# Hand Pose Recognition Using Parallel Multi Stream CNN

Subjects: Computer Science, Artificial Intelligence | Others

Contributor: Iram Noreen

Recently, several computer applications provided operating mode through pointing fingers, waving hands, and with body movement instead of a mouse, keyboard, audio, or touch input such as sign language recognition, robot control, games, appliances control, and smart surveillance. With the increase of hand-pose-based applications, new challenges in this domain have also emerged. Support vector machines and neural networks have been extensively used in this domain using conventional RGB data, which are not very effective for adequate performance.

hand posture classification deep learning 2D CNN multi stream depth data

## 1. Introduction

Innovative technology has enabled us to communicate with computers and machines through touchless mechanisms in today's fast-paced digital world. This mechanism involves postures such as waving hands, pointing fingers, and moving different body parts instead of touching screens, pressing switches, and raising voices [1]. This new paradigm of posture-based communication mechanics has evolved into a new era of smart applications and challenges as well. This mechanism uses machine learning, pattern recognition, and computer vision technology. Posture recognition has concerns about the accurate identification of meaningful postures. Further, posture recognition has great significance in human-computer interfacing and intelligent systems. Posture recognition has various applications, e.g., smart surveillance, sign language recognition, human-robot interaction, vision-based robotic surgery [2], handicapped medical assistance [3], TV control, gaming, and robot control [4]. The hand poses recognition problem has received the attention of researchers in the field of machine learning and computer vision [5] due to convenience and a wide range of applications. Hand-posture-based interfacing in applications has introduced convenience, flexible products, and efficiency for users to control devices without physical contact [6]. However, being an evolving domain, it confronts many issues and challenges. Traditional hand posture recognition methods began at the beginning of the 1980s [7]. Until recently, most posture recognition approaches were focused mainly on RGB intensity images. The main limitation in such RGB image datasets is that they are highly sensitive to illumination conditions, cluttered backgrounds, camera resolutions, and diverse points of view. This is problematic for segmentation, motion analysis, and detection of points of interest, and can perform well only in limited scenarios.

With the introduction of Kinect in 2010, easily available depth data has improved the domain. Firstly, with recent advances in in-depth camera technology, depth data can be acquired for much less cost. Secondly, depth data

provide better information descriptions for the hand poses. Third, previous approaches are only effective with a limited number of gestures. Most of the work reported in the literature focuses on using conventional machine learning approaches with RGB data and depth data as well.

Recently, deep learning approaches have achieved state-of-art results as compared to conventional handcrafted feature-based techniques; however, they have also not investigated depth data much in comparison to RGB data. In recent years, CNNs have gained popularity and attention. Most of the CNN-based approaches involve 3D data representation of 2D depth images. Depth images are converted into 3D voxels to utilize 3D spatial information followed by an application of a 3D CNN for 3D hand pose classification. The methods using 3D point cloud-based inputs have performed well in capturing the geometric features of depth images; however, they suffer from complex data-conversion phases and also involve heavy parameter management, increasing the time complexity. Three-dimensional CNNs have a complex network structure with a larger number of parameters and high computational costs. In the same way, when a depth image is mapped to a 3D space, it may cause information loss or add unnecessary information (noise), resulting in errors. Therefore, 3D CNN-based methods not only waste convolution process calculations but also move the neural network from learning effective features. Additionally, it is reported in the literature that multiple streams are helpful to improve recognition performance [8]. Further, little work is also reported with two-stream CNN using the optical flow method, but the use of the optical flow method increases the process complexity [9]. Hand posture recognition remains a challenging problem and needs improvement. In this study, we designed a four-stream 2D CNN (2-dimensional convolutional neural network) to resolve the aforementioned issues.

## 2. Hand Pose Recognition Using Parallel Multi-Stream CNN

The research community is active in the domain of hand pose recognition and a number of researchers have addressed this problem. Prominent work in the domain is described in this section. Zhu et al. [1] proposed a method to discuss the problem of HRI (human-robot interaction) and SAIL (smart assisted living) for disabled and elderly people. They used a neural network for posture spotting and HMM for context-based identification of posture. They collected data from the foot and waist joints. Cheng et al. [10] presented a survey on 3D posture estimation and presented a state-of-the-art analysis for 3D hand modeling, hand trajectory, continuous hand postures, and static hand posture estimation.

Plouffee et al. [11] proposed a method for static and dynamic posture recognition. They developed a natural posture UI (user interface) to track real-time hand postures using depth data of the ASL (American Sign Language) data set. They extracted the area of interest of the hand using a segmentation process with the assumption that the user is the closest object or entity in the scene to the camera. They improved scanning time on hand contour. They identified fingerprints using a k-curvature algorithm to recognize a person performing posture using the DTW (dynamic time warping) algorithm. They achieved a recognition rate of 92.5% for fifty-five static and dynamic postures.

Liu et al. [7] proposed a method for 3D posture recognition to accord the skeleton tracking of a person. They collected skeleton data from depth images generated through Kinect. Wu et al. [12] presented a method known as DDNN (deep dynamic neural network) for posture recognition using both depth and RGB data. They used a semi-supervised framework based on HMM for posture segmentation and DBN (deep belief network) to handle skeleton dynamics. Further, they adopted 3D CNN to fuse batches of depth and RGB data. Bao et al. [13] proposed a deep convolutional neural network method to classify postures without segmentation. Their method was able to classify seven types of hand postures in real-time and user-independent ways. Kumar et al. [14] proposed a robust position invariant SLR (sign language recognition) framework to observe occluded sign postures. Nunez et al. [5] used Kinect sensors and HMM to obtain and process skeleton data. They presented a method based on CNN and LSTM (long short-term memory) for posture and human activity recognition. They used 3D sequences of the full-body scan. Saha et al. [15] used an ensemble of tree classifiers with a bagging mechanism for an efficient two-person interaction detection system.

Supancic et al. [16] introduced a novel test set to perform the challenging segmentation of the active hand. Jun et al. [17] presented a vision-based hand posture recognition system to accord high security in an IoT (Internet of Things) application. The system was able to interact with users by recognizing postures in the captured images through a monocular camera installed on a terminal device. The first module of the system used an edge repair-based hand segmentation algorithm and the second module located the user's position using an adaptive method. Avola et al. [18] used RNN (recurrent neural network) to train angles formed through the bones of fingers. Features were acquired through a LMC (leap motion controller) sensor and were selected based on hand joint movement. They achieved 96% accuracy on the ASL dataset and acquired more than 70% accuracy on the SHREC dataset. Tai et al. [19] proposed a many-to-many LSTM-based scheme for hand posture recognition. They applied maximum posteriori estimation on sensory data received by a gyroscope and also implemented a smartphone application to measure the performance. Mirehi et al. [20] presented a method based on fine-tuned inception V3 for static posture recognition from RGB-D data and obtained nearly 90% accuracy using ASL and NTU hand digit datasets. They also presented the SBU-1 dataset [21] that comprises multiple variations and deformation of hand postures.

Sanchez-Riera et al. [22] presented a method for posture recognition of multiple persons using the ICP (iterative closest point) algorithm. It is a minimization function that is initialized through parameters obtained through a dataset. Pinto et al. [23] used a combination of segmentation, polygonal approximation, contour detection, and morphological filters during preprocessing for better feature extraction followed by a CNN. Zhang et al. [24] proposed a robust feature descriptor based on path signature. They proposed the AOH (attention on hand) principle to identify single joint and pair joint features. They achieved an accuracy of 82% on the NVIDIA hand dataset. Hu et al. [25] presented a hand gesture recognition system using CNN with eight layers to control unmanned aerial vehicles (UAV). They achieved an average accuracy of 96%. However, the model is applicable only for non-scaled datasets. Okan et al. [26] also presented a CNN-based hand gesture recognition approach for video data. They applied a sliding window approach and evaluated the performance efficiency on two public datasets, NVIDIA, and EgoGesture. Their main contribution was to improve memory and computing requirements. Jinxian et al. [27] proposed an approach using PCA and generalized regression neural network (GRNN). They obtained classification results of hand gestures from the model and applied them further to extract human emotions

with 95% accuracy. However, their approach was applicable to only nine static human gestures. Chen et al. [28] used CNN to recognize hand gestures through surface electromyography signals (sEMG). However, for classification, they used the traditional machine learning classification model. They trained the model on the Myo dataset and achieved an accuracy of 90%.

Kolivand et al. [29] acquired 96% accuracy on depth data of American Sign Language (ASL) by an artificial neural network (ANN) and support vector machine (SVM) using radial basis function (RBF). Their main contribution was to devise a rotation-invariant procedure of geometric feature extraction from hands received by the depth camera. However, the limitation is that this scheme is manual. Similarly, Kapuściński et al. [30] trained near-neighbor and SVM methods using a static depth dataset of Polish Sign Language (PSL). They proposed a distance-based descriptor to recognize static gesture images. Though their method showed good performance but inherits all the limitations of the manual procedure.

