Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 3,January 2023

To Implement Hand Gesture Recognition Using Convolutional Neural Networks

Ms KANNAIAHGARI SRAVYA[1], Ms CHENNA NAVYA[2]
Mr YELISHALA YUDHISTER[3], KANDE ARCHANA[4]

Department of Computer Science and Engineering,
Malla Reddy Institute of Engineering and Technology,
Hyderabad, Telangana state, India

Abstract

Automatic human gesture recognition from camera images is an interesting topic for developing intelligent vision systems. Here, we propose a convolution neural network (CNN) method to recognize hand gestures of human task activities from a camera image. To achieve the robustness performance, the skin model and the calibration of hand position and orientation are applied to obtain the training and testing data for the CNN. Since the light condition seriously affects the skin color, we adopt a Gaussian Mixture model (GMM) to train the skin model which is used to robustly filter out non-skin colors of an image. The calibration of hand position and orientation aims at translating and rotating the hand image to a neutral pose. Then the calibrated images are used to train the CNN. In our experiment, we provided a validation of the proposed method on recognizing human gestures which shows robust results with various hand positions and orientations and light conditions. Our experimental evaluation of seven subjects performing seven hand gestures with average recognition accuracy around 95.96% shows the feasibility and reliability of the proposed method.

Keywords

Convolutional Neural Networks (CNN), Gaussian Mixture Model (GMM), Robust, Calibration, Hand Gestures.

1. INTRODUCTION

Sign language, as one of the most widely used communication means for hearing-impaired people, is expressed by variations of hand- shapes, body movement, and even facial expression. Since it is difficult to collaboratively exploit the information from hand-shapes and body movement trajectory, sign language recognition is still a very challenging task. This paper proposes an effective recognition model to translate sign language into text or speech in order to help the hearing impaired communicate with normal people through sign language.

Technically speaking, the main challenge of sign language recognition lies in developing

descriptors to express hand-shapes and motion trajectory. In particular, hand-shape description involves tracking hand regions in video stream, segmenting hand-shape images from complex background in each frame and gestures recognition problems. Motion trajectory is also related to tracking of the key points and curve matching. Although lots of research works have been conducted on these two issues for now, it is still hard to obtain satisfying result for SLR due to the variation

and occlusion of hands and body joints. Besides, it is a nontrivial issue to integrate the hand-shape features and trajectory features together. To address these difficulties, we develop a CNNs to naturally integrate hand- shapes, trajectory of action and facial expression. Instead of using commonly used colour images as input to networks like [1, 2], we take colour images, depth images and body skeleton images simultaneously as input which are all provided by Microsoft Kinect.

Kinect is a motion sensor which can provide colour stream and depth stream. With the public Windows SDK, the body joint locations can be obtained in real-time

as shown in Fig.1. Therefore, we choose Kinect as capture device to record sign words dataset. The change of colour and depth in pixel level are useful information to discriminate different sign actions. And the variation of body joints in time dimension can depict the trajectory of sign actions. Using multiple types of visual sources as input leads CNNs paying attention to the change not only in colour, but also in depth and trajectory. It is worth mentioning that we can avoid the difficulty of tracking hands, segmenting hands from background and designing descriptors for hands because CNNs have the capability to learn features automatically from raw data without any prior knowledge [3].

CNNs have been applied in video stream classification recently years. A potential concern of CNNs is time consuming. It costs several weeks or months to train a CNNs with million-scale in million videos. Fortunately, it is still possible to achieve real-time efficiency, with the help of CUDA for parallel processing. We propose to apply CNNs to extract spatial and temporal features from video stream for Sign Language Recognition (SLR). Existing methods for SLR use hand- crafted features to describe sign language motion and build classification model based on these features. In contrast, CNNs can capture motion information from raw video data automatically, avoiding designing features. We develop a CNNs taking multiple

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 3,January 2023

ypes of data as input. This architecture integrates colour, depth and trajectory information by performing convolution and subsampling on adjacent video frames. Experimental results demonstrate that 3D CNNs can significantly outperform Gaussian mixture model with Hidden Markov model (GMM-HMM) baselines on some sign words recorded by ourselves.

1.1 EXISTING SYSTEM

This work is a CNN-based human hand gesture recognition system. CNN is a research branch of neural networks. Using a CNN to learn human gestures, there is no need to develop complicated algorithms to extract image features and learn them. Through the convolution and sub-sampling layers of a CNN, invariant features are allowed with little dislocation. To reduce the effect of various hand poses of a hand gesture type on the recognition accuracies, the principal axis of the hand is found to calibrate the image in this work. Calibrated images are advantageous to a CNN to learn and recognize correctly.

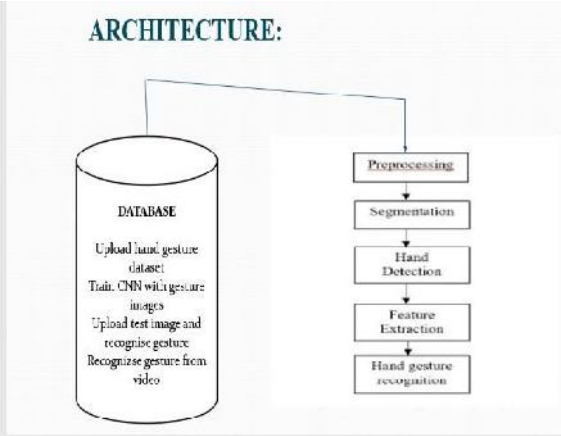
1.2 PROPOSED SYSTEM

From the camera image input, the hand is extracted by skin color segmentation. The skin model is trained by a Gaussian Mixture model to classify skin color and non-skin color. After that, the calibration of hand position and orientation is used to translate and rotate the hand image to a neutral pose.

The calibrated image is fed to the CNN to train or test the network. For continuous hand motion, the post-processing is used to filter out the noises of the results from the CNN.

2. LITERATURE SURVEY

Kamal et al. [29] proposed a pattern recognition method for static recognition, which is able to handle the low variability among the different gestures. Authors have used shape geodesics and robust registration for calculating the accelerated time. The proposed system is evaluated by considering three distances of the shape geodesics, and the experiment results showed that the proposed model is efficient than the other related methods.



Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 3,January 2023

Fig.1 Proposed Architecture

Wei et al. [51] proposed a multi-view deep learning model by relating classical surface electromyography (sEMG) feature sets with a CNN-based deep learning model to recognize the gestures. The multi-view model mainly emphasized on the parallel functioning of CNN multi-streams and training of the network with deep feature sets of sEMG gestures. Experiments were conducted with
11 different databases of sEMG, and results
shown that the multi-view model performs exceptionally well on the dissimilar data streams of sEMG.

Tan et al. [47] proposed a static gesture recognition model using electromagnetic fields. This model primarily focuses on vision-based recognition and provides training with CNN by an end-to-end recognizer. The proposed model was tested with the various datasets of static hand gesture images and achieved 99% recognition rate for full aperture, and for one-eight aperture, the accuracy is 95.32%. Results outperformed even for the limited aperture and also had improved scalability on the gesture images.

Hu et al. [14] proposed a hand gesture recognition system to control the unmanned aerial vehicles (UAV). The entire model has been trained and tested with the various layers of deep learning neural networks like 2-layer and 5-layer fully connected neural networkand a CNN of 8 layers. The experimental results proved that the efficiency is better than the existing systems and achieved an average accuracy of 96.7% for 2 layers and 98% for 5 layers. Finally, CNN with 8 layers attained
89.6% and 96.9% on scaled and non-scaled datasets.

Okan et al. [22] proposed a model that works for the video hand gesture recognition. CNN is used to classify and detect the number of gestures and also in evaluating single-time activations. Two datasets NVIDIA and Ego Gesture were used in calculating the efficiency of the gestures and achieved an accuracy of 94.03 %. The model was very well extended for the sliding window approach, and the results are outperformed compared to the existing video recognition systems.

Sruthy et al. [45] proposed a CNN-based hand gesture recognition framework for capturing various hand gestures. The deep convolutional neural network [25, 26, 50] used in this work to classify the gestures and train the two spectrograms of the Doppler radar capable. The proposed model got trained by CNN, and testing was done in two phases in producing the quadrature components. The experimental results proved that the proposed architecture has a good accuracy of 95% compared to the other models.

Pinto et al. [33] proposed a gesture recognition-based model using convolutional neural networks. This method mainly focuses on the preprocessing steps like polygon filter and segmentation process of the various gestures. Using convolutional neural networks, the training and testing part has been carried out by 60% and 40%. The results are analyzed both in the testing and training processes, and the calculated metrics show that the proposed

model is robust than the existing methodologies.

Li et al. [24] proposed CNN-based hand gesture recognition framework where the number of gestures is characterized by the neural network and error backpropagation algorithm. In this model, the recognition of gestures and extracting its features were labeled by unsupervised learning approaches. Further, support vector machine was considered to examine the best possible gestures from the optimized dataset. It has proved that the proposed system shows a high accuracy by means of classification of gestures in static and dynamic representation.

Ahmed et al. [2] proposed a novel method of recognition of gestures by finger counting using convolutional neural networks. It provides an immersive experience to the gesture handling people, and researchers used it for an alternative approach in accessing the optimal location of a gesture recognizer. Proposed model impulses the finger counting and labels to the sensors and motions of a human body. This model gives better accuracy over the other frameworks and performs a stable recognition for real-world applications.

Jiang et al. [18] proposed a vision-based recognition method using convolutional neural networks. It aims to perform the best possible hand gestures of a human body by means of keletonization algorithm and CNN. Here, the gesture recognition process was carried out by

the spatial coordinate system and sparse representation. The model has been trained and tested by the American Sign Language database and the results showed that the proposed model is having a high recognition rate of 96.01% to the existing frameworks.

Jinxian et al. [35] proposed a system to identify hand gestures using EMG signals and PCA and Generalized regression neural network (GRNN). The model is processed with nine static gestures and extracted the important human emotions. It is further improvised for the real-time recognition of human emotions and reduced the signal dimension. Finally, the proposed model showed overall recognition rate as 95% after dimensionality reduction and training with neural network and gave the better average recognition compared to the existing approaches.

Chen et al. [7] proposed a deep neural network model for recognizing the hand gestures using CNN through surface electromyography signals. As the proposed model progress the accuracy in classification and also diminishes the various parameters compared to the existing hand gesture recognition methods. Classification accuracy process was done by the classical machine learning methods and executed on the Dataset Myo. Further, the model provides better results with sEMG signals and also provides the classification of sEMG signals along with the CNN architecture.

3.MODULES

The overall system consists of two parts, back
end and front-end. The back-end system consists of three modules: Camera module, Detection module and Recognition module .

3.1 Web Camera

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 6- Issue 1, Paper 3,January 2023

This module is responsible for connecting
and capturing input through the different types of image detectors and sends this image to the detection module for processing in the form of frames.

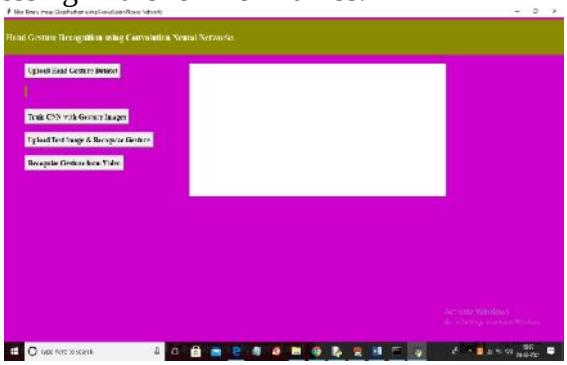


Fig: 2 Capture the image

The commonly used methods of capturing input are data gloves, hand belts and cameras. In our system, we use the webcam inbuilt which is cost efficient to recognize both static and dynamic gestures. The system has suitable provision to allow input from a USB based webcam as well but this would require some expenditure from the user. The image frames obtained are in the form of a video.

3.2 Detection gesture

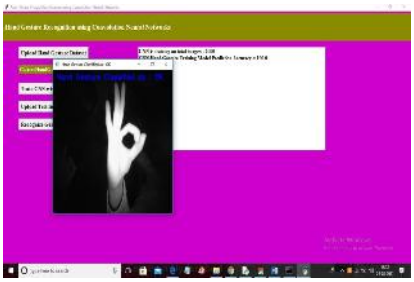


Fig: 3 Detect the Image

This module is responsible for the image processing. The output from camera module is subjected to different image processing techniques such as colour conversion, noise removal, thresholding following which the image undergoes contour extraction. If the image contains defects, then convexity defects are found according to which the gesture is detected. If there are no defects, then the image is classified using Haar cascade to detect the gesture.

In the case of dynamic gestures, the detection module does the following; If Microsoft PowerPoint has been launched with a slideshow being enabled and the webcam detects palm in movement, for 5 continuous frames then the dynamic gesture swipe is detected.

3.3 Recognition

This module is responsible for mapping the detected hand gestures to their associated

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 3,January 2023

actions. These actions are then passed to the appropriate application. The front end consists of three windows. The first window consists of the video input that is captured from the camera with the corresponding name of the gesture detected. The second window displays the contours found within the input images. The third window displays the smooth thresholded version of the image. The advantage of adding the threshold and contour window as a part of the GraphicalUser Interface is to make the user aware of the background inconsistencies that would affect the input to the system and thus they can adjust their laptop or desktop web camera in order to avoid them. This would result in better performance.

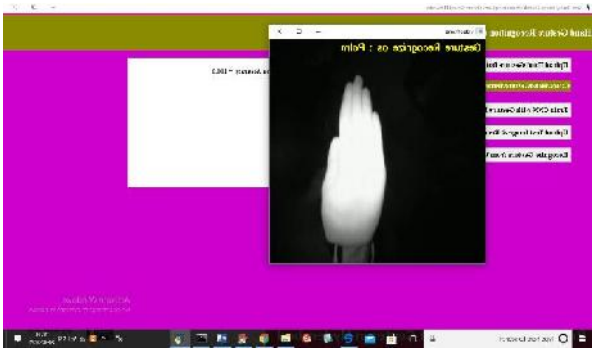


Fig: 4 Recognize the Image

Conclusion

we developed a CNN-based human hand gesture recognition system. The salient feature of the system is that there is no need to build a model for every gesture using hand features such as fingertips and contours. To have robust performance, we applied a GMM to learn the skin model and segment the hand area for recognition. Also, the calibration of the hand pose was used to rotate and shift the hand on the image to a neutral pose. Then, a CNN was trained to learn seven gesture types in this paper. In the experiments, we conducted 4-fold cross-validation on the system where 600 and 200 images from a subject were used to train and test, respectively and the results showed that the average recognition rates of the seven gesture types were around 99%. To test the proposed method on multiple subjects, we trained and tested the hand images of the seven gesture types from seven subjects. The average recognition rate was 95.96%. The proposed system also had the satisfactory results on the transitive gestures in a continuous motion using the proposed rules. In the future, a high- level semantic analysis will be applied to the current system to enhance the recognition capability for complex human tasks.

References

[1]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[2]. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, “Large-scale video classification with convolutional neural networks,” in CVPR, 2014. [3] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick ´ Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6- Issue 1, Paper 3,January 2023

[3]. Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio, "A biologically inspired system for action recognition," in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. Ieee, 2007, pp.
1–8. [5] Shuiwang Ji, Wei Xu, Ming Yang,
and Kai Yu, "3D convolutional neural networks for human action recognition," IEEE TPAMI, vol. 35, no. 1, pp. 221–231,
2013.

[4]. Kirsti Grobel and Marcell Assan, "Isolated sign language recognition using hidden Markov models," in Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on. IEEE, 1997, vol. 1, pp. 162–167.

[5]. Thad Starner, Joshua Weaver, and Alex Pentland, "Realtime American sign language recognition using desk and wearable computer-based video," IEEE TPAMI, vol.
20, no. 12, pp. 1371–1375, 1998.

[6]. Christian Vogler and Dimitris Metaxas, "Parallel hidden Markov models for American sign language recognition," in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. IEEE, 1999, vol. 1, pp. 116–122.

[7]. Kouichi Murakami and Hitomi Taguchi, "Gesture recognition using recurrent neural networks," in Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 1991, pp. 237–242.

[8]. Chung-Lin Huang and Wen-Yi Huang, "Signlanguage recognition using model-based tracking and a 3D Hopfield neural network," Machine vision and applications, vol. 10, no.5-6, pp. 292–307, 1998.[9]. Jong-Sung Kim, Won Jang, and Zeungnam Bien, "A dynamic gesture recognition system for the Korean sign language (ksl)," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 26, no. 2, pp. 354–359, 1996.

[10]. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," arXiv preprint arXiv:1311.2524, 2013.

[11]. Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks withmultitask learning," in ICML. ACM,2008, pp. 160–167.

[12]. Clement Farabet, Camille Couprie, Laurent Najman, ´ and Yann LeCun, "Learning hierarchical features for scene labelling," IEEE TPAMI, vol. 35, no. 8, pp. 1915– 1929, 2013.

[13]. Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Homesteader, Kevin Briggman, Winfried Denk, and H Sebastian Seung, "Convolutional networks can learn to generate affinity graphs for image segmentation," Neural Computation, vol. 22, no. 2, pp. 511– 538, 2010.

[14]. Ao Tang, Ke Lu, Yufei Wang, Jie Huang, and Houqiang Li, "A real-time hand posture recognition system using deep neural networks," ACM Transactions on Intelligent Systems and Technology, 2014.

[15]. Moez Baccouche, Franck Mama let, Christian Wolf, Christophe Garcia, and Atilla Bozkurt, "Sequential deep learningfor human action recognition," in Human Behaviour Understanding, pp. 29–39. Springer, 2011.