rao *Gundavarapu*[1]

[1]Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, India.
[2]Uttaranchal Institute of Management, Uttranchal University, Dehradun, India.
[3]KG Reddy College of Engineering, Hyderabad, India.
[4]Department of CSE, MLR Institute of Technology, Hyderabad, India.

Abstract. Creating visuals from words may appear to be a complex process, but it is achievable with today's technological advancements in Information Systems. Naturally, all the human-centric actions and assumptions may lead to visualization using Artificial Intelligence. In today's Information Systems technological world, any item or a thing can be best described in pictorial form as a human person. Our paper aims to focus on providing machines with this intelligence. To complete this challenge, we used Natural Language Processing with Deep Learning. Our primary focus is on Generative Adversarial Networks. GANs will generate data based on word labels that are provided. NLP is also important since it helps to translate the provided speech into embedding vectors that the model can use. Our study is on the CUB dataset, which comprises bird photos. In today's world, there are text-to-image generating models accessible. The authors investigated all of them, extending text-to-image generation to voice-to-image generation.

1 Introduction

Several types of research in Artificial Intelligence (AI) are making the human lifestyle utterly different from previous generations. Some of the inventions are automated cars, voice assistants, and robots in the workplace. For all of them, neural networks are the core concepts and are achieved by Supervised Machine Learning Techniques using neural networks like DNN, CNN (vision), and RNN (voice). While coming to Generative Adversarial Networks (GAN) [1, 2], it is an unsupervised machine learning modules [23] whose key point is to create new things from the observed patterns. Most of the time, humans envision numerous things in the world that do not exist and photographing it is impossible. Painting and VFX design are two options for presenting what we have in mind, and need a significant amount of time and work. Authors' intend to build an AI system that

*Corresponding author: prasanna.yeluri@gmail.com

outputs picture by specifying their attributes. For example, if we want to see a purple bird, we may tell the model and it will generate the image for us.

This is employed by using generative adversarial networks [1], where they can create visuals that are as excellent as our words [10,11,13,14]. It takes a long time, almost days, for a single individual to sketch or produce something. So, if the model is correctly trained, it can create fictitious graphics in seconds. The model should be thoroughly trained so that our task becomes more straightforward, and thus the result is expected to be more accurate. As a result, our final model will serve as a virtual picture creator or painter for everyone that will come in handy in this creative world. We anticipate that it will be helpful for the manufacturing departments [26,27] in developing new styles and designing new animations for planning and testing alternative wall colors in the bedroom, etc.

Our technology has advanced significantly and there are several text-to-image generation models [10,11,13,14] that uses text as input to produce images. Recently, speech-based work [16] has been developed that construct a person's visage based on their voice. Though it is a decent one, it just deals with the acoustics of the voice and does not consider the language. Our study focuses on the language of spoken text in the English language. We added the voice element to one of the text-to-image generating methods, AttnGan. We are converting the voice given by user into text and feeding the text as the input to our revised AttnGAN model for the speech, as shown in section 3

2 Related Works

As previously said, there are several efforts on text-to-image generation [10,11,13,14]. All of these are accomplished through generative adversarial networks [1,2], as proposed by Ian Goodfellow in his paper Generative Adversarial Networks [1]. Generative Adversarial Networks are neural networks that comprise many different neural networks, such as Convolutional, Dense, Recurrent, Upsampling, and Downsampling networks. It is unsupervised learning extension to autoencoders in which we produce high-dimensional images in the low-dimensional latent area called encoding and from the latent space to the image called decoding. Both encoder and decoder are present in GAN's Generator, one of two Competitive neural Networks in the GAN. Another neural Network, discriminator determines if a created image is genuine or fake. Another study [2] that enhanced GAN training approaches by employing alternative loss functions for the generators and discriminator is improving techniques for training the networks. Initial GAN's are extremely difficult to train because they lack a precise moment as a result of which cease training. Convolutional neural networks failed to execute unsupervised learning despite performing exceptionally well in supervised image processing and computer vision. Unsupervised Representational Learning with Deep Convolutional Generative Adversarial Networks [7] is another paper that presented DCGAN, which is a CNN for unsupervised data. It categorized the unlabeled images efficiently. Because the GAN train is crucial, they modified the Convolutional GAN to stabilize the training. They also utilized pre-trained image classification models in Discriminators. They also demonstrate that Generator learns certain filters to create visual output.

GANs can generate pictures without labels, but the image generation techniques utilizing text descriptions are not as accurate as a two-stage process. As an alternate, StackGAN [9,10,11] is suggested which can generate photo-realistic visuals. The initial step of stackGAN or StackGAN-1 sketches rudimentary shapes and colors from the provided text. However these are low-resolution images with poor quality. As a result, the second step of StackGAN-2 creates high-quality images based on the images created in the first stage and utilizing text descriptions. A color-consistency regularization for the multi-distribution

approximation is used in this study to enhance both conditioned and unconditioned picture production.

Another work on text-to-image generation is carried-out by using the Attentional Generative Adversarial Networks [12], where along with the text there are separate word vectors generated and checked with each image latent space of three generators in the architecture. It has three generators connected one after another same like the stackGAN and each generator is associated with a discriminator that produce 64*64*3,128*128*3,256*256*3 images in the three stages. First generator can draw the image as given in the text and the next two uses word labels and attentional neural networks to improve the quality of images. They have excellent inception scores compared to previous text-to-image generations. It has a Deep Attentional Multimodal Similarity Model (DAMSM), which compares word labels with image labels because the exact image according to the text is formed.

Style-based Generative Adversarial Networks [29] are another text-to-image generation model that uses the AttnGAN model in their work. They improved the quality of images produced by the AttnGAN with adding neural network after the AttnGAN. The new neural networks take the final output of the attentional generator's generated image and after some processing and styling the image quality is high in comparison to the prior high-quality image of the AttnGAN Generator.

All the text-to-image generation models discussed above are multi-stage models, including stackGAN and AttnGAN. This results in increasing the training time and quality of the images created in the early stages is ignored. Another aspect to consider is that training more discriminators also requires more time. Dual Attentional Generative Adversarial Networks (DTGAN) [13] uses one generator with six levels and one discriminator. It also includes channel-based and pixel-based attentional models to give more weight to essential words, such as color in text descriptions.

Even though single-stage models produce high-quality images from text images, they suffer from a lack of variation. They repeatedly produced the exact output for the same comparable text. So DiverGAN [14] is developed to deal with diversity difficulties by using a fully connected layer to minimize the diversity problem.

Recently speech-based [16] work produced the person's visage depending on the given voice. However, it is entirely depends on the tendency of the voice rather than natural words. Speech-to-image generation model [15] has also been developed recently by using the CUB and Oxford-102 datasets as well as text descriptions [19], to generate the images. Using the tacotron2 module [21], they transformed the collected text descriptions of the photos into speech which is used in the training process and also includes a speech embedding network.

3 The Proposed Method

Our task is to create high-quality images for the supplied speech input. Providing voice as an input and converting it to text initially utilizes speech-to-text processes and next passes the resulting text to the text-encoder that will encode the voice to the relevant single dimensional vectors. In order to produce high-quality photographs [24,25], authors' employed Attentional Generative Adversarial Networks in the work.

