iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38206

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 IX Sep 2021- Available at www.ijraset.com

Deep Neural Networks for Human Action Recognition

Prof. Rajeshwari. J. Kodulkar[1], Prof. Mrunalini. S. Chakote[2]
[1, 2]*Department of Computer Science, DKTE Society's Yashwantrao Chavan Polytechnic, Ichalkaranji, India*

Abstract: ***In deep neural networks, human action detection is one of the most demanding and complex tasks. Human gesture recognition is the same as human action recognition. Gesture is defined as a series of bodily motions that communicate a message. Gestures are a more natural and preferable way for humans to engage with computers, thereby bridging the gap between humans and robots. The finest communication platform for the deaf and dumb is human action recognition. We propose in this work to create a system for hand gesture identification that recognizes hand movements, hand characteristics such as peak calculation and angle calculation, and then converts gesture photos into text.***
Index Terms: Human action recognition, Deaf and dumb, CNN.

I. INTRODUCTION

Recognition of human activity is important in human-to-human interaction and interpersonal relationships. It is tough to extract since it contains information about a person's identity, personality, and psychological condition. One of the primary themes of research in the scientific fields of computer vision and machine learning is the human ability to identify another person's activity. As a result, a multimodal activity recognition system is required in many applications, including video surveillance systems, human-computer interface, and robots for human behavior characterization. Human activity recognition is an important study field in image and video analysis. A vast number of articles on human activity recognition in video and picture sequences have been published in the past. Methods, techniques, and quantitative evaluations of human activity recognition performance are included in this overview of current developments in human activity recognition. The steps in activity recognition are as follows. The first step is to record human video pictures. Second, to distinguish between the many sorts of human actions. Previous attempts to video action detection have always used similar concepts to those used in image recognition. Human activities, on the other hand, consist of ever-changing motions with diverse target objects, and distinct things have different looks in different settings. As a result, exploring a variety of spatial-temporal characteristics is critical for action recognition. Due to its capacity to automatically learn features from huge datasets, one deep neural network (DNN) has achieved significant success in several domains, including object identification, recognition, and picture classification, by making full use of motion information. Convolution layers of a Convolution Neural Network (CNN) with orientation-sensitive filters retrieve spatial information of image scans. We can recognize human movements in images using a Convolution Neural Network. As we all know, the number of deaf and dumb people has increased dramatically as a result of birth abnormalities, accidents, and mouth illnesses. Because deaf and dumb persons are unable to interact with regular people, they must rely on visual communication. The finest communication platform for hearing impaired and deaf people to communicate with normal people is sign language. The goal of this study is to create a real-time hand gesture recognition system that recognizes hand movements, hand characteristics such as peak calculation and angle calculation, and then converts gesture photos into text.

II. LITERATURE REVIEW

The first attempt was made by Koller et al. [1] in 1991, who developed a framework that could be used to describe vehicle movement in real-world traffic situations using natural language action words. The tuples-based SVO (Subject, Object, Verb) approaches were among the first effective tactics used specifically for video representation. It took a long time for researchers to figure out how to translate visual content into everyday English

Later in 1997, Brand et al. [2] Getting back to SVO tuple-based techniques, which handle the video portrayal age task in two phases. The first arrange known as substance identification centers around visual acknowledgment and classification of the primary items in the video cut. The second stage involves sentence age which maps the items identified in the first stage to Subject, Verb and Object for syntactically stable sentences.

Hanckmann et al. [3] introduced a method to automatically describe multiple actions at a time. Human-human interactions are considered in addition to human-object interactions. Action detectors for detecting and classifying actions in a video. The description generator subsequently describes the verbs relating the actions to the scene entities. It finds the appropriate actors among objects or persons and connects them to the appropriate verbs.

Donahue et al. [4] were the first to utilize a profound neural system to take care of the video describing issue. They proposed three models for video depiction. Their model expects to have CRF based expectations of subjects, items, and action words after full go of complete video. This enables the design to watch the total video at each time step.

Kovashka and Grauman [5] propose a discriminative representation by learning the shapes of space-time feature neighborhoods and forming a hierarchy of words that capture space-time configurations at successively broader scales. Although the state-of-the-art methods have showed their success on the task of human action recognition, there still exist two severe problems. From the view point of model learning, the current single-task learning methods seldom consider the following three aspects together. First aspect is the consistence between the body-based classification and the part-based classification. The current single- task learning methods that usually map the low level (LL) feature to one class directly. Second aspect is correlation among multiple action categories. Although different actions have diverse characteristics, they can still be highly correlated by sharing similar partwise motion patterns. And last and third aspect is correlation among multiple views. The identical action capturing in different views can natur ally have different visual features. However, they can still have strong correlation with each other and discovering their latent correlation can benefit to Multiview information for classification.

