iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** I **Month of publication:** January 2021

DOI: https://doi.org/10.22214/ijraset.2021.32873

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue I Jan 2021- Available at www.ijraset.com

A Review on Gesture-To-Voice Recognition System based on Facial Expressions & Hand Gestures

Rohit Jain[1], Saurav Kabra[2], Tapendu Bera[3], Prof. Stevina Dias[4]
[1, 2, 3, 4]*Department of Information Technology, D. J. Sanghvi College of Engineering, Mumbai, India*

Abstract: ***Voice is a medium through which one communicates ideas, interest, personality and alot more. Deaf-mute people use sign-language to communicate with others but some people don't know sign-language in order to communicate back. Gesture recognition is a growing field which helps those who aren't able to talk but rather use gestures in order to communicate with people around them. In this paper, an in depth research is performed on various systems that are developed for the various features for making a robust system that helps mute people to communicate more freely and easily using computer vision and natural language processing.***

Keyword: CNN, LSTM, SVM, Hand Landmarks, Face Landmarks, Emotion Detection

I. INTRODUCTION

One's voice is one of the most powerful tools. One can share big ideas, thoughts, and even speak with people across the world by merely speaking with one another. However, some of us aren't able to use their voices to carry out these essential tasks. However they can communicate easily with writing but this creates a problem and suffers a lot of face to face communication. This problem can be minimized using sign language. We realized that people who use sign language as their primary means of communication could be given a method to talk with those who don't know the complicated gestures behind this language.

There are between 138 to 300 recognized sign languages across the world. Such diversity in languages has its own challenges when it comes to communicating across different communities, villages, cities and states. Indian Sign Language is one of the living languages in India used by deaf and mute communities. The number of deaf people in India accounts to 1.3 million (5.8 percent of total 21.9 million people with disabilities in India according to census). Educational system in India first focused on oral-aural methods. But now situations are improving with use of Indian Sign Language (ISL).

II. LITERATURE SURVEY

A. Smart Glove-based Hand Gesture

This system is an IOT based system. It uses data gloves to recognize the gestures made by hands to denote the English alphabets. The data gloves used in this system provides highly accurate responses as described in the paper[1], but the overall system is a bit expensive. The data glove is embedded with flex sensors along each finger as shown in Fig. 1. The output of the flex sensors measures the degree of bend of the fingers. The analogue outputs from the sensors are then transmitted to the microcontroller which processes the signals and performs analogue to digital signal conversion. This is followed by gesture recognition and further corresponding text identification.

The user needs to know the signs of particular alphabets and he needs to stay with the sign for two seconds. There are no limitations for signs but the new sign introduced should be supported by the software used in the system. This pair of gloves along with sensors enables mute people or old people to interact with the public which is very much helpful for them.

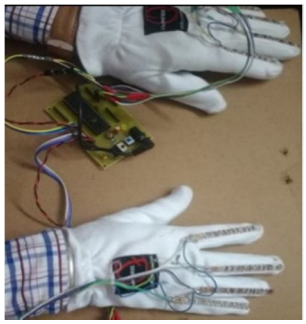


Fig. 1: Smart Gloves

B. Gesture Recognition Using Color Gloves

This method presents a real-time hand gesture recognizer based on a colour glove as shown in Fig. 2. The recognizer is formed by three modules - the first module, the frames are acquired by a webcam followed by segmentation which involves converting image to HSI color-space, identifying glove-colors & morphological operations resulting in identifying the hand image in the scene; the second module, a feature extractor, represents the hand-image by a nine-dimensional feature vector which are invariant wrt rotation and translation of image plane; and the third module, the classifier, is performed by means of Learning Vector Quantization (LVQ) which is a prototype-based supervised classification algorithm - a precursor to self-organizing maps (SOM).

Fig. 2: Color-glove

As you can see in Fig. 2, the gloves used in the paper[2] have been coloured the palm by magenta and the fingers by cyan and yellow. Further investigations seem to show that the above mentioned choice of colors does not affect remarkably the performances of the recognizer but if colors used in gloves are also present in the background, it will affect the performance. In terms of costs, they are way cheaper and affordable.

C. Bare-Hand Gesture Recognition

This method makes use of Region-based Convolutional Neural Network which focuses on the recognition and localization of hand gestures. The paper[3] majorly focusses on 2 types of gestures: open and closed hand & recognition of these gestures in dynamic backgrounds as shown in Fig. 3.