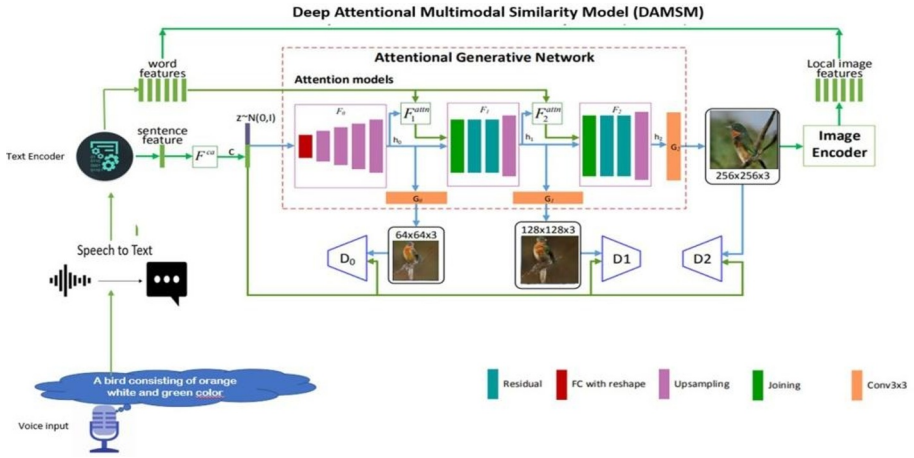

Fig. 1. Architecture of the proposed model using AttnGAN framework.

3.1 Dataset

We are working on the CUB [8] dataset. It contains 11,788 images of 200 subcategories belonging to birds, 8,855 for training and 2,933 for testing. Preprocessing on the dataset has done as [11]. Each image has ten labeled descriptions from the work [18]. Since there is no speech data for the dataset, we are converting voice into the text at initial stages.

3.2 Voice to Text Conversion Process

There are many models that can translate text to images; however, not every language in the world has text. As a result, this paper focuses on the voice-to-image conversion mechanism in one of the world languages (English). Using Python's speech recognition techniques, the specified audio frequencies are converted into text descriptions using built-in APIs. As stated in the work suggested using the tacotron2 [21] module to translate all of the text to voice, this paper focuses in the reverse manner.

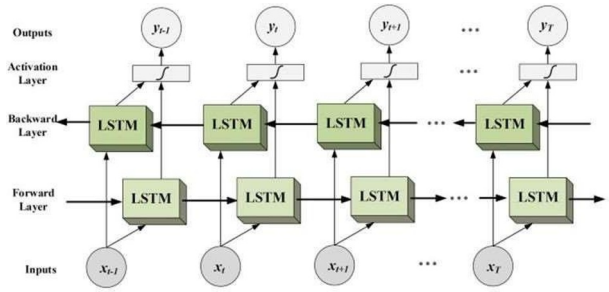

Fig. 2 Bi-directional Long Short-Term Memory Neural Network Structure

We captured the speech of all the labels and visual descriptions and used it to validate our model. The model is trained on the text to save time as speech recognition modules can be used for validation. Deep learning is the internal notion involved in speech-to-text

conversion in a sequence and recurrent neural networks are the ideal choice. However long sequences results in issues such as vanishing and exploding gradients.

Long Short-Term Memory (LSTM) is a recurrent neural network that deals with the aforementioned issues. Using bi-directional Long Short-Term Memory (BiLSTM) for accurate speech-to-text conversion helps the model in text generation by preserving the data for long time. As human voice depends on its frequency and each letter of the English alphabet has phonetic sounds and its frequency, there is a chance of overlap such as red and read. So while converting to text, the model has also concentrated on the natural language along with the acoustics of speech and the base of sound. Thus combining the acoustic model with a language model improves the efficiency of our model.

3.3 Text Encoder

In this phase, the text generated by the speech-to-text method is taken and converted into an encoded single-dimensional vector. The text encoder is made up of a bi-directional Long Short Term Memory that has a forward and backward layer for each phase, as shown in Fig.3. There are two hidden states for each layer, one from the direct input voice description and the other from the reverse of the given voice description. To construct the encoding vector, we merged the two hidden states and a global word vector is used as direct input to the Attentional Generative Adversarial Network to generate the images. The deep attentional multimodal similarity model utilizes a word vector (DAMSM).

