Call For Help American Sign Language Recognition Using Deep Learning Models

1st Kushagra Srivastava
School of Computing
Dublin City University
Dublin, Ireland
kushagra.srivastava3@mail.dcu.ie
Date of completion : 31/07/2023

2nd Cheong Hao Io
School of Computing
Dublin City University
Dublin, Ireland
hao.cheong3@mail.dcu.ie
Date of completion : 31/07/2023

***Abstract*—Currently the Social Media platform does not offer sign language detection model that fits online deaf-mute community,in this paper we deliver a groundbreaking Sign language recognition model for call for help ASL Signs. Despite the fast growing user volumes on social media platforms, the existing safety detection is based on the written or spoken words and not focused on body or hand gestures. In our study we focus on call for help signs detection which also can be used to detect call for help related body gestures from video and ensure users are safe.**

We have focused on 9 words from the American Sign Language.Our dataset comprised of 431 videos out of which 6985 were extracted. We have used various machine learning models for classification and found that the model where the Feature Extraction was done using InceptionV3 architechture and classification using Dilated CNN gave the best performance with 95% accuracy on validation data and 98% accuracy on test data. When using ensemble of predictions from LSTM and DCNN models , we were able to achieve accuracy of 98.7% on sample data.

***Index Terms*—DCNN, CNN, InceptionV3, LSTM, GRU, Gesture Recognition, Social Media, Platform and Online Safety**

I. Introduction

Communication technology is a fast-emerging topic in the last decades, from the embark of Social Media platforms from mid-2000 until now, it is now becoming part of our daily life, this technology helps connect billions of users to their loved ones and not limited to geographical region and time restriction. In April 2023, there is 4.8 billion active social media platform users, and it is equally to nearly 60% of the global population. [1]

As the popularity grows, the concerns for user safety, truth worthiness, community support quality are growing simultaneously. Each social media platform holds a set of community guidelines under topics of cyber-bullying, nudity, hate speech, regulated good and intellectual property right etc to ensure only safe content will expose to platform users, since there is excessive user created data generated in every second, Computer Vision- deep learning algorithm is now a dominate method to mass detect and remove violated content, as opposed to traditional content moderation method which carry impractical issues such as: Expensive, interpretation bias, low scalability and efficiency. However, existed detection method does not cover all community user condition.

Sign Language is like any language it has own dialect and region difference. In 2023, there is approximate three hundred different Sign Languages around the globe. American Sign Language is the largest English based division globally and many countries which outside United States are using this sign language dialect, which is why ASL is selected here [2]

Based on World Health Organisation statistic data in February 2023 [3], there is 70 million deaf-mute population globally. According to research, American Sign Language call for help purpose specific recognition project is still not yet developed, hence this project is highly beneficial to the deaf-mute and enhance the overall online safety standard.

Project objective to help to bridge this community support leakage, once the model detection suspicious wording will then automatically get generate a ticket and get review by Emergency response team, as to develop a safe space for deaf-mute Social Media platform users to extract with.

All videos are sourced from two open sourced ASL Sign Language datasets WSASL and MSASL.In this paper we have also added new video demonstration of sign lanaguage and collected data from 29 volunteers.The videos were captured by smart phones, data collection consent form is provided prior filming process, all signers are friends or family members of project owners.

This is a multi-class video gesture recognition project, and it is followed by four steps: Data Collection, prepossessing, Feature Extraction and Classification. We have distinguished 9 words which related to call for help in sign languages, and which are 'I', 'Help', 'Hurt', 'Me','My','Sad','Save','Stuck' and 'Trapped.

Various models have been used for classification. It was found DCNN model where the feature extraction was done using InceptionV3 architecture gave the best performance.Other models that were used for classifying sign language were LSTM, GRU, InceptionV3 and CNN.

II. Related Work

Sign Language AI inspired applications are still relatively new in the domain but here are several breakthrough developments since 2020.