Human activities are categorized into group collaborations, interactions, actions, and gestures according to different complexity levels. Warchoł et al. [31] proposed a bone pair descriptor and distance descriptor for skeletal data. Their method is beneficial as it is light and positions invariant; however, it is designed for ‘action’ recognition and not for hand ‘gesture’ recognition. Garcia et al. [32] presented a real-time segmentation method for hand gesture recognition using RGB frames. They demonstrated its performance on the IPN hand dataset comprising thirteen different gestures for touchless screen interaction. Sarma et al. [9] presented 2D CNN and 3D CNN using an optical-flow-based motion template. The main issue with recent 2D or 3D approaches based on optical flow modalities is that they are computationally expensive and not suitable for real-time application due to increased complexity and computation cost. Though several researchers have addressed the hand posture recognition problem, very little work is reported using depth data in proportion to RGB data. Moreover, much of the work with depth data is focused either on action recognition or applies traditional machine learning methods, which are constrained to manual or handcrafted feature extraction. Manual procedures are undesirable because they are prone to errors, are skill-dependent, and time-consuming, and involve tedious labor.

# 3.Conclusion

Hand posture recognition using RGB intensity images has challenges due to lighting conditions and background clutter. However, depth image data have added value in the domain for better identification of postures than color images. 2D CNN with 4 parallel streams is proposed for posture identification using benchmark Kaggle, Dexter, and First Person datasets. Testing accuracy of the proposed method using the Kaggle dataset is 99.99%, using the Dexter dataset is 99.48%, and using the First Person dataset is 98%. Moreover, other evaluation matrices such as precision, recall, F1 score, AUC score, and root mean square error are also applied to measure performance. Comparison with other state-of-art approaches shows that our model has improved performance and is sufficiently robust. The future work plan is to extend the proposed approach as a multiple stream 3D CNN, and conduct ablation studies between 2D and 3D network behavior. Another future direction is to train the proposed model using a hand gesture dataset of emergency/crisis vocabulary to develop an assistive application. Another future direction could be hand posture estimation, which is the next level after gesture recognition. It requires a skeletal joint

mapping process and identification of mean distance error for joint identification. Light, position, and rotation invariance capacity building is an open challenge to address by the community in this domain. Research in this direction will open new horizons of research and development due to the increased use of smart devices in the near future.

## References


1. Zhu, C.; Sheng, W. Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 2011, 41, 569–573.
2. Su, H.; Ovur, S.E.; Zhou, X.; Qi, W.; Ferrigno, G.; De Momi, E. Depth vision guided hand gesture recognition using electromyographic signals. Adv. Robot. 2020, 34, 985–997.
3. Van Amsterdam, B.; Clarkson, M.J.; Stoyanov, D. Gesture Recognition in Robotic Surgery: A Review. IEEE Trans. Biomed. Eng. 2021, 68, 2021–2035.
4. Danafar, S.; Gheissari, N. Action recognition for surveillance applications using optic flow and SVM. In Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4844, pp. 457–466.
5. Núñez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Vélez, J.F. Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. 2018, 76, 80–94.
6. Zhang, Z.; Tian, Z.; Zhou, M. Latern: Dynamic Continuous Hand Gesture Recognition Using FMCW Radar Sensor. IEEE Sens. J. 2018, 18, 3278–3289.
7. Liu, Y.; Dong, M.; Bi, S.; Gao, D.; Jing, Y.; Li, L. Gesture recognition based on Kinect. In Proceedings of the 6th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2016, Chengdu, China, 19–22 June 2016; pp. 343–347.
8. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.-F. Large-scale video classification with convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
9. Sarma, D.; Kavyasree, V.; Bhuyan, M.K. Two-stream fusion model for dynamic hand gesture recognition using 3d-cnn and 2d-cnn optical flow guided motion template. arXiv 2020, arXiv:2007.08847.
10. Cheng, H.; Yang, L.; Liu, Z. Survey on 3D Hand Gesture Recognition. IEEE Trans. Circuits Syst. Video Technol. 2016, 26, 1659–1673.

11. Plouffe, G.; Cretu, A.M. Static and dynamic hand gesture recognition in depth data using dynamic time warping. IEEE Trans. Instrum. Meas. 2016, 65, 305–316.
12. Wu, D.; Pigou, L.; Kindermans, P.J.; Le, N.D.H.; Shao, L.; Dambre, J.; Odobez, J.M. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 1583–1597.
13. Bao, P.; Maqueda, A.I.; Del-Blanco, C.R.; Garciá, N. Tiny hand gesture recognition without localization via a deep convolutional network. IEEE Trans. Consum. Electron. 2017, 63, 251–257.
14. Kumar, P.; Saini, R.; Roy, P.P.; Dogra, D.P. A position and rotation invariant framework for sign language recognition (SLR) using Kinect. Multimed. Tools Appl. 2018, 77, 8823–8846.
15. Saha, S.; Ganguly, B.; Konar, A. Gesture recognition from two-person interactions using ensemble decision tree. Adv. Intell. Syst. Comput. 2018, 518, 287–293.
16. Supančič, J.S.; Rogez, G.; Yang, Y.; Shotton, J.; Ramanan, D. Depth-Based Hand Pose Estimation: Methods, Data, and Challenges. Int. J. Comput. Vis. 2018, 126, 1180–1198.
17. Xu, J.; Zhang, X.; Zhou, M. A High-Security and Smart Interaction System Based on Hand Gesture Recognition for Internet of Things. Secur. Commun. Netw. 2018, 2018, 11.
18. Avola, D.; Bernardi, M.; Member, S.; Massaroni, C.; Member, S. Exploiting Recurrent Neural Networks and Leap Motion Controller for the Recognition of Sign Language and Semaphoric Hand Gestures. IEEE Trans. Multimed. 2018, 21, 234–245.
19. Tai, T.-M.; Jhang, Y.-J.; Liao, Z.-W.; Teng, K.-C.; Hwang, W.-J. Sensor-Based Continuous Hand Gesture Recognition by Long Short-Term Memory. IEEE Sens. Lett. 2018, 2, 1–4.
20. Mirehi, N.; Tahmasbi, M.; Targhi, A.T. Hand gesture recognition using topological features. Multimed. Tools Appl. 2019, 78, 13361–13386.
21. Marcon, M.; Paracchini, M.B.M.; Tubaro, S. A framework for interpreting, modeling and recognizing human body gestures through 3D eigenpostures. Int. J. Mach. Learn. Cybern. 2019, 10, 1205–1226.
22. Sanchez-Riera, J.; Srinivasan, K.; Hua, K.L.; Cheng, W.H.; Hossain, M.A.; Alhamid, M.F. Robust RGB-D Hand Tracking Using Deep Learning Priors. IEEE Trans. Circuits Syst. Video Technol. 2018, 28, 2289–2301.
23. Pinto, R.F.; Borges, C.D.B.; Almeida, A.M.A.; Paula, I.C. Static Hand Gesture Recognition Based on Convolutional Neural Networks. J. Electr. Comput. Eng. 2019, 2019, 4167890.
24. Li, C.; Zhang, X.; Liao, L.; Jin, L.; Yang, W. Skeleton-Based Gesture Recognition Using Several Fully Connected Layers with Path Signature Features and Temporal Transformer Module. Proc. AAAI Conf. Artif. Intell. 2019, 33, 8585–8593.

25. Hu, B.; Wang, J. Deep learning based hand gesture recognition and UAV flight controls. Int. J. Autom. Comput. 2020, 17, 17–29.
26. Kopuklu, O.; Gunduz, A.; Kose, N.; Rigoll, G. Online Dynamic Hand Gesture Recognition Including Efficiency Analysis. IEEE Trans. Biom. Behav. Identity Sci. 2020, 2, 85–97.
27. Qi, J.; Jiang, G.; Li, G.; Sun, Y.; Tao, B. Surface EMG hand gesture recognition system based on PCA and GRNN. Neural Comput. Appl. 2020, 32, 6343–6351.
28. Chen, L.; Fu, J.; Wu, Y.; Li, H.; Zheng, B. Hand gesture recognition using compact CNN via surface electromyography signals. Sensors 2020, 20, 672.
29. Kolivand, H.; Joudaki, S.; Sunar, M.S. An implementation of sign language alphabet hand posture recognition using geometrical features through artificial neural network. Neural Comput. Appl. 2021, 33, 13885–13907.
30. Kapuściński, T.; Warchoł, D. Hand Posture Recognition Using Skeletal Data and Distance Descriptor. Appl. Sci. 2020, 10, 2132.
31. Warchoł, D.; Kapuściński, T. Human Action Recognition Using Bone Pair Descriptor and Distance Descriptor. Symmetry 2020, 12, 1580.
32. Benitez-Garcia, G.; Prudente-Tixteco, L.; Castro-Madrid, L.C.; Toscano-Medina, R.; Olivares-Mercado, J.; Sanchez-Perez, G.; Villalba, L.J.G. Improving Real-Time Hand Gesture Recognition with Semantic Segmentation. Sensors 2021, 21, 356.