International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Multi Modal RGB D Data based CNN Training with UNI Modal RGB Data Testing for Real Time Sign Language Recognition

Sunitha Ravi, M. Suman, P.V.V. Kishore, E. Kiran Kumar, M. Teja Kiran Kumar and D. Anil Kumar

Abstract*: At present, the accuracy of translating video based sign language into text or voice remains indistinct and is therefore an interesting and challenging problem for computer scientists. Higher accuracies can now be achieved by applying deep learning models for sign language recognition (SLR), which was done successfully for human action recognition problem. This inspired us to investigate convolutional neural networks (CNN) for translating 2D sign videos into text. To this end, we propose a novel four steam CNN architecture with multi modal training (MT) with RGB and depth data; and unimodal testing (UT) with only RGB data on RGB – D sign language video data. The four streams cluster into two native modal streams, RGB and depth. Based on the domain characteristics, native modals were divided into two modal specific spatial and temporal streams. The major drawback of feeding raw sign video for training can be ineffective due to the small variations in sign language data compared to large background variations in the video sequence. We have observed this overfitting problem by preliminary experimentation, where the CNNs learn the noisy background rather than the foreground signer. The overfitting problem was solved by feature sharing mechanism between RGB and depth modals. Experimental results show that the proposed CNN is capable of predicting the class labels with unimodal data (RGB) only. We tested the performance of the proposed MTUTCNN architecture on our own RGB – D sign language (BVCSL3D) and three RGB -D based action datasets for scale, subject and view invariance. Results were validated against current state – of – the – art deep learning based sign language (or action) recognition models. Our study shows a recognition rate of 91.93% on BVCSL3D dataset.

***Index Terms*: 3D sign language recognition, Feature sharing CNNs, Convolutional neural nets, Multimodal training, Unimodal testing.**

Revised Manuscript Received on 30 May 2019.

***** Correspondence Author

Sunitha Ravi*, Department of ECE, K.L.E.F., Green Fields, Vaddeswaram, Guntur, A.P., INDIA - 522502

M. Suman, Department of ECE, K.L.E.F., Green Fields, Vaddeswaram, Guntur, A.P., INDIA – 522502.

P.V.V. Kishore, Department of ECE, K.L.E.F., Green Fields, Vaddeswaram, Guntur, A.P., INDIA - 522502

E. Kiran Kumar, Department of ECE, K.L.E.F., Green Fields, Vaddeswaram, Guntur, A.P., INDIA – 522502.

M. Teja Kiran Kumar, Department of ECE, K.L.E.F., Green Fields, Vaddeswaram, Guntur, A.P., INDIA - 522502

D. Anil Kumar, Department of ECE, K.L.E.F., Green Fields, Vaddeswaram, Guntur, A.P., INDIA – 522502.

I. INTRODUCTION

Due to its locale hand motion attributes, which appear small in respect of the background content of the sign video, sign language recognition is still used with color videos with a simple background. This highlighted the hand-tracking and segmentation of form algorithm that used RGB video sign interpretation color information. These methods have produced good segmentation and accuracy in non-complex video contexts. In real world scenarios like anatomical human variations, self-occlusion, light variations and camera vibrations are challenged, in addition to the complex video backgrounds, gestures recognition.

In human activity recognition (HAR) videos, colour, depth and temporal information have been extensively investigated. In order to identify efficiently activities under complex circumstances, LST (Local Spatio Temporal) methods are becoming increasingly popular in addition to spatial and zeital methods. LST features characterized textures and forms in which the underlying activities in video sequences are distinctly modeled. LST character maps describe a human pose in the neighborhood of certain pre-select interest points by means of local form and motion information. Invariance of LST features to the size of the image, direction, translation, lighting and partial occlusions. In RGB videos, LST features are commonly utilized for human representation. The LST features have limited the removal of meaningful information to distinguish close correspondence from them, although they have been successful. The availability of low-cost RGB-D-sensors, such as Microsoft Kinect, also increased the use of depth data to increase poorly captured color information. In addition, the measurements using the depth information represent the signator by distinguishing human subjects from the background. The LST feature extraction now uses the depth data to locate interest points in the RGB data, which is again a spatial – depth image registration problem. This problem exists due to the scale variations in depth and RGB data. Convolutional Neural Network (CNN) is the most widely used deep learning algorithm for recognizing and classifying human actions from video data. CNNs use multiple layers of convolution operations during training by iterative filter optimizations for generating multi band features.

Iterative parameter optimization method such as error back propagation learns feature localization for image classification, object detection and video classification applications. CNNs have always been a popular choice with researchers for human action recognition. A large component of work has focused on using the traditional CNNs with RGB video input and its variants such as temporal and saliency maps. However, availability of depth information, the trend seems to have changed: CNNs evolved into multi stream structures with RGB and depth modal input [1] in each stream. Inspired by the previous works on multi stream architectures for human action classification, a unique CNN architecture for SLR is investigated subsequently.

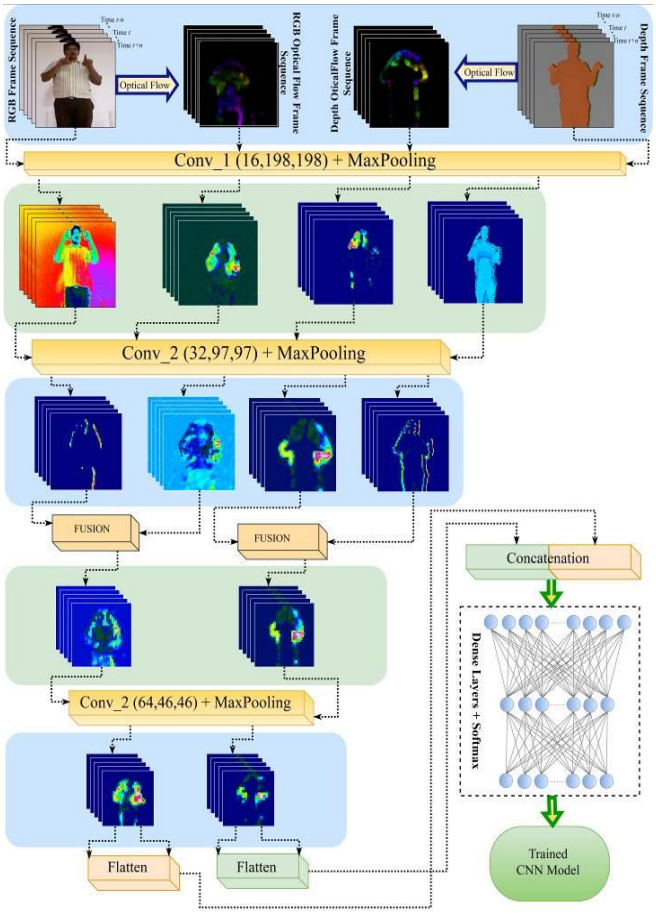


Fig. 1. Multi modal spatio temporal Co-Trained CNN architecture.

Sign language is a sub and a very complicated domain of human action. In this paper, we explore the relationship between multi modal data on a multi stream CNN. In this proposed CNN architecture, we have used four streams with two modal specific inputs in a spatio temporal framework. In each modal specific stream, two CNNs operate with inputs from spatial and temporal i.e. colored optical flow maps. Thus, the four streams get inputs from RGB spatial and temporal maps; depth spatial and temporal maps.

Three convolutionary layer, two dense layer and one Soft Max layer are constructed into this proposed CNN architecture. We introduced two additional fusion layers in RGB specific and depth specific streams for generating a region of interest specific features (ROISF). RIOSFs share spatial and temporal features in two modal streams as a part of feature sharing mechanism. ROISFs merges spatial and temporal streams in each modal specific stream into one. Now, each modal specific stream has only one CNN stream. The output features from each modal specific stream are shared in the flattening layer, which creates a distinctive feature vector representing the sign. Finally, the flattened multi modal shared feature vector flows into the dense layers and a predictive SoftMax layer. Fig.1 shows the proposed CNN architecture.

In the sign language video data, the human subject has a 20% density of pixels compared to the remaining 80% of redundant background pixels. This greatly reduces the chance that raw RGB data with CNNs will detect and recognize gestures. Therefore, RGB and the depth data used in past works are used as two separate CNN spatial streams. The scores of all streams were then combined to create a predictive recognition probability score. However, the RGB and the depth streams of one of the layers do not have information which requires both RGB and depth inputs in order to test for a good grade. There are uncertain results due to the lack of any modal data during tests.

The proposed design of the multi modal trained and uni modal tested CNN (MTUTCNNs) solves these two problems. Firstly, it localizes the human signer in the RGB specific and depth specific video data using spatio temporal fusion in the ROISF layers. This procedure extracts the signer action feature by eliminating the background and other redundant information.

The multi modal features are then mixed to generate a nonlinear feature vector to feed the dense layers. This novel CNN design, ensures speedy convergence during training, with no dropouts in the final layers, which is used to create a nonlinear feature to avoid overfitting. Secondly, for a real-time sign language recognizer on portable platforms, only RGB camera is available for capturing the input. Hence, the proposed MTUTCNN takes only RGB and its motion map as input for recognition. The feature fusion blocks use average fusion of the spatio temporal features in RGB and depth modes. The 2nd multi modal fusion of features use sparse concatenation to generate a nonlinear feature vector for the dense layers.

The remaining paper is arranged accordingly. The following section presents a brief review of multi-modal RGB-D symbols with state-of - the-art methodologies. Section 3 details the architecture, training and testing modules proposed by the MTUTCNN. Section four provides experiments on the results and analysis on our sign language and RGB-D datasets of the proposed CNN architecture.

II. RELATED WORK

1D, 2D and 3D data are being explored using multiple signal, image and video processing algorithms for feature representation to identify sign language action (SLR). In order to transmit finger movements to a microcontroller for detection, the 1D SLR uses wireless radio frequency technology controlled glov. When the accent is placed only on the faster 1D models, good recognition rates occur.

The sign language, in addition to the hands, includes expressions of the head, torso and face as they are developed as a visual language. Compared with 1D data gloves, 2D sign video data generates more information to improve recognition.

The two-dimensional capture explores all the elements of a VLT with speed and accuracy constraints. In addition, 2D SLRs are widely investigated in Hidden Markov models with continuous and discreet sign language versions. HMMs are faster if features removed from the 2D sign video accurately represent the sign, but because of blurrence, illumination, background, shadows and occlusions this is difficult. However, isolating a hand movement in a video sequence with extracting finger shapes under the influence of light, fluidity and occlusions is still very difficult. With the help of additional information, the solution is found that is somewhat immune to brightness, blurring and occlusion.