3.4 Attentional Generative Model

All previous efforts considered the global sentence vector but ignored the impact of solid terms. However, in the attentional generative model, all of the essential terms are assigned to the word vector, and we use the word vector to create all of the image's subfields by giving more importance to the terms. As we can see in Fig.2, there are three generator's (G0, G1, G2) in the architecture of our model and each generator has hidden states (h0, h1, h2) that are used for generating images at the three stages respectively. There are three discriminators in architecture (D0, D1, D2). Each generator produces the image (x0, x1, x2). F0 is the first neural network in the model that takes random noise z as input along with the output of Fca, which is conditioning augmentation. Fca takes sentence vector $\bar{e}$ as input and converts it into the conditioning vectors as shown in the equation (1).

$$h_0 = F_0\left(z, F^{ca}(\bar{e})\right) \quad (1)$$

For the neural networks F1 and F2, hidden output from the previous stage is taken as the input along with the output of F_i^{attn} where i=1,2. F_i^{attn} take two inputs hi-1 and e, where e is the matrix of essential words and hi-1 is output from the previous neural networks. Fi-1 is as shown in equation (2).

$$h_i = F_i\left(h_{i-1}, F_i^{attn}(e, h_{i-1})\right) \qquad for\ i = 1,2 \quad (2)$$

F_i^{attn} is the attentional model. The word features e converted to the image features by adding a new perceptron layer. Here the word context vector is assigned for each sub-region of image vector h. Both the image vector and word context vector are used in generation of the image. Each column of the image vector is associated with a sub-region of the image. The hidden states hi obtained at each stage from equation (2) are provided as input to the generator as shown in the equation (3).

$$\widehat{x^i} = G_i(h^i) \quad for\ i = 1,2,3 \quad (3)$$

All the images are generated by the specified conditions at sentence-level and word-level vectors. So, the final objective function of the attentional generative adversarial network is as shown in equation (4).

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM} \quad where\ \mathcal{L}_G = \sum_{i=0}^{2} \mathcal{L}_G \tag{4}$$

Here, hyperparameter λ balances the $\mathscr{G}$G and $\mathscr{D}$DAMSM in equation (4). The loss at each generator Gi in the Attentional Generative Network is a combination of conditional and unconditional losses, which is given in equation (5).

$$\mathcal{L}_{G_i} = -\frac{1}{2}\mathbb{E}_{\hat{x}\sim\rho G_i}[\log D_i(\hat{x}_i)] - -\frac{1}{2}\mathbb{E}_{\hat{x}\sim\rho G_i}[\log D_i(\hat{x}_i, \vec{e})] \tag{5}$$

Here the conditional loss determines whether the images and the words are matched or not and the unconditional loss decides whether the image is actual or not. Now to classify whether the image is real or fake, discriminators are also to be trained. The loss function for each discriminator Di is given in equation (6).

$$\mathcal{L}_{D_i} = -\frac{1}{2}\mathbb{E}_{\hat{x}\sim\rho data_i}[\log D_i(x_i)] \ -\frac{1}{2}\mathbb{E}_{\hat{x}\sim\rho G_i}[1 - \log D_i(\hat{x}_i)] +$$
$$-\frac{1}{2}\mathbb{E}_{\hat{x}\sim\rho data_i}[\log D_i(x_i, \vec{e})] - \frac{1}{2}\mathbb{E}_{\hat{x}\sim\rho G_i}[1 - \log D_i(\hat{x}_i, \vec{e})] \tag{6}$$

where $\hat{x}_i$ is taken from the model distribution ρG_i at the same scale as x_i from the genuine image distribution $\rho data_i$ at the scale i. Discriminators are fundamentally separate, allowing them to be trained concurrently and with a single image scale as their primary emphasis.

The second term of equation (4) is the loss function from the DAMSM model for comparing the generated image with given text. To calculate a fine-grained loss for image production, the Deep Attentional Multimodal Similarity Model first learns two neural networks that map sentence words and image sub-regions to a shared semantic space. The text encoder and image encoder are the two neural networks. The text encoder with a bi-directional LSTM text encoder extracts semantic vectors from the text description is used. The global sentence vector is created by concatenating the bi-directional LSTM's final hidden states. The image encoder converts images into meaningful vectors by using CNN using the ImageNet-pretrained Inceptionv3[32] model and the global feature vector is retrieved from the average pooling layer. Finally a perceptron layer is added to transform the image features into a shared semantic space of text features.

