iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.56672

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: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Intuitive Perception - Speech Recognition using Machine Learning

Mr. Kumar K[1], Sanket G Hegde[2], Shashikanth N G[3], Vishal N Korabu[4], Vummaneni Charan[5]
[1, 2, 3, 4] *Department of Computer Science and Engineering, K S Institute of Technology, Bengaluru, India*

Abstract: ***Machine Learning is widely used to detect the movement of lips. It has been observed that the data generated through visual stir of mouth and corresponding audio are largely identified. This fact has been exploited for lip reading and for perfecting speech recognition. We propose a system that uses a CNN(Convolutional Neural Network) which is trained and used to descry the movement of lips and prognosticate the words being spoken. This trained CNN will be suitable to descry the words that are spoken within the videotape and display it in a textbook format. The CNN may also calculate on fresh information handed by the environment, knowledge of the language, and any residual hail. We hope to learn whether the application of machine literacy, more specifically the DNN(Deep Neural Network), could also be an applicable seeker for working the problem of lip reading. The main end of our design is to directly honor the expressions being spoken through automated lip reading***
Keywords: Visual Speech Recognition, Lip reading, OpenCV, neural network, CNN, DNN, 3D convolutions, object detection, data pre-processing, Python, Keras.

I. INTRODUCTION

Visual lip- reading plays an important part in mortal- computer commerce in noisy surroundings where audio speech recognition may be delicate. The art of lip reading has colorful operations, for illustration it can be used to help people with hail disabilities, or conceivably by security forces in situations where it's necessary to identify a person's speech when the audio records aren't available. still, like audio speech recognition, lip- reading systems also face several challenges due to dissonances in the inputs, similar as with facial features, skin colours, speaking pets, and intensities, it's insolvable to manually produce a computer algorithm that will be reading from the lips. Indeed mortal professionals in this field can rightly estimate nearly every other word and can do so only under ideal conditions. thus, the complex task of lip reading is suitable seeker for expansive exploration in the field of deep literacy.

Lip reading is also an extremely delicate task because several different words can be spoken with nearly indistinguishable lip movements. thus, the problem of lip reading provides unique challenges. This has led to multitudinous advancements in the field of automated speech recognitions systems using machine literacy.

Several models have been developed to ameliorate hail aids, for silent dictation in noisy public surroundings, identification for security purposes etc. still not until the use of Deep Learning did the delicacy of these models increase. The use of Deep Learning and deep neural networks has revolutionised the quality of automated lip- reading systems due to the large quantities of data sets that can be used.

There are mainly four stages involved in the technique used to perform automated lip reading [1][6]. Namely, face detection, cropping module, feature extraction and text decoding. The primary goal of face detection algorithms is to check whether there is any face in an image or not. The cropping module is used to crop out the region of interest (in this case the lips) and feature extraction helps in extracting the required features. This is represented in the figure (Fig 1) given below:

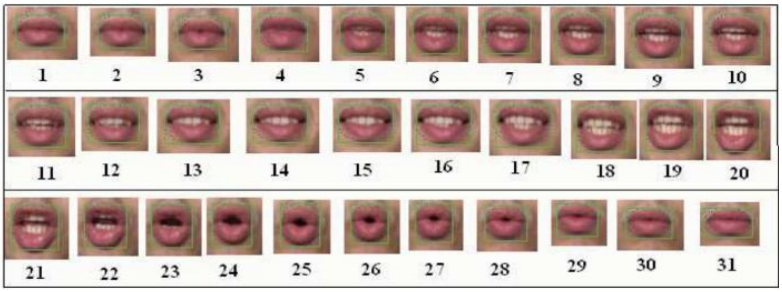


Fig. 1. Visualization of Lip Reading

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

II. RELATED WORK

The task of decoding text based on the movement of lips is complex. It requires complete and extensive training of the model to be able to recognise lip movements.

Wand and Jurgen Schmidhuber teamed up to create a lipreading system that's pretty cutting-edge. This system is end-to-end trainable, which means it can learn and improve on its own. What's cool is that it doesn't need a ton of frames with transcribed target data—just a tiny number does the trick. They've amped up the recognition accuracy for the target speaker by training it to be speaker-independent. They've thrown in a domain-adversarial twist to make the lipreader even more advanced. the target data—just a tiny number does the trick. They've amped up the recognition accuracy for the target speaker by training it to be speaker-independent. They've thrown in a domain-adversarial twist to make the lipreader even more advanced. The foremost goal was to push the network to an intermediate data representation which is domain-agnostic that is it should be independent whether data file is obtained from target speaker or a source speaker. TensorFlow's Momentum Optimizer is applied with the help of the stochastic gradient descent so on attenuate the multi-class cross- entropy hereby achieving optimization.