Kinect sensor captures RGB-D video frames that support the traditional RGB color video data. Signator's images of Almeida et al. [2], RGB-D (rough, green, blue and deep), for seven different sign-linguistic features from Brazil. Multi-class support vector classifier classifies the extracted multiple features. In 10,000 sign frames this approach recorded a 80 percent recovery rate. Shao-ZiLi et al. [3] developed a sparse self-encoder (SAE) and RGB-D input principle language sign language analysis features. The characteristics of RGB and Depth are learned via a convolutions neural network and then the extracted features are combined with the principle component analysis (PCA). Experimental ASL results showed an acknowledgement rate between 75 and 99.2 %. The Kinect sensors 3D data are manual paths, guidelines and speeds from a single image of depth. For sign classification features such as 3D body joint locations and finger earth movers distances (FEMD) are used. Sutarman et al. [4] suggested the Dynamic Malaysian sign language recognition system, using skeletal data traceability to capture 3D Image data from the Kinect sensor. The x, y, z coordinates in relation to the head and spine positions are used for extracting the feature. The spherical process of co-ordinate conversion is achieved using dimension matching segmentation. For classification with node variations in the hidden layers the back propagating neural networks are used. With an accuracy of 80.54 %, the system identifies 15 Malaysian sign language gestures. This system offers an advantage in image acquisition through profound images and uses infrarot light to make the light conditions dependent. In the present scenario, deep learning offers a solution to many image and video classification problems to pattern recognition from end to end. Conv networks will only be used in different scales in previous work to find solutions to 3D-HAR problems [5]. The RGB-D data or the combination of the both skeletal data is used in several streams in the case of CNN. The RGB-video frames in the MSR action data set are used by single-stream CNN models as inputs with multiple convolutionary strata which extract features for the SoftMax layer classification [6]. Various previous works on video actions datasets for KTH and MSR 2D [7] demonstrate meager performance because temporal data about a action object does not exist. Most of the CNN single stream uses only RGB spatial data to process in 2D areas, which lead to large errors. RGB D-based inputs with their functions provide the network with additional information for learning in different events. However, for each type of input map this CNN job requires several CNN streams and a fusion model. Multi-channel CNN's with RGB stream and profound stream were separately introduced in this era of deep networks as a CNN input [8]. In a similar way, two and three CNN streams were implemented, one focusing on spatial and the other temporal data [9], on video data. The HAR performance has been enhanced by the late 6-stream architecture of 2D RGB videos on the CNN. The correct position is identified precisely through three spatial streams with multi-scaling input RGB data and their opticalflow maps. Inspired by Convolutionary networks for RGB–D data for the recognition of sign language and human actions, this document offers a new improved architecture in order to build a real-time model of recognition of sign languages. The past CNNs have been used in several modes, i.e. In multi-stream mode, the RGB Color and Depth. Two or more CNNs are fed in multi-streams with color and depth data in space and time. Finally, the results achieved by each stream are fused to produce an input-related class score. These types of CNN architectures share the features from one stream to another in multi modal data. This leads to two constraints: one, the issue of data overfit nd two, unable to handle missing data in one mode. The solution to these issues is the new CNN architecture in figure. 1. Correct learning occurs when the network attempts to generalize the data sequence characteristics. However there are multiple instances of overfitting data to model due to large noise in action or sign language video sequences compared to the human item, resulting in false positives in the testing. To avoid it, dropouts in the dense layers are introduced in [10] to reduce noise in the background by randomly dropping 50 %. In many cases, this had a significant impact on the results. This work, however, suggests to isolate the RGB Modal Stream signature using the features obtained from deep modal steam in multiple layers as shown in fig. 1. The RGB stream, RGB motion maps and depth motion maps share the features to cancel redundancies in RGB streams. This is done at two sites throughout the flow. Finally, RGB data in dense layers and SoftMax layers are used by the past two layers. From beginning to finish, only spatial modal data flow from RGB. By sharing their clean features with the RGB data, the other streams act as support systems. Our proposed CNN architecture is co-trained on different modal data while only RGB spatial and movement data are being tested. A series of experiments in our sign language RGB–D (BVCSL3D) and other action recognition benchmarks, for example UT–Kinect, G3D and MSRDailyAction3D with clean RGB–D data, are conducted in order to validate the proposed architecture. We tested RGB video data set from Indian sign language as well. The proposed network and similar state-of - the-art multiple-stream CNNs with two, three, four and six streams are used for cross subject and cross-data assessment. Only RGB data in the same dataset, different datasets and missing RGB frame datasets are validated.

With the proposed CNN architecture, the results are encouraging, over the current multi-stream CNNs in terms of accuracy and training time.

III. PROPOSED SIGN RECOGNITION FRAMEWORK

ConvNets are deep neural networks with simple architectures that perform complicated computations for visual data learning. In this section, we present CNN architecture, training and testing related issues for RGB-D spatio-temporal data multi-stream CNNs.

A. The Proposed Multi Stream CNN Architecture (MTUTCNN)

In previous works, the multi modal CNN architectures were designed to operated separately on RGB and depth data. In the end, the probability class scores are fused after the SoftMax layer. The CNNs did not make decisions based on common color-and depth characteristics, although this was successful. In SLR, past and present operations use only functional concatenation in the dense layers without investigating the interconnection between various modes to enhance the likelihood. We thus propose a novel CNN 4 stream architecture that takes advantage of the RGB semantic character and depth data in training, leading to better recognition scores only using RGB mode testing.

The CNN architecture is trained to predict sign (or action) labels from one or two spatial data input variations, finally–to finish. The proposed RGB-D Sign Language Recognition (MTUTCNN) four-stream architecture is displayed in fig. 1. With four inputs: RGB and its motion-maps, depth and move-maps, our proposed CNN model is trained. However, only one or two inputs from only RGB video data have been tested. The CNN architecture is designed to share spatial and temporal characteristics in order to create ROISF maps representing only the signing object. Thus, the attributes in the final layers are focused on the human signature rather than the background information.

The input video frames for four streams for a sign 'Art' is shown in fig. 2. There are three convolutional layers in each stream with two pooling layers after the 2nd and 3rd convolutional layers. The pool has a stride of 2, which means the image gets down sampled by 2 after 2nd the 3rd conv net layers. All layer's use filters of size 3×3. Feature fusion is performed after the 2nd conv net in RGB and depth streams. Here we apply average fusion on RGB spatial and temporal maps to generate a region of interest specific features that flow to the next convolutional layer in the RGB stream. Similar operation happens in depth stream. Here, the 4 stream CNN mergers into a 2 – stream CNN. After the 3rd convolution layer, flattening is used to generate multi modal features. These multi modal features are concatenated to construct a complete feature vector representing a sign frame in both RGB and depth modes. This CNN architecture results in a sparse feature in spatio temporal domain. Finally, the multi model feature vector passes through two dense fully connected layers and one output SoftMax layer. The proposed MTUTCNN is different in two aspects from regular CNNs used for sign (or action) recognition. 1) RGB steam and depth stream uses spatio temporal information to generate a ROISFs for early detection of signer information. 2) The architecture is a shallow CNN, and needs less training and testing computations.

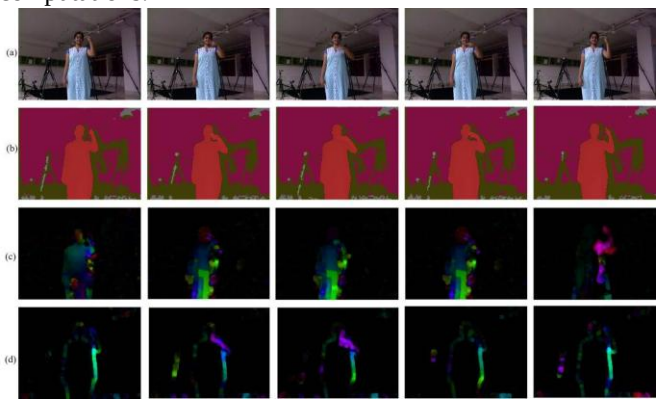


Fig. 2. Inputs to 4 – stream CNN for Sign (action) recognition, (a) RGB spatial, (b) Depth spatial, (c) RGB temporal and (d) Depth temporal.

The visualization of the above two advantages is shown in fig. 3. The input to the proposed CNN in fig. 1 is shown in the first row as four multi modal inputs to 4 – streams. RGB spatial and temporal maps; depth spatial and temporal maps are fed to the 1st convolutional layer. The RGB streams features are visualized under the heading 'RGB Stream – Conv+Pool Layer 1 Features' in the 2nd row. The depth streams are under 'Depth Stream – Conv+Pool Layer 1 Features'. The 3rd row shows features from 2nd convolution layer with max pooling.

The 4th row is the averaged RGB spatial and temporal features to select the region of interest specific features (ROISF). This eliminates 80% of the background noise in both the streams. Now, the 4 – stream CNN transforms into a 2 – stream CNN. The 5th row points to features obtained after the 3rd convolution and pooling layers. No more convolutional layers are required, as the background noise is eliminated.

The outputs of the 3rd layer is flattened into a vector in the RGB and depth streams as visualized in 6th row of fig. 3. Here, we apply multiple fusion mechanisms to identify a nonlinear feature vector. In fig. 3, 7th layer, we show concatenated feature from both streams, which flows into the 1st dense layer. The output of 1st dense layer is visualized in the 8th row of fig. 3. The last row is the output of the 2nd dense layer and the SoftMax layers. The 9th row is obtained after testing the trained CNN. The 'Yellow' color in the last row has the highest-class probability score of all the signs (or actions) in the test set.

B. Learning

The proposed CNN trains on datasets: the BVCSL3D, MSRDailyActivity3D [11], UT Kinect [12] and G3D D dataset [13]. Our BVCSL3D has 200 Indian sign language classes. In both RGB, depth and skeleton mode, each sign class is captured for 100 frames at 30fps. 10 signers performed 15 times each sign, with a total of 150 video signs per class, at multiple scales, views and hand speeds. The entire data package consists of 30000 RGB videos, with 3000000 frames each in depth and skeleton modes.

Three benchmark data sets have been used for education and testing to validate the CNN architecture. MSRDailyActivity3D is made of 20 actions, 10 subjects and 567 RGB-D action videos. UT Kinect contains ten actions, ten topics and 200 action videos. G3D is a 16 actions, 16 subjects and 1280 videos game action dataset. Based on these action videos, we have developed our own action dataset by mixing all 3 classes of data sets.

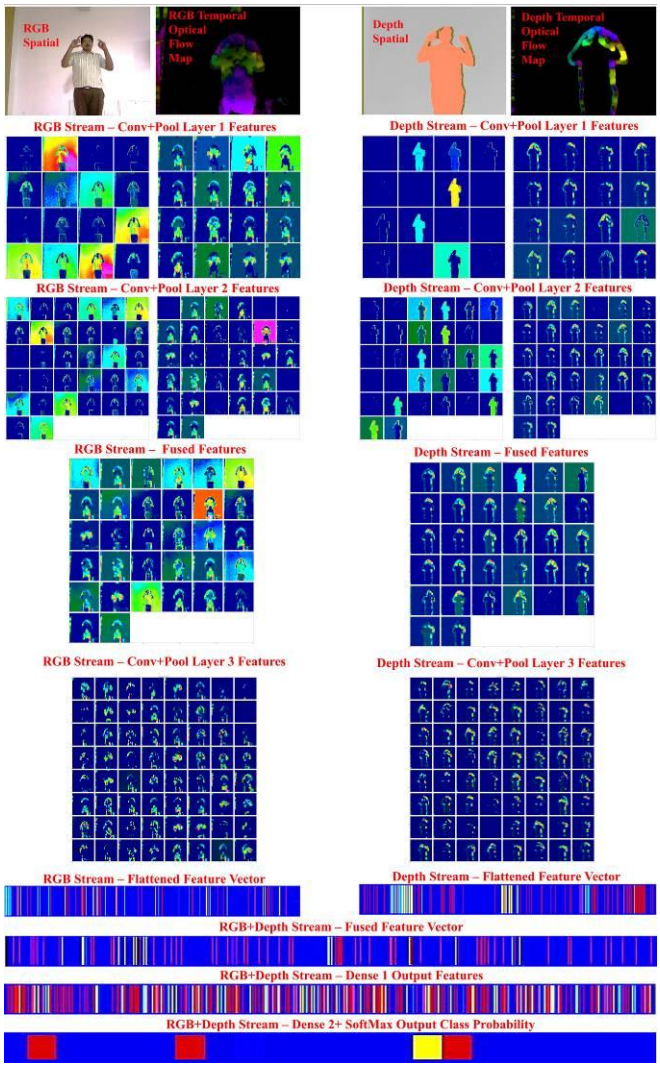


Fig. 3. Multimodal Training and Unimodal Testing CNN Feature Maps.

Each sign (or action) is labeled for training with the word that represents the right sign (or action). There are 75 video clips with 150 frames each training batch. For RGB spatial stream, 0.003 for depth spatial stream and 0.02 for both motion streams, the original training rate is chosen as 0.05. The learning rate for MSRAction3D, UT Kinect and G3D benchmark data sets was reduced by three times by 0.1. Nevertheless, after 48 K iterations and 0.1 after 31 K iterations, the learning rate was decreased twice to.02. The first decrease is following 59 K iterations for the MSRAction3D data set, and the second decrease is after 44 K iterations. After 42 K iterations, the third decrease was followed, and the network converged with a further 15K. The first drop is after 24 K for UT Kinect, the second after 18 K and the third after 16K. After 20 K iterations, convergence occurred. With projected decreases in learning rates, the G3D data converged more quickly after 22 K, 30 K and 15K. However, the decrease in our BVCSL3D data set is 2 times the 28 K, and then 24K. After 17 K iterations, the networks converged.

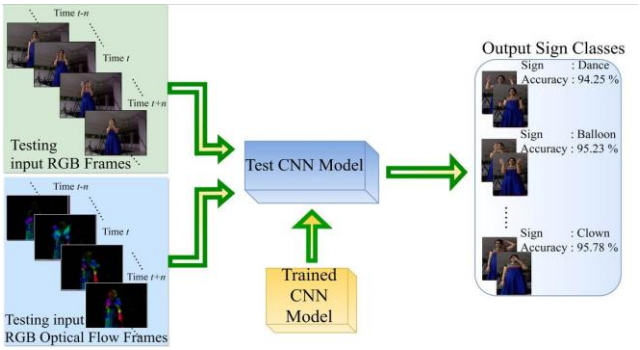


Fig. 4. Testing process for the proposed sign language recognition system with only RGB and its optical flow maps.

There is no common practice in CNNs with the proposed experimental setup after convolution layers before the dense layer. The networks do not generalize and converge further to noisy data in the image, known as overfitting, without a stop. The CNN architecture shown in fig.1 prevents the overfitting of spatial and temporal steams by using selective dumping after 2nd layer. This procedure leads to non-linearity before flattening or dense layers in the final characteristics. Non-linearity is a prerequisite for the network to generalize the data input. After every drop in learning rate, the weight decay is initialized to 0.05 and decreased to 0.1.

In order to find a potential match to the trained Sign (or Action) class, statistical measurements of the outputs of Softmax level are given. At the output of the SoftMax a high class score points to the input test sign class. In this case, the test does not use the input depth mode. The model of the test is shown in fig. 4.

IV. EXPERIMENTAL CLASSIFICATION RESULTS AND ANALYSIS

The four-stream CNN deep architecture is tested by BVCSL3D, MSRDailyActivity3D and G3D datasets from the RGB-D datasets UT Kinect. Datasets are first described and network parameter effects are then investigated during training. Secondly, on the data kits to validate the proposed CNN architecture, several networks are compared. Finally, we will present a detailed review of the past work on RGB D and the proposed CNN architecture on the recognition of CNN-based sign language. Various action datasets were considered for the experimentation along with our sign language data. Figure 5 show our sample database of RGB and depth captured from the University of K.L.E.F. Biomechanics and Vision Computing Research Centre. Our method, (MTUTCNN), was tested using four datasets: 3D Indian Sign Language, 3D Indian Sign Language, 3D Indian Sign Language and the three available public RGB-D action datasets.

With our dataset, we have created ten CNNs for validating the CNN architecture proposed against state-of-the-art SLR CNN methods. We also use the existing action recognition methods on benchmark data sets to test our MTUTCNN architecture. For testing of the trained CNN, the test stage uses RGB spatial and RGB+RGB optical flux. During testing with equal, cross, scaled subjects and cross views, we study the effects of the missing deep data.

A. *Testing for Fusion strategies after 2nd and 3rd layers of the proposed method*

In our proposed CNN architecture, we first compared multiple fusion strategies on various layers (Fig. 1). The proposed fusion rules were based on 2nd and 3rd layer characteristics. In order to integrate functionalities generated through spatial and temporeal streams in the RGB and depth streams, five fusion strategies were tested: sum, product, averaging, principle component analysis and concatenation. The conclusions of Table 1 and Table 2 were drawn by considering only 50 BVCSL3D sign classes. Each class has 50 samples trained and 50 samples tested. The fusion strategy decision is based on the rate of recognition from the class labels obtained. The course took 25 epochs, which was sufficient to decide on the strategy for fusion.

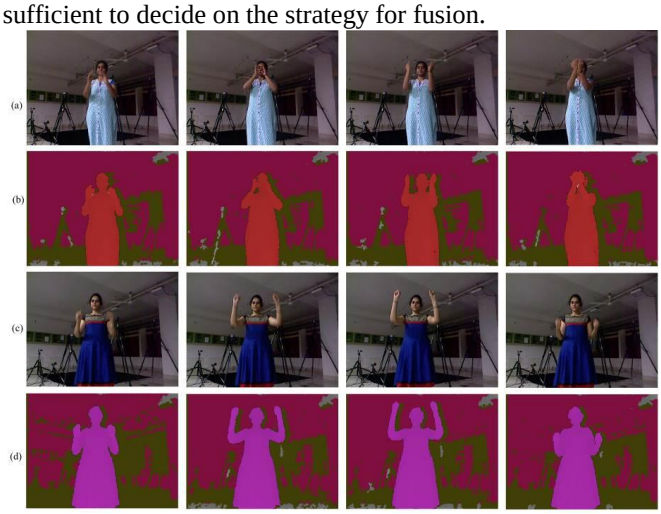


Fig. 5. BVCSL3D RGB D sign dataset, (a) Sign 'Balloon' in RGB and its (b) Depth, (c) RGB 'Circus' and (d) its depth map.

From now on we use averaging fusion and concatenation after 2nd and 3rd layers respectively for our proposed MTUTCNN. Now we have tested on our own BVCSL3D and the three benchmarking action datasets the performance of our proposed CNN (MTUTCNN). The skeleton of 500, 23,70 and 60 classes were tested on a total of 15000 BVCSL3D, 750 HDM05, 350 CMU and 1200 NTU RGB–D skeletons, respectively.

Table 1. Selection Criterion of different spatio-temporal fusion strategies after 2nd layer.

Fusion Strategy	Implementing Layer	Recognition rate (%)
Product Fusion	2nd	72.76
Sum Fusion	2nd	76.72
PCA Fusion	2nd	75.49
Concatenation	2nd	69.53
Averaging Fusion (Ours)	2nd	85.14

Table 2. Selection Criterion of different spatio-temporal fusion strategies in the flattening.

Fusion Strategy	Implementing Layer	Recognition rate (%)
Product Fusion	3rd	69.76
Sum Fusion	3rd	79.72
PCA Fusion	3rd	78.49
Concatenation (Ours)	3rd	89.53
Averaging Fusion	3rd	82.14

B. *MTUTCNN testing and validation*

The MTUTCNN deep model training offered lasted for 585 epochs with 15,000 sample signs covering five subjects, five scales and five points of view. In the network there is no drop-out. The CNN architectures with a score fusion are validated by our model against two [8] and three CNN stream [1] [9] of standard multiple stream models. The average recognition rates for all data sets for this work are shown in Table 3. The RGB+RGB optical flow videos on our MTUTCNN only are tested without feeding depth data to the two other streams.

The networks of [1], [8] and [9] were reconstructed and trained from scratch on the given datasets. The results show that our four-stream model features shared better than the other three modes with fewer training. In [8] the2-stream CNN with multiplicative score fusion includes RGB and profile spatial data. The fixed error rate of 0.0001 is the stop criteria for all networks. The other three stream networks are located in its three streams [9] and have spatial, temporal and spatial maps of RGB.

Table 3. Recognition rates of our proposed MTUTCNN against the state of the art multi stream models.

Dataset		Recognition rate (%)			
		Gao [8]	Liu [9]	Wang[1]	MTUTCNN
Same Subject with multiple orientations	BVCSL3D	81.51	84.35	86.69	88.94
	MSRDailyAction3D	84.93	85.81	89.15	91.43
	UT Kinect	82.87	85.33	88.67	90.92
	G3D	86.88	89.27	92.54	95.79
Cross Subject	BVCSL3D	75.96	81.25	85.59	88.89
	MSRDailyAction3D	85.53	86.58	88.92	91.17

	UT Kinect	84.14	84.93	88.27	90.52
	G3D	84.98	86.42	91.76	95.01
Cross View	BVCSL3D	63.96	68.25	62.59	76.89
	MSRDailyAction3D	74.53	74.58	77.92	70.17
	UT Kinect	73.14	73.93	77.27	79.52
	G3D	74.98	77.42	70.76	73.01
Cross Scale	BVCSL3D	53.96	58.25	52.59	66.89
	MSRDailyAction3D	64.53	64.58	67.92	60.17
	UT Kinect	63.14	63.93	67.27	69.52
	G3D	64.98	67.42	60.76	63.01

Table 4. Precision and Recall averages of our MTUTCNN against the methods in [14], [34] and [36].

Dataset		Gao [8]		Liu [9]		Wang [1]		MTUTCNN	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Same Subject with multiple orientations	BVCSL3D	81.51	75.81	84.35	86.65	86.69	91.44	88.94	92.94
	MSRDailyAction3D	84.93	88.66	85.81	87.42	89.15	92.14	91.43	93.25
	UT Kinect	82.87	87.24	85.33	87.73	88.67	92.71	90.92	93.90
	G3D	86.88	88.79	89.27	88.33	92.54	93.60	95.79	94.66
Cross Subject	BVCSL3D	75.96	82.57	81.25	85.25	85.59	84.01	88.89	87.56
	MSRDailyAction3D	85.53	86.96	86.58	85.59	88.92	90.33	91.17	91.98
	UT Kinect	84.14	88.54	84.93	88.41	88.27	92.25	90.52	93.33
	G3D	84.98	89.86	86.42	89.44	91.76	93.45	95.01	94.51
Cross View	BVCSL3D	63.96	62.51	68.25	65.24	62.59	70.85	76.89	71.99
	MSRDailyAction3D	74.53	66.57	74.58	65.98	77.92	70.99	70.17	71.85
	UT Kinect	73.14	68.11	73.93	68.65	77.27	72.29	79.52	73.84
	G3D	74.98	69.15	77.42	69.52	70.76	73.71	73.01	74.59
Cross Scale	BVCSL3D	53.96	72.74	58.25	75.57	52.59	75.15	66.89	79.96
	MSRDailyAction3D	64.53	76.38	64.58	75.91	67.92	79.72	60.17	78.68
	UT Kinect	63.14	78.43	63.93	78.31	67.27	72.75	69.52	73.43
	G3D	64.98	79.61	67.42	79.36	60.76	73.37	63.01	74.91

In [9] and [1] there are differing numbers of layers and fusion strategies and the number of ConvNet layers is 6 and 8 respectively. Feature concatenation is used before dense layers are started in [1] while average score fusion was utilized in [9] after layer SoftMax. The three nets in [1], [8] and [9] used 50 percent drop out for our network before the dense layers.

In the course of tests with 0.96 score thresholds, accuracy and recall metrics were calculated to further validate the MTUTCNN. Table 4 shows the average accuracy and recall values. Table 4 values show the precise way to predict a sign (or an action) for a single RGB sign video during testing. As shown in fig, in our test system. 4, as inputs for the above CNNs only RGB and RGB optical flux are supplied.

Fig. 6. Cross views and multi scale train and test data in BVCSL3D.

Despite the inclusion of cross scaled and cross view data in the training set, the false positives detected during testing are more compared to cross subject and same subject with different hand orientations. However weak may be the precision and recall values are for cross view and cross scale, they are compatible with the global average of 70%. Fig. 6 and fig. 7 shows the cross views and cross subject samples in the BVCSL3D Indian sign language dataset.

Fig. 7. BVCSL3D Cross Subject data.

C. *Validating the BVCSL3D dataset*

Here, our RGB D sign language dataset is being validated with the above4-stream model in other CNN network architectures, such as AlexNet, VGG16, Google LeNet and ResNet. During the training or testing, the number of layers and the initialization parameter were not affected. The nets have been trained with only 15000 videos and the remaining 15000 video clips with indian sign language datasets.

The average recognition rate for multi-modal network tests is reported in Table 5. We conclude that different networks can accept the data set universally. The four-stream architecture proposed for the construction of these four networks was used in black box models.

Table 5. Validation of AlexNet, VGG16, GoogLeNet, ResNet and the proposed CNN on image data.

Multi Modal Inputs	Networks				
	AlexNet	VGG16	GoogLeNet	ResNet	Proposed
RGB Only (1-Stream)	82.73	86.91	86.28	86.22	86.62
Depth Only (1-Stream)	84.44	86.19	87.91	87.17	87.05
RGB + Depth (2-Stream)	84.49	86.59	87.77	87.08	91.59
RGB + RGB Optical Flow (2-Stream)	84.94	89.96	89.92	89.96	91.93
Depth + Depth Optical Flow (2-Stream)	83.91	86.73	86.35	87.44	90.97
RGB + RGB Optical Flow + Depth (3-Stream)	85.02	89.79	89.99	90.03	91.98
RGB + Depth + Depth Optical Flow (3-Stream)	84.86	86.93	88.76	88.79	91.08
RGB + RGB Optical Flow + Depth + Depth Optical Flow (4-Stream)	85.03	90.11	90.25	90.24	92.36

D.*MTUTCNN Failure Modes Effects Analysis on BVCSL3D dataset*

This section highlights the analysis of the effects on simple, medium and complex signs of the proposed CNN architectures in the Indian sign language. We also present the error analysis, which explains why the classification mode is failing to predict a sign and how the error can be reduced in complicated signs. A set of simple, medium, and complex signs from the Indian sign language is displayed in Fig.8 (a-b), fig.8(c-d, and fig.8(e-f).

The test results with different orientations for the same subject are shown as a matrix of confusion in Fig.9. The results indicate that simple and medium signs are well-known for most signs at a recognition rate of approximately 91%. The results are however distributed over close matches for complex signs. For instance we have 'Bite' and 'Balloon' signs in Fig.8 (e-f), which are 90% similar to each other. They differ in the hand clutch part, the only RGB data training or testing that can not be properly detected. The improvement in the recognition rate of our previous methods based on RGB can be attributed to the inclusion of space-related details and time maps in the CNN training. However, it has been difficult to test complicated signs with only RGB frames and other signs in the dataset to detect them poorly.

Fig. 8. Simple signs from ISL with only one hand movement (a) Art, (b) Watch; Medium signs of ISL with two non-intersecting hands (c) Doll, (d) Stage; Complex signs of ISL with two intersecting hands and with respect to other upper body parts (e) Bite, (f) Balloon.

It improves the situation using the RGB and temporal maps of RGB as input to the CoT4CNN trained. It has been stressed that CoT4CNN network consists of 4 input streams, RGB and space depth; RGB and time depth. With two input tests on the same subject, the confusion matrix in Fig9 is generated. The rates in Fig.9 show that the network does not exceed, since we did not find 100% matches with the same subject and the same sign. The threshold for the prediction is 0.9 for every test. We showed the rest of the five unseen topics as RGB space and temporal video frames in order to confirm the MTUTCNN network trained. Fig.10 shows the resulting matrix of confusion on a few signs for two topics for cross-subject testing. Most of the failures or misclassifications occurred in signs using hands with other upper body part such as head and double hands intersecting or occluding each other. The signs by subject-2 in fig.10(b) shows higher misclassification rate due to the change in video background.

The average cross-subject test recognition for the whole BVCSL3D dataset is 74.58 %. However, we have seen the recognition rate increase to 87.56% when cross-subject test data have been mixed with training data. Cross view and cross scale test on the BVCSL3D dataset is shown in fig. 11(a) and fig. 11(b) for a few signs in the dataset. For cross scales, the trained CNN performed well across all known subjects in the dataset.

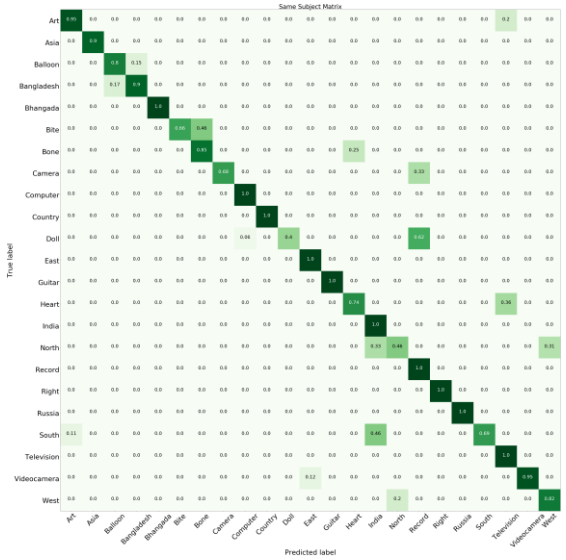
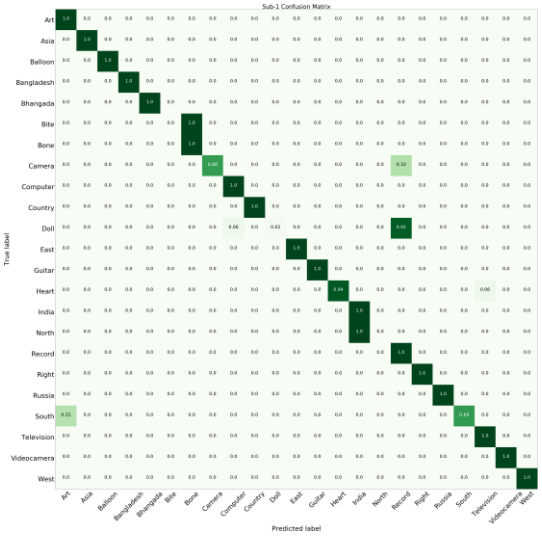
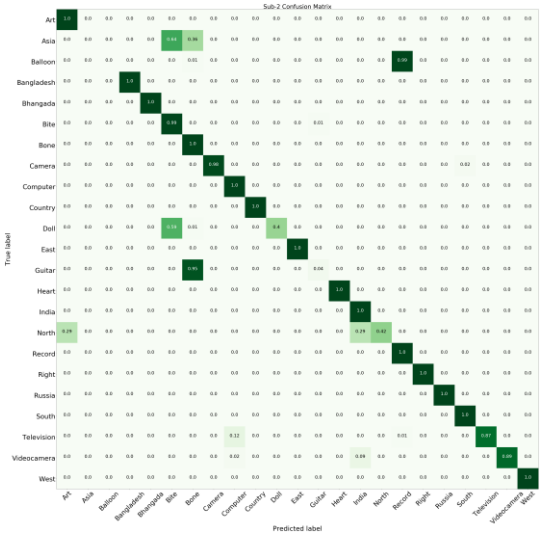


Fig. 9. Testing MTUTCNN on same subject in different orientations of hand movements.

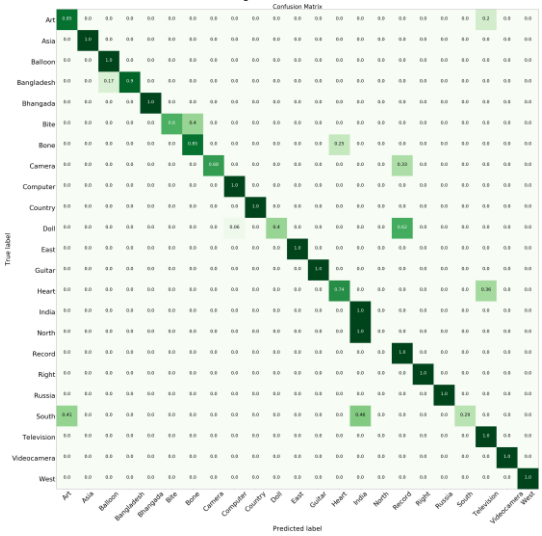


(a)

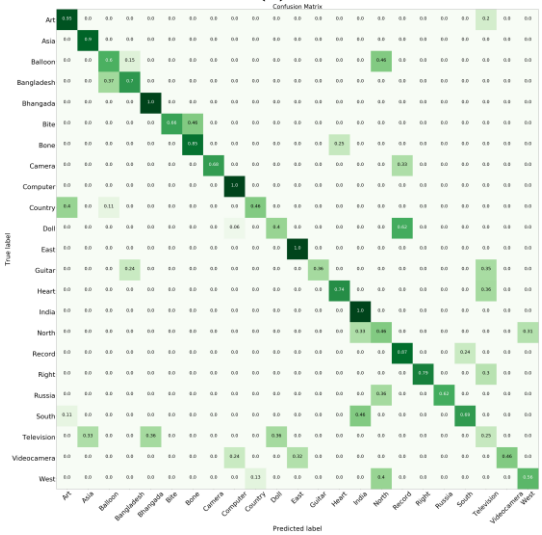


(b)

Fig. 10. Cross subject testing: a) with subjects 1 and b) with subject 2.



(a)



(b)

Fig. 11. Validation results on (a) Cross Scale and (b) Cross View data.

However, cross view testing for the same subjects resulted in mediocre recognition accuracies as shown in fig.11(b). Our analysis shows that, signs involving double hand movements and multiple body parts fail to get recognized even with an additional depth information used for training. Further, cross view and cross scale test failed on these complex signs. The recognition accuracies can be improved by considering a lower precision recall values during testing. For complex signs in cross view and cross scale we found that, a prediction threshold of 0.45 worked for all signs in the dataset giving 91.23% recognition rate. However, such lower prediction thresholds in real time will fail to classify a sign from a non-sign gestures.

We have found that multi-modal inputs from RGB+RGB Opticalflow provide better results than those from RGB during testing alone. In order to establish a general weight matrix, RGB and depth are poorly correlated with the CNN. In our CNN design, multi-modal features emulate each other during training to create the best area of interest. This CNN is trained and can be used in real time environments with RGB+RGB opticalflow streams. We show a few frames of real time tests with RGB+RGB Opticalflow in fig. 12 for the input sign 'Art' for cross-subject, scale and display data. In fig. 12 with high class scores we included some miscarriages. In terms of multiple training programs and tests, the average recognition achieved is 91.93% for MTUTCNN. The fig.13 shows an example set of MTUTCNN outputs.

Fig. 12. Cross subject, view and scale test with sign 'Art'.

Fig. 13. Example Results of random testing.

This shows that, due to the lack of real time available specialized video data sources, a trained Multi Stream Deep Model like MTUTCNN can be used for classification with an available data source.

V.CONCLUSION

During this study, the CNN architecture with multi-modal data will be trained and tested with the same model data to implement deep sign language recognition models in real-time using the RGB-D datasets. The CNN architecture is divided into RGB and depth streams, each of which is spatially and temporally divided into input streams. Both RGB and profundity streams share spatial and time characteristics in the 2nd layer to extract region of particular interest characteristics (ROISF). These ROISFs are again fused into RGB and depth streams with concatenation to produce a discriminatory feature combined with RGB and sign depth details. This CNN is used for all four streams, but only two RGB streams are tested. This design ensures that the lack of depth modal data during the testing phase allows sign language recognition in real time. In our own Indian sign language RGB–D data set, BVCSL3D and three reference RGB–D action datasets, MSRDailyActivity 3D, UT Kinect, and G3D we test our proposed Multimodal Trained Unimodal Tested CNN model. In all these data sets, the network performed better without the drop-out in one of these levels. The recognition rate on RGB + RGB Opticalflow input data is 91.93 % for the sign language data set. The results point to a new class of deep learning architectures, which with missing profound modal data can be used in real time. The MTUTCNN also improved against some state-of-the-art RGB-D recognition of sign languages based on CNN.

REFERENCES

1. Wang, Pichao, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O. Ogunbona. "Action recognition from depth maps using deep convolutional neural networks." IEEE Transactions on Human-Machine Systems 46, no. 4 (2016): 498-509.
2. Almeida, Sílvia Grasiella Moreira, Frederico Gadelha Guimarães, and Jaime Arturo Ramírez. "Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors." Expert Systems with Applications 41, no. 16 (2014): 7259-7271.
3. Li, Shao-Zi, Bin Yu, Wei Wu, Song-Zhi Su, and Rong-Rong Ji. "Feature learning based on SAE–PCA network for human gesture recognition in RGBD images." Neurocomputing 151 (2015): 565-573.
4. Sutarman and Jasni Binti Mohamad Zain, Mazlina Binti Abdul Majid, Arief Hermawan "Recognition of Malaysian Sign Language Using Skeleton Data with Neural Network" in 2015 International Conference on Science in Information Technology (ICSITech), 2015, pp. 231–236.
5. G. Yu and T. Li, "Recognition of Human Continuous Action with 3D CNN," Computer Vision Systems, pp. 314–322, 2017.
6. Z. Tu, Jun Cao, Yikang Li, and B. Li, "MSR-CNN: Applying motion salient region based descriptors for action recognition," 2016 23rd International Conference on Pattern Recognition (ICPR), Dec. 2016.
7. G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-Based CNN Features for Action Recognition," 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015.
8. X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length," IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 634–644, Mar. 2018.
9. D. Liu, Y. Wang, and J. Kato, "Evaluation of Triple-Stream Convolutional Networks for Action Recognition," 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Nov. 2017.
10. E. K. Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. T. K. Kumar, and D. A. Kumar, "Training CNNs for 3-D Sign Language Recognition With Color Texture Coded Joint Angular Displacement Maps," IEEE Signal Processing Letters, vol. 25, no. 5, pp. 645–649, May 2018.

11. Wang, Jiang, Zicheng Liu, Ying Wu, and Junsong Yuan. "Mining actionlet ensemble for action recognition with depth cameras." In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1290-1297. IEEE, 2012.
12. L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Jun. 2012.
13. V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Jun. 2012.