Seminar-II Report
Real time rotation scaling invariant Hasta Mudra classification using deep learning framework

Gautam Sagar, 21CS60R15
Department of Computer Science and Engineering
Indian Institue of Technology Kharagpur
West Bengal, IN
gautamsagarnsit@kgpian.iitkgp.ac.in

Under the supervision of
Dr. Partha Pratim Das
Department of Computer Science and Engineering
Indian Institue of Technology Kharagpur
West Bengal, IN
ppd@cse.iitkgp.ac.in

Abstract

Automatic Hand Gesture Recognition(HGR) has become more important in recent years. It is is a natural way of Human Machine Interaction and has been applied on different areas. The Indian classical dance hasta mudra is a part of gesture or action recognition, but the dance hasta mudras are complex and hard to detect when compared with general action gestures. The complexity arises due to the fact that dance are performed at many different places with different background, illumination. Also the camera recording the dance can be present at any location with respect to the performer and distance from the stage. There are many segmentation algorithms like color clustering, edge detection and thresholding, these methods are fails in extracting features from complex hand gestures and hasta mudras. In recent years various deep learning based algorithms were proposed for general hand gesture recognition which are shown to be fast, illumination invariant, rotation invariant and scaling invariant. In this report several Machine learning and Deep Learning based hand gesture recognition methods are discussed briefly.

1 INTRODUCTION

Gestures reproduce ideas powerfully through non-vocal communication among human beings. It is one of the most effectual and implemented modes of message transmission even while we normally speak in our daily life. The fact that we humans are socials animals, is almost entirely based on our ability to communicate effectively. And effective communication between a people and recording this communication in form of historical archives comprises of both verbal and non-verbal forms. Several Indian classical dance forms such as Bharatnatyam, Kathak, Kathakali,Kuchipudi, Manipuri, Mohiniattam and Odissi use well-choreographed mudras(hand gestures) to communicate the story line. These mudras as well as different body postures and facial expressions have been codified in the famous book 'Natya Shastra'. Gesture or mudra is considered the soul of Indian classical dance. According to the Natyashastra, 'the experts are to use the mudras according to the popular practice and in this matter, they should have an eye on their movement, objects, sphere, quantity, appropriateness and mode'. Bharatanatyam gives more importance to body postures and hand gestures, latter known as hastas or mudras. There are a total of 52 hand gestures, out of which, 28 are single hand gestures (asamyukta hastas), while the remaining 24 are double hand gestures (samyukta hastas). Fig. 1 shows some of the Indian Classical dance mudras.

In recent years, Bailey, H. et. Al [1], [2] was ultimately brought down by his work on e-dance project, it draws an importance of e-science tools, by using practice-led research in dance. In This

