2021 The 14th International Conference on Machine Vision (ICMV 2021)
November 8 – 12, 2021

Paper ID : T066

Bangla Sign Digits Recognition using Depth Information

S. M. Rayeed[1], Gazi Wasif Akram[1], Sidratul Tamzida Tuba[1], Golam Sadman Zilani[1], Hasan Mahmud[1], Md. Kamrul Hasan[1]

[1] Systems and Software Lab, Department of Computer Science and Engineering, Islamic University of Technology (IUT), Bangladesh

Outline

- Introduction
- Objective
- Background Study
- Proposed Dataset
- Proposed Approach
- Classification
- Result Analysis
- Future Work

Introduction

- Sign language is a visual language that uses predefined hand-shapes, facial expression, gestures and body language, in order to ease the communication of hearing-impaired people with others.
- Sign language recognition (SLR) is a renowned topic in Computer Vision and Human Computer Interaction. Now-a-days, using depth information in sign language recognition has become a trending research topic all over the world.
- Not much work has been done in Bangla sign language. The existing dataset of Bangla sign digits (Ishara-Bochon) contains RGB images only – no work has been done based on depth information.
- In this paper, we have proposed a dataset for Bangla sign digits using depth information, and then evaluated the performance.

Objective

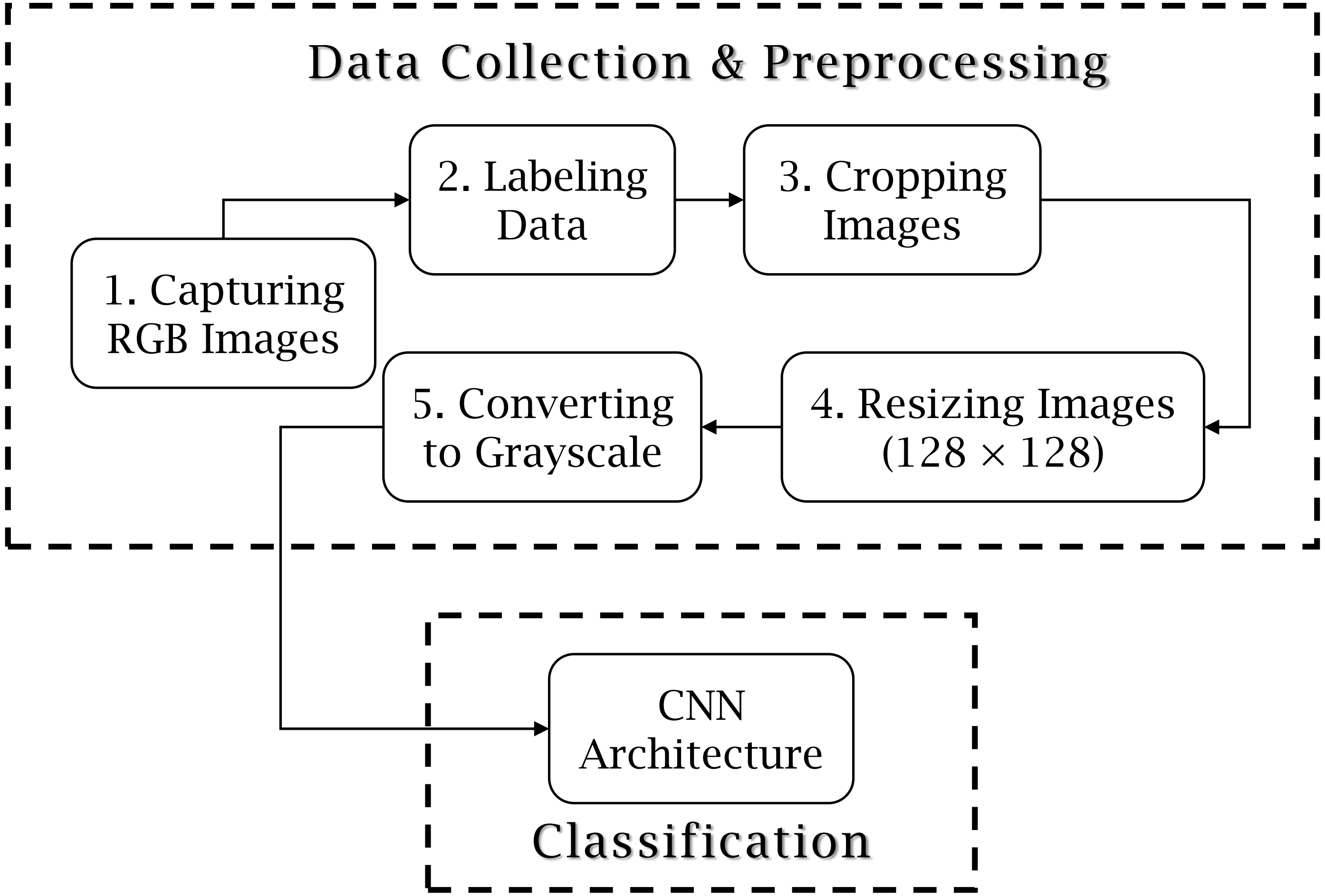
We intended to develop an open access dataset for 10 isolated sign digits from Zero (Shunno in Bangla) to Nine (Noy in Bangla) in Bangla Sign Language using depth information (i.e. x, y, and z co-ordinates of the hand key-point), in order to achieve better accuracy than the existing RGB dataset and then evaluate the performance of the proposed dataset on different classification models.

Background Study

Existing dataset for Bangla Sign Digits (Ishara-Bochon)

Overview of Ishara-Bochon

Sample type	RGB Images
Sign type	Static
Image size	128 × 128
Number of images	1000
Number of signs	10
Background	White
Classifier	CNN
Accuracy	92.87%



Background Study

Works on other sign languages using depth information

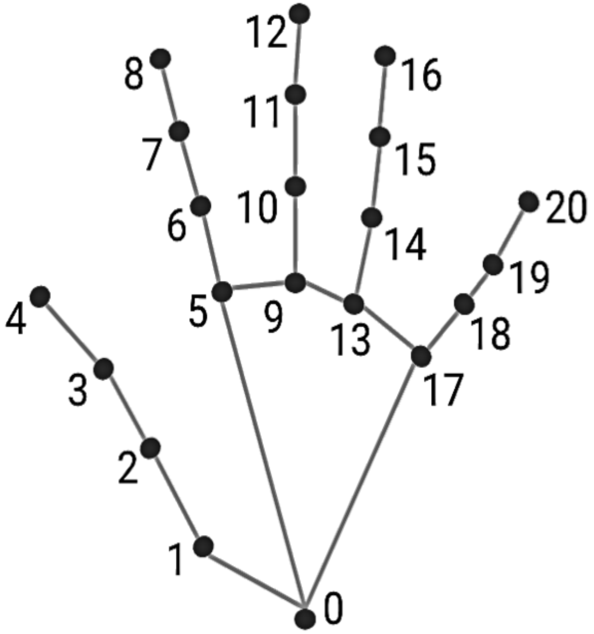
	ArSL (Arabian)	**GSL (German)**	**CSL (Chinese)**	**ASL - 1 (American)**	**ASL – 2 (American)**	**ISL (Indian)**
Type of signs	Sign Letters	Sign Letters	Sign Letters	Sign Letters	Sign Letters	Sign Digits
No. of signs	28	25	26	14	26	10
Device	Kinect, Leap Motion Controller	Kinect	Kinect, RealSense	Kinect, Leap Motion Controller	RealSense	RealSense
Classifier	NN	HMM	DNN	SVM	SVM, ANN	SVM, DTC
Dataset Size (sign*user*sample)	$28 \times 4 \times 2 = 224$	…	65000 (3 users video)	$14 \times 10 \times 100 = 14000$	$26 \times 10 \times 100 = 26000$	$10 \times 40 \times 100 = 40000$
Accuracy	99.10%	97.00%	K – 97.80% RS – 98.9%	91.28%	SVM – 95.00% ANN – 92.10%	SVM – 93.82% DTC – 85.30%
Paper	1	2	3	4	5	6

Background Study

3D hand key-point detection framework – MediaPipe

Module for hand-tracking
- MediaPipe Hand

MediaPipe Hand
underlying models –
1. Palm Detection Model
2. Hand Landmark Model



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Figure : MediaPipe Hand – 21 Hand Landmarks

Background Study

Other research works using MediaPipe

Related Work	Datasets/ Sign Gestures	Sample Type	Dataset Size	Classifier	No. of Classes	Accuracy	Max. Accuracy with Previous Approach
1	American (alphabet)	RGB Images	156000	SVM	26	99.15%	98.05%
	Indian (alphabet)	RGB Images	4972	SVM	24	99.29%	97.50%
	Italian (alphabet)	RGB Images	12856	SVM	22	98.19%	91.70%
	American (numbers)	RGB Images	1400	SVM	10	99.18%	98.05%
	Turkey (numbers)	RGB Images	4124	SVM	10	96.22%	…
2	American (hello, no, sign, understand)	RGB Videos	7200 Videos	RNN	4	92.00%	91.82%

Background Study

Generating Hand Landmarks from ASL

Dataset Statistics of the benchmark ASL dataset

Sample type	RGB Images
Sign/Hand-gesture type	Static
Number of input images	2524
Number of signers	5
Number of digit signs	10 (from 0 to 9)
Number of letter signs	24 (exp. J and Z)
Image Background	Black

Classification Results on the benchmark ASL dataset (Digits)

Classification Model	Accuracy
KNN	97.70%
SVM (Linear)	**100.00%**
SVM (Polynomial)	98.70%
SVM (RBF)	99.90%
XGB	99.70%
RFC	98.10%
DTC	96.80%
ANN	97.70%

Background Study

Comparing MediaPipe with other approaches on benchmark ASL dataset

Reference	ASL Dataset Type	Feature Extraction Method	Classifier	Accuracy (%)
1	Sign Letters (A - J)	Discriminative Zernike moments	KNN	98.51
2	Sign Digits	Convolutional Neural Network	CNN	97.00
3	Sign Digits	Convolutional Neural Network	CNN	**100.00**
4	Sign Letters	Spatial Pyramid Pooling	CNN	99.64
5	Both	YOLOv3 and VGG16	CNN	97.41
6	Sign Letters	Inceptionv3	CNN	98.81
7	Sign Digits	Gabor-Filter	CNN	99.90
	Sign Letters	Gabor-Filter	CNN	99.06
8	Both	AlexNet	SVM	97.00
	Both	VGG16	SVM	97.00
9	Sign Letters	Histogram of Gradient (HOG)	SVM	97.00
	Sign Letters	HOG-LBP (Local Binary Pattern)	SVM	97.00
	Sign Letters	Generalized Search Tree (GIST)	SVM	97.00
Ours	**Sign Digits**	**MediaPipe**	**SVM**	**100.00**

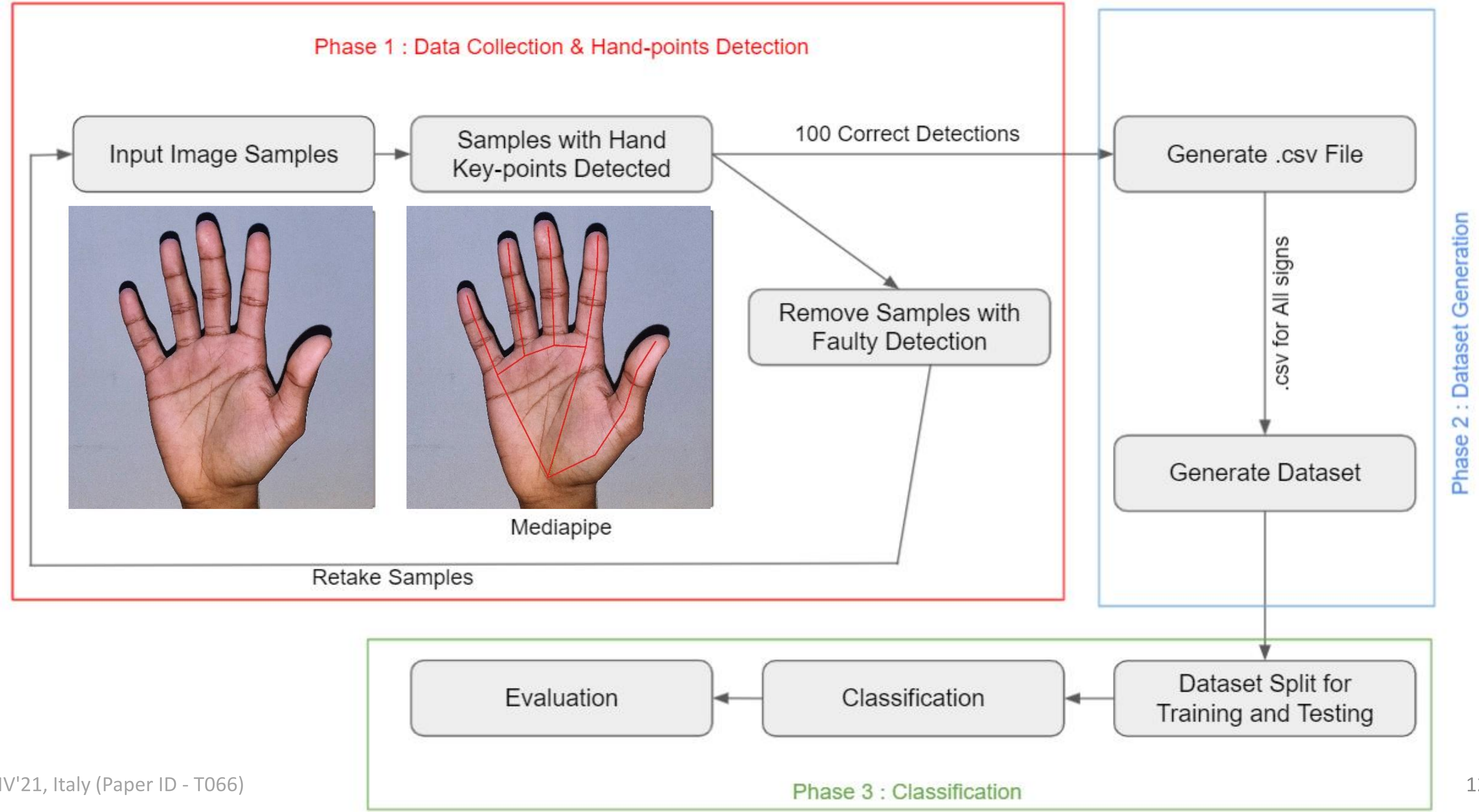
Proposed Dataset

Dataset Statistics of the proposed dataset

Type of samples	RGB Images
Type of signs/hand-gestures	Static
Size of image samples	640×480
Number of signer	10
Number of signs	10 (from 0 to 9)
Number of samples per signer	100
Total number of samples	$10 \times 10 \times 100 = 10{,}000$
Number of input features	63
Type of input features	Spatial
Number of output labels	10

Proposed Approach

Implementation is divided into 2 main parts – **Dataset Generation & Classification**



Steps of Dataset Generation

▪ Image Samples Collection

- From each user, we have collected more than 100 samples
- Images were taken using general-purpose camera

- Variants considered in sample collection
  - User Variation
    - Age, Sex, Hand-shape, Skin-color
  - Environment Variation
    - Luminance, Rotation, Scaling, Translation, Hand-facing, Background etc.

Image Sample Collection

Examples of user and environment variation

a

b

c

d

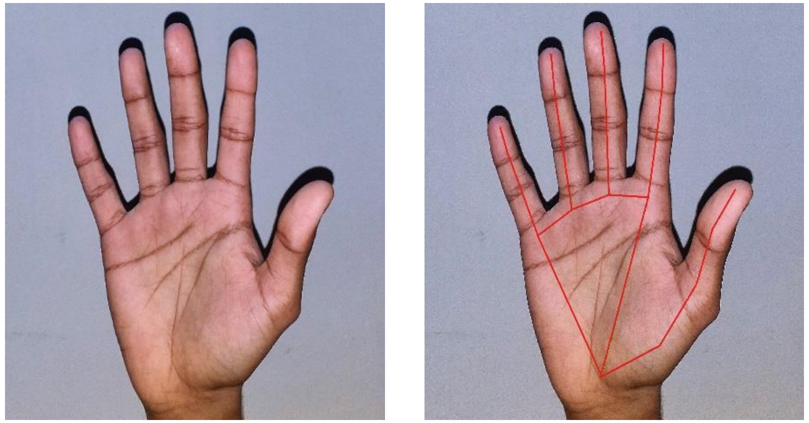
e

Dataset Variation : (a) background, (b) hand-orientation, (c) angle,
(d) luminance, (e) user (hand-shape, age, sex)

Steps of Dataset Generation

▪ Hand Key-points Detection and Preprocessing

Collected samples were processed via MediaPipe, for hand key-points detection



Original images, with corresponding images with detected hand key-points

Pre-processing

Remove samples with faulty detection

- Faulty detection may occur due to
  - Darker lighting ambience
  - Noisy/complex background
  - Sample being too hazy
  - Occlusion of certain fingers
- We removed such images manually

Steps of Dataset Generation

▪ CSV File Generation

- One .csv file per user per sign, therefore **10 users × 10 signs = 100 .csv files**
- A .csv file contains –
  - Name of the image sample (in the 1st column)
  - 3D coordinate values of 21 hand key-points (in 21 × 3 = 63 columns)
  - Label : to which sign the sample belongs (in the last column)

Single Image Frame	x00	y00	z00	...	...	...	x20	y20	z20	Label
Sign 5 - Sample (1).jpg	0.5268982053	0.7076662183	-0.0001521946833	...	...	...	0.6330968142	0.3030059636	-0.02904015779	5
Sign 5 - Sample (2).jpg	0.5247818232	0.6949129105	-0.0001708280179	...	...	...	0.6325411797	0.2845495939	-0.02152325585	5
Sign 5 - Sample (3).jpg	0.529882133	0.7028650045	-0.0001707728225	...	...	...	0.6289576292	0.2685497999	-0.03530567512	5
...	...	...	...	...	...	...	...	...	...	5
...	...	...	...	...	...	...	...	...	...	5
...	...	...	...	...	...	...	...	...	...	5
...	...	...	...	...	...	...	...	...	...	5
Sign 5 - Sample (98).jpg	0.4832853675	0.7142745256	-3.69E-05	...	...	...	0.3733857572	0.3342831135	-0.06373260915	5
Sign 5 - Sample (99).jpg	0.4827418327	0.7284878492	-6.04E-05	...	...	...	0.3589046001	0.3403604031	-0.02973021567	5
Sign 5 - Sample (100).jpg	0.4893758297	0.7470804453	-6.00E-05	...	...	...	0.3513546586	0.3379026055	-0.03841268271	5

Structure of a sample .csv file

Classification

Classification Results on proposed Bangla sign digits dataset

- 10-class classification problem
- Cross-validation applied : 10 fold
- Train-test ratio - 0.80 : 0.20
- Number of spatial features : 63
    - 3D coordinates of 21 points
- Classifiers used :
    - K-Nearest Neighbor
    - Support Vector Machine
    - Extreme Gradient Boosting
    - Random Forest Classifier
    - Decision Tree Classifier

Classification Model	Accuracy
KNN	94.60%
SVM (Linear)	98.00%
SVM (Polynomial)	93.35%
SVM (RBF)	98.65%
XGB	98.00%
RFC	97.75%
DTC	96.25%

Classification

Classification Results on individual sign digits

Classification Model	Sign 0 (%)	Sign 1 (%)	Sign 2 (%)	Sign 3 (%)	Sign 4 (%)	Sign 5 (%)	Sign 6 (%)	Sign 7 (%)	Sign 8 (%)	Sign 9 (%)
KNN	98.09	95.91	99.50	99.47	97.57	97.99	99.50	96.48	99.03	100
SVM (Lin)	99.04	97.76	98.51	100	98.52	98.51	100	97.93	99.03	100
SVM (Poly)	99.03	97.75	95.69	97.42	93.43	94.95	99.50	98.96	97.09	100
SVM (RBF)	**99.52**	**99.09**	**99.01**	**100**	**99.51**	**98.99**	**99.50**	**100**	**99.51**	**100**
XGB	99.52	97.29	99.50	99.47	98.54	98.49	100	97.44	100	100
RFC	99.04	98.20	98.52	99.47	96.63	98.49	100	97.93	98.54	100
DTC	95.22	98.64	99.50	97.93	95.67	92.72	99.50	97.93	98.54	100

Result Analysis

Incorrect classification :

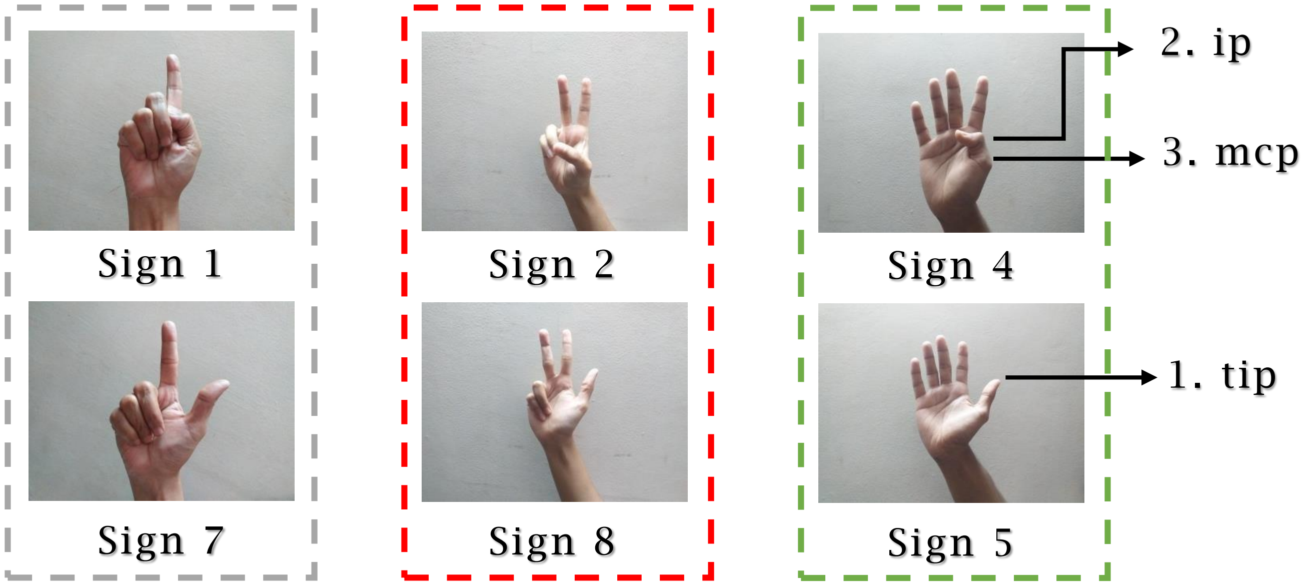
Mostly between –

- Sign 1 and 7
- Sign 2 and 8
- Sign 4 and 5

Reason of incorrect classification

Co-ordinate values are similar except for 3 thumb finger landmarks

1. Thumb tip
2. Thumb inter phalangeal
3. Thumb meta-carpo phalangeal



Comparison with existing Bangla sign digits dataset

	Ishara-Bochon Dataset	Proposed Dataset
Sample Type	Images	Images
Image Type	RGB	RGB
Sign Type	Static	Static
Dataset Size	1000 Images	10000 Images 100 .csv files
Image Size	128×128	640×480
Image Background	White	Various
Feature Extraction Method	CNN	MediaPipe
Classification Model	CNN	SVM
Accuracy	92.87%	98.65%

Proposed dataset –

- Larger in size
- More challenging – with user and environment variation
- Takes less computation time
- Yields better accuracy

Comparison with existing Bangla sign digits dataset

For further ground level comparison, we have evaluated the performance of our dataset on the same CNN architecture, that was applied on Ishara-Bochon –

- 10000 RGB Images were used
- Images were resized to 128 × 128, same as the Ishara-Bochon
- Then images were converted from RGB to Grayscale
- Same CNN architecture was used, with same hyper-parameter values

Proposed dataset provides better accuracy in lesser time (epochs)

Datasets on Same CNN architecture

	Ishara-Bochon Dataset	Proposed Dataset
Sample Type	Images	Images
Image Type	Grayscale	Grayscale
Sign Type	Static	Static
Dataset Size	1000	10000
Image Size	128 × 128	128 × 128
Background	White	Various
Epochs	500	100, 50
Optimizer	Adam	Adam, SGD
Accuracy (%)	92.87	97.70 (Adam, 50) 99.10 (Adam, 100) 99.30 (SGD, 100)

Conclusion and Future Work

- In this paper, we have only focused on 10 sign digits
- Apart from sign digits, in Bangla Sign Language, there are –
  - 38 One-handed signs of sign alphabet
  - 36 Two-handed signs of sign alphabet
  - Dynamic signs of Bangla sign alphabet

- Our next focus is on developing a complete dataset on One-handed signs of Bangla sign alphabet using depth information, followed by another dataset on Two-handed signs of Bangla sign alphabet

- We also intend to continue our research on Hand Gesture Recognition by working on other sign languages, as well as other static and dynamic sign gestures

Questions ?

References

[1] M.S. Islam, S.S.S. Mousumi, N.A. Jessan, A.S.A. Rabby and S.A Hossain (2019). Ishara Bochon: The First Multi-purpose Open Access Dataset for Bangla Sign Language Isolated Digits. 420–428. https://doi.org/10.1007/978-981-13-9181-1_37

[2] M.S. Islam, S.S.S. Mousumi, N.A. Jessan, A.S.A. Rabby and S.A Hossain (2018). Ishara Lipi: The First Complete Multipurpose Open Access Dataset of Isolated Characters for Bangla Sign Language. In 2018, International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE, 1–4.

[3] A. Halder and A. Tayade (2021). Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning. Journal homepage: https://www.ijrpr.com ISSN 2582 ([n. d.]), 7421.

[4] A.L.C. Barczak, N.H. Reyes, M. Abastillas, A. Piccio and T. Susnjak (2011). A New 2D Static Hand Gesture Colour Image Dataset for ASL Gestures. Research Letters in the Information and Mathematical Sciences 15 (2011), 12–20. http://www.massey.ac.nz/massey/fms/Colleges/College%20of%20Sciences/IIMS/RLIMS/Volume%2015/GestureDatasetRLIMS2011.pdf

[5] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard and V. Athitsos (2011). Comparing gesture recognition accuracy using color and depth information. In Proceedings of the 4th international conference on Pervasive technologies related to assistive environments. 1–7.

[6] N. Pugeault and R. Bowden (2011). Spelling it out: Real-time ASL fingerspelling recognition. In 2011 IEEE International conference on computer vision workshops (ICCV workshops). IEEE, 1114–1119.

[7] K.O. Rodriguez and G.C. Chavez (2013). Finger spelling recognition from RGB-D information using kernel descriptor. In 2013 XXVI Conference on Graphics, Patterns and Images. IEEE, 1–7.

[8] W. Nai, Y. Liu, D. Rempel and Y. Wang (2017). Fast hand posture classification using depth features extracted from random line segments. Pattern Recognition 65 (2017), 1–10.

[9] H. Liang and J. Yuan (2014). Hand parsing and gesture recognition with a commodity depth camera. In Computer Vision and Machine Learning with RGB-D Sensors. Springer, 239–265.

[10] S. Mudduluru (2017). Indian Sign Language Numbers Recognition using Intel RealSense Camera. (2017).

[11] H. Al-Nuaim and M.A. Almasre (2016). A Real-Time Letter Recognition Model for Arabic Sign Language Using Kinect and Leap Motion Controller v2. International Journal of Advanced Engineering, Management and Science 2 (2016).

[12] S. Lang, M. Block and R. Rojas (2012). Sign language recognition using kinect. In International Conference on Artificial Intelligence and Soft Computing. Springer, 394–402.

[13] J. Mistry and B. Inden (2018). An Approach to Sign Language Translation using the Intel RealSense Camera. In 2018 10th Computer Science and Electronic Engineering (CEEC). 219–224. https://doi.org/10.1109/CEEC.2018.8674227

[14] B. Liao, J. Li, Z. Ju and G. Ouyang (2018). Hand gesture recognition with generalized hough transform and DC-CNN using realsense. In 2018 Eighth International Conference on Information Science and Technology (ICIST). IEEE, 84–90.

[15] A.D. López (2020). Sign Language Recognition - ASL Recognition with MediaPipe and Recurrent Neural Networks. B.S. thesis. Universitat Politècnica de Catalunya.

[16] M.A. Aowal, A.S. Zaman, S.M.M. Rahman and D. Hatzinakos (2014). Static Hand Gesture Recognition Using Discriminative 2D Zernike Moments. IEEE Region 10 Annual International Conference, Proceedings/TENCON 2015. https://doi.org/10.1109/TENCON.2014.7022345

[17] V. Bheda and D. Radpour (2017). Using Deep Convolutional Networks for Gesture Recognition in American Sign Language. ArXiv abs/1710.06836 (2017).

References

[18] S.R. Kalbhor and A.M. Deshpande (2018). Digit Recognition Using Machine Learning and Convolutional Neural Network. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). 604–609. https://doi.org/10.1109/ICOEI.2018.8553954

[19] Y.S. Tan, K.M. Lim, C. Tee, C.P. Lee and C.Y. Low (2021). Convolutional neural network with spatial pyramid pooling for hand gesture recognition. Neural Computing and Applications 33 (05 2021), 1–13. https://doi.org/10.1007/s00521-020-05337-0

[20] H. Luqman, M.E. El-Alfy and G. Binmakhashen (2021). Joint space representation and recognition of sign language fingerspelling using Gabor filter and convolutional neural network. Multimedia Tools and Applications 80 (03 2021), 1–22. https://doi.org/10.1007/s11042-020-09994-0

[21] P. Das, T. Ahmed, M.F. Ali (2020). Static Hand Gesture Recognition for American Sign Language using Deep Convolutional Neural Network. 1762– 1765. https://doi.org/10.1109/TENSYMP50017.2020.9230772

[22] M.M. Hasan, A.K. Srizon, A. Sayeed and M.A.M. Hasan (2020). Classification of American Sign Language by Applying a Transfer Learned Deep Convolutional Neural Network. https://doi.org/10.1109/ICCIT51783.2020.9392703

[23] S. Sharma, H.P.J. Dutta, M. Bhuyan and R. Laskar (2020). Hand Gesture Localization and Classification by Deep Neural Network for Online Text Entry. https://doi.org/10.1109/ASPCON49795.2020.9276713

[24] A.A. Barbhuiya, R.K. Karsh and R. Jain (2021). CNN based feature extraction and classification for sign language. Multimedia Tools and Applications 80 (01 2021), 1–19. https://doi.org/10.1007/s11042-020-09829-y

[25] P.M. The and M.T. Yu (2021). Static and Dynamic Hand Gesture Recognition Using GIST and Linear Discriminant Analysis. Volume 14 (06 2021), 123. https://doi.org/10.22266/ijies2021.0831.12

[26] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran and M. Grundmann (2019). Blazeface: Sub-millisecond neural face detection on mobile gpus. arXiv preprint arXiv:1907.05047 (2019).

[27] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.L. Chang and M. Grundmann (2020). MediaPipe Hands: On-device Real-time Hand Tracking. arXiv:2006.10214 [cs.CV]

[28] T.Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and G. Belongie (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2117–2125.

[29] T. Simon, H. Joo, I. Matthews and Y. Sheikh (2017). Hand key-point detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 1145–1153.

[30] W.U. Wickramaarachchi and S. Vasanthapriyan, "Multi-Layer Framed Offline Signature Recognition Algorithm," Journal of Image and Graphics, Vol. 3, No. 1, pp. 11-15, June 2015. doi:10.18178/joig.3.1.11-15

[31] A.M. Ceniza, T.K.B. Archival and K.V. Bongo, "Mobile Application for Recognizing Text in Degraded Document Images Using Optical Character Recognition with Adaptive Document Image Binarization," Journal of Image and Graphics, Vol. 6, No. 1, pp. 44-47, June 2018. doi:10.18178/joig.6.1.44-47

[32] M. Liu, G. Liu, Y. Liu and Q. Jiao, "Oracle Bone Inscriptions Recognition Based on Deep Convolutional Neural Network," Journal of Image and Graphics, Vol. 8, No. 4, pp. 114-119, December 2020. doi:10.18178/joig.8.4.114-119

Thank You