Eriglen Gani and Alda Kika [6] proposed a continuous sign language recognition captured from signers both hands. Kinect device is used to construct depth map. To classify signers, hand a k means clustering algorithm is used to partition pixels into two groups. After extracting the hands contour pixels, centroid distance is calculated and Fourier descriptors is obtained which is used for hand shape representation.

Rashmi B. Hiremath and Ramesh M. Kagalkar [7] presents work on image processing techniques such as frame extraction, erosion, dilation, edge detection, blur elimination, noise. elimination, wavelet transform and image fusion techniques. Fourier descriptors are used for feature extraction and extracted features with hindi text are stored in the database and compared with given input testing video of the signer.

Cao Dong, Ming C.Leu, Zhaozheng Yin [8] presents the work on American Sign Language Alphabet Recognition Using Microsoft Kinect. Kinect is only Microsoft movement sensor which comprises of profundity sensor, RGB camera. For removing various highlights separation versatile plan was utilized and bolster vector machine is utilized for arrangement reason. This framework has one inconvenience that it gives constrained exactness.

III. RELATED WORK

In this part, we give a broad overview of how we build the system.

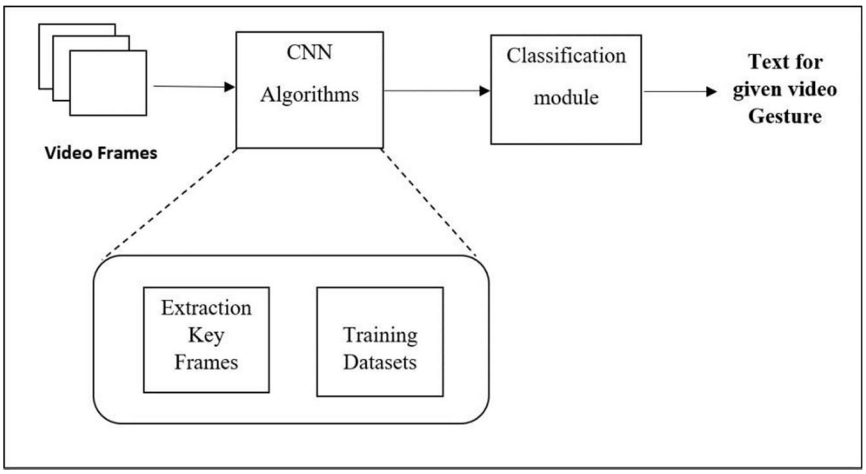


Fig. 1. Architecture of the System

The suggested system's system architecture is depicted in Figure 1. In the first step, key frames are extracted from provided video frames, after which frames are separated and the Determining Key Frames phase is completed. The second stage involves extracting key characteristics from key frames by comparing the similarity of an input picture to a reference image of an item. After that, a test dataset is used to train the CNN algorithm. The categorization of frames for hand gesture recognition is the final stage. The end result is a text for the provided hand motion.

A. Conversion of Video To Equivalent Human Natural Language Text

Figure 2 illustrates this. The following steps make up the system functioning architecture, which includes the process of extracting motions from video and translating them to natural language.

1) Frame formation from video.
2) Pre-processing and noise removal.
3) Extract only key frames.
4) Applying CNN algorithm.
5) Classification of gesture.

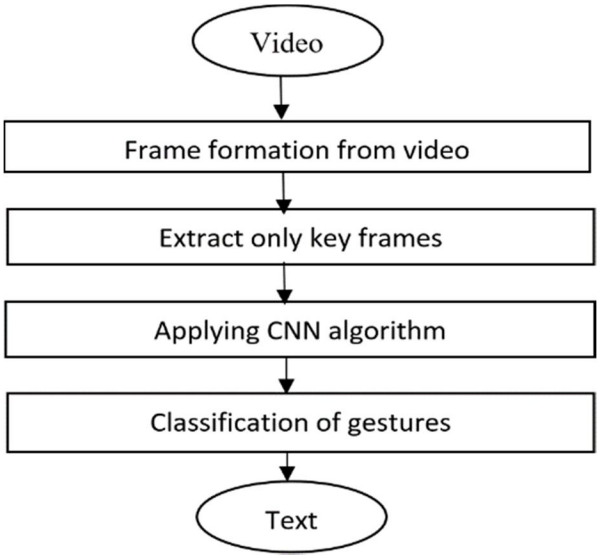


Fig. 2. System Working Architecture

a) *Frame Formation from Video:* We caught human gestures with the camera device for video sequences or frame separation. When it comes to processing movies with a computer, it is a more difficult operation than processing photos. As a result, we streamline the process by dividing the movie into picture frames. A video is essentially a large sequence of pictures recorded at a high data rate. As a result, generating frames from video entails reducing the video down to its simplest form, which is pictures.

b) *Extract only key Frames*

- *Separation of Frames:* A video consists of a series of frames. It's important to separate each frame in order to track the specific item from them. One of the most critical parts for object detection, identification, and tracking in video is the detection and separation of background and foreground components. Moving human body detection is an essential element of the ana- lysis of human body motion recognition. The most challenging challenge in video sequences is to recognize a moving human body against a backdrop element.
- *Determine Key Frames:* Because there will be significant changes in security camera recordings, such as high-frequency background objects, camera oscillations, and other disruptions. When separating foreground and background objects, these disruptions might create a lot of problems. The goal of frame processing is to extract as much information as possible from each frame before preparing the changed video frames.

IV. APPLYING CNN ALGORITHMS

Convolution layers are crucial for feature extraction in a Convolution Neural Network (CNN). CNNs are utilised in a variety of applications, such as image processing and pattern identification, natural language processing, and the recently added video analysis, among others. CNNs were created to translate visual data to a variable output.

Several films are gathered here, and the item to be recognized is trained against them. In this case, edge detection is crucial. The system's actual operation is examined to see if it is capable of extracting the above-mentioned properties accurately.

V. CLASSIFICATION OF GESTURE

Image categorization is an important stage in pattern recognition since it determines the efficiency and accuracy. Feature extraction is used to ensure that datasets are successfully trained. The rate of recognition is determined by all of the classifier's parameters. Pattern recognition benefits from include categorization. In one way or another, they are all significant. Any of these can be used to efficiently classify information. The following characteristics may influence classifier selection:

A. Edge detection
B. Eigen values and eigen vectors of a certain object's face
C. Texture lighting
D. Brightness
E. Histogram derived from retrieved data

