Video Based Hand Gesture Detection System Using Machine Learning

Article · June 2020

CITATIONS
9

READS
2,489

4 authors:

Manjunath R Kounte
REVA University
92 PUBLICATIONS **802** CITATIONS
SEE PROFILE

E. Niveditha
REVA University
3 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Sai Sudeshna
REVA University
3 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Kalaigar Afrose
REVA University
4 PUBLICATIONS **12** CITATIONS
SEE PROFILE

International Journal of Advance Science and Technology
Vol. 29, No. 10S, (2020), pp.3801-3810

Video Based Hand Gesture Detection System Using Machine Learning

Manjunath R Kounte[1], E Niveditha[2], A Sai Sudeshna[3], Kalaigar Afrose[4]

School of ECE, REVA University, Banglore, India

[1]manjunath.kounte@gmail.com, [2]niveditha.e.nt@gmail.com, [3]saisudeshnaa@gmail.com, [4]kalaigarafrose@gmail.com

Abstract

Machine Learning and convolution networks have achieved great success in image recognition. However, for action recognition in videos, their advantage over traditional methods is not so evident. In this paper, a general and flexible video-level framework for learning action models in videos is presented by considering the Temporal Shift Module (TSM) and Convolutional Neural Networks (CNN) Using Jetson Nano KIt. The main objective is to develop a prototype for implementing a low latency and highly efficient real-time online hand gesture recognition system with low computational cost. The problem is very challenging when it comes to processing speed and efficiency/accuracy. To solve this problem, we are using a multi-modal algorithm (TSM+2D CNN). In this algorithm, TSM combines with 2D Convolutional Neural Networks (CNN) to achieve high efficiency. In the prototype we trained a hand action recognition model capable of detecting around 15 different actions from online videos which can be an efficient and natural human-computer interaction. The design procedure of the trained model has been explained completely. The result obtained in the design and implementation of the system has been quite satisfactory.

Keywords: *Convolutional Neural Networks (CNN), Gesture Recognition, Machine Learning, Temporal Segment Networks (TSN), Temporal shift Model (TSM), Jetson Nano Kit.*

1. INTRODUCTION

Hand gesture recognition is very significant for human-computer interaction. Gesture recognition is the procedure of figuring out the gestures through the computer which is made by the user. Till now, many hand gesture recognition technologies have been evolved, but Video-Based Hand Gesture Detection with high efficiency and low computational cost is a very challenging task.

Computer vision plays a crucial role in video understanding over the years. Computer vision is a new field where the computers are given the ability to extract high-levels of understanding from digital images and videos. It can be applied to hand gesture recognition systems to provide input to the computer where we can effortlessly move and manipulate virtual objects by simply moving and turning the hands resulting in a low-cost human-computer interaction device substituting the need of keyboards and mouse in laptops and computers. Temporal modeling can't be done by computer vision which is a major disadvantage[10-14].

Over the last few years, there has been research going on machine learning. Machine Learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data. There are various techniques for machine learning. Convolutional Neural Networks (CNN) is one of the techniques of machine learning, specially meant for image processing and object detection applications. CNN works by breaking an image down into smaller groups of pixels called frames and does a series of calculations on the pixels comparing against another pixel in a specific pattern[9].

To make accurate predictions each time a CNN predicts a labeled data it uses an error function to compare how close its prediction was to the actual image label. Based on this error function the CNN updates its filter values and starts the process again ideally, each iteration performs with slightly more accuracy. Conventional 2D CNNs are computationally cheap but cannot capture temporal relationships. One key difference between video recognition and image recognition is the need for temporal modeling. For example, we have a picture of a cardboard box filled with things, we want to label it whether we are packing or unpacking the box, using CNN we can label the image as open box or closed box but not the action is done, this is where CNN lacks they can only take into account spatial features of the image but can't handle temporal or time features. To detect the action, we must compare how a frame is related to the before it, to solve this issue we have to feed the output of the CNN model to another model (TSM) that can handle the temporal nature of the videos[15-18].

