Real Time Sign Language Translation

Tejal Deshpande[1], Amit Marathe[2], Ronak Singh Rajput[3], Shubham Maurya[4] and Ninad Tawade[5]
[1]*Assistant Professor, Dept. of Electronics and Telecommunication Engineering, Xavier Institute of Engineering, Maharashtra, India*
[2,3,4,5]*Student, Dept. of Electronics and Telecommunication Engineering, Xavier Institute of Engineering, Maharashtra, India*

Abstract **- Sign language is a very significant non-verbal communication medium for people with hearing impairment. Real time translation will be very useful tool for such people since it will ease communication and save some time. In this paper, we present a concept of hand movement detection and identification for sign language in real time. Proposed system will detect the gestures or hand movements and identify its meaning in sign language and convert it into words and thus form meaningful English sentences. With this we can narrowed down communication gap for people with hearing disability with hearing disability.**

Index Terms **- Convolutional Neural Networks, Dataset Augmentation, Generative Pre-trained Transformer, Natural Language Generation.**

I.INTRODUCTION

Sign languages are authentic languages that have evolved among the individuals with hearing or speech impairment. When the impediments to verbal communication are permanent and lifelong, oral means of communication cripples. While the spoken language follows a definite syntactical order, which is determined by the different parts of speech involved in an expression, sign language expressions are more morphologically oriented with the order being determined by the semantically based material classification, as well as the shape of the signs.

Two-dimensional human pose estimation is a visual recognition task dealing with the autonomous localization of anatomical human joints or "key points" in RGB images and videos. Identification of the human body parts involves many challenges especially when socially engaged individuals are involved. It depends on the activity being performed. Identification becomes easier for a single person performing as compared to a group of 10. The limbs of the people performing a particular action might overlap with the other, making the association of parts quite rigorous. This challenge intensifies with the involvement of multiple people in the frame. Bottom-up approaches are efficient and have the potential to decouple runtime complexity by detecting key points from the number of individuals.

In this paper, we propose a system that would help identify and teach Indian Sign Language to the masses. Once the human body is identified, parts of the images containing the hand would be extracted and run through a model. The system detects the key points of the hands and the output would be the coordinates that would be fed to classification models followed by language models. The vocabulary list will have more than 7,000 signs that deal with words used in medical, technical, academic, legal, and routine conversations by the deaf in India. This system has the potential of teaching sign language and help all the hearing and speech impaired people.

II. LITERATURE REVIEW

A. Sign Language Recognition techniques with Machine learning

Sign language is the mode of communication which uses visual ways like expressions, hand gestures, and body movements to convey meaning.[1]. Sign language recognition refers to the conversion of these gestures into words or alphabets of existing formally spoken languages. Thus, conversion of sign language into words by an algorithm or a model can help bridge the gap between people with hearing or speaking impairment and the rest of the world.

B. Natural Language Generation

NLG is the process of producing a human language text response supported by some data input. In its

essence, it automatically generates narratives that describe, summarize or explicate input structured data in a humanoid manner at the very high speed of thousands of pages per second. The most common use of natural language generation technology is to create computer systems that present information to people in a representation that they and easy to comprehend. Internally, computer systems use techniques which are straightforward for them to manipulate, such as airline schedule databases, accounting spreadsheets, expert system knowledge bases, grid-based simulations of physical systems, and so forth [2]. In many cases, however, these delineations of information require a distinguished amount of expertise to interpret. This means that there is often a requirement for a mechanism which can present such data in a meaningful form to amateur user.

C. Study of Pose Detection Techniques

SLAM -simultaneous localization and mapping, camera calibration and SfM- Structure from Movement. Keypoint discovery has a long history originating before profound learning, and numerous incredible calculations in wide industry applications (like ORB, SIFT, and FAST) depend close by made highlights. As in numerous other PC vision undertakings, individuals have been investigating profound figuring out how to beat hand-created calculations.

Deep learning has overwhelmed cutting edge semantic keypoint identification. Cover R CNN (ICCV 2017) and PifPaf (CVPR 2019) are two delegate strategies for identifying semantic central issues. The two cycles are regulated learning and need broad and costly human comment. This makes the utilization of keypoints location testing since interest focuses are semantically poorly characterized; hence a human annotator can't dependably distinguish similar arrangement of keypoints. Subsequently, it is difficult to define revenue point location as a directed learning issue.

