INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 12, Issue 6, (2022) E-ISSN: 2222-6990

Translating Hand Gestures Using 3D Convolutional Neural Network

Farah Yasmin Abdul Rahman[1], Amirul Asyraf Kamaruzzaman[1], Shahrani Shahbudin[1], Roslina Mohamad[1], Nor Surayahani Suriani[2] and Saiful Izwan Suliman[1]

[1]Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia, [2]Department of Electronic Engineering, Faculty of Electrical and Electronics Engineering, Universiti Tun Hussein Onn, 84600 Batu Pahat, Johor, Malaysia.

Corresponding Author Email: farahy@uitm.edu.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v12-i6/13989 DOI:10.6007/IJARBSS/v12-i6/13989

Published Date: *06* June 2022

Abstract

Hand gestures are one of the mediums that many people use to communicate with each other. The use of gesture recognition applications has become increasingly popular in recent years especially in computer vision areas. Typically, gestures can easily be recognized from a single image frame (i.e. alphabet from sign language), however the ability to recognize complex gestures with subtle differences between movement requires more works and larger datasets. In this work, we introduce a simple gesture recognition system that translates 5 different hand gestures, namely "doing other things", "swiping down", "swiping left", "zooming out with two fingers" and "drumming fingers". We used datasets obtained from Jester dataset. The inputs were processed in 'RGB' format during the pre-processing phase and a spatiotemporal filter were used as a feature extraction method, which were also the main building block in this system. Next, we trained the features using 3D Convolution Neural Network (3D-CNN). Further, we used real-time video to test the developed recognition system with 5 different actors. Findings show that the developed model can translate hand gestures with accuracy of 85.70% and 0.4% losses.

Keywords: Gesture Recognition, 3D Convolutional Neural Network, Hand Gesture, Translate

Introduction

The growing number of individuals suffer from the inability to speak due to biological and/or physical disabilities that encourage them to choose to use sign language in order to communicate with others. Most normal people find difficulties in learning sign language to communicate with the disable people. Gesture recognition system is ongoing research in computer vision. The gesture recognition system can be used to increase human to human interaction because hand gestures have become the same part of communication as language and expression.

Besides encouraging sign language and its automatic processing, hand recognition system has a wide range of usage scope in various industries. It has the potential to be used to control devices in human interfaces and applications in sectors such as automation, home automation, public transit and etc.

Over the years, gesture recognition has been used to develop various technologies. General recognition (Haba et al., 2018), sign language usages (Bhaskaran, 2017), and usage in gaming and gaming features (Wilk et al., 2018), have been proposed and carried out through the use of wearable sensor devices. These devices combine several precise and efficient built-in sensors detecting various types of formations, speed, hand positioning, etc. The disadvantage of using these approaches are that there is a need for devices that have these sensors incorporated into them, as well as the capital requirements to buy these sensors and capable of integration between sensors and devices. The computer vision approach removes the requirement for such devices (besides the use of a camera) but needs large amounts of data to be able to accurately and efficiently train systems that can generalize to scenarios that it has not seen before. There is a lot of work involving complex hand segmentations (Sharp, et al., 2015; Raheja et al., 2017) or joint segmentation (Tkach et al., 2016; Smedt et al., 2017; Zhi, 2018).

One of the earliest works in this particular field was carried out by A. Utsumi, T. Miyasato, and F. Kishino, published in 1995, where they utilized multiple cameras to create a hand pose recognition system using skeleton hands (Utsumi et al., 1995). A renowned work that appeared and became recognized and acknowledged in most papers up to this day is the use of curvature scale space in a hand pose recognition system (Chang et al., 2002), which became the start of the use and idea of spatial features. Their work achieved about 98.3% recognition rate to identify 6 different hand poses. In subsequent years, more people explored the idea of hand recognition on various features. An approach of using a new feature type and HMM gave recognition rate of 96% was proposed by Bao et al., published in 2009 (Bao et al., 2009), while the used of Motion History Histograms was proposed in (Meng et al., 2009). The usage of optical flow and motion vector in moving object tracking was proposed in (Kale et al., 2015), which claimed to have accurate and robust results over different types of real-time and standard inputs.

