

## Sign Language to Sentence Formation: A Real Time Solution for Deaf People

Muhammad Sanaullah<sup>1,\*</sup>, Muhammad Kashif<sup>2</sup>, Babar Ahmad<sup>2</sup>, Tauqeer Safdar<sup>2</sup>, Mehdi Hassan<sup>3</sup>,  
Mohd Hilmi Hasan<sup>4</sup> and Amir Haider<sup>5</sup>

<sup>1</sup>Bahauddin Zakariya University, Department of Computer Science, Multan, 60,000, Pakistan

<sup>2</sup>Air University, Department of Computer Science, Multan, 60,000, Pakistan

<sup>3</sup>Air University, Department of Computer Science, Islamabad, 44,000, Pakistan

<sup>4</sup>Centre for Research in Data Science, Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia

<sup>5</sup>Department of Intelligent Mechatronics Engineering, Sejong University, Seoul, 05006, Korea

\*Corresponding Author: Muhammad Sanaullah. Email: drsanaullah@bzu.edu.pk

Received: 23 July 2021; Accepted: 29 September 2021

**Abstract:** Communication is a basic need of every human being to exchange thoughts and interact with the society. Acute peoples usually confab through different spoken languages, whereas deaf people cannot do so. Therefore, the Sign Language (SL) is the communication medium of such people for their conversation and interaction with the society. The SL is expressed in terms of specific gesture for every word and a gesture is consisted in a sequence of performed signs. The acute people normally observe these signs to understand the difference between single and multiple gestures for singular and plural words respectively. The signs for singular words such as I, eat, drink, home are unlike the plural words as school, cars, players. A special training is required to gain the sufficient knowledge and practice so that people can differentiate and understand every gesture/sign appropriately. Innumerable researches have been performed to articulate the computer-based solution to understand the single gesture with the help of a single hand enumeration. The complete understanding of such communications are possible only with the help of this differentiation of gestures in computer-based solution of SL to cope with the real world environment. Hence, there is still a demand for specific environment to automate such a communication solution to interact with such type of special people. This research focuses on facilitating the deaf community by capturing the gestures in video format and then mapping and differentiating as single or multiple gestures used in words. Finally, these are converted into the respective words/sentences within a reasonable time. This provide a real time solution for the deaf people to communicate and interact with the society.

**Keywords:** Sign language; machine learning; conventional neural network; image processing; deaf community



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1 Introduction

Communication is the primary source of sharing and transferring any knowledge in any society. Generally, human do this task through speaking/talking whereas deaf people cannot hear and talk. Therefore, the medium of communication for such kind of persons is the Sign Language (SL) to interact and participate into a society. The statistics of such people is very alarming according to the World Health Organization (WHO) which is around 466 million people in the world [1,2] and 0.6 million people only from United States [3]. The SL is based on hand gestures to transfer and share their ideas with the society. Although, the understanding of common used gestures such as hungry, drink, go and study are easy but for the formal or professional talks, there must be a proper knowledge of SL is required. These talks contain on such kind of sentences as what is your name? How old are you? What is the specification of this mobile phone?

Hence, the deaf people face many difficulties in sharing their thoughts in their professional career and as a result, they feel lonely and go in isolation. A technical solution is demanded to overcome the information transferring barrier which must provide the answer of the following concerns.

In SL the movements of the hand gestures can be divided into two categories, “Single Sign” (e.g., I, eat, drink) which contains the sign of a single gesture and “Multiple Sign” (e.g., she, hot, outside, etc.) which accommodates the multiple gestures for a single concept. This categorization is further divided into “Single Hand” and “Both Hand” gestures on the basis of hand movement. For example, the gestures of “I”, “You”, “go”, “drink” etc. are performed by a single hand and have a single sign. On the same site, the gestures of “she”, “hot”, “outside” etc. are performed by a single hand and have multiple signs. Whereas some gestures are performed by both hands and have single signs e.g., “home”, “love” etc. and both hands with multiple signs for e.g., play, football, car, drive, etc.

According to the categorization and grouping mentioned above, a summary of the found literature is shown in [Tab. 1](#). Most of the literature is about “single sign by single hand”. Unfortunately, no literature is found on “multiple signs with both hands”. The capturing and recognition of hand movement and rotation add complexities in the computed solutions design; therefore, the existing research mainly focuses on the primary category.

**Table 1:** State-of-the-art of gesture recognition

Gesture type/Performing hand	Word		Sentence	
	<i>Multiple sign</i>	<i>Single sign</i>	<i>Single sign</i>	<i>Multiple sign</i>
Single hand	Literature	Specific	Specific	Not found
Double hand	Specific	Not found	Not found	Not found

[Fig. 1](#) the left side gesture is from Britain Sign Language (BSL), which shows that the right hand with the first finger in up-word direction is used to perform “What”. The middle gesture is from American Sign Language(ASL), which shows that both hands with open palms are used to perform “What” and the right side gesture is from Pakistan Sign Language (PSL), which shows that the right hand with first two out-word pointing fingers is used to perform “What”. Hence, a holistic solution for all cultural societies is not possible and each country has to design and develop its own solution according to its society gestures.



**Figure 1:** Sign of “What” in different sign languages

Moreover, in the current proposed solutions the following concerns are also addressed:

- **Mobility Issue:** Most of the researchers use Kinect 3D camera, which can easily detect hand movements. But in a real world, it is difficult to carry out such hardware to every place where ever the special person move (e.g., market, school and hospital). There is another solution in form of a sensory glove with Flex sensor, in which contact sensor and accelerometer is used to detect hand movement and rotation. However, this solution also required to carry out the extra hardware at every-time. The proposed methodology is offering a solution without any extra hardware for mobility and movement of such special persons.
- **Sign Capturing Issue:** In a real environment, the person can wear any color and type of shirt/t-shirt and the detection of hand and their movement is not an easy task (skin segmentation of face, arms and hand). Therefore, most researchers assume to use colored gloves or different color strips attached to white color gloves. To wear these gloves at everywhere and every time which is a difficult job. This issue is resolved in the proposed solution using the video capturing for signs identifications.
- **Special Environment Issue:** For the recognition of a person, researchers design a special environment in which some specific colored dress with a specific color background is required. They used the black dress with black background (easy detection of skin), but in a real environment, it is quite challenging to maintain such an environment everywhere. Hence, there must be a solution for real time world in which the person and sign detection is much easier.
- **Sign Recognition Issue:** Most researchers work on recognizing words with “Single sign” gesture, but in real life, we speak complete sentences. The SL sentences consist of a sequence of gestures, recognizing each words from a single sign is also a difficult task.
- **Human Experimental Issues:** Experimental issues are also found in literature where solutions are validated with a limited number of persons and a recorded dataset of videos.

Thanks to Conventional Neural Network (CNN) and advanced Image Processing techniques to facilitate us for providing a solution to translate the SL without any special environment and fixed specific hardware. It also facilitates us in recognition of the gestures of sentences involving both hands with multiple signs. The validation of the proposed solution is confirmed with ten different males and females in a different real environment. After resolving all the mentioned issues, 94.66% accuracy is achieved.

In the rest of the paper, Section 2 present the literature review, Section 3 explains the proposed solution and the results are presented in Section 4. The discussions on results are presented in Section 5. The conclusion and future work is given in Section 6.

## 2 Literature Review

In this section, a literature review of the existing research work is presented, due to the space limitation only the paper which are mostly cited are presented. The literature is evaluated on the basis of the following parameters: their research assumptions, considered gestures, used hardware, numbers of verified signs, number of participants, learning and testing techniques and the accuracy. A summary in the form tabular view, of the evaluation, is presented in [Tabs. 2](#) and [3](#).

**Table 2:** Different SL translation techniques and their limitations

Ref.	Year	Algorithm	Gestures	Dataset collection method/ Device	Sign language	Limitation	Real environment
[4]	2015	PCA	Alphabets	Arduino sensors	American	Hardware mobility	No
[5]	2016	–	Words	Arduino sensors	American	Hardware mobility	No
[6]	2020	NN	23 alphabets 67 words 10 digits	Static images dataset	Indian	Specific background color No sentences	No
[7]	2014	Euclidean distance	Words (Urdu)	Arduino sensors	Pakistani	Hardware mobility	No
[8]	2018	–	Words	Arduino sensors	American	Hardware mobility	No
[9]	2011	ANN	Finger spellings	Webcam	–	Webcam fixed	No
[10]	2015	CNN	Alphabets digits	3D sensor	American	3D device fixed	No
[11]	2019	Deep NN	Sentences	Recorded videos	German	Background and dress color specific, tested in controlled environment	No
[12]	2007	ANN	10 words alphabets numbers	Webcam	Malaysian	Color gloves fixed webcam	No
[13]	2008	RNN	Arabic alphabets	Digital camera	Arabian	Color gloves	No
[14]	2018	CNN	Words	iPhone 6 mobile camera	American	Same background	No

(Continued)

**Table 2:** Continued

Ref.	Year	Algorithm	Gestures	Dataset collection method/ Device	Sign language	Limitation	Real environment
[15]	2019	CNN	Words, sentences	Leap motion sensor	Indian	Mobility	No
[16]	2019	MKNN GT2K RSAR	Words, sentences	Motion detector camera data glove	Arabian	Mobility	No

**Table 3:** Different sign recognition techniques

Ref.	Gesture type			Training data	Testing data	Image processing	Machine learning	Accuracy
	Performing hands	Words	Sentences					
[4]	Single hand	26	✗	26	–	✗	✓	92%
[5]	Single hand	2	✗	–	40	✗	✗	82.5%
[6]	Both hands	100	✗	–	–	✓	✓	99.90%
[7]	Both hands	10	✗	300	10	✗	✗	90%
[8]	Single hand	9	✗	–	90	✗	✗	74%
[9]	Single hand	14	✗	84	–	✓	✓	80%
[10]	Single hand	26	✗	50%	25%	✓	✓	85.49%
[11]	Both hands	✗	603	1809	531	✓	✓	91.93%
[12]	Single hand	26	✗	–	–	✓	✓	90%
[13]	Single hand	30	✗	900	300	✓	✓	95.11%
[14]	Both hands	150	✗	1800	–	✓	✓	91%
[15]	Single hand	35	942	–	–	✗	✓	89.50%
[16]	Both hands	–	40	–	–	✗	✓	97.78%

Bukhari et al. [4] designed a sensory glove having flex sensors for capturing the movement of fingers, accelerometer for capturing the rotation of hands and contact sensors for bending of palm. They used 26 gestures for recognition. Each sign is recorded 20 times. They got 92% accuracy for the recognition of gestures. This work has issues of mobility, sign recognition and Human Experimental.

Helderman et al. [5] designed a sensory glove for the translation of SL. They used flex sensor, contact sensor and gyroscope for capturing the contact between fingers and rotation of hand. They used Arduino for controlling sensors. Blue-tooth module was used to transmit signal from Arduino to smart phone. They recognized only two signs from ASL. The first sign was “apple” and second was “party”. They test each sign 20 times. Their glove recognized “apple” 19 times and 14 times “party” sign. Their glove shows 95% accuracy for “apple” and 70% accuracy for “party” sign. This work has issues of mobility, sign recognition and Human Experimental.

Wadhawan et al. [6] proposed a deep learning based SL recognition system. They used 100 static words for the recognition and they achieved 99.90% accuracy. Their system has the limitation of mobility and real environment.

Kanwal et al. [7] designed a sensory glove for the translation of Pakistani SL. They tested ten signs of PSL and their glove recognized only nine signs accurately. The accuracy of their sensory glove is 90%. This work has issues of mobility, sign recognition issues and Human Experimental issues.

Ambar et al. [8] designed a sensory glove to recognize the words of American SL. They got an accuracy of 74% for translating SL. They used the sensor for capturing the movement of fingers and hands. This work has issues of mobility, sign recognition issues and Human Experimental issues.

Lungociu [9] proposed a neural network approach for the recognition of SL. They used 14 finger spellings for the recognition. Their accuracy was 80%. They used webcam for data acquisition and only captured the hand shapes. This work has issues of mobility, sign recognition, Human Experimental, Sign capturing and special environment.

Kang et al. [10] used CNN for the recognition of SL. They recognized alphabets and digits taken from ASL. They used 3D sensor for the capturing of sign. They got 85.49% accuracy. This work has issues of mobility, sign recognition, Human Experimental, Sign capturing and special environment.

Cui et al. [11] proposed a framework for SL recognition. They recognized 603 recorded sentences of German SL. They got 91.93% accuracy for using Deep NN. This work has human experimental issues.

Akmeliawati et al. [12] translated the Malaysian SL using image processing technique. A webcam is used for data capturing and color gloves are used to capture the sign. In this approach, the author translated A–Z alphabets, 0–9 numbers and ten words. Their approach gained 90% accuracy for recognition. This work has mobility issues, sign recognition issues, human experimental issues, Sign capturing and special environment issues.

Maraqqa et al. [13] proposed a system for Arabic SL recognition. They used the Digital camera for capturing images and captured 900 images of 30 signs training data-set. Three hundred more images are captured for testing data. They used the white color glove with different color patches on the fingertips and a wrist color band. They gained an accuracy of 95% for sign recognition. This work has mobility issues, sign recognition issues, human experimental issues, Sign capturing and special environment issues.

Bantupalli et al. [14] American SL using RNN and CNN. They recorded videos using iPhone 6 with same background. They tested 150 signs and achieved accuracy of 91%. Their work has specific background issue.

Mittal et al. [15] used Leap Motion sensor to capture Indian SL CNN was used to recognize it. They tested words and sentences. CNN was trained using 35 isolated words and model is tested using 942 sentences. Average accuracy for words 89.50% and for sentences 72.30%. This work has mobility issues.

Hassan et al. [16] produced three data sets of Arabian SL. They also used different techniques of recognition. Data sets have words and sentences. Two data sets are produced using motion detector and camera and one data set is produced using sensory glove. Tools used for classification of SL are MKNN, RASR and GT2K.

Ullah et al. [17] used Wi See technology that can detects the gestures using multiple antennas. They used the gesture to control the movement of a car. This work has mobility issues, sign recognition issues, human experimental issues, Sign capturing and special environment issues.

### 3 Proposed Solution

The proposed solution is divided in three components: Image Processing–in which video signs are captured and Key Frames are extracted–Classification–of gestures with the use of CNN–and Sentence Formation–where the words are arranged in a semantic form with the use of Natural Language Processing (NLP). The framework of the proposed solution is presented in Fig. 2 and explanation of each component is given in the following subsections:

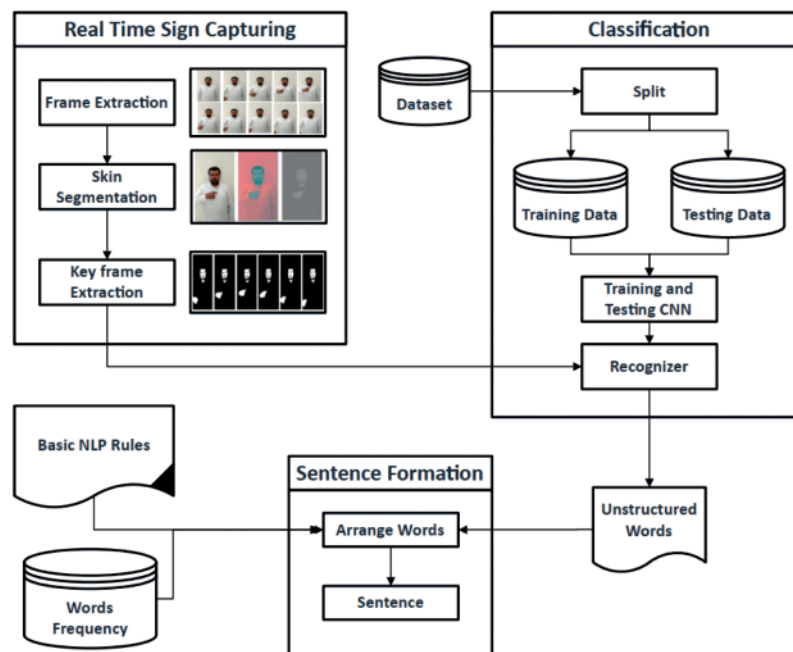


Figure 2: Framework of the proposed solution

### 3.1 Real Time Sign Capturing

The work performed in this component is to identify those frames in which some gestures are performed from the real time recording then frames are extracted. Further key frames are filtered from extracted frames. The key frames contain useful information of gesture sign. So that instead of working on all extracted frames, only filtered key frames are considered for further processing. This component consists of the following subcomponents.

#### 3.1.1 Signs' Frame Extraction

Real Time signs are captured in a video format using any digital camera. The real time video is accessed frame by frame and parsed to detect focused person movement by comparing these frames. In detecting any movement in a focused person, this component stores the frame identity and continues its working to detect the frame in which the movement stops. A set of these frames, from starting to end the movement detected frames, is sent to the "Skin Segmentation" component for further processing.

#### 3.1.2 Skin Segmentation

The skin Segmentation process works to identify the hands of the focused person. For this purpose, Otsu threshold method, which iterates through all the possible threshold values, calculates the spread for the pixels that fall in foreground or background. LAB color space is selected because it is an effective color space in Otsu thresholding. The activities performed for this purpose is sequenced in Algorithm 1. In which, the set of frames, extracted in the previous section and are in RGB format, are work as input and also shown in Fig. 3 part (A).

---

#### Algorithm 1: Skin Detection Algorithm

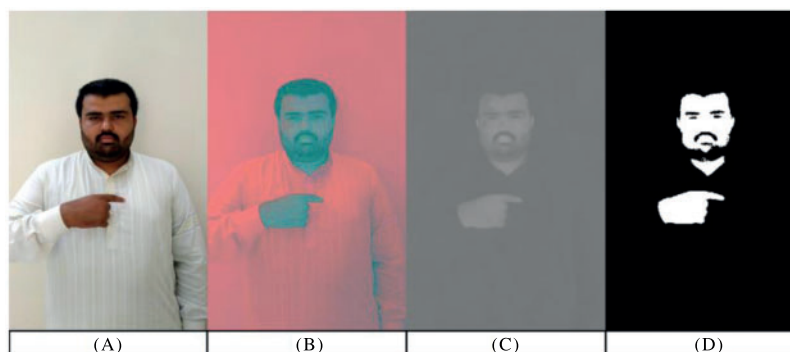
---

```

1 RGB color space image Skin Segmented Binary Image Get RGB sign image
2 Convert RGB image to LAB color space image
3 Apply Otsu threshold
4 Apply Gray threshold
5 Binaries the image to convert the background pixels to black and foreground pixels to white
return segmented image

```

---



**Figure 3:** Image of signer showing hand and face detection



### 3.1.3 Key Frame Extraction

A set of these binary images consists of many frames. Most of the frames represent the sequence of moving hands and do not contain any factual information, which is required for the recognition of gestures. The inclusion of these frames increases time and space complexity along with the increase of error rate. We need to exclude these frames. The rest of the frames are considered as key frames, which have some specific information. The step performed for key frames identification is presented in Algorithm 2.

Firstly, it calculates the entropy of the considering frame based on the frames histogram values and then compares these values. The base value from which the comparison is made is the base frames entropy in which hands are in the rest position. This comparison is performed based on Eq. (1) in which,  $\alpha$  represents the entropy of the first frame,  $\beta_j$  represents the entropy of other frames and  $\mu$  represents the threshold value.

$$\mu = (\beta_j - \alpha) / (\beta_j + \alpha) \quad : j < 1 \tag{1}$$

From a lot of experiments, it is observed that the frames having threshold value less than 0.8 are just containing hands movements and therefore can be excluded. Hence, the frames having threshold value greater than 0.8 are considered as key frames of this sign. Let consider the sign “I”, as shown in Fig. 3 the entropy and threshold values against each frame is presented in Tab. 4. In which the entropy of the first/base frame is 6.8423 which will remain same in the whole process. This “I” sign consist of 77 frames and after the key frame identification process only 38 frames are considered as key frames, rest of the frames are excluded, these 38 frames are passed to the classification component for further processing.

---

**Algorithm 2:** Key Frames Extraction Algorithm

---

```

A ← Acquire first Image (A ∈ Segmented Image)
I ← Calculate Entropy(A)
for each, B ∈ (Segmented Image) do
    J ← Calculate Entropy(B)
    M ← Calculate Difference(I, J)
    N ← Calculate Sum(I, J)
    Threshold ← Calculate Division(M, N)
    If (Absolute(Threshold * 100) ≥ 0.8) then
        Keyframe ← Image B is
    end
end
return KeyFrames
    
```

---

**Table 4:** Keyframe selection based on entropy and threshold

Frame Id	Entropy	Threshold	Key frames	Frame Id	Entropy	Threshold	Key frames
2	6.8439	0.011461	N	40	6.9545	0.81353	Y
3	6.846	0.026801	N	41	6.9586	0.84253	Y

(Continued)

**Table 4:** Continued

Frame Id	Entropy	Threshold	Key frames	Frame Id	Entropy	Threshold	Key frames
4	6.8455	0.023309	N	42	6.9538	0.80865	Y
5	6.8349	-0.05414	N	43	6.9683	0.91256	Y
6	6.8405	-0.0133	N	44	6.9697	0.92222	Y
7	6.8385	-0.02778	N	45	6.9649	0.88807	Y
8	6.8409	-0.00996	N	46	6.9653	0.8907	Y
9	6.8385	-0.02742	N	47	6.9704	0.92729	Y
10	6.8377	-0.03331	N	48	6.9694	0.92069	Y
11	6.8406	-0.01234	N	49	6.971	0.9316	Y
12	6.8412	-0.00779	N	50	6.9739	0.95259	Y
13	6.8441	0.012979	N	51	6.9798	0.99498	Y
14	6.8438	0.011086	N	52	6.984	1.0252	Y
15	6.8464	0.030202	N	53	6.9839	1.0242	Y
16	6.8467	0.032572	N	54	6.9823	1.013	Y
17	6.8455	0.023811	N	55	6.9867	1.0443	Y
18	6.8509	0.063093	N	56	6.9892	1.0619	Y
19	6.8491	0.049826	N	57	7.0023	1.1557	Y
20	6.8514	0.066274	N	58	7.0115	1.2216	Y
21	6.8528	0.076557	N	59	7.01	1.2109	Y
22	6.8732	0.22537	N	60	7.0198	1.2806	Y
23	6.9035	0.44548	N	61	7.0438	1.4514	Y
24	6.946	0.75219	N	62	7.0408	1.4296	Y
25	6.973	0.94649	Y	63	7.0305	1.3564	Y
26	6.9837	1.0228	Y	64	7.0112	1.2194	Y
27	6.9881	1.0546	Y	65	6.9798	0.995	Y
28	6.9902	1.0695	Y	66	6.9484	0.76933	N
29	6.9873	1.0487	Y	67	6.9142	0.52295	N
30	6.992	1.082	Y	68	6.8936	0.37379	N
31	6.9759	0.96668	Y	69	6.8855	0.31473	N
32	6.9724	0.94206	Y	70	6.8785	0.26426	N
33	6.9633	0.87633	Y	71	6.8794	0.27034	N
34	6.9628	0.87285	Y	72	6.8739	0.23041	N
35	6.9554	0.81979	Y	73	6.8685	0.19125	N
36	6.9507	0.78614	N	74	6.8683	0.18951	N
37	6.9407	0.7137	N	75	6.8711	0.21003	N
38	6.9448	0.74389	N	76	6.8689	0.19395	N
39	6.9524	0.79862	N	77	6.8697	0.19986	N

### 3.2 Classification

The classification component takes the key frames as input and predicts the label of gesture that belongs to that key frames. The overall working of the classification component is explained in the following subsections.

### 3.2.1 Sign Repository

PSL case study is used to implement the developed methodology. A repository of PSL signs is created using pre-recorded gestures performed by different people at different places of different age and gender groups. Some of the videos from the official PSL website [18] are also included in the repository. Firstly, a database of 300 daily life sentences are created, in which it is found that some words are repeatedly used, among which 21 high-frequency words are selected and shown in Tab. 5 with their occurrence frequencies. These words gesture is recorded three times by three different people and their videos are stored in a signed repository. Moreover, 15 gestures videos are downloaded from the PSL website and also stored in the signed repository. Hence, the total recorded videos are 204. The total number of Key frames that are extracted from 204 videos is 1882. The key frames of each gesture are labeled with the appropriate name (e.g., we, Today, Drive, etc.).

PSL case study is used to implement the developed methodology. A repository of PSL signs is created using pre-recorded gestures performed by different people at different places of different age and gender groups. Some of the videos from the official PSL website [18] are also included in the repository. Firstly, a database of 300 daily life sentences are created, in which it is found that some words are repeatedly used, among which 21 high-frequency words are selected and shown in Tab. 5 with their occurrence frequencies. These words gesture is recorded three times by three different people and their videos are stored in a signed repository. Moreover, 15 gestures videos are downloaded from the PSL website and also stored in the signed repository. Hence, the total recorded videos are 204. The total number of Key frames that are extracted from 204 videos is 1882. The key frames of each gesture are labeled with the appropriate name (e.g., we, Today, Drive, etc.).






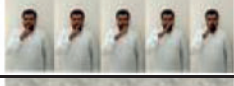













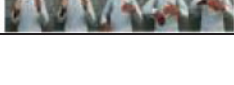
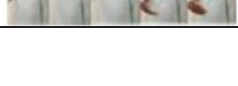
### 3.2.2 Training and Testing with CNN Model

CNN is a type of Deep Learning algorithm which belongs to machine learning. CNN is best for recognition because it implicitly calculates the features, whereas, in other techniques, the features are required to be calculated explicitly. The input image layer is  $200 \times 200$  pixels in size, we have 1882 images in over sign repository. Convolution filters are applied to the inputted image using a 2-D convolution layer. This layer convolves the image vertically and horizontally by moving the filters and computes the dot product of CNN weights to the inputted image. Moreover, the convolutional layer has 20 filters of size  $5 \times 5$  and has a ReLU layer, which automatically works for rectifying the error. The max-pooling layer is created, which performs down-sampling on the input and dividing into rectangular regions by computing the maximum value for every region. The layer has pool size  $2 \times 2$  and a stride of 2, followed by the Fully Connected Layers. The architecture of the Designed neural network is shown in Fig. 4. From the signed repository, 75% data is used for training purposes and 25% data is used for testing purposes.

### 3.2.3 Recognizer

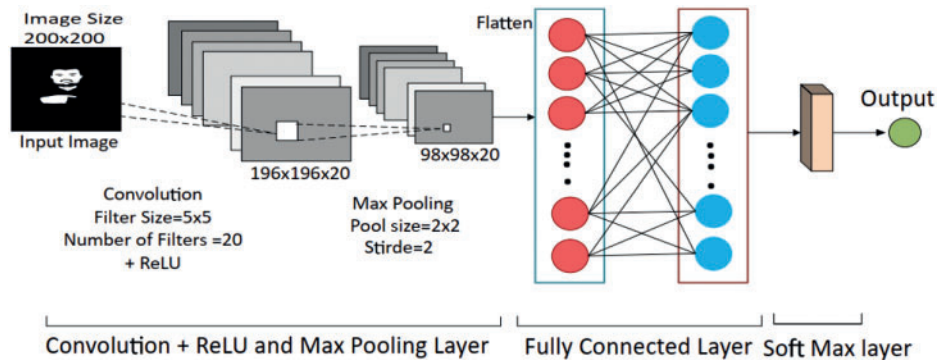
After training and testing, recognition is done through CNN. Extracted Key frames of a gesture are given to the Recognizer component, which works with the trained CNN model, discussed in Section 3.2.2. The recognizer identifies the total number of gestures in the given set of key frames and returns the gestures labels. In the case of gesture “I” 38 frames are passed to the recognizer. It takes 4 s to recognize and label it. In the case of “She” and “beautiful”, it takes 13 s to recognize and label it.

**Table 5:** Bag of Words

Gesture	Frequency	Extracted key frames	Some key frames	Gesture	Frequency	Extracted key frames	Some key frames
I	66	38		Beautiful	81	73	
Car	36	96		City	15	90	
Drink	42	66		Hot	24	72	
Lahore	30	184		Milk	30	195	
Play	36	179		Today	27	84	
Went	24	60		See	30	64	
you	75	68		Go	30	68	
Home	21	74		Rain	15	63	
She	75	98		Like	45	68	
Drive	27	70		Football	36	103	
Outside	42	72					

### 3.3 Sentence Formation

The identified labels are sent to this component. Firstly, the nature of labels are identified like subject, verb or object using NLP and a dictionary in which most terms are classified with their nature. The frequency in which these terms are used that their labels are arranged in Subject–verb–object format. Although it is not fully satisfied the English grammar rules but some extent it is able to convey the meaning of the sentence. For example when a person perform gestures as shown in [Fig. 5](#) after all the processing the sentence formation component returns “I go Home”.



**Figure 4:** CNN architecture



**Figure 5:** Frames from the sentence “I go to home”

## 4 Results

To measure the validity of the proposed solution, a set of 300 daily used sentences is considered. A “bag of words” file is generated from these sentences, which contains the different words with their occurrence frequencies, as 21 high-frequency words are shown in [Tab. 5](#). The dataset for training and testing purposes is created by using these 21 gestures. Each gesture is performed three times by three different males and females of different age groups at different locations. Moreover, 15 videos or SL expert from the Pakistan Sign Language website [18] is also added to the repository. Overall, we have 204 videos, these videos are processed for the key frame extraction component and in-result 1882 frames were generated. The CNN considers 75% for training and takes approx. 4 min (3 min and 37 s) for this purpose.

Real Time Sign Capturing identifies the focused person body movement and parses it for the key frame extraction. These key frames are further passed to the recognizer component, which performs recognition of the gesture(s) based on the trained module and returns the label(s) of the gesture(s).

The time required of key frame identification from real time video to movement identification, frame by frame parsing and then key frame extraction is given in [Tab. 8](#) under Capturing time parameter and the recognition time shows the time taken for the recognition and labeling of the gestures and the accuracy is the values provided by CNN model against each sentence. Overall, the achieved accuracy is 94.66%.

[Tab. 7](#), shows the time spend in the case of single words, in which, Gesture capturing time is that time which is required for the Key frames extraction from a video of a gesture and the recognition time is the time which recognizer takes to recognize a gesture.

Precision, recall, false-negative rate, false discovery rate and f-score are the measure used to measure a classification algorithms performance. For this, the standard formulas are given below. In which, True Positive (TP): Actual value is positive and predicted is also positive value, False Negative (FN): Actual value is positive but predicted is a negative value, True Negative (TN): Actual value is negative and predicted is also negative value and False Positive (FP): Actual value is negative, but the predicted value is positive.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$\text{FN Rate} = \text{FN}/(\text{TP} + \text{FP}) \quad (4)$$

$$\text{FP Rate} = \text{FP}/(\text{TN} + \text{FP}) \quad (5)$$

$$\text{F1 - Score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (6)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (7)$$

These formulas are used to calculate the binary classification results. Our dataset has multi-class labels and results are calculated using confusion matrix which is given in [Fig. 6](#). The following [Tab. 6](#) is playing the calculated values for precision, recall, false positive rate, false negative rate, F1-score and accuracy of the above mentioned gestures.

The rows presents the predicted class and the columns presents to the actual class. The diagonal cells presents to observations that are correctly classified. The off-diagonal cells presents incorrectly classified observations. Both the number of observations and the percentage of the total number of observations are shown in each cell. The last column shows the percentages of all the examples predicted to belong to each class that are correctly and incorrectly classified. These metrics are often called the precision and false discovery rate, respectively. The row at the bottom shows the percentages of all the examples belonging to each class that are correctly and incorrectly classified. These metrics are often called the recall and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy.

## 5 Discussion

[Tab. 7](#) the first column displays the Gesture label and the second column displays gesture capturing time. The “Real Time Sign Capturing” component takes gesture capturing time. The “Real Time Sign Capturing” component has three sub-components. The total time of three sub-components is given in that column. This time is depending on the number of key frames in a gesture or gesture performing time. As gesture performing time increases, the number of key frames also increases, so capturing time increases. One more observation is that which gesture consists of multiples signs also has a large capturing time. The gesture, which consists of a single gesture, has the lowest capturing time. As shown in table “Lahore” gesture has the highest capturing time, 171 s, because it consists of multiple gestures and has a greater performing time. Similarly, a gesture “I” has the lowest capturing time because it consists of a single sign, has the lowest number of key frames and has the lowest-performing time. This time can be reduced if we can build a method that finds the minimum number of key frames from a gesture.

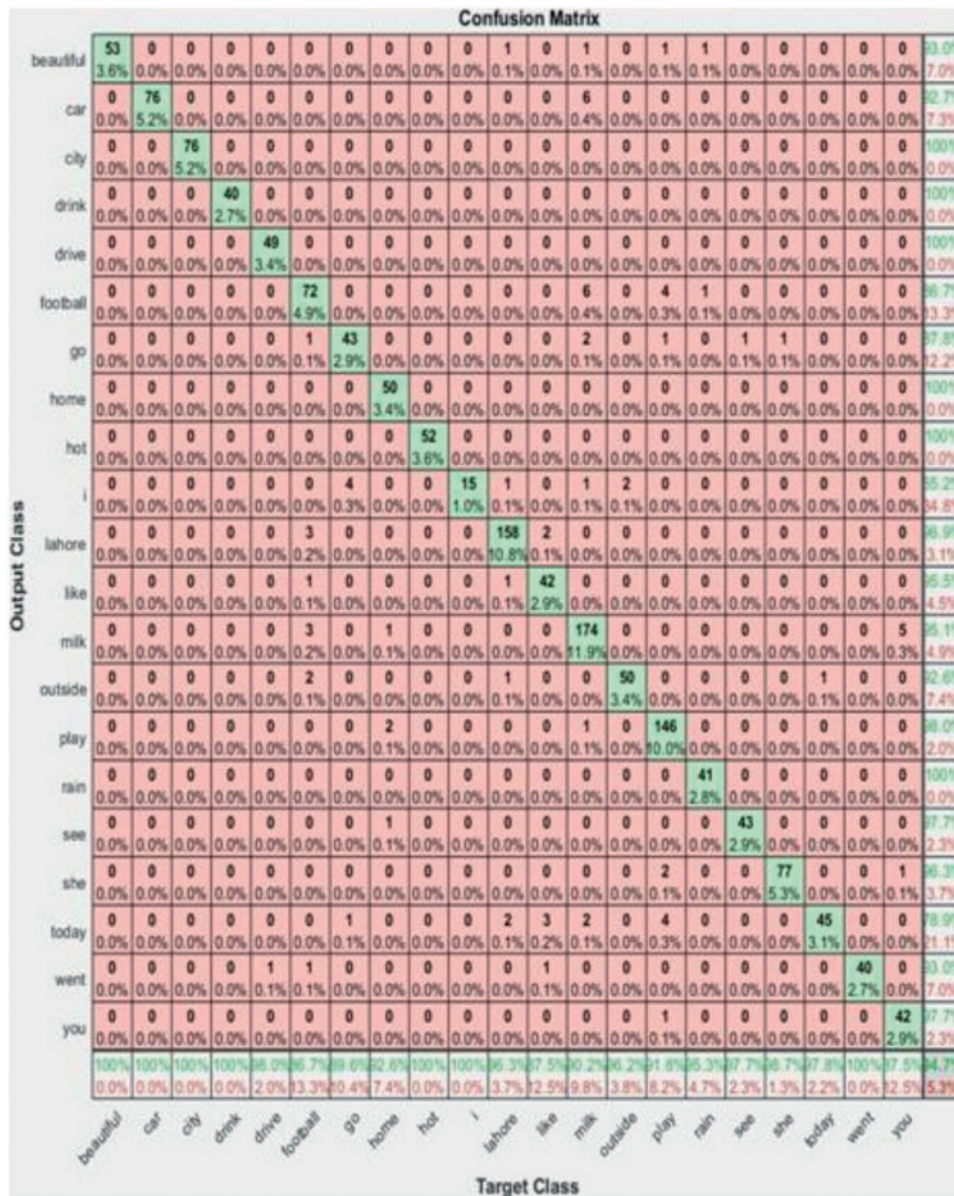


Figure 6: Confusion matrix

Table 6: Performance measures for trained model

Gesture	Recall	False negative rate	Precision	False positive rate	F1-score	Accuracy (%)
Beautiful	1	0	0.93	0.07	0.963731	92.98
Car	1	0	0.927	0.073	0.962117	92.68
City	1	0	1	0	1	100
Drink	1	0	1	0	1	100

(Continued)

**Table 6:** Continued

Gesture	Recall	False negative rate	Precision	False positive rate	F1-score	Accuracy (%)
Drive	0.98	0.02	1	0	0.9898	100
Football	0.867	0.133	0.867	0.133	0.867	86.74
Go	0.896	0.104	0.878	0.122	0.886909	87.75
Home	0.926	0.074	1	0	0.961578	100
Hot	1	0	1	0	1	100
I	1	0	0.652	0.348	0.789346	65.21
Lahore	0.963	0.037	0.969	0.031	0.965991	96.93
Like	0.875	0.125	0.955	0.045	0.913251	95.45
Milk	0.902	0.098	0.951	0.049	0.925852	95.08
Outside	0.962	0.038	0.926	0.074	0.943657	92.59
Play	0.918	0.082	0.98	0.02	0.947987	97.98
Rain	0.953	0.047	1	0	0.975934	100
See	0.977	0.023	0.977	0.023	0.977	97.72
She	0.987	0.013	0.963	0.037	0.974852	96.25
Today	0.978	0.022	0.789	0.211	0.873392	78.94
Went	1	0	0.93	0.07	0.963731	93.02
You	0.875	0.125	0.977	0.023	0.923191	97.67

**Table 7:** Recognized bag of words

Gesture	Gesture capturing time (s)	Recognition time (s)	Total time (s)
Beautiful	78	5	83
Drink	99	4	103
Drive	73	3	76
Football	93	9	102
Go	72	6	78
Home	78	7	85
I	37	4	41
Lahore	171	15	186
Like	60	5	65
Milk	115	6	121
Outside	64	4	68
Play	93	14	107
Rain	69	5	74
See	54	4	58
She	105	8	113
Today	78	5	83

(Continued)



**Table 7:** Continued

Gesture	Gesture capturing time (s)	Recognition time (s)	Total time (s)
Went	41	4	45
You	56	5	61

The third column in [Tab. 7](#) is displaying the gesture recognition time. A trained model recognizes a frame in milliseconds, but our recognizer takes some seconds to recognize a gesture because it recognizes the whole number of key frames extracted from a gesture. So recognition time is depending on the number of key frames that are given to the recognizer. As shown in the table “I” gesture takes 4 s, which is the lowest recognition time in our case because gesture “I” has the lowest number of extracted key frames. Similarly, the gesture “Lahore” has the highest recognition time, which 15 s in our case. This time is highest due to the highest number of key frames. After the recognition of key frames, the highest frequency label is considered as a final label of gesture. The highest frequency label approach is adopted due to similar key frames in gesture. It is observed that many of the gesture has similar key frames e.g., (Drink and Milk, Drive and Car, etc.). Similarly, [Tab. 8](#) shows the capturing and recognition time of the sentence gestures.

**Table 8:** Sentence capturing and recognition time

Sentence	Capturing time (s)	Recognition time (s)	Accuracy (%)
She is beautiful	183	13	94.61
I go home	187	17	84.32
You see football	203	18	94.04
I like rain	166	14	86.88
I drink milk	252	14	86.73
I play football	223	27	83.31
She goes Lahore	348	29	93.64
You drive	129	8	98.84
Today rain outside	211	14	90.51
She went home	224	19	96.42
You see Lahore	179	14	97.44
She like milk	280	19	95.59
I play outside	194	22	85.26
You like home	194	17	97.70
I went Lahore	249	23	85.13
Today rain Lahore	318	20	91.95
You like Lahore	287	20	96.68
You go outside	192	15	92.66
I see rain	160	13	87.64

## 6 Conclusion and Future Work

Deaf people are part of society and have the right to live in society and participate in every aspects of life. They need communication way to transfer and interact with other people participating in society. In this research, an automated gestures/signs recognition system is designed and developed. The computer-based Sign Language (SL) recognition solution is efficiently implemented to help the deaf people of Pakistan. We tried to remove mobility and gestures limitations of single vs. multiple signs by creating a special environment for SL translation in proposed computer-based solution. The proposed solution is even applicable for the signer to communicate through multiple gestures with the deaf people. For this purpose, he does not need to wear any extra hardware such as special gloves for SL translation. The multiple signs/gestures of individual words as well as the complete sentences can be recognized in the proposed solution as it works on the basis of sign videos captured from real environment and translate it into text. The results are verified by adding 204 videos which consist on 1882 key frames with an accuracy of 94.66%. In the future, the computational optimization for smart phone is recommended. On the other hand, computation power of mobile technology can be enhanced to enable the processing of the complex image processing and machine learning tasks.

**Acknowledgement:** The work presented in this paper is part of an ongoing research funded by Yayasan Universiti Teknologi PETRONAS Grant (015LC0-311 and 015LC0-029).

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>. Accessed: 2020-02-02.
- [2] World Health Organization, <https://www.yesprograms.org/stories/sign-language-accessibility-for-the-deaf-in-pakistan>. Accessed: 2020-02-02.
- [3] U.S deaf population, <https://www.gallaudet.edu/office-of-international-affairs/demographics/deafemployment-reports>. Accessed: 2020-10-07.
- [4] J. Bukhari, M. Rehman, S. I. Malik, A. M. Kambogh and A. Salman, "American sign language translation through sensory glove," *International Journal of u- and eService, Science and Technology*, vol. 8, no. 1, pp. 131–142, 2015.
- [5] A. M. Helderman, P. M. Cloutier and R. Mehilli, "Sign language glove," in *BS Project*, US: Worcester Polytechnic Institute, 2016.
- [6] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7957–7968, 2020.
- [7] K. Kanwal, S. Abdullah, Y. B. Ahmed, Y. Saher and A. R. Jafri, "Assistive glove for Pakistani sign language translation," in *Proc. 17th IEEE Int. Multi Topic Conf.*, Karachi, Pakistan, pp. 173–176, 2014.
- [8] R. Ambar, C. K. Fail, M. H. A. Wahab, M. M. A. Jamil and A. A. Ma'radzi, "Development of a wearable device for sign language recognition," *Journal of Physics: Conference Series*, vol. 1019, no. 1, pp. 012017, 2018.
- [9] C. Lungociu, "Real time sign language recognition using artificial neural networks," *INFORMATICA*, vol. 56, no. 4, pp. 75–84, 2011.
- [10] B. Kang, S. Tripathi and T. Q. Nguyen, "Real-time sign language finger spelling recognition using convolutional neural networks from depth map," in *Proc. 3rd IAPR Asian Conf. on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia, pp. 136–140, 2015.

- [11] R. Cui, H. Liu and C. Zhang, “A deep neural framework for continuous sign language recognition by iterative training”, *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019.
- [12] R. Akmeliawati, M. P. L. Ooi and Y. C. Kuang, “Real-time Malaysian sign language translation using colour segmentation and neural network,” in *Proc. IEEE Instrumentation & Measurement Technology Conf.*, Warsaw, Poland, pp. 1–6, 2007.
- [13] M. Maraqa and R. Abu-Zaiter, “Recognition of arabic sign language (arsl) using recurrent neural networks,” in *Proc. First Int. Conf. on the Applications of Digital Information and Web Technologies*, Ostrava, Czech Republic, pp. 478–781, 2008.
- [14] K. Bantupalli and Y. Xie, “American sign language recognition using deep learning and computer vision,” in *Proc. IEEE Int. Conf. on Big Data*, Seattle, WA, USA, pp. 4896–4899, 2018.
- [15] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian and B. B. Chaudhuri, “A modified lstm model for continuous sign language recognition using leap motion,” *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7056–7063, 2019.
- [16] M. Hassan, K. Assaleh and T. Shanableh, “Multiple proposals for continuous arabic sign language recognition,” *Sensing and Imaging*, vol. 20, no. 1, pp. 1–23, 2019.
- [17] S. Ullah, Z. Mumtaz, S. Liu, M. Abubaqr, A. Mahboobet *et al.*, “An automated robot-car control system with hand-gestures and mobile application using arduino,” *Sensing and Image*, preprint, 2019.
- [18] Pakistan sign language. <https://www.psl.org.pk/signs>. Accessed: 2020-02-02.