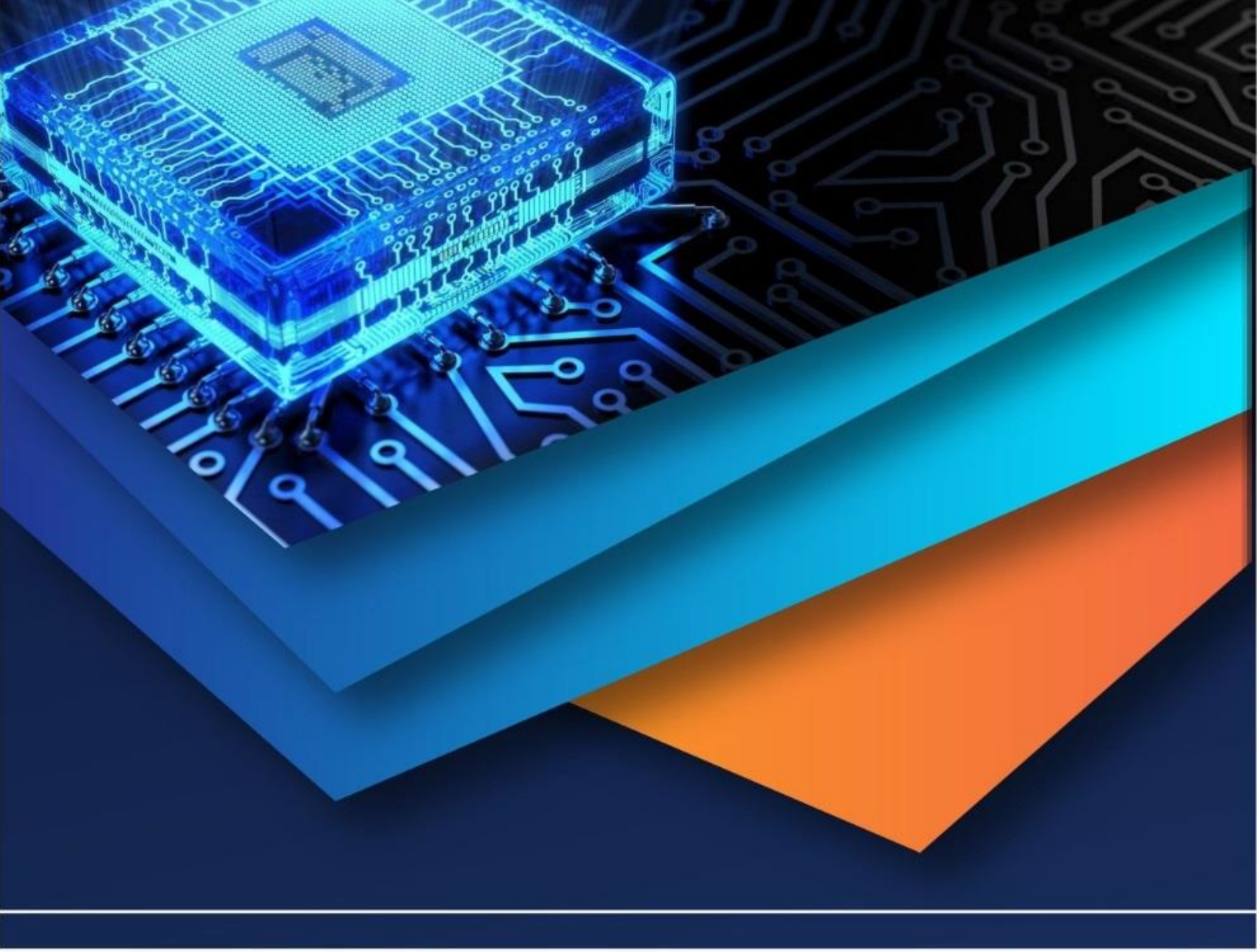
We provide the classifier one frame as input, which is the image of the person we're looking for. The aforementioned four modules are used by the classifier to carry out its tasks. The use of classifiers is used to match patterns. Finally, the anticipated output will be the frame that matches the video to the reference image.

VI. CONCLUSION

The suggested method is designed to bridge the communication gap between Deaf and Dumb people and normal people. Deaf-dumb individuals can utilize hand gestures as their primary language, which will be translated into text, thanks to this initiative. As a result, communication between them will be simple. In order for the system to become more dependable, additional research in the areas of feature extraction and lighting is required. The suggested technology translates video of hand gestures. frames into writing so that non-deaf and deaf people may understand what the deaf and dumb people are saying As a result, these deaf and dumb persons can communicate with their peers. Gestures play a significant role in human interactions, both interpersonally and in man-machine interfaces. Human motions can be recognized in a variety of ways. As a result, it must identify a vital essential component in action. The motions are interpreted by the CNN algorithm, which creates a statement from the video. The meaning of such movements is this statement or textual content.

REFERENCES

[1] D. Koller, N. Heinze, and H. Nagel. 1991. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. In IEEE Computer Society Conference on CVPR. 90-95.
[2] M. Brand. 1997. The" Inverse Hollywood problem": from video to scripts and storyboards via causal analysis. In AAAI/IAAI. Citeseer, 132-137.
[3] A. Kojima, T. Tamura, and K. Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. IJCV 50, 2 (2002), 171-184.
[4] F. Nishida and S. Takamatsu. 1982. Japanese-English translation through internal expressions. In Proceedings of the 9th conference on Computational Linguistics-Volume 1. Academia Praha, 271-276.
[5] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729, (2014).
[6] Yu-Ting Su, Ping-Ping Jia, Zan Gao, Tong Hao, and Zhao-Xuan Yang. Multipe/Single-View Human Action Recognition via Part- Induced Multitask Structural Learning. In IEEE Transaction 2015-16.
[7] Eriglen Gani , Alda Kika, "Albanian Sign Language (AlbSL) Number Recognition from Both Hand's Gestures Acquired by Kinect Sensors" International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016.
[8] Rashmi. B. Hiremath, Ramesh. M. Kagalkar, "Methodology for Sign Language Video Interpretation in Hindi Text Language" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 5, May 2016.
[9] Cao Dong, Ming C.Leu, Zhaozheng Yin, " American sign language Alphabet Recognition Using Microsoft Kinect", Computer Vision and pattern Recognition workshop, IEEE conference, pp ,2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY