International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 06th Feb 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=ISSUE-02

DOI: 10.48047/IJIEMR/V12/ISSUE 02/15

Title **3DCNN-GRU Based Video Controller through Hand Gestures**

Volume 12, Issue 02, Pages: 105-109

Paper Authors
Godavarthi Deepthi, Beebi Naseeba, Mahalakshmi Palvadi

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

3DCNN-GRU Based Video Controller through Hand Gestures

Godavarthi Deepthi [1][0000-0003-0712-6899], **Beebi Naseeba**[2] , **Mahalakshmi Palvadi**[3]
[1]Assistant Professor, School of Computer Science and Engineering (SCOPE), VIT-AP University, Andhra Pradesh, India.
[2] Assistant Professor, School of Computer Science and Engineering (SCOPE), VIT-AP University, India.
[3]Student, School of Computer Science and Engineering (SCOPE), VIT-AP University, India.

Abstract

The goal of dynamic hand gesture recognition framework is to create a natural interaction between human being and a machine. Existing systems are not so efficient in providing accurate outcome to users. Hence this work presents a new dynamic hand recognition framework based in deep learning that helps to operate a video through hand gestures and performs operations namely: Play, Pause, Volume Up, Volume Down for a video. Initially, divide each video input into number of clusters and then select a frame from each cluster randomly. The result is then fused and fed into GRU to predict classification result. CNN with GRU helps in better long-term feature extraction compared to other methods with an accuracy of 98.5. The proposed model helps the users with disability to control the video player from anywhere in the room just with captured hand gestures.

Keywords: Deep learning, Computer Vision, Hand Gesture Recognition, Gated Recurrent Unit(GRU)

Introduction

Computer vision is being used mostly now a days where a small program can be used for feature identification without any human work. Computer vision is used to recognize facial features, for Color detection and even automation. Computer vision is used in Optical mouse creation using hand gestures. Different gestures that are performed by an individual's hand are read using the camera of the computer and based on gestures movement, the computer's cursor will move, also performs right and left clicks by making use of different gestures. The only hardware requirement is the web cam. In this we use GRU algorithm for classification, processing and prediction. These hand gestures can be used in replacement of spoken language for communication. On the basis of input devices, these existing hand gesture recognition approaches are categorized into vision based and non- vision based. The original image information provided through the input devices includes or mal camera, stereo camera, and Time of flight camera. Features can be extracted from the region of interest (ROI), when it is detected. Most widely used features are Colour, brightness, and gradient. Hand gestures provides a best way for human robot interaction. Due to various applications, Vision based dynamic hand gestures recognition became a most awaited research topic. n this work, deep learning-based framework which is an integration of 3D-CNN and GRU has been used to recognize gestures and to control the video using them.

Related Work

Muneeb Ur Rehman et.al [1] proposed a framework for Dynamic Hand Gesture Recognition based on 3D-CNN and LSTM in which 3D-CNN is used for extracting both spatial and spectral features which are then provided as input to LSTM .Then classification is done using LSTM.

Zhang et al [2] proposed Dynamic Hand Gesture Recognition model in which the identification of gestures is done through meaningful body motions that involves hands, head, body to communicate with the environment. A wearable arm hand based real-time gesture recognition system using electromyography sensors, machine learning algorithms like k-nearest neighbour and decision tree algorithm has been used in this work.

Molchanov et al.[3]developed an algorithm for classification through hand gestures whichutilizes3D-CNNcombiningvarious augmentation techniques to reduce over fitting in spatio-temporal domain. When the developed model has been tested on VIVA challenge dataset, it produced a precision of 78% in classification .

Pigou et al.[4] proposed a system that integrates residual network along with batch normalization and other units for RGB dataset. They observed that LSTM alone fails to capture low level information and hence 3D-CNN has been coupled with LSTM in their study.

John et al.[5] developed a long term recurrent convolution network based model for classification of hand gestures from video sequences. They extracted few frames of video sequences and provided them as input to LRCN network thus improved computational efficiency and accuracy.

Lai and Yanushkevich et al.[6] proposed a model where convolutional neural networks (CNN) is coupled with recurrent neural networks (RNN) to recognize hand gestures automatically from depth and skeleton data and achieved 85.5 % accuracy on dynamic and gesture –14/28 dataset.

Ula Tarik Salim and Shefa Abdulrahman Dawwd[7] Proposed vision based hand gesture recognition system which has two categories namely static and dynamic. The developed model can be used for static hand shape detection along with classification in one image or at the beginning, within or at the end of the gesture in video sequences.

Rihem Mahmoud et al. [8] Proposed a new recognition system to resolve the issue of large scale continuous geture recognition from input videos. The proposed system divides geture sequences into individual gestures using mean of velocity information. Then extracted set of suitable descriptors termed deep signature features for each isolated segment and the constructed features for depth and gray scale sequences are provided as input into linear SVM.

Adithya and Rajesh[9] Proposed the CNN architecture which eliminated the necessity of hand detection and segmentation from images captured through webcam. Hence the computational burden during recognition of hand postures has been reduced. The proposed model also derives important features that identifies the hand gestures having small interclass variations also.

Dushyant Kumar Singh[10] Used 3D-CNN to model the mostly used gestures of Indian Community. The developed model provides natural language output for signs of ISL. It helps in effective communication with deaf and dumb people. A total of 20 gestures has been taken from Indian Sign Language to train the model.

Cleison Correia de Amorim et al.[11] Developed Spatial-Temporal Graph Convolutional Network to recognise sign language of human skeletal movements. In this method graphs are used to capture signs in spatial and temporal features. Authors proposed a noval dataset with skeletons of humans based on ASLLVD for sign language as a contribution to future studies.

Jun Wanet al.[12] Addressed the challenges in collecting annotations of gestures and provided an analysis for gesture recognition of isolated and continuous RGB-D video sequences. Authors introduced corrected segmented rate metric for performance evaluation for continuous gesture recognition and proposed Bi-LSTM method to discover video separation points contingent on points retrieved from convolutional pose machine.

Ying Ma and co-workers [13]Proposed two stream mixed method to improve feature expression correlation between images for dynamic gestures with two operations namely feature extraction and fusion. The proposed system consists of pre-processing, TSM block and CNN classifiers. In the pre-processing stage, two consecutive images in gestures are taken as input and operations such as resizing, transformation and augmentation are performed. The fusion feature map is obtained in TSM block by summation and concatenation which is used as input to classifier. The classifier then classifies the images.

Proposed Work

Dataset

In this paper, to train a model,20BNjester dataset has been used which has around 148,092 labelled video clips with hand gestures of different people. The videos in the dataset are of different length from 27 to 46 frames. During pre-processing, number of frames in video are restricted to 36 frames which are used for training a model and the frames of the input are firster sized to 112*112 pixels.

Methodology

Learning low level information having temporal and spatial details to recogniz hand gestures with a single model is a challenging task. To address this issue, a novel framework with 3D-CNN followed by GRU and a Soft Max layer has been proposed as shown in Fig 1. The developed model combines 3D convolutional neural networks(3D-CNN) with feature fusion. It's necessary to learn both spatial and temporal features to classify dynamic gestures. Therefore, using a 3D-CNN model is not sufficient to recognize the dynamic hand gestures as it cannot learn more about spatial and temporal data from video sequences. Hence, a novel network is required to learn more about this information. In thiswork,3D convolutional neural networks (3D-CNN)has been integrated with GRU network. The pipeline consists of fusing models namely a 3D-CNN coupled with LSTM and 3D-CNN coupled with GRU. To train a model,20BNjester dataset has been used which has around 148,092 labelled video clips with hand gestures of different people. The videos in the dataset are of different length from 27 to 46 frames. During pre-processing, number of frames in video are restricted to 36 frames which are used for training a model and the frames of the input arefirstresizedto112×112pixels.The spatial and temporal features are represented using original and optical flow key frames respectively which are then fed to the 3D-CNN for feature fusion and final recognition. The 3D-CNNproduces the output as feature maps having spatial and temporal features, which is provided as input to the GRU model. The exact information can be obtained from video frames for which GRU has been used to classify the hand gestures.

After training a model, the image captured lively through web cam is provided as input to the developed model which results in the class label prediction together with gesture. A modified VLC is developed from python whose operations are managed using various python libraries. The commands such as play, stop, volume up ,volume down are provided as arguments to the actions carried out by the gestures. Whenever user provides one specific gesture as input through web cam while the video is in play mode, inter related action invoked from VLC file.

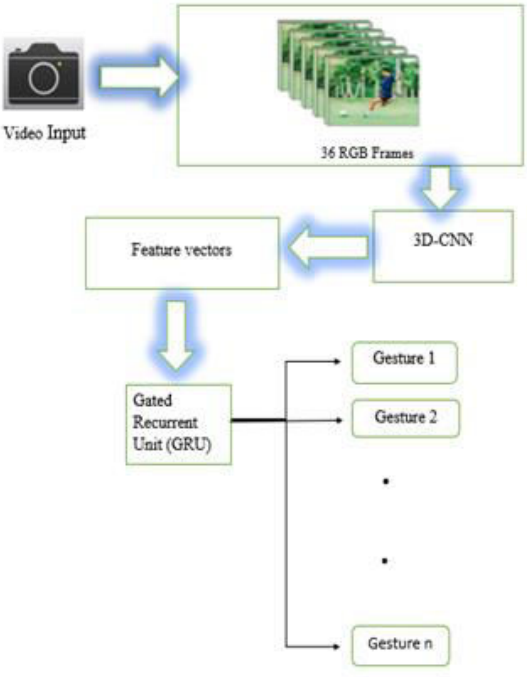

Figure 1: Architecture of the proposed model

We compared the proposed frame work acquired an accuracy of 95.4% on unseen gestures taken through web cam.

Results

The developed model has been tested on image captured lively through web cam and the obtained validation accuracies along with training results are shown in Figure. 2.

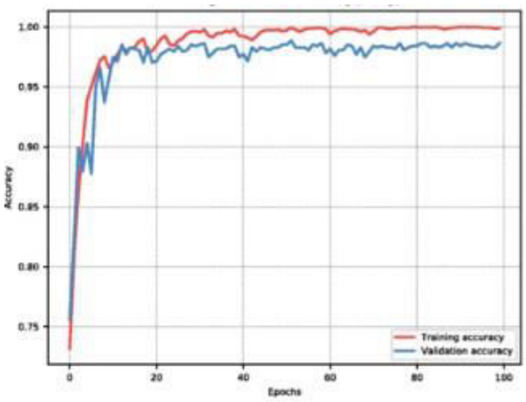

Figure 2: Categorical Accuracy

From the Figure 3, it is clear that hand is detected from webcam which is then converted into RGB. Video will play when both palms are open as shown in the Figure below.

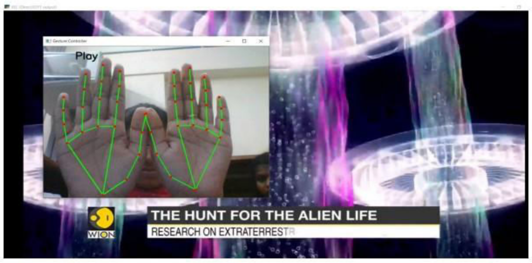

Figure:3 'Play' Hand Gesture for VLC Player

When both the palms are closed video will stop or pause which is shown in Figure 4.

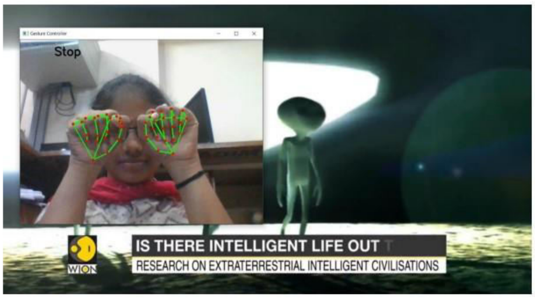
Gesture Controller
Stop
WION
IS THERE INTELLIGENT LIFE OUT
RESEARCH ON EXTRATERRESTRIAL INTELLIGENT CIVILISATIONS

Figure 4: 'Stop' Hand Gesture for VLC Player

Moving index fingers of both hands upwards instruct the video player to increase the volume as shown in Figure5 while moving index fingers downwards decrease the volume of the video player.

