Indian Sign Language Translator Using Residual Network and Computer Vision

Shrinidhi Gindi[1], Amina Bhatkar[2], S Hasan Haider[3], Ramsha Ansari[4]

[1]Professor, Dept. of Information Technology, MH Saboo Siddik College of Engineering, Maharashtra, India
[2]Student, Dept. of Information Technology, MH Saboo Siddik College of Engineering, Maharashtra, India
[3]Student, Dept. of Information Technology, MH Saboo Siddik College of Engineering, Maharashtra, India
[4]Student, Dept. of Information Technology, MH Saboo Siddik College of Engineering, Maharashtra, India

---***---

Abstract - *Sign Language communication consists of movements of fingers and hands, which may be affected by noise such as lighting effects. Sign Language is used by the hearing and speech impaired as a means of communication and is in fact the only medium through which they can communicate. This means that people who do not know the Sign Language will not be able to directly communicate with them. Moreover, sign language varies according to countries and regions. This paper proposes a novel approach for two way Indian Sign Language (ISL) communication. The Sign Language video captured is pre-processed to detect motion frames and reduce the number of frames to be processed. Then, frames captured are forwarded to the trained model which predicts the sign in the image. To find the sign of a given word or phrase, it is typed as input. The phrase is split into words. Since articles and helping words do not have signs they are ignored. Sign images are stored on cloud. Images for words are fetched from the cloud and are then displayed. If the word doesn't exist in the list of supported words, it is split into letters and images of letters are fetched from the cloud and are displayed.*

***Key Words*: Indian Sign Language, Sign Language Translator, Sign to Text Conversion, Text to Sign Conversion.**

1. INTRODUCTION

The hearing and speech impaired form a large as well as a significant part of the population. Not being able to communicate has isolated them as they communicate only with the people who can understand sign language or there arises a need of a translator. The approach proposed in this paper consists of two modules: Sign-to-Text (STT) and Text-to-Sign (TTS). The STT module faces a lot of challenges that includes appropriate pattern matching, the reason being the fact that images suffer from noise, occlusions etc. The Indian Sign Language (ISL) is derived from the British Sign Language (BSL) and the French Sign Language (FSL) and thus there exists no standard dataset for ISL and there are no existing systems for two-way communication with a speech impaired person that could process data in real time.

Sign Language Recognition (SLR) has always been a subject of extensive research. Researchers in the past have used 1D, 2D & 3D sensors. Although, 1D and 2D sensors are cheaper and thus popular, the systems are not as efficient due to various reasons. 1D data captures very little information, i.e. only the finger movements and 2D data suffer from a lot of noise. The major limitations of using 2D sensors for sign language recognition are discussed in the literature [1].

Researchers later proposed fusion of 3D depth and RGB data to capture sign language with increased efficiency [2][3]. Although the efficiency increases, the cost increases significantly. This literature is an example of 3D depth and RGB fusion model [4]. The system captures sign language gestures as 3D motionlets differentiated on body position and the number of hands used. Researchers have also considered shape, texture and local movement along with YCbCr skin color model for sign language recognition [5].

Other approach includes Dynamic Sign Language Recognition [6]. This comprises of two main sub modules: the Image Processing module and the Stochastic Linear Formal Grammar module. This is applied on a robot, Pumpkin.

Through this paper, we propose a novel approach for two-way communication with the hearing and speech impaired. The two main modules are Sign-to-Text (STT) and Text-to-Sign (TTS). The TTS accepts input sentence, splits it into words. Helping words and other grammar are simply ignored. Then the word is searched for in the database, if an image result exists, it is displayed. Otherwise, the word is split into letters, and images for the letters are displayed. The STT is again divided into three main submodules: Motion Frame Detection, Frame Reductor and Pattern Matching. The motion frame detection module captures only the moving parts in the video, successfully reducing the length of the video to be processed. The frame reductor saves only one frame per second, which is the mid-frame.The last submodule is pattern matching, making use of ResNet50 of ImageAI. It matches the input frame with the existing ones in the dataset and returns text as output.

The rest of the paper is organized as follows: Section 2 discusses work that has been done in this field by various researchers. Section 3 discusses the proposed methodology of the system. Section 4 highlights the results obtained followed by a conclusion in Section 5.

2. RELATED LITERATURE

Sign language consists of human gestures that denote a word. For example, showing thumbs up means good in ISL. Sign language is perceived by the eyes. In the past years, a lot of research has been done in this field. Researchers have used 1D, 2D and 3D sensors, leap motion controllers, devices such as Microsoft Kinect and also used a depth and RGB fusion model.

Leap motion controllers use infrared cameras and mathematical algorithms to translate hand and finger movements to 3D input. This technology is utilized in the literature [7][8]. Leap motion sensors are used in [9] for finger-spelling using American Sign Language with recognition rate of 82.71%. Leap motion data was combined with Microsoft Kinect or RGB depth data to improve the recognition rate. In [10], the recognition rate is 98% for Arabic alphabet. Leap motion sensor along with depth sensor are also used in [11].

Literature in [12] suggests the use of transition movement models along with dynamic time warping and temporal clustering algorithm for Chinese Sign Language Recognition with an average accuracy of 91.9%. In [13], discussion is done on the various techniques used for hand detection. They have identified the techniques as: appearance based, model based, soft computing and other approaches. Soft computing approach includes Artificial Neural Network, fuzzy logic and genetic algorithm. The paper analyses all the different approaches. In [14], focus is laid on a large number of 3D hand postures, the system developed is GREFIT. In [15] finite state machines and fuzzy logic are used to extract fingertips.

A model based approach discussed in the paper is [16] which makes used of a histogram for skin color observation. The artificial neural network approaches [17], fuzzy logic [18][19], and genetic algorithm-based approaches [20] are also discussed in the paper.

[21] have proposed a novel approach for New Zealand Sign Language without using any marker. Haar features and AdaBoost are used [22][23] for dynamic hand classification. Bag of Features (BOF) was used for hand gesture detection and recognition [24]. Fuzzy neural networks were used in [25] for hand gestures. Inverse projection matrices and inverse kinematics were used to calibrate a hand model [26]. Use of Markov Random Field is seen in [27] to remove noise component in processed figure.

Text to Sign translation for Arabic Sign Language [28] is done. A mobile application was developed that took text as input and output was displayed as animated avatar. Architecture for translation of English into a DRS-based intermediate semantic representation for BSL is discussed in [29][30].

Use of latest technology, X3D is done to build a mobile application for text to sign language translation [31]. It allows receiving short text messages and sign language animation.

Speech to sign translation for Spanish is discussed in [32]. The output is displayed as an animated avatar.

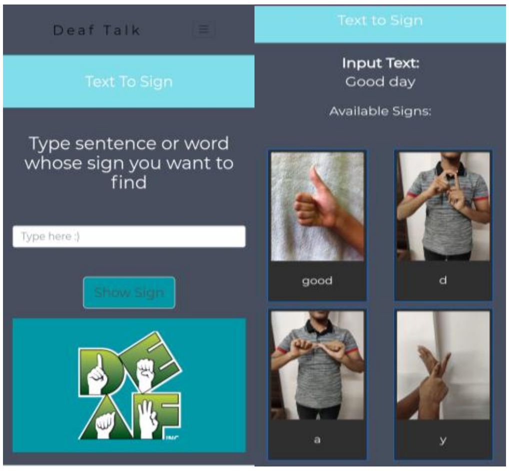
3. PROPOSED METHODOLOGY

We have divided the project into two main modules: STT & TTS. The STT has three sub-modules: Motion detection, Frame Reduction & Pattern Matching.

3.1 Text to Sign Module

For TTS, the user is asked to add a text input. The input sentence is split into words. The words used for grammar such as 'is', 'are', 'am', 'a', 'an', 'the' are ignored. For the rest of the sentence, the words are considered one by one; if the action for the word exists in the database the image is returned. If no image exists, the word is split into letters and the image of the letter is displayed.

The text entered by the user is also printed alongside the image and the image is displayed for a few seconds; the time can be changed.



Type sentence or word whose sign you want to find

Input Text:
Good day

Fig -1: Displaying Signs

3.2 Sign to Text Module

The STT is divided into three submodules: Motion Frame Detection, Frame Reductor & Pattern Matching. The aim of Motion Frame Detection is to identify the frames that contain gesture information and discard the still frames in order to reduce the number of frames to be processed. Frame Reductor algorithm is used to reduce the number of redundant frames to save processing time and cost.

3.2.1 Motion Frame Detection

The aim is to identify frames that contain non-stationary movement/object and discard the ones that are stationary frames. Background/foreground segmentation algorithm is used for this. It compares the current frame with the previous one of the video and estimates changes and saves only the changes, not the whole video. This is done for real time tracking of moving objects [33][34]. Background subtraction involves calculating a reference image, subtracting each new frame from this image and discarding the ones that are below the threshold.

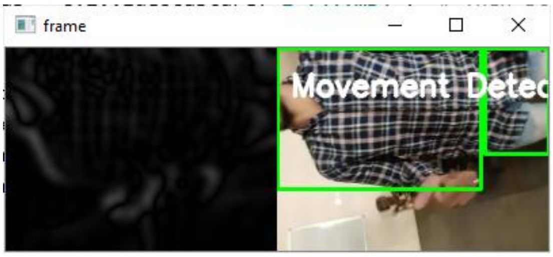


Fig -2: Motion Frame Detection

The first frame is taken as the initial frame and the absolute difference is calculated which will give the value that is used for defining the movement in frame. Contours are calculated on thresh value, having same color or intensity. When there is movement in the video, that particular frame is captured and stored in another video which is then used for further processing.

Gaussian mixture model can also be used for the background subtraction of the video frames to get the frames which automatically select the number of components per pixel [35].

3.2.2 Frame Reductor

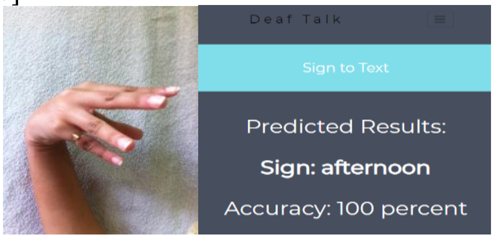
With the advancement in technology and the availability of high quality video frames, the frame rate has increased to about 60 frames per second (fps) [36]. This leads to a large number of redundant frames which increases the time and cost of processing. The aim of Frame Reductor is to reduce the number of redundant frames by saving only one frame per second. Firstly, the frame rate is calculated and only the mid-frame is saved for further processing.

Motion detection and frame reduction are background processes, not visible to the user.

3.2.3 Pattern Matching

Pattern Matching and model training allow for identification of similarities in an image. The model is trained to identify gestures and based on the trained data; the sign of the input word is recognized.

We have made use of the ResNet algorithm from ImageAI[37]. ResNet is a fast algorithm of size 98MB, with less prediction time and high accuracy. The main innovation of ResNet is the skip connection. The skip connection in the diagram below is labeled “identity.” It allows the network to learn the identity function, which allows it to pass the input through the block without passing through the other weight layers [38].



Predicted Results:

Sign: afternoon

Accuracy: 100 percent

Fig -3: Predicted Output for sample image

Our model uses Residual Network with 50 layer having channel depth of 256, 512, 1024, 2048 for pooling entire model and filters of 64,128,256,512 values in 2D convolution layer.

For the dataset, 4000 images per sign, including no sign class (taking 5 to 6 class/sign for new sample model). Accuracy of trained model for sign prediction of 15 words is 99.09% with batch size 8 and 50 number of experiments/epochs for training the model.

4. RESULTS AND DISCUSSION

Table -1: Result of Sample1 for STT Module

Sign To Text	Sample 1
Video length (in sec)	0.35
Number of frames before motion detection	745
Number of frames after motion detection	331
Reduced frames	66
Predicted signs of extracted frames	Good
Maximum accuracy of predicted signs	100

Table -2: Result of Samples for TTS Module

Text To Sign	Sample 1	Sample 2	Sample 3
Sign input	good	good morning	bla bla
Type of input	1 word present in storage	Sentence having words present in storage	Sentence not having meaning or not present in storage
Output Image shown	Good for 5 sec	Good for 5 sec and then morning for 5 sec	Sign of each alphabets shown in a sequence of 5 sec
Accurate	Yes	Yes	Yes

5. CONCLUSIONS

The system allows two way translation of Indian Sign Language in a smooth and efficient manner. Use of free and open-source technologies has enabled us to build a system that is inexpensive and can be widely used by ordinary people in India.

The TTS module is a complete subsystem and can be easily used by an ordinary person to convey a message to the hearing and/or speech impaired. The use of cloud storage ensures that the image files of signs are not stored locally on the user's device and no local storage memory is used.

The problem with ISL is that it is ambiguous and there exists no standard database. Due to this, the STT module suffers a lot and using a custom dataset takes up a lot of time. As of now, only 15 words can be supported for STT. The preprocessing videos and images are not stored locally; eliminating the use of local storage memory.

Our near future perspective work includes addition of more words and dynamic gestures in order to support the complete ISL.

REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311–324, 2007.

[2] C. Sun, T. Zhang, B.-K. Bao, C. Xu, and T. Mei, "Discriminative exemplar coding for sign language recognition with kinect," IEEE Transactions on Cybernetics, vol. 43, no. 5, pp. 1418–1428, 2013.

[3] G. Plouffe and A.-M. Cretu, "Static and dynamic hand gesture recognition in depth data using dynamic time warping," IEEE transactions on instrumentation and measurement, vol. 65, no. 2, pp. 305–316, 2016.

[4] P. V. V. Kishore, D. A. Kumar, A. S. C. S. Sastry and E. K. Kumar, "Motionlets Matching With Adaptive Kernels for 3-D Indian Sign Language Recognition," in IEEE Sensors Journal, vol. 18, no. 8, pp. 3327-3337, 15 April15, 2018.

[5] J. Rekha, J. Bhattacharya and S. Majumder, "Shape, texture and local movement hand gesture features for Indian Sign Language recognition," 3rd International Conference on Trendz in Information Sciences & Computing (TISC2011), Chennai, 2011, pp. 30-35.

[6] M. R. Abid, E. M. Petriu and E. Amjadian, "Dynamic Sign Language Recognition for Smart Home Interactive Application Using Stochastic Linear Formal Grammar," in IEEE Transactions on Instrumentation and Measurement, vol. 64, no. 3, pp. 596-605, March 2015.

[7] M. Mohandes, S. Aliyu, and M. Deriche, "Arabic sign language recognition using the leap motion controller," in Industrial Electronics (ISIE), 2014 IEEE 23rd International Symposium on. IEEE, 2014, pp. 960–965.

[8] C.-H. Chuan, E. Regina, and C. Guardino, "American sign language recognition using leap motion sensor," in Machine Learning and Applications (ICMLA), 2014 13th International Conference on. IEEE, 2014,pp. 541–544.

[9] M. Funasaka, Y. Ishikawa, M. Takata, and K. Joe, "Sign language recognition using leap motion controller," in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015, p. 263.

[10] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to arabic sign language recognition," IEEE transactions on human-machine systems, vol. 44, no. 4, pp. 551–557, 2014.

[11] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with jointly calibrated leap motion and depth sensor," Multimedia Tools and Applications, vol. 75, no. 22, pp. 14 991–15 015, 2016.

[12] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," IEEE transactions on systems, man, and cybernetics-part a: systems and humans, vol. 37, no. 1, pp. 1–9, 2007.

[13] A. Chaudhary, J. L. Raheja, K. Das, and S. Raheja, "A survey on hand gesture recognition in context of soft computing," in Proc. CCSIT, vol. 133. Jan. 2011, pp. 46–55.

[14] C. Nolker and H. Ritter, "Visual recognition of continuous hand postures," IEEE Trans. Neural Netw., vol. 13, no. 4, pp. 983–994, Jul. 2002.

[15] Verma, R., Dev, A.: Vision based Hand Gesture Recognition Using finite State Machines and Fuzzy Logic. In: International Conference on Ultra-Modern Telecommunications & Workshops, October 12-14, pp. 1–6 (2009)

[16] A. El-Sawah, N. D. Georganas, and E. M. Petriu, "A prototype for 3-D hand tracking and posture estimation," IEEE Trans. Instrum. Meas., vol. 57, no. 8, pp. 1627–1636, Aug. 2008.

[17] Lee, D., Park, Y.: Vision-Based Remote Control System by Motion Detection and Open Finger Counting. IEEE Transactions on Consumer Electronics 55(4), 2308–2313 (2009)

[18] Trivino, G., Bailador, G.: Linguistic description of human body posture using fuzzy logic and several levels of abstraction. In: IEEE Conference on Computational Intelligence for Measurement Systems and Applications, Ostuni, Italy, June 27-29, pp. 105–109 (2007)

[19] Schlomer, T., et al.: Gesture recognition with a Wii Controller. In: Proceedings of the 2nd International Conference and Embedded Interaction, Bonn, Germany, February 18-20, pp. 11–14 (2008)

[20] C. Hu, Q. Yu, Y. Li, and S. Ma, "Extraction of parametric human model for posture recognition using genetic algorithm," in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., Mar. 2000, pp. 518–523.

[21] R. Akmeliawati et al., "Towards real-time sign language analysis via
markerless gesture tracking," in Proc. Instrum. Meas. Technol. Conf.,
May 2009, pp. 1200–1204.

[22] Q. Chen, N. D. Georganas, and E. M. Petriu, "Real-time vision-based hand gesture recognition using Haar-like features," in Proc. Instrum.
Meas. Technol. Conf., May 2007, pp. 1–6.

[23] Q. Chen, N. D. Georganas, and E. M. Petriu, "Hand gesture recognition using Haar-like features and a

stochastic context-free grammar," IEEE Trans. Instrum. Meas., vol. 57, no. 8, pp. 1562–1571, Aug. 2008.

[24] N. H. Dardas and N. D. Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," IEEE Trans. Instrum. Meas., vol. 60, no. 11, pp. 3592–3607, Nov. 2011.

[25] A. R. Varkonyi-Koczy and B. Tusor, "Human–computer interaction for smart environment applications using fuzzy hand posture and gesture models," IEEE Trans. Instrum. Meas., vol. 60, no. 5, pp. 1505–1514,
May 2011.

[26] C. Joslin, A. El-Sawah, Q. Chen, and N. Georganas, "Dynamic gesture recognition," in Proc. IEEE IMTC, May 2005, pp. 1706–1711.

[27] Zhou, H., Ruan, Q.: A Real-time Gesture Recognition Algorithm on Video Surveillance. In: 8th International Conference on Signal Processing (2006)

[28] Halawani, S.M., 2008. Arabic sign language translation system on mobile devices. IJCSNS International Journal of Computer Science and Network Security, 8(1), pp.251-256.

[29] Sáfár, É. and Marshall, I., 2001, September. The architecture of an English-text-to-Sign-Languages translation system. In Recent Advances in Natural Language Processing (RANLP) (pp. 223-228). Tzigov Chark Bulgaria.

[30] Marshall, I. and Sáfár, É., 2003, July. A prototype text to British Sign Language (BSL) translation system. In The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (pp. 113-116).

[31] Boulares, M. and Jemni, M., 2012, April. Mobile sign language translation system for deaf community. In Proceedings of the international cross-disciplinary conference on web accessibility (pp. 1-4).

[32] San-Segundo, R., Barra, R., Córdoba, R., d'Haro, L.F., Fernández, F., Ferreiros, J., Lucas, J.M., Macías-Guarasa, J., Montero, J.M. and Pardo, J.M., 2008. Speech to sign language translation system for Spanish. Speech Communication, 50(11-12), pp.1009-1020.

[33] Friedman, N. and Russell, S., 2013. Image segmentation in video sequences: A probabilistic approach. arXiv preprint arXiv:1302.1539.

[34] Zivkovic, Z. and Van Der Heijden, F., 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern recognition letters, 27(7), pp.773-780.

[35] Zivkovic, Z., 2004, August. Improved adaptive Gaussian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 2, pp. 28-31). IEEE.

[36] Wikipedia contributors. Frame rate. Wikipedia, The Free Encyclopedia. January 1, 2020, 19:15 UTC. Accessed on: Feb. 4, 2020. Available at: https://en.wikipedia.org/w/index.php?title=Frame_rate&oldid=933559738.

[37] M. Olafenwa, J. Olafenwa. ImageAi 2.0.3. Accessed on: Feb. 1, 2020. [Online]. Available at: https://imageai.readthedocs.io/en/latest/

[38] Wikipedia contributors. Residual neural network. Wikipedia, The Free Encyclopedia. January 7, 2020, 14:35 UTC. Accessed February 4, 2020. Available at: https://en.wikipedia.org/w/index.php?title=Residual_neural_network&oldid=934615040.