The overall prediction of expectation, different models predict keypoints and this machine performs regressions [3]. Gaining from the past model, some unacceptable could be rectified after some time. Thus, for the following one, the overall thought is something very similar from the past however now we are utilizing convolutional neural organizations is more qualified.

Essentially, stacked AE on top of each other to perform relapse toward the finish of the layer [4]. Here, the Liking field is utilized to perform relapse on the human posture. There is an execution of this technique and by and large, it is sufficient [5]. They joined the posture data into division since, in a great deal of recordings, when an individual is covered by another, the presentation corrupts.

III. METHODOLOGY

A. Dataset:

1. Datasets for Alphabet: The Alphabet dataset is in place as the users of sign language occasionally require to convey words by the spelling, in case of proper nouns. The Model must have a provision for detecting the alphabet individually. The task is not particularly challenging as the alphabets are not prolonged gestures and just one frame will suffice. A simple convolutional classifier would do the trick.

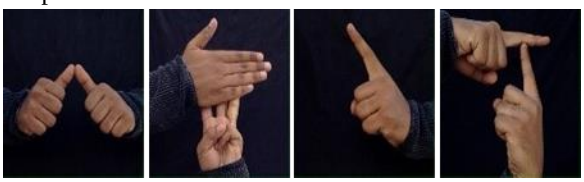

Figure1: The ISL Alphabet

2. Scraping through videos for Gestures: Due to absence of any dataset for ISL words and gesture detection the data will have to be scraped from the websites. Every word/phrase/gesture has videos available. A piece of code to scrape through the words and download these videos and then have them run through the OpenPose model with their coordinates save with their corresponding labels are required.

3. Scraping through videos for Gestures: Due to absence of any dataset for ISL words and gesture detection the data will have to be scraped from the websites. Every word/phrase/gesture has videos available. A piece of code to scrape through the words and download these videos and then have them run through the OpenPose model with their coordinates save with their corresponding labels is required. Augmentation for word Gestures: After obtaining the limited dataset from both sources we need to augment it. Dataset Augmentation is a common practice while training over a limited dataset to make it more generalized and robust. This is achieved by adding random noise to the data so that the model is made subject to variance in the training data. Adding

Gaussian noise to the dataset increases the variance of the dataset giving more examples for training.

4. Dataset for Language Model RNN: GPT-2 (Generative Pre-trained Transformer) is a large transformer-based Language model with 1.5 billion parameters, trained on a dataset of 8 million web pages [6]. GPT-2 is trained with a simple objective to predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains.

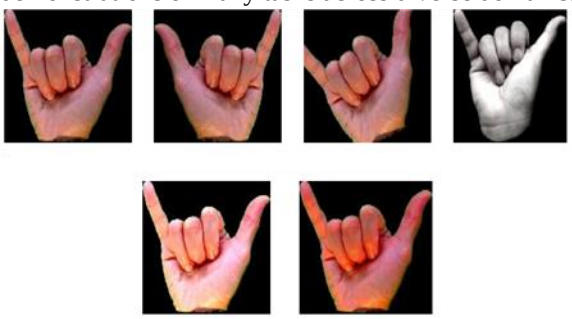

Figure2: Image Augmentation with inversion, tilt, dialing saturation, colors and contrast

GPT-2 displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of unprecedented quality. On language tasks like question answering, reading comprehension, summarization and translation. GPT-2 begins to learn these tasks from the raw text, using no task-specific training data.

B. The Models:

The following section covers the model required for detecting the key points of the gestures and translating the information into natural spoken language.

1. OpenPose Model: This model is required for detecting the key points of the hands, the output of which would be the coordinates that would be fed to classification models followed by language models.

2. Model Architecture: The model that we have chosen was proposed by [7][8].

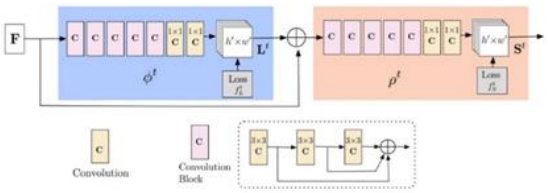

Figure3: OpenPose model architecture

It's a multimodal approach, two multi staged Convolutional Neural Networks that are responsible for detecting the points and the lines joining them. The first stage is a classic CNN with increased depth that refines the prediction over time. It includes several convolutional kernels of size 7x7 with an architecture resembling the DenseNET. The second stage also has a similar architecture. The outputs from both are ultimately concatenated and passed on as a final output.3. Detecting the points: The model uses two techniques for detecting the significant points for constructing the skeleton of the individual: confidence maps and part affinity fields (PAFs) [10]. The model referenced in the paper delivers a output that is a conglomeration of the two. The primary half predicts a heatmap over the picture where it thinks the keypoint lie. The regions on the picture where there is a high probability of the event of these focuses are communicated with splendid color. Part affinity fields are responsible for creating a vector field over the section of the image where the model detects the existence of a link joining the keypoints points. These too are expressed as heatmaps with an elongated shape. These vectors likewise propose the directions, to forestall disarray within the sight of numerous subjects. The test happens when there are numerous individuals in the casing since it would bring about a few arrangements of similar keypoints and sorting out the associations isn't so natural. A unit vector along the direction of the anticipated lengthened shapes help in distinguishing right arrangements of focuses that have a place with a specific person.

4. Model Extension for Hands: OpenPose is a popular human body pose detector which is able to detect up to 125 body key points between bodies, faces, foot, and hands. It uses a 21 keypoint model for hands, four points for each finger plus one for the wrist. OpenPose started as the code repository for the published paper "iPOSE–Real Time Movement Tracking Application" [10], and has since grown into a very useful platform performing people tracking, Unity integration, etc. OpenPose uses different networks to detect the body and the hand keypoints. On each image, OpenPose will first run the body keypoint location network, which using a bottom-up approach will first detect body keypoints and then wire them together on a reasonable way to construct the different human bodies. Thanks to this approach, the running time is constant, it doesn't depend on how many people are seen on the image, because OpenPose's Neural Network will detect them all at once. But for the hands

the mechanism is different, instead of feeding the image just once into a Neural Network and detecting all the seen hands, parts of the image containing hands candidates are extracted and sequentially, one at a time, fed into the hand keypoint detector. Therefore, if ten persons are seen on the image, for each one two hand candidate areas of the image are going to be extracted and the hand detector will have to look on each of these twenty areas for keypoints. If you are running OpenPose with the hand detector turned on, the more people on the scene the slower it will run. These hand candidate areas are just the parts of the image likely to contain a hand, and OpenPose proposes these candidates based on the position of seen wrist, elbows and shoulders. Basically, if it sees these body parts are close together, it assumes where the hand should be. You can check it on the source code on the getHandFromPoseIndexes method. The heuristic has sense, as the people's hand's location is strongly related to the position their wrists, elbows and shoulders, but it means OpenPose won't be able to detect a hand not joined to any of these body keypoints. While floating hands are not that common, it may happen on scenarios with occlusions or if we spot the camera just directly to a hand on a close-up.

The key points taken into consideration entirely come from the upper body, i.e. arms, hands shoulders and the face, as all these features are highly relevant to the problem. The coordinates of the detected points instead of the image file to save space and later to be sent for further processing.

Figure4: OpenPose model architecture

C. RNN models for image data to words:

1. Convolutional Classifier: The augmented alphabet dataset contains images of the English alphabet in the ISL set of gestures. A simple network for multiclass classification would suffice, taking in the coordinate data of the key points from the OpenPose model as the input and performing classification among 26 classes.

2. Recurrent Encoder Decoder: The words and phrases part of the dataset requires a Recurrent model as the gestures last for a long time, the videos of the gestures are sampled at 15 frames per second and then fed to the model. The model is a classic encoder decoder pair, with the job of the encoder being getting the context out of the frames fed to it and the decoders job being classifying the context rich vector into one the words of the dictionary.