Brendan Shillingford and his team created a massive visual speech recognition dataset, clocking in at a whopping 3,886 hours of video featuring faces speaking. Their integrated lip reading system is a powerhouse, with a video processing pipeline that turns raw footage into stable lip videos and phoneme sequences. The scalable deep neural network then maps these lip videos to sequences of phoneme distributions, and a production-level speech decoder outputs sequences of words.

Lele Chen and team took a different route, fusing audio and image embeddings to generate multiple lip images simultaneously. They introduced a novel correlation loss to synchronize lip and speech changes, making their model robust to various factors like lip shapes, view angles, and facial characteristics. Their approach, outlined in "Lip Reading at a Glance," leverages the observation that speech correlates with lip movements across different individuals.

Now, onto Joon Son Chung and crew, who delved into sequence-to-sequence translator architectures for speech recognition. Their Watch, Listen, Attend, and Spell (WLAS) network transcribes mouth motion videos to characters. They implemented a curriculum learning strategy to speed up training and mitigate overfitting. Their dataset, "Lip Reading Sentences" (LRS), boasts over 100,000 natural sentences from British television. The unique aspect here is the model's ability to operate with dual attention mechanisms—over visual input only, audio input only, or both. With image and audio encoders and a character decoder, their model tackles the intricate task of lipreading, aiming to recognize phrases whether spoken with or without accompanying audio.

III. PROPOSED SYSTEM

We propose a system that will take in a videotape input from the stoner. This videotape is to bepre-processed and divided into frames of images. This is done to have non inclined values and to help honor the face in a better manner. The coming step will be to be suitable to descry the region of interest that's the mouth and crop it out. This cropped ROI is to be passed to the sophisticated neural network (CNN) for farther processing. Then the visual features are uprooted, and the model is trained, grounded on which the spoken words are decrypted. Figure 2 represents the inflow illustration of our proposed system.

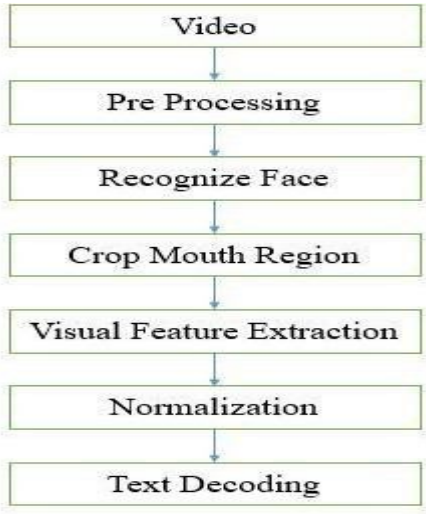


Fig. 2. Flow diagram

A. *Pre-processing*

Initially, the video is to be divided into frames of images. These frames of images obtained will most likely be in the RGB format. These images should then be converted to grayscale from RGB to avoid additional count of parameters present in an RGB image which is just an overhead to the system. The obtained set of frames from the video is then passed onto further processing [1][3].

B. *Face Detection and Cropping*

Once the frames have been obtained from the video, proposed system will detect the face in the frame if it exists and for the simplicity of our project, we are assuming that our system will be able to detect faces with full frontal view only discarding the possibility of having partial or side views of a human. We plan to make use of the DLib face detector and landmark predictor with 68 landmarks making use of the Hear features to be able to detect a face in the frame. After the detection of the face, the frames with no face will be discarded [6].

The coming step will be to be suitable to identify our Region of Interest(ROI) which is the lips and the mouth region in this case. It's to be linked with help of the haar waterfall classifier itself. Once the mouth region has been linked we will need to crop out the mouth region to be suitable to descry the mouth and the lip moment and for farther processing and training of our system. The RGB channels need to be standardised to have zero mean and unit friction.. After the process is done the images will be saved as a NumPy array with the cropped region images as values, it might look like the representation shown in Figure 4. The whole process of face detection and cropping is represented in Figure 3.