Our proposed paper focus on temporal modeling, we are using a multi-model algorithm that enables us to do real-time video recognition with high efficiency and low computational cost. This algorithm combines TSM with 2d CNN which can handle temporal features with high accuracy. In our prototype, using a jetson nano kit, we trained a model, capable of detecting around 15 different actions from online videos.

2.LITERATURE SURVEY

Recently there has been extensive research in the field of hand gesture recognition and video recognition models and has made great progress and achieved high efficient recognition rates in almost all the domains.

Primarily in the hand gesture recognition was implemented using a wearable gesture recognition systems as proposed in the following [1],[2] used flex sensors coupled with accelerometers detect the gestures based on the resistance generated due to the finger movements.

Paper [3] proposed switch-based digital gloves embedded with 5 light-emitting diodes of each finger to identify the gesture. Wearable technology is limited to small range applications and requires hardware such as gloves to be worn constraining and uncomfortable to use. Deep convolution networks are known for efficient gesture recognition from images.

ISSN: 2005-4238 IJAST

The papers [4][5] have implemented video-based action detection however this model focus on short-term motions and lacks the capacity to incorporate long-range temporal structure which are important for dynamic action understanding. Also these ConvNets require large sets of labeled data to achieve high efficiency.

2D Convolution Neural Networks(2D CNN) is proposed in [6] using Temporal Segment Networks(TSN) which extracts short snippets randomly in a long video sequence using a sparse sampling scheme along the temporal dimension and aggregate the information to enable long-range temporal modeling.TSN are computationally cheap compared to 3D CNN but not great at inferring complex temporal relations.

In paper [7,8] proposes 3D CNN model where it has high spatial and temporal modeling characteristics however 3D CNN suffers from high computational cost as the numbers of parameters used are more compared to 2D CNN and are under the risk of over fitting.

Hence to implement a good video based action recognition system that has both high accuracy as that of 3D CNN and having low computational cost as that of 2D CNN, we used a Temporal shift Model (TSM) presented in [8] which is inserted into the 2D CNN back bone. Uni-directional TSM algorithm enables to do real time video recognition with low latency and high temporal modeling on high computation devices[19-23].

3.DESIGN

Contemporary researches in the field of hand gesture recognition have proven to have many practical implementations. The present desktop monitors and projectors can't coexist with upcoming 3D applications on augmented and virtual reality applications that create immersive experiences for the users. Virtual realities have a wide range of applications in architecture, military, medicine, etc. hand gestures based inputs can assist the VR technology to make the user environment realistically as it enables hardware-free control mechanism. As a corollary the system can be used to implement a smart power pointing system, just by using a camera user can interact and control the PowerPoint wirelessly with simple hand gestures. The prototype is capable of providing a solution to the hearing impaired people with automatically recognizing the hand gestures and converts it into text so making it easy to communicate with normal people. . Hand gestures are interfaced to many smart-home applications, such as control of TV, lighting systems, the modern gamming modules to provide a joyful experience to the users[24-27].

The main objective of this project is to develop a state of art prototype for implementing a low latency and highly efficient real time online hand gesture recognition system. At any point of time an user can perform an hand gesture in front of web camera connected to high end processing, which acquires gestures to analyze and identify the same. The gesture identifies is displayed on the screen connected to the process. In the prototype we aim to train an hand action recognition model capable of detecting around 15 different actions from online videos which can be an efficient and natural human-computer interaction. The recognition process has to be fast and accurate for online action detection. An enhance 2D convolution neural networks(2D CNN) is used for extracting the special features and followed by Temporal shift module(TSM) for extracting the temporal features by maintain the spatial density of the images and has been used for hand action detection.

The input used by the gesture recognition system is provided by a Logitech C270 HD web camera that is fixed on a table. This camera can capture 60 frames per second and 1280x720 high-quality images and is installed on to the NVIDIA Jetson Nano board through USB 2.0port. the Jetson Nano has been opted over Raspberry pi 4 has as it has higher processing speed which runs on the QUAD-core ARM cortex-A57 processor specially designed to run heavy graphics for machine learning algorithms. The Jetson nano board is powered using a 5V 4A USB Micro B adapter plugged into the 240v main socket. A keyboard and mouse are interfaced through one of the USB ports on the board which is used to provide input commands to stop and start the gesture recognition process. A monitor is connected using an HDMI cable where the live camera video is displayed along with the hand gesture prediction for the specific gesture made by the user is also displayed as an output from the jetson nano on processing the video.

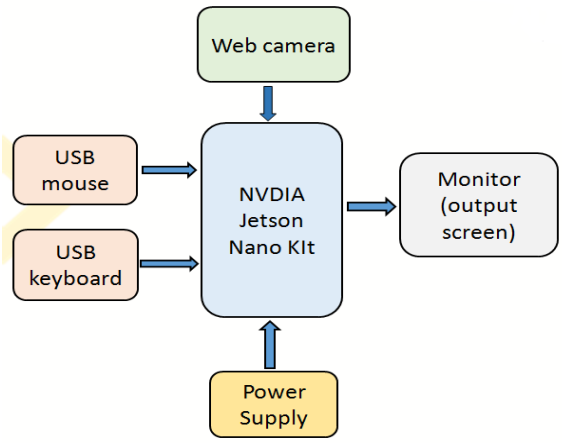

Fig 1. Block Diagram of Gesture Recognition System

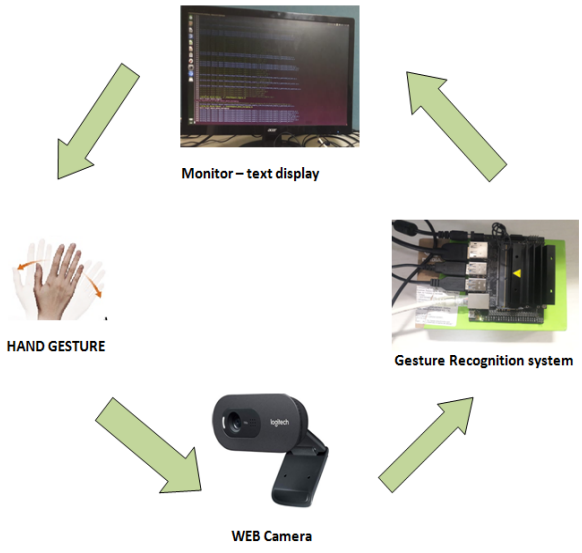

Fig 2: Workflow of dynamic hand gesture recognition

The hand gestures are of two types namely static and dynamic hand gestures, the static hand gesture is recognized by defining the posture of the hand whereas dynamic gestures are formed by the movement of the hand in the space. To implement an online video-based hand gesture recognition system dynamic hand recognition is needed this is done by using a multimodal algorithm 2D CNN +TSM. The video acquired through the web camera is sampled and are individually preprocessed in the 2D CNN phase where spatial data is extracted; the output of this is feed into TSM for temporal feature extraction which is very essential for action recognition. TSM works by shuffling the feature maps of the adjacent frames individually along the temporal dimensions without any additional computation cost.

The present application is implemented in python 3 using the open CV 4.0 to access the camera feed, open-source frameworks PyTorch v1.2 - torchvision v0.4.0 are installed for manipulating and preprocessing the images with standard image processing algorithms. Tensor Virtual Machine (TVM) is installed to increase the compatibility between various frontend modules(TSM) and backend hardware(Jetson Nano, web camera). Transfer learning made use instead of training the model from the scratch and optimizing it, with the help of Open Neural Network Exchange (ONNX) pre-trained models is imported integrated into our model seamlessly. To achieve our goal of low latency in gesture recognition computer Unified Device Architecture (CUDA) is used for speeding up heavy computational tasks such as video recognition by applying parallel processing techniques.

4. RESULTS

The experimental prototype for the proposed video based hand gesture recognition system using Jetson Nano for which we obtained a satisfactory result that aimed at recognizing highly efficient real time hand gesture recognition. This project was successful in cutting down high computation costs and improved the processing speed resulting in accurate and fast gesture recognition by utilizing a hybrid algorithm constituting a 2D CNN followed by Temporal Shift Module.

The following parameters can be considered to witness the performance of the proposed hand gesture recognition system:

High processing speed: Low latency and high through put are vital for video understanding; inserting TSM into 2D CNN baseline has improved the video understanding capability without any additional computational cost and latency. In paper [9] an Efficient Convolutional Network for Online Video Understanding (ECO) is proposed using both 2D and 3D CNN architectures, ECO use multi- level temporal fusion for each prediction which may lead to latency in gesture prediction. While TSM model uses less expensive 2D backbone where it catches only 1/8th portion of the frame and are replaced with adjacent frames for each prediction thus reducing the processing time.

Low memory utilization: As the model caches only 1/8th features in each frame it only consumes 0.9MB of memory for storing the intermediate features which is comparatively very less compared to other similar models.

ISSN: 2005-4238 IJAST

Multi-level temporal fusion: TSM enables temporal fusion at every level while paper [10] proposes model which only does late temporal fusion after the feature extraction and [9] does mid level temporal fusion.

The system was capable of accurately detect the hand gestures even under low light, complex back ground or even when the body is present in the frames. The model rigorously predicted 15 different hand gestures made by the user before the web camera with least delay and a processing speed of about 30vid/s. The gesture detected is displayed on the monitor in the form of text.

Following figures shows the results obtained from different gestures made by the user.

When the hand gesture thumbs up is made the system immediately displaied the prediction to be “Thumb Up” as shown in Fig.4.

Fig 4:Thumb up

When the all fingers are moved continuously up and down the system predicted that the gesture was “Drumming Fingers” as shown in Fig. 5

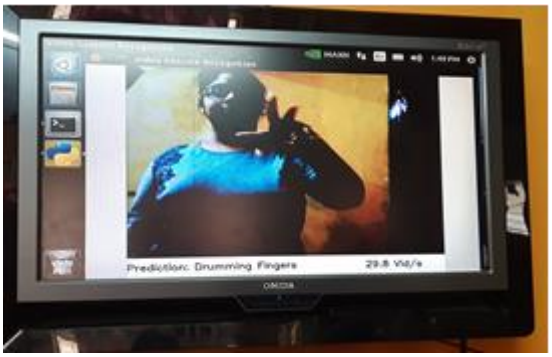

Fig 5: Drumming Fingers

Fig.6: Zooming in with full hand

When all the four fingers are together moved closer towards thumb finger it was depicted as “zooming in with full hand” as shown in fig.6

When the four finger and the middle fingers alone are moved from one side to the other the system predicted as “sliding two fingers left” and “sliding two fingers right” as shown in Fig.7 and Fig.8.

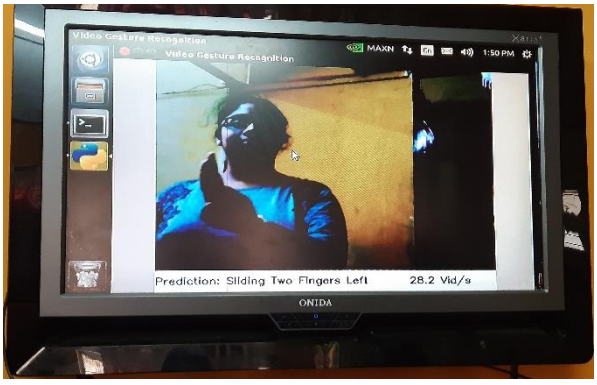

Fig 7: sliding two fingers left

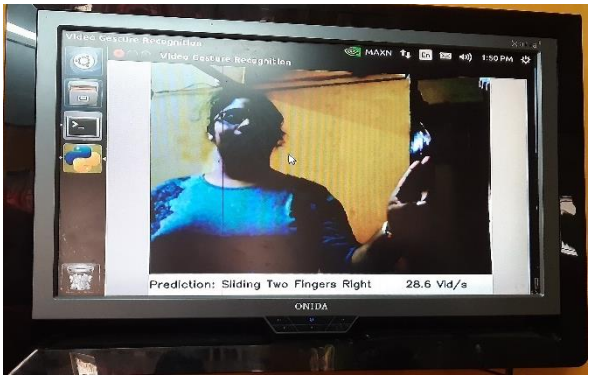

Fig 8: sliding two fingers right

Further the model can be developed and integrated to various real time applications such as it can substitute the traditional keyboard and mouse with natural hand gesture command inputs predicted by our model also can help greatly to speech, hearing retarded people by acting as an intermediate aid between them and normal people.

5.CONCLUSION

In the present world there are numerous was of providing input to the smart devices such as a smart tv, virtual environments used for various purposes, home automation systems etc although these are not feasible to the trending and upcoming technologies, thus to meet this standards in this paper we put forward a hardware efficient dynamic gesture recognition system, Which could be a boon to the differently abled people as it lays a natural form to communications with hand movements to the smart devices. The model uses temporal shift model combined with CNN techniques to extract spatio-temporal

features for efficient action detection. A data set of around 16 hand gestures can be predicted by the proposed prototype, also we are further planning to implement this model on a PowerPoint presentation where the presenter can control the movement of the slides it just by using hand gestures without any other support

REFERENCES

[1] El Hayek, H., Nacouzi, J., Kassem, A., Hamad, M., & El-Murr,"Sign to letter translator system using a hand glove", In The Third International Conference on e-Technologies and Networks for Development April,2014 IEEE.

[2] Tubaiz, Noor, Tamer Shanableh, and Khaled Assaleh. "Glove-based continuous Arabic sign language recognition in user-dependent mode." IEEE Transactions on Human-Machine Systems 45, no. 4 (2015)

[3] Praveen, Nikhita, Naveen Karanth, and M. S. Megha. "Sign language interpreter using a smart glove." In 2014 International Conference on Advances in Electronics Computers and Communications, pp. 1-5. IEEE, 2014.

[4] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Largescale video classification with convolutional neural networks. In: CVPR, pp. 1725–1732 (2014)

[5] Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H."Real-time action recognition with enhanced motion vector CNNs," In CVPR, pp. 2718–2726 (2016)

[6] Limin Wang, YuanjunXiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. ,"Temporal segment networks: Towards good practices for deep action recognition," In European Conference on Computer Vision, pages 20–36. Springer, 2016.

[7] Du Tran, LubomirBourdev, Rob Fergus, Lorenzo Torresani, and Manohar Palur,"Learning spatiotemporal features with 3d convolutional networks," In Proceedings of the IEEE inter- national conference on computer vision, pages 4489–4497, 2015.

[8] Lin, Ji, Chuang Gan, and Song Han. "Tsm: Temporal shift module for efficient video understanding." In Proceedings of the IEEE International Conference on Computer Vision, pp. 7083-7093. 2019..

[9] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," 2013.

[10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, 2010.

[11] Hakim, NoorkholisLuthfil, Timothy K. Shih, KasthuriArachchi, SandeliPriyanwada, Wisnu Aditya, Yi-Cheng Chen, and Chih-Yang Lin. "Dynamic Hand

Gesture Recognition Using 3DCNN and LSTM with FSM Context-Aware Model." Sensors 19, no. 24 (2019): 5429.

[12] Dubey, Kundan Kumar, Ajitanshu Jha, Akshay Tiwari, and K. Narmatha. "Hand Gesture Movement Recognition System Using Convolution Neural Network Algorithm." International Research Journal of Computer Science (IRJCS) Issue 4.

[13] Zhou, Bolei, Alex Andonian, Aude Oliva, and Antonio Torralba. "Temporal relational reasoning in videos." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 803-818. 2018.

[14] Phuong, Huy Nguyen, and Mai Thuong Duong Thi. "An Approach in Building a Vision-Based Hand Gesture Recognition System."

[15] Rautaray, Siddharth S., and Anupam Agrawal. "Vision based hand gesture recognition for human computer interaction: a survey." Artificial intelligence review 43, no. 1 (2015): 1-54.

[16] Panwar, Meenakshi, and Pawan Singh Mehra. "Hand gesture recognition for human computer interaction." In 2011 International Conference on Image Information Processing, pp. 1-7. IEEE, 2011.

[17] Haria, Aashni, Archanasri Subramanian, NivedhithaAsokkumar, Shristi Poddar, and Jyothi S. Nayak. "Hand gesture recognition for human computer interaction." Procedia computer science 115 (2017): 367-374.

[18] Manresa, Cristina, Javier Varona, Ramon Mas, and Francisco J. Perales. "Hand tracking and gesture recognition for human-computer interaction." ELCVIA Electronic Letters on Computer Vision and Image Analysis 5, no. 3 (2005): 96-104.

[19] Soumyalatha Naveen, Manjunath R Kounte, “Key Technologies and Challenges in IoT Edge Computing”, 3rd International Conference on IoT in Social, Mobile, Analytics and Cloud (ISMAC 2019), Palladam, India, 12-14 Dec 2019, pp 178-183.

[20] Shridevi Jeevan Kamble, Manjunath R Kounte, “On Road Intelligent Vehicle Path Prediction and Clustering Using Machine Learning Approach”, 3rd International Conference on IoT in Social, Mobile, Analytics and Cloud (ISMAC 2019), Palladam, India, 12-14 Dec 2019, pp 487-491.

[21] Helen K Joy, Manjunath R Kounte, “ An Overview of Traditional and Recent Trends in Video Processing”, 2nd International Conference on Smart Systems and Inventive Technology, Thirunelveli, India 27-29 Nov 2019, pp. 848-851.

[22] Teja K, Shravani M B, C YashwanthSimha, Manjunath R Kounte, ”Secure Voting Through Blockchain Technology” 3rd International Conference on Trends in Electronics and Informatics (ICOEI 2019), Tirunelveli, Tamil Nadu, India, 23-25 April 2019.

[23] Harshini V M, Shreevani Danai, Usha H R, Manjunath R Kounte, ”Health Record Management through Blockchain Technology” 3rd International Conference on Trends in

Electronics and Informatics (ICOEI 2019), Tirunelveli, Tamil Nadu, India, 23-25 April 2019.

[24] ChintarlapallireddyYaswanthSimha, Harshini V M, L V S Raghuvamsi, Manjunath R Kounte, "Enabling Technologies for Internet of Things & It's Security issues", Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018), Madurai, India, 14-15 June 2018, pp 1849-1852

[25] Naveen S., Kounte M.R. (2020) In Search of the Future Technologies: Fusion of Machine Learning, Fog and Edge Computing in the Internet of Things. In: Pandian A., Senjyu T., Islam S., Wang H. (eds) Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018). ICCBI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 31. Springer, Cham

[26] Kamble S.J., Kounte M.R. (2020) Enabling Technologies for Internet of Vehicles. In: Pandian A., Senjyu T., Islam S., Wang H. (eds) Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018). ICCBI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 31. Springer, Cham

[27] Niveditha. E, Chaitra. H, K. Afrose and S. Thomas, "A Novel Circularly Polarized Multiple Elliptical Printed Monopole Antenna," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 382-385, doi: 10.1109/ICSSIT46314.2019.8987799