Hybrid FiST_CNN Approach for Feature Extraction for Vision-Based Indian Sign Language Recognition

Akansha Tyagi
Department of Computer Science and Engineering,
Maharishi Markandeshwar
(Deemed to be) University, India
akan7sha@gmail.com

Sandhya Bansal
Department of Computer Science and Engineering,
Maharishi Markandeshwar,
(Deemed to be) University, India
sandhya12bansal@gmail.com

Abstract: *Indian Sign Language (ISL) is the commonly used language by the deaf-mute community in the Indian continent. Effective feature extraction is essential for the automatic recognition of gestures. This paper aims at developing an efficient feature extraction technique using Features from Fast Accelerated Segment Test (FAST), Scale-Invariant Feature Transformation (SIFT), and Convolution Neural Networks (CNN). FAST with SIFT are used to detect and compute features, respectively. CNN is used for classification with the hybridization of FAST-SIFT features. The system is implemented and tested using the python-based library Keras. The results of the proposed techniques have been tested on 34 gestures of ISL (24 alphabets set and 10 digit sets) and then compared with the CNN and SIFT_CNN, and it is also tested on two publicly available datasets on Jochen Trisech Dataset (JTD) and NUS-II dataset. The proposed study outperformed some existing ISLR works with an accuracy of 97.89%, 95.68%, 94.90% and 95.87% for ISL-alphabets, MNIST, JTD and NUS-II, respectively.*

Keywords: *Sign language, indian sign language, fast accelerated segment test, scale-invariant feature transformation, convolution neural networks.*

Received September 9, 2020; accepted March 10, 2021
https://doi.org/10.34028/iajit/19/3/15

1. Introduction

Communication is the most fundamental part of our daily life. Without communication, the interaction became difficult. However, language is an integral part of communication. Sign language (SL) is the gestural means of communication between hearing human beings and hard hearing people. It is a collection of various signs performed through facial expressions, hands, and other parts of our body [35]. Every country uses its SL system with its own syntactical and grammatical meaning like American Sign Language, British Sign Language, Chinese Sign language. Indian Sign Language (ISL) is the most predominant SL used in India [13]. Although SL is a boom for the deaf and mute community, still they are isolated from the common society because common people do not understand SL properly as ISL is not the official sign language despite the six million deaf and mute community [29].

Moreover, it is also a significant concern for deaf mute parents and the government of any country. The deaf school in any country urges to learn SL, which is not an easy task for a normal person. So, developing an Automatic Sign Language Recognition System (ASLR) may bridge the communication gap between the deaf and non-deaf communities [7]. It is an active research area in computer vision because of various challenges and importance.

ASLR involves four necessary steps: image acquisition, image pre-processing, feature extraction, and image classification [36, 41, 47]. Good feature affects the classification accuracy. Features are the key points in the image, and feature extraction is extracting prominent features [1, 23] from the image. A taxonomy of these techniques has been found in [41]. Various computer vision techniques like Scale-invariant Feature Transformation (SIFT), Histogram of Oriented Gradient (HOG) [18], Oriented FAST and Rotated Brief (ORB) [1], Fast Accelerated Segment Test (FAST), Speeded up Robust Feature (SURF) have been effectively used for feature extraction. A comparative analysis of the performance of these techniques on various metrics can be found [6, 10, 43]. proposes a novel approach that relies on multiple representations of HOG, Global Image Descriptor (GiST), and Binarized Statistical Image Feature (BSIF) to recognise English alphabets of ISL. Random forest is used as a classifier. The proposed technique achieved an accuracy of 92.20 % on 1300 ISL images and concluded that applying feature extraction before classification enhances the system's accuracy. Tao *et al*. [38] uses a fuzzy neural network for the development of hand gesture translator gloves. An accuracy of 92.58 % was achieved. Dudhal *et al*. [9] developed hybrid SIFT with adaptive thresholding for feature extraction and uses CNN as a classifier trained on 5000 images and achieves an accuracy of 92.78 %. In [3] 50 classes of Batik images, a traditional Indonesian fabric, were classified using Bag of Visual

Words (BoVW) and SIFT. Further [8] also uses BoVW for ISL gestures. Tareen *et al*. [39] makes use of invariant characteristics of SIFT for feature extraction of various ISL gestures. 30 signs of Arabic Sign Language (ArSL) are recognized using a skin-blob technique [15]. Although in [10, 30] SIFT has been shown as a promising technique for feature extraction but slow in processing high-resolution images [39] resulting in slow speed. Tyagi and Bansal [42] used the FAST and SIFT for recognition of ISL with more computational complexity. Extracting highly informative features reduces the computational burden of the classifier. Due to its importance, extensive studies on feature extraction techniques involving computer vision, soft computing, and deep learning are in the literature.

The contribution of the proposed model is named FiST_CNN. The proposed model has been validated for 24 ISL alphabets and 10 digits of ISL with 200 images for each gesture. For the experimental results, 80:20 of the dataset was used for training and validation of the CNN. The novelty of the presented work is developing a hybrid approach for fast feature extraction that will help in the efficient recognition of ISL gestures. The CNN model consists of seven convolution layers, and max-pooling layers, followed by two dense, fully connected layers that have been constructed. The performance of the proposed model has been evaluated using a confusion matrix.

The rest of the paper is organized as follows: Section 2 points up the related work. Section 3 gives a brief overview of FAST, SIFT, and CNN. Section 4 presents the working of a hybrid approach. Section 5 shows the effectiveness of the proposed approach by testing it on different sets of ISL gestures. Finally, the conclusion and future work is given in section 6.

2. Related Work

ISL is a developing language in the Asian continent. However, a lot of work has been done on the recognition of ASL and ISL, but our study shows that reducing features' space and time complexity is still a challenging task in ISLR.

2.1. Literature Review

In the present study, the fast localization of key points in the gestures has been done by the FAST computer vision technique. To improve the recognition efficiency, [21, 25, 26, 33] uses a FAST technique that detects accurate and fast key points even in low-resolution images. Despite fast detecting key points by FAST [14], SIFT has proved its effectiveness as a descriptor [2, 9]. FAST, which is an improved version of SIFT, is considered a very fast detection technique [43]. But because of its computing disability, SIFT has been used for computing features, making analysis very efficient and effective [21, 39]. Then these localized key points are passed to the SIFT technique for the computation of values. Finally, CNN is then used for the classification of gestures [16, 24, 28, 36, 47]. The incredible results from CNN in image processing and image classification have inspired researchers to apply it to SLR [5, 12]. The main idea behind using CNN for SLR is its automatic feature extraction, unlike the traditional soft computing methods such as neural networks [13], genetic algorithms, and fuzzy clustering mean [22, 44], which require handcraft feature extraction. There exist a lot of SLR systems making use of CNN. Wangchuk *et al*. [48] uses CNN for the recognition of 10 different real-time Bhutanese Sign Language digits. The network is trained on 20,000 images, and an accuracy of 97.62 is achieved during the testing phase. Islalm *et al*. [17] uses CNN to recognize Bangla Sign Language for 35 characters and 10 numbers. Similarly, in [20] CNN with learning, residual learning is utilized to recognise 700 digit signs. The developed model is independent of rotation. Kishore *et al*. [24] utilizes the ability of 4-layer CNN for recognizing the ISLR recognition on mobile-based applications. The average accuracy of 92.88 % is achieved by the system for 200 ISL signs with different orientations. In [38], a max-pooling CNN is proposed for vision-based hand gesture recognition. However, the proposed system is trained for six gestures only, and a dataset of 6000 images was collected. In addition to this, Wadhawan *et al*. [46] uses CNN with different window sizes followed by Relu and max-pooling layers used to recognise daily usage words, digits, and alphabets of ISL. A dataset of 35,000 images captured from a web-based camera has been created, and the highest validation accuracy of 98.80 % is achieved. Sarkar *et al*. [35] use the efficient working of CNN to develop a real-time ISLR system with a dataset of 52000 images of 26 ISL symbols captured by using a USB camera with an accuracy of 99.40%. An analytic comparison among three deep learning models such as Very deep convolution neural network for large scale image recognition (VGG16), Natural Language Processing (NLP) model and hierarchical model has been made in for identifying ISL gestures. The hierarchal model got an accuracy of 98.52% for one-hand gestures and 97.0 % for double-handed gestures. The proposed model is cost-effective as no hardware is needed. In [32] a hand pose model for the isolated sign language gesture using discriminant spatiotemporal features is proposed to solve the challenges faced in using deep learning approaches. Further, in [37] KAZE descriptors and BoVW are used for the recognition of sign language based on the skin segmentation method. The overall accuracy acquired by the proposed model is 99.23% on the MLP classifier. Thereafter for the recognition of continuous sign language words [49] model is proposed. The proposed multi-level Connectionist Temporal Classification (CTC) model uses frame feature extraction at the fully connected layer and the loss part is used for training to obtain recognition results.

2.2. Discussion on Related Work

From the mined literature following points have been discovered:

1. In ISLR, few studies focused on efficient feature extraction on images in different occlusion [6, 33].
2. In [10, 19], it is concluded that SIFT is stable in most conditions, but it is slow.
3. Furthermore, most of the work uses techniques like SIFT, SURF, HOG, or deep networks to recognise gestures.
4. However, time demands that these two techniques (Computer vision, Deep neural networks) be combined so that more accurate models for recognition of ISL with less computation time can be achieved.

3. A Brief Overview of Methods Used-FAST, SIFT, CNN

Feature extraction is extracting relevant information from the input hand gesture and then transforming it into the compact vector form. Various techniques with their pros and cons are proposed in the literature for this purpose. In this paper, FAST, SIFT and CNN are used. This section presents an overview of these methods.

3.1. Features from Accelerated Segment Test (FAST)

FAST, a corner detection algorithm was proposed by Rosten and Drummond [34] in 2006. The basic idea behind the working of FAST is that corners show more intensity change than edges. For feature detection, it places a circular mask over a pixel (p) to classify whether it is a corner or not. For a set of N contiguous pixels $y \in (y_1, y_2, .., y_n)$ on a circle, pixel p can be classified as a corner if it meets either of the following two conditions:

- Condition 1: ∀y_i ∈y, the intensity of$I_y > I_p + t$
- Condition 2: ∀y_i ∈y, the intensity of $I_x < I_p - t$

Whether I_p is the intensity of a pixel p and t any threshold value. The most promising advantage of FAST is its computational efficiency over SIFT and SURF.

3.2. Scale Invariant Feature Transformation (SIFT)

Lowe [26] introduced a scale and rotation invariant feature extraction technique, SIFT. It works in four phases: Detection of scale-space, keypoint localization, keypoint orientation, and keypoint computation. The first phase increases computation speed; it uses the Difference-Of-Gaussian (DOG) function instead of Gaussian to compute key points. In the second step, low contrast and poorly localized key points are removed. Then each key point is assigned an orientation depending upon local image properties in the orientation assignment stage. Finally, the SIFT keypoint descriptor is created in the phase by padding up the key points and balancing the orientation.

3.3. Convolution Neural Network (CNN)

Convolutional Neural Network CNN is a deep learning approach based on a feed-forward neural network. CNN is highly recommended for computer vision tasks. A typical CNN has three layers: a convolution layer, a max-pooling layer, and a Fully Connected (FC) layer. The first layer is the convolution layer, where the list of ‘filters’ such as ‘blur’, ‘sharpen’, and ‘edge-detection’, are all done with a convolution of kernel or filter with the image. The results from each convolution are placed into the next layer in a hidden mode. The output of convolved layer is then passed to the Pooling layer. The Pooling layer merges the pixel regions in the convolved image (shrinking the image) before learning kernels on it. The next layer is fully connected to a convolution network used to flatten the feature matrix into a vector and feed it into an FC layer for class classification. The FC layer follows the backpropagation method to find out the most accurate weights. Finally, a dropout is added after the FC layer to avoid the overfitting issue in the model [12]. Dropout means drop units out randomly with a probability p, which can be set to zero during feedforward and back-propagation in the network.

4. The Proposed Hybrid of FAST-SIFT based CNN (FiST_CNN)

4.1. Motivation

Although traditional feature extraction techniques such as SIFT, FAST, HOG, etc., perform exceptionally well in one situation but may underperform in other situations, they are intended to extract specific features from an image. Therefore, the lack of generalization is the main drawback of traditional feature extraction techniques. For instance, FAST has high computational efficiency [26] and high-speed performance for detecting key points making it more suitable for real-time vision-based applications. However, it is not stable to rotation, blurring, and illumination. It has also been noticed that SIFT performs well in these conditions but with wrong timings [10].

On the other hand, CNN has good generalization capability but is computationally expensive. This motivated us to hybridize the traditional feature extraction techniques with each other and with CNN. Moreover, the hybridization of these has rarely been used in sign language recognition systems. Along with this, the recognition of different kind of gestures separetly may reduce the model computational complexity [42]. Therefore, we combine FAST, SIFT, and CNN for effective and efficient features extraction of ISL gestures.

4.2. Proposed Hybrid Approach

Figure 1 shows the overall architecture of the hybrid approach FiST_CNN for ISL. It consists of three major phases: data preprocessing, feature extraction, and training and testing of CNN. In the first phase, the stored static single-handed images are resized to 224*224, and then data augmentation is done on resized images. Then in the next phase, key points are localized by FAST techniques. Then the value of these localized key points is computed using SIFT in the third phase. Finally, these values are passed to CNN for training. After this classification of images into various classes is done by CNN.

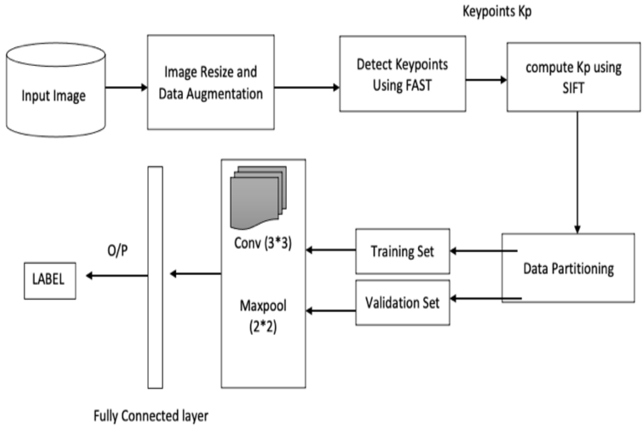


Figure 1. Architecture of FiST_CNN for ISL.

4.2.1. Image Resize and Data Augmentation

The vector with localized magnitudes and gradients computed using FiST is passed for training and testing groups using data augmentation. However, the following approach has extracted only the essentialkeypoints from the image, making other pixel value 0. Data augmentation is applied to make the system more robust in terms of image orientations, occlusions and transformation at different angles and lighting conditions.

4.2.2. Feature Extraction

In this phase firstly the key points are localized by using the FAST computer vision technique.

To identify a pixel p as an interesting point, Bresenham's circle of 16 pixels is used as a test mask. Every pixel y in the considered circle may have one of the following three states [14]:

$$S_{p\to y}=\begin{cases} d, & I_y \le I_p - T\ (darker)\\ b, & I_y \ge I_p + T\ (brighter)\\ s, & I_p - T < I_y < I_p + T\ (similar)\end{cases} \tag{1}$$

Where, I_y is the intensity value of pixel, I_p is the intensity value of nucleus (pixel p) and T is the threshold parameter that controls the number of corner responses.

A pixel p is identified as a corner if there exist 12 contiguous points on the segment that have an intensity value brighter and darker than p. This process is repeated for all the pixels in the image.

After this, a vector of localized pixels as the output of the FAST algorithm is passed to SIFT technique for computation of their values. The magnitude and direction of the localized points are computed by SIFT using the equation [43].

4.2.3. Data Partioning

The FiST approach has saved the model from the chances of overfitting as, overfitting is generated when the data contains noise.To validate the performance of the FiST_CNN model after the data augmentation dataset is also divided in a ratio of 80:20.

4.2.4. Model Training Using CNN

Thereafter, the group with training images (T_r) is passed into CNN for training. After this various convolution functions and max pooling, functions are performed on $T_r$using Equations (2) and (4) respectively.

$$I_j(x,y)=K*((x-m+1)*(y-m+1)) \tag{2}$$

$$I_j(x,y)=K*(x_m)*(y_m) \tag{3}$$

$$I_j(x,y)=K*\frac{x_m}{n}*\frac{y_m}{n} \tag{4}$$

Here $I_j(x,y)\in T_r$ is the input image from the training set. A kernel (K) with a size of (m,m) and a stride of (n,n) is used.

After this normalization is performed using the 'RELU' function on $Ij(x,y)$:

$$I_j(x,y)=max(0,x_s) \tag{5}$$

Then this normalized output is flatted into a single vector and fed to the dense layer. A dropout ratio of 0.5 is further added at a fully connected layer to avoid over-fitting. A dense layer with 124 neurons is linked as a fully connected layer.

Leaky Rectifier Linear Unit (Leaky ReLU) is used to introduce the non-linearity of CNN. A categorical cross-entropy is used as the cost function given in Equation (6):

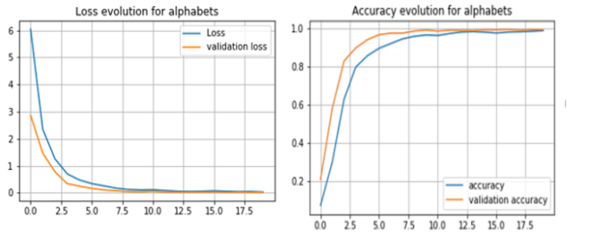
$$CE=-log\left(\frac{e^{S_p}}{\sum_{j=1}^{C}e^{S_j}}\right) \tag{6}$$

Where Sp is the CNN score for the positive class, C is the class andSj is the class score for each class j in C. The model is then optimized using Adam, which is an adaptive gradient-based optimization method.

Probabilities are calculated by using the softmax function at the final layer using Equation (7)

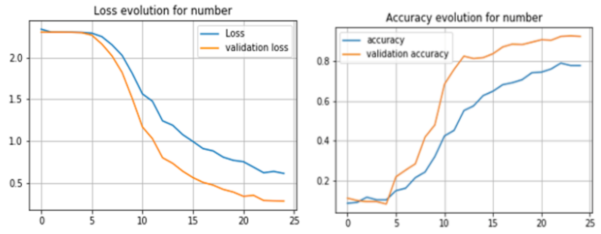
$$f(CT)_{ij} = \frac{e^{CT_{ij}}}{\sum_{j=1}^{C} e^{CT_{ij}}} \qquad (7)$$

Then the trained model FiST_CNN is saved for the predictions. The training and validation, accuracy and loss for ISL alphabets and numbers are shown in Figure 2-a), Figure-b) and Figure 3-a), Figure-b) respectively.



a) Loss evaluation for alphabets. b) Accuracy evaluation for alphabets.

Figure 2. Training accuracy graph for ISL alphabets.



a) Loss evaluation for numbers. b) Accuracy evaluation for numbers.

Figure 3. Training accuracy graph for ISL numbers.

4.2.5. Testing Phase

In this phase firstly the accuracy and time taken by FiST_CNN for trained images are calculated. Then the testing images from the testing folder are passed to the FiST_CNN for prediction. Finally, the confusion matrix, accuracy, and error rate are generated for the testing folder.

5. Experimental Results

To evaluate the performance of FiST_CNN, the approach has been tested on MNIST [27], JTD [40], and NUS-II [31] dataset. As no standard dataset for ISL alphabet gestures is available so dataset from a GitHubproject [11]. ISL dataset is consist of 4962 images with more than 200 images for 24 gesture. The algorithm has been implemented on Python 3-jupyter notebook and the simulation was done Intel® core™, 8 GB RAM and 256 caches per core, 3MB cache in total. Graphics with GPU type with Video Random Access Memory (VRAM) 1536 MB. The main objective of the performance analysis of FiST_CNN is to maximize the accuracy of the model with reduced computation complexity.

5.1. Performance Metrics

For evaluation of FiST_CNN following performance metrics are considered:

1. *Accuracy*: accuracy is the number of correct predictions made by the model over all the predictions made. The accuracy of the FiST_CNN is computed based on correct gestures predictions.
2. *Confusion Matrix*: the confusion matrix here is used to summarize the performance at the classification stage, on a set of validation data whose value is mapped from training data.
3. *Computational Time*: it is the total processing time of the model computed from image pre-processing to the predictions of the label.

Findings come from FiST_CNN are compared from existing CNN, SIFT_CNN[9] based on the above performance parameters.

5.2. Comparison based on Prediction Accuracy

Figure 4 shows the obtained accuracy comparison of CNN, SIFT_CNN [9], and FiST_CNN. FiST_CNN has achieved 97.89% accuracy for ISL alphabets, while the accuracy of CNN and SIFT_CNN is 94.64% and 95.58% respectively. It clearly shows that FiST_CNN has obtained higher accuracy with an improvement of 3% over CNN, and 2% over SIFT_CNN [9]. For number dataset accuracy achieved by FiST_CNN is 95.68%. Comparison of accuracy is also done at each epoch as shown in Figures 5 and 6 for ISL alphabets and numbers respectively. FiST_CNN has achieved 97.89% accuracy in only 10 epochs, however, CNN and SIFT_CNN [9] have an iteration of 20 epochs.

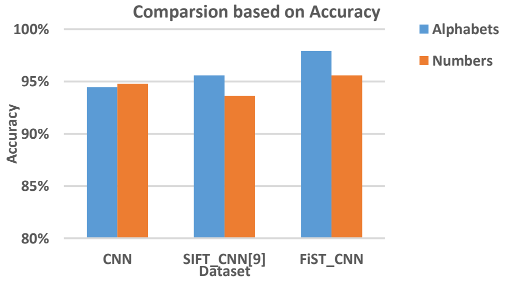


Figure 4. Accuracy Comparison of FiST_CNN, CNN and SIFT_CNN [9].

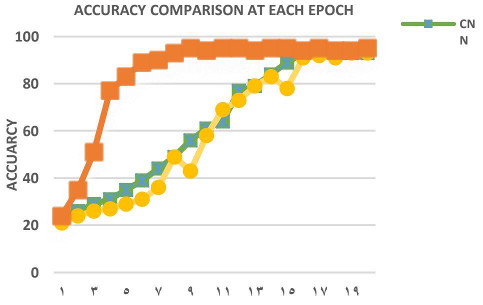


Figure 5. Accuracy Comparison per epochs for alphabet set.

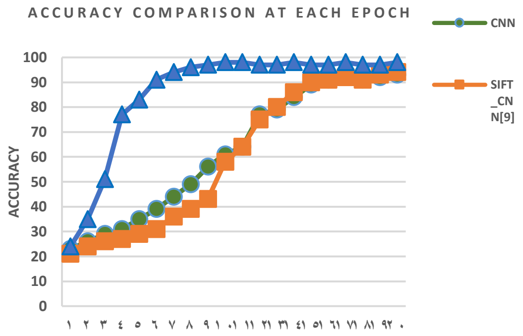


Figure 6. Accuracy comparison per epochs for numbers.

5.3. Comparison based on Computational Time

The time taken by FiST_CNN for alphabet set and number set is 2985.87 and 2593.55 secondsrespectively which is less than the CNN and SIFT_CNN [9]. The computational time analysis is shown in Figure 7 justifies the benefit of applying the FAST technique.

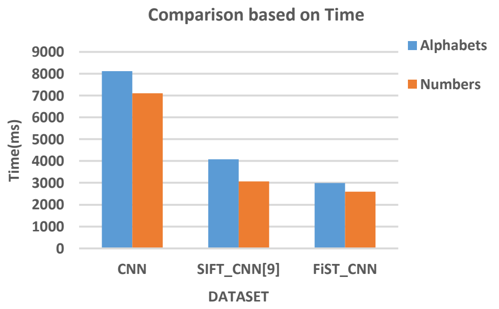


Figure 7. Time comparison of FiST_CNN, CNN and SIFT_CNN [9].

Tables 1 and 2 shows the total number of features extracted from all three approaches for the two datasetsrespectively. FiST_CNN has extracted reduced features compared to CNN, SIFT_CNN [9].

Table 1. Feature vector for ISL alphabet.

Technique	Feature Vector	Trainable Feature vector
CNN	6,13,84,870	4,12,84,857
SIFT_CNN [9]	4,46,03,558	3,26,03,446
FiST_CNN	3,43,05,158	1,25,03,100

Table 2. Feature vector for ISL number.

Technique	Feature Vector	Trainable Feature vector
CNN	3,12,43,170	2,10,52,172
SIFT_CNN [9]	2,23,12,135	1,34,26,263
FiST_CNN	1,55,67,271	90,23,729

5.4. Comparison based on Gesture Classification

The FiST_CNN is compared to other existing works in terms of classification accuracy by testing on Jochen Trisech Dataset (JTD) [40] and NUS-II dataset [31]. A comparison of the prediction of gestures is shown in Table 3, which proves the effectiveness of the FiST_CNN.

Table 3. Comparison of work with JTD [40] and NUS-II [31].

Dataset	Author name/ Approach used	Classifier	Accuracy
JTD	Trisech and Von Der Malsburg [40]	Gabor edge filter	86.2%
	Cubic kernel [4]	CNN	91%
	Joshi *et al*. [18]	SVM	92%
	FiST_CNN	CNN	94.90%
NUS-II	Kaur *et al*. [23]	SVM	92.50%
	Pisharady *et al*. [31]	SVM	94.36%
	Vishwakarma [45]	SVM	94.6%
	FiST_CNN	CNN	95.87%

5.5. Confusion Matrix

The confusion matrix obtained for FiST_CNN is in normalized form, Figures 8 and 9 represents the confusion matrix for ISL, ISL alphabets (A-Y) and number (0-9) respectively. The X-axis of the graph represents the predicted label while Y-axis represents the true label. Precision, Recall and F1 score were also calculated for the above dataset as shown in Table 3.

Table 3. Precision, recall and F1 score for FiST_CNN.

Sign	Precision	Recall	F1 score	Sign	Precision	Recall	F1 score
A,B,C,E,F,G, H,I,K,M,O,P, Q,R,S,X,Y	100	100	100	ONE	98	100	99
D	98	100	99	TWO	98	88	92
L	100	97	98	THREE	100	96	98
N	99	100	99	FOUR	90	94	92
T	100	99	99	FIVE	95	100	97
V	86	100	93	SIX	87	94	90
W	100	95	97	SEVEN	92	87	89
U	100	92	96	EIGHT	90	96	93
ZERO	98	98	98	NINE	100	96	98

6. Conclusions and Future Work

The Proposed study developed a hybrid framework FiST_CNN to reduce the pre-processing computation of images in the ISLR environment. The FiST_CNN has experimented on one-hand ISL static gestures. Features are detected by using FAST, which detects key points very rapidly. Further, to compute key points in invariant and distinctive conditions SIFT is used. Classification is done by using CNN. The performance of the proposed FiST_CNN has been compared with other techniques CNN and SIFT_CNN [9]. Results in section 4 conclude that FiST_CNN overpowers the concern of accuracy and computation time compares to both CNN and SIFT_CNN [9]. The FiST_CNN has achived an accuracy of 97.89%, 95.68%, 94.90% and 95.87% for ISL-alphabets, MNIST, JTD and NUS-II respectively. The performance parameter computed by confusion matrix shown in Figure 8 and 9 for ISL alphabets and numbers respectively, shows the recognition effectiveness of FiST_CNN. In future, this approach can pertain to the prediction of two hand gestures, dynamic gestures and some real-life gestures of the ISL dictionary.

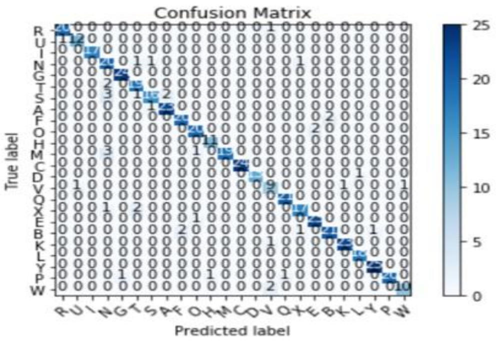


Figure 8. Confusion matrix of FiST_CNN for ISL alphabets.

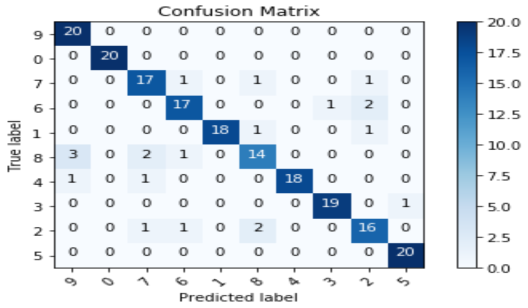


Figure 9. Confusion matrix of FiST_CNN for ISL numbers.

References

[1] Adel E., Elmogy M., and Elbakry H., “Image Stitching System based on ORB Feature-Based Technique and Compensation Blending,” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, pp. 55-62, 2015.

[2] Agrawal S., Jalal A., and Bhatnagar C., “Recognition of Indian Sign Language Using Feature Fusion,” *in Proceedings of the4th International Conference on Intelligent Human Computer Interaction*, Kharagpur, pp. 1-5, 2012.

[3] Azhar R., Tuwohingide D., Kamudi D., and Suciati N., “Batik Image Classification Using SIFT Feature Extraction, Bag of Features and Support Vector Machine,” *Procedia Computer Science*, vol. 72, pp. 24-30, 2015.

[4] Barros P., Maciel-Junior N., Fernandes B., Bezerra, B., and Fernandes S., “A Dynamic Gesture Recognition and Prediction System Using the Convexity Approach,” *Computer Vision and Image Understanding*, vol. 155, pp. 139-149, 2017.

[5] Bheda V. and Radpour D., “Using Deep Convolutional Networks for Gesture Recognition in American Sign Language,” arXiv preprint arXiv:1710.06836, 2017.

[6] Bora R., Bisht A., Saini A., Gupta T., and Mittal A., “ISL Gesture Recognition Using Multiple Feature Fusion,” *in Proceedings of theInternational Conference on Wireless Communications, Signal Processing and Networking*, Chennai, pp. 196-199, 2017.

[7] Cheok M., Omar Z., and Jaward M., “A of Hand Gesture and Sign Language Recognition Techniques,” *International Journal of Machine Learning and Cybernetics*,” vol. 10, no. 1, pp.131-153, 2019.

[8] Dardas N., Chen Q., Georganas N., and Petriu E., “Hand Gesture Recognition Using Bag-of-Features and Multi-Class Support Vector Machine,” *in Proceedings of theIEEE International Symposium on Haptic Audio Visual Environments and Games*, Phoenix, pp. 1-5, 2010.

[9] Dudhal A., Mathkar H., Jain A., Kadam O., and Shirole M., “Hybrid Sift Feature Extraction Approach for Indian Sign Language Recognition System Based on Cnn,” *in Proceedings of the International Conference on ISMAC in Computational Vision and Bio-Engineering*, Palladam, pp. 727-738, 2018.

[10] El-Gayar M., Soliman H., and Meky N., “A Comparative Study of Image Low Level Feature Extraction Algorithms,” *Egyptian Informatics Journal*, vo1. 4, no. 2, pp. 175-181, 2013.

[11] GitHub project [Licensed by MIT] (https://github.com/imRishabhGupta/Indian-Sign-Language-Recognition [Unpulished raw data], 2017.

[12] He K., Zhang X., Ren S., and Sun J., “Deep Residual Learning for Image Recognition,” *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 770-778, 2016.

[13] Hore S., Chatterjee S., Santhi V., Dey N., Ashour A., Balas V., and Fuqian S., *Indian Sign Language Recognition Using Optimized Neural Networks*, Springer International Publishing, 2017.

[14] Huijuan Z. and Qiong H., “Fast Image Matching Based-on Improved SURF Algorithm,” *in Proceedings of theInternational Conference on Electronics, Communications and Control*, Ningbo, pp. 1460-1463, 2011.

[15] Ibrahim N., Selim M., and Zayed H., “An automatic Arabic Sign Language Recognition System (ArSLRS),” *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 470-477, 2018.

[16] Islam M., Mitu U., Bhuiyan R., and Shin J., “Hand Gesture Feature Extraction Using Deep Convolutional Neural Network for Recognizing American Sign Language,” *in Proceedings of the4th International Conference on Frontiers of Signal Processing*, Poitiers, pp. 115-119,2018.

[17] Islalm M., Rahman M., Rahman M., Arifuzzaman M., Sassi R., and Aktaruzzaman M., “Recognition Bangla Sign Language Using Convolutional Neural Network,” *in Proceedings of theInternational Conference on Innovation and Intelligence for Informatics, Computing, and*

Technologies (3ICT), Sakhier, pp. 1-6, 2019.

[18] Joshi G., Singh S., and RenuV., "Taguchi-TOPSIS based HOG Parameter Selection for Complex Background Sign Language Recognition," *Journal of Visual Communication and Image Representation*, vol. 71, no. 4, pp. 102834, 2020.

[19] Juan L. and Oubong G., "A Comparison of Sift, Pca-Sift and Surf," *International Journal of Image Processing,* vol. 3, no. 4, pp. 143-152, 2009.

[20] Kalam M., Mondal M., and Ahmed B., "Rotation Independent Digit Recognition in Sign Language," *in Proceedings of theInternational Conference on Electrical, Computer and Communication Engineering*, Cox'sBazar, pp. 1-5, 2019.

[21] Karami E., Prasad S., andShehata M., "Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images," *arXiv preprint arXiv:1710.02726*, 2017.

[22] Karima K. and Nacera B., "A Dynamic Particle Swarm Optimisation and Fuzzy Clustering Means Algorithm for Segmentation of Multimodal Brain Magnetic Resonance Image Data," *The International Arab Journal of Information Technology*, vol. 17, no. 6, pp. 976-983, 2020.

[23] Kaur B., Joshi G., and Vig R., "Indian Sign Language Recognition Using KrawtchoukMoment-Based Local Features," *The Imaging Science Journal*, vol. 65, no. 3, pp. 171-179, 2017.

[24] Kishore P., Rao G., Kumar E., Kumar M., and Kumar D., "SelfieSign Language Recognition with Convolutional Neural Networks," *International Journal of Intelligent Systems and Applications*, vol. 10, no. 10, pp. 63-71, 2018.

[25] Loncomilla P., Ruiz-del-Solar J., and Martínez L., "Object Recognition Using Local Invariant Features for Robotic Applications: A Survey," *Pattern Recognition*, vol. 60, pp. 499-514, 2016.

[26] Lowe D., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[27] Mavi A., "A New Dataset and Proposed Convolutional Neural Network Architecture for Classification of American Sign Language Digits," *arXiv preprint arXiv:2011.08927*, 2020.

[28] Nagi J., Ducatelle F., Di Caro G., Cireşan D., Meier U., Giusti A., and Gambardella L., "Max-pooling Convolutional Neural Networks for Vision-Based Hand Gesture Recognition,"*in Proceedings of the IEEE International Conference on Signal and Image Processing Applications*, Kuala Lumpur, pp. 342-347, 2011.

[29] Nandy A., Prasad J., Mondal S., Chakraborty P., and Nandi G., "Recognition of Isolated Indian Sign Language Gesture in Real-Time," *in Proceedings of the International Conference on Business Administration and Information Processing*, Trivandrum, pp. 102-107,2010.

[30] Patil S. and Sinha G., "Distinctive Feature Extraction for Indian Sign Language (ISL) Gesture Using Scale Invariant Feature Transform (SIFT)," *Journal of The Institution of Engineers (India): Series B,* vol. 98, no. 1, pp. 19-26, 2017.

[31] Pisharady P., Vadakkepat P., and Loh A., "Attention Based Detection And Recognition of Hand Postures Against Complex Backgrounds," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 403-419, 2013.

[32] Rastgoo R., Kiani K., and Escalera S., "Hand Pose Aware Multimodal Isolated Sign Language Recognition," *Multimedia Tools and Applications* vol. 80, pp. 127-163, 2021.

[33] Rosten E., Porter R., and Drummond T., "Faster and Better: A Machine Learning Approach to Corner Detection," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105-119, 2008.

[34] Rosten E. and Drummond T., "Machine learning for High-Speed Corner Detection," *in Proceedings of the European Conference on Computer Vision*, Graz, pp. 430-443, 2006.

[35] Sarkar A., Talukdar A., and Sarma K., "Cnn-Based Real-Time Indian Sign Language Recognition System," *in Proceedings of the International Conference on Advances in Computational Intelligence and Informatics*, Hyderabad, 2020.

[36] Sharma A., Sharma N., Saxena Y., Singh A., and Sadhya D., "Benchmarking Deep Neural Network Approaches for Indian Sign Language Recognition," *Neural Computing and Applications*, vol. 33, pp. 6685-6696, 2020.

[37] Sharafudeen M., David S., and Simon P., "Visual Words based Static Indian Sign Language Alphabet Recognition using KAZE Descriptors. *in Proceedings of Evolution in Signal Processing and Telecommunication Networks*, Singapore, pp. 93-101, 2022.

[38] Tao W., Leu M., and Yin Z., "American Sign Language Alphabet Recognition Using Convolutional Neural Networks with Multiview Augmentation and Inference Fusion," *Engineering Applications of Artificial Intelligence,* vol. 76, pp. 202-213, 2018.

[39] Tareen S. and Saleem Z., "A Comparative Analysis of Sift, Surf, Kaze, Akaze, Orb, and Brisk," *in Proceedings of the International Conference on Computing, Mathematics and Engineering Technologies*, Sukkur, pp. 1-10,

2018.

[40] Triesch J., and Von Der Malsburg C., "Robust Classification of Hand Postures Against Complex Backgrounds," in *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, Killington, pp. 170-175, 1996.

[41] Tyagi A. and Bansal S., "Feature Extraction Technique for Vision-Based Indian Sign Language Recognition System: A Review," *in Proceedings of Computational Methods and Data Engineering*, Singapore, pp. 39-53, 2021.

[42] Tyagi A. and Bansal D., "Hybrid FAST-SIFT-CNN (HFSC) Approach for Vision-Based Indian Sign Language Recognition," *International Journal of Computing and Digital System*, 2022.

[43] Tyagi A., Bansal S., and Kashyap A., "Comparative Analysis of Feature Detection and Extraction Techniques for Vision-based ISLR System," *in Proceedings of the 6th International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, pp. 515-520, 2020.

[44] Villagomez E., King R., Ordinario M., Lazaro J., and Villaverde J.,"Hand Gesture Recognition for Deaf-Mute using Fuzzy-Neural Network," *in Proceedings of the IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, Bangkok, pp. 30-33, 2019.

[45] Vishwakarma D., "Hand Gesture Recognition Using Shape and Texture Evidences in Complex Background," *in Proceedings of theInternational Conference on Inventive Computing and Informatics*, Coimbatore, pp. 278-283, 2017.

[46] Wadhawan A. and Kumar P., "Deep Learning-Based Sign Language Recognition System for Static Signs," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7957-7968, 2020.

[47] Wadhawan A.and Kumar P., "Sign Language Recognition Systems: A Decade Systematic Literature Review," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 785-813, 2021.

[48] Wangchuk K., Riyamongkol P., and Waranusast R., "Real-Time Bhutanese Sign Language Digits Recognition System Using Convolutional Neural Network," *ICT Express*, vol. 7, no. 2, pp. 215-220, 2021.

[49] Zhu Q., Li J., Yuan F., and Gan Q., "Multi-Scale Temporal Network for Continuous Sign Language Recognition," arXiv preprint arXiv:2204.03864, 2022.

AkanshaTyagi is a Ph.D. Research Scholar at Maharishi Markandeswar University Mullana. She received her M. Tech degree from Maharishi Dayanand University Rohtak, (Haryana) and B. Tech degree from Punjab Technical University (Punjab). Her areas of interest are Soft-computing, Sign language recognition, and computer vision.

Sandhya Bansal is an Associate Professor at Maharishi Markandeswar University Mullana. She received her PhD degree from the same University. B. Tech degree from Kurukshetra University (Haryana). She has supervised 8 M. Tech candidates. Her areas of interest are WSN, Metaheuristics and VRP. She has about 25 research papers in international journals (SCI, Web of Science, Scopus, IGI and Elsevier etc..). Currently, she is supervising 2 PhD, research scholars.