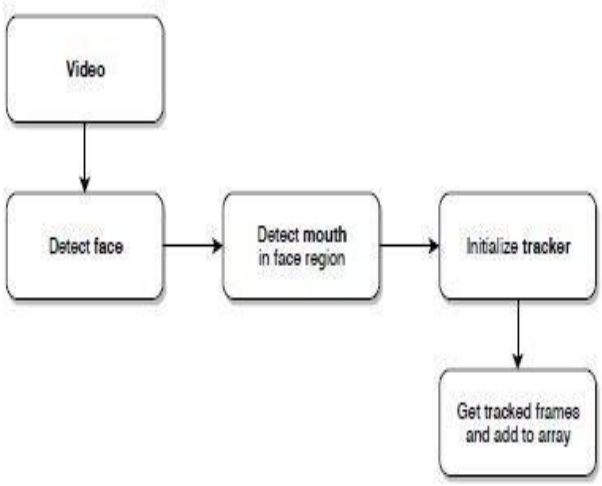


Fig. 3. Face detection and cropping process

Fig. 4. Saved NumPy array sample

C. Feature Extraction and Normalization

After the images are stored as an array the features from the ROI need to be extracted. The spatio temporal features need to be extracted and fed into the CNN as an input for training of the model [2].

Normalization of the image frames is necessary to avoid any irregularities in the dataset. For illustration, a person might take one alternate to gasp a word, while another existent may take two seconds to gasp the same word. Leaving similar irregularities unattended may beget disagreement in training and the results. So, we make use of normalization to be suitable to have an even training data.Normalization of the image frames is necessary to avoid any irregularities in the dataset. For example, a person might take one second to pronounce a word, while another individual may take two seconds to pronounce the same word. Leaving such irregularities unattended may cause discrepancies in training and the results.

D. Text Classification and Decoding

Once the normalization is done, the data will be fed into the CNN for training and text decoding. The CNN learns on its own by having many epochs and passing the information learnt among the multiple hidden layers. The decoding will be done by matching the lip movement with the image data and the given dataset used for training, the word spoken will be predicted [4].

The words spoken will then be embedded together for the whole video. The words prognosticated need to be put together to form the original sentence which was spoken by the existant in the dataset.

E. Architecture

The proposed system architecture is designed based on working of a Convoluted Neural Network(CNN). It is designed with an input layer, three hidden layers and an output layer. It also uses SoftMax subcaste as a probability classifier and maxmum pooling to reduce the number of parameters for the successive layers. The hidden layers consist of 32, 64 and 96 neurons in consecutive layers respectively. The system is tested using both 3 hidden layer architecture as well as the 5 hidden layer architecture but the 3-layer architecture is given more priority keeping in mind the computation problems for 5-layer architecture.

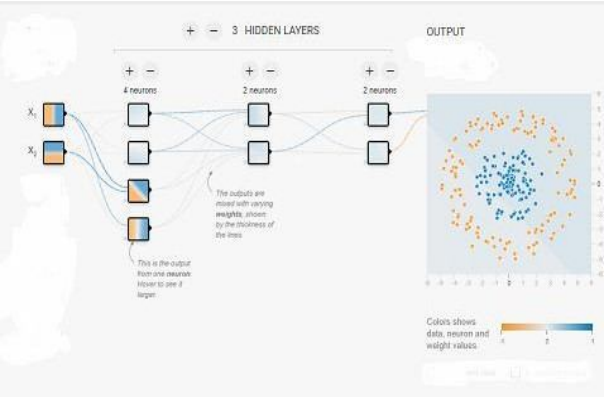


Fig. 5. CNN Basic Architecture

IV. EXPERIMENT

A. Development Environment

We have enforced the system on an Intel(R) core(TM) i7 CPU 2.6 GHz with 8 GB RAM and NVIDIA GeForce GTX 1650 (4 GB VRAM). The system ran OS Windows 10 Home. We have enforced the system in python 3.6 programming language.

OpenCV is the computer vision operation used for image processing and bracket. We have also used Keras, Microsoft Cognitive Toolkit, Theano.

If not specified otherwise, the model is trained with the following parameters:

1) Number of ages- 30 or ends if confirmation delicacy doesn't ameliorate after 4 consequent ages..
2) Literacy rate - 1 × 10-4.
3) Optimizer- Adam15.

The prognosticated set of words is displayed in resemblant with the videotape sample as subtitles which is shown in Figure 6.

B. Dataset

The GRID dataset consists of 34 subjects, each uttering 1000 phrases. The utterance of every word may be represented within the sort of verb (4) + color (4) + preposition(4) + alphabet (26) + digit (0-9) + adverb (4) ; e.g. 'put blue at A 1 now'. the full vocabulary size is 51, but the volume of possibilities at any given point within the affair is effectively constrained to the figures within the classes over. The videos are recorded during a controlled lab terrain, shown in Figure 7 [11].

Evaluation protocol. The evaluation follows the quality protocol and therefore the data is erratically divided into train, evidence and test sets, where the ultimate contains 255 utterances for every speaker. We report the word error rates. a number of the former factory report word rigor, which is defined as (WAcc = 1 - WER) [2].

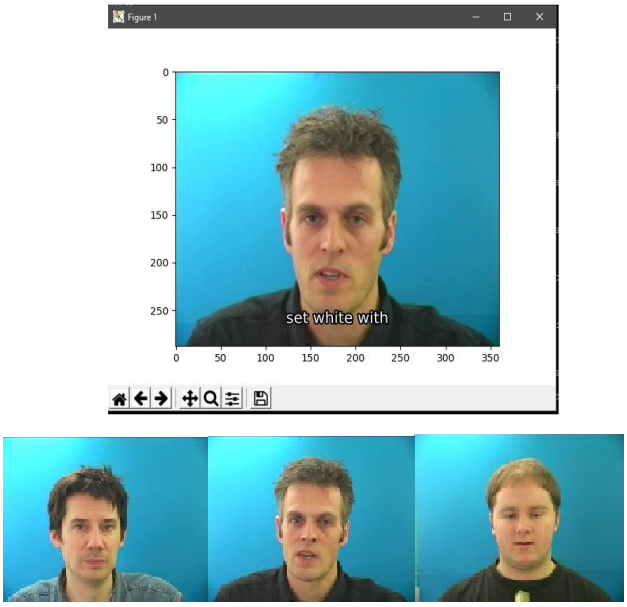


Fig. 7. Still images from GRID dataset

V. RESULTS

The system has been trained within GRID CORPUS dataset. The system shows variable delicacy between 70- 80 you choose the test dataset. The delicacy achieved is depicted in Figure 8 while comparing the kernel sizes. It's apparent from Table I that while adding the kernel size of CNN from 3X3X3 to 5X5X5 the delicacy increases significantly subject to number of ages.

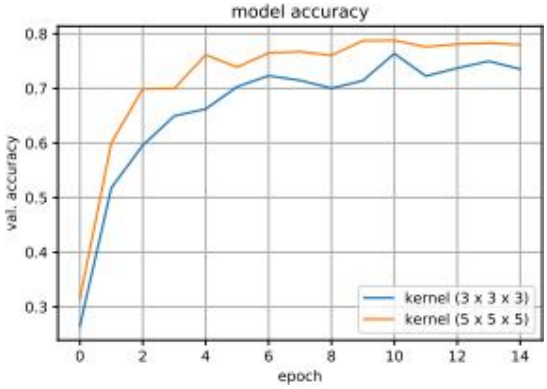


Fig. 8. Model Accuracy

TABLE I. KERNEL SIZE AND ACCURACY

Kernel Size	***Dropout***	***Epochs***	***Accuracy***
3X3X3	NO	14	75.84
5X5X5	NO	18	79.52

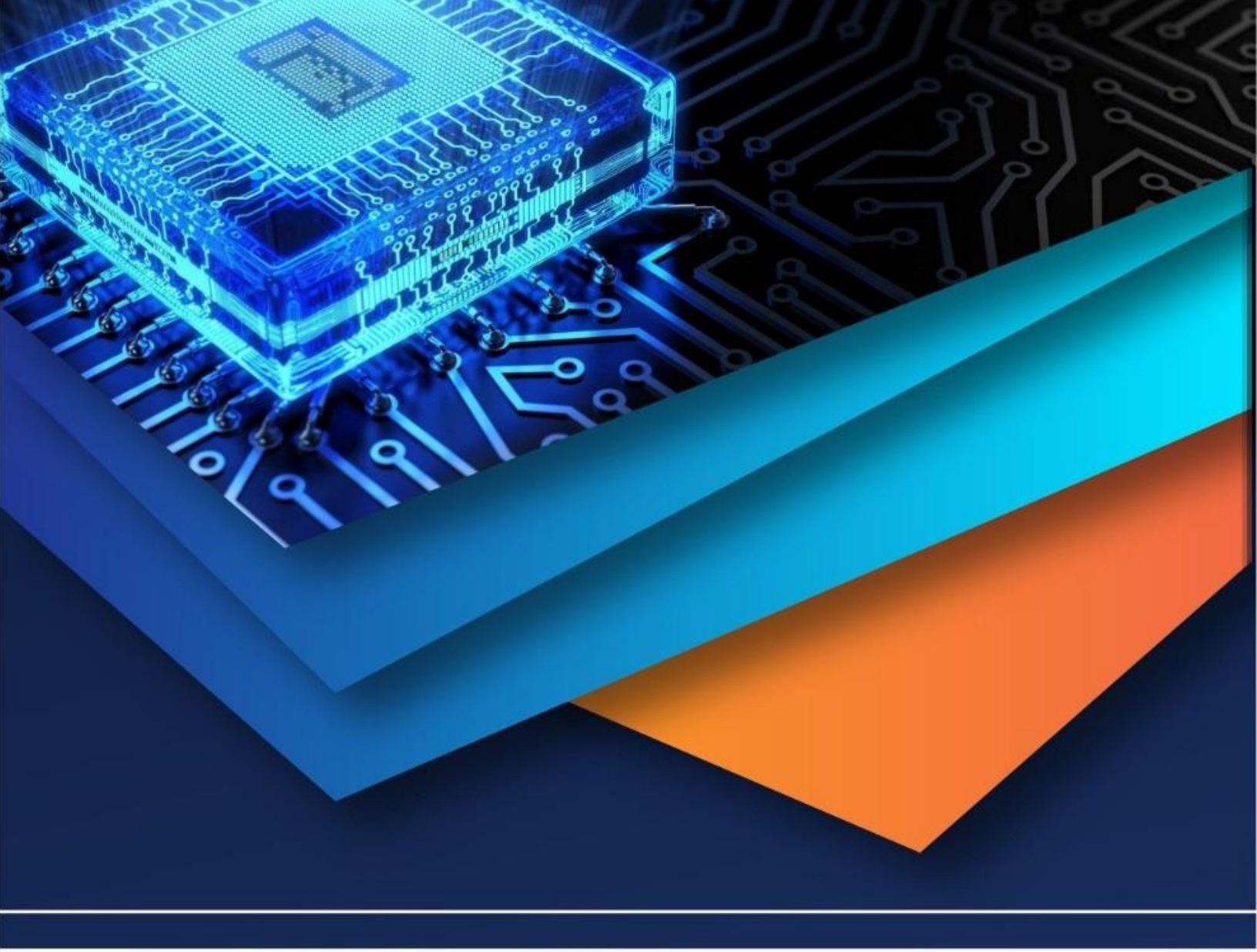
VI. CONCLUSION AND FUTURE SCOPE

We've proposed Perception, a trained model which uses some ways of AI to restate the silent videotape sample to a subtitled videotape. employing a trained CNN, the delicacy would change between 70 to 80 supported different videotape samples and also it showed advanced delicacy while using 5X5X5 kernel. This system may be employed in colorful fields like forensics, film processing, aid to the deaf and dumb, and lots of further naturally.

To further enhance this fashion within the future, we could descry different views of the content piecemeal from the anterior view so on apply it to a CCTV terrain. We could also extend it to other language datasets and other extended datasets(5).

REFERENCES

[1] Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. Under submission to ICLR 2017, arXiv:1611.01599 (2016)
[2] Almajai, S. Cox, R. Harvey, and Y. Lan. Improved speaker independent lip-reading using speaker adaptive training and deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2722–2726, 2016.
[3] Michael Wand and Jurgen Schmidhuber, Improving SpeakerIndependent Lipreading with Domain Adversarial Training. The Swiss AI Lab IDSIA, USI & SUPSI, MannoLugano, Switzerland, arXiv:1708.01565v1 [cs.CV] 4 Aug 2017.
[4] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016).
[5] Chung, J. S.; Zisserman, A. Lip Reading in the Wild. In Asian Conference on Computer Vision, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY