A Contemporary Model of Real-Time Indian Sign Language Recognition (ISLR) system using segmentation and image classification

Mrs.P.Priya[1], Dr.P.Kumari[2]

P.G Student[1], Associate Professor [2],

Excel Engineering College, Namakkal, Tamilnadu [1,2]

Abstract:

With the multitude of potential uses, hand signs are the useful method of communication among challenged people and particularly speech impaired people. People with speech impairments use the sign language all over the world for communication. Actually this section of people comprises roughly 1% of Indian citizens. The main justification for incorporating this model that could realize Indian Sign Language would be extremely beneficial to these people. This paper presents a novel technique that uses the Bag of Visual Words model (BOVW) to recognise the Indian sign language alphabets (A–Z) and numbers (0–9) in real-time video streaming. Next, both text and speech output of the expected labels are produced. Segmentation is further performed using background subtraction.

Key words : Bag of Visual Words model , expected labels, background

1.Introduction:

Communication has been essential part of our life. It is fundamental to us to be able to communicate and engage with others. But our perspective and the way interact with others can be very different from those around us due to a variety of factors such as education, society, and so forth. Furthermore, it is crucial to make sure that the intentions of the speaker are fully understood among speech impaired people.

Normal people have little trouble relating to communication and are able to express themselves with ease through speech, writing, gestures, body language, reading, and other common human abilities. It is more challenging for those with speech impairments to interact with others because they are constrained within the boundary of sign language.

Deaf people in India use Indian Sign Language (ISL), a complex and colourful means of communication. It is a visual language that expresses meaning through body language, facial emotions, and hand gestures. The development of technologies capable of recognising and interpreting ISL has garnered increasing attention in recent times. The Indian Sign Language Recognition System (ISLRS) has the potential to improve accessibility and inclusivity in daily interactions by bridging the communication gap between hearing and deaf people. India is a diversified country that is home to around 17.7% of the world's population, yet compared to other countries, it has produced comparatively little study in this sector, and it is highly

inconsistent [1-3]. There is proof of delayed standardisation to corroborate this. The first studies on Indian Sign Language are conducted in India in 1978. However, due to the lack of a uniform form, ISL could only be utilised in short-term courses. In addition, only around 5 percent of the deaf community attended these schools, and most of these establishments used very different gestures. ISL was standardised in 2003, and following standardisation, research intensifies [4]. Indian Sign Language (ISL) has both static and dynamic signs, single- and double-handed signs, and numerous signs for the same alphabet in different regions of the nation. This makes the introduction of such a plan extremely difficult. Furthermore, there's no shared dataset available.

The two primary techniques that are frequently employed in sign language recognition are sensor-based and vision-based [5]. Web cameras are used to take pictures or videos. Vision-based gesture detection has gained appeal due to its spontaneity and lack of specialised hardware needs, according to its proponents [6]. Nonetheless, segmentation of the hands in a complex environment is crucial for identification. Thus, there is a framework that can get around this issue. The current study proposes an approach to build a big, diversified, and dependable real-time system for the recognition of Indian Sign Language (ISL) digits (0-9) and alphabets (A-Z). The system has identified signs that are in front of the camera using photos that are retrieved through a webcam.

2.Literature survey

Eigen value-weighted Euclidean distance was used by I. J. Singha et al. [7] to classify signs in a real-time recognition system. In order to classify the indicators, H. P. Kishore et al. [8] introduced a recognition method that makes use of artificial neural networks (ANNs) to recognise active contours from boundary edge maps [9]. The work done by Otsu is based on the algorithm yielded a fairly high rate of accuracy [10]. Segmentation is the first and most crucial step in manual processing. An effort was made to bypass the initialization and segmentation steps by using the moving block distance parameterization technique [11]. In this paper, 33 elementary word units and highly accurate static symbols were employed.

Pattern recognition, feature extraction, and similar techniques formed the basis of the majority of these works [12]. But, a single feature is insufficient for most of the recognition systems. Therefore, in order to address this issue, hybrid approaches were implemented. A.Nandy et.al. orientated histogram features to classify gestures using hybrid approaches combining Euclidean distance and K-Nearest Neighbor (KNN). Poor performance was evidenced in cases of similar gestures. Manjushree K.et.al. used feature matching and a histogram of oriented gradients and that are employed with single-handed sign classification. S. Kanade et.al. created a system based on SVM and PCA features, with a custom dataset and achieved accurate results. Both Single-handed and double handed signs could be recognized in ISL, according to Sahoo. In contrast, the B-Spline approximation was used in the system put forth by Geetha M et al. to match the shapes of the static gestures that represented the ISL alphabets and numbers. Natural language processing (NLP) technology and a neuro-fuzzy

technique were presented in article [18] as a way to categorise word symbols and show the finished word. As of [19], AdaBoost algorithm was presented a technique for hand gesture recognition. Complete gesture recognition can be achieved by using arbitrary context-free grammars.

Higher calculation accuracy was obtained using the PCA and local coordinate system combination, outperformed in the algorithm-based approach. To solve this issue more quickly for real-time systems, image recognition using different models can now be automated due to the recent advancements in Deep Learning technologies. In works proposed in [21,22], deep learning has proved significantly by using convolutional neural networks(CNN). Once the features in the photos were extracted by a CNN (Convolutional Neural Network), Jayadeep et al.[23] used an LSTM (Long Short Term Memory) to classify and translate these motions into text. The InceptionV3 model, as proposed in paper [24], proposed to use Motion segmentation and feature extraction with depth sensors to identify static indications. A methodology for categorising photos for each American Sign Language letter and digit (0–9) using deep convolutional neural networks and a mini-batch supervised learning technique of stochastic gradient descent was described by Vivek Bheda et al. [25]. After a thorough analysis of all of these studies, the suggested approach seeks to create a unique dataset and a highly effective algorithm for accurate video detection.

3.Proposed work

Fig. 1 shows the data flows for the several stages of sign language recognition, i.e., data set, image capture, feature extraction, data pre-processing, and sign classification. Fig. 2 depicts the overall system architecture of the suggested system.

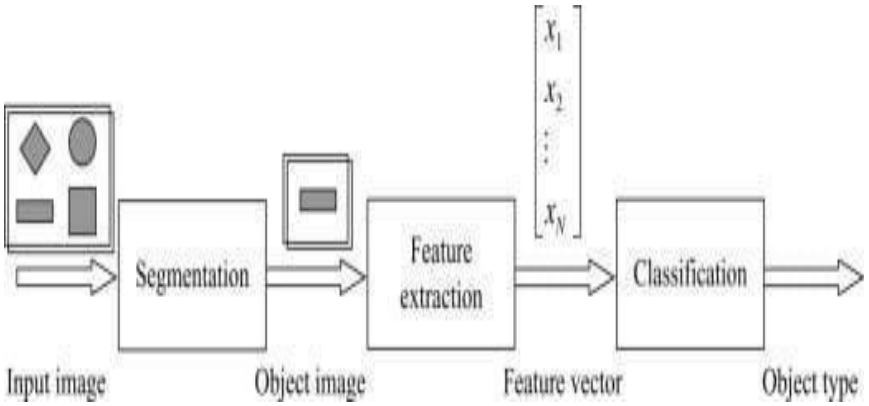

Fig1.1: Phases of sign language recognition

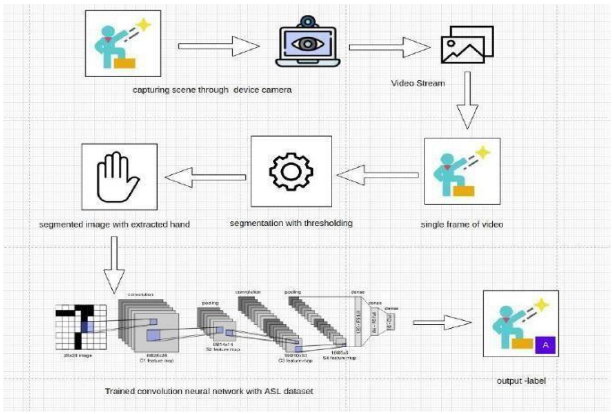

Fig.2 Process of the proposed work

4.Dataset collection

Dataset collection is a vital component of research projects in all fields since it is essential to the advancement of deep learning and machine learning models. The lack of standard datasets for Indian sign language presented our largest obstacle during the data collection process. So, as part of this work, a customized dataset was used. First of all, a webcam is used to record the videos, for getting different signs. The inputs ranges from 26 distinct alphabets (A-Z) and 10 numerical signs (0–9) for consideration. This is shown in Fig.3.

Fig 3: ISL signs used for training

The camera's position is critically important for both picture quality and background noise reduction. There were two methods for taking the pictures in order to add diversity to the dataset. The first technique is the standard method which can be applied to an image with a plain color background and second method needs to segment the skin on it.

5.Comprehending aspect ratios

The magnitude of width and height of the image plays crucial role in detection, which describes the shape of an image. The formula for calculating aspect ratios is width to height. Even if the

image were 500 pixels by 500 pixels or 1500 pixels by 1500 pixels, the aspect ratio would remain at 1:1. An image in the portrait orientation, for instance, might have a ratio of 2:3. The height is 1.5 times longer than the width when the aspect ratio is this one. Therefore, the image's dimensions could be 500 × 750, 1500 × 2250, etc. While Cropping an image to an angle along with using the built-in style options, must be done manually by crop an image to an angle. Use the crop tool to select from the available aspect ratios after launching the editor. Another option is to customize the dimensions by using a third-party editor to crop images to a custom aspect ratio that our built-in image editor does not support.

6.Preprocessing

For the sake of keeping the aspect ratio constant, all images have the same dimensions. The default setting options are utilised to transform the captured video frame into the HSV colour space for images taken against plain backgrounds. The colour contrast from the background makes it easy to distinguish the skin tone. After that, the frame is exposed to an experimental threshold value that establishes the image's hue value and eliminates any pixels that correspond to the colour of skin. This phase involves getting the image ready for feature extraction and detection. To ensure consistency in aspect ratio, all images have the same dimensions. Because of the background's differing colour, the skin tone may be plainly distinguished. The frame is then exposed to an experimental threshold value, which establishes the image's hue value and eliminates the skin-colored pixels. In order to generate a mask during the hand segmentation procedure, the hand's most linked region must be removed. To further eliminate noise, morphological processes like dilation and erosion are employed.

7.Build a Bag of Visual Words

The procedure of feature extraction, feature clustering, codebook creation for the model, and histogram generation can be used to create a bag of visual words (BOVW). NLP's (Natural Language Processing) Bag of Words (BOW) and data retrieval are the foundations of the Bag of Visual Words (BOVW) picture categorization model [26]. Here, the frequency of keywords is used to account for the presence of each word in a text, and a frequency histogram is created using the collected data. By using the image's attributes as words instead of words, this concept is changed. A vocabulary is constructed where each image is represented as a frequency histogram of the acquired features using the picture descriptors and key points. This frequency histogram can be used to determine the type of future image that is similar.

The features that are obtained are then clustered, which is the next stage in the feature extraction process. The K-means algorithm can be used for clustering, because the data is high volume, mini batch K-means method was chosen in proposed work. The whole set of data does not need to be in memory at once because the method works by employing small random batches of fixed-size data at a time. Until convergence, this process is maintained. The value of k has always been

180.

8.Classification

After the feature extraction and detection phase is completed, the classification phase gets initiated. This supervised model is capable of handling both linear and non-linear problems in the domains of regression and classification, that works by using the concept of decision planes, define decision boundaries. SVM with a linear kernel has been used for this classification. In order to classify and recognize ISL signs, SVM the visual word histograms have fed as feature vectors. The training process uses a total of 28,800 photos. The testing set, which totals 7236 photos, is used to evaluate the classifier's performance. The accuracy, precision, recall, and other metrics are used to gauge the classifier's performance. The customised dataset used for training is displayed in Fig.4

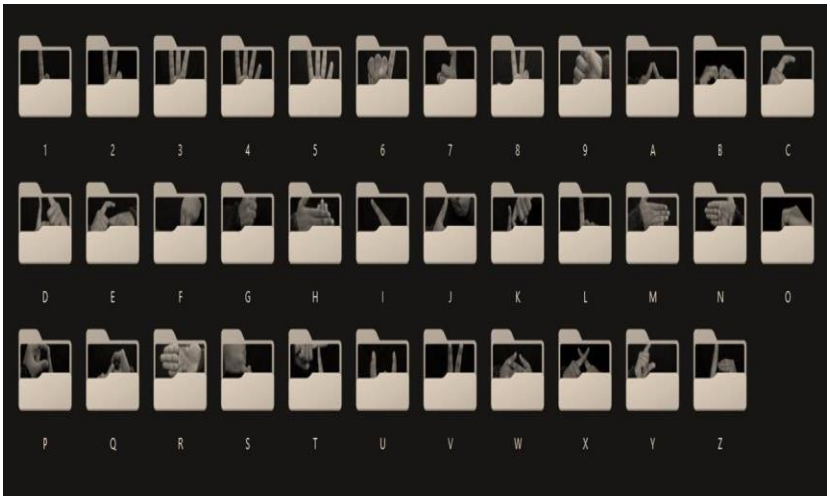

Fig.4 Data set using for training

The visual cortex of the human brain served as the model for functional extraction used by convolutional neural networks. When a filter map is applied to specific areas of an image, CNNs analyze the images piece by piece. These segments are referred to as features, and they are used to compare two images by identifying nearly identical features at nearly identical locations. When the image recognition and classification are taken into consideration, CNNs outperforms than the other neural networks. The overall architecture of the proposed system consists of several convolutional and dense, is a fairly typical CNN architecture and is shown in Fig. 5

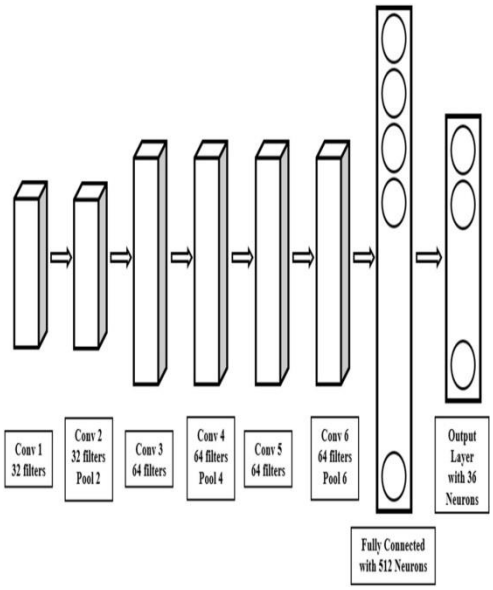

Fig.5. CNN Architecture

9.Output sign

The technology automatically translates text and speech based on predicted class labels that are returned as numerical vectors. The main advantages of this approach are its ease of use and improved communication. After the label has been recognised by the classifier, it is given as a key to a dictionary, which returns the sign that corresponds to it as value. This is then shown to the user. Text to voice is produced using the Python text to speech library PyttsX3. Threading is used because it causes the live video stream to lag by processing frames slowly. This is accomplished by simultaneously doing speech-to-text translation and sign prediction. This ensures that the sound will always be continuously played. The suggested system's snapshots are displayed in Fig. 6.

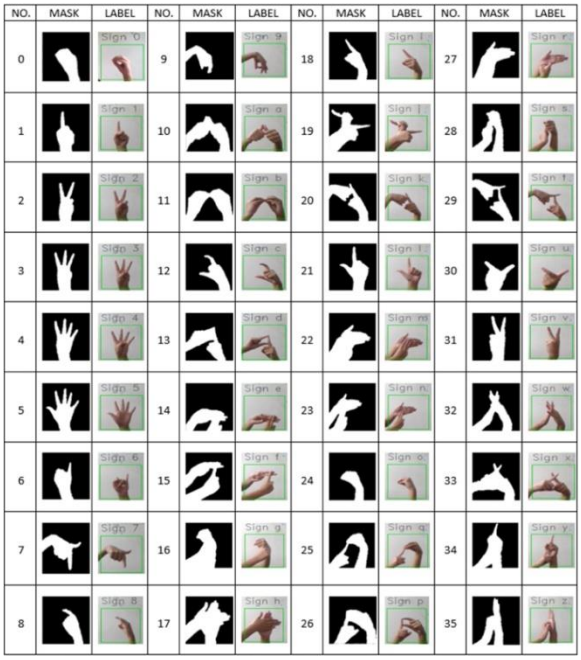

Fig.6 Snapshot of the proposed system

10.Results and Outcomes

The dataset is divided into two sets: the training set contains 80% of the total data, while the testing set contains the remaining 20%. The two classifiers that are utilised to determine which image has the highest accuracy are SVM and CNN. CNN, on the other hand, has performed better with less features. The system has been trained to recognise 36 signs, which consist of 26 alphabets and 10 digits. While there is always space for improvement, the present outcomes are promising.

10.1 Performance of SVM: With the testing data sets provided, SVM yielded a 99.14% accuracy rate. high level of overall accuracy as determined by the SVM-classified alphabet and digit computed precision and recall scores. By class, the accuracy is shown in Table 1.

CNN broadcast: An overall accuracy of 94% is attained on the training set during the latest epoch, while testing accuracy exceeded 99%. In total, there are fifty epochs. The softmax function was used as the activation function and the categorical cross entropy loss function was used to train the suggested model. The ultimate era produced testing losses of 0.0184 and training losses of 0.1748 as a result. Table.1 shows each sign's accuracy. The accuracy graph for the suggested system is shown in Fig. 7. Fig 8 shows the loss graph.

Table.1 Accuracy of the proposed system

Label	SVM %	CNN %	Label	SVM %	CNN %
0	100	100	I	100	100
1	100	100	J	100	100
2	100	100	K	96	100
3	98	100	L	100	97
4	100	100	M	100	100
5	100	100	N	100	100
6	98	100	O	97	98
7	100	99	P	100	100
8	98	100	Q	100	100
9	100	100	R	100	100
A	100	100	S	94	99
B	98	100	T	100	100
C	100	97	U	100	100
D	100	100	V	100	100
E	98	100	W	95	100
F	100	97	X	100	99
G	100	100	Y	100	100
H	98	100	Z	96	100

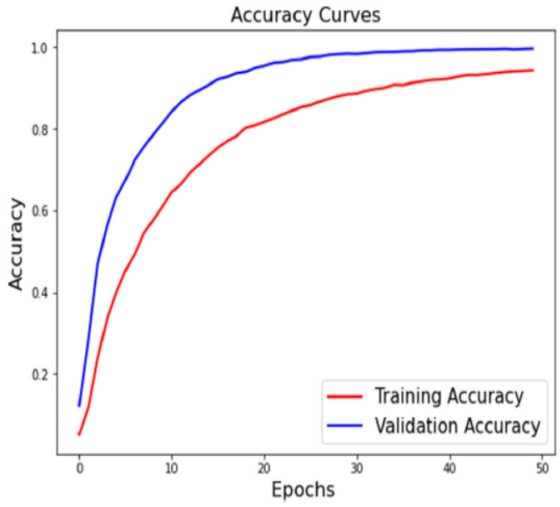

Fig.7 Accuracy rate using CNN

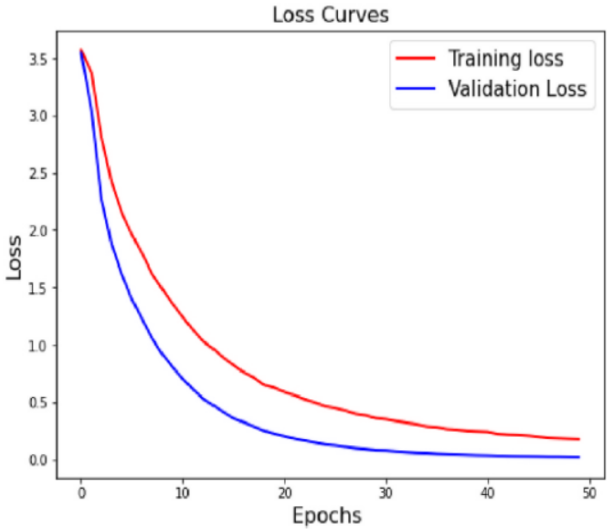

Fig.8 Loss graph of the proposed system using CNN

11.Conclusion and future work

The primary objective of the proposed system is to develop an enhanced real-time recognition utility that can be utilised anyplace. It is accomplished by creating a unique data set, solving the background dependence issue, and making the system rotationally invariant. With 99 percent accuracy, the system has been successfully trained on all 36 ISL static alphabets and digits. In the future, more signs in other languages spoken in different nations can be added to the collection, making the framework more beneficial for real-time applications. Using this method to form basic phrases and expressions can be used for both isolated and continuous recognition tasks. Reaction time acceleration is crucial for real-time applications.

References:

1. Aziz, R., Banerjee, S., Bouzefrane, S., & Vinh, T. L. (2023). Exploring Homomorphic Encryption and Differential Privacy Techniques towards Secure Federated Learning Paradigm. *Future Internet*, *15*(9), 310
2. Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., & Ryo, M. (2023). Ten deep learning techniques to address small data problems with remote sensing. *EarthArXiv (California Digital Library)*. Khetavath, S., Sendhilkumar, N. C., Mukunthan, P., Jana, S., Gopalakrishnan, S., Malliga, L., Chand, S. R., & Farhaoui, Y. (2023). An intelligent heuristic Manta-Ray foraging optimization and adaptive extreme learning machine for hand gesture image recognition. *Big Data Mining and Analytics*, *6*(3), 321–335.
3. Parthasarathy, N. S., & Yogesh, P. (2023). Novel video benchmark dataset generation and Real-Time recognition of symbolic hand gestures in Indian dance applying deep learning techniques. Journal on Computing and Cultural Heritage, 16(3), 1–19.

4. Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. https://doi.org/10.1109/msp.2017.2738401
5. Eid, A. H., & Schwenker, F. (2023). Visual static hand gesture recognition using convolutional neural network. *Algorithms*, *16*(8), 361.
6. Bantupalli Kshitij, Xie Ying. American sign language recognition using machine learning and computer vision. Master of Science in Computer Science Theses 2019; 21.
7. Shadman Shahriar, Ashraf Siddiquee, Tanveerul Islam, Abesh Ghosh, Rajat Chakraborty, Asir Intisar Khan, Celia Shahnaz and Shaikh Anowarul Fattah. Real- time American sign language recognition using skin segmentation and image category classification with convolutional neural network and deep learning. In TENCON, IEEE Region 10 International Conference.
8. Shivashankara S, Srinath S. A comparative study of various techniques and outcomes of recognizing American sign language: a review. In: International Journal of Scientific Research Engineering & Technology (IJSRET); 2017. ISSN 2278 – 0882. 6(9).
9. Viswanathan Daleesha M, Idicula Sumam Mary. Recent developments in Indian sign language recognition: an analysis. Int J Comput Sci Inf Technol 2015;6(1): 289–93.
10. Nair Anuja V, Bindu V. A review on Indian sign language recognition. Int J Comput Appl 2013;73(22).
11. Athira K, Sruthi CJ, Lijiya A. A signer independent sign language recognition with Co-articulation elimination from live videos: an Indian scenario. J King Saud Univ Comput Inf Sci 2022;34(3):771–8.
12. Singha J, Das K. Recognition of Indian sign language in live video. Int J Comput Appl 2013;70(19):17–22.
13. Kishore PVV, Kumar DA. Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. In: IEEE 6th international conference on advanced Computing; 2016.
14. Swamy Shanmukha, Chethan MP, Gatwadi Mahantesh. Indian sign language interpreter with android implementation. Int J Comput Appl 2014:975–8887.
15. Agrawal SC, Jalal AS, Bhatnagar C, Ieee. Recognition of Indian sign language using feature Fusion. 2012.