Human Action Recognition Using Deep Neural Networks

Miss. Rashmi R. Koli, Prof. Tanveer I. Bagban

Department of Computer Science, D.K.T.E. society's Textile and Engineering Institute, Ichalkaranji, India.
Department of Information Technology, D.K.T.E. society's Textile and Engineering Institute, Ichalkaranji, India.

Abstract : Human action recognition is one of the most difficult and challenging things in deep neural networks. Human action recognition is nothing but the human gesture recognition. Gesture shows a movements of body prats which convey some meaningful message. Gestures are more preferable and natural to interact with computers for human thus it builds bridge be- tween humans and machines. Human action recognition provides best communication platform for the deaf and dumb person. In this paper we propose the to develop a system for hand gesture recognition, which recognize hand gestures, features of hands such as peak calculation and angle calculation and angle calculation and then convert gestures images into text.

IndexTerms - **Human action recognition, Deaf and dumb, CNN.**

1. Introduction

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations. Because it provides information about the identity of a person, their personality, and psychological state, it is difficult to extract. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning. As a result of this many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In image and video analysis, human activity recognition is an important research direction. In the past, a large number of papers have been published on human activity recognition in video and image sequences. The survey of the recent development in human activity recognition includes methods, systems, and quantitative evaluation of the performance of human activity recognition. Following steps are performed in activity recognition. First, to capture the human video images. Second, to identify the different types of action performed by human. For video action recognition, previous approaches always take similar ideas with that of image recognition. But different from still images, human actions consist of everchanging motions with different target objects, and different objects have various appearances in different scenes. So, it's indispensable to explore diverse spatial-temporal features for action recognition. To make full use of motion information, one deep neural networks (DNNs) have obtained great achievement in many areas such as object detection, recognition, and image classification, due to its ability of automatically learning features from large datasets. Spatial features of image scan be extracted by convolution layers in Convolution Neural Network (CNN), which contains orientation-sensitive filters. By using Convolution Neural Network, we can identify human gestures in image. As we know there has been rapid increase in the number of deaf and dumb victims due to birth defects, accidents and oral diseases. Since deaf and dumb people cannot communicate with normal person so they have to depend on some sort of visual communication Sign language provide best communication platform for the hearing impaired and dumb person to communicate with normal person. The objective of this research is to develop a real time system for hand gesture recognition which recognize hand gestures, features of hands such as peak calculation and angle calculation and then convert gesture images into text.

2. Literature Review

The first ever endeavor returns to Koller et al. [1] in 1991, who built up a framework that had the option to describe movement of vehicles in genuine traffic scenes utilizing characteristic language action words. The SVO (Subject, Object, Verb) tuples-based techniques are among the first effective strategies utilized specifically for video depiction. Research endeavors were made well before to depict visual substance into common language.

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.

3. Related Work

In this section, we present an overall description of how we develop the system.

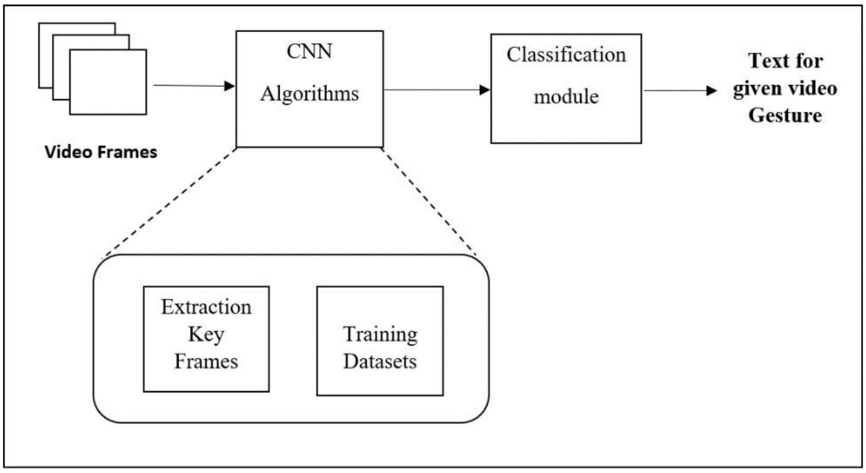

Fig. 1. System Architecture

Fig.1 shows the system architecture of proposed system. In first stage extraction key frames from given video frames, in that separation of frames and Determining Key Frames part is done. After that second stage is key features extraction on key frames, by measuring the similarity between an input image and a reference image of an object. Then CNN algorithm is trained using test dataset. Then last stage is classification of frames for hand gesture recognition. The final outcome is text for given hand gesture.

3.1 Conversion of video to equivalent human natural language text.

As shown by Fig.2. System working architecture, the process of drawing out gestures in video and converting it to natural language comprises of the fol- lowing phases.

- Frame formation from video.
- Pre-processing and noise removal.
- Extract only key frames.
- Applying CNN algorithm.
- Classification of gesture.

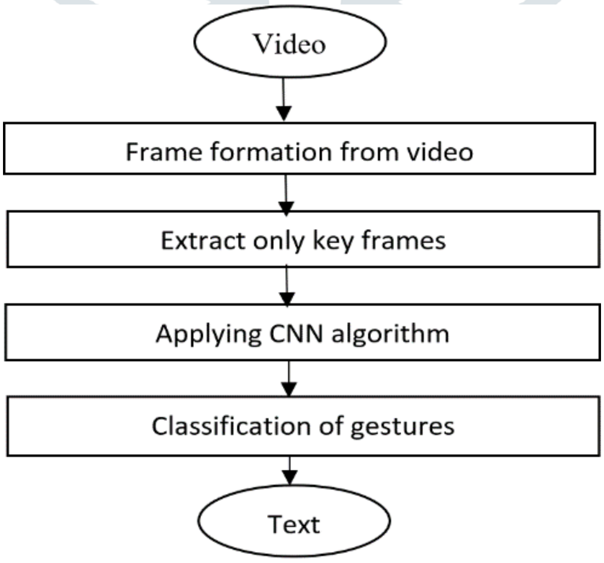

Fig. 2. System Working Architecture

I.Frame formation from video.

Using the camera device, we captured human gesture for video sequences or frame separation. When we talk about processing videos with computer, it is complicated task as compared to processing of images. Thus, we simplify this process by breaking the video into image frames. Actually, a video is nothing but huge sequence of images captured at high data rate. So, formation of frames from video is breaking the video into its original base from that is images.

II.Extract only key frames.

1) Separation of Frames:

Video is a collection of many frames. To trace the particular object from these frames it is necessary to separate each frame. Video background compo- nent detection and foreground component separation is one of the most im- portant elements for object detection, identification, and tracking. In the anal- ysis of human body motion recognition, the moving human body detection is an important part. The moving human body is detected from the background element in video sequences is the most difficult task.

2) Determine Key Frames:

Since there will be important changes in videos captured by surveillance cameras, high-frequencies background objects, camera oscillations, and other disturbances. These disturbances can cause a lot of trouble when we separate foreground object from the background. The purpose of the frame processing is to gained maximum amount of information from frame, then prepare the modified video frames.

III. Applying CNN algorithms.

In a Convolution Neural Network (CNN), convolution layers play the im- portant role of feature extraction. CNN's are used in a different type of areas, including image processing and pattern recognition, natural language pro- cessing, and newly introduced video analysis, etc. CNNs, were designed to map image data to an output variable.

Here several videos are collected and object to be identified is trained against these videos. Here edge detection plays vital role. Actual working of system is analyzed whether system is able to extract above mentioned features correctly.

IV. Classification of gesture.

Image classification is a necessary step in pattern recognition, the efficiency and accuracy mainly depend on the classification. To do the successful Train- ing of datasets, feature extraction is carried out. Recognition rate depends on all the steps of classifier parameters. Including classification has its importance in pattern recognition. All have their importance in one or the other way. An- yone of them can be used to perform efficient classification. The selection of classifier may depend upon the following parameters:

(1) Edge detection

(2) Eigen values and eigen vectors of face of particular object

(3) Texture illumination

(4) Brightness

(5) Histogram obtained from extracted

We give one frame as input i.e. image of the person which we are looking for is given to classifier. classifier carries out its working based on the above four modules. Pattern matching is done with the help of classifiers. Finally, the frame which matches from video to reference image will be expected output.

4. Conclusion

The proposed system is introduced for Deaf and Dumb People for removing communication gap between Normal People. With this project the deaf-dumb people can use the hand gestures as their primary language and it will be converted into text. So, the communication between them can take place easily. There is need of research in the area feature extraction and illumination so the system becomes more reliable. The proposed system converts gesture video. frames into text so the normal person knows that what the deaf and dumb person can said. Hence, this deaf and dumb people can connect with their society. Gestures are important aspect of human interaction, both interpersonally and in the context of man-machine interfaces. There are many ways to the recognize the human gestures. So, it needs to identify valuable key element in action. CNN algorithm interprets the gestures and builds a statement from the video. This statement or textual information is the meaning of that gestures.

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.