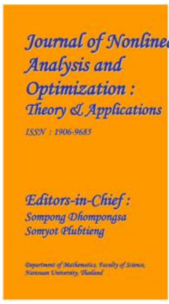
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 12, No. 2, (2021),
ISSN : **1906-9685**



HAND GESTURE RECOGNITION AND VOICE COMMUNICATION FOR DEAF AND DUMB

1Rohan Kokate,2Chetan Padole,3Umesh Samarth,4 Rajat Khawaskar 123Assistant Professor,4Student Department Of Information Technology J D College of Engineering & Management, Nagpur

ABSTRACT

Sign Language Recognition (SLR) targets on interpreting the sign language into text or speech, so as to facilitate the communication between deaf-mute people and ordinary people. This task has broad social impact, but is still very challenging due to the complexity and large variations in hand actions. Existing methods for SLR use hand-crafted features to describe sign language motion and build classification models based on those features. However, it is difficult to design reliable features to adapt to the large variations of hand gestures. To approach this problem, we propose a novel convolutional neural network (CNN) which extracts discriminative spatial-temporal features from raw video stream automatically without any prior knowledge, avoiding designing features. To boost the performance, multi-channels of video streams, including color information, depth clue, and body joint positions, are used as input to the CNN in order to integrate color, depth and trajectory information. We validate the proposed model on a real dataset collected with Microsoft Kinect and demonstrate its effectiveness over the traditional approaches based on hand-crafted features

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 color images as input to networks like [1, 2], we take color 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 color 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 color 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 color, 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 types of data as input. This architecture integrates color, 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.

SYSTEM ANALYSIS

Existing System

Creating a desktop application that uses a computer's webcam to capture a person signing gestures for American Sign Language (ASL), and translate it into corresponding text and speech in real time. The translated sign language gesture will be acquired in text which is farther converted into audio.

Disadvantage of existing system

1. Less efficiency.

Proposed system

In this manner we are implementing a finger spelling sign language translator. To enable the detection of gestures, we are making use of a Convolutional neural network (CNN). A CNN is highly efficient in tackling computer vision problems and is capable of detecting the desired features with a high degree of accuracy upon sufficient training.

Advantage of Proposed system

1. More efficiency.

MODULES DESCRIPTION:

User:

The User can start the project by running run.py file. User has Upload Hand Gesture Dataset, Train CNN with Gesture Images User has to open cv class VideoCapture(0) means primary camera of the system, VideoCapture(1) means secondary camera of the system. VideoCapture(Videfile path) means with out camera we can load the video file from the disk. Vgg16, Vgg19 has programitaically configured. User can change the model selection in the code and can run in multiple ways.

HSR System:

Video-based Hand Sign recognition can be categorized as vision-based according. The vision based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the hand. Therefore, this methodology is getting more consideration nowadays, consequently making the HSR framework simple and easy to be deployed in many applications. We first extracted the frames for each activities from the videos. Specifically, we use transfer learning to get deep image features and trained machine learning classifiers.
Hand Gesture Recognition In the past decade, the computational power of computers has doubled, while the human computer interface (HCI) has not changed too much. When we work with a computer, we are constrained by intermediary devices (keyboards and mice). However, these are inconvenient and have become a bottleneck in human-computer interaction. In our daily life, we use speech to communicate with each other, and use gestures to point, emphasize and navigate. They are the more natural and preferable means to interact with computers for human beings. To make computers understand this however is not an easy task. Gesture recognition is a topic in computer science and

language technology with the goal of interpreting human gestures via mathematical algorithms. Gesture s can originate from any bodily motion or state but commonly originate from the face or hand. Gesture recognition can be seen as a way for computers to begin to understand human body language, thus building a richer bridge between machines and humans. Gesture recognition enables humans to communicate with the machine and interact naturally without any mechanical devices. Gesture recognition can be conducted with techniques from computer vision and image processing.

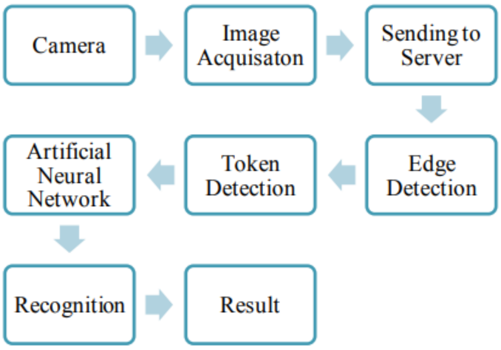


Figure 1: Block diagram of System

Hand and Fingers and Palm Segmentation

The original images used for hand gesture recognition in the work are demonstrated. These images are captured with a normal camera. These hand images are taken under the same condition. The background of these images is identical. So, it is easy and effective to detect the hand region from the original image using the background subtraction method. However, in some cases, there are other moving objects included in the result of background subtraction. The skin color can be used to discriminate the hand region from the other moving objects. The color of the skin is measured with the HSV model. The HSV (hue, saturation, and value) value of the skin color is 315, 94, and 37, respectively. The image of the detected hand is resized to to make the gesture recognition invariant to image scale.

The output of the hand detection is a binary image in which the white pixels are the members of the hand region, while the black pixels belong to the background. Then, the following procedure is implemented on the binary hand image to segment the fingers and palm.

Palm Point. The palm point is defined as the center point of the palm. It is found by the method of distance transform. Distance transform also called distance map is a representation of an image. In the distance transform image, each pixel records the distance of it and the nearest boundary pixel.

Inner Circle of the Maximal Radius. When the palm point is found, it can draw a circle with the palm point as the center point inside the palm. The circle is called the inner circle because it is included inside the palm. The radius of the circle gradually increases until it reaches the edge of the palm.

Wrist Points and Palm Mask. When the radius of the maximal inner circle is acquired, a larger circle the radius of which is 1.2 times of that of the maximal inner circle is produced.

Hand Rotation. When the palm point and wrist point are obtained, it can yield an arrow pointing from the palm point to the middle point of the wrist line at the bottom of the hand. Then, the arrow is adjusted to the direction of the north. The hand image rotates synchronously so as to make the hand gesture invariant to the rotation. Meanwhile, the parts below the wrist line in the rotated image are cut to produce an accurate hand image that does not enclose the part of the arm.

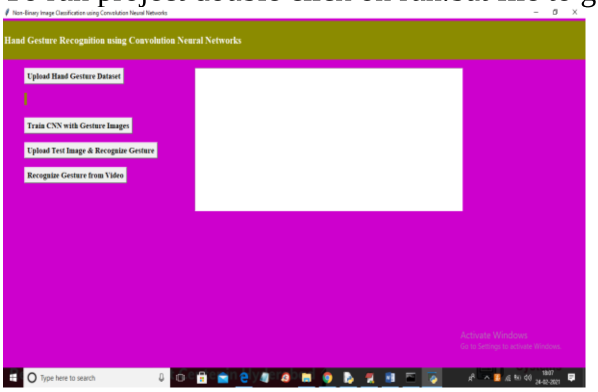
Convolutional Neural Network (CNN)

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.

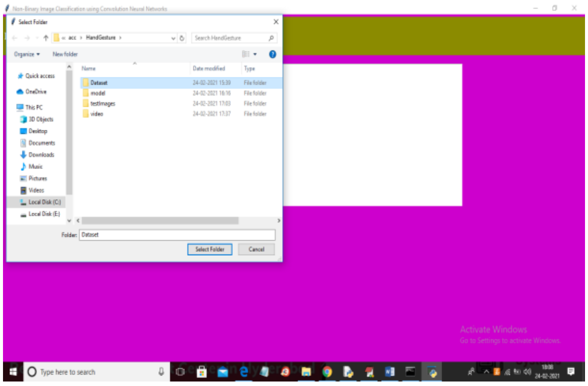
Technically speaking, the main challenge of sign language recognition lies in developing descriptors to express hand-shapes and motion trajectory. In particular, hand-shape description involves tracking hand regions in video stream, segmenting hand-shape images from complex background in each frame and gestures recognition problems. Motion trajectory is also related to tracking of the key points and curve matching. To address these difficulties, we develop a CNNs to naturally integrate hand-shapes, trajectory of action and facial expression. Instead of using commonly used color images as input to networks like [1, 2], we take color 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 color 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 color 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 color, 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 types of data as input. This architecture integrates color, 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.

SCREEN SHOTS

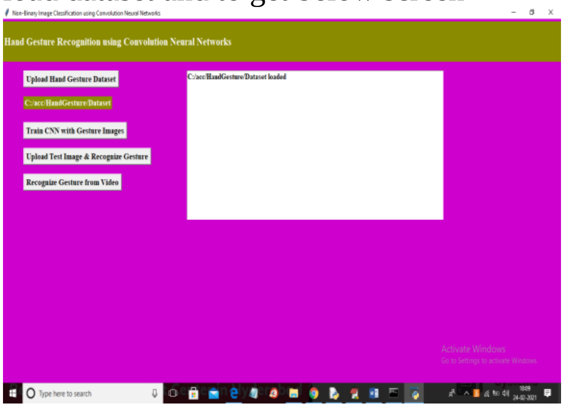
To run project double click on run.bat file to get below screen



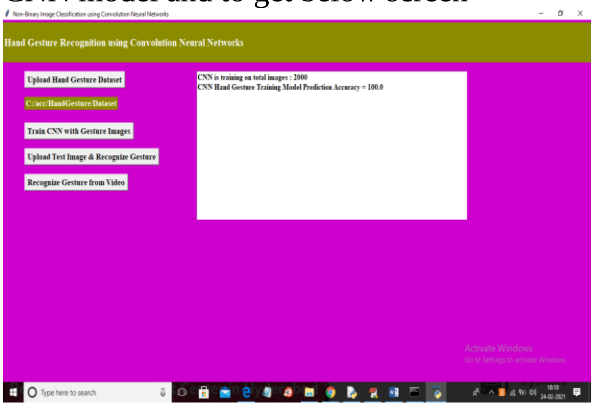
In above screen click on ‘Upload Hand Gesture Dataset’ button to upload dataset and to get below screen



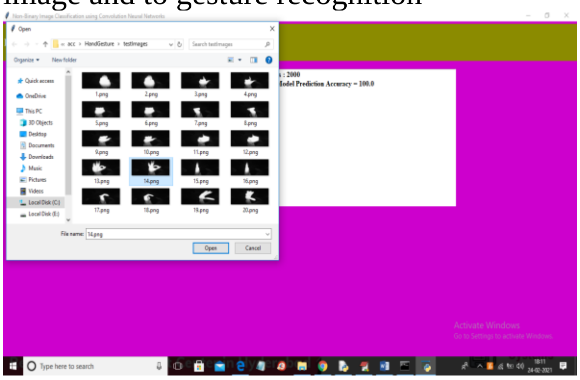
In above screen selecting and uploading ‘Dataset’ folder and then click on ‘Select Folder’ button to load dataset and to get below screen



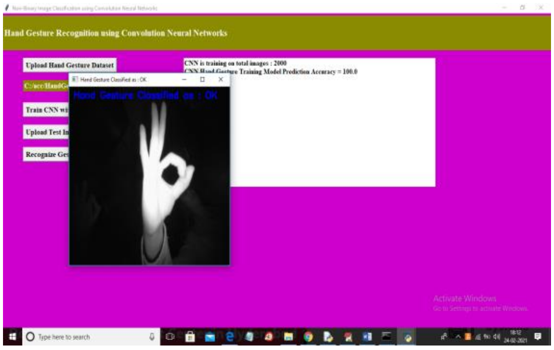
In above screen dataset loaded and now click on ‘Train CNN with Gesture Images’ button to trained CNN model and to get below screen



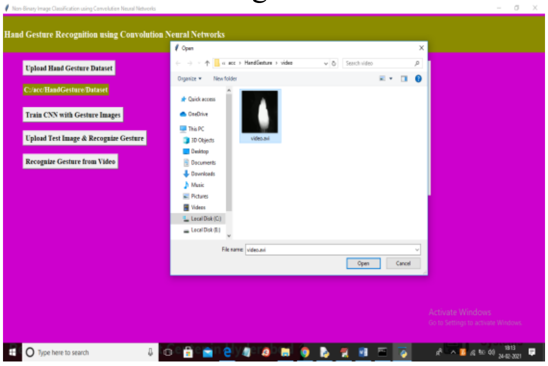
In above screen CNN model trained on 2000 images and its prediction accuracy we got as 100% and now model is ready and now click on ‘Upload Test Image & Recognize Gesture’ button to upload image and to gesture recognition



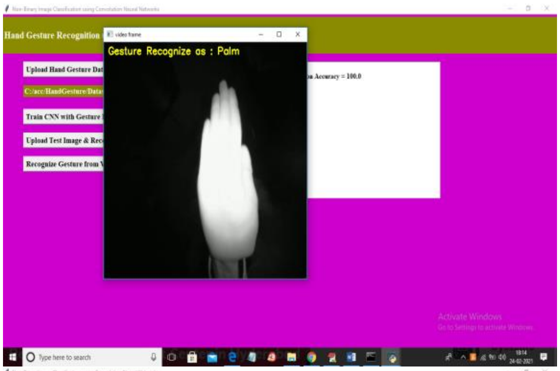
In above screen selecting and uploading ’14.png’ file and then click Open button to get below result

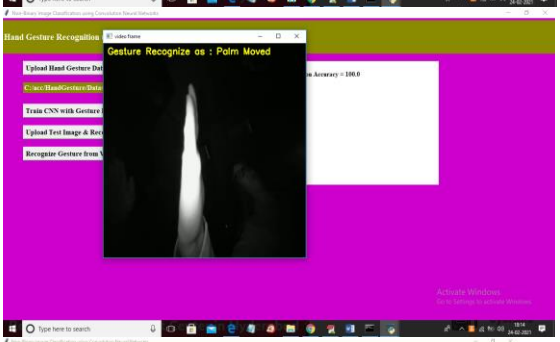


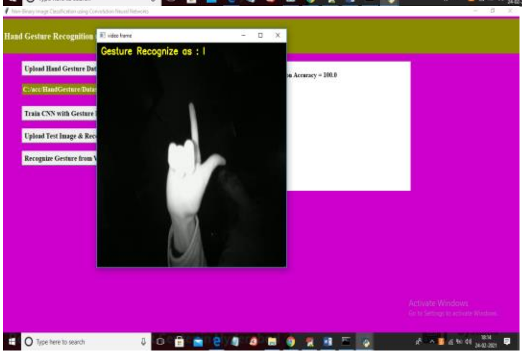
In above screen gesture recognize as OK and similarly you can upload any image and get result and now click on ‘Recognize Gesture from Video’ button to upload video and get result.



In above screen selecting and uploading ‘video.avi’ file and then click on ‘Open’ button to get below result







In above screen as video play then will get recognition result.

CONCLUSION

We developed a CNN model for sign language recognition. Our model learns and extracts both spatial and temporal features by performing 3D convolutions. The developed deep architecture extracts multiple types of information from adjacent input frames and then performs convolution and subsampling separately. The final feature representation combines information from all channels. We use multilayer perceptron classifier to classify these feature representations. For comparison, we evaluate both CNN and GMM-HMM on the same dataset. The experimental results demonstrate the effectiveness of the proposed method.

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.

[4] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Chung-Lin Huang and Wen-Yi Huang, “Sign language recognition using model-based tracking and a 3D hopfield neural network,” Machine vision and applications, vol. 10, no. 5-6, pp. 292–307, 1998.

[11] 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.

[12] 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.

[13] Ronan Collobert and Jason Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in ICML. ACM, 2008, pp. 160–167.

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

[15] Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, 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.