4 RESULTS AND DISCUSSION

4.1 Qualitative Results

The output of the model using various color combinations with supplied voice input such as "A small color-1 bird with color-2 belly". Here color-1 and color-2 are replaceable colors in Fig. 3. The model uses six different colors (Black, Blue, Red, Yellow, White, and Orange) to generate the images of nine birds with different color combinations. The images generated have the same quality as that of an actual photograph. As the birds are generated using inputs and weights from the model, they cannot be distinguished from real-world pictures.

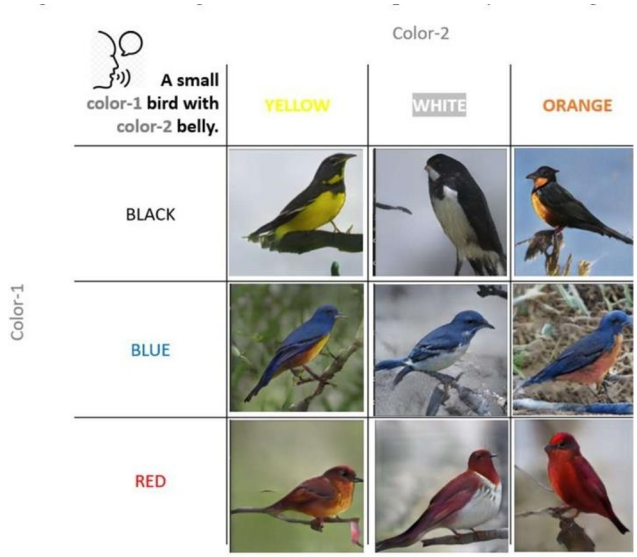

Fig. 3. Outputs generated by the model for the given speech with the different combinations of colors

4.2 Comparative Results

The proposed model outputs are generated and compared with the previous works as shown in Fig. 4. The previous works includes images generated by models such as StackGAN (T2IG, S2IG) and S2IGAN. The results depicted proves that our model utilizes an enhanced voice recognition system to convert the inputs received to text and produces high-quality images that cannot be distinguished from the real-world photographs clicked by a person.

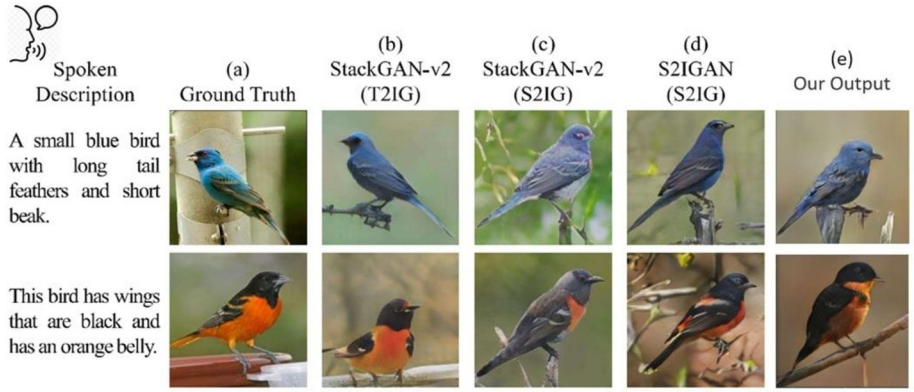

Fig. 4 Outputs compared with the previous works.

4.3 Quantitative Results

The model is compared with previous state-of-art GAN models like text-to-image and voice-to-image generation. The inception score for our model is 4.36 + .03 which is 3.82+ .06 for stackGAN-V2, 3.70 + .04 for stackGAN-V1, 4.04 +.04 for StackGAN++ [11], 3.62 + .07 for GAWWN [17] and 2.88 + .04 for GAN-INT-CLS [19].

