International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (March-2022)

http://dx.doi.org/10.12785/ijcds/110199

Hybrid FAST-SIFT-CNN (HFSC) Approach for Vision-Based Indian Sign Language Recognition

Akansha Tyagi[1] and Sandhya Bansal[2]

[1, 2]*Department of Computer Science and Engineering, Maharishi Markandeshwar, (Deemed to be) University, Ambala, Haryana, India.*

Received 9 Dec. 2020, Revised 19 Jan. 2022, Accepted 4 Feb. 2022, Published 31 Mar. 2022

Abstract: Indian Sign Language (ISL) is the conventional means of communication for the deaf-mute community in the Indian subcontinent. Accurate feature extraction is one of the prime challenges in automatic gesture recognition of ISL gestures. In this paper, a hybrid approach, namely HFSC, integrating FAST and SIFT with CNN has been proposed for automatic and accurate recognition of ISL's static and single-hand gestures. Features from accelerated segment test (FAST) and scale-invariant feature transform (SIFT) provides the basic framework for feature extraction while CNN is used for classification. The performance of HFSC is compared with existing sign language recognition approaches by testing on standard benchmark (MNIST, Jochen-Trisech, and NUS hand postureII datasets. The HFSC algorithm's efficiency has been shown by comparing it with CNN and SIFT_CNN for a uniform dataset with an accuracy of 97.89%. Furthermore, the Computational results of the HFSC on complex background dataset achieve comparable accuracy of 95%.

Keywords: ISL, SIFT, FAST, CNN, Soft-Computing, Computer-Vision.

1. Introduction

In everyday life, communication is carried out in spoken form by speech and non-verbal form through gestures. Gestures that are being used consciously and unconsciously in almost all communication perspectives between human beings form the base of the languages used by hard of hearing people termed as sign language. Sign language is the non-verbal mode of communication used by the deaf-mute community. ISL is a natural language that serves as the predominant sign language for deaf and mute communities. However, the deaf-mute community contributes 1.1 million to the Indian population, and they had a literacy rate of 98%[1]. ISL has been removing hurdles in India's personal, educational, and social domains, giving many such people a lifeline. ISL is composed of static and dynamic gestures precisely. To recognize these signs, it is required to solve numerous difficulties as signs involving both hands, the two hands moving sometimes distinctive hand shapes, and so on[2].

Indian Sign Language Recognition (ISLR) system is an approach towards developing a vision-based gesture recognition system that can bridge the communication gap. ISLR is an active research area in computer vision as it has been proved a boon to the deaf and mute community. There are four necessary steps in ISLR: image acquisition, image pre-processing, feature extraction, and image classification. Feature extraction is essential for better image classification. It is a process of extracting critical features or key points from the image. Feature extraction is needed to reduce the redundant features in ISLR, preventing the loss of information in gesture recognition. This process helps in reducing the computation load during the training of the model. Due to its importance, various feature extraction and soft computing algorithms have been deployed. The taxonomy of these techniques has been found in [3]. Most of these techniques are deployed in content-based image retrieval (CBIR) features followed by a classifier such as support vector machine(SVM)[4][5], linear discriminant analysis(LDA)[6], neural network [7][8][9], and convolution neural network (CNN) [10][11][12][13]. These issues motivate the conception of an efficient feature extraction technique, which might be a fundamental challenge to prove. Extensive studies have been conducted in the literature on all phases of the ISLR system. Many researchers had addressed the feature extraction phase of ISLR. Various computer vision techniques like SIFT, SURF, HOG, ORB have been used for effective feature extraction from images. Dudhal et al. [10] implemented hybrid SIFT with adaptive thresholding for feature extraction and CNN as a classifier on a dataset of 5000 images. This method yields an accuracy of 92.78%. Bora et al. [1] implemented the various feature extraction on 1300 ISL images and concluded that applying feature extraction before classification enhances the sys-

E-mail address: akan7sha@gmail.com, sandhya12bansal@gmail.com

tem's accuracy. Bhumika et al. [14] extracted features using HOG on a 26 ISL gesture yielding 78.84% accuracy using K-Nearest correlated Neighbors for classification. SIFT has also been used on the same dataset, which progresses by up to 80%. Azhar et al., [15] implemented Bag of Words(BOW) and SIFT to classify 50 classes of Batik images which is an Indonesian a traditional Indonesian et al., [16] used SIFT due to its invariant characteristic over-illumination, rotation, translation, scaling, and slightly to the viewpoint and then implemented it on various ISL gestures for feature extraction. Each image has more than 400 features, and it reached up to 80% in BOW, providing a reliable matching among disrupted images. Ibrahim et al.,[17] implemented a dynamic skin detector to detect hand using skin-blob tracking technique on 30 signs of Arabic sign language(ARSL) with an accuracy of 97%. Tharwat et al.,[18] used SIFT to make the ARSL system robust against rotation with an accuracy of 99%. SIFT has evolved as the most promising technique in terms of feature extraction[14]. However, SIFT has been slow in processing high-resolution images[19][20], which also degrades the computation speed. The FAST technique has the advantage of being a fast keypoint detector, even in low-resolution images, which improves the recognition efficiency[21][19][22]. Instead, FAST can only detect key points, while SIFT has proved its effectiveness as a descriptor[23][9].FAST, an improved version of SIFT is considered a high-speed feature detection technique[21]. But because of its computing disability, SIFT has been used for computing features which made analysis very efficient and effective[20].

CNN's recent success in image classification tasks[24] has been extended to sign language recognition[4]. Unlike other traditional soft computing methods such as neural network, KNN, or genetic algorithm (GA), features were extracted manually, while CNN learns features from the training database. These networks preserve the spatial structure and can be used for object recognition tasks such as handwritten digit recognition. Like an ordinary neural network, CNN accepts and holds the pixels by learning internal feature representations. Kishore et al.,[13]use CNN's ability to extract features from the smaller portion of the image, making it more effective in ISLR recognition on mobile-based applications. A CNN architecture of four layers comprising a dataset of 200 ISL signs of different orientations is used for training. The system has achieved an average recognition rate of 92.88%.

Further, Wadhawan et al.,[2]developed an ISLR system for static gesture recognition with a dataset of 35000 images of 100 fixed signs. The proposed system attained the peak training accuracy of 99.72% and 99.90% on coloured and grayscale images, Sarkar et al.,[25] used a 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. The system achieves an accuracy of around 99.40 % tested on four signers in a real-time environment.

Another work has been done on complex background dataset. Extracting efficient features from the background with noise is a difficult task. Jochen et al. l[26], used hand posture against a complex background using the Gabor-edge filter method for ten ASL alphabets. The proposed system has achieved an accuracy of 86.2% for dark, light, and complex background hand postures. Agnes et al.[27] approach for Modified Census Transform(MCT) based on feature space classification to enhance vision-based hand gesture recognition. The proposed system has acquired 99.2% and 89.8% accuracy for uniform and complex background image.

Shao et al.[28], used Multi-objective Genetic Programming(MOGP) for feature extraction from complex background images. The approached technique experimented on MIT natural scene, and Jochen Trisech's hand posture dataset got an accuracy of 91.4%. To enhance feature extraction, Kaur et al.[29] use the krawtchouk moment-based shape feature for the ISL recognition system. The proposed method has achieved 97.9% accuracy on the ISL database. Following the same approach, Joshi et al.[30], design and ISLR system for complex background by extracting features using Taguchi and Technique for Order of Preference by Similarity to Ideal Solution(TOPSIS) based on decision-making technique. Experiments results on ISL, Jochen Trisech's, and ASL dataset acquired an accuracy of 92%, 92.8%, and 99.2% respectively.

In recent times for real-time systems, much research has been going on high-resolution hand postures. Several approaches were used in classifying complex gesture images. Pisharady et al.[31], proposed a Bayesian model to produce a saliency map for hand gesture recognition. For feature extraction, YCbCr colour space and skin mapping segmentation have been used. The proposed approach has achieved an accuracy of 94.36% on ten complex background hand postures. Mei et al.[32] used multiple threshold sets for each stump classifier that enhances its dimensionality power to reduce feature dimension and computational cost. Multi-classifier stumps covered each dimension of the hand posture, which helps extract gestures in a complex background. Zhang et al. [33], proposed the fusion of HOG and LBP to minimize the complication of feature selection and time in a hand gesture recognition system. The proposed approach is trained using SVM with Radial Basis Function (RBF) and acquired an accuracy of 95.09% on the NUS hand posture-II dataset. Adithya et al.[34], approached visual recognition of large sign language vocabulary in unrestrained background conditions using deep learning. The proposed method results are evaluated on NUS hand posture-II and ASL datasets, acquiring 98.13% and 97.89% respectively.

From the drilled literature, it has been observed that only 16% of work has been done on ISL [25]. Out of these, 45% of work has been done on static gestures with 48% on single-handed gestures. Moreover, most of the work uses a neural network for the classification of gestures of ASL. However, there is a requirement for fusing traditional computer vision techniques with deep learning models for ISL to develop more accurate models with less computation time. It has been hybridized with FAST for fast localization of key gestures' key points. In the present study, the SIFT has been applied to computation localized key points provided by FAST. CNN is then used for the classification of gestures. This paper referred to these enhancements in the form of HFSC. The Experimental results are evaluated on both uniform (34-ISL gestures) and complex (Jochen Trisech's and NUS hand posture-II) datasets.

This paper's main contribution is developing a hybrid approach for accurate and fast recognition of ISLR with fewer features to help the deaf and mute community. The rest of the paper is structured as follows: Section 2 gives a brief overview of the basic techniques. In section 3 proposed a hybrid approach (HFSC) is discussed. The proposed hybrid algorithm's effectiveness is tested on the set of gestures of alphabets; ISL digits are presented in section 4. Finally, section 5 concludes the present study with future work.

2. Brief overview of FAST, SIFT, and CNN

Feature extraction in ISLR is a process of extracting the feature vector set from images. Reducing the number of irrelevant features reduces the learning algorithms' running time, yields a more general classifier, and ensures a better understanding of the data and the classification rule. There are various methods available for solving this task. In this section, a brief overview of FAST, SIFT, and CNN is given.

A. *Features from Accelerated Segment Test (FAST)*

Edward Rosten and Tom Drummond[35] in 2006 proposed FAST, a corner detector algorithm. The concept inspired that corners are more vital points to show intensity changes than edges. It uses a circular mask over a pixel for testing. Every pixel point on the circle $y\epsilon(y_1, y_2, .., y_n)$ has three states S_y:

$$S_y = \begin{cases} d, I_y \leq I_p - T & (darker) \\ b, I_y \geq I_p + T & (brighter) \\ s, I_p - T \leq I_y \leq I_p + T & (similar) \end{cases} \quad (1)$$

Where,$I_y$is the intensity value of pixel y,I_p the intensity value of the nucleus and T is the threshold parameter that controls the number of corner responses. Hence, point pis classified as a corner by FAST if there is a segment with at least twelve contiguous points with intensity value brighter and darker than pixel p. Interest points are detected from 'N' adjacent pixels from the circle with value either above or below the intensity I_p. This process is repeated for all the image pixels until we find the contiguous feature vector set. To get rid of adjacent corners, non-max suppression is applied.

B. *Scale Invariant Feature Transform (SIFT)*

In 2004, D. G. Lowe [36] came up with SIFT for feature extraction. It extracted the distinctive local features, which are robust to occlusion and clutter. It can be used both as a feature detector and a descriptor[37]. Firstly, a scale-space of images has been created to extract the potential features from different locations [14]. Gaussian blur operator is then used to create the blurred image octaves. The previous step produces a lot of key points. Key points are then found using LOG approximations which are scale-invariant. To get a more accurate feature Taylor series expansion of space scale at local extrema is used to check the intensity of pixels at the taken threshold value. The next step is to computer the matching features by performing a descriptor operation on the local image region. SIFT descriptors have a random number of descriptors with 128 dimensions.

C. *Convolution Neural Network (CNN)*

CNN, a feed-forward artificial neural network, can perform various tasks with even better time and accuracy. A typical CNN has three layers: a convolution layer, a max-pooling layer, and a fully connected 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 image. The results from each convolution are placed into the next layer in a hidden mode. The convolved layer's output is then passed to the pooling layer, as shown in Figure 1. The pooling layer merges the pixel regions in the convolved image (shrinking the image) before attempting to learn kernels on it.

The next layer is fully connected in a convolution network used to flatten the feature matrix into a vector. This layer is responsible for the classification of labels. Further, back-propagation is done by the fully connected layer to determine labels with maximum true weights. The fully connected layer is usually followed by a dropout on the hidden layer[38]. Dropout means drop units out randomly with a probability p, which can be set zero during feed-forward and back-propagation in the network.

3. PROPOSED hybrid of FAST-SIFT-CNN (HFSC)

3.1 Motivation

Although most of the existing approaches for feature extraction, generally perform well in vast situations where they are being used, there is still an opportunity to develop new systems further because the existing techniques have some limitations in computational complexity or detecting gestures accurately. For instance, FAST has high computational efficiency[30] and high-speed performance for detecting key-points making it more suitable for real-time vision-based applications [39] however not stable to the rotation, blurring, and illumination. It has also been noticed that SIFT performs well in these conditions but with bad timings[40]. This motivated us that appropriate hybridization of these techniques might prove more effective and efficient for recognising ISL gestures. Keeping this, we combine FAST and SIFT for fast and accurate recognition of ISL gestures.

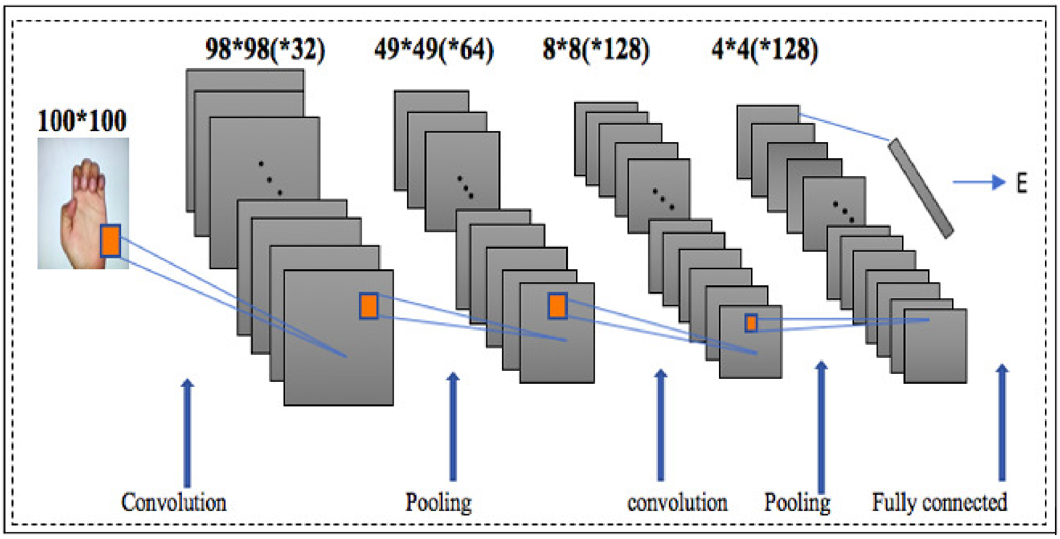

Figure 1. Architecture of CNN

3.2 Proposed Hybrid HFSC

In this paper, a hybrid of FAST and SIFT has been done for feature extraction of various gestures of ISL. Then CNN is trained on these features for classification of the given gestures. In our proposed approach, firstly, fast key-point localization is done by the FAST computer vision technique. Then this localized key point image is passed to SIFT for computation of the values. Finally, these values are given to CNN for training that classifies the gestures into various classes whose results are predicated through a confusion matrix.

The overall architecture of the HFSC model is shown in Figure 2. Firstly, all the images from each class are resized to 224*224. Then data augmentation of images is performed. For fast feature extraction, localization of key points will be performed by a FAST technique using equation (1) of sub-section 2.1. After this, the magnitude and direction of located key points will be calculated by the SIFT technique using the equation [36]. The images with located magnitude and gradients have been divided into training and testing groups. After that, images in the training group are passed into CNN, where various convolution functions and max-pooling functions are applied. The output of the convolution layers is flatted and fed to the dense layer. To avoid over-fitting, a dropout ratio of 0.5 is added at a fully connected layer. The dense layer consists of 124 neurons linked as a fully connected layer. To introduce CNN's non-linearity, we used Leaky Rectifier Linear Unit (Leaky ReLU) to solve this. A categorical cross-entropy is used as the cost function, as shown in equation 2:

$$CE = -\log \frac{e^{S_p}}{\sum_{j=1}^{c} e^{S_j}} \tag{2}$$

Where $S_p$is the CNN score for the positive class, C is the class and S_j 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. The trained model, after that, has been utilized for the prediction of gestures in the testing group. The overall procedure is shown in Algorithm 1 and 2.

Algorithm 1: Training Phase

Step 1: Load images of gestures for each class C_i.

Step 2: Resize the images.

Step 3: Perform data augmentation.

Step 4: Repeat steps 5-8 for each class C_i.

Step 5: Repeat steps 6-8 for each image $I_i \in C_i$.

Step 6: Construct a vector of key-points k_p we are using

$$\text{FAST\_DETECT}(I_i) \tag{3}$$

Step 7: Construct key-points descriptors and construct a vector of key-points k_p and their values v

$$[k_p, value] = \text{SIFT\_COMPUTE}([k_p]) \tag{4}$$

Step 8: Save images obtained from Step 7 to each class C_j

Step 9: Split C_j into training T_r and testing T_t

Step 10: Repeat steps 10-17 for each class CT_r

Step 11: Repeat step 12-15 for each image $I_j \in CT_r$

Step 12: Repeat step 13-18 for each epoch ep

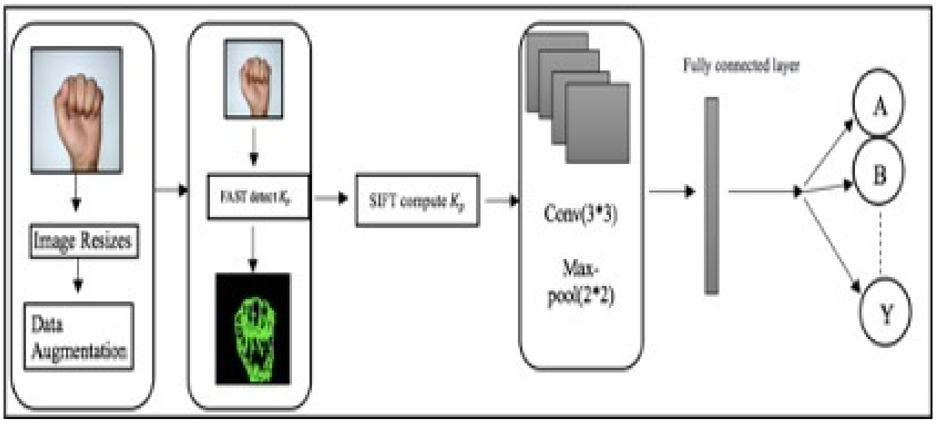

Figure 2. Architecture of HFSC

Step 13: Apply convolution function with input $I_j(x,y) \in C_T r$ with a kernel(K) size of (m,m):

$$I_j(x,y) = K * ((x - m + 1) * (y - m + 1)) \tag{5}$$

$$I_j(x,y) = K * (x_m) * (y_m) \tag{6}$$

Step 14: Apply max pooling on $I_j(x,y)$ with a stride of (n,n):

$$I_j(x,y) = K * \frac{x_m}{n} * \frac{y_m}{n} \tag{7}$$

Step 15: Normalization is done using the 'RELU' function on $I_j(x,y)$:

$$I_j(x,y) = max(0, x_s) \tag{8}$$

Step 16: Flatten each image $I_j(x,y) \in CT_r$ into a single vector I_j.

Step 17: Construct feature vector by using a fully connected layer, the sum of bias value x[j], layer weight w[i][j], and activation function:

$$FC = \sum_{\theta}^{x_s} I_j(x,y) * w[i][j] * x[j] \tag{9}$$

Step 18: Find the probabilities using the softmax function at the final layer:

$$fCT_{ij} = \frac{e^C T_{ij}}{\sum_{j=1}^{c} e^C T_{ij}} \tag{10}$$

Step 19: Predict the accuracy and time of the model for training gestures.

Step 20: Save the trained model (HFSC) for the predictions.

Algorithm 2: Testing Phase

Step 1: Load testing images from T_t.

Step 2: Use a saved model (HFSC) obtained from Step 20 of Algorithm 1 for predictions of testing images.

Step 3: Generate confusion matrix and obtain accuracy and error rate using

$$Accuracy(\%) = \frac{TruePositive + TrueNegative}{\text{Total images in } T_t * 100\%} \tag{11}$$

$$\text{Error Rate} = 1 - Accuracy \tag{12}$$

4. EXPERIMENTAL RESULTS

The algorithm has been implemented on Python 3-jupyter notebook, and the simulation is done using Intel® core™, 8 GB RAM and 256 caches per core, 3MB cache in total. Graphics with GPU type with VRAM 1536 MB. The dataset is split into two parts training (70%), and testing (30%) as per industry standards. The main objective of the performance analysis of HFSC is to maximize the accuracy of the model with reduced computation complexity.

4.1 DATASETS

The proposed work has been set for the uniform datasets (ISL and MNIST) and then tested on publicly available Jochen Trisech's and NUS hand posture-II complex background datasets. To enhance model performance, Data-augmentation is applied in both uniform and complex datasets. The dataset features are described below:

4.1.1 UNIFORM DATASET

(i) MNIST

The numeric dataset has been taken from MNIST [41]. It consists of 2062 images with 206 images for each gesture from 0 to 9, as shown in Figure 3. Both MNIST and ISL datasets were together used for training the model.

(ii) ISL

As no standard dataset for ISL alphabet gestures is available, so dataset from a GitHub project https://drive.google.com/drive/folders/1wgXtF6QHKBuXRx3qxuf-o6aOmN87t8G- which consists of 4962 images with more than 200 images per gesture has been used as shown in Figure 4. Alphabets dataset comprised of 24 classes except for J and Z, because they require motion.

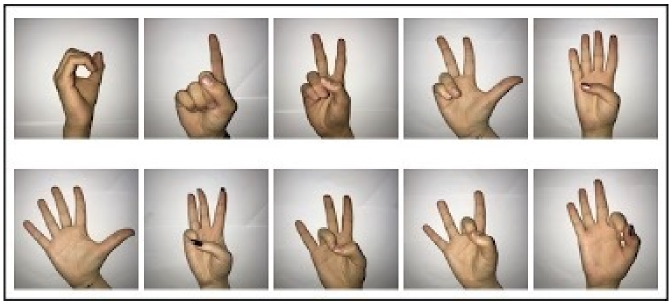

Figure 3. Sample images from MNIST dataset

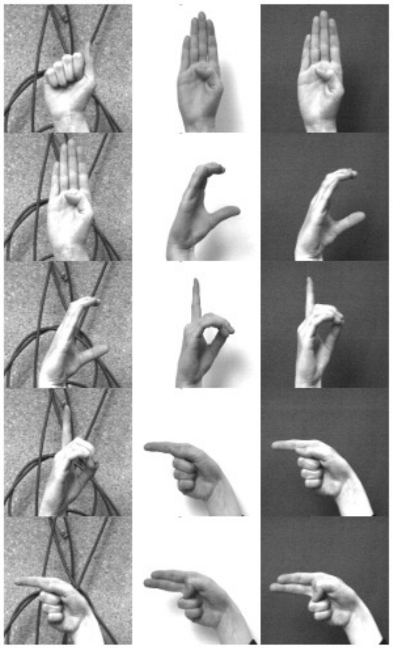

Figure 4. Sample images from JTD

4.1.2 COMPLEX DATASET

(i) JTD

Jochen Trisech dataset (JTD)[26] consists of single-hand static gestures collected from 24 subjects in dark, light, and complex backgrounds, as shown in Figure 5. The images are already converted in greyscale before applying feature extraction. There is a total of 2127 images in 10 classes.

(ii) NUS

NUS hand posture-II dataset[42] as shown in Figure 6. NUS hand posture-II consists of a training and testing set. The training set contains 2000 images from 40 subjects of 10 classes in a complex background. The test set includes 750 images from 15 subjects of 10 classes in different lighting conditions.

4.2 PERFORMANCE METRICS

For evaluation of HFSC following performance metrics are considered:

1) Accuracy

It is defined as the percentage ratio of correctly classified gestures to the total number of gestures in a particular class during the testing phase. It is calculated using equation (10).

2) Computational Time

It is defined as the algorithm's total time to calculate the feature vector and the classifier's time to classify the gestures.

3) Confusion Matrix

The confusion matrix here is used to summarize the performance at the classification stage, on a set of test data whose value is mapped from training data.The performance of the proposed HFSC has been compared with basic CNN, `SIFT_CNN` [10] based on the above parameters.

4.3 ANALYSIS OF HFSC ON THE UNIFORM DATASET

The HFSC is compared with three existing models on the uniform dataset. Table 1 shows the obtained accuracy

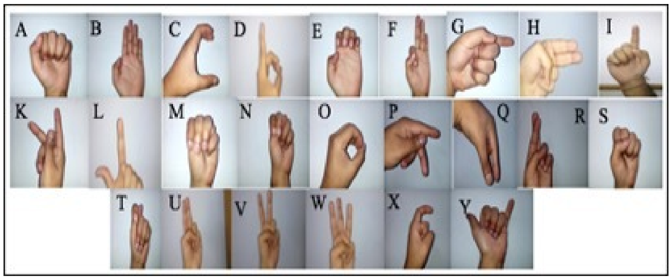

Figure 5. Sample images of ISL dataset

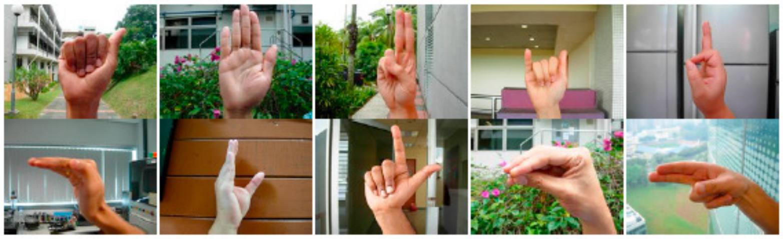

Figure 6. Sample images from NUS hand posture-II dataset

TABLE I. Results on ISL uniform background dataset

Technique	Image Size	Epoch	Feature vector#	Time#	Accuracy
CNN	224*224	20	6,13,84,870	8118.80 seconds	94.64%
SIFT_CNN[10]	224*224	20	4,46,03,558	21976.38 seconds	95.58%
HFSC	224*224	10	3,43,05,158	10578.51 seconds	97.89%

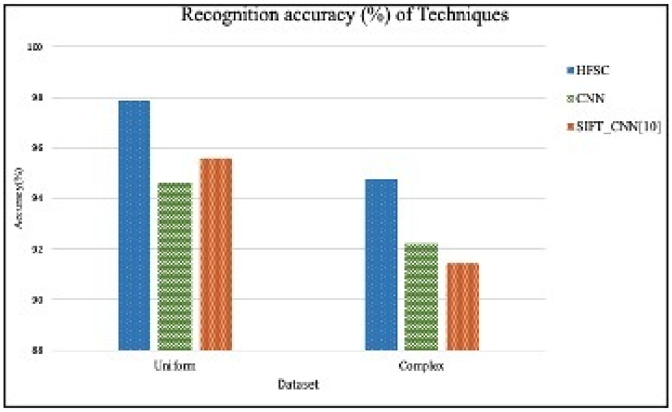

Figure 7. Recognition accuracy of Techniques

comparison of CNN, SIFT_CNN [10], and HFSC. HFSC has achieved 97.89% accuracy for the ISL dataset, while the accuracy of CNN and SIFT_CNN is 94.64% and 95.58% respectively. It clearly shows that HFSC has obtained higher accuracy with an improvement of 3% over CNN, and 2% over SIFT_CNN [10].

4.4 ANALYSIS OF HFSC ON COMPLEX BACKGROUND

To prove the robustness and effectiveness of HFSC on complex background dataset. Table 2 shows the performance of HFSC on JTD and NUS hand posture-II datasets compared to other approaches based on accuracy.

4.5 RECOGNITION ANALYSIS OF TECHNIQUES

A comparison of recognition accuracy is shown in Figure 7. It offers a significant accuracy in the case of the HFSC on both uniform and complex datasets compared to CNN and SIFT_CNN [10].

4.6 CONFUSION MATRIX

The confusion matrix obtained for HFSC is in normalized form. Figures 8, 9, and 10 represent the confusion

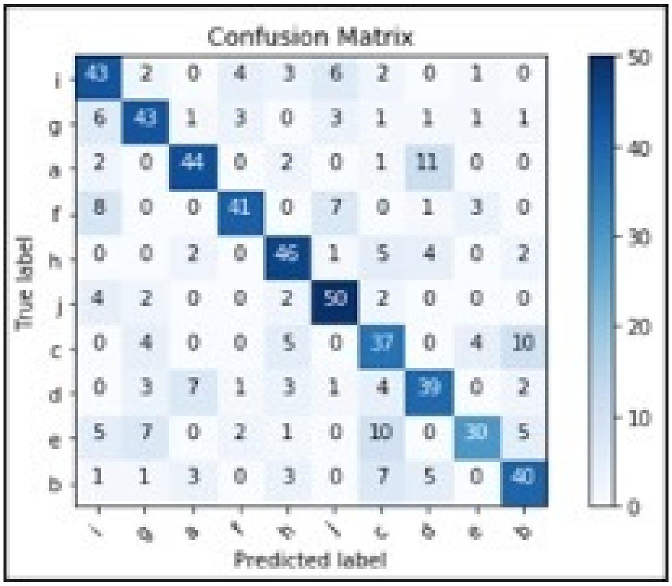

Figure 8. Confusion matrix for NUS hand posture-II dataset

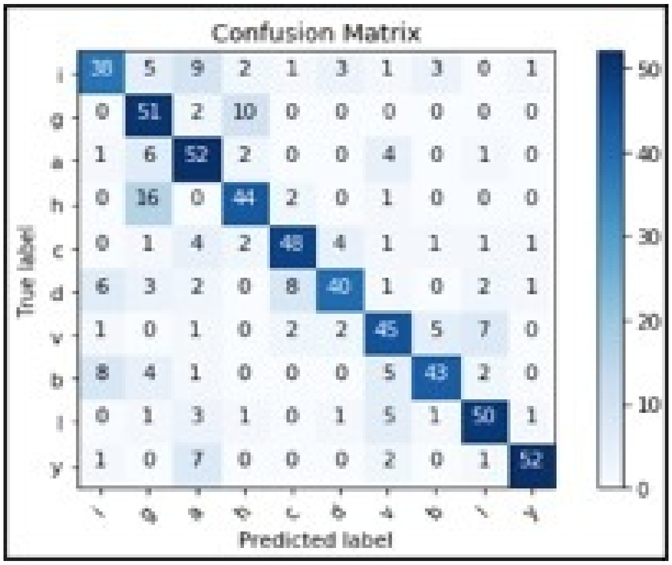

Figure 9. Confusion matrix for Jochen Trisech's dataset

matrix for JTD, NUS hand posture-II, and ISL dataset. The X-axis of the graph represents the predicted label, while Y-axis represents the true label.

5. CONCLUSION AND FUTURE WORK

The study developed a hybrid approach HFSC, in which FAST and SIFT are hybridized for extracting features for automatic and accurate recognition of static single hand ISL gestures. HFSC reduces the pre-processing time of images by detecting features using FAST, which detects keypoint very speedily. Further, SIFT known as the best feature descriptor with highly distinctive and invariant view-points is used to compute descriptors. CNN is used for classification. Table 1 shows that HFSC is superior to CNN and SIFT_CNN [10] both in terms of accuracy and time-computation for ISL gestures with an accuracy of 97.89% in a uniform background. The system proves to be robust against complex datasets (JTD and NUS hand posture-II) with an accuracy of 94.78% and 95.56% respectively. The confusion matrix shown in Figures 8, 9, and 10 proves HFSC effectiveness for recognizing sign language gestures. Future work aims adaptation of this hybrid approach to dynamic ISL gestures and more real-life gestures.

6. COMPLIANCE WITH ETHICS REQUIREMENTS

This paper does not contain any studies with human or animal subjects, and all authors declare that they have no conflict of interest. Any institute did not fund this study.

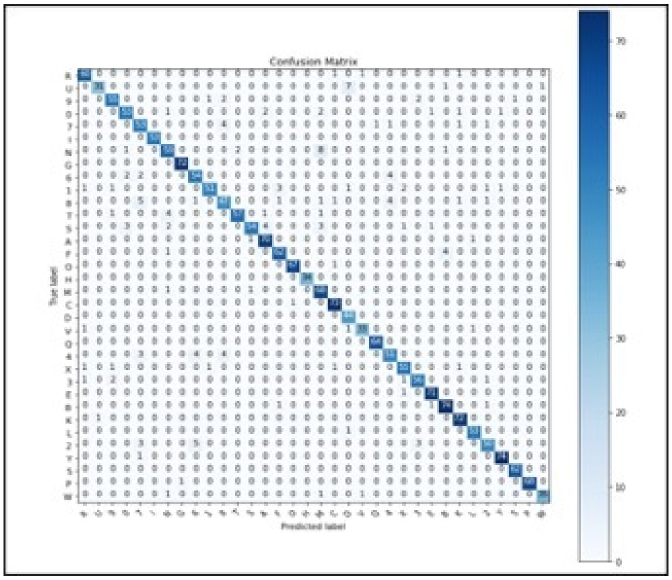

Figure 10. Confusion matrix for ISL alphabets

TABLE II. Performance of HFSC on Complex background dataset

Dataset	Author Name/ Approach Used	Classifier	Accuracy
JTD	Trisech et al,[26]	Gabor edge filter	86.2%
JTD	EGM[43]	-	82.6%
JTD	MCT[27]	Adaboost	98%
JTD	MOGP[28]	SVM	91.4%
JTD	LHFD[44]	SVM	95.2%
JTD	Cubic kernel[42]	CNN	91%
JTD	Joshi et al.,[30]	SVM	92%
JTD	Kelly et al.,[45]	SVM	93%
JTD	X. Y. Wu [46]	CNN	98.02%
JTD	HFSC	CNN	94.78%
NUS	Kaur et al.,[29]	SVM	92.50%
NUS	Adithya et al.,[34]	SVM	92.50%
NUS	Pisharady et al.,[31]	SVM	94.36%
NUS	Haile et al.,[47]	RTDD	90.66%
NUS	Zhang et al.,[33]	-	95.07%
NUS	DSPF[48]	-	96.53%
NUS	HFSC	CNN	95.56%

References

[1] R. Bora, A. Bisht, A. Saini, T. Gupta, and A. Mittal, “Isl gesture recognition using multiple feature fusion,” in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2017, pp. 196–199.

[2] A. Wadhawan and P. Kumar, “Deep learning-based sign language recognition system for static signs,” *Neural Computing and Applications*, vol. 32, no. 12, pp. 7957–7968, 2020.

[3] A. Tyagi and S. Bansal, “Feature extraction technique for vision-based indian sign language recognition system: A review,” *Computational Methods and Data Engineering*, pp. 39–53, 2021.

[4] D.-Y. Huang, W.-C. Hu, and S.-H. Chang, “Vision-based hand gesture recognition using pca+ gabor filters and svm,” in *2009 fifth international conference on intelligent information hiding and multimedia signal processing*. IEEE, 2009, pp. 1–4.

[5] J. Raheja, A. Mishra, and A. Chaudhary, “Indian sign language recognition using svm,” *Pattern Recognition and Image Analysis*, vol. 26, no. 2, pp. 434–441, 2016.

[6] N. Kumar, “Sign language recognition for hearing impaired people based on hands symbols classification,” in *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017, pp. 244–249.

[7] M. Sharma, R. Pal, and A. K. Sahoo, “Indian sign language recognition using neural networks and knn classifiers,” *ARPN Journal of*

Engineering and Applied Sciences, vol. 9, no. 8, pp. 1255–1259, 2014.

[8] H. Kondhalkar and P. Mukherji, "Speech recognition using novel diatonic frequency cepstral coefficients and hybrid neuro fuzzy classifier," in *International Conference on ISMAC in Computational Vision and Bio-Engineering*. Springer, 2018, pp. 775–788.

[9] S. C. Agrawal, A. S. Jalal, and C. Bhatnagar, "Recognition of indian sign language using feature fusion," in *2012 4th international conference on intelligent human computer interaction (IHCI)*. Ieee, 2012, pp. 1–5.

[10] A. Dudhal, H. Mathkar, A. Jain, O. Kadam, and M. Shirole, "Hybrid sift feature extraction approach for indian sign language recognition system based on cnn," in *International Conference on ISMAC in Computational Vision and Bio-Engineering*. Springer, 2018, pp. 727–738.

[11] X. Sun and M. Lv, "Facial expression recognition based on a hybrid model combining deep and shallow features," *Cognitive Computation*, vol. 11, no. 4, pp. 587–597, 2019.

[12] G. A. Rao, K. Syamala, P. Kishore, and A. Sastry, "Deep convolutional neural networks for sign language recognition," in *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*. IEEE, 2018, pp. 194–197.

[13] P. Kishore, G. A. Rao, E. K. Kumar, M. T. K. Kumar, and D. A. Kumar, "Selfie sign language recognition with convolutional neural networks," *International Journal of Intelligent Systems and Applications*, vol. 11, no. 10, p. 63, 2018.

[14] B. Gupta, P. Shukla, and A. Mittal, "K-nearest correlated neighbor classification for indian sign language gesture recognition using feature fusion," in *2016 International conference on computer communication and informatics (ICCCI)*. IEEE, 2016, pp. 1–5.

[15] R. Azhar, D. Tuwohingide, D. Kamudi, N. Suciati *et al.*, "Batik image classification using sift feature extraction, bag of features and support vector machine," *Procedia Computer Science*, vol. 72, pp. 24–30, 2015.

[16] S. B. Patil and G. Sinha, "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.

[17] N. B. Ibrahim, M. M. Selim, and H. H. Zayed, "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.

[18] A. Abraham, P. Krömer, and V. Snasel, *Afro-European Conference for Industrial Advancement*. Springer, 2015.

[19] E. Karami, S. Prasad, and M. Shehata, "Image matching using sift, surf, brief and orb: performance comparison for distorted images," *arXiv preprint arXiv:1710.02726*, 2017.

[20] E. Adel, M. Elmogy, and H. Elbakry, "Image stitching system based on orb feature-based technique and compensation blending," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, 2015.

[21] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[22] P. Loncomilla, J. Ruiz-del Solar, and L. Martínez, "Object recognition using local invariant features for robotic applications: A survey," *Pattern Recognition*, vol. 60, pp. 499–514, 2016.

[23] N. Dardas, Q. Chen, N. D. Georganas, and E. M. Petriu, "Hand gesture recognition using bag-of-features and multi-class support vector machine," in *2010 IEEE International Symposium on Haptic Audio Visual Environments and Games*. IEEE, 2010, pp. 1–5.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] A. Sarkar, A. K. Talukdar, and K. K. Sarma, "Cnn-based real-time indian sign language recognition system," in *International Conference on Advances in Computational Intelligence and Informatics*. Springer, 2019, pp. 71–79.

[26] J. Triesch and C. Von Der Malsburg, "Robust classification of hand postures against complex backgrounds," in *Proceedings of the second international conference on automatic face and gesture recognition*. IEEE, 1996, pp. 170–175.

[27] A. Just, Y. Rodriguez, and S. Marcel, "Hand posture classification and recognition using the modified census transform," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 2006, pp. 351–356.

[28] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 7, pp. 1359–1371, 2013.

[29] B. Kaur, G. Joshi, and R. Vig, "Indian sign language recognition using krawtchouk moment-based local features," *The Imaging Science Journal*, vol. 65, no. 3, pp. 171–179, 2017.

[30] G. Joshi, S. Singh, and R. Vig, "Taguchi-topsis based hog parameter selection for complex background sign language recognition," *Journal of Visual Communication and Image Representation*, vol. 71, p. 102834, 2020.

[31] P. K. Pisharady, P. Vadakkepat, and A. P. Loh, "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] K. Mei, J. Zhang, G. Li, B. Xi, N. Zheng, and J. Fan, "Training more discriminative multi-class classifiers for hand detection," *Pattern Recognition*, vol. 48, no. 3, pp. 785–797, 2015.

[33] F. Zhang, Y. Liu, C. Zou, and Y. Wang, "Hand gesture recognition based on hog-lbp feature," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2018, pp. 1–6.

[34] V. Adithya and R. Rajesh, "An efficient method for hand posture recognition using spatial histogram coding of nct coefficients," in *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2018, pp. 16–20.

[35] E. Rosten, R. Porter, and T. Drummond, "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.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[37] F. Alhwarin, C. Wang, D. Ristić-Durrant, and A. Gräser, "Improved sift-features matching for object recognition," in *Visions of Computer Science-BCS International Academic Conference*, 2008, pp. 179–190.

[38] W. Tao, M. C. Leu, and Z. Yin, "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] Z. Huijuan and H. Qiong, "Fast image matching based-on improved surf algorithm," in *2011 International conference on electronics, communications and control (ICECC)*. IEEE, 2011, pp. 1460–1463.

[40] M. El-Gayar, H. Soliman *et al.*, "A comparative study of image low level feature extraction algorithms," *Egyptian Informatics Journal*, vol. 14, no. 2, pp. 175–181, 2013.

[41] A. Mavi, "A new dataset and proposed convolutional neural network architecture for classification of american sign language digits," *arXiv preprint arXiv:2011.08927*, 2020.

[42] P. Barros, N. T. Maciel-Junior, B. J. Fernandes, B. L. Bezerra, and S. M. Fernandes, "A dynamic gesture recognition and prediction system using the convexity approach," *Computer Vision and Image Understanding*, vol. 155, pp. 139–149, 2017.

[43] I. Abasova, P. Rustamova, and R. Tahmazov, "About the relationship of structure-functional states of the teeth and internal organs," , no. 4, pp. 9–12, 2007.

[44] D. A. Reddy, J. P. Sahoo, and S. Ari, "Hand gesture recognition using local histogram feature descriptor," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2018, pp. 199–203.

[45] H. Josiński, D. Kostrzewa, A. Michalczuk, A. Świtoński, and K. Wojciechowski, "Feature extraction and hmm-based classification of gait video sequences for the purpose of human identification," in *Vision Based Systemsfor UAV Applications*. Springer, 2013, pp. 233–245.

[46] X. Y. Wu, "A hand gesture recognition algorithm based on dc-cnn," *Multimedia Tools and Applications*, vol. 79, no. 13, pp. 9193–9205, 2020.

[47] U. H. Hernandez-Belmonte and V. Ayala-Ramirez, "Real-time hand posture recognition for human-robot interaction tasks," *Sensors*, vol. 16, no. 1, p. 36, 2016.

[48] Q. Zhang, M. Yang, K. Kpalma, Q. Zheng, and X. Zhang, "Segmentation of hand posture against complex backgrounds based on saliency and skin colour detection," *IAENG International Journal of Computer Science*, vol. 45, no. 3, pp. 435–444, 2018.

Akansha Tyagi Akansha Tyagi 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 Sandhya Bansal is an Associate Professor at Maharishi Markandeswar University Mullana. She received her Ph.D. 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 6 Ph.D. research scholars.