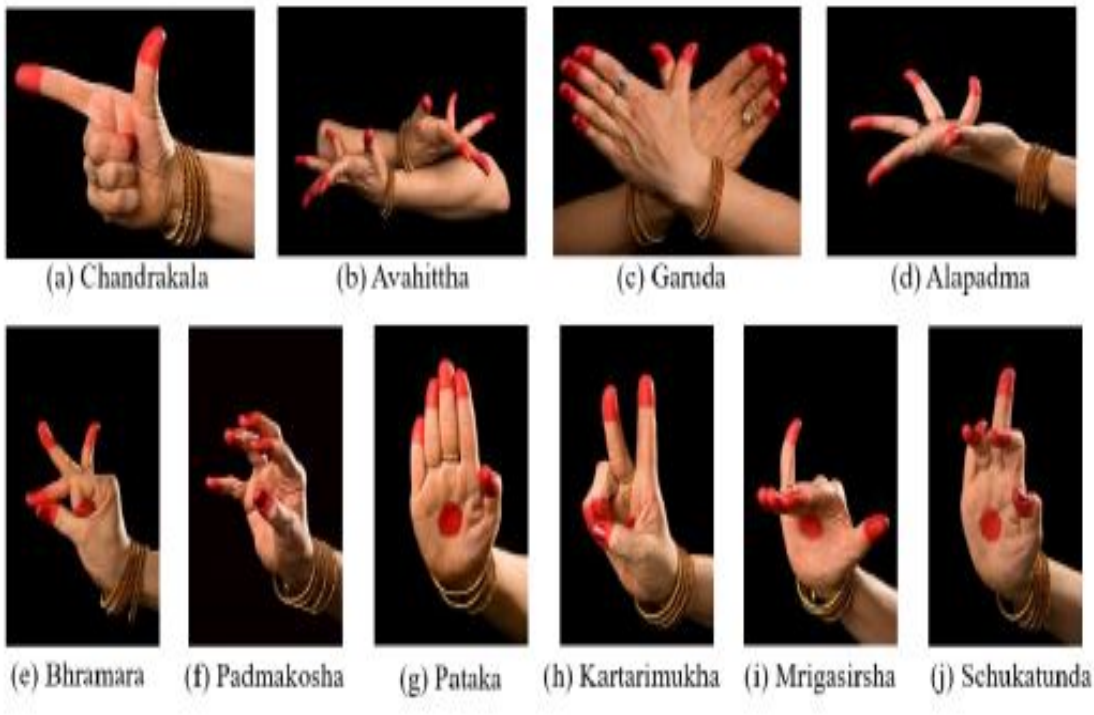

Figure 1: **Hasta mudras of indian classical dances**

project, the researchers explained about new research aspects like choreography, video conferencing, human-computer interaction by using e-dance project. The Indian classical dance hasta mudra is a part of action recognition, but the dance hasta mudras are more complex gestures when we compared with general action gestures. There are many segmentation algorithms like color clustering, edge detection and thresholding, these methods are fails in extracting features from complex hand gestures and hasta mudras [3]–[5]. In the literature, the traditional feature extraction approaches are Speed Up Robust Features (SURF), Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Haar Wavelet Features (HAAR) and Local Binary Patterns (LBP). In last few decades, the feature extraction and classification on 2D images were good, but the classification on complex gestures are more misclassifications due to the limitations of 2D images are capturing angle, lighting condition, different background when capture gestures.

Nowadays, depth sensors, like Intel RealSense and Microsoft Kinect, have become easily accessible, so have the body and hand skeletons [6]–[8]. Accordingly, hand gesture recognition based on pose/skeleton has attracted more and more interests[9]–[14]. In the past few years, hand gesture recognition based on handcrafted features has been widely reported [9], [15]–[17]. Generally, shape of connected joints, hand orientation, surface normal orientation are often used as spatial features, and dynamic time warping or hidden Markov model are then used to process temporal information for classification. Recently, deep learning has also been explored for hand gestures recognition. One approach is to encode joint sequences into texture images and feed into Convolutional Neural Networks (CNNs) in order to extract discriminative features for gesture recognition. Several methods [13], [20]–[22] have been proposed along this approach. However, these methods cannot effectively and efficiently express the dependency between joints, since hand joints are not distributed in a regular grid but in a non-Euclidean domain. To address this problem, graph convolutional networks (GCN) [22]–[24] expressing the dependency among joints with a graph have been proposed. For instance, the method in [23] sets four types of edges to capture relationship between nonadjacent joints. However, the topology of the graph is fixed which is ineffective to deal with varying joint relationships for different hand gestures. A typical example is that, the connection between tip of thumb and tip of forefinger in gesture “Write” is likely to be strong, but it is not the case for gestures

"Prick" and "Tap". Modelling gesture dependent collaboration among joints is especially important for robust hand gesture recognition. Furthermore, conventional ST-GCN has limited receptive field in temporal domain, hence, longterm temporal information cannot be effectively learned.
Another approach is to apply Recurrent Neural Networks (RNNs) to a hand skeleton sequence for classification [12], [18], [19]. However, due to the problem of gradient exploding and vanishing in the temporal direction or gradient decay along layers in training a RNN, shallow RNNs are often adopted. This paper adopts the Independently Recurrent Neural Network (IndRNN) [25], [26] as the basic component to develop a deep residual bidirectional IndRNN (RBi-IndRNN) for effective extraction of long-term temporal features.

2 Methods for Gesture Recognition

Early computer vision techniques involves several challenges, including Variation of illumination conditions: where any change in the lighting condition affects badly on the extracted hand skin region . Rotation problem: this problem arises when the hand region rotated in any direction in the scene. Background problem: refers to the complex background where there is other objects in the scene with the hand objects and these objects might contain skin like color which would produce misclassification problem. Scale problem; this problem appears when the hand poses have different sizes in the gesture image. Finally, Translation problem; the variation of hand positions in different images also leads to erroneous representation of the features. These methods were mostly based on colour recognition and appearance recognition. In traditional Machine learning techniques, most of the applied features need to be identified by an domain expert in order to reduce the complexity of the data and make patterns more visible to learning algorithms to work. Deep Learning has an advantage of learning high-level features from data in an incremental manner. This eliminates the need of domain expertise and hard core feature extraction. In the following sub-sections some machine learning and modern deep learning methods for gesture recognition are discussed.

2.1 Machine learning based Gesture Recognition

Machine learning based methods generally include extraction of traditional computer vision features such as Histogram of oriented Gradients(HOG), Speed Up Robust Features (SURF), Scale Invariant Feature Transform (SIFT), Haar Wavelet Features (HAAR) and Local Binary Patterns (LBP) followed by a robust classifier such as Support Vector Machines.

In [27], The images are first passed through a segmentation stage. Segmentation helps to retrieve accurate features from a image. Segmentation is followed by features extraction(SIFT, SURF, LBP,HOG). These extracted features were sent to the SVM classifier. Performance on all the features is compared with two classifiers i.e. KNN and SVM. HOG feature and SVM classifier is shown to have superior performance than other features with either SVM or KNN. Because number of mudra are more than 2, thus a multiclass SVM is used. For a training set $(p1, q1) \ldots (p_n, q_n)$ with labels q_j in $[1 \ldots z]$, it finds the solution of the following optimization problem during training.

$$min \frac{1}{2} \sum_{i=1}^{z} w_j * wj + \frac{C}{n} \sum_{i=1}^{n} \xi i \tag{1}$$

With similar process to [27], using HOG and SVM, dance mudra are classified in the images obtained from depth sensor[28]. Microsoft's Depth Sensor was used to capture three different types of data like RGB, RGB-D and 3D skeleton data. Only RGB-D data was used for classification. The depth data were used to eliminate 2D image processing limitations and easily remove backgrounds from original image using histogram algorithms and the hands are separated from the rest of the body. Fig. 2 shows the Visualization HOG features of Dance mudras.

The typical flowchart of dance mudra recognition using machine learning is shown in fig.3. Table 1 shows characterization of several features based on the parameters required for good feature descriptor.

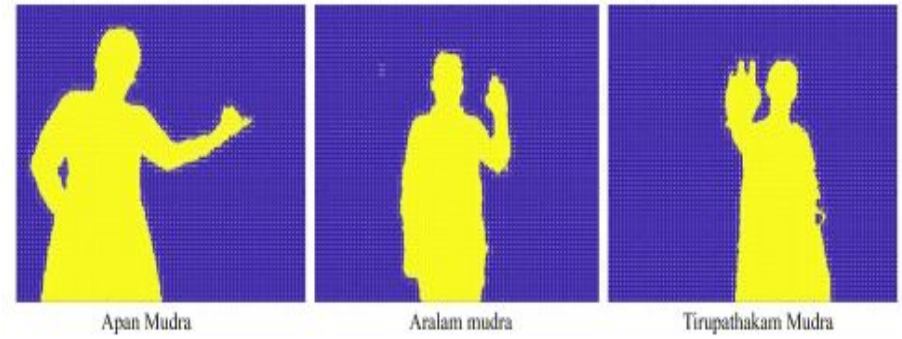

Figure 2: **Visualization HOG features of Dance mudras**

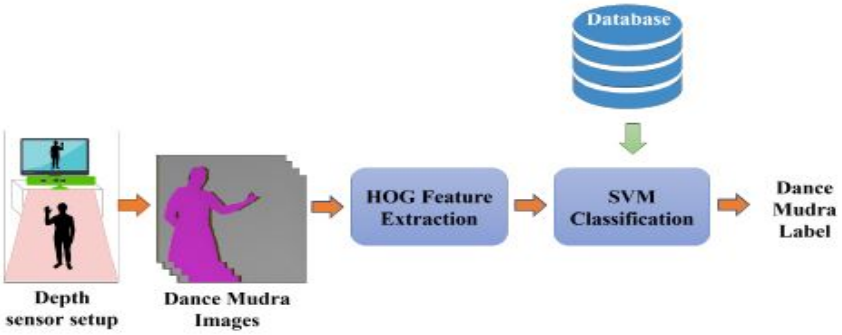

Figure 3: **Overview of a typical dance mudra recognition system**

2.2 Deep Learning based methods

Machine learning based used few handpicked features selected by domain experts. This leads to performance degradation of ML models when tested on data of different domains. Gesture recognition using ML methods suffer of the problem of background and illumination change. This leads the researcher to focus on modern Deep Learning based methods. CNN based architecture for illumination invariant and local scaling invariant facial gesture recognition in a video shows to have good performance[29] but it was tested only on frontal perspectives thereby imposing a constraint on the input facial orientations. Thus this method can not be adopted to hand gesture recognition which requires it to be rotation invariant.

Hand detection can be very helpful in segmentation of hand thereby making the model less susceptible to changes in other parts of data which can be very useful in hand gesture recognition. FOANet Network Architecture[3027 introduces a separate channel for every focus region (global, right hand, left hand) and modality (RGB, depth, and two types of flow fields). The result is 12 channels processing different types of localized data, as shown in Figure 4.

Table 1: **Feature Descriptor Characterization for Dance Mudra Classification**

Methods	Timing	Transformation	Scaling	Rotation	Blurring	Illumination
FAST	Common	Common	Common	Bad	Bad	Bad
SIFT	Bad	Good	Good	Best	Best	Best
SURF	Good	Best	Best	Good	Good	Good
HOG	Good	Best	Best	Best	Best	Best
LBP	Good	Good	Good	Good	Common	Good

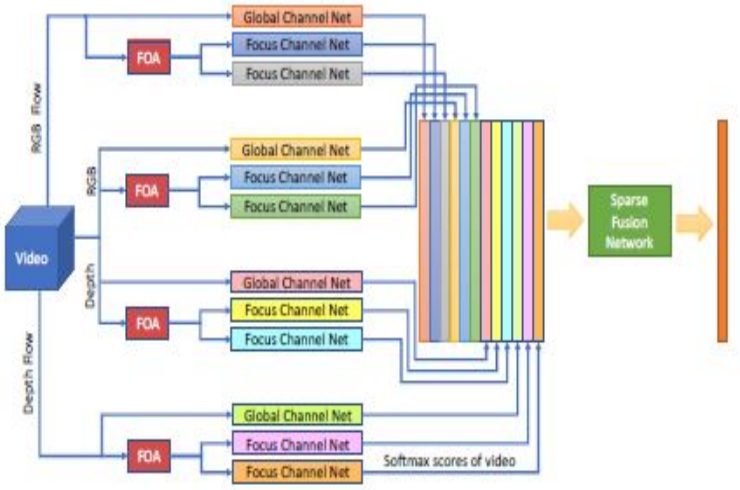

Figure 4: **The FOANet Network Architecture. The architecture consists of a separate channel for every focus region (global, left hand, right hand) and modality (RGB, depth, RGB flow and depth flow). FOA module is used to detect hands. The video level softmax scores from 12 channels are stacked together. Sparse fusion combines softmax scores according to the gesture type.**

Skin segmentation based hand detection[27,31] was shown to perform better than previous method in cluttered environment. Popular object detection method such as RCNN and Faster RCNN was used[31] for obtaining the initial estimate of the spatial location of hands in the input image.To reduce false positives in hand detection, a patch-based convolutional neural network skin detector is used which is robust to background clutter and detects skin pixels accurately. A more robust and complex method for hand detection used multi-scale deep learning algorithm to detect hands in unconstrained scenarios as well as frames from driving videos[27]. The pipeline of this multi-scale faster R-CNN based hand detection algorithm is shown in fig. 5. A set of 4 shallow Faster-RCNN architectures tackles different scales which is a prominent problem in hand detection literature. Each specialized Faster-RCNN architecture is responsible detection for a specific scale range. Note that a deeper Faster-RCNN architecture due to its larger stride and multiple down-sampling operations often does not generate the necessary features which are essential for small object detection. For example, if we use VGG-16 as a backbone network, then output of the last feature layer would be $7 * 7$ due to multiple down-sampling operations which would be insufficient for detection of small objects. The first shallow Faster-RCNN network is built over a backbone network that consists of two convolution layers. The second shallow network's backbone network consists of three convolutional layers. Similarly, the backbone networks of third and fourth Faster-RCNN consist of four and five convolutional layers, respectively. To remove faces, all the Faster-RCNN networks are trained to detect faces and hands simultaneously.Given an image, this algorithm first detects hands as well as faces. The detected faces are then removed from the results to reduce false positives. Because many hand gesture methods were tested only on the frontal perspective of hand images, hand detection can be useful to enhance the performance of these hand gesture recognition methods.

Pose based hand gesture recognition is another widely studied method in the recent years. Compared with full body action recognition, hand gesture involves joints that are more spatially closely distributed with stronger collaboration. This nature requires a different approach from action recognition to capturing the complex spatial features. One recently introduced pose based hand gesture recognition method is a two-stream neural network with one stream being a self-attention based graph convolutional network (SAGCN) extracting the short-term temporal information and hierarchical spatial information, and the other being a residual-connection enhanced bidirectional Independently Recurrent Neural Network (RBi-IndRNN) for extracting long-term temporal information[28]. The self-attention based graph convolutional network has a dynamic self-attention mechanism to adaptively exploit the relationships of all hand joints in addition to the fixed topology and local feature

extraction in the GCN. On the other hand, the residual-connection enhanced Bi-IndRNN extends an IndRNN with the capability of bidirectional processing for temporal modelling. The two streams are fused together for recognition. Fig. 6 shows the framework of the two-stream network.

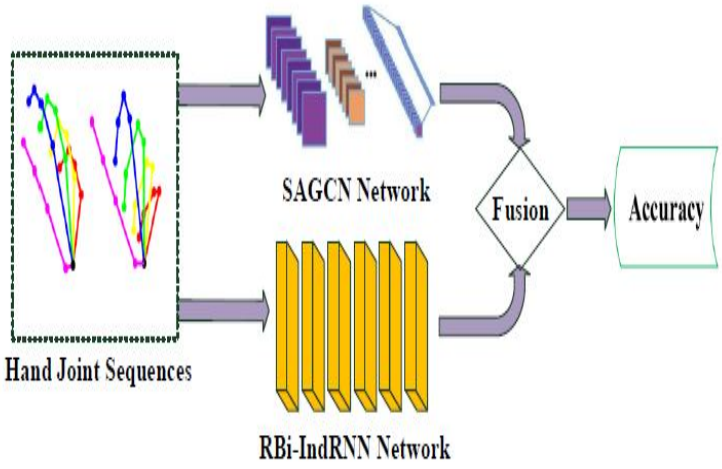

Figure 5: **Framework of the proposed two-stream network. SAGCN module focuses on hierarchical spatial information and short-term temporal information, and RBi-IndRNN focuses on long-term temporal information.**

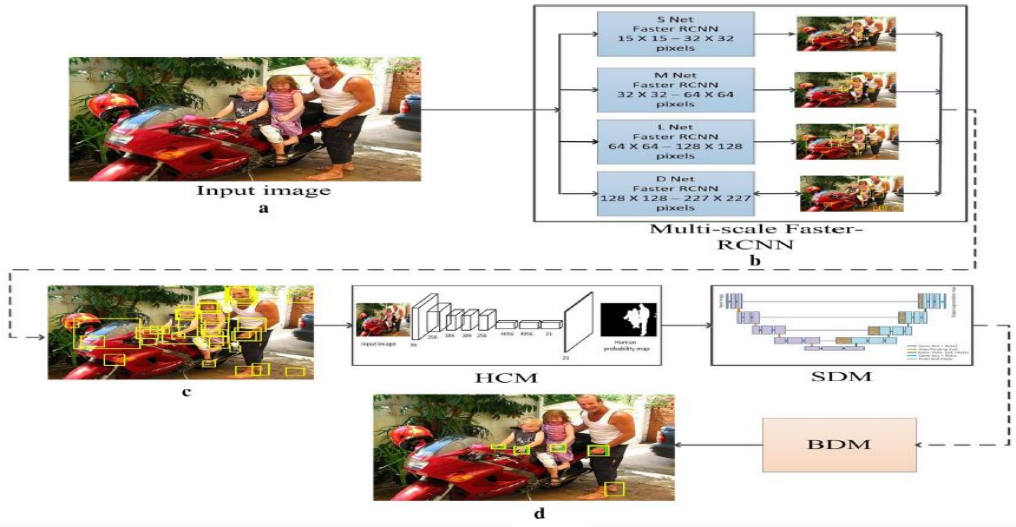

Figure 6: **The pipeline of the proposed multi-scale Faster-RCNN hand detection algorithm. (a) Input image. (b) Multi-scale Faster-RCNN stage consisting of four Faster-RCNNs with different architectures. (c) Detected bounding boxes obtained from each Faster-RCNN overlaid on top of input image. (d) Final hand detection results after post-processing with HCM, human context module; SDM, skin detection module; BDM, blob detection module.**

3 Conclusion

Traditional computer vision algorithms for hand gesture recognition had many limitations such as illumination invariance or their usability was limited to certain background and images from specific camera angle. Machine learning based methods tried to tackle these problems with robust

classifiers like SVM. But they needed domain experts for selecting and hand crafting the features which were then extracted from the image. With the use of CNN based methods several Deep Learning architectures were able to solve the problem several types of invariant and increased the scope of their usability in different environments. Many deep learning architectures were proposed for hand gesture recognition, but very few focused on mudra detection in Indian class dance such as Bharatnatyam. As these dances are performed in various different environments and recorded from several camera angle, the need arises for real time hasta mudra recognition methods which are also rotation and scaling invariant unaffected by illumination conditions.

References

[1] Alexandra Okada, S Buckingham Shum, & Tony Sherborne. Knowledge cartography. Software Tools and Mapping Techniques, 2008.

[2] Helen Bailey, Michelle Bachler, Simon Buckingham Shum, Anja Le Blanc, Sita Popat, Andrew Rowley, & Martin Turner. Dancing on the grid: using e-science tools to extend choreographic research. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1898):2793–2806, 2009.

[3] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43(1):177–214, 2015.

[4] Larry S Davis. A survey of edge detection techniques. Computer graphics and image processing, 4(3):248–270, 1975.

[5] Mehmet Celenk. A color clustering technique for image segmentation. Computer Vision, Graphics, and image processing, 52(2):145–170, 1990.

[6] C. Li, B. Zhang, C. Chen, Q. Ye, J. Han, G. Guo, and R. Ji, "Deep manifold structure transfer for action recognition," IEEE Transactions on Image Processing, vol. 28, pp. 4646–4658, 2019.

[7] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu, and Y. Zhang, "Skeletonbased action recognition with gated convolutional neural networks," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, pp. 3247–3257, 2019.

[8] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatiotemporal attention model for human action recognition from skeleton data," in AAAI, 2017.

[9] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 1–9.

[10] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. Le Saux, and D. Filliat, "Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset," in Eurographics Workshop on 3D Object Retrieval, 2017.

[11] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog for action recognition," in IEEE conference on computer vision and pattern recognition workshops, 2013, pp. 465–470.

[12] X. Chen, H. Guo, G. Wang, and L. Zhang, "Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition," in IEEE International Conference on Image Processing, 2017, pp. 2881–2885.

[13] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data," in IEEE International Conference on Automatic Face Gesture Recognition, 2018, pp. 106–113.

[14] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," IEEE transactions on cybernetics, vol. 45, no. 7, pp. 1340–1352, 2015.

[15] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera,ACM international conference on Multimedia, 2011, pp. 1093–1096.

[16] H. Wang, Q. Wang, and X. Chen, "Hand posture recognition from disparity cost map," in Asian Conference on Computer Vision, 2012.

[17] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in IEEE International Conference on Image Processing, 2014, pp. 1565–1569.

[18] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, "Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3120–3128.

[19] S. Shin and W. Kim, “Skeleton-based dynamic hand gesture recognition using a part-based gru-rnn for gesture-based interface,” IEEE Access, vol. 8, pp. 50 236–50 243, 2020.

[20] P. Narayana, J. R. Beveridge, and B. A. Draper, “Gesture recognition: Focus on the hands,” in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5235–5244.

[21] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, “Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4207–4215.

[22] X. S. Nguyen, L. Brun, O. Lezoray, and S. Bougleux, “A neural network based on spd manifold learning for skeleton-based hand gesture recognition,” in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 036–12 045.

[23] Y. Li, Z. He, X. Ye, Z. He, and K. Han, “Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition,” Eurasip Journal on Image and Video Processing, vol. 2019, no. 1, pp. 1–7, 2019.

[24] G. Hu, B. Cui, Y. He, and S. Yu, “Progressive relation learning for group activity recognition,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 977–986, 2020.

[25] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, “Independently recurrent neural network (indrnn): Building a longer and deeper rnn,” in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5457–5466.

[26] S. Li, W. Li, C. Cook, and Y. Gao, “Deep independently recurrent neural network (indrnn),” ArXiv, vol. abs/1910.06251, 2019.

[27] Roy, K., Sahay, R.R. (2021). A robust multi-scale deep learning approach for unconstrained hand detection aided by skin segmentation. The Visual Computer, 1-25.

[28] Chuankun Li, Shuai Li, Yanbo Gao, Xiang Zhang, Wanqing Li: A Two-stream Neural Network for Pose-based Hand Gesture Recognition. CoRR abs/2101.08926 (2021)

[29] Otkrist Gupta, Dan Raviv, and Ramesh Raskar. 2018. Illumination invariants in deep video expression recognition. Pattern Recogn. 76, C (April 2018), 25–35. DOI:https://doi.org/10.1016/j.patcog.2017.10.017

[30] P. Narayana, J. R. Beveridge and B. A. Draper, "Gesture Recognition: Focus on the Hands," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5235-5244, doi: 10.1109/CVPR.2018.00549.

[31] K. Roy, A. Mohanty and R. R. Sahay, "Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation," 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 640-649, doi: 10.1109/ICCVW.2017.81.