Sepehri et al. used gesture recognition in a virtual environment. They recommend hand usage as an interface device for controlling and drawing (Sepehri, Yacoob, & Larry, 2006). A novel interactive method of virtual reality systems was proposed by Zhao et al., which achieved a recognition rate of 96% (Zhao et al., 2009),

Further, there are studies that have yielded satisfactory results and mostly involve the use of expensive and high-quality technologies. For example, Smedt et al (Smedt et al., 2017) used Intel RealSense short-range depth camera, and wielded good recognition rates. Microsoft Kinect sensor was used in Ravi et al, (Ravi, et al., 2019).

There are also studies that propose sign language recognition based on spoken language. Mahmood et al. (Mahmood et al., 2018) proposed a hand gestures recognition system for Kurdish Sign Language using two lines of features. Ahuja et al (Ahuja, et al., 2019)

and Pardasani et al. (Pardasani et al., 2018) presented their sign language recognition studies for American Sign Language.

In this work, we have proposed a system of hand gesture recognition with usage of only one camera that making the system cost-efficient. We used large datasets that encompasses gestures videos as opposed to the static images used in the proposed sign language recognition. We also used features, such as spatio-temporal filters and mean pixel value functions, that do not require complex alterations and dataset processing. 3D convolutional neural network was used to train, test, and validate the features.

Methodology

This section discusses the method used in this research. It is divided into four main stages, which are data acquisition, pre-processing, feature extraction, and finally recognition and classification.

A. Data Acquisition

In this work, we used the Jester Dataset (Materzynska et al., 2019) - a large-scale gesture recognition real-world video dataset – obtained from 20BN. It comprises of 148,092 short clips of videos. The videos show people conducting gestures in front of the camera containing set of 27 gestures that was recorded by 1,376 actors. In this study, we only focused on 5 actions to reduce the load during training. Fig. 1 shows the list of classes used in this work. The datasets were divided into train, validation, and test sets with a ratio of 8:1:1. Split datasets were created to ensure that video occurrences in training and fractional tests from the same actors involved in Jester's datasets did not occur. The total number of videos for each class used in this work from the Jester dataset is also depicted in Fig. 1. Meanwhile, the testing datasets used were real-time videos with 5 actors performing the hand gestures. The videos were taken at frontal view using 32 Mega pixel digital camera. In this study, 5 students were selected as actors to perform the hand gestures in the real-time video. User 1, User 3 and User 5 were male actors aged between 20 – 23 years old and User 2 and User 4 were female actors aged between 20 – 23 years old as well.

Fig.1: Overview of Gesture Classes Taken from Jester Dataset

B. Pre-processing

The pre-processing procedure involved having to transform the data into a form that could be more effectively and effortlessly processed, which the model can benefit from. The aim

was to suppress unwanted distortions and/or enhance some important image features. We converted the video into frames and processed them in 'RGB' format as opposed to Greyscale images. This was mainly due to the fact that the multiple videos and images in the dataset were from various people who took the videos in various environments (i.e. difference in skin colour, darker and/or lighter backgrounds, etc.). This presented the solution to get detailed information as to the nuances of the movements that were being made by the actors. The images were then cropped to standardize the images size throughout the dataset.

C. Feature Extraction

Feature extractions for image processing were divided into two stages; temporal feature extraction and spatial feature extraction. The main building block for this network, as previously described in (Tran et al., 2015) used spatio-temporal filters. A natural representation of spatio-temporal data was provided by these operations. Other than that, a mean pixel value technique on the 'RGB' channels were applied to the 'RGB' images. This allows the input images to be smoother and makes the identifying of gestures more efficient for the model.

D. Recognition and Classification

In this work, we proposed a 3D convolutional neural network (3D-CNN) to be used as the baseline model as proven to be efficient and fast in recognition (Li et al., 2015). In the following, a 3D convolutional layer refers to a convolutional block, which was followed by ELU non-linearity and batch normalization layer. The model consists of four 3D convolutional blocks and a max-pooling layer. A spatial max-pooling layer was then applied in the end. Then, the output of the last layer was then passed through a fully connected layer, an ELU activation factor layer, and finally another fully connected layer. The 3D-CNN architecture used in this study is shown in Fig. 2.

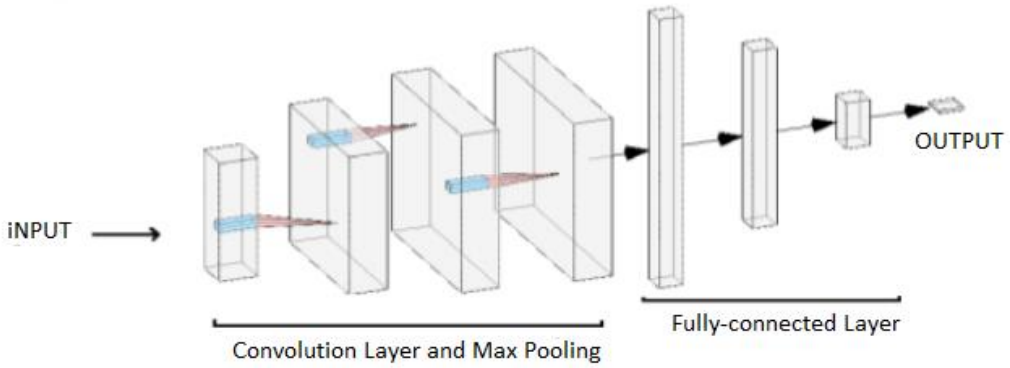

Fig.2: 3D-CNN Architecture Used in This Study

Our model was trained using SGD with a learning rate of 0.001. Top-k function used to determine the accuracy of models and the results of the top 1 accuracy were recorded. The network architecture of the system can be seen in Table I. Note that all layers used filter size of (3,3,3).

Table I
Network Architecture Model the Proposed 3D-CNN Model

Layer	Layer type	Hyperparameters
1	conv3D	64
2	max pool	(1, 2, 2)
3	conv3D	128
4	max pool	(2, 2, 2)
5	conv3D	256
6	max pool	(2, 2, 2)
7	conv3D	256
8	max pool	(2, 2, 2)
9	fully connected	12800
10	ELU	512
11	fully connected	512

Results and Discussion

This section presents the results obtained in this study. The training for each model used 1,500 videos per class. Top-k function was applied to determine the accuracy of the model; if the predictions were in the first (top-1) of predictions or amongst the first five (top-5) of the predictions being made by the model. The highest accuracy for the top-1 prediction was recorded. Fig. 3 shows the accuracy of the developed CNN model during training stages. Based on Fig. 3, we can see that the developed CNN network can achieve more than 80% accuracy. Therefore, the model was kept for testing and validation stage.

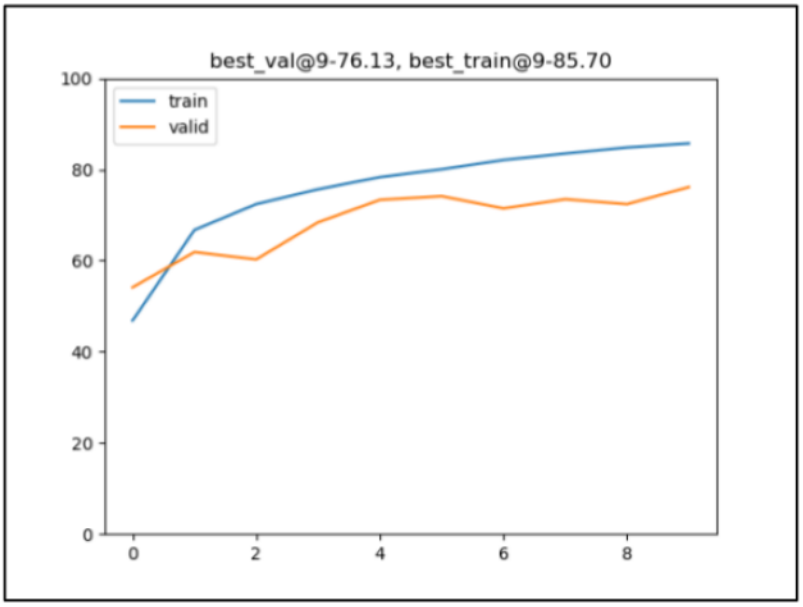

Fig.4: Initial prototype setup

Next, this study tested and validated the developed system with the real-time videos. The actors performed hand gestures in front of a camera that attached to the computer, and the system will inform the type of hand gestures that the actors have performed. Actors were allowed to try each gesture 5 times until the system correctly detects gestures, before switching to the next gesture. Scores were given based on the number of times the user had

to try the gesture before the system could detect the gesture being performed (Score 1 means has tried only once, 1/2 has tried twice, 1/3 after trying three times, etc.). Table II depicts the scoring used in this study.

Table II
Scoring System Used in Real-Time Testing

Number of Tries	Score
1	1
2	1/2
3	1/3
4	1/4
5	1/5

The results of the experiment are tabulated in Table III. Based on Table III, the hand gesture “Doing Other Things” is the highest recognition rate among other hand gestures by 100%, followed by “Zooming Out with Two Fingers” with 90% and “Swiping Down” with 80%. The system can only recognize the hand gesture "Zooming Out with Two Fingers" by 40%, the lowest rate of recognition among 5 hand gestures tested.

Table III
Network Architecture Model for the Proposed 3D-CNN Model

User/ Gesture	Doing Other Things	Swiping Down	Swiping Left	Zooming Out with Two Fingers
1	1	1/2	1	1/2
2	1	1	1	1/3
3	1	1/2	1/2	1/3
4	1	1	1	1/2
5	1	1	1	1/3
Total (%)	**100**	**80**	**90**	**40**

Fig. 4 shows the Loss Plot of the developed 3D-CNN system to validate the findings. The results show that there are descending values for both training and validation losses, with validation loss having a gap with the training one, and both stabilized. This proved that the developed CNN model is good and suitable to use (Stack Exchange, 2019)

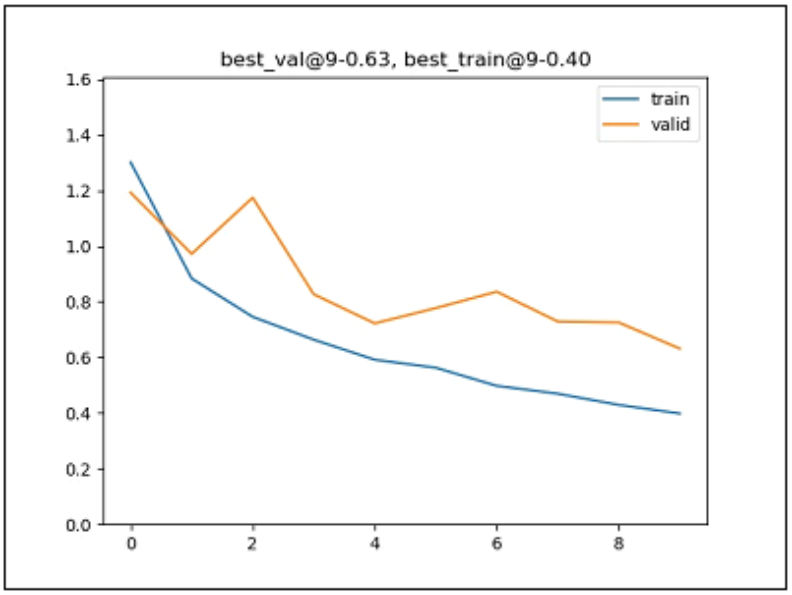

Fig.5: Loss Plot of the Developed 3D-CNN Model

Further, Table IV shows the overall performance of the developed 3D-CNN system to translate real-time videos that consist of 5 types hand gestures based on offline datasets. From the table, the system can provide accuracy with 85.70% and a minimal loss of 0.40%.

Table IV
Overall Performance of the Developed System

Accuracy (%)	Loss (%)
85.70	0.40

Conclusion

In conclusion, the 3D-CNN system has been successfully developed to detect and recognize 5 types of hand gestures. The training datasets were obtained from Jetset Dataset and the testing datasets were real-time video. The system was developed by processing all the images into 'RGB' images and extracting spatio-temporal features from the 3D convolution of the images. These two methods allow us to retain most of the information and data from various compilation images (video). The system was successfully analyzed by the 3D-CNN model, where the accuracy during training using offline datasets and testing in real-time were recorded, as well as the loss of the system. Findings show that the developed 3D-CNN system successfully translated the real-time hand gestures with a recognition rate of 85.70% and 0.4% loss. For future studies, the extracted features will be trained and compared with other recognition tools such as Artificial Neural Network and Support Vector Machine.

References

Ahuja, R., Jain, D., Sachdeva, D., Garg, A., & Rajput, C. (2019). Convolutional Neural Network Based American Sign Language Static Hand Gesture Recognition. *International Journal of Ambient Computing and Intelligence*, 60-73.

Bao, P. T., Binh, N. T., & Khoa, T. D. (2009). A New Approach to Hand Tracking and Gesture Recognition by a New Feature Type and HMM. *Sixth International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 3-6). 2009: IEEE.

Chang, C.-C., Chen, I.-Y., & Huang, Y.-S. (2002). Hand pose recognition using curvature scale space. *Object recognition supported by user interaction for service robots. 2*, pp. 386-389. Quebec City, Quebec, Canada: IEEE.

Zhi, D. T. E. (2018). Teaching a Robot Sign Language using Vision-Based Hand Gesture Recognition. *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)* (pp. 1-6). Ottawa, ON: IEEE.

Haba, C.-G., Breniuc, L., Ciobanu, R., & Tudosa, I. (2018). Development of a wireless glove based on RFID Sensor. *International Conference on Applied and Theoretical Electricity ICATE 2018.* Craiova, Romania.

Bhaskaran, A. G. K. A. (2017). Smart gloves for hand gesture recognition: Sign language to speech conversion system. *International Conference on Robotics and Automation for Humanitarian Applications (RAHA).* Kollam, India: IEEE.

Kale, K., Pawar, S., & Dhulekar, P. (2015). Moving object tracking using optical flow and motion vector estimation. *4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)* (pp. 1-6). Noida: IEEE.

Li, H., Lin, Z., Shen, X., & Brandt, J. (2015). A convolutional neural network cascade for face detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5325-5334). IEEE.

Mahmood, M. R. A. M. (2018). Dynamic Hand Gesture Recognition System for Kurdish Sign Language Using Two Lines of Features. *International Conference on Advanced Science and Engineering (ICOASE)* (pp. 42-47). Duhok: IEEE.

Materzynska, J., Berger, G., Bax, I., & Memisevic, R. (2019). The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 2874-2882). Seoul, Korea (South): IEEE.

Meng, H., Pears, N., Freeman, M., & Bailey, C. (2009). *Motion History Histograms for Human Action Recognition.*

Pardasani, A., Sharma, A. K., Banerjee, S., Garg, V., & Roy, D. S. (2018). Enhancing the Ability to Communicate by Synthesizing American Sign Language using Image Recognition in A Chatbot for Differently Abled. *7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 529-532). Noida, India: IEEE.

Raheja, J., Chandra, M., & Chaudhary, A. (2017). 3D Gesture based Real-time Object Selection and Recognition. *Pattern Recognition Letters*.

Ravi, S., Suman, M., Kishore, P., Eepuri, K., Maddala, T., & Kumar, A. D. (2019). Multi Modal Spatio Temporal Co-Trained CNNs with Single Modal Testing on RGB – D based Sign Language Gesture Recognition. *Journal of Computer Languages, 52*, 88–102.

Sepehri, A., Yacoob, Y., & Larry, D. (2006). Employing the Hand as an Interface Device. *Journal of Multimedia, 1*, 18-29.

Sharp, T., Wei, Y., Freedman, D., Kohli, P., Krupka, E., Fitzgibbon, A., Vinnikov, A. (2015). Accurate, robust, and flexible realtime hand tracking. *The 33rd Annual ACM Conference* (pp. 3633-3642). ACM.

Smedt, Q., Wannous, H., Vandeborre, J., Guerry, J., & Le Saux, B. (2017). 3D hand gesture recognition using a depth and skeletal dataset: SHREC'17 track. *3Dor '17: Proceedings of the Workshop on 3D Object Retrieval* (pp. 33 - 38). ACM.

Stack Exchange. (2019). *Data Science*. Retrieved Sept 1st, 2020, from Understanding Training and Test Loss Plot: : https://datascience.stackexchange.com/questions/52028/understanding-training-and-test-loss-plots

Tkach, A., Pauly, M., & Tagliasacchi, A. (2016). Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics, 35*(6).

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV)* (pp. 4489-4497). Santiago: IEEE.

Tsai, C.-Y., & Lee, Y.-H. (2011). The parameters effect on performance in ANN for hand gesture recognition system. *Expert Syst. Appl.*, 7980-7983.

Utsumi, A., Miyasato, T., & Kishino, F. (1995). Multi-camera hand pose recognition system using skeleton image. *4th IEEE International Workshop on Robot and Human Communication* (pp. 219-224). Tokyo, Japan: IEEE.

Wilk, M. P., Torres-Sanchez, J., Tedesco, S., & O'Flynn, B. (2018). Wearable Human Computer Interface for Control Within Immersive VAMR Gaming Environments Using Data Glove and Hand Gestures. *IEEE Games, Entertainment, Media Conference (GEM).* Galway, Ireland.

Zhao, S., Tan, W., Wu, C., Liu, C., & Wen, S. (2009). A novel interactive method of virtual reality system based on hand gesture recognition. *Chinese Control and Decision Conference* (pp. 5879-5882). Guilin: 2009.