Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 1061 – 1076
© TÜBİTAK
doi:10.3906/elk-2005-57

Learning multiview deep features from skeletal sign language videos for recognition

Ashraf Ali SHAIK, Venkata Durga Prasad MAREEDU, Venkata Vijaya Kishore POLURIE[*]
Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, India

Received: 08.05.2020 • **Accepted/Published Online:** 11.09.2020 • **Final Version:** 30.03.2021

Abstract: The most challenging objective in machine translation of sign language has been the machine's inability to learn interoccluding finger movements during an action process. This work addresses the problem of teaching a deep learning model to recognize differently oriented skeletal data. The multi-view 2D skeletal sign language video data is obtained using 3D motion-captured system. A total of 9 signer views were used for training the proposed network and the 6 for testing and validation. In order to obtain multi-view deep features for recognition, we proposed an end-to-end trainable multistream convolutional neural network (CNN) with late feature fusion. The fused multiview features are then inputted to a two-layer dense and a decision making softmax. The proposed CNN employs numerous layers to characterize view correspondence to generate maximally discriminative features. This study is important to understand the effects of multiview data processing by CNNs for sign language recognition in decoding joint spatial information. Further, deeper perspectives were developed into multiview processing of CNNs by applying skeletal action data.

Key words: Skeletal sign language, multiview learning, deep learning, pattern recognition

1. Introduction

Sign language recognition (SLR) has been attempted rigorously in the past using a wide range of sensors that capture bodily movements in one-, two- and three-dimensional spaces [1]. Despite extensive research on all data spaces with multiple methodologies, the successful translation of these models into real time SLR systems is scarce. This is because of the uncharacteristic human hand movements in space, which cause sensors to capture ambiguous data. For example, the 2D video data [2] of finger movements is blurred during motion, which results in poor recognition. The 1D sensor data[3] from flex and gyro sensors are a great alternative to 2D videos but has failed to recover signs with respect to other body parts. Consequently, the 3D sensors such as Kinect[4], leap motion[5], and motion capture[6] have improved the SLR results in the past decade. However, only motion capture had the capability to model near perfect biomechanics of human body to recreate sign language 3D data[7].

Machine interpretation of sign language has been challenging due to its intuitive nature of human hand and finger movements. Specifically, these machine translated movements should provide meaningful arguments for a smooth conversation between two classes of humans. However, the human body biomechanics makes it difficult to capture a particular sign in its entirety. Figure 1 exemplifies the above statement by displaying video frames of some signs that require a different orientation by the algorithm to recognize correctly. Computer

[*] Correspondence: pvvkishore@kluniversity.in

vision combined with machine learning algorithms for sign language recognition have achieved good outcomes in the past decades [8–10]. Despite successful outcomes, most of the research focused on data captured using only front view, and orientations of the signer position were not considered for evaluation.

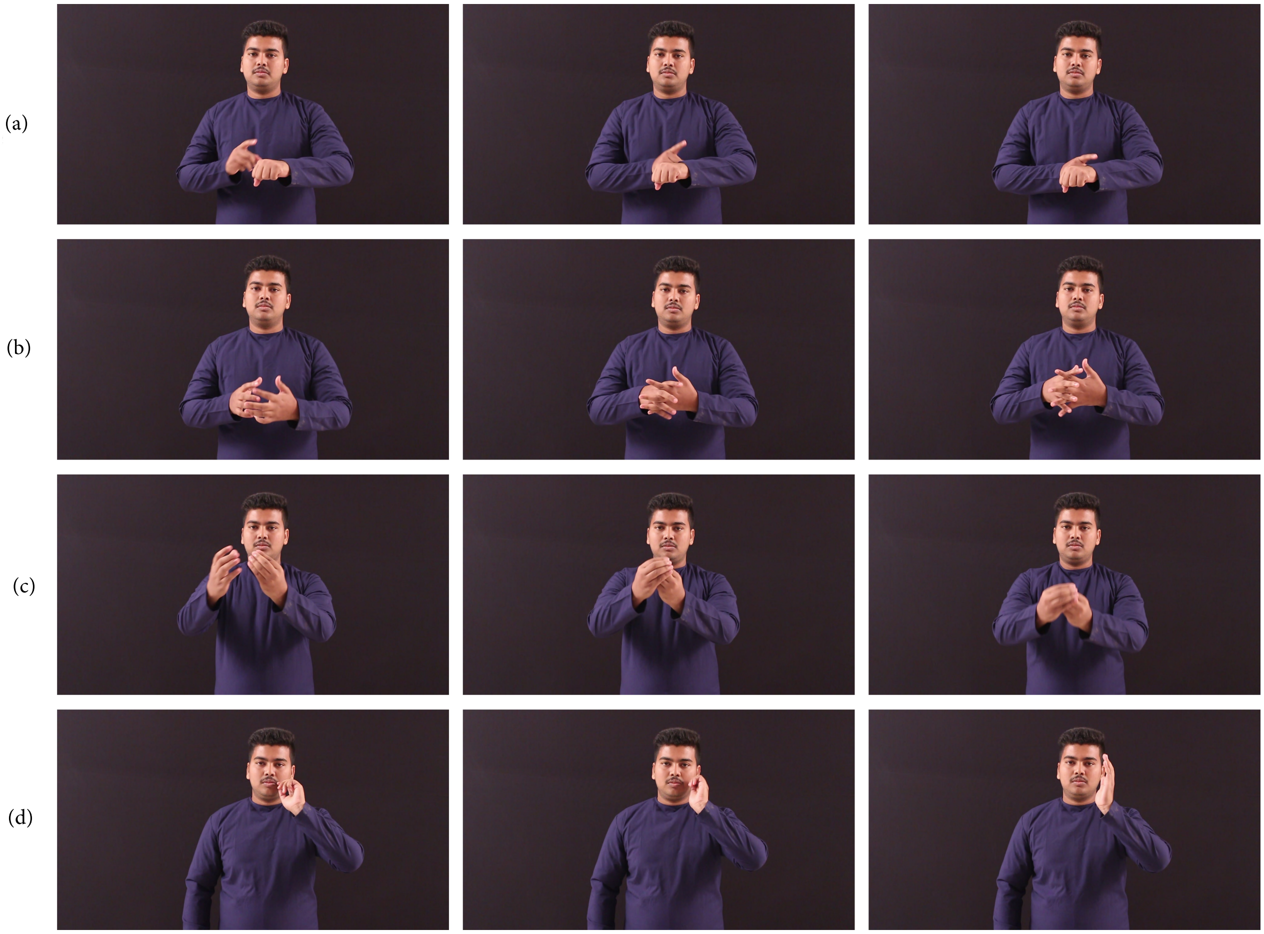


Figure 1. Camera orinetation complications during capturing of sign language for machine interpretation. Video frames of Indian sign language for the words (a) time, (b) together, (c) with, and (d) forward.

The hand and finger movements performed by humans are bounded by individuals' body biomechanics. Hence, there will always be variations in same signs performed by different signers or even in the signs done by the same signer in different times. A complete SLR requires hand and fingers to reflect visual properties of joints and their orientations as input to a machine interpreter. Past works, which focused mainly on multi-view SLR [1], gesture recognition [11], face [12], and action recognition problems [13] have brought some insight into this problem. There are very few works in the area of multiview skeletal sign language recognition to understand the view effects on the performance of the deep learning frameworks. Conventional SLR models are based on hand crafted feature extraction methods and machine learning [8]. Unfortunately, the hand-crafted features using computer vision algorithms seldom model all the attributes of a sign language in multiple views. Figure 2 presents a 9-view big camera positions from our dataset used for training and 6-view small camera position for testing. Consequently, Figure 3 shows the video frames captured from 9 views for training. It clearly shows the need for a multiview processing approach as there are selfocclusions from hands during the signing process.

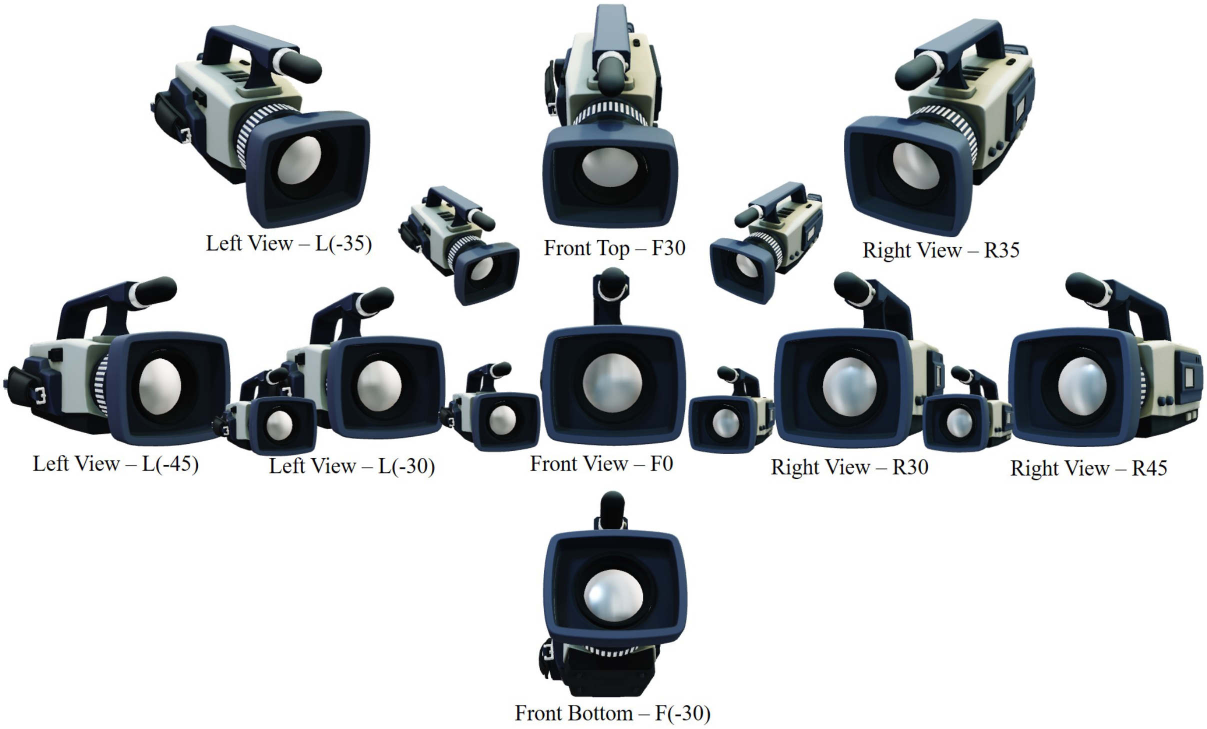


Figure 2. Camera orientations used in capturing multiple views of 3D sign language. Larger cameras represent the training views and smaller ones represent the test views

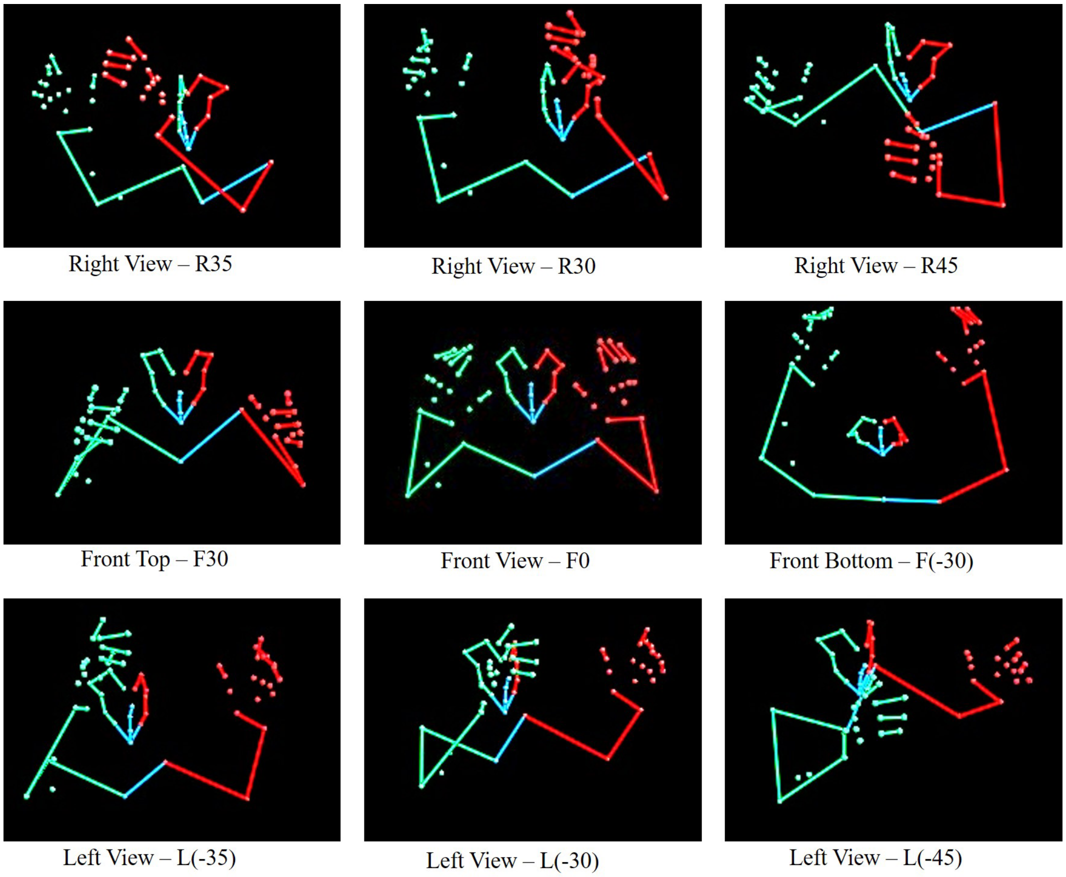
Multiview learning enhances the degrees of freedom of camera orientation in the field, and thus improves real time performance of the recognition algorithm. It is useful to explore how skeletal sign language video data can be integrated into deep learning framework. The major contributions of this work are:

1. We propose to capture a new 2D skeletal sign language video dataset with 200 signs that are recorded with 3D motion capture technology. The 2D skeletal videos can be obtained in any orientation from 0 to 360 degrees on all three axis. Our dataset is named as KLEF3DSL_2Dskeletal, which has 15 videos per sign in multiple views.

2. We propose a 9 stream CNN and a novel fractal view based feature fusion mechanism which generates a maximally discriminative multiview features for recognition. In this categorized pooling model, fractal views having maximum uncorrupted information is combined into one feature vector. We have three such fractal views that generate a high-quality noise free multiview feature vector.

3. Intensive evaluations on our $200 \times 15 \times 5 = 15000$ video skeletal sign language video datasets as well as 4 multi modal action datasets have been used in validating the proposed CNN architecture. The experimentation has shown that the proposed multiview learning through deep features and categorized pooling network achieves higher accuracies than the single view and other multimodel sign language recognition methods.

The rest of the paper is organized as follows. Section 2 serves to provide an insight into the past works; Section 3 gives the machine construction, training and testing procedures; results and related analysis on the



Right View – R35

Right View – R30

Right View – R45

Front Top – F30

Front View – F0

Front Bottom – F(-30)

Left View – L(-35)

Left View – L(-30)

Left View – L(-45)

Figure 3. Figure showing a time frame of captured training frames in 9 multiple views for the sign 'basketball'.

outcomes of the proposed method are presented in Section 4; Section 5 draws conclusions and gives future directions.

2. Related background

SLR has been investigated in the past three decades exclusively using methods in computer vision, image processing, and machine learning [8] However, very little was ever explored from multiview machine perception for SLR. The multiview perception of sign language has been possible with leap motion [10], Kinect [4], and 3D motion capture [6]. The previous works in [1] has pointed the use of SLR with multicamera model and indicated the need for such models as the hand movements cause selfocclusions which are difficult to be represented for recognition. Hence, it becomes necessary for machined translators of sign language to learn from multiple views for constructing a rich feature space for recognition in real time.

The concept of view invariance has been explored exhaustively in human action recognition models. These methods can be divided into two categories based on the type of data being used: generative models and non-generative models. In generative models, synthetic multiview data will be generated by the program for training and testing. Contrastingly, the nongenerative models use multiple sensors in different orientations to capture real multiview data for experimentation. The generative models use deep generative models that apply adversarial training to obtain information from multiple views [14]. These methods try to extract correlations between the views through GANs which has shown promising results on 3D skeletal data [15]. These methods are largely focused on 3D models, where a single image 3D model is applied to GANs to generate a multiple views for recognition using deep networks. The input to generative models is a 3D object image, which generates learned representations of the disentangled and oriented 3D output synthesized objects in the view manifold.

The generative models used geometric reasoning to derive multiple views from single view representations. These methods focused on morphed prewrapping and postwrapping the interpolation images to generate multiple views [16]. These hand-crafted methods were unsuccessful in capturing the required information. With the advent of deep learning, the geometric reasoning to obtain multiple views has been achieved through learning on a convolutional neural network (CNN)[17]. These models require two or more images for synthesis of multiple views and at times reproduced information only seen in the images. Others such methods applied CNN decoders which used implicit/explicit graphics code to generate transformed view pixels [18]. Recent models show a deep generative encoder-decoder CNN network to generate an appearance flow vector in the input view to the output [19]. All these methods generate views that are derived using some form of transformation, which works under certain constraints set by the algorithm.

Conversely, the nongenerative models acquire multiview data from differently oriented sources to learn the derived features for recognition [20]. Multi view CNNs represent features from multiple views of available data from 3D object space [21], video action recognition [22], and person reidentification [23]. A comprehensive feature representations are attained through iterative learning from different views. These CNNs models have multiple streams that feed into data from each view which generates a view specific features. These features are fused together to form a discriminative mixed feature vector that represent all views. Finally, the flattened multiview features are processed by a fully connected neural network for classification[24]. Despite their higher accuracies, there are several shortcoming in the above model.

The biggest limitation is in the multiview data acquisition. The network model is efficient for views below 10 for 3D object recognition [25]. Similarly, if the depth of the network is not hyperparameterized accordingly, it becomes computationally inefficient. Although the above two are iteratively manageable, the recognition accuracy depends on the capabilities of the fusion network. Mostly, the previous multi view CNN algorithms have differed in the fusion network. View pooling is the commonly used fusion in multi view recognition [24]. Enhancements to view polling are group pooling [20] and intraview group pooling [26] that has shown considerable accuracy in 3D object recognition. However, most of the models approached the problem of multi view 3D object recognition, which resulted in improved recognition over single view models [27].

The concept of multiview deep learning is being extended to human action recognition with different modal inputs from RGB video [22], RGB D Kinect data [28], and 3D skeletal data [29, 30]. Other than deep models, random forests [31], fuzzy distances [32], linear discriminant analysis [33], and bag of key poses [34] were some of the popular methods that were applied with multiple variations across human action recognition. Non deep learning models extract features manually from each view and are ensembled using a fusion model to be classified by a machine learning algorithm. The success of the above models has inspired researches to apply these models for gesture recognition [35]. It uses 3D hand gesture cloud in multiple views for recognition. Similarly, skeletal action data in multiple views were used extensively for a large group of action recognition models due its complexity in terms of data representation in multiple views [36].

Inspired by the preceding discussion, we propose to apply the multiview deep model for skeletal sign language recognition. We propose a 4-layer CNN with 9 multi view streams whose features are pooled categorically to form a multivariate discriminating feature vector. Our pooling network is called categorized pooling network (CPN), which pools features based on views to generate a nonoverlapping features to be processed by the next CNN layers. The following section describes the proposed network architecture with training and testing procedures.

3. Multiview CNN for sign language

This section presents the architecture of the proposed network for multiview skeletal sign language recognition from 2D video data. The section begins with problem statement, which is followed by the proposed architecture, training, and testing procedures.

3.1. Problem statement

Here, we define multiview skeletal signs as V_x^{γ}, which represents 2D skeletal video data. The parameter γ denotes the view orientation and takes values in the range of 1 to 9, indicating the viewing class. Similarly, γ is ranged between 1 and 6 for testing data. The orientation of cameras was shown in Figure 2. The trained CNN with learned weights w is represented as C_w. For the input multi view data $V_x^{\gamma(test)}$ on the learned CNN, C_w is given as

$$Y_s = C_w.V_x^{\gamma(test)} \tag{1}$$

Where Y_s is the softmax output showing probabilistic class representations. Previous approaches on multiview CNN applied 2D skeletal and RGB video data of human actions to train the model. However, this is the first time a 2D skeletal sign language video data is being used for multi-view recognition. The 2D skeletal videos are generated for 3D sign language data captured with 3D motion capture technology [6, 7]. The 3D sign language data is recorded with an 8 camera Vicon motion capture system at Biomechanics and Vision Computing Research Centre, KLEF University. Thereupon, the 3D data in 15 viewing orientations are scaled down to 2D skeletal videos. Hence, each sign class will have 15 orientations; 9 of them will be used for training and 6 for testing. The raw 2D skeletal videos are used for assembling the multiview CNN from input video data x as

$$V_x^{\gamma} = \left[v_x^1, v_x^2, v_x^3, ..., v_x^{\gamma}\right] \tag{2}$$

Subsequently, assembled views are used for training the multiview CNN architecture C_{w_v} to generate sign language recognition system as

$$Y_{class} = C_{w_v} \cdot \left[v_x^1, v_x^2, v_x^3, ..., v_x^{\gamma}\right] \tag{3}$$

The following subsection details the proposed multiview CNN architecture.

3.2. Multiview CNN architecture

This work proposes a 9-stream shallow CNN architecture for decoding the information embedded in the multiple views of skeletal sign language 2D video data. The architecture is based on our previous works on 3D sign language [6]. Figure 4 presents the multiview CNN framework proposed in this work. Since the pixel values are similar in all the video data, the batch normalization and ReLU activations were discarded. It is also found through experimentation that there is no need for batch normalization between intermediate layers for stabilizing the leaning process and, therefore, reducing the training epochs. There are 4 convolutional layers operated with 3×3 filter structures and a 2×2 pooling with stride 1 after each convolutional layer. There are no dropouts in the final layers of the proposed network.

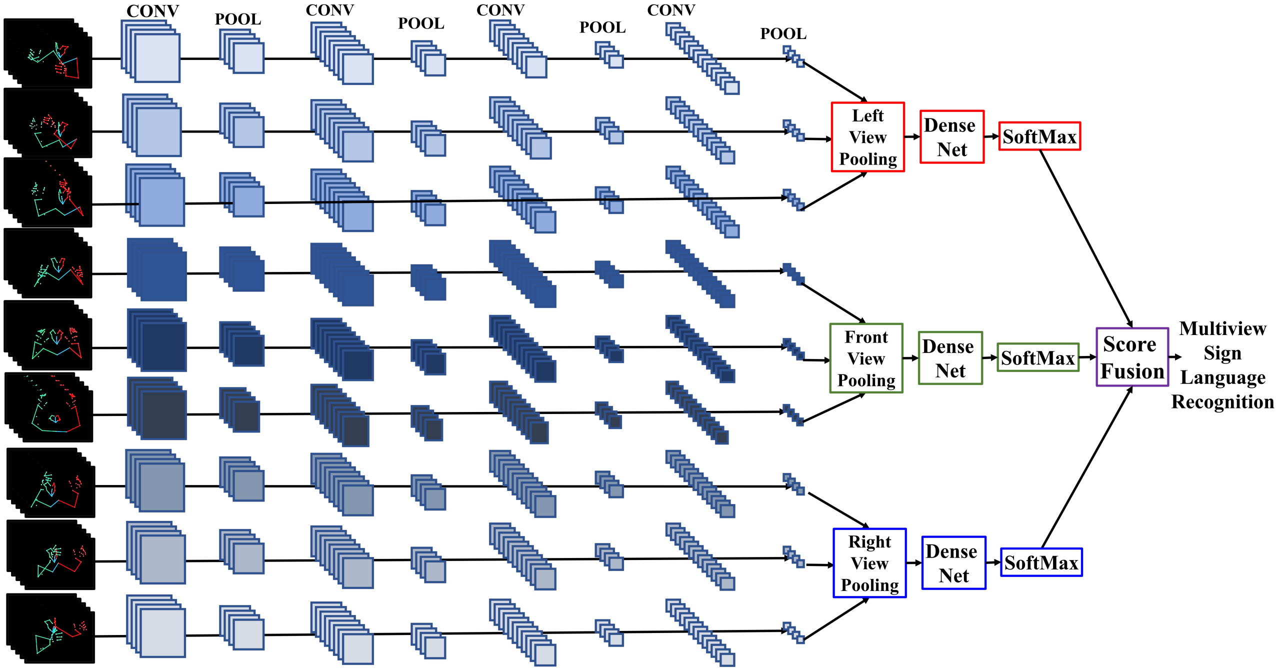


Figure 4. Proposed multi stream CNN architecture for multi view 2D skeletal sign language recognition.

In the previous works, similar architectures were designed with feature fusion at the last convolutional layers or in the early layers. The fusion in last layers is called late fusion and in early layers is called early fusion. The end of the network fusion is termed as score or decision level fusion. In contrast to the previous works in [20–23], we propose categorized pooling network (CPN), which pools features using view profiling. The view profiling is performed with three views: Left, Right, and Front. The top and bottom views also fall in these categories. The pooling network performs a location wise intersection $f^{(\bigcap)}$, which generates a unique set of features in one orientation given as:

$$H_p^R = f^{(\bigcap)}\left(v_x^{R35}, v_x^{R45}, v_x^{R30}\right) \tag{4}$$

$$H_p^L = f^{(\bigcap)}\left(v_x^{L(-35)}, v_x^{L(-45)}, v_x^{L(-30)}\right) \tag{5}$$

$$H_p^F = f^{(\bigcap)}\left(v_x^{F0}, v_x^{F30}, v_x^{F(-30)}\right) \tag{6}$$

The orientations in the above equations can be understood using the Figure 3. Following the CPN network are the two-layer fully-connected dense and softmax layers. The output decision level forging from three views is obtained for recognition of a particular sign as:

$$R_s = f_{use}\left(p_L, p_R, p_F\right), \tag{7}$$

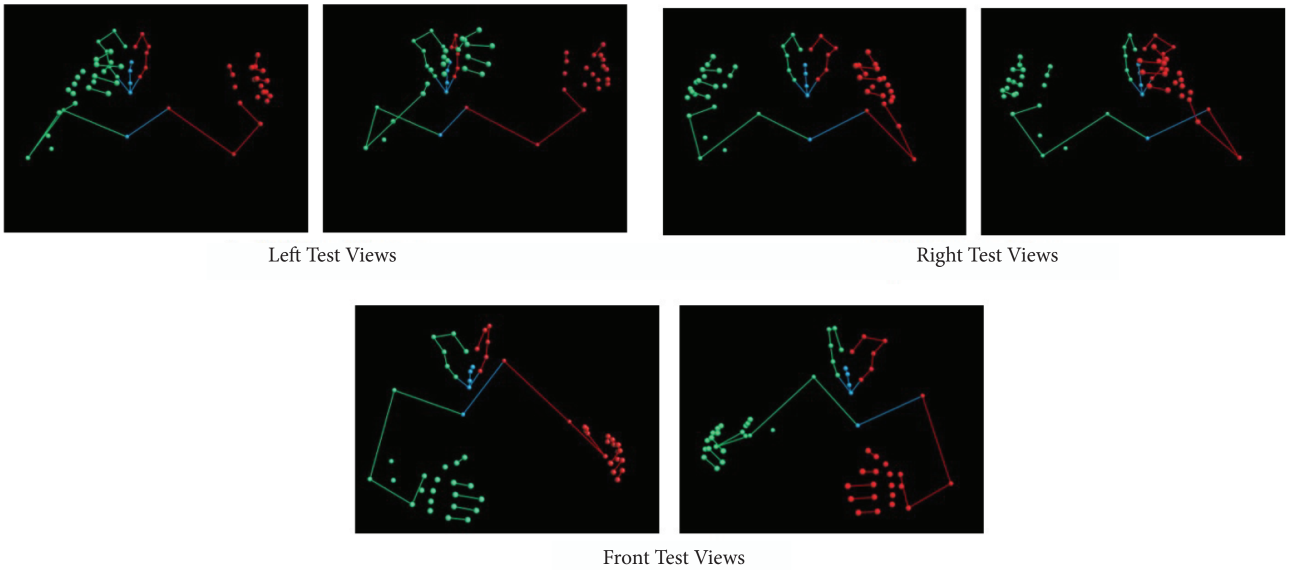
where, R_s is the final fused score for recognition. The f_{use} is the fusion rule and (p_L, p_R, p_F) are decision level probabilities of softmax layers in left, right, and front views. The following subsection presents a details the training the proposed architecture.

3.3. Training

The proposed multiview CNN is designed and trained in python 3.6 using keras as the frontend and tensorflow2.0 as the backend. The hyperparameters such as batch size, initializations, and video frame sizes were kept constant across training on all datasets. However, the learning rate was found initially across a given dataset and was decremented by a factor of 0.01, whenever the validation error became constant for four consecutive epochs. At the start of training, the weights and biases were initialized randomly with a zero mean and 0.01 variance gaussian distribution model. The proposed CNN structure learned multiview video frames by updating the weights and biases using backpropagation gradient descent algorithm. A batch size of 16 was iteratively identified based on image resolution and GPU memory. The number of classes trained were 200. For each sign class, a set of 27 videos were used for training and 18 were used for validation. The total number of 9 (views) × 4 (subjects) × 96 (frames) × 200 (classes) = 691,200 frames trained the entire network and 172800 for validation. The network is trained multiple times during which it reached an average validation error of 2.74% at 139 epochs. During training, the learning rate was decreased twice from 0.01 to 0.001 at 43 epochs and 0.001 to 0.005 at 88 epochs. There was no fine tuning applied to improve the validation error. Apart from the proposed data KLEF3DSL_2Dskeletal, we also validated the networks performance on other benchmark multiview datasets such as NTU RGB+D [37], SBU kinect interaction [38], KLYoga3D [39], and MMA [40]. From each dataset, 40 action classes were used in the orientations as shown in Figure 2 by rotating the skeletons and extracting the video actions. The number of frames across all networks are kept constant at 9 × 4 × 96 × 40 = 13,8240, for training. The validation set has 34560 frames distributed across 40 classes. The NTU RGB+D dataset trained network reached an average validation error of 3.45% at 98 epochs. Similarly, KLYoga3D and MMA attained average validation error of 2.95% at 105 epochs and 3.18% at 109 epochs, respectively. Apart from the proposed CNN, there are other multistream models that were experimented with using the same datasets from [20, 22, 24, 26, 27, 30]. All these models were redesigned to accommodate the skeleton video datasets used in this work are trained from scratch with the same hyperparameters and variable parameters. The results of this experimentation will be presented in section 4. The next subsection shows the testing procedures followed for proposed CNN model and the other similar networks.

3.4. Testing

After training on the KLEF3DSL_2D skeletal dataset, the testing is conducted on a 6- view test dataset. The test data is of different orientation from the training set. The capture angles of test data are represented with small cameras in Figure 4. Figure 5 shows the test data frames. A total of 6×5×96×200 = 576000 frames are utilized for testing the trained CNN. Similarly, the test data for other skeletal action datasets accounts to 6×2×96×40=46080 frames. Two parameters, mean recognition accuracy (mRA) and mean F1-score (mF1s) are computed to measure the performance of the proposed multi view CNN and the other models. The next section discusses the results obtained with analysis to generate critical insights into the dynamics of multi view sign language recognition.



Left Test Views

Right Test Views

Front Test Views

Figure 5. Figure showing a time frame of testing video sequences in 6 multiple views for the sign 'basketall'.

4. Results and analysis

Experiments were conducted to understand the influence of multi view learning with respect to skeletal sign language video dataset. The skeletal sign video dataset is most challenging one compared to other 2D video and 3D sign datasets. First, the performance of the proposed CNN is analysed for single view and multiview testing. Second, we validate the proposed categorized pooling against other forms of pooling approaches to prove its usefulness towards multiview recognition. Third, we compare the proposed network with other state-of-the-art methods on our skeletal sign language dataset. Finally, a through comparison on benchmark skeletal action datasets is performed to ensure its competence against the state-of-the-art action recognition methods.

4.1. Evaluating single and multiview test samples

In this subsection, we evaluate our model's performance on single and multiview test videos. The test set is comprising of 6 views. First, we test the trained multiview CNN using one single view video per class at a time. For each view, we compute the performance measures mRA and mF1S. During this test, all the 9 streams are shown the same view. None of the test views are not part of the training set. The results of this single view testing are presented in Table 1. However, these test views come from the subjects used for training. Hence, the 1st evaluation is from a test dataset with single view same subject. The second part of Table 1 shows performance measures from single view different subject. Subsequent comparisons on Table 1 show that the proposed CNN architecture is capable of recognizing signs in views that are not seen earlier by the network. Despite good scores, the cross-subject test inputs performed slightly lesser than the same subject testing.

Second, the proposed CNN was offered to test multiple views concurrently in both same and cross subject modes. The results of this experiment are shown in Table 2. In contrast to previous single view test, the multiview has performed exceedingly well. In multiview test, the left view is presented to the CNN streams that were trained in left and this protocol was followed for remaining views also. The results show that the features learned by the individual streams close to the one particular view have higher accuracies than the

Table 1. Performance of the our trained multiview CNN with only one test view as input.

Views	Same Subject Cross View						Cross Subject Cross View					
Test Views/Scores	Ltv1	Ltv2	Rtv1	Rtv2	Ftv1	Ftv2	Ltv1	Ltv1	Rtv1	Rtv2	Ftv1	Ftv2
mRA	0.6819	0.6639	0.6471	0.6837	0.7182	0.7328	0.6472	0.6236	0.6745	0.6454	0.6524	0.6412
mF1S	0.6124	0.6185	0.6218	0.6185	0.6845	0.6842	0.6219	0.6188	0.6249	0.6148	0.6179	0.6182

others. For example, the left view test video has given maximum softmax decision score on the streams that have learned left view data and vice versa. The mRA and mF1S in Tables 1 and 2 show the same. However, the cross-subject test has slightly lower accuracies than the same subject test inputs. Overall, the results show that the proposed multiview CNN has the ability to learn multiple views, simultaneously. The following Figures 6a and 6b show the confusion matrices of two views Ltv1 and Ftv1 in cross subject cross view testing.

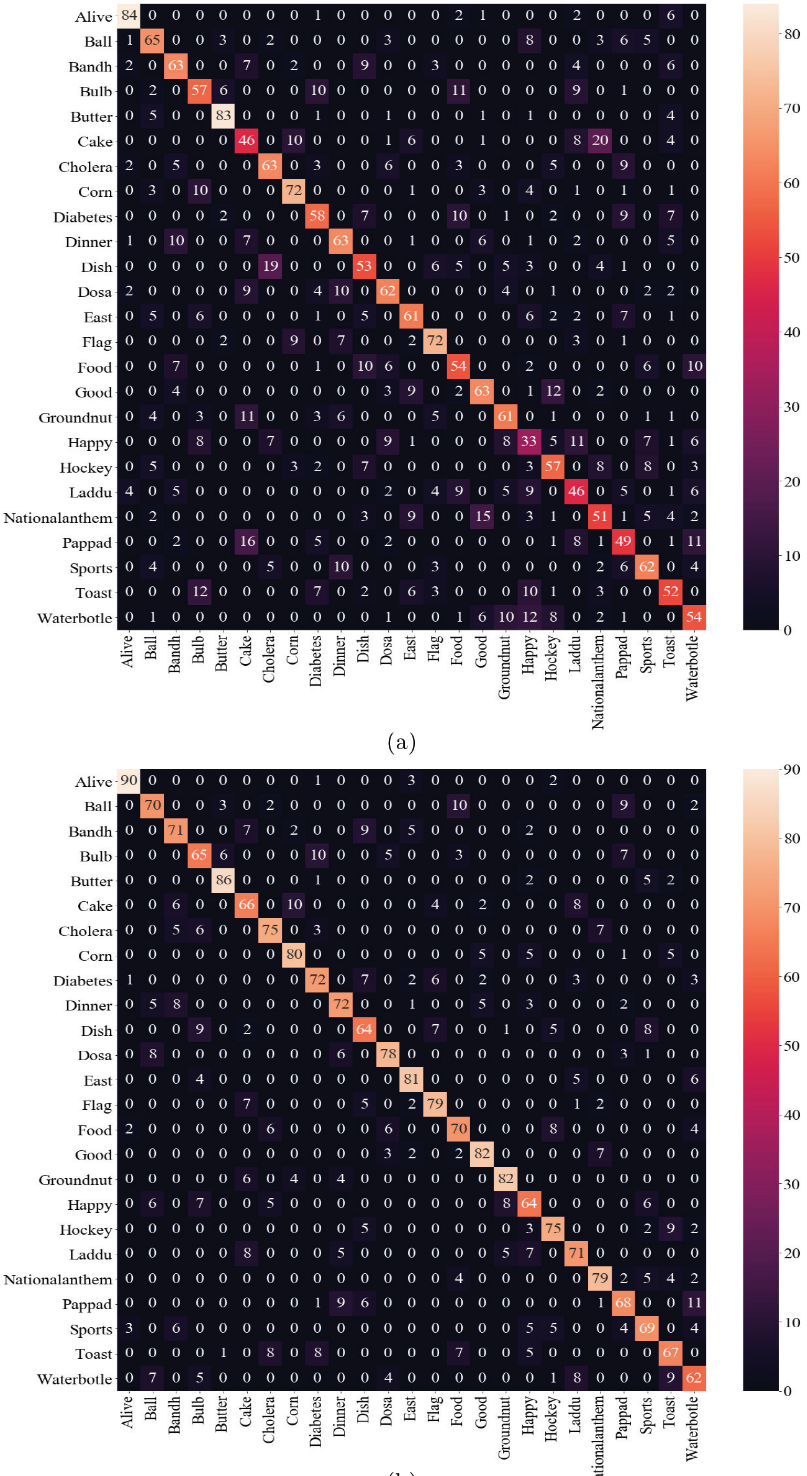
Table 2. Performance of the our trained Multiview CNN with two test views as input.

Views	Same Subject Cross View						Cross Subject Cross View					
Test Views/Scores	Ltv1– Ltv2	Rtv1– Rtv2	Ftv1– Ftv2	Ltv1– Rtv1– Ftv1	Ltv2– Rtv2- Ftv2	All Six Views	Ltv1– Ltv2	Rtv1– Rtv2	Ftv1– Ftv2	Ltv1– Rtv1– Ftv1	Ltv2– Rtv2- Ftv2	All Six Views
mRA	0.7519	0.7439	0.7721	0.7798	0.7682	0.8197	0.7309	0.7324	0.7273	0.7511	0.7497	0.7805
mF1S	0.7244	0.7102	0.7374	0.7485	0.7419	0.7859	0.6912	0.7043	0.6972	0.7108	0.7075	0.7204

4.2. Evaluation of categorized polling network

This subsection presents the effectiveness of the proposed categorized pooling network against similar pooling techniques. The other pooling networks used by our model before arriving at the proposed network are max pooling, average pooling, grouped pooling, and concatenation. Pooling is an operation, where element-wise operations are performed on the multiple feature tensors. The proposed categorized pooling involves a view defined multistage local pooling approach in which the operation is performed categorically on a particular view. This generates a feature vector that is maximally distant from globally and minimally close locally in the vicinity of particular view. Consequently, this process has guaranteed a highly discriminant feature space for closely mapped signs. The challenge in sign language recognition is identifying closely matched signs, and, to our astonishment, most of the signs have very similar hand movement. Figure 6 presents the confusion matrix to show the similarity between signs. In single-view sign language recognition systems, these similarities are augmented considerably due to feature matching. In addition to our dataset, we also tested the proposed multiview CNN on the benchmark datasets for validating the proposed CPN against other pooling forms. To maintain uniformity across datasets, only 40 classes in all datasets were used for testing. The results of our testing are presented in Table 3. In grouped pooling, a set of views are concatenated as it performs better than the concatenated as shown in Table 3. The categorized pooling has outperformed the others due to its ability to select features piece wise based on viewing orientation.

The decision level fusion in all the experiments was max fusion model, which by far proved to be the best across different architectures. In the next subsection, we train similar multi view state-of -the-art models on our dataset to evaluate the performance of our proposed CNN architecture. However, we test these models with all six-view sign language skeletal videos.



(a)

(b)

Figure 6. Confusion matrices for 25 signs in cross subject cross view testing on (a) Left view 1(Ltv1) and (b) Front view 1(Ftv1).

Table 3. Performance of CPN against other feature pooling methods on different multi view datasets trained and tested with our proposed CNN.

Datasets	Pooling	max	Average	Grouped	Concatenated	Categorized (ours)
NTU RGB+D	mRA	0.5896	0.5125	0.6258	0.5485	0.6895
	mF1S	0.5512	0.4856	0.5798	0.5245	0.6485
SBU Kinect Interaction	mRA	0.5248	0.4881	0.5592	0.5149	0.6254
	mF1S	0.4972	0.4258	0.4258	0.4673	0.5863
KLYoga3D	mRA	0.5421	0.5103	0.5723	0.5358	0.6211
	mF1S	0.5213	0.4992	0.5296	0.5013	0.6052
MMA	mRA	0.5845	0.5196	0.6189	0.5473	0.7125
	mF1S	0.5698	0.4923	0.5994	0.5291	0.6951
KLEF3DSL_2Dskeletal	mRA	0.5853	0.5099	0.6196	0.5488	0.7519
	mF1S	0.5522	0.4765	0.5841	0.5123	0.7244

4.3. Comparison with other multiview CNN models

This section highlights the advantages of the proposed CNN model adjacent to other state -of -the -art multiview deep learning models. We designed the programmes as proposed in [20, 22, 24, 26, 27, 30]. All are trained from scratch using our dataset and tested with our test set on the same machine. Table 4 gives the performance parameters mRA and mF1S.

The proposed network decodes each view into corresponding feature maps that are view specific. In most of the previous works, the views are pooled randomly without any underlying phenomenon. In this work, we propose to pool features that are view neighbouring features such as right sided, left sided and centre with top and bottom views. This proposed network computes scores that are specific to a particular set of views and tags them with the remaining views to generate a more comprehensive recognition scores on unseen views in the test set. This is done to ensure the system operated on the unknown views when introduced for real time operations. The Table 4 presents mRAs that are computed with one set of views as input during testing. For all the training views that are presented to the network as test views, the mRA was recorded around 93.43%. However, we restrained from testing using all the training views, and therefore, used a test set that was having different views unseen by the network. This is the reason behind the slightly less mRAs when compared to other scores reported in the previous works

The experiments on our skeletal sign language dataset reveal the complexity of sign language. The toughness lies in the joint representations which occupy around 10 pixels in a video frame of 256×256. Hence convolutional layer filter sizes should be minimum for producing optimal features. Therefore, instead of working the original filter sizes specified by the previous works, we used for all networks a filter size of 3×3. The performance of the proposed model is higher than the previous models due to the presence of CPN, which selectively manufactured a high discriminating feature vector for recognition. The large quantity of misclassifications occur when the hands and face overlap making it difficult for the machine to judge its inference for sign recognition. Categorized pooling takes the best features from a set of similar looking views to build a feature vector that best represents all those views. However, a lot of fine tuning has be done to further improve the recognition of skeletal sign language. The following subsection gives a more generalized comparison of the proposed and other multiview CNNs on benchmark action datasets.

Table 4. Shows comparison of our proposed multi view CNN with previous works on the KLEF3DSL_2D skeletal dataset.

Multiview CNN models	Views	Same subject cross view						Cross subject cross view					
	Test views/ Scores	Ltv1	Ltv2	Rtv1	Rtv2	Ftv1	Ftv2	Ltv1	Ltv1	Rtv1	Rtv2	Ftv1	Ftv2
[27]	mRA	0.6604	0.6424	0.6256	0.6622	0.6967	0.7113	0.6257	0.6021	0.653	0.6239	0.6309	0.6197
	mF1S	0.5909	0.597	0.6003	0.597	0.663	0.6627	0.6004	0.5973	0.6034	0.5933	0.5964	0.5967
[28]	mRA	0.6694	0.6514	0.6346	0.6712	0.7057	0.7203	0.6347	0.6111	0.662	0.6329	0.6399	0.6287
	mF1S	0.5999	0.606	0.6093	0.606	0.672	0.6717	0.6094	0.6063	0.6124	0.6023	0.6054	0.6057
[29]	mRA	0.6394	0.6214	0.6046	0.6412	0.6757	0.6903	0.6047	0.5811	0.632	0.6029	0.6099	0.5987
	mF1S	0.5699	0.576	0.5793	0.576	0.642	0.6417	0.5794	0.5763	0.5824	0.5723	0.5754	0.5757
[30]	mRA	0.6677	0.6497	0.6329	0.6695	0.704	0.7186	0.633	0.6094	0.6603	0.6312	0.6382	0.627
	mF1S	0.5982	0.6043	0.6076	0.6043	0.6703	0.67	0.6077	0.6046	0.6107	0.6006	0.6037	0.604
[31]	mRA	0.6418	0.6238	0.607	0.6436	0.6781	0.6927	0.6071	0.5835	0.6344	0.6053	0.6123	0.6011
	mF1S	0.5723	0.5784	0.5817	0.5784	0.6444	0.6441	0.5818	0.5787	0.5848	0.5747	0.5778	0.5781
[34]	mRA	0.6765	0.6585	0.6417	0.6783	0.7128	0.7274	0.6418	0.6182	0.6691	0.64	0.647	0.6358
	mF1S	0.607	0.6131	0.6164	0.6131	0.6791	0.6788	0.6165	0.6134	0.6195	0.6094	0.6125	0.6128
Ours	mRA	0.6819	0.6639	0.6471	0.6837	0.7182	0.7328	0.6472	0.6236	0.6745	0.6454	0.6524	0.6412
	mF1S	0.6124	0.6185	0.6218	0.6185	0.6845	0.6842	0.6219	0.6188	0.6249	0.6148	0.6179	0.6182

4.4. Comparing multiview models on skeletal action data

An extensive comparison is required to decide the performance of the proposed CNN against state-of-the-art multiview CNNs from literature. All the methods from previous works are designed for action recognition. The skeleton in action sequences have 24 joints as compared to 57 joints in our sign language dataset. Hence, the models for action recognition are computationally less rigorous than the sign language. The performance parameter mRA is obtained using various algorithms on the skeletal action datasets in Table 5. Accordingly, the number of epochs for action recognition are down by 36%. The multiview skeletal data for training and testing have been created by rotating the skeletal joints in the directions of the cameras as shown in Figure 2.

The proposed model has shown better performance on action datasets captured under various conditions when compared to other multiview CNN methods. The higher performance of the proposed multiview CNN is due to the polling network, which categorically pools the features based on the view profiling. This process created a more robust feature vector the preserved common features related to a view along with highly discriminant ones that are unique to a particular orientation. Finally, we compare the proposed multi view CNN based on computational complexity and ablation studies.

5. Conclusions

This work enriches the importance of multiview sign language recognition on skeletal data and provides new insights on categorized pooling of features for improved learning. The categorized pooling provides a set of uncommon features that are oriented in a particular direction and are uniquely represented in that orientation. This has provided a rich feature vector for 2D skeletal sign language video data in 9 different training views. Six different test views are used to evaluate the proposed multiview CNN model with categorized pooling network. The results have shown a network with better performance on 2D multiview sign language skeletal video dataset as well as four benchmark action datasets.

Table 5. Shows recogntion accuracies of signle view test on our proposed and previous works across benchmark datasets.

		Same Subject Cross View						Cross Subject Cross View					
		Ltv1	Ltv2	Rtv1	Rtv2	Ftv1	Ftv2	Ltv1	Ltv1	Rtv1	Rtv2	Ftv1	Ftv2
NTU RGB+D	[27]	0.6504	0.6324	0.6156	0.6522	0.6867	0.7013	0.6157	0.5921	0.643	0.6139	0.6209	0.6097
	[28]	0.6594	0.6414	0.6246	0.6612	0.6957	0.7103	0.6247	0.6011	0.652	0.6229	0.6299	0.6187
	[29]	0.6294	0.6114	0.5946	0.6312	0.6657	0.6803	0.5947	0.5711	0.622	0.5929	0.5999	0.5887
	[30]	0.6577	0.6397	0.6229	0.6595	0.694	0.7086	0.623	0.5994	0.6503	0.6212	0.6282	0.617
	[31]	0.6318	0.6138	0.597	0.6336	0.6681	0.6827	0.5971	0.5735	0.6244	0.5953	0.6023	0.5911
	[34]	0.6665	0.6485	0.6317	0.6683	0.7028	0.7174	0.6318	0.6082	0.6591	0.63	0.637	0.6258
	ours	0.6719	0.6539	0.6371	0.6737	0.7082	0.7228	0.6372	0.6136	0.6645	0.6354	0.6424	0.6312
SBU Kinect Interaction	[27]	0.6704	0.6524	0.6356	0.6722	0.7067	0.7213	0.6357	0.6121	0.663	0.6339	0.6409	0.6297
	[28]	0.6794	0.6614	0.6446	0.6812	0.7157	0.7303	0.6447	0.6211	0.672	0.6429	0.6499	0.6387
	[29]	0.6494	0.6314	0.6146	0.6512	0.6857	0.7003	0.6147	0.5911	0.642	0.6129	0.6199	0.6087
	[30]	0.6777	0.6597	0.6429	0.6795	0.714	0.7286	0.643	0.6194	0.6703	0.6412	0.6482	0.637
	[31]	0.6518	0.6338	0.617	0.6536	0.6881	0.7027	0.6171	0.5935	0.6444	0.6153	0.6223	0.6111
	[34]	0.6865	0.6685	0.6517	0.6883	0.7228	0.7374	0.6518	0.6282	0.6791	0.65	0.657	0.6458
	Ours	0.6919	0.6739	0.6571	0.6937	0.7282	0.7428	0.6572	0.6336	0.6845	0.6554	0.6624	0.6512
KLYoga3D	[27]	0.6804	0.6624	0.6456	0.6822	0.7167	0.7313	0.6457	0.6221	0.673	0.6439	0.6509	0.6397
	[28]	0.6894	0.6714	0.6546	0.6912	0.7257	0.7403	0.6547	0.6311	0.682	0.6529	0.6599	0.6487
	[29]	0.6594	0.6414	0.6246	0.6612	0.6957	0.7103	0.6247	0.6011	0.652	0.6229	0.6299	0.6187
	[30]	0.6877	0.6697	0.6529	0.6895	0.724	0.7386	0.653	0.6294	0.6803	0.6512	0.6582	0.647
	[31]	0.6618	0.6438	0.627	0.6636	0.6981	0.7127	0.6271	0.6035	0.6544	0.6253	0.6323	0.6211
	[34]	0.6965	0.6785	0.6617	0.6983	0.7328	0.7474	0.6618	0.6382	0.6891	0.66	0.667	0.6558
	Ours	0.7019	0.6839	0.6671	0.7037	0.7382	0.7528	0.6672	0.6436	0.6945	0.6654	0.6724	0.6612
MMA	[27]	0.6404	0.6224	0.6056	0.6422	0.6767	0.6913	0.6057	0.5821	0.633	0.6039	0.6109	0.5997
	[28]	0.6494	0.6314	0.6146	0.6512	0.6857	0.7003	0.6147	0.5911	0.642	0.6129	0.6199	0.6087
	[29]	0.6194	0.6014	0.5846	0.6212	0.6557	0.6703	0.5847	0.5611	0.612	0.5829	0.5899	0.5787
	[30]	0.6477	0.6297	0.6129	0.6495	0.684	0.6986	0.613	0.5894	0.6403	0.6112	0.6182	0.607
	[31]	0.6218	0.6038	0.587	0.6236	0.6581	0.6727	0.5871	0.5635	0.6144	0.5853	0.5923	0.5811
	[34]	0.6565	0.6385	0.6217	0.6583	0.6928	0.7074	0.6218	0.5982	0.6491	0.62	0.627	0.6158
	Ours	0.6619	0.6439	0.6271	0.6637	0.6982	0.7128	0.6272	0.6036	0.6545	0.6254	0.6324	0.6212

References

[1] Kishore PVV, Prasad MVD, PCR, Rahul R. 4-Camera model for sign language recognition using elliptical fourier descriptors and ANN. In: 2015 International Conference on Signal Processing and Communication Engineering Systems; Guntur, India; 2015. pp. 34-38.

[2] Kishore PVV, Sastry ASCS, Kartheek A. Visual-verbal machine interpreter for sign language recognition under versatile video backgrounds. In: 2014 First International Conference on Networks & Soft Computing (ICNSC 2014); Guntur, India; 2014. pp. 135-140.

[3] Oz C, Leu MC. American Sign Language word recognition with a sensory glove using artificial neural networks. Engineering Applications of Artificial Intelligence 2011; 24(7):1204-1213.

[4] Ravi S, Suman M, Kishore PVV, E KK, M TKK, D AK. Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–D based sign language gesture recognition. Journal of Computer Languages 2019; 52: 88-102.

[5] Mittal A, Kumar P, Roy PP, Balasubramanian R, Chaudhuri BB. A modified LSTM Model for continuous sign language recognition using leap motion. IEEE Sensors Journal 2019; 19(16): 7056-7063.

[6] Kumar EK, Kishore PVV, Sastry ASCS, Kumar MTK, Kumar DA. Training CNNs for 3-D Sign language recognition with color texture coded joint angular displacement maps. IEEE Signal Processing Letters 2018; 25(5): 645-649.

[7] Kishore PVV, Kumar DA, Sastry ASCS, Kumar EK. Motionlets Matching With Adaptive Kernels for 3-D Indian Sign Language Recognition. IEEE Sensors Journal 2018; 18(8): 3327-3337.

[8] Cheok M, Jin Z, Omar MH, Jaward . A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics 2019; 10(1): 131-153.

[9] Kumar P, Gauba H, Roy PP, Dogra DP. A multimodal framework for sensor based sign language recognition. Neurocomputing 2017; 259: 21-38.

[10] Mohandes M, Deriche M, Liu J. Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition. IEEE Transactions on Human-Machine Systems 2014; 44(4): 551-557.

[11] Wei W, Wong Y, Du Y, Hu Y, Kankanhalli M, Geng W. A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface. Pattern Recognition Letters 2019; 119: 131-138.

[12] Farfade S, Sudhakar MJ, Saberian LJ, Li . Multi-view face detection using deep convolutional neural networks. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR 2015); Shanghai China; 2015; 1: 643-650.

[13] Gao Z, Zhang H, Xu GP, Xue YB, Hauptmann AG. Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. Signal Processing 2015; 112: 83-97.

[14] Wang L, Ding Z, Tao Z, Liu Y, Fu Y. Generative multi-view human action recognition. Proceedings of the IEEE International Conference on Computer Vision. 2019; Seoul, South Korea; 1: 6212-6221.

[15] Cui J, Li S, Xia Q, Hao A, Qin H. Learning multi-view manifold for single image based modeling. Computers & Graphics. 2019; 82: 275-285.

[16] Chaurasia G, Sorkine O, Drettakis G. Silhouette-Aware Warping for Image-Based Rendering. Computer Graphics Forum. 2011; 30(4): 1223-1232.

[17] Flynn J, Neulander I, Philbin J, Snavely N, Deepstereo . Learning to predict new views from the world's imagery. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; Las Vegas, NV, USA; 1: 5515-5539.

[18] Zhou T, Tulsiani S, Sun W, Malik J, Efros AA. View synthesis by appearance flow. Proceedings of the European Conference on Computer Vision. 2016; Amsterdam, Netherlands; 1: 286-301.

[19] Hinton GE, Krizhevsky A, Wang SD. Transforming auto-encoders. Proceedings of the International Conference on Artificial Neural Networks. 2011; Espoo, Finland; 1: 44-51.

[20] Gao Z, Wang DY, Xue YB, Xu GP, Zhang H, Wang YL. 3D object recognition based on pairwise multi-view convolutional neural networks. Journal of Visual Communication and Image Representation 2018; 56: 305-315.

[21] He T, Mao H, Yi Z. Moving object recognition using multi-view three-dimensional convolutional neural networks. Neural Computing and Applications 2017; 28(12): 3827-3835.

[22] Wang D, Ouyang W, Li W, Xu D. Dividing and aggregating network for multi-view action recognition. Proceedings of the European Conference on Computer Vision (ECCV). 2018; 1: 451-467.

[23] Zhou Y, Shao L. Aware attentive multi-view inference for vehicle re-identification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018; 1: 6489-6498.

[24] Su H, Maji S. Multi-view convolutional neural networks for 3d shape recognition. Proceedings of the IEEE International Conference on Computer Vision 2015; 1: 945-953

[25] Setio AAA, Ciompi F, Litjens G, et al. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. IEEE Transactions on Medical Imaging 2016; 35(5): 1160-1169.

[26] Feng Y, Zhang Z, Zhao X, Ji R, Gao Y. GVCNN: Group-view convolutional neural networks for 3D shape recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018; 1: 264-272.

[27] Li XX, Cao Q, Wei S. 3D object retrieval based on multi-view convolutional neural networks. Multimedia Tools and Applications 2017; 76(19), 20111-20124.

[28] Ijjina EP, Chalavadi KM. Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recognition 2017; 72: 504-516.

[29] Li C, Wang P, Wang S, Hou Y, Li W. Skeleton-based action recognition using LSTM and CNN. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW); Hong Kong; 2017. pp 585-590.

[30] Liu M, Liu H, Chen C. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition 2017; 68: 346-362.

[31] Zhu F, Shao L, Lin M. Multi-view action recognition using local similarity random forests and sensor fusion. Pattern Recognition Letters 2013; 34(1): 20-24.

[32] Iosifidis A, Tefas A, Pitas I. Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis. Signal Processing 2013; 93(6): 1445-1457.

[33] Yan Y, Liu G, Ricci E, Sebe N. Multi-task linear discriminant analysis for multi-view action recognition. In: IEEE International Conference on Image Processing; Melbourne, Australia; 2013. pp 2842-2846.

[34] Chaaraoui A, Andre P, Climent-Pérez F, Flórez-Revuelta . An efficient approach for multi-view human action recognition based on bag-of-key-poses. In International Workshop on Human Behavior Understanding. Springer; Vilamoura, Portugal; 2012. pp 29-40.

[35] Ge L, Liang H, Yuan J, Thalmann D. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; Las Vegas, NV, USA; 1: 3593-3601.

[36] Liu AA, Xu N, Su YT, Lin H, Hao T, Yang ZX. Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 2015;151:544-553.

[37] Shahroudy A, Liu J, Ng TT, Wang G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; Las Vegas, NV, USA; 1: 1010-1019.

[38] Li M, Leung H. Multiview Skeletal Interaction Recognition Using Active Joint Interaction Graph. IEEE Transactions on Multimedia 2016; 18(11): 2293-2302.

[39] Maddala TKK, Kishore PVV, Eepuri KK, Dande AK. YogaNet: 3-D Yoga Asana Recognition Using Joint Angular Displacement Maps With ConvNets. IEEE Transactions on Multimedia 2019; 21(10): 2492-2503.

[40] Gao Z, Han T, Zhang H, Xue Y, Xu G. MMA: a multi-view and multi-modality benchmark dataset for human action recognition. Multimedia Tools and Applications 2018; 77(22): 29383-29404.