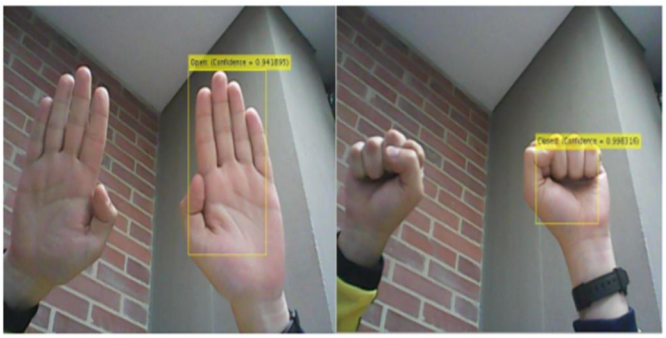


Fig. 3: Open & Close Hand Gesture

In this method, first the pre-recorded video is converted into frames or images and each frame is processed for detecting hands in it. If hands are found in the frame, it is then worked upon to check if the hands are closed or open. The recognition of hand gestures are done using RCNN which makes use of segmentation of different parts of the image by means of a region proposal algorithm. The results of RCNN are passed to a SVM classifier and linear regression model which are trained for detecting the class (hand gesture) & getting tighter bounding boxes respectively. This approach can be used to detect static gestures (which can be recognized using a single frame) but not dynamic gestures (which requires multiple continuous frames).

D. Emotion Recognition Model Based on Facial Recognition

The paper[4] proposes an Emotion recognition model using HAAR Cascades - a machine learning object detection algorithm used to identify objects in an image based on the concept of features to detect mouth and eyes identify six emotion classes through the deep neural networks. The paper[4] compares this HAAR cascades (HC) method with another method of sobel edge (SE) detector and proves to get better results as shown in Table 1.

Table 1: Avg. accuracy(%) for each emotion

	Sad	Surprise	Happy	Anger	Disgust	Fear
HC	78.54	93.26	95.25	91.22	84.32	82.58
SE	76	87.72	94	87.66	82.76	79.73

E. Facial Expression Recognition Using Attentional Convolutional Network

The method makes use of a new framework for facial expression recognition i.e., an attentional convolutional network. The model proposed in this paper[5] makes use of a localization network for attention mechanism as shown in Fig. 4. Here, attention plays an important role for detecting facial expressions as it enables neural networks with not more than 10 layers to compete with much deeper networks for emotion recognition. This is because not all parts of the face image plays an important role in emotion recognition as shown in Fig. 5. This method showed promising results on four popular facial expression recognition databases described in the paper[5] with extensive experimental analysis.