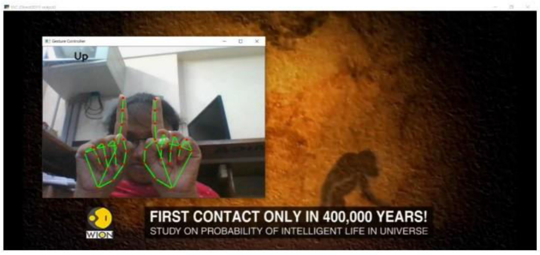
Gesture Controller
Up
WION
FIRST CONTACT ONLY IN 400,000 YEARS!
STUDY ON PROBABILITY OF INTELLIGENT LIFE IN UNIVERSE

Figure 5: Volume Up Hand Gesture for VLC Player

Conclusion and Future Scope

In this project, we took various images and trained the model to recognize and detect different hand gestures. This application capture hand gestures as an input through a basic 720p camera which serves as a direct command to control the video player. Using the proposed work basic operations such as play, pause could be performed and one can also adjust the volumes with different hand gestures assigned to it, which makes it easy to operate. Here we combinedConv3D+GRU. GRU can be trained on human hand gestures in a direct fashion, without gesture segmentation. The developed model will be enhanced further by applying a high-level semantic analysis to the present system thus improving the recognition capability for more complicated human tasks. And, a part from controlling the video player we will try to implement it for the entire screen controlling which will be much more beneficial in real time.

References

[1] M. Ur Rehman et al., "Dynamic hand gesture recognition using 3D-CNN and LSTM networks," Comput. Mater. Contin., vol. 70, no. 3, pp. 4675–4690, 2022, doi: 10.32604/cmc.2022.019586.

[2] W. Zhang, J. Wang, and F. Lan, "Dynamic hand gesture recognition based on short-term sampling neural networks," IEEE/CAA J. Autom. Sin., vol. 8, no. 1, pp. 110–120, 2021, doi: 10.1109/JAS.2020.1003465.

[3] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional

neural networks," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2015-Octob, pp. 1–7, 2015, doi: 10.1109/CVPRW.2015.7301342.

[4] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and Sign Language Recognition with Temporal Residual Networks," Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017, vol. 2018-Janua, no. Figure 1, pp. 3086–3093, 2017, doi: 10.1109/ICCVW.2017.365.

[5] V. John, A. Boyali, S. Mita, M. Imanishi, and N. Sanma, "Deep Learning-Based Fast Hand Gesture Recognition Using Representative Frames," 2016 Int. Conf. Digit. Image Comput. Tech. Appl. DICTA 2016, no. October 2017, 2016, doi: 10.1109/DICTA.2016.7797030.

[6] K. Lai and S. N. Yanushkevich, "CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition," Proc. - Int. Conf. Pattern Recognit., vol. 2018-Augus, pp. 3451–3456, 2018, doi: 10.1109/ICPR.2018.8545718.

[7] U. T. Salim and S. A. Dawwd, "Systolic hand gesture recognition/detection system based on FPGA with multi-port BRAMs," Alexandria Eng. J., vol. 58, no. 3, pp. 841–848, 2019, doi: 10.1016/j.aej.2019.05.018.

[8] R. Mahmoud, S. Belgacem, and M. N. Omri, "Deep signature-based isolated and large scale continuous gesture recognition approach," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 5, pp. 1793–1807, 2022, doi: 10.1016/j.jksuci.2020.08.017.

[9] V. Adithya and R. Rajesh, "A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition," Procedia Comput. Sci., vol. 171, no. 2019, pp. 2353–2361, 2020, doi: 10.1016/j.procs.2020.04.255.

[10] D. K. Singh, "3D-CNN based Dynamic Gesture Recognition for Indian Sign Language Modeling," Procedia CIRP, vol. 189, pp. 76–83, 2021, doi: 10.1016/j.procs.2021.05.071.

[11] C. C. de Amorim, D. Macêdo, and C. Zanchettin, "Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11731 LNCS, pp. 646–657, 2019, doi: 10.1007/978-3-030-30493-5_59.

[12] J. Wan et al., "ChaLearn Looking at People: IsoGD and ConGD Large-Scale RGB-D Gesture Recognition," IEEE Trans. Cybern., vol. 52, no. 5, pp. 3422–3433, 2022, doi: 10.1109/TCYB.2020.3012092.

[13] Y. Ma, T. Xu, and K. Kim, "Two-Stream Mixed Convolutional Neural Network for American Sign Language Recognition," Sensors, vol. 22, no. 16, 2022, doi: 10.3390/s22165959.