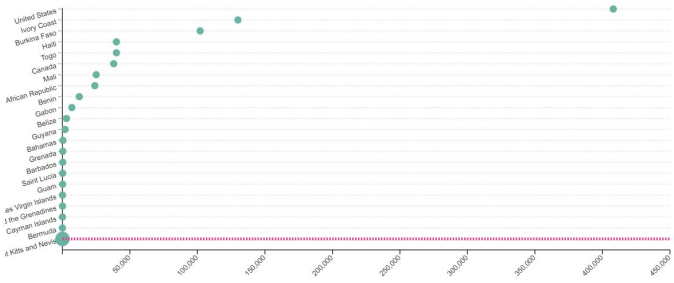

Fig. 1. 2023 ASL global usage

HandTalk is Sign Language detection AI inspired application that uses camera to pick up visual images, model offer gesture recognition and translate information into written text format, according to World Health Organisation data there are 80% of deaf population are semi-literate or illiterate, this development helps to bridge communication barriers significantly and ability to operates in numerous working environment such as Google Translate and Mobile phone. [4]

A. Recognition Basics Recognition process [5] can simply split into four steps which included Data acquisition, hand segmentation, feature extraction is commonly to apply HSC brightness matching and Gradient Hough transform (GHT) SIFT techniques, lastly body gesture recognition [6] can be captured through :

Convolution Neural Network: CNN models can be used to extract both spatial and temporal features from a video stream, follow to transfer the encoded image data into neural network for classification usage.

- Colour Recognition: Model apply segmentation technique to capture divergences between hand skin colour and the backgrounds.
- Apply Skeletal Recognition: This technique focuses on the alignments of the joints and hands, ability to monitor skeletal angle points in real time. This method uses hardware Kinet to track the body movements.
- Dynamic and Static Gesture Recognition: Dynamic gesture depends on the movement of hands and facial expression on classification, whereas in static gesture is a configuration of the hand is used to convey the message. [7]

B. Pre-processing of Data

The primary technique is coming from gaming theory method, established model first is to calculate and sum up the screen pixels which fall in the threshold values of HSV and YCbCr which reflects on exposed skin colour. This method can detect and extract five images which contain skin colour with the highest pixel counts. [8]

Secondly model has the identify and extract hand gesture frames which contain important information, it leverages combination of image entropy and density to cluster potential informative frames, this development has significantly improved the operation time and efficiency. [9]

C. Dataset and Evaluation

Sign Language Recognition is a new and emerging development area in AI. In fact, there is limited open data resources available in this domain. All reviewed papers had faced dataset limitation issue simultaneously and the finalised video dataset number is still relatively small compared to other cross domain projects.

The largest dataset within American Sign language space are "Word-Level American Sign Language (WSASL)" and " MS-American Sign Language (MSASL), where will be used in this project, which each covers over 2000 word counts from each, video contents were predominately captured for educational content, each sign language division only contain number of video samples only. As deep learning technique relies on big data, hence it is essential to establish additional self-created datasets, to have greater diversity and variety quality, allowing models to develop a robust models with high classification accuracy rate.

Although Sign Language recognition is a challenging task but there are several models have achieved considerable accuracy level:

- CNN: This deep learning method works specifically well with small dataset where content less feature counts.[10]
- Ensemble learning technique: Feature is primary extracted by InceoptionResNetV2, result will then pass through the combination of LSTM and GRU layers to achieved 97 precent of accuracy rate.[11]
- SVM: Support Vector Machine(SVM) is a supervised machine learning technique, this learning methodology is commonly used for binary or Multi-class image pattern detection, with the accuracy rate to acheive to 90 precent in image classification task. [12]
- PCA and K-NN: Statistical models can also helps to work on multi-class video classification issues, result showed Participant Component Analysis(PCA) got 95% and K-Nearest Neighbours(K-NN) got 98 % for accuracy rate. [13]

E. Comparison to this project

According to existed field research result, 'Call for help' purposed sign language detection model is still not yet developed in the current social media platform, where highlights the gaps and limitation from the existed community violation detected methods. Hence this project is highly beneficial to the deaf-mute and online safety community.

III. Methodology

In this section we discuss the methodology used for the Sign Language Detection. Here we have described the process used for Data Collection , prepossessing , Feature Extraction and Classification.