5 CONCLUSION

In this paper, authors developed a speech-to-image generation model based on attentional generative adversarial networks that creates high-quality images for the input words provided. Authors worked solely on the CUB dataset which comprises photos of birds. For speech-to-text conversation, bi-directional LSTM model is used concentrating on both acoustic and natural language models and resulted in generating high-quality bird photos for the input voice provided. The main limitation of the model is that it focuses only in generating visuals in English language not for other languages. Another limitation is that there is a chance of random image generation in case of improper voice input. In future the work can be expanded by including various languages by training a language translator model and a language recognition model to recognise the input language and convert it to English. Another enhancement is to improve the model's performance for a generalized data provided as input.

References

1. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. *In NIPS,* (2014).
2. T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *In NIPS,* (2016).
3. K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. *In ICML,* (2015).
4. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *In CVPR*, (2017).
5. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super resolution using a generative adversarial network. *In CVPR*, (2017).
6. A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR,* (2017).
7. A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *In ICLR,* (2016).
8. C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS- TR2011-001, *California Institute of Technology,* (2011).
9. Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. *In CVPR,* (2016).
10. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *In ICCV*, (2017).
11. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916, (2017).

12. T Xu, P Zhang, Q Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. *In CVPR,* (2018).
13. Zhenxing Zhang, Lambert Schomaker. DTGAN: Dual Attention Generative Adversarial Networks for Text-to-Image Generation. arXiv:2011.02709,(2020).
14. Zhenxing Zhang, Lambert Schomaker. DiverGAN: An Efficient and Effective Single-Stage Framework for Diverse Text-to-Image Generation. arXiv:2111.09267,(2021).
15. Xinsheng Wang, Tingting Qiao, Jihua Zhu, Alan Hanjalic, Odette Scharenborg: S2IGAN: Speech-to-Image Generation via Adversarial Learning. arXiv:2005.06968, (2020).
16. Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Wojciech Matusik: Speech2Face: Learning the Face Behind a Voice. arXiv:1905.09773,(2019).
17. S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. *In NIPS*, (2016).
18. S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. *In CVPR*, (2016).
19. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. *In ICML*, (2016).
20. S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. *In ICML,* (2017).
21. J Shen, R Pang, Ron J. Weiss, M Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv:1712.05884
22. A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. VQA: visual question answering. IJCV, 123(1):4–31, (2017).
23. G. Ramesh et al., "Feature Selection Based Supervised Learning Method for Network Intrusion Detection", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-**8**, Issue-1, May (2019).
24. Y. Sara, J. Dumne, A. Reddy Musku, D. Devarapaga and R. Gajula, "A Deep Learning Facial Expression Recognition based Scoring System for Restaurants," *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, pp. 630-634, doi: 10.1109/ICAAIC53929.2022.9793219.* (2022)
25. Ramesh, G., Anugu, A., Madhavi, K., Surekha, P.. Automated Identification and Classification of Blur Images, Duplicate Images Using Open CV. *In: Luhach, A.K., Jat, D.S., Bin Ghazali, K.H., Gao, XZ., Lingras, P. (eds) Advanced Informatics for Computing Research. ICAICR 2020. Communications in Computer and Information Science, vol 1393. Springer, Singapore.* https://doi.org/10.1007/978-981-16-3660-8_52, (2021)
26. G. Ramesh, J. Praveen, Artificial Intelligence (AI) Framework for Multi-Modal Learning and Decision Making towards Autonomous and Electric Vehicles, *E3S Web Conf. 309 01167, DOI: 10.1051/e3sconf/202130901167* (2021)
27. Parameswari, D.V.L., Rao, C.M., Kalyani, D. et al. Mining images of high spatial resolution in agricultural environments. Appl Nanosci ,(2021). https://doi.org/10.1007/s13204-021-01969-3

28. Somasekar, J Ramesh, G "Beneficial Image Preprocessing by Contrast Enhancement Technique for SEM Images", IJEMS Vol.**29**(6) [December 2022], NIScPR-CSIR,India, (2022)