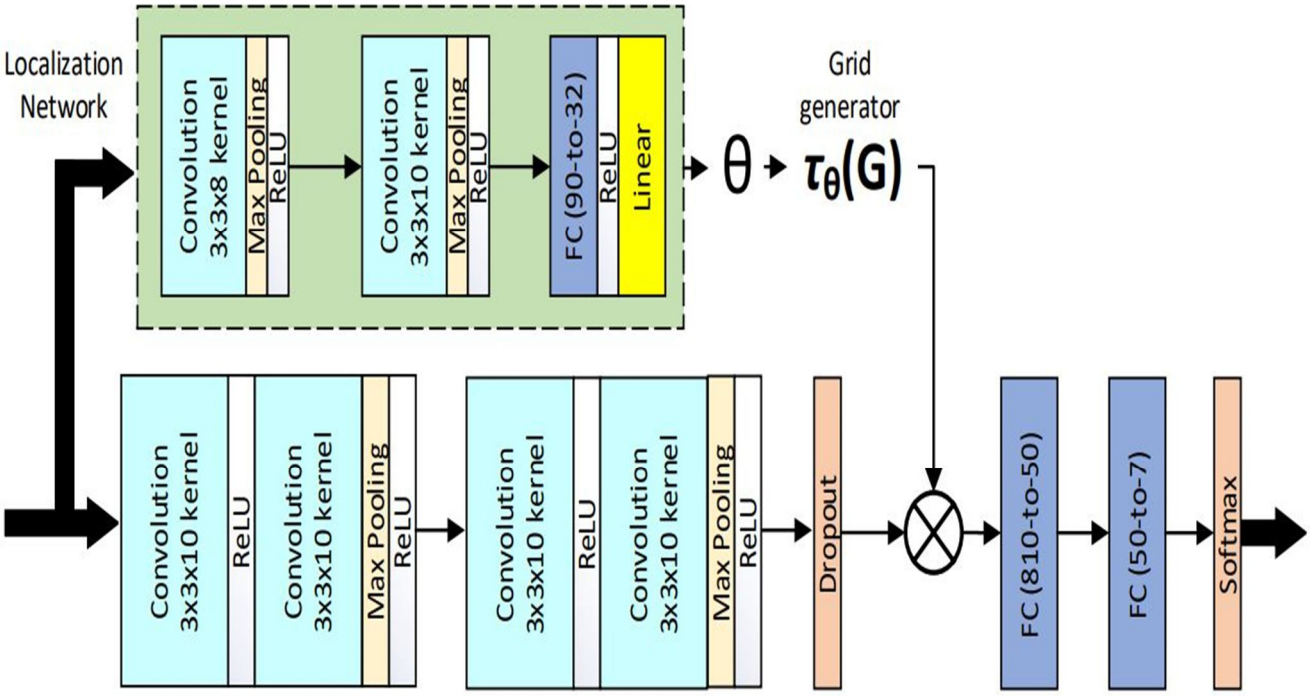


Fig. 4: Attention Convolutional Network

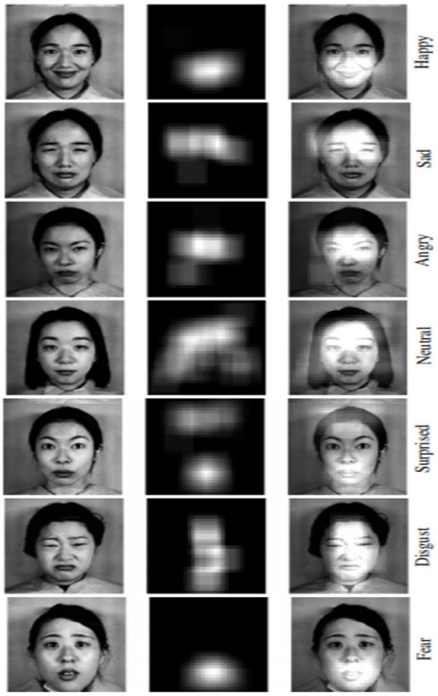


Fig. 5: Important face regions for detecting emotions

III. PROPOSED ARCHITECTURE

Human communication through gestures are not constrained only to a single instance of hands. Hand gestures include a sequence of instances of hand movements along with their facial expressions for communication. As seen in the literature survey, hand gestures were the only way one could communicate in those systems, which curtails the human expression as a whole.

The purpose of the project is to make gesture based communication more natural by considering sequences of hand gesture instances and facial expression together as shown in Fig. 6.

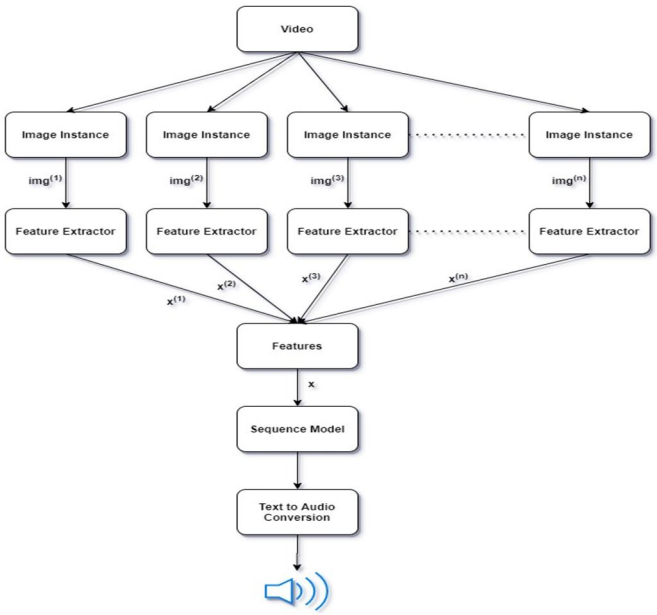


Fig. 6: Proposed System Architecture

As shown in the Fig. 6, first the video is divided into sequence of image instances at a certain interval rate and then each extracted image instance is passed to a feature extractor module for extracting features.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue I Jan 2021- Available at www.ijraset.com

The feature extractor module as shown in the Fig. 7, takes image as an input and passes it into two different modules - first is hand recognition followed by hand landmarks detection module and second is face recognition followed by face landmarks detection module. Both these modules can be implemented using an object detection module like Single-Shot MultiBox Detector (SSD)[6] with some minute changes to extract landmarks i.e., pre-defined distant unique points of a specific object.