A. Data Collection

.We have identified 9 words that can be used as call for help in sign language. These are 'I', 'Help', 'Hurt', 'Me', 'My','Sad','Save','Stuck' and 'Trapped'. Figure1 one shows that ASL signs for words.The dataset that is used in this project

is gathered from various sources.There are three sources of the data :

- Word-Level American Sign Language (ASL)[14] dataset.This dataset contains sign language covering 2000 common words.
- MS-American Sign Language[15] dataset. This dataset has been released by Microsoft which has 25000 videos and is created by over 200 people.
- The third source of data is the data we have collected by our selves. There are 29 people who have recorded 100 videos demonstrating the sign languages in the scope of this project

Table I shows the distribution of the data that as been collected for the call for help words across the three datasets.The data was collected in the form of videos. Smart phone cameras were used to capture videos.

TABLE I
DATASET BREAKDOWN

	help	hurt	I	me	my	sad	save	stuck	trapped	Total
WSASL	0	16	5	7	0	11	11	5	0	55
MSASL	0	22	8	16	0	21	5	7	0	79
Own Data	36	33	27	31	36	30	37	20	47	297
Total	36	71	40	54	36	62	53	32	47	431

B. Preprocessing

The vidoes which were demonstrated the sign language that are part of the scope of this project were extracted and downloaded from the public datasets WSASL and MS-ASL.There were various preprocessing task that was involved in getting data ready for consumption for the machine learning model.

- The first part was extracting frames from the dataset. In all of the collected videos the signers used sign language with bare hands ie. they did not have gloves or any other form of clothing covering their hands.When extracting frames our goal was to ensure that we extract frames where we are capturing maximum hand movements.To achieve this we first do the skin tone detection for each frame by first calculating the pixels which fall in the threshold values of HSV and YCbCr colour spaces for skin colour. We first find out the pixels which falls in the range of the human skin tone for HSV and YCbCr colour space. The results of both the HSV and YCbCr are then combined to get the area of the image which reflects skin tone [3]. Figure3 shows the skin tone detected for various colour spaces.
  Using this method we can compare the skin tone exposure across frames. When we have detected the skin tone exposure for all the frames a particular video we extract five frames having maximum skin exposure. When extracting frames there were conditions set so that frames with blurry images are not extracted.There were some videos where the extracted frames did not capture any hand movements at all.The frames for these videos were extracted manually.
- Each of the extracted frames went through a series of augmentation. Each extracted frame was horizontally and vertically flipped, rotated by a random angle and blurred.
- After extraction and augmentation it was found that there were imbalances in the dataset where the classes from 'hurt', 'sad' and 'me' had about 1000 images where as some classes had lesser number of frames extracted. Such imbalances may cause bias towards a class. To address this issue frames were dropped from classes with high number of frames and more frames were extracted for classes with lower number of frames. At the end of this exercise each of the class had approx 780 frames each.

C. Feature Extraction

Once the frames have been extracted from the video , the next step is to extract the features from the images using the InceptionV3 architechture.The extracted frames are resized and passed to the InceptionV3 model which has been weights pre-trained on ImageNet dataset.The InceptionV3 model extracts the features from the images in the form of a feature vector. The extracted features are then classification models to predict the correct sign language. The feature extraction using InceptionV3 was used by LSTM , GRU and DCNN models.

D. Classification

In this paper we have used several deep learning models for classification. We have experimented with GRU(Gated Recurrent Network) and LSTM(Long Short-Term Memory) models which are type of Recurrent Neural Networks to predict the sign language. Since the frames extracted from videos are in a sequence and have temporal relation with one another RNN models can be used to learn the temporal dependencies and and patterns in the sequential data.Since the RNN models processes the data sequentially and have more parameters to train they generally take a longer time to train.

We have also used two types of CNN models for prediction one with a dilation rate and the other one without a dilation rate.Dilated convolutions allow for an expanded receptive field without increasing the number of parameters, making them efficient for capturing global dependencies in a sequence [16].The dilated CNN is also resolves the problem of vanishing gradient which is faced when using CNN networks with many layers.Using Dilation in CNN model also makes the model more efficient as we can get more information because of the increased receptive area without increasing the number of parameters [17].DCNN when compared to RNN models are not recursive and have fewer parameters and hence take lesser time to train and not performance intensive.

Our DCNN network consists two convolution 2D Layers. The input to the model is a 8X8X2048 tensor, which represents a sequence of images. The first Conv2D layer takes the input tensor and applies a convolution operation with a kernel size of 3x3 and a dilation rate of 2. The dilation rate of 2 means that the kernel will skip over one element in the input tensor for every two elements that it covers. This allows the kernel to capture long-range temporal dependencies between the