3. The Encoder: The Encoder part of this model is an LSTM that takes in the coordinate data from the OpenPose model and generates a context rich vector. The coordinate data is put in the form of a vector from the images sampled at 15 frames per second. The keypoints are concatenated and fed to the LSTM.

4. The Decoder: The context rich vector from the Encoder is then fed to the decoder which is a multiclass classifier which would classify the input into one the words of the dictionary. The encoder decoder model is trained on the augmented gesture and words dataset where the classes are vocabulary of the ISL.

D. GPT-2 for words to sentences:

For this situation, we will utilize the little form of GPT-2 with 12 layers of decoders. The model was prepared on 8 million site pages and is as of now very incredible in language errands. To hold its overall force in language demonstrating while at the same time adjusting to our informational index, we will in part retrain the model by freezing half of the layers by setting. This will likewise accelerate the preparation since the quantity of in reverse passes are decreased.

1. GPT-2 Architecture: The accompanying covers the engineering of the GPT-2 language model and what permits it to create normal sentences. The GPT-2 design was a variety of the well-known Transformer architecture. At its center, the Transformer design gives a conventional instrument dependent on encoder-decoders to identify conditions among information sources and yields. In the Transformer model, the encoder maps an information arrangement of image portrayals x(x1-xn) to a grouping of persistent portrayals z (z1-zn). Given z, the decoder then, at that point produces a yield grouping (y1-ym) of images each component in turn. At each progression the model is auto-backward, burning-through the recently produced images as extra information while creating the following [11].

While the Transformer engineering distinguishes long haul conditions between printed information, it does

nothing as far as learning explicit undertakings. The GPT-2 design expands the center Transformer model by infusing enhancements for explicit NLU errands. Also, GPT-2 enhances information move between the various layers getting more-strong across the whole range of NLU undertakings.

The data it trains on will be basic proclamations with the data sources being only the watchwords of the assertion and the yield would develop the total assertion.

This sort of pre preparing is exceptionally pertinent to our case as the signals and words that make sentences in the ISL are for the most part catchphrases and it's up to the subjects for translation. The GPT adjusted model will give us totally framed sentences from the words and motions got by our Encoder Decoder Model.

IV. CONCLUSION

This paper has presented an approach for adapting existing pose-detection models to assist people in teaching sign language. The system detects significant points of the hands, that gives the output coordinates to be fed to classification models followed by language models. This would help in understanding and teaching sign language and would enable hearing and speech impaired people to communicate.

REFERENCE

[1] Sharma A., Sharma N., Saxena, Y. et al.- Benchmarking deep neural network approaches for Indian Sign Language recognition. https://doi.org/10.1007/s00521-020-05448-8

[2] Reiter, E., & Dale, R.- Building applied natural language generation systems. Natural Language Engineering. doi:10.1017/S1351324997001502

[3] Yunji Kim, Seonghyeon Nam, In Cho and Seon Joo Kim- Unsupervised keypoint learning for guiding class-conditional video prediction. Yonsei University.

[4] Daniel DeTone, Tomasz Malisiewicz and Andrew Rabinovich- SuperPoint: self-supervised interest point detection and description. Cornell University.

[5] Shangze Wu, Christian Rupprecht and Andrea Vedaldi- Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. Visual Geometry Group, University of Oxford.

[6] Nagesh Singh Chauhan, Feb 2021, Hugging Face Transformers Package – What Is It and How to Use It, KDnuggets. https://www.kdnuggets.com /2021/02/hugging-face-transformer-basics.html

[7] Tomas Simon, Hanbyul Joo, Iain Matthews and Yaser Sheikh- Hand keypoint detection in single images using multiview bootstrapping.

[8] Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh- Convolutional pose machines.

[9] Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh- Realtime multi-person 2d pose estimation using part affinity fields. The Robotics Institute, Carnegie Mellon University.

[10] Tejal Deshpande, Amit Marathe, Ronak Singh Rajput, Shubham Maurya and Ninad Tawade- iPOSE – Real Time Movement Tracking Application. IRJET, Volume: 08 Issue: 05, May 2021.https://www.irjet.net/archives/V8/i5/IRJET-V8I599.pdf

[11] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin- Attention Is All You Need. https://arxiv.org/abs/1706.03762