Sign Language Live Captioning Using Deep Learning

Saiyam Arora	–	35111502720
Mahira Puri	–	35311502720
Rohan Singh	–	07311502720
Vinayak	–	09811502720

Mentor

Dr. Preeti Nagrath

Introduction

- Only means of communication for hearing-impaired individuals
- Problems faced by people.
- Sign Language is the solution.
- Do different languages have different Sign Languages?

- How can one **COMMUNICATE** through Sign Language ?
- Is it possible to **TEACH** Sign Language to everyone?
- How can **TECHNOLOGY** help us in this ?
- What **IMPACT** does this create ?

Objective

- ➔ This proposal is based on **continuous detection of image frames** in real-time using **action detection** so as to detect the action performed by the user.

- ➔ Using **LSTM neural network model** after identifying keypoints using **mediapipe holistic** which includes face, pose and hand features.

- ➔ To **develop and compare** the performance of different deep learning models in the recognition and translation of sign language and human actions for effective communication.

- ➔ Finally, Making **Sign Language Decoding feasible** without learning all Hand Gestures of different Sign Language.

Literature Review

Author Name(s)	Paper Title	Accuracy	Methodology	Research Gap
[1] B. Bauer and K. Karl-Friedrich.	Towards an automatic sign language recognition system using subunits. (2001)	**95.4%**	The authors have used **coloured gloves** for data acquisition and hand segmentation. The recognition process was carried out using HMM-based language modelling where the accuracy of 95.4% and 93.2% were recorded. However, the system had restrictions on signer's clothing and required a uniform coloured background for correct segmentation of hands.	The system had restrictions on signer's clothing and required a uniform coloured background for correct segmentation of hands.
[2]C. Vogler and D. Metaxas.	A framework for recognizing the simultaneous aspects of american sign language. (2001)	**84.84%**	A framework for the continuous-SLR system using **three orthogonal cameras** to capture 3D hand movements using **parallel-HMM** on 99 ASL sentences with an accuracy of 84.84% that **outperforms the conventional HMM-based recognition.**	Although it outperforms the conventional HMM model in terms of processing, its accuracy was lower.
[3] H. Li and M. Greenspan	Model-based segmentation and recognition of dynamic gestures in continuous video streams (2011)	**82%**	Model-based framework for segmentation and recognition of continuous SLR using the **video sequence**. The authors presented three different approaches for **endpoint localization of gestures** that include multi-scale search, Dynamic Time Warping (DTW) and dynamic programming, where a recognition rate of 82% was recorded using early-decision dynamic programming with correlation and mutual information based similarity measures.	Tested only for 12 continuous sign gestures. Further research must be conducted to measure Accuracy using a bigger dataset.

Literature Review

Author Name(s)	Paper Title	Accuracy	Methodology	Research Gap
[4] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos.	Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. (2012)	**86%**	A framework for segmentation, tracking and modelling of hand shapes from video sequences was proposed. It offers a **compact and descriptive representation** of the hand configuration.The hand-shape features have been extracted using the affine modelled hand that was used to **construct an unsupervised set of sub-units which constitute the signs.**	Most handshapes are recognized correctly since there is a small number of clusters. However, because of the single subunit sequence map to multiple signs there is no sign discrimination.
[5] W. Gao, G. Fang, D. Zhao, and Y. Chen.	Transition movement models for large vocabulary continuous sign language recognition. (2004)	**90.8%**	SLR system for Chinese Sign Language (CSL) using **data gloves and three position trackers to extract the hand appearance and position**. The authors have extracted 48-dimensional feature vector that includes **hand shape, position and orientation vector.** A modified **k-means clustering algorithm** was used with DTW based distance measuring technique to cluster the transition movements between two signs.	It was assumed that the transition movement between two signs is always similar in different sentences. However, such transitions vary in real-world applications.
[6] N. Tubaiz, T. Shanableh, and K. Assaleh.	Glove-based continuous arabic sign language recognition in user-dependent mode. (2015)	**98.9%**	.A similar approach was taken by some researchers working on the Arabian Sign Language dataset. They have used a **modified version of k-NN algorithm** for classification of 40 signed sentences. However, the evaluation was performed in user-dependent mode.	Sensor gloves can be expensive.

Summarising the Literature Review

- Existing SLR systems are based on **2D video cameras, expensive and colored gloves , sensor-gloves, etc.**
- Recently, the SLR research is shifting to a novel **3D environment using depth cameras/Leap Motion sensors.**
- Most of the work on the recognition of sign gestures are based on **HMMs, Artificial Neural Networks (ANN)**

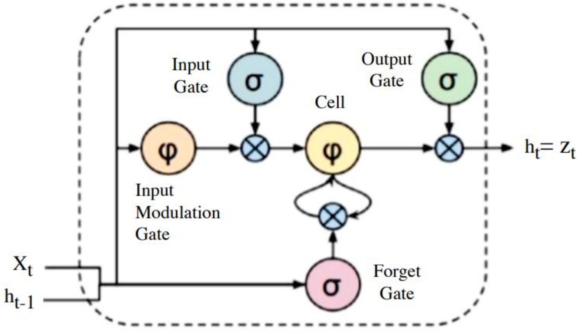
  and rule-based modelling techniques.

Methodology

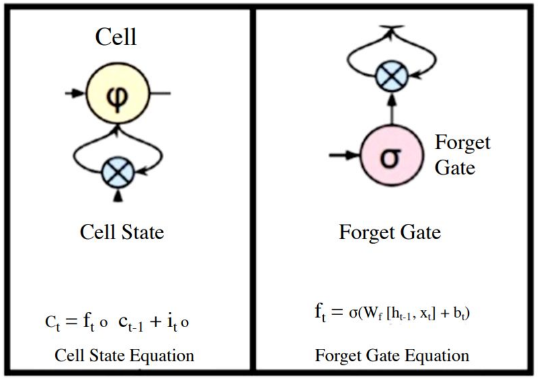
The main goal of this experiment was to develop an LSTM model to predict the American sign language using multiple frames and to predict the action being demonstrated in real-time.

What is LSTM?

LSTM Architecture Long short-term memory, commonly known as LSTM, is a type of RNN architecture that is capable of remembering values at arbitrary intervals. They are developed for the classification, processing, and prediction of time series of particular time lags with durations that are unknown. Unlike other sequence learning models such as Hidden Markov models (HMM) or other RNNs, LSTMs have relative intensity gaps, which provide an advantage over the alternatives.



LSTM



Cell
φ
Cell State
$C_t = f_t \circ c_{t-1} + i_t \circ$
Cell State Equation
σ
Forget Gate
Forget Gate
$f_t = \sigma(W_f [h_{t-1}, x_t] + b_t)$
Forget Gate Equation

Methodology Explained Step by Step

1. Install and Import Dependencies
2. Keypoints using MP Holistic
3. Extract Keypoint Values
4. Setup folders for collection
5. Collect keypoint values for training and testing

Methodology Explained Step by Step

6. Preprocess data and create labels and features
7. Build and train LSTM Neural Network
8. Make Predictions
9. Evaluate Confusion Matrix and Accuracy
10. Test in Real Time

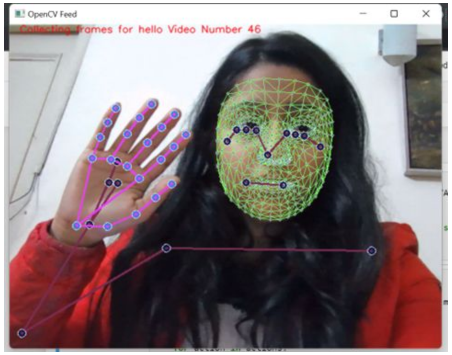
Performance Comparison

We **trained** our model with three continuous sign sentences

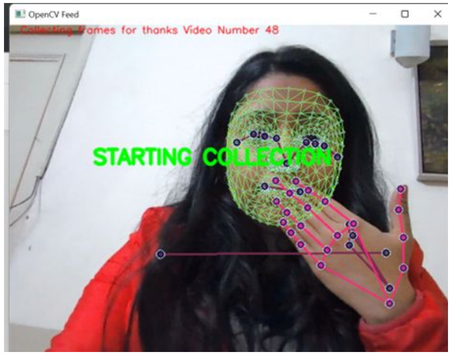
Hello

Thank You

I love you

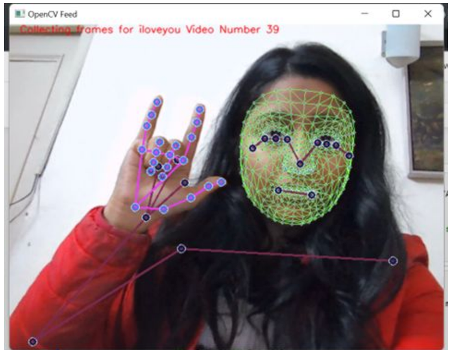


Collecting frames for hello Video Number 46



Collecting frames for thanks Video Number 48

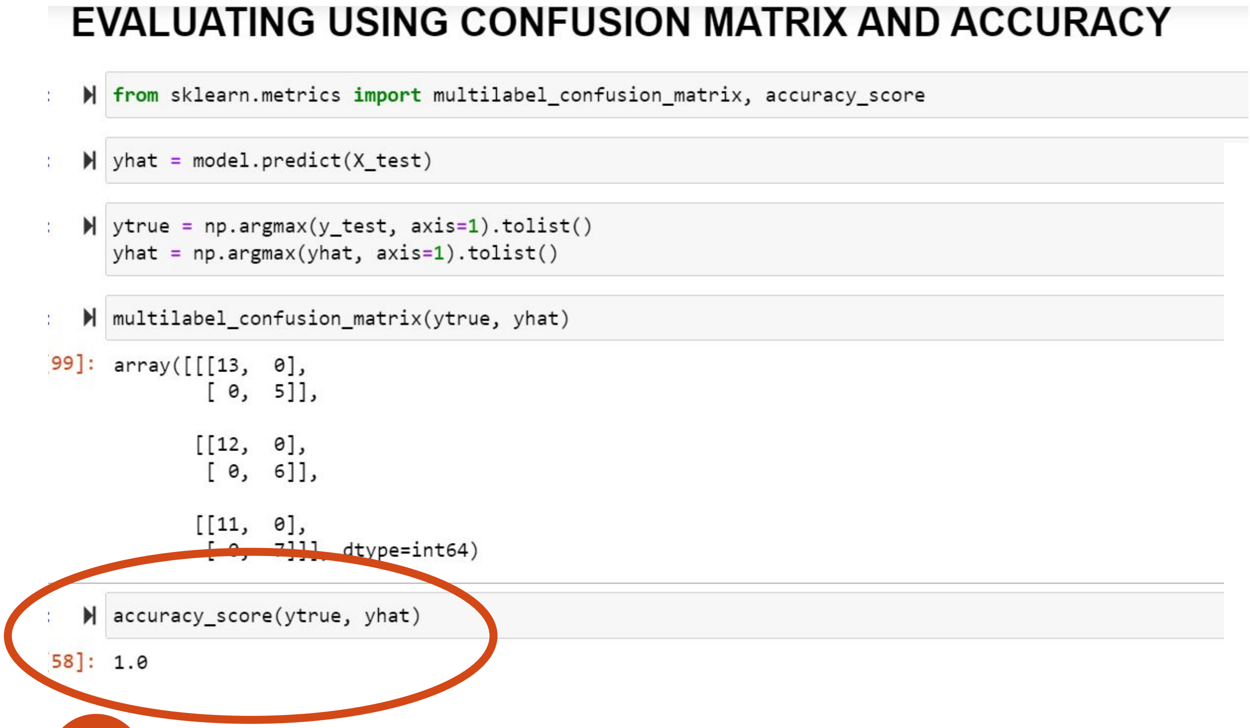
STARTING COLLECTION



Collecting frames for iloveyou Video Number 39

Performance Comparison

The performance comparison of the proposed method with the state-of-the-art methods shows that our method outperforms all other methods in terms of accuracy. Below is the accuracy and confusion matrix to support our claim:



EVALUATING USING CONFUSION MATRIX AND ACCURACY

```
from sklearn.metrics import multilabel_confusion_matrix, accuracy_score
```

```
yhat = model.predict(X_test)
```

```
ytrue = np.argmax(y_test, axis=1).tolist()
yhat = np.argmax(yhat, axis=1).tolist()
```

```
multilabel_confusion_matrix(ytrue, yhat)
```

```
99]: array([[[13,  0],
             [ 0,  5]],

            [[12,  0],
             [ 0,  6]],

            [[11,  0],
             [ 0,  7]]], dtype=int64)
```

```
accuracy_score(ytrue, yhat)
```

```
58]: 1.0
```

Fig 1. Accuracy

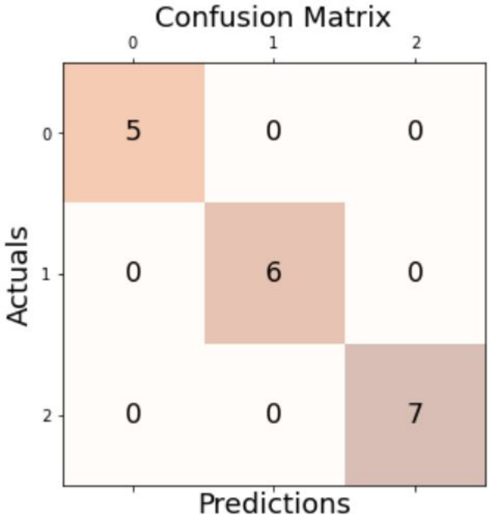
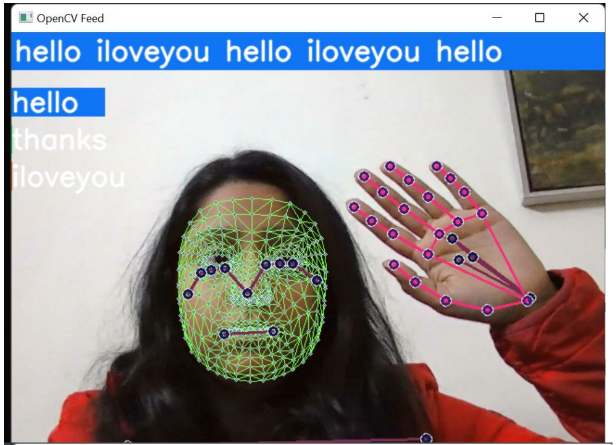


Fig 2. Confusion Matrix

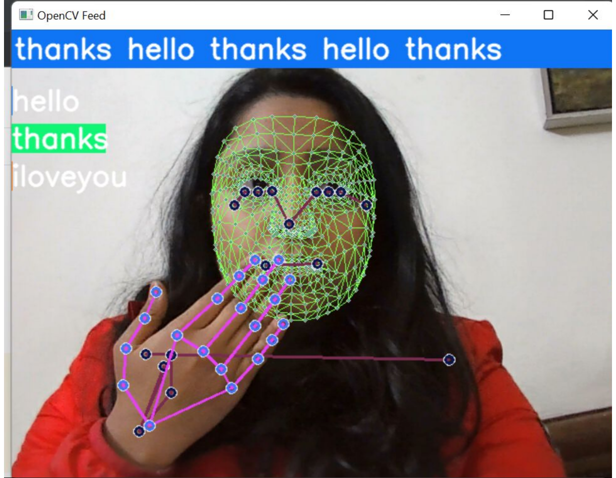
Testing our LSTM model

Table 1. Comparison of the accuracy between the suggested method and the methods in the other studies

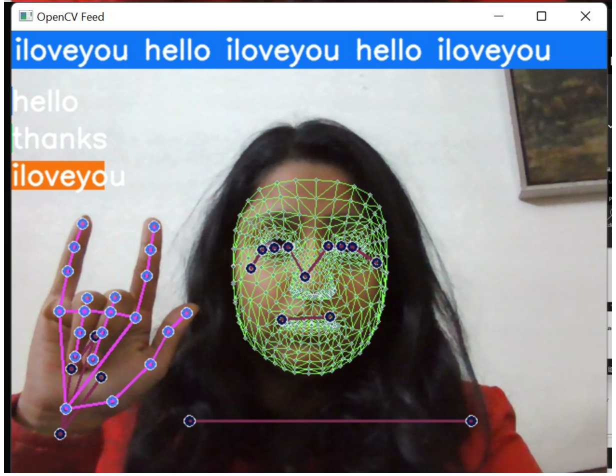
Method	Accuracy
Action Detection (LSTM)	100%
CNN [7]	92.88%
SVM [8]	81.49%
k-means clustering [5]	90.8%
Parallel-HMM [2]	84.48%



```
hello iloveyou hello iloveyou hello
hello
thanks
iloveyou
```



```
thanks hello thanks hello thanks
hello
thanks
iloveyou
```



```
iloveyou hello iloveyou hello iloveyou
hello
thanks
iloveyou
```

FUTURE PROSPECTS

In the future

- the recognition performance can be improved by **increasing more training data** for better model learning, by introducing **more than one sign languages (like BSL, ISL, CSL etc).**
- further research to be conducted for integration with other major **sign language parameters** either manual, such as **movement and place-of-articulation, or facial expressions.**
- an **API** with this model can be deployed which can be **integrated with video conferencing applications** like Zoom, Google Meet etc.

References

[1] B. Bauer and K. Karl-Friedrich. Towards an automatic sign language recognition system using subunits. In International Gesture Workshop, pages 64–75. Springer, 2001. 1, 2

[2] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. Computer Vision and Image Understanding, 81(3):358–384, 2001. 2

[3] H. Li and M. Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. Pattern Recognition, 44(8):1614–1628, 2011. 2

[4] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In ECCV, pages 258–272. Springer, 2010. 2

[5] W. Gao, G. Fang, D. Zhao, and Y. Chen. Transition movement models for large vocabulary continuous sign language recognition. In International Conference on Automatic Face and Gesture Recognition, pages 553–558. IEEE, 2004. 2

[6] N. Tubaiz, T. Shanableh, and K. Assaleh. Glove-based continuous arabic sign language recognition in user-dependent mode. IEEE Transactions on Human-Machine Systems, 45(4):526–533, 2015. 2

[7]Ahmed KASAPBAŞI, Ahmed Eltayeb AHMED ELBUSHRA, Omar AL-HARDANEE, Arif YILMAZ, DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals, Computer Methods and Programs in Biomedicine Update, Volume 2, 2022.

[8] V. Jain, A. Jain, A. Chauhan, S.S. Kotla, A. Gautam American sign language recognition using support vector machine and convolutional neural network Int. J. Inf. Technol., 13 (2021)