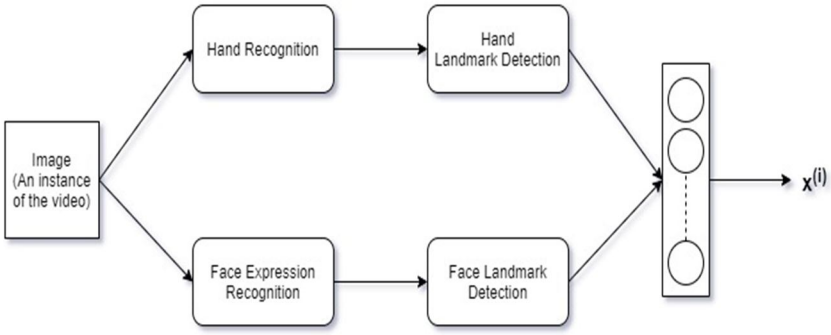


Fig. 7: Feature Extractor Module

The features of both the modules are stacked together to get a feature vector $x^{(i)}$ as shown in Fig. 7, which forms the output of the feature extractor module.

The outputs of the feature extractor module for each image instance are stacked together and passed to the next module i.e., sequence model.

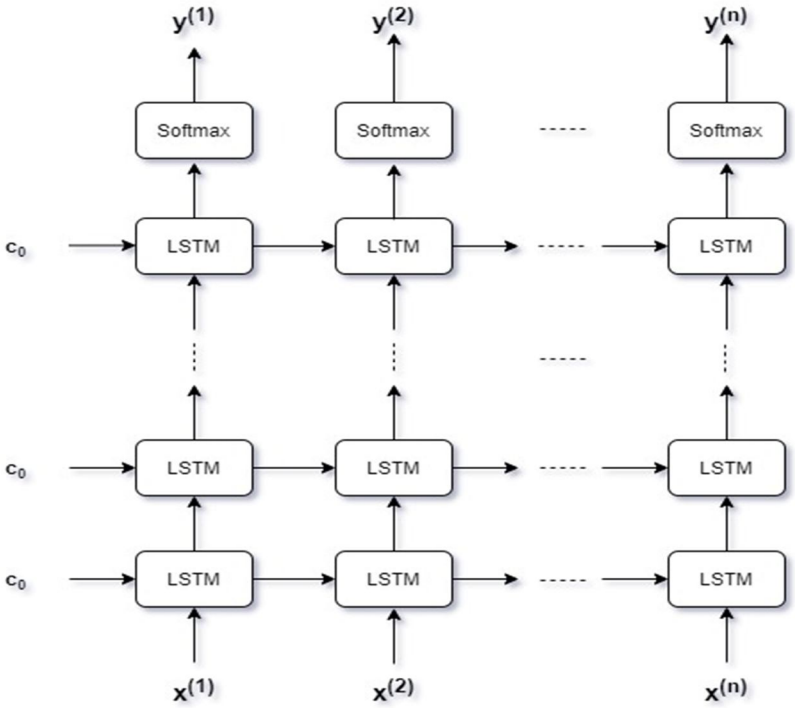


Fig. 8: Sequence Model

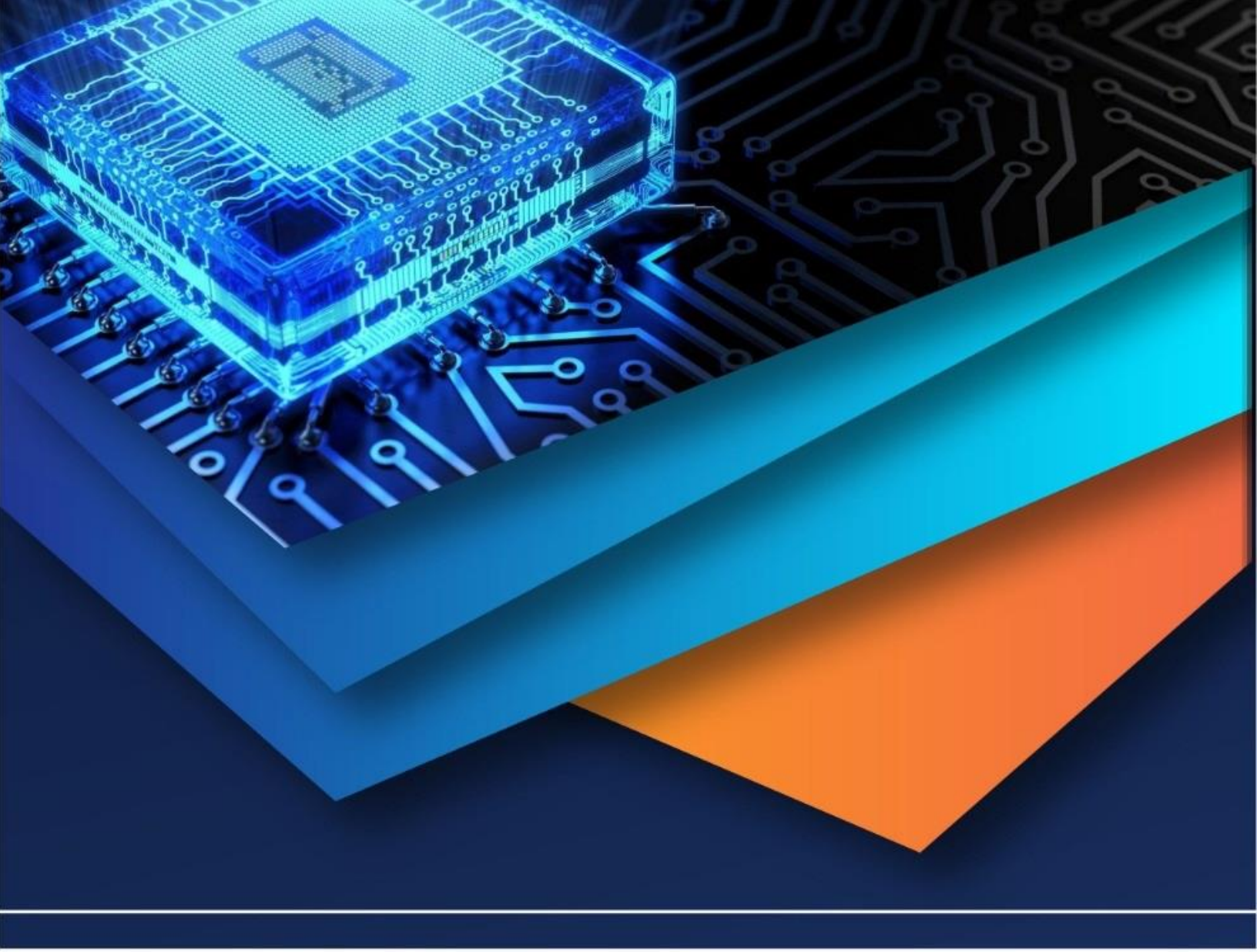
As shown in Fig. 8, the sequence model consists of Deep-LSTM neural networks - a set of sequences of Long Short-Term Memory (LSTM)[7] stacked over each other. LSTM has feedback connections unlike standard feedforward neural networks. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). The final layer is used as a softmax layer to get the probabilities of the text labels as the output i.e., $y^{(i)}$. The text label associated with maximum probability will be converted to audio to get our project's final result using Google Translate's text-to-speech API (gTTS).

IV. CONCLUSION

The report gives an analysis of various recognized systems, their approaches and methodologies, algorithms, technologies that have been used earlier and also provides their corresponding results. All these systems had some limitations in terms of special hardware requirements like IOT-based gloves or color-gloves or were restricted to static gestures. After having the detailed analysis of the systems, the paper proposes a system that provides a better way to make a gesture to voice recognition system which doesn't have any special hardware requirement. This review paper helps us in understanding the basic building blocks of the gesture recognition system that we would be developing in future, using our own approach and some of the existing solutions.

REFERENCES

[1] Purohit, Kunal. (2017). A Wearable Hand Gloves Gesture Detection based on Flex Sensors for disabled People. International Journal for Research in Applied Science and Engineering Technology. V. 1825-1833. 10.22214/ijraset.2017.8261.
[2] Lamberti L., Camastra F. (2011) Real-Time Hand Gesture Recognition Using a Color Glove. In: Maino G., Foresti G.L. (eds) Image Analysis and Processing – ICIAP 2011. ICIAP 2011. Lecture Notes in Computer Science, vol 6978. Springer, Berlin, Heidelberg
[3] J. Sun, T. Ji, S. Zhang, J. Yang and G. Ji, "Research on the Hand Gesture Recognition Based on Deep Learning," 2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE), Hangzhou, China, 2018, pp. 1-4.
[4] Yang, D. & Alsadoon, Abeer & Prasad, P.W.C. & Singh, A.K. & Elchouemi, A.. (2018). An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment. Procedia Computer Science. 125. 2-10. 10.1016/j.procs.2017.12.003.
[5] Minaee, S., & Abdolrashidi, A. (2019). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. ArXiv, abs/1902.01019.
[6] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., & Berg, A. (2016). SSD: Single Shot MultiBox Detector. ECCV.
[7] Greff, Klaus et al. "LSTM: A Search Space Odyssey." IEEE Transactions on Neural Networks and Learning Systems 28.10 (2017): 2222–2232. Crossref. Web.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY