INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Performance Evaluation of Feature Fusion Approaches for Indian Sign Language Recognition System

B V Poornima[1]*, S Srinath[1], S Rashmi[1], R Rakshitha[1]

1 Department of CSE, SJCE, JSSSTU, Mysore, Karnataka

OPEN ACCESS

Received: 14-07-2023

Accepted: 12-09-2023

Published: 12-11-2023

Citation: Poornima BV, Srinath S, Rashmi S, Rakshitha R (2023) Performance Evaluation of Feature Fusion Approaches for Indian Sign Language Recognition System. Indian Journal of Science and Technology 16(41): 3691-3703. https://doi.org/10.17485/IJST/v16i41.1767

* **Corresponding author**.

poornimabv.85@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To recognize and analyze the Indian sign language (ISL) gestures in simple background using various features and Machine learning classifiers. **Methods:** a) Data pre-processing: Contour matching approach for hand region segmentation b) Feature extraction: Local features like gradient & key point descriptors were extracted using HOG, SIFT, SURF, LBP, FAST, feature fusion is done by concatenating features of HOG with LBP, SIFT with FAST, BOVW model with SURF. c) Model development: SVM, Random Forest, Logistic regression, Naïve Bayes were trained on large dataset and was experimented with hyper parameter tuning. The experiment was performed on 2 standard datasets which consist of alpha numerals (A-Z & 1-9) in simple black background. **Findings:** Our research demonstrates that our model consistently achieved a remarkable 100% accuracy rate when utilizing feature fusion techniques on both of the datasets employed in this study. The research findings underscore the need to consider a more comprehensive approach for gesture recognition. Relying solely on distinct features extracted from a single algorithm is shown to be insufficient in addressing the challenges posed by the diverse nature of sign shapes, varying illumination conditions, and different orientations. This emphasizes the importance of exploring hybrid or multi-algorithmic strategies to enhance the accuracy and robustness of gesture recognition systems. **Novelty:** This research introduces a novel perspective on ISL gesture recognition by emphasizing gestures against a simple black background. The innovative application of feature fusion, combining various elements like hand shape and orientation enhances accuracy. The inclusion of hand segmentation using contour matching algorithm and experimentation on benchmark data adds another layer of novelty, highlighting practical applicability. The experimental results show that SVM has given better results when used with different combinations of feature extractors.

Keywords: ISL; Feature Fusion; Keypoint Descriptors; HOG; SIFT; SURF; LBP; FAST

1 Introduction

Sign language[1], a visual mode of communication for the deaf or hard-of-hearing, utilizes intricate hand motions, facial expressions, and bodily cues. Capitalizing on machine learning's potential, automated recognition of sign language gestures is a promising avenue with implications spanning assistive tech, education, and communication realms. Our focus centers on the intricacies of ISL[2], a rich linguistic entity imbued with distinct syntax, grammar, and vocabulary. We propose harnessing machine learning algorithms to discern ISL gestures, emphasizing a simplified dataset featuring signs against a uniform black backdrop. This strategic dataset choice inherently diminishes the requisite intricacy for capturing underlying patterns. Despite the advantageous prowess of deep learning models in discerning complex patterns in vast datasets, for our specific context, the simplicity of machine learning models circumvents the need for deep learning's elevated complexity and computational demands. Through our methodological innovation, we endeavor to rectify extant shortcomings, yielding heightened recognition accuracy and efficiency. Our study strives to engineer a steadfast recognition system tailored for ISL alphabets (A-Z) and digits (1-9), emphasizing image-based sign recognition techniques to bypass reliance on sophisticated equipment like gloves or the Kinect. Our objectives encompass locating the region of interest (ROI), extracting manifold features from this ROI, and achieving sign recognition through diverse classifiers. This research focuses on a big advancement: combining different aspects like gradient details, key descriptors, texture features, and corner points. These are all important in understanding sign language gestures, which can have various shapes and angles & orientations. This makes recognizing each gesture accurately very hard. So, the main goal of this study is to find a good way to pick out these important aspects in a smart way. By blending all these methods together, we hope to improve how well we can tell apart different signs in sign language. This work overcomes drawback of traditional feature extraction techniques. Though they perform exceptionally well in one situation but may underperform in other situations, as they are intended to extract specific features from an image. Since sign language gestures have different shape and orientation and is not fixed to a certain feature, recognizing each sign with good accuracy is a challenging task. Hence this research work aims to extract the features in an effective way which helps in achieving accurate recognition.

SLR holds significance across various fields beyond its role in assistive technology. It notably contributes to education, communication improvement, and the broader landscape of societal integration. Here are the key highlights.

- **Education:** SLR technology can be leveraged to develop interactive educational tools and resources, including sign language learning applications and interactive sign language textbooks.
- **Communication services:** Real-time sign language interpretation services can be implemented using sign language recognition algorithms. These services can be integrated into various communication platforms, such as video conferencing software and customer service chatbots, enabling seamless communication between deaf and hearing individuals. In healthcare settings, it can facilitate accurate communication between healthcare providers and patients who use sign language.
- **Entertainment:** It can be integrated into entertainment platforms, allowing users to enjoy multimedia content like movies, television shows, and live performances with sign language interpretation.
- **Accessibility in public spaces:** Public spaces, including transportation hubs and government buildings, can implement SLR systems to provide visual information, emergency alerts, and announcements in sign language. This ensures that vital information is accessible to all individuals, including those who rely on sign language for communication.
- **Gesture-based interfaces:** Beyond sign languages, gesture recognition technology can be employed to create gesture-based interfaces for controlling devices, playing video games, and interacting with digital systems. This expands the scope of accessible technology and enhances the user experience for individuals with various communication needs.
- **Accessibility in smart homes:** Smart home systems can integrate SLR to allow deaf individuals to control home appliances, lighting, security systems, and other devices through sign language commands, promoting independent living and accessibility.

In[3], real-time SLR system, utilizing Fuzzy Cognitive Maps (FCM), has achieved an impressive accuracy rate of 75% in accurately identifying gestures. The system focuses on recognizing words and has demonstrated the capability to promptly recognize 40 distinct ISL words in real-time. Compared to alternative clustering algorithms, FCM has proven its efficiency and reliability in various applications, showcasing superior performance. While the system yielded promising results with an accuracy of 75% in real-time recognition of 40 words from Indian Sign Language using Fuzzy Cognitive Maps, a limitation lies in the relatively small training dataset comprising only eight samples per sign, and testing on just two samples, which could potentially impact its generalization and robustness. In[4], the MediaPipe framework facilitates data preprocessing and feature extraction by capturing facial, hand, and body key points from webcam input frames. These features undergo data augmentation.

Key points from the first stage are stored, enabling removal of null entries before data labeling. In the third stage, our MOPGRU model trains and classifies labeled gestures, converting them into on-screen text. Despite its accurate and efficient performance, it's crucial to acknowledge that the study is constrained by a limited dataset. One drawback of the MediaPipe framework is that its predefined key points and landmarks might not cover all possible variations of gestures or poses, leading to limitations in accurately representing certain complex movements or positions. In their study, Ms. Greeshma Pala et al[5] conducted a comparative analysis of KNN, SVM, and CNN algorithms to determine the most accurate approach. They utilized approximately 29,000 images, which were divided into test and train data sets and preprocessed for compatibility with the KNN, SVM, and CNN models. The obtained accuracies were 93.83% for KNN, 88.89% for SVM, and 98.49% for CNN, respectively. Based on these results, the CNN algorithm demonstrated the highest accuracy among the three algorithms studied. In their paper, Ashish Sharma et al[6] introduced a technique for feature extraction using ORB (Oriented FAST and Rotated BRIEF) and evaluated its performance against other pre-processing techniques such as Histogram of Gradients (HOG), Local Binary Patterns (LBP), and Principal Component Analysis (PCA) on the same dataset. The results showed that the proposed ORB feature extraction technique outperformed the other pre-processing techniques when used with Naïve Bayes, Logistic Regression, and KNN classifiers. On the other hand, PCA performed better than the other techniques when used with MLP, Random Forest, and SVM classifiers. These findings highlight the effectiveness of the proposed ORB technique and PCA in improving the performance of specific classifiers in the context of the study. In their paper, Rajesh B. Mapari et al[7] proposed an approach for classifying alpha-numeric characters in the Indian Sign Language using an RGB camera. The method involved extracting features from the images, including the Discrete Cosine Transform (DCT) of the grayscale image and regional properties of black and white images. These features formed a feature vector with 74 values. The dataset used in the study consisted of 33 signs performed by 60 signers, resulting in a total of 1980 signs. Various Neural Network classifiers, including MLP, GFFNN, and SVM, were trained and tested on the dataset. The highest classification accuracy achieved was 86.27% on the cross-validation dataset using the MLP Neural Network. Dhivyasri S et al.[8] concluded that the combination of SVM classifier with K-means clustering and Bag of Visual Words (BoV) classifiers is the most suitable for gesture recognition. Based on this finding, they developed a user-friendly application capable of interpreting ISL. The application utilizes the highly efficient SVM classifier for gesture-to-text conversion, while Google Speech Recognition API is employed for speech to gesture conversion. By integrating these technologies, the application offers a comprehensive solution for bridging the communication gap between sign language and spoken language. Purva Chaitanya Badhe et al.[9] proposed an approach for gesture classification that involves handcrafted feature extraction techniques and utilizes an Artificial Neural Network (ANN) for classification. The achieved model accuracy is remarkably high, reaching 98%. In this approach, Fourier descriptors are applied to the pre-processed gestures as part of the feature extraction process. The classification is performed using a multi-class neural network classifier, which demonstrates a training accuracy of 98% and a validation accuracy of 63%. For the purpose of feature extraction, 28 Fourier descriptors are extracted per frame of the video gesture after conducting multiple trials. These 28x28 Fourier descriptors are quantized using Vector Quantization (VQ) technique with the aid of a created codebook. Vector quantization is a non-uniform and many-to-one mapping lossy compression method. This approach showcases the effectiveness of handcrafted feature extraction and ANN-based classification in achieving high accuracy in gesture recognition tasks. Amrutha K et al.[10] developed a vision-based isolated hand gesture detection and recognition model. The performance of the machine learning-based Sign Language Recognition (SLR) model was evaluated using four candidates in a controlled environment. The model employed a convex hull for feature extraction and utilized the KNN algorithm for classification. The model achieved an accuracy of 65%.

1.1 Discussion on related work

Following observations are made from the literature review:

Limited scope of gestures: Some studies have focused solely on alphabetic gestures, omitting crucial signs from their experiments, which impacts the system's overall applicability.

Preprocessing challenges: The application of minimal preprocessing on non-uniform background images can lead to increased computation time and potentially compromise accurate hand region extraction, which is vital in SLR. Explicit hand segmentation is often necessary for machine learning approaches on non-uniform backgrounds.

Feature Extraction: Several studies do not comprehensively highlight their feature extraction techniques. This lack of transparency can hinder the reproducibility and effectiveness of their methods.

Algorithm limitations: The use of 2D DCT may not effectively handle complex scenarios or diverse backgrounds, potentially leading to inaccuracies and reduced performance in real-world environments.

Insufficient dataset: Few researchers have utilized a relatively small dataset, limiting the robustness and generalizability of their findings. The exclusion of certain signs further diminishes the applicability of their results.

Discriminative power: Relying solely on a single feature extraction algorithm may not suffice for challenging sign language recognition (SLR) scenarios.

Limited spatial information: The application of fourier descriptors which mainly captures the shape characteristics, can be restrictive in distinguishing objects with similar shapes but differing textures, potentially affecting the accuracy of the system.

1.2 Research Gap

The existing approaches highlighted in our paper clearly indicates that particular feature from a specific feature extractor algorithm is not sufficient to recognize ISL gestures as the signs will have different shapes, illumination and angles.

The empirical evidence presented in our paper underscores the insufficiency of relying solely on distinct features extracted from a specific algorithm for effective recognition of gestures. This inadequacy stems from the inherent variability in the shapes, illumination conditions, and orientations of the signs.

1.3 Contribution of the proposed work

Efficient way of extracting the region of interest i.e. the hand region by using contour matching approach which reduces the computational cost.

Worked on different features and combined the features which is the primary contribution of the proposed work to efficiently recognize the ISL signs.

2 Methodology

In comparison to existing approaches in the literature, our proposed system introduces a novel and impactful methodology for SLR. While many prior methods tend to rely on individual features or simplistic combinations, our approach represents a paradigm shift by integrating a comprehensive range of features including gradient magnitude and orientation, key descriptors, local features, and corner detectors. This holistic feature fusion enables our model to capture the intricate variations in sign gestures that are often missed by traditional techniques. Unlike conventional fixed-feature models, our system adapts flexibly to the diverse shapes and orientations intrinsic to sign language gestures, leading to heightened accuracy in recognition. This departure from single-feature-centric approaches aligns with the complex nature of sign language and sets our method apart as a promising solution for robust and accurate gesture recognition.

The experiment is conducted on the Kaggle dataset (Dataset-I) which consists of 35 gestures (1-9 digits and A-Z alphabets), each gesture having 1200 RGB images which is of the size 128*128 and it has a total of 42000 images. The second set of data considered is the dataset published by shagun katoch (Dataset II)[11]. The dataset has images in different hand orientations. It consists of digits (1-9) and the alphabets(A-Z), each gesture having 1000 RGB images where each image is of the size 225*225. The total volume of the dataset-II is 36000. Figure 1 shows the alpha-numeric chart of Dataset-I and Dataset-II captured in the uniform black background. Certain pre-processing activity has been done on the datasets which further reduces the size of the images and it has been explain in the section 2.2.

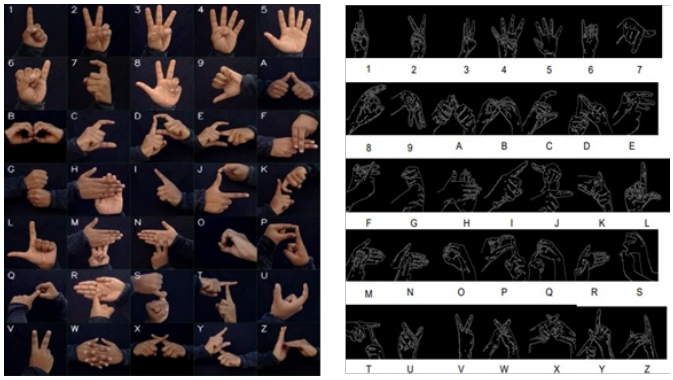

Fig 1. Alphanumeric chart of the benchmark Datasets I & II

2.1 Hardware/Software requirement

- Hardware configuration

GPU: We conducted our experiments on a system equipped with an NVIDIA GeForce GTX1650 GPU with 4GB memory.

RAM: The system is equipped with 8GB of DDR4 RAM.

- Software Environment

Operating System: We conducted our experiments on a machine running Windows 10.

Python Version: Python3.8 was used for all the experiments. Jupyter Notebook: We employed Jupyter Notebook, version 6.1.4, as the primary development environment for code prototyping, data analysis, and experimentation.

2.2 Pre-processing

The first and most crucial phase in the SLR system is the hand segmentation. Segmentation reduces the processing time and increases the precision of recognizing the signs.

Contour matching: It is applied to the images to detect and extract the contour of the hand region. Contours are continuous curves that represent the boundaries of objects in an image. This algorithm identifies the contour indicating the presence of ROI in an image. This technique is applied on Dataset-I. The contour matching algorithm [12] works well on black background images and so the foreground objects can be extracted accurately. This approach works well with binary images because edge clarity will be good as it has only 0 or 255 pixel values. In other words, finding contours is like finding a white object from a black background. Figure 2 shows the ROI obtained after applying the contour matching algorithm.

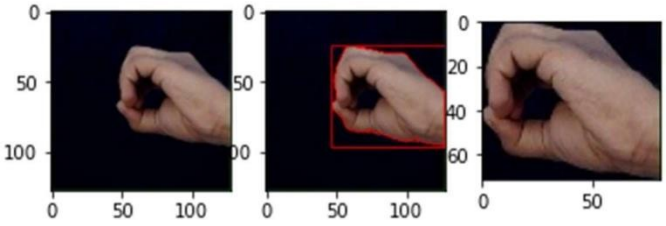

Fig 2. Extracted ROI

Cropping equally on all the sides: After obtaining the edge-detected image by using canny edge detection algorithm [13], cropping is performed to remove the unwanted regions and focus solely on the ROI. Starting from the edges of the image, the cropping algorithm traverses towards the centre in each direction (top, bottom, left, and right) until it encounters a region of white pixels. The white pixel region indicates the boundary of the hand, and the algorithm crops the image to exclude the unwanted area. This technique is applied on Dataset –II.

2.3 Feature extraction

This section explains about the different feature extractors and the fusion of features employed for the SLR. Our work is based on local features like texture, key points, gradient-based features.

a) Histogram of oriented gradients (HOG) extracts contour or edge features and texture features by calculating the distribution of gradient orientations within an image. It captures local variations in intensity and is often used in object detection and texture analysis. It has several benefits in SLR: Robust to variations in appearance, captures shape and motion information, localized feature representation, invariant to translation and scaling computational efficiency. Table 1 shows the variation in the parameters of HOG [14] used in the experiment.

Algorithm

1. Data Preparation
2. For each image compute the gradient magnitudes and orientations
3. For each cell, gradients in X and Y coordinates are calculated Eqs. (1) and (2).

Hx and Hy are derivatives in X and Y coordinates, respectively.

$$\text{Magnitude, } M = \text{sqrt}(Hx + Hy2) \tag{1}$$

$$\text{Orientation}, \theta = \tan -1(\text{Hx}/\text{Hy}) \quad (2)$$

4. Divide the image into small cells, e.g., 8x8 pixels, and compute the gradient magnitudes and orientations within each cell.

5. Create a histogram of gradient orientations for each cell by binning the orientations into predefined angular bins, e.g., 9 bins ranging from 0 to 180 degrees.

6. Concatenate the histograms from all the cells in the image to form a feature vector for that image.

Table 1. Different levels of HOG parameters

S.No.	Parameters(Control Factors)	Levels		
		1	2	3
1	Image Size	50*99	110*50	225*225
2	Bin Size	6	9	12
3	Cell Size	6*6	8*8	10*10
4	Block size	2*2	2*2	2*2

b) Scale-Invariant Feature Transform (SIFT)

The algorithm is applied to images to identify key points [15], which are useful to represent the sign language gestures. Figure 3 shows the images with the detected key points. The following are the steps to find out the key point descriptors in an image.

Scale-space extrema detection: SIFT looks for key points or interest points in the image that are invariant to scale changes.

Key point localization: SIFT refines the key points obtained by localizing. It uses a process called ”subpixel refinement” to accurately locate the key points based on the gradients and curvature of the intensity values around the key points.

Orientation assignment: SIFT assigns an orientation to each key point to make it invariant to rotation.

Key point descriptor: SIFT generates a descriptor for each key point that encodes the local appearance and gradient information around the key point.

Fig 3. SIFT Key points

The input image is converted to grayscale, key points are detected to compute the descriptors. The crucial parameters for SIFT in our study were:Sigma (Gaussian Blur): We carefully set this parameter to 2.0. Sigma controls the extent of Gaussian blurring applied to the image. It plays a pivotal role in determining the scale size and dimensions of the key points identified by SIFT.Max Features: To optimize our approach, we thoughtfully configured the ”Max Features” parameter to 500 key points. Additionally, we experimented with various values, ranging from 100 within our dataset. The dimensional descriptor obtained for each of these key points is consistently 128 dimensions. These 128-dimensional feature vectors describe the distinctive characteristics of each key point, and each row of the descriptors array corresponds to a descriptor vector for a specific key point in the image.

c) Speeded up robust fast (SURF)

General overview of how the SURF [9] algorithm works in the context of SLR:

Scale-space extrema detection: SURF detects interest points in an image at multiple scales. The algorithm analyzes the image at different scales by convolving it with Gaussian filters of various sizes. Key points are then detected as local extrema in this scale space.

Key point localization: Once the key points are detected, SURF performs subpixel localization to improve the accuracy of key point positions.

Orientation assignment: SURF computes the dominant orientation for each key point to achieve rotation invariance. It constructs a local neighborhood around each key point and calculates the gradient orientations and magnitudes within that

region. A histogram of orientations is created, and the dominant orientation is determined. The key points are then assigned the dominant orientation.

Descriptor extraction: After orientation assignment, SURF extracts feature descriptors that capture the local image information around each key point. It constructs a square region around each key point called a ”descriptor window.” This window is divided into smaller sub-regions or cells. Within each cell, SURF computes the Haar wavelet responses in horizontal and vertical directions. These responses are then used to calculate the descriptor vector, which represents the local image structure around the key point. Figure 4 shows the SURF representation. The following are the key parameters considered for the experimentation:

Response threshold: The value was set to 500. This threshold controls where key points are detected based on image intensity variations.Number of octaves: It determines how many scales of an image are used for feature detection. The value considered was 4. Upright: Key points consider orientation when set to ”False” and ”Upright” to ”True” to make key points orientation-invariant, Descriptor Size: Typically we obtained 128 dimensions, which specifies the dimensionality of feature vectors for key points.

Both the algorithms i.e. SIFT and SURF work on key point-based features in images. Specifically, the focus is on extracting robust and distinctive features known as key points or descriptors. These features are designed to be invariant to changes in scale, rotation, affine transformations, and partial occlusions, making them suitable for various computer vision tasks such as object recognition

Fig 4. SURF Key points

d) Local binary pattern (LBP)

It is a texture descriptor that characterizes local patterns of pixel intensities in an image. The basic idea behind LBP [16] is to compare the intensity value of a center pixel with its neighboring pixels and encode the result as a binary pattern.

Consider a grayscale image patch of size 3x3 pixels: 50 45 60

30 70 80

90 20 75

Center pixel selection: Choose a center pixel within the patch. Ex. pixel with intensity value 70. Neighboring pixel comparisons: Compare the intensity value of the center pixel with its neighboring pixels in a circular manner, usually clockwise or counter clockwise.

45 60

20 75

Binary pattern generation: For each neighbor, if its intensity value is greater than or equal to the center pixel value, assign a binary value of 1; otherwise, assign a value of 0. In this case, we can generate the binary pattern as follows: 0110

Final LBP value: The LBP value for the center pixel is the decimal value obtained from the binary pattern. In this case, the LBP value for the center pixel (70) is 6. In our experiment, the LBP radius is set to 1 and the number of points to sample in the neighborhood is set to 8 and the method parameter is set to uniform to ensure rotation-invariant LBP patterns.

e) Features from accelerated segment test (FAST)

This algorithm is a popular corner detection algorithm [17]. It is designed to efficiently and robustly identify corners in images by exploiting the characteristic of corners being areas of rapid intensity change. Steps carried out are as follows:

Apply the FAST algorithm to detect corners/key points in the pre-processed image frames.

For each detected corner, extract relevant information such as the location (x, y coordinates) and the corner intensity or response.

Feature Representation:

Represent each image as a set of extracted FAST key points or corners.

Each key point/corner can be represented by its location coordinates, and intensity/response value. Key points are detected, for each detected corner relevant information such as location (x, y coordinates) and corner intensity (response is extracted and

stored in the list). For each detected key point, relevant information such as the location coordinates (x, y) and the intensity/ response values are extracted and stored in a dictionary. Here are values for the key parameters of the FAST algorithm:

- Threshold: The threshold was set to 20. The threshold value helps in the stringent corner detection.
- Non-maximal Suppression: This parameter controls how closely spaced corners are handled. A common choice is 9 contiguous pixels, which means that to be considered a corner, a pixel must have 9 contiguous pixels with intensity differences greater or less than the threshold.

f) Feature fusion approaches

While traditional feature extraction methods like SIFT, FAST, HOG, SURF, and others demonstrate impressive performance in certain scenarios, they may exhibit limitations in different situations. These techniques are designed to extract specific features from images, leading to a lack of generalization as their main drawback. Consequently, their effectiveness can vary across different contexts, and they may underperform when faced with diverse or unfamiliar data.

FAST is widely recognized for its high computational efficiency and rapid key point detection, rendering it well-suited for real-time vision-based applications. However, it tends to exhibit instability when faced with transformations, blurring, and variations in illumination. This implies that its performance may degrade when the input images undergo significant transformations or exhibit poor lighting conditions. While FAST excels in terms of speed and efficiency, its limitations in handling such challenges should be taken into consideration. SIFT and SURF have proven to perform well in conditions where images undergo transformations, blurring, or variations in illumination. Since sign language gestures have different shape and orientation and is not fixed to a certain feature, recognizing each sign with good accuracy is a challenging task. Hence this research work aims to extract the features in an effective way which helps in achieving accurate recognition.

HOG+LBP

The proposed approach for accurate hand gesture recognition involves the combination of HOG and LBP features. By leveraging the texture information provided by LBP and the contour information provided by HOG, we aim to enhance the recognition performance. To address the presence of both non-textured and textured regions in images, cascading HOG and LBP features is done. By integrating texture and contour information, our method aims to achieve more robust and discriminative hand gesture recognition. The combined features provide a comprehensive representation that captures both the fine-grained texture details and the contours of the hand gestures. The concatenation operation appends the elements of lbp_features to the end of hog_features, resulting in a single combined feature vector. Sign language recognition systems often need to perform reliably under varying lighting conditions, orientations, and backgrounds. HOG and LBP are both relatively robust to such variations, making them suitable candidates for feature fusion. This is the reason, in this experiment we decided to combine these two features specifically.

FAST+SIFT

Firstly, FAST key point localization is done and then this localized key point image is passed to SIFT for computation of the values. The magnitude and direction of located key points will be calculated by the SIFT technique. FAST algorithm detects key points very speedily. Further, SIFT known as the best feature descriptor with highly distinctive and invariant viewpoints is used to compute descriptors. The fusion of SIFT and FAST offers flexibility in handling various types of sign language gestures. SIFT provides the ability to capture detailed hand shapes, while FAST can efficiently detect features like fingertips or areas with rapid texture changes, which are relevant for sign language recognition.

BOVW features integrated with SURF

The Bag of Visual Words (BOVW) technique encompasses several steps, including feature extraction, feature clustering, codebook construction, and histogram generation. Initially derived from the Bag of Words (BOW) model in data retrieval and Natural Language Processing (NLP), BOVW has gained widespread usage in image classification [18]. Instead of counting the occurrence of words in a text, BOVW adapts this concept by considering image features as ”words.” The process involves constructing a vocabulary by representing each image as a frequency histogram of its characteristic descriptors and key points. This modified approach enables the generation of histograms based on the frequency of image features, thereby facilitating effective image classification. Cluster all the SURF features which are similar, to make a visual vocabulary. The K-Means clustering algorithm plays a crucial role in the Bag of Visual Words (BOVW) technique. It facilitates the partitioning of a given set of n features into k clusters, assigning new features to the cluster based on the mean (centroid) of each cluster. By clustering similar features together, K-Means helps create a collection of visual words or visual vocabulary.

Histogram can be calculated by finding the frequency of occurrence of each visual word that belongs to image in total visual words. Figure 5 shows the representation of BOVW.

SURF with BOVW is advantageous due to SURF's robust key point detection, distinctive descriptors, effective aggregation, enhanced recognition accuracy, scalability, and flexibility. This combination is particularly well-suited for tasks that involve

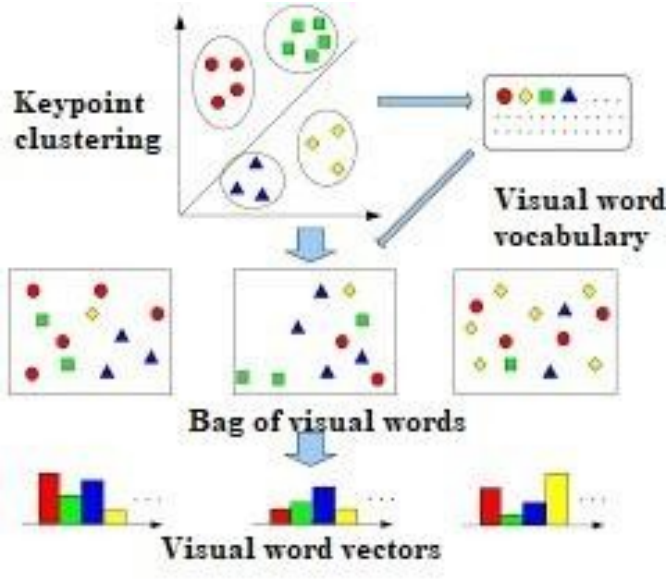

Fig 5. BOVW representation

recognizing and classifying objects or patterns within images.

An overview of the machine learning model proposed in this paper is represented in Figure 6.

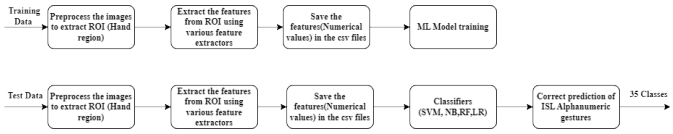

Fig 6. Overview of the proposed model

Previous approaches in SLR have primarily relied on directly extracting texture or key point features from images, which may not provide sufficient information and can result in degraded recognition rates. In contrast, the proposed method addresses this limitation by introducing a feature concatenation technique. This approach combines multiple features, enabling the extraction of extensive information and significantly improving the recognition rates of classifiers in accurately classifying the gestures.

3 Results and Discussion

In this section, the performance evaluation of local features like gradient based and key point descriptors for SLR is presented. The experiment is conducted on Dataset I and Dataset II, which is split into training and test sets with a ratio of 80:20. Training a machine learning model can be computationally intensive, especially with large datasets or complex models. Using an 80:20 split allows us to allocate a reasonable portion of our data for training while keeping the training time manageable. The hand-extracted regions from the images are used as input for feature extraction methods including HOG, SIFT, SURF, LBP, and FAST. To enhance the model's performance, LBP is integrated with HOG, SIFT is combined with FAST, and a bag of visual words is created for SURF key points. In addition to these feature extractors, the Support vector machine(SVM), Random Forest(RF), Naïve Bayes (NB) and Logistic regression(LR) classifiers are used for the sign prediction. Among the classifiers used, SVM demonstrates superior performance for all the features compared to LR, RF, and NB. In our work we have used linear kernel and multi class SVM. Experiment showed that the performance of the model improved with HOG features than other feature extractors used for the research work. Notably, RF achieves the lowest accuracy in recognizing ISL signs. The parameters of RF classifier i.e. the max_depth and random state was varied to get the better recognition rate. The system has undergone comprehensive training to accurately identify and interpret 35 distinct signs. These signs encompass 26 alphabets and 9 numerical symbols (1-9). The performance of the model varies with the combination of feature extractors and the classifiers considered for the experiment. Dataset-I has 33600 images for training and 8400 images for testing. Dataset-II: 28800 images

for training and 7200 images for testing.

3.1 Quantitative analysis

Accuracy: It is the most important performance measure and is a ratio of correctly predicted observations to the total observations. In terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), the formula of the accuracy can be written as,

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$$

Performance metrics: Precision is the ratio of correctly predicted positive observations to the total observations which are positive. The recall is the ratio of correctly predicted positive labels to the total no. of labels which are positive whereas the F1 score is the weighted average of precision and recall. The results can be seen in Table 5 .

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F1 Score} = \frac{2^{*(\text{ Recall} * \text{Precision })}}{\text{Recall} + \text{Precision}}$$

Results comparison of other methods on Dataset-I & II is shown in Table 2.

Table 2. Results comparison of Dataset-I & II

DATASET-I		**DATASET-II**	
Method	Accuracy	Method	Accuracy
Shravani K et al., 2020 [19]	0.99	Shagun Katoch et al., 2022 [11]	0.9917
Proposed	1.0	Proposed	1.0

The accuracy of the classifiers on Dataset I and II without using the feature fusion approach is shown in Table 3.

Table 3. Accuracy obtained without concatenating the features

Dataset	**Models used**	**Accuracy**
Dataset- I	SVM	0.98
	RF	0.78
	NB	0.88
	LR	0.92
Dataset- II	SVM	0.99
	RF	0.70
	NB	0.96
	LR	0.94

The accuracy of the classifiers on Dataset I and II by using the feature fusion approach is shown in Table 4.

The average prediction rate for Dataset I & II achieved based on the metrics like precision, recall and F1 Score is mentioned in the Table 5.

Result analysis show that SVM has given better accuracy and Random forest has given the least accuracy in recognizing the ISL gestures when worked with various features. Table 6 presents the latest advancements in ISL recognition specifically designed for simple backgrounds. These state-of-the-art methods showcase the most advanced techniques and approaches that have achieved significant progress in accurately interpreting and understanding ISL gestures within a simplified background environment.

Table 4. Accuracy of the classifiers (with features fusion approach)

Dataset	Models used	Accuracy
Dataset-I	SVM	1.0
	RF	0.88
	NB	0.98
	LR	0.97
Dataset-II	SVM	1.0
	RF	0.85
	NB	0.96
	LR	0.98

Table 5. Performance evaluation of various classifiers with different feature extractors

Classifier	Feature Extractors	Precision	Recall	F1 Score
SVM	HOG	0.97	0.98	0.96
	SIFT	0.96	0.97	0.98
	SURF	0.82	0.84	0.86
	HOG+LBP	0.99	1.0	0.99
	SIFT+FAST	1.0	0.99	0.99
	BOVW	0.99	0.98	0.98
LR	HOG	0.92	0.93	0.92
	SIFT	0.94	0.95	0.97
	SURF	0.78	0.81	0.76
	HOG+LBP	0.96	0.97	0.99
	SIFT+FAST	0.98	0.97	0.98
	BOVW	0.95	0.94	0.99
NB	HOG	0.96	0.97	0.98
	SIFT	0.98	0.95	0.96
	SURF	0.74	0.78	0.79
	HOG+LBP	1.0	0.99	0.97
	SIFT+FAST	1.0	0.99	0.99
	BOVW	0.96	0.96	0.97
RF	HOG	0.78	0.76	0.79
	SIFT	0.76	0.76	0.73
	SURF	0.79	0.74	0.78
	HOG+LBP	0.82	0.86	0.83
	SIFT+FAST	0.89	0.84	0.82
	BOVW	0.82	0.86	0.83

Table 6. State of art methods

Year	Dataset volume	Gesture	Models used	Average Accuracy
2020 [20]	24624 images	A-Z,1-9 and word gestures	HOG,KNN, HMM chain	98%
2021 [21]	5000 images	0-9	KNN, NB	98%
2022 [22]	24000 images	A-Z, 0-9	KAZE, KNN, SVM, NB	96%
Proposed	Dataset-I= 42000 images Dataset-II= 36000 images	A-Z & 1-9	Feature fusion approach	100%

3.2 State of art result comparison

The proposed method gives the best recognition rate by concatenating various features from the images. Feature fusion approaches in SLR provide a powerful means to combine diverse information leading to improved performance, robustness, and adaptability of the recognition system. They are particularly valuable in complex and challenging scenarios where a single feature extractor may not be sufficient to capture all the relevant information. It has several benefits in SLR:

1. Robust to variations in appearance
2. Captures shape and motion information
3. Localized feature representation
4. Invariant to translation, rotation, and scaling
5. Computational efficiency
6. Distinctiveness of key points
7. Invariance to affine transformations

4 Conclusion

This study introduces a pioneering method for classifying and recognizing ISL signs, specifically focusing on the alphabets (A-Z) and digits (1-9) with a consistent black background. The proposed system was extensively trained on a comprehensive dataset containing 35 static ISL alphabets and digits, resulting in an outstanding average accuracy of 100%. Among the different approaches, the SVM classifier with feature fusion exhibited the highest accuracy, surpassing other methods. Our approach incorporates texture features, including gradient features, local binary patterns, and key point descriptor features, for the classification of ISL gestures. We considered images captured from different angles, orientations, and under uniform illumination conditions to evaluate the system's performance. Remarkably, our model achieved a perfect accuracy of 100% on both the Dataset A and Dataset B. Overall, our study demonstrates the effectiveness of our feature fusion approach with various classifiers in achieving high recognition rates and accurate prediction of ISL signs. Our research brings innovation to static sign language gesture recognition. By combining diverse features like gradient attributes, key descriptors, local features, and corner points, we address the complexity of recognizing static signs' varied shapes and orientations. Our fusion strategy boosts accuracy, particularly benefiting communication aids for people with hearing impairments. In a field mostly focused on dynamic gestures, our work fills a critical gap by refining recognition for static signs, expanding support for a wider user range.

4.1 Future research direction and challenges

For future work, it would be beneficial to expand the dataset by incorporating a wider range of signs, including those in complex backgrounds, as suggested in reference [23]. This expansion would contribute to the development of a more comprehensive framework suitable for real-time applications. Additionally, there's a need to recognize not just individual signs but also entire sentence and phrases in sign language, necessitating models capable of understanding the grammar and syntax of sign language for more natural communication. Neural network algorithms are at the forefront of driving progress in SLR. They hold the potential to significantly elevate accuracy and proficiency, particularly in dealing with intricate signs, including time series data a task that conventional machine learning algorithms are not as adept at handling. SLR faces challenges in data diversity, encompassing various sign languages and demographics, privacy and ethics in data collection, real-time processing for practical applications, cross-platform compatibility, scalability, signer proficiency variability, standardization, noise tolerance, robustness to lighting conditions, affordability, ethical considerations, and standardized evaluation protocols. Tackling these challenges is crucial for advancing the field and improving accessibility for individuals with hearing impairments.

References

1) Bhattacharya A, Zope V, Kumbhar K, Borwankar P, Mendes A. Classification of Sign Language Gestures using Machine Learning. *International Journal of Advanced Research in Computer and Engineering*. 2020;8(12):97–103. Available from: https://doi.org/10.17148/IJARCCE.2019.81219.
2) Kumbhar S, Landge A, Kulkarni A, Solanki D, Kurtadikar V, Karad V. Indian Sign Language Recognition System. *International Journal of Innovative Science and Research Technology*. 2021;6(6). Available from: https://ijisrt.com/assets/upload/files/IJISRT21JUN1179.pdf.
3) Mariappan HM, Gomathi V. Real-Time Recognition of Indian Sign Language. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. 2019. Available from: https://doi.org/10.1109/ICCIDS.2019.8862125.
4) Subramanian B, Olimov B, Naik SM, Kim S, Park KHH, Kim J. An integrated mediapipe-optimized GRU model for Indian sign language recognition. *Scientific Reports*. 2022;12(1):1–16. Available from: https://doi.org/10.1038/s41598-022-15998-7.
5) Pala G, Jethwani JB, Kumbhar SS, Patil SD. Machine Learning-based Hand Sign Recognition. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. 2021;p. 356–363. Available from: https://doi.org/10.1109/ICAIS50930.2021.9396030.

6) Sharma A, Mittal A, Singh S, Awatramani V. Hand Gesture Recognition using Image Processing and Feature Extraction Techniques. *Procedia Computer Science*. 2020;173:181–190. Available from: https://doi.org/10.1016/J.PROCS.2020.06.022.
7) Mapari RB, Kharat GU. Indian Sign Language Alpha-Numeric Character Classification using Neural Network. *International Journal of Latest Research Engineering and Technology*. 2022. Available from: https://www.academia.edu/28423478/Indian_Sign_Language_Alpha_Numeric_Character_Classification_using_Neural_Network.
8) Dhivyasri S, Hari KB, Akash M, Sona M, Divyapriya S, Krishnaveni V. An efficient approach for interpretation of Indian sign language using machine learning. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. 2021;2021:130–133. Available from: https://doi.org/10.1109/ICSPC51351.2021.9451692.
9) Badhe PC, Kulkarni V. Artificial Neural Network based Indian Sign Language Recognition using hand crafted features. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2020;p. 1–6. Available from: https://doi.org/10.1109/ICCCNT49239.2020.9225294.
10) Amrutha K, Prabu P. ML Based Sign Language Recognition System. *2021 International Conference on Innovative Trends in Information Technology (ICITIIT)*. 2021. Available from: https://doi.org/10.1109/ICITIIT51526.2021.9399594.
11) Katoch S, Singh V, Tiwary US. Indian Sign Language recognition system using SURF with SVM and CNN. *Array*. 2022;14:100141. Available from: https://doi.org/10.1016/J.ARRAY.2022.100141.
12) Bello RWW, Mohamed ASA, Talib AZ. Contour Extraction of Individual Cattle From an Image Using Enhanced Mask R-CNN Instance Segmentation Method. *IEEE Access*. 2021;9:56984–57000. Available from: https://doi.org/10.1109/ACCESS.2021.3072636.
13) Munnaluri V, Pandey V, Singh P. Machine Learning based Approach for Indian Sign Language Recognition. *2022 7th International Conference on Communication and Electronics Systems (ICCES)*. 2022;p. 1128–1132. Available from: https://doi.org/10.1109/ICCES54183.2022.9835908.
14) Sreemathy R, Turuk M, Kulkarni I, Khurana S. Sign language recognition using artificial intelligence. *Education and Information Technologies*. 2023;28(5):5259–5278. Available from: https://doi.org/10.1007/s10639-022-11391-z.
15) Savant R, Nasriwala J, Bhatt P. Static Gesture Recognition for Indian Sign Language Alphabets and Numbers using SVM with ORB Keypoints and Image Pixel As Feature. 2023. Available from: https://www.ijert.org/static-gesture-recognition-for-indian-sign-language-alphabets-and-numbers-using-svm-with-orb-keypoints-and-image-pixel-as-feature.
16) Nguyen HBD, Do HN. Deep Learning for American Sign Language Fingerspelling Recognition System. *2019 26th International Conference on Telecommunications (ICT)*. 2019;p. 314–318. Available from: https://doi.org/10.1109/ICT.2019.8798856.
17) Arulkumar V, Prakash SJ, Subramanian EK, Thangadurai N. An Intelligent Face Detection by Corner Detection using Special Morphological Masking System and Fast Algorithm. *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. 2021;p. 1556–1561. Available from: https://doi.org/10.1109/ICT.2019.8798856.
18) Dhivyasri S, Hari KB, Akash M, Sona M, Divyapriya S, Krishnaveni V. An efficient approach for interpretation of Indian sign language using machine learning. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. 2021;2021:130–133. Available from: https://doi.org/10.1109/ICSPC51351.2021.9451692.
19) Shravani K, Lakshmi A, Geethika S, Sapna B. Indian Sign Language Character Recognition. *IOSR Journal of Computer Engineering*. 2020;22(3):14–19. Available from: https://www.iosrjournals.org/iosr-jce/papers/Vol22-issue3/Series-1/B2203011419.pdf.
20) Joshi G, Singh S, Vig R. Taguchi-TOPSIS based HOG parameter selection for complex background sign language recognition. *Journal of Visual Communication and Image Representation*. 2020;71:102834. Available from: https://doi.org/10.1016/j.jvcir.2020.102834.
21) Sahoo AK. Indian Sign Language Recognition Using Machine Learning Techniques. *Macromolecular Symposia*. 2021;397(1). Available from: https://doi.org/10.1002/masy.202000241.
22) Manikandan J, Krishna BV, Narayan SS, Surendar K. Sign Language Recognition using Machine Learning. *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*. 2022;p. 3235–3240. Available from: https://doi.org/10.1109/ICSES55317.2022.9914155.
23) Venugopalan A, Reghunadhan R. Applying deep neural networks for the automatic recognition of sign language words: A communication aid to deaf agriculturists. *Expert Systems with Applications*. 2021;185:115601. Available from: https://doi.org/10.1016/j.eswa.2021.115601.