3D Hand Gesture Representation and Recognition through Deep Joint Distance Measurements

P. Vasavi[1], Suman Maloji[2], E. Kiran Kumar[3], D. Anil Kumar[4], N. Sasikala[5]
Department of ECM, Koneru Lakshmaiah Education Foundation, Guntur (DT), Andhra Pradesh, INDIA[1]
Department of ECE, Koneru Lakshmaiah Education Foundation, Guntur (DT), Andhra Pradesh, INDIA[2,3,5]
Department of ECE, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, INDIA[4]

***Abstract*—Hand gestures with finger relationships are among the toughest features to extract for machine recognition. In this paper, this particular research challenge is addressed with 3D hand joint features extracted from distance measurements which are then colour mapped as spatio temporal features. Further patterns are learned using an 8-layer convolutional neural network (CNN) to estimate the hand gesture. The results showed a higher degree of recognition accuracy when compared to similar 3D hand gesture methods. The recognition accuracy for our dataset KL_3DHG with 220 classes was around 94.32%. Robustness of the proposed method was validated with only available benchmark 3D skeletal hand gesture dataset DGH 14/28.**

***Keywords*—Gesture recognition; 3D motion capture; deep learning; joint relational distance maps**

I. Introduction

Hand gestures were considered to be one of the most powerful form of communication known to humans. It has evolved with the progression of generations which has now been regarded as the formidable communication between humans and machines. Hand gestures have now become a part of natural language processing in the current scenario. Hence, hand gestures have become an increasingly important part of human computer interaction (HCI) [1].

There are only three sensors that are exclusively available for capturing 3D hand and fingers. They are Kinect [2], leap motion [3] and Time of Flight (ToF) [4] sensors. Kinect 2 has the capabilities to capture fingers abstractly though noticeably imperfect at times. Leap motion is a good choice for hand capture but the factors for quality depends on the precision movements on the sensor, which at times attracts failures. The ToF sensor reconstructs 3D images from time series data captured by the sensors which however are quite complex to effectively predict hand gestures. Apart from the above, the most popular currently are based on 3D depth sensing technologies [5].

The depth-based hand gestures used 3D modelling for finger relationships for recognition [6]. Moreover, to 3D hand gesture recognition has been the most sought after for its challenging nature. Recent studies point towards static, trajectory and continuous 3D hand gesture recognition for many applications such as human robot interaction, daily assistance, gaming and sign language recognition [7].

In contrast to the above sensors for 3D hand capture, we propose a 3D motion capture technology-based hand gesture recognition. In this work, we used an 8-camera motion capture technology to extract hand gestures for representation of Indian sign language. Here 3D hand gestures are Modelled as a time series 3D joints on the hands. Two hands are used in cohesion as against the existing separation techniques.

The 3D hand joint across frames is Modelled as a time series position vectors that change over frames. This data from all 3D joints is converted into a spatio temporal image representing the varying hand gestures. Hence, the 3D hand gesture recognition problem translates into a spatio temporal RGB image recognition problem. This RGB image recognition problem is handled efficiently using deep networks. An 8-layer CNN is built for this purpose which is based on VGG-16 architecture. However, these networks showed resistance to inter hand variations which resulted in non-discriminatory features at the end of the network. In this work, we propose a multi layered CNN network that preserves the long-term spatial relationships among actions thus generating discriminatory features that facilitate better performance.

To test the proposed multi layered CNN architecture, we intend to use our own 3D hand gesture dataset (KL_3DHG) in skeletal form along with only available skeletal DGH 14/18 [8]. The rest of the paper is organized as follows. Section 2 describes the literature review related to the proposed framework. Section 3 gives the methodology of the proposed framework that has been followed for 3D hand gesture recognition. It is then followed by results and discussion in Section 4. Finally, Section 5 concludes the work.

II. Literature Review

Hand gestures are an important part of human communication. It's classified as a natural language processing tool when comes to interactions between humans and machines. Numerous studies have been successfully conducted in the last few decades to develop a framework for hand recognition using multiple sensors for data capturing with subsequent experimentation to improve recognition performances. This section describes the methods and their findings with gaps towards development of a 3D hand gesture recognition system.

Hand gesture recognition has been attempted visually through video data captured using 2D sensors. However, the operations on this 2D video data has been a series of steps such as pre-processing, segmentation, feature extraction and finally classification [7], [9]. Consequently, the methods used have generated interest mildly, but could not create an impact on the applications related to 3D hand gestures. The underlying reason for poor performance lies in the input sensors ability to capture real time hand gestures effectively [10].

Consequently, sensors such as Kinect and ToF were instrumental in capturing 3D human hand gesture recognition to a new dimension involving depth and skeletal data [11]. The 3D hand gestures recognition problem has been approached in two ways: 1) Static hand poses and 2) Dynamic poses. The static 3D hand shapes are represented as original 3D depth data or using some transform domain data. The 3D hand features are projected as a pixel wise depth features in different hand positions accounting for a large feature space with computational complexities in [12]. In contrast, ensemble of shape function has been proposed to represent 3D shapes as a point cloud which greatly reduced feature space [13]. Apart from spatial domain, the transform domain used Haar [14], Gabor [15], invariant moments [16] as features to model intensity and orientation of 3D hand shapes.

More efficient methods were proposed for representing 3D hand gestures using histogram of 3D Facets as features that modelled surfaces on 3D point clouds [17]. However, the most successful features were SIFT [18], SURF [19] and BOW [20] which achieved highest classification accuracies on a large contingent of classifiers. Moreover, the hybrid features such as bag of words (BoW) has improved the performance of the 3D hand gesture recognition methods effectively. Apart from BoW, other hybrid methods that have shown promising improvement in the recognition accuracies are feature fusion [21] and sensor fusion methods [22]. After feature extraction, an efficient classifier is necessary for producing highly accurate 3D hand gesture recognition. The most widely employed classifiers for 3D static hand gesture recognition are, support vector machines (SVM), artificial neural networks (ANN), random forests (RF) and template matching (TM) [5].

However, dynamic hand gestures were a set of time varying hand representations which need trajectories and orientations for efficient recognition. Two most exclusively used methods for dynamic hand recognition are hidden Markova models (HMM) [23] and dynamic time warping [24]. Besides the above models for continuous 3D hand gesture recognition, condition random fields (CRF) [25] and windowed DTW [26] has proved to achieve higher accuracies.

In the last couple of years, the hand gesture recognition has shifted gears to accommodate real time application capabilities using deep learning models. The most widely employed deep learning model being convolutional neural network (CNN) [27] for 3D human action recognition. Deep learning has been popular on 2D hand gesture video data with 3D CNNs at the learning core to estimate gestures [28]. These are two stream models that are quite popular than the single stream methods. Depth and skeletal data were being exploited simultaneously for recognition with multi stream CNNs [8]. The SoftMax scores from skeletal and depth stream are fused together to generate a class score. However, the most challenging dataset for 3D hand recognition has been the skeletal data. This is due to joint occlusions and overlapping that are hard to analyse on the CNN [29], [30]. Moreover, these methods directly operate on the raw positional vectors as inputs to the CNNs. The results point to a poor recognition accuracy due to inconsistences in the data during the signing process with joint many possible joint interactions.

Apart from CNNs, other deep learning methods used for 3D skeletal hand recognition are memory based deep learning architectures called recurring neural networks (RNNs) and its derived models such as Long Short-Term Memory (LSTMs). The most accurate are a mixture of both spatial and temporal feature learning models that used CNNs for spatial features and RNNs or LSTMs for temporal features. The Recurrent CNN (R-CNN) [31] used 3D convolutional neural networks to extract spatial features which are learned in time by RNNs to generate a complete spatio temporal learning. However, RNNs are slow and could not handle long sequence of data streams making them sluggish for real time operation. These shortcomings were handled efficiently by using long short term memory networks (LSTMs) and there are a multitude of CNN – LSTM [32], [33], [34] combinations with different network architectures that have shown their might in learning spatio temporal features in 3D hand gesture recognition. The sad part is that these hybrid recurrent CNNs are not end – to – end trainable, which limits their capacity for real time modelling. The solution is to develop a complete spatio temporal features which represent spatial and time series variations in 3D hand gestures.

This is however is managed effectively by extracting features on the raw time series positional data as motion maps [35]. The problems in raw 3D joint data has been effectively regulated by transforming the joint time series positional data into spatio temporal feature maps such as joint distance maps (JDMs) [36], joint angular displacement maps (JADMs) [37], joint velocity maps (JVM) [38], joint quad maps (JQM) [39] and joint trajectory maps (JTM) [40]. There are joint surface maps and joint acceleration maps [36] proposed on skeletal data. All the coded maps represent spatio temporal information in the joints with a colour coded image maps which can be effectively learned by a deep convolutional neural network. The key objectives of this work are

1) To generate a 3D hand skeletal dataset with 36 joints on both hands using 3D motion capture technology, which is first of its kind dataset with highest number of joint representations.
2) To extract features from the 3D skeletal hand gesture data for characterizing then using a maximally discriminant spatio temporal colour coded feature maps.
3) To design and train an end – to – end deep learning model to learn the 3D gesture characterizations from spatio temporal maps to accurately recognize gestures of Indian sign language.

The proposed work is different from the existing 3D hand recognition models in three aspects:

1) Most joints on the hands till now for modelling accurately the real time 3D hand motions.
2) A colour coded feature map to characterize the spatio temporal variations in the 3D hand skeletal data, which have not been explore fully for hand gesture recognition.
3) A fast training CNN architecture which can estimate gestures accurately on the proposed features.

The following section describes in detail the methodology for 3D hand gesture recognition framework with datasets, maps creation and CNN operation.

III. Proposed Methodology

The section presents a detailed description of the methods used in 3D hand gesture recognition with deep CNNs. The 3D data describes the hand to hand communication in Indian sign language. The data is captured using 3D motion capture system with 8 cameras. The captured 3D data is a time series representations of hand joints as shown in Fig. 1. Consequently, joint distance features of hands are computed which are then transformed into spatio temporal RGB images. Finally, a deep CNN is inputted with these images to estimate a class label pertaining to the sign. This section contains information regarding 3D hand gesture datasets, joint distance measurements, colour coding joint distance to features, CNN training and testing procedures.

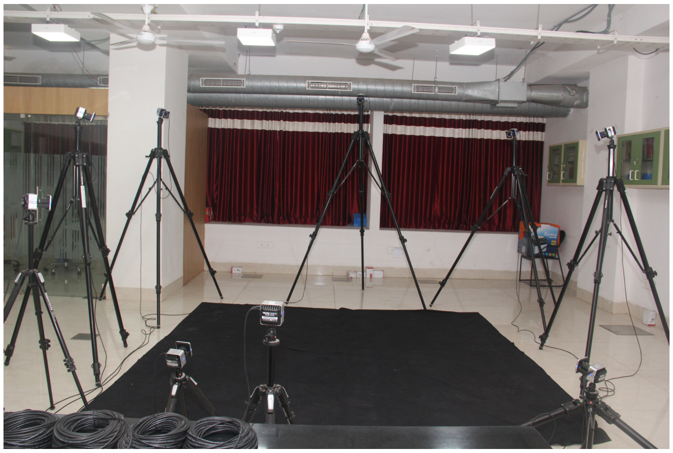


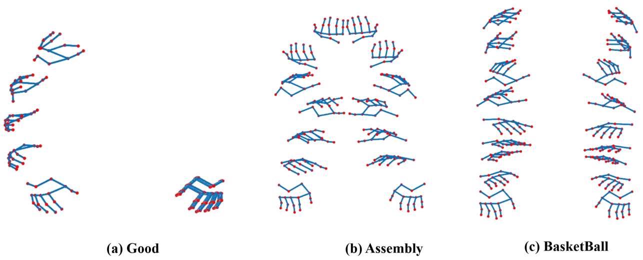
Fig. 1. 3D motion capture system for hand gesture capture

A. 3D Hand Gesture Datasets

The 3D hand gesture skeletal data for sign language is the most complex dataset and hence a challenging task to learn features for recognition. Since Indian sign language is a two-hand system, both hands are used in this work to generate data. Each hand is marked with 18 joints, taking the total number of joints in both hands to 36. This is currently the highest number of joint representations for 3D hand gesture recognition in sign language application. The recorded 3D data gives positional information of each of the finger joints individually with in a video frame. For a particular sign these 3D hand joints are variable across frames in a video sequence.

The time series 3D positional values of hand joints represent a spatio temporal information of a particular class of signs. To construct an entire dataset for training and testing the proposed CNN, we capture 220 sign classes with 10 subjects in 4 views. Fig. 2 shows 3D hand gesture of Indian sign language. Each 3D video frame is recorded for 280 frames, which is considered as a hyper parameter for optimal capture of all signs. A total of $220 \times 280 \times 10 \times 4 \times 3 = 73,92,000$ 2D tensors or 24,64,000 3D video frames are available for processing by the proposed CNN.

Apart from our 3D hand gesture datasets (KL_3DHG), we test the proposed network on benchmark skeletal dataset captured using Intel's real sense technology is DGH 14/28 [8]. This is the only skeletal dataset that is available for hand gesture recognition. It consists of 22 joints in a single hand pose to record 3D skeletal data. The system has a resolution of 640×480 and captures hand poses at 30fps. Each 3D skeletal video in the dataset has 20 to 50 frames per gesture. There are around 2800 samples with 14 or 28 class labels in the DGH 14/28 hand gesture dataset. Comparatively, our KL_3DHG is quite advanced than the DGH dataset with highest sign gestures with full HD resolution with a recording frame rate of 120fps. Our dataset has a greater number of frames per class than the DGH 14/28 dataset. Next section presents the feature calculation and colour coded map generation.



(a) Good (b) Assembly (c) BasketBall

Fig. 2. 3D Hand gestures for Indian sign language

B. JRDM Feature Calculations

Inspired from the methods in [27], [35], [36], [37], [38], [39], [40], we propose to calculate joint relational distance maps (JRDM) between the two hands separately and combine them into a single mapping entity. Here, we calculate joint distances of each hand separately in each frame and further calculate the distance between the two hands from the distances of individual distances of corresponding joint pairs. This JRDM is calculated between joint distance of paired joints on individual hands.

The location p_i of the joint J can be represented in 3D space using 3D coordinates as $p_i\,(x_i, y_i, z_i)\,\forall i = 1$ to $J \in R^{3\times J}$. We then have the combined position vector for the full set of J joints on a N-frame hand sign can therefore be expressed as $S_h = \{p_1, p_2, ..., p_N\}\,\forall R^{J\times 3\times N}$, where h is the hand pointer which takes two variables such as l for left and r for right hand. The intra frame hand distances between i^{th} and j^{th} joint is

$$d_{ij\_h}^{n} = \left\| P_{ih}^{n} - P_{jh}^{n} \right\|_2 \tag{1}$$

For left hand pair (i, j), the distance becomes $d_{ij\_l}^{n}$ and it is $d_{ij\_r}^{n}$ for right hand in the n^{th} video frame, respectively. The two-hand joint relative distance (JRD) that gives the relationship between hands is formulated as

$$D_{ij}^{n} = \left\| d_{ij\_l}^{n} - d_{ij\_r}^{n} \right\|_2 \tag{2}$$

Where D_{ij}^{n} characterizes the hand relationships between joint pairs in an entire video sequence. However, if only one hand is present during a signing process, only intra hand distances are used as feature vector. The final feature matrix for an entire 3D hand sign sequence of N frames is given as

$$D_{ij}^{N} = \left[D_{ij}^{1}, D_{ij}^{2}, ..., D_{ij}^{N} \right] \forall D_{ij\_l}^{N}, D_{ij\_r}^{N} \tag{3}$$

else

$$D_{ij}^{N} = \left[D_{ij\_l}^{1}, D_{ij\_l}^{2}, ..., D_{ij\_l}^{N}\right] \forall h = l \qquad (4)$$

Or

$$D_{ij}^{N} = \left[D_{ij\_r}^{1}, D_{ij\_r}^{2}, ..., D_{ij\_r}^{N}\right] \forall h = r \qquad (5)$$

The JRD matrix captures three types of motion details, namely the intra hand joint distances, inter hand relational joint distances and the time. Finally, the JRD matrix is transformed into a JRDM mapped entity that represents 3D hand movements in Indian sign language.

In contrast to previous studies [27], we simply encode the JRD matrix into an image, using a standard mapping procedure [36] with the "Jet" colour map. Combining the three RGB colour planes into one produces a JRD image which consists of intensity values only. Previous methods have encoded distance maps into colour images [27], but these are affected by the subject's dimensions, leading to an increased number of misclassifications. We used the inter hand relationships between hand joints in the present study to account for the differences in their unrelated features, thus making our approach resistant to subject to subject dimensionality differences. Fig. 3 shows how the JRDM is encoded for a 3D hand sign video.

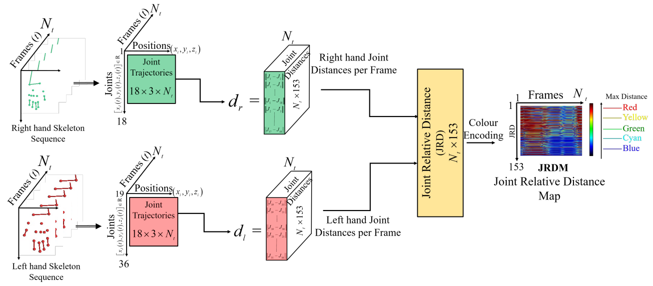


Fig. 3. JRDM color encoding process

C. Proposed 3DH_CNN

The proposed 3DH_CNN is inspired by the modified signet VGG architecture developed in [36], a moderately deep CNN model that demonstrated state-of-the-art classification and precision for 3D sign language recognition. The architecture of 3DH_CNN is shown in Fig. 4. It has 8 convolutional layers followed by a max pooling and ReLu layers. Drop out of 0.5 was introduced at the end of 8^{th} layer for inducing nonlinearity into the feature vectors. Two dense layers and a SoftMax were present at the end of the network to assign class probabilities during training and testing. The filter sizes in each layer are kept constant at with an increasing filter numbers every two layers. The dual constant filter layers are 16, 32, 64 and 128. The image resolution of 256 × 256 is considered for both training and testing to match the filter resolutions and their number which avoided vanishing gradients.

D. Training 3DH_CNN

Python 3.7, with a Keras frontend and a TensorFlow backend is used for implementing 3DH_CNN on our KL_3DHG dataset with 220 class labels. We used the same hyperparameters for all datasets, except for the learning rate, which was reassigned during training for benchmark dataset DGH 14/18. Specifically, we decreased the learning rate exponentially from 0.001 until the error became constant. At the start of the training phase for each dataset, we set the network's weights and bias parameters randomly using a zero-mean Gaussian distribution function with variance 0.01.

The 3DH_CNN learned by updating its weights and bias parameters using the back propagation gradient descent algorithm. We applied ReLu and SoftMax hyperparameter activations in the convolutional and dense layers, respectively. Finally, we used a fixed batch size of 64 for training, based on the image resolution and amount of GPU memory available. During training, we used k-fold cross validation, setting the k value at 20% of the training set. After training on each dataset, the trained model was saved, and then its hyperparameters were tuned based on feedback acquired through layer visualizations. Later, we compared our model's performance against those of several state-of-the-art DNNs used for 3D hand gesture recognition in [27], [28], [8], [29], [30], [35]. The training accuracy and loss functional plots are shown in Fig. 5 from the proposed 3DH_CNN on KL_3DHG.

E. Testing and Performance Evaluation

After training on each dataset, the CCNN and the other DNNs were tested on the test sets described in Table I. Table I shows the recognition accuracies obtained for each of the two skeletal datasets for hand gesture recognition which are averaged over the entire set. Further, video-based 3D hand gesture recognition based on CNNs with datasets in [41] and [42] were also tested with our network. These results show that our 3DH_CNN recognition accuracies were higher than those of the state-of-the-art DNNs. The proposed 3DH_CNN showed no signs of disappearing gradients, and weight decay was relatively smooth in the dense layers. The promising results for the 2D video hand gesture datasets inspired us to look into the more difficult question of identification of 3D human action skeletal dataset such as NTU RGB D, HDM05 and CMU [27].

IV. Experimental Evaluations and Discussions

Firstly, the proposed method is being evaluated for 3D hand gesture skeletal data characterizations using JRDMs with

TABLE I. Prediction accuracies achieved for two hand skeleton datasets

Datasets	Recognition Rates (%)						
	VGG	CNN+ LSTM	CNN+ RNN	Multi-Stream CNN	GoogLeNet	Connived ResNet	3DH_CNN (Proposed)
DGH 14/28	86.23	88.82	88.31	91.52	93.07	93.86	96.07
KL_3DHG	84.36	86.48	86.37	88.96	91.16	91.75	94.32

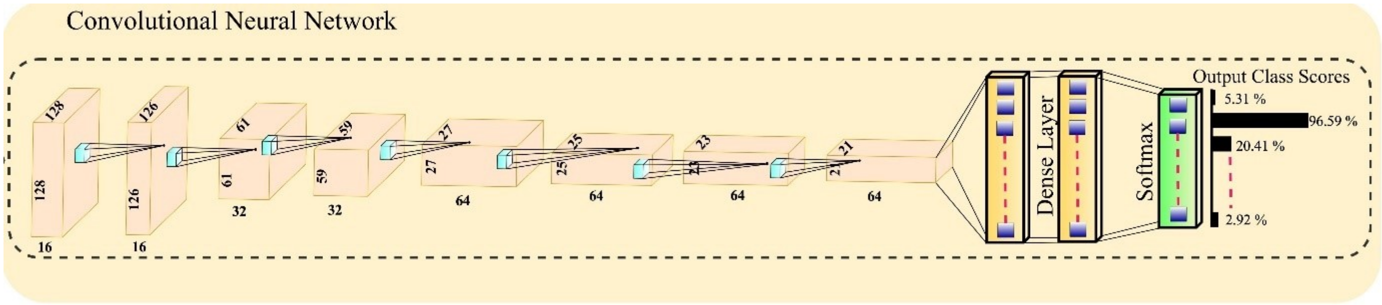
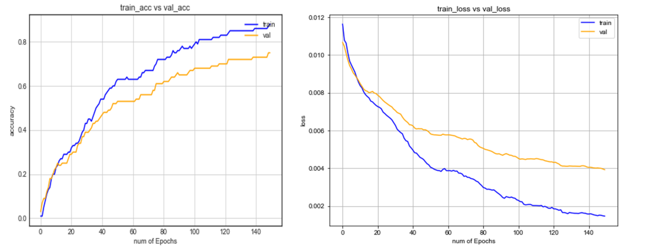


Fig. 4. Proposed 3DH_CNN for 3D skeletal hand gesture recognition



(a) (b)

Fig. 5. (a) accuracy Vs epochs and (b) loss Vs epochs.

3DH_CNN. Second, various colour coded maps will be tested with 3DH_CNN on KL_3DHG and DGH 14/28 hand skeletal datasets. Thirdly, different DNNs gauge the performance of the proposed 3DH_CNN on the two hand gesture datasets. Finally, we test the performance of the proposed JRDMs on 3D skeletal action datasets with 3DH_CNN and other popular models.

A. Evaluation on KL_3DHG with 3DH_CNN

The implementation is derived from Keras and TensorFlow toolboxes available in python 3.6 with considerable adjustments during training and testing. The training is accomplished on an 8GB GPU from NVIDIA with model number GTX1080. The proposed 3DH_CNN is tested on KL_3DHG mocap data on the above GPU system. Performance of each network with the proposed JRDA encoding format is evaluated with respect to mean average recognition (mAR) on the entire training set. The 3DH_CNN is shown examples from 8 subjects in 2 views during training and the remaining 2 subjects with 2 views are applied during testing. Table II shows the mAR for both same and cross subject test results. It also shows results of same and cross view testing.

In this part, we plot confusion matrices of the proposed JRDM's on our 3DH_CNN architecture resulted from cross subject and cross view testing of the trained network. Fig. 6 and 7 shows the confusion matrix for 30 hand gestures in Indian sign language. The confusion matrices clearly show the influence of putting relational information between hands into distance maps together with the help of Eq. (4). The overall recognition accuracies achieved are around 94.32% for cross subject and 91.28% for cross view testing respectively.

TABLE II. MAR FOR FEW SIGNS IN OUR KL_3DHG DATASET

3D Hand Gestures	Same subject	Cross subject	Same View	Cross View
Eat	0.9867	0.9738	0.9845	0.9692
Read	0.9899	0.9756	0.9894	0.9711
Hi	0.9946	0.9912	0.9969	0.9893
Good	0.9904	0.9824	0.9891	0.9723
North	0.9975	0.9889	0.9985	0.9862
East	0.9994	0.9918	0.9865	0.9812
Biscuit	0.9912	0.9893	0.9817	0.9734
Breakfast	0.9704	0.9671	0.9661	0.9417
Curd	0.9927	0.9827	0.9914	0.9751
Puri	0.9819	0.9698	0.9775	0.9576
Food	0.9964	0.9911	0.9924	0.9727
Cake	0.9918	0.9865	0.9867	0.9687
Ball	0.9345	0.9294	0.9259	0.9176
Sports	0.9934	0.9833	0.9989	0.9871
Trophy	0.9899	0.9817	0.9962	0.9815
Games	0.9845	0.9798	0.9808	0.9572
Badminton	0.9264	0.9175	0.9159	0.8973
Lose	0.9891	0.9795	0.9711	0.9669
Volleyball	0.9847	0.9768	0.9835	0.9714
Assembly	0.9221	0.9115	0.9152	0.8917
Power	0.9843	0.9721	0.9814	0.9656
Strike	0.9795	0.9689	0.9786	0.9512
Leader	0.9449	0.9412	0.9412	0.9256
Flag	0.9528	0.9465	0.9573	0.9487
Corn	0.9733	0.9663	0.9721	0.9617

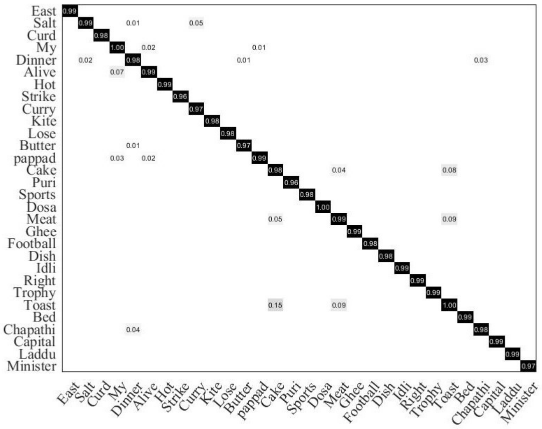


Fig. 6. Cross subject Confusion matrix for 30 class 3D hand gesture data

B. Evaluating Colour Coding on 3D Hand Gestures

The proposed JRDM's based colour texture encoding method is tested on our KL_3DHG mocap dataset and one publicly available 3D hand gesture dataset DGH 14/28. Two CNN's are built with the proposed architecture on images

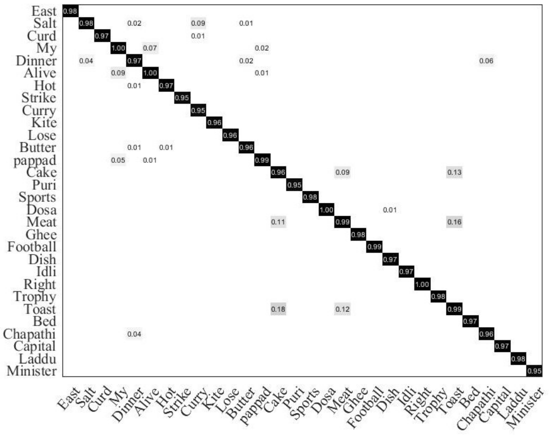


Fig. 7. Confusion matrix for 30 hand gestures tested with cross view data

encoded with our JRDM's and other popular maps from joint distance maps (JDMs) [36], joint angular displacement maps (JADMs) [37], joint velocity maps (JVM) [38], joint quad maps (JQM) [39] and joint trajectory maps (JTM) [40]. We present the average recognition accuracies for JRDM and other encoded images for front view, cross view and cross subject evaluation on the two datasets in Table III on both the datasets.

The superior performance registered by JRDM's over other maps on 3D hand gesture datasets can be attributed to joint relational information that provides relationships between joints on both hands. All the values are averaged over the number of test subjects used for testing the proposed 3DH_CNN.

C. Performance of Hand Gesture Recognition on state – of – the – art CNNs.

The image encoding model is further evaluated on popular state of the art single stream CNN architectures, to prove that the encoding mode is universal across architectures. Training for all architectures is given from scratch by keeping the network attributes such as learning rate, learning momentum and stopping criterion as common.

From Table III, the recognition accuracies for cross subject and cross view show that JRDM type colour texture encoding is better than all other encoding on our KL_3DHG data. All 220 class labels are tested with an encoded image size of as input for each deep net architecture. Table IV gives the mRA of the networks trained with JRDM's and other maps on benchmarked deep learning models. The cross-view scores are a little less than cross subject scores in all the cases due to inter finger occlusions in joints during the signing process.

D. Performance on 3D Skeletal Action Recognition

This section evaluates the advantages of using our JRDMs across different 3D skeletal action datasets with multiple DNN classifiers. Table V lists the recognition accuracies on HDM05 [46], CMU [47] and NTU RGB-D [48] 3D action datasets. The JRDMs were generated on the positional vectors using the process described in Section 3. All the maps were normalised and resized to 256 × 256, irrespective of number of joints in the skeletons. Since, the performance of other maps has already been reported in earlier works [36], [37], [38], we recommend the reader to refer them for drawing conclusions with the present relative geometric maps.

V. Conclusion

This work proposes Joint Relational Distance maps (JRDM's) for representing spatio temporal information in 3D mocap hand gesture recognition data. Unlike, Joint other previously proposed maps for action recognition, the proposed JRDM maps to rich colour coded images with local information is computed using paired joint distances of left- and right-hand joint distances. Further, a 3DH_CNN architecture is proposed for classifying the encoded images. The CNN's are trained from scratch with KL_3DHG and DGH 14/28 hand gesture datasets. The results show the JRDM encoded images generate unique representations of 3D mocap hand gesture data which are recognized with deep CNN frameworks.

TABLE III. Comparing Different Formats of Colour Texture Encoding on 3D hand gesture datasets for performance evaluation.

Dataset		% mRA					
		JDM	JQM	JADM	JVM	JTM	JRDM
Front View	KL_3DHG	0.9148	0.937	0.9412	0.9185	0.8841	0.9523
	DGH 14/28	0.9426	0.9661	0.9689	0.9556	0.9118	0.9796
Cross View	KL_3DHG	0.8464	0.8832	0.8871	0.8626	0.8294	0.9128
	DGH 14/28	0.8691	0.9029	0.9101	0.8836	0.8525	0.9325
Cross Subject	KL_3DHG	0.8899	0.9212	0.926	0.9035	0.8637	0.9432
	DGH 14/28	0.9021	0.9338	0.943	0.9208	0.884	0.9607

TABLE IV. Recognition rates for State-of-the-art CNN models

Architecture	JDM		JADM		JVM		JQM		JTM		JRDM	
	Cross Subject	Cross View	Cross Subject	Cross View	Cross Subject	Cross View	Cross Subject	Cross View	Cross Subject	Cross View	Cross Subject	Cross View
VGG [37]	0.7788	0.7363	0.8352	0.7834	0.8147	0.7486	0.834	0.7733	0.7713	0.7203	0.8436	0.8009
CNN+LSTM [43]	0.803	0.7554	0.8533	0.7965	0.8337	0.783	0.8518	0.8058	0.7867	0.7465	0.8648	0.8188
CNN+RNN [44]	0.8043	0.7635	0.8553	0.807	0.8372	0.7906	0.8592	0.8149	0.7916	0.7454	0.8637	0.8204
Multi-Stream CNN [40]	0.8249	0.779	0.8755	0.836	0.8558	0.8104	0.8738	0.8329	0.8213	0.7665	0.8896	0.8393
GoogLeNet [45]	0.8514	0.7959	0.8983	0.8583	0.8715	0.8282	0.8915	0.8522	0.8332	0.7844	0.9116	0.8635
Connived ResNet [38]	0.8558	0.8098	0.8972	0.8628	0.8744	0.8365	0.8933	0.8545	0.8358	0.7931	0.9175	0.8655
3DH_CNN (Proposed)	0.8899	0.8464	0.926	0.8871	0.9035	0.8626	0.9212	0.8832	0.8637	0.8294	0.9432	0.9128

TABLE V. Performance of JRDMs across public 3D skeletal datasets

Architecture	Datasets	Validation Error (%)	Cross Subject	Cross View
VGG	HDM05	5.54	85.35	83.79
	CMU	6.67	79.92	78.42
	NTU RGB-D	5.92	82.17	80.52
CNN+LSTM	HDM05	4.84	87.53	84.92
	CMU	5.55	82.15	80.26
	NTU RGB-D	4.96	84.27	82.11
CNN+RNN	HDM05	4.59	87.61	84.45
	CMU	5.29	83.35	81.21
	NTU RGB-D	4.77	84.27	81.92
Multi-Stream CNN	HDM05	4.21	90.41	87.71
	CMU	5.03	84.32	81.27
	NTU RGB-D	4.42	86.43	85.57
GoogLeNet	HDM05	4.18	91.24	89.14
	CMU	4.93	86.12	84.46
	NTU RGB-D	4.37	88.11	86.19
Connived ResNet	HDM05	3.53	92.54	90.35
	CMU	4.95	91.34	89.63
	NTU RGB-D	3.26	91.01	88.41
3DH_CNN (Proposed)	HDM05	2.06	97.52	96.37
	CMU	3.21	94.29	92.75
	NTU RGB-D	2.17	96.68	95.24

References

[1] Xiaoming Yin and Ming Xie. Estimation of the fundamental matrix from uncalibrated stereo hand images for 3d hand gesture recognition. *Pattern Recognition*, 36(3):567–584, 2003.

[2] Zhou Ren, Junsong Yuan, Jingjing Meng, and Zhengyou Zhang. Robust part-based hand gesture recognition using kinect sensor. *IEEE transactions on multimedia*, 15(5):1110–1120, 2013.

[3] Wei Zeng, Cong Wang, and Qinghui Wang. Hand gesture recognition using leap motion via deterministic learning. *Multimedia Tools and Applications*, 77(21):28185–28206, 2018.

[4] Tomasz Kapuściński, Mariusz Oszust, and Marian Wysocki. Hand gesture recognition using time-of-flight camera and viewpoint feature histogram. In *Intelligent Systems in Technical and Medical Diagnostics*, pages 403–414. Springer, 2014.

[5] Hong Cheng, Lu Yang, and Zicheng Liu. Survey on 3d hand gesture recognition. *IEEE transactions on circuits and systems for video technology*, 26(9):1659–1673, 2015.

[6] Kang Ling, Haipeng Dai, Yuntang Liu, and Alex X Liu. Ultragesture: Fine-grained gesture sensing and recognition. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2018.

[7] Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, 43(1):1–54, 2015.

[8] Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerry, Bertrand Le Saux, and David Filliat. Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset. 2017.

[9] William T Freeman. Dynamic and static hand gesture recognition through low-level image analysis, September 26 1995. US Patent 5,454,043.

[10] Pragati Garg, Naveen Aggarwal, and Sanjeev Sofat. Vision based hand gesture recognition. *World Academy of Science, Engineering and Technology*, 49(1):972–977, 2009.

[11] Xia Liu and Kikuo Fujimura. Hand gesture recognition using depth data. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 529–534. IEEE, 2004.

[12] Myoung-Kyu Sohn, Sang-Heon Lee, Dong-Ju Kim, Byungmin Kim, and Hyunduk Kim. 3d hand gesture recognition from one example. In *2013 IEEE international conference on consumer electronics (ICCE)*, pages 171–172. IEEE, 2013.

[13] Yidan Zhou, Jian Lu, Kuo Du, Xiangbo Lin, Yi Sun, and Xiaohong Ma. Hbe: Hand branch ensemble network for real-time 3d hand pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 501–516, 2018.

[14] Qing Chen, Nicolas D Georganas, and Emil M Petriu. Real-time vision-based hand gesture recognition using haar-like features. In *2007 IEEE instrumentation & measurement technology conference IMTC 2007*, pages 1–6. IEEE, 2007.

[15] Samy Bakheet and Ayoub Al-Hamadi. Hand gesture recognition using optimized local gabor features. *Journal of Computational and Theoretical Nanoscience*, 14(3):1380–1389, 2017.

[16] R Grzeszcuk, Gary Bradski, Michael H Chu, and J-Y Bouguet. Stereo based gesture recognition invariant to 3d pose and lighting. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 826–833. IEEE, 2000.

[17] Chenyang Zhang, Xiaodong Yang, and YingLi Tian. Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[18] Wei-Syun Lin, Yi-Leh Wu, Wei-Chih Hung, and Cheng-Yuan Tang. A study of real-time hand gesture recognition using sift on binary images. In *Advances in Intelligent Systems and Applications-Volume 2*, pages 235–246. Springer, 2013.

[19] Peter Sykora, Patrik Kamencay, and Robert Hudec. Comparison of sift and surf methods for use on hand gesture recognition based on depth map. *Aasri Procedia*, 9:19–24, 2014.

[20] Nasser Dardas, Qing Chen, Nicolas D Georganas, and Emil M Petriu. Hand gesture recognition using bag-of-features and multi-class support vector machine. In *2010 IEEE International Symposium on Haptic Audio Visual Environments and Games*, pages 1–5. IEEE, 2010.

[21] Saba Jadooki, Dzulkifli Mohamad, Tanzila Saba, Abdulaziz S Almazyad, and Amjad Rehman. Fused features mining for depth-based hand gesture recognition to classify blind human communication. *Neural Computing and Applications*, 28(11):3285–3294, 2017.

[22] Manuel Caputo, Klaus Denker, Benjamin Dums, and Georg Umlauf. 3d hand gesture recognition based on sensor fusion of commodity hardware. *Mensch & Computer 2012: interaktiv informiert–allgegenwärtig und allumfassend!?*, 2012.

[23] A Safaei and M Jahed. 3d hand motion evaluation using hmm. *Journal of Electrical and Computer Engineering Innovations*, 1(1):11–18, 2013.

[24] Hong Cheng, Zhongjun Dai, Zicheng Liu, and Yang Zhao. An image-to-class dynamic time warping approach for both 3d static and trajectory hand gesture recognition. *Pattern recognition*, 55:137–147, 2016.

[25] Hee-Deok Yang. Sign language recognition with the kinect sensor based on conditional random fields. *Sensors*, 15(1):135–147, 2015.

[26] Hong Cheng, Jun Luo, and Xuewen Chen. A windowed dynamic time warping approach for 3d continuous hand gesture recognition. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[27] Teja Kiran Kumar Maddala, PVV Kishore, Kiran Kumar Eepuri, and Anil Kumar Dande. Yoganet: 3-d yoga asana recognition using joint angular displacement maps with convnets. *IEEE Transactions on Multimedia*, 21(10):2492–2503, 2019.

[28] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand gesture recognition with 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–7, 2015.

[29] Guillaume Devineau, Fabien Moutarde, Wang Xi, and Jie Yang. Deep learning for hand gesture recognition on skeletal data. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 106–113. IEEE, 2018.

[30] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre. Skeleton-based dynamic hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2016.

[31] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016.

[32] Juan C Nunez, Raul Cabido, Juan J Pantrigo, Antonio S Montemayor, and Jose F Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, 2018.

[33] Chinmaya R Naguri and Razvan C Bunescu. Recognition of dynamic hand gestures from 3d motion data using lstm and cnn architectures. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1130–1133. IEEE, 2017.

[34] Chunyong Ma, Anni Wang, Ge Chen, and Chi Xu. Hand joints-based gesture recognition for noisy dataset using nested interval unscented kalman filter with lstm network. *The visual computer*, 34(6-8):1053–1063, 2018.

[35] Reza Azad, Maryam Asadi-Aghbolaghi, Shohreh Kasaei, and Sergio Escalera. Dynamic 3d hand gesture recognition by learning weighted depth motion maps. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6):1729–1740, 2018.

[36] Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24(5):624–628, 2017.

[37] E Kiran Kumar, PVV Kishore, ASCS Sastry, M Teja Kiran Kumar, and D Anil Kumar. Training cnns for 3-d sign language recognition with color texture coded joint angular displacement maps. *IEEE Signal Processing Letters*, 25(5):645–649, 2018.

[38] Eepuri Kiran Kumar, PVV Kishore, Maddala Teja Kiran Kumar, Dande Anil Kumar, and ASCS Sastry. Three-dimensional sign language recognition with angular velocity maps and connived feature resnet. *IEEE Signal Processing Letters*, 25(12):1860–1864, 2018.

[39] D Anil Kumar, ASCS Sastry, PVV Kishore, E Kiran Kumar, and M Teja Kiran Kumar. S3drgf: Spatial 3-d relational geometric features for 3-d sign language representation and recognition. *IEEE Signal Processing Letters*, 26(1):169–173, 2018.

[40] Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43–53, 2018.

[41] Hong Cheng, Zhongjun Dai, and Zicheng Liu. Image-to-class dynamic time warping for 3d hand gesture recognition. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.

[42] Alexey Kurakin, Zhengyou Zhang, and Zicheng Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *2012 Proceedings of the 20th European signal processing conference (EUSIPCO)*, pages 1975–1979. IEEE, 2012.

[43] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 585–590. IEEE, 2017.

[44] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3461–3470, 2017.

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[46] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database hdm05. 2007.

[47] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58. IEEE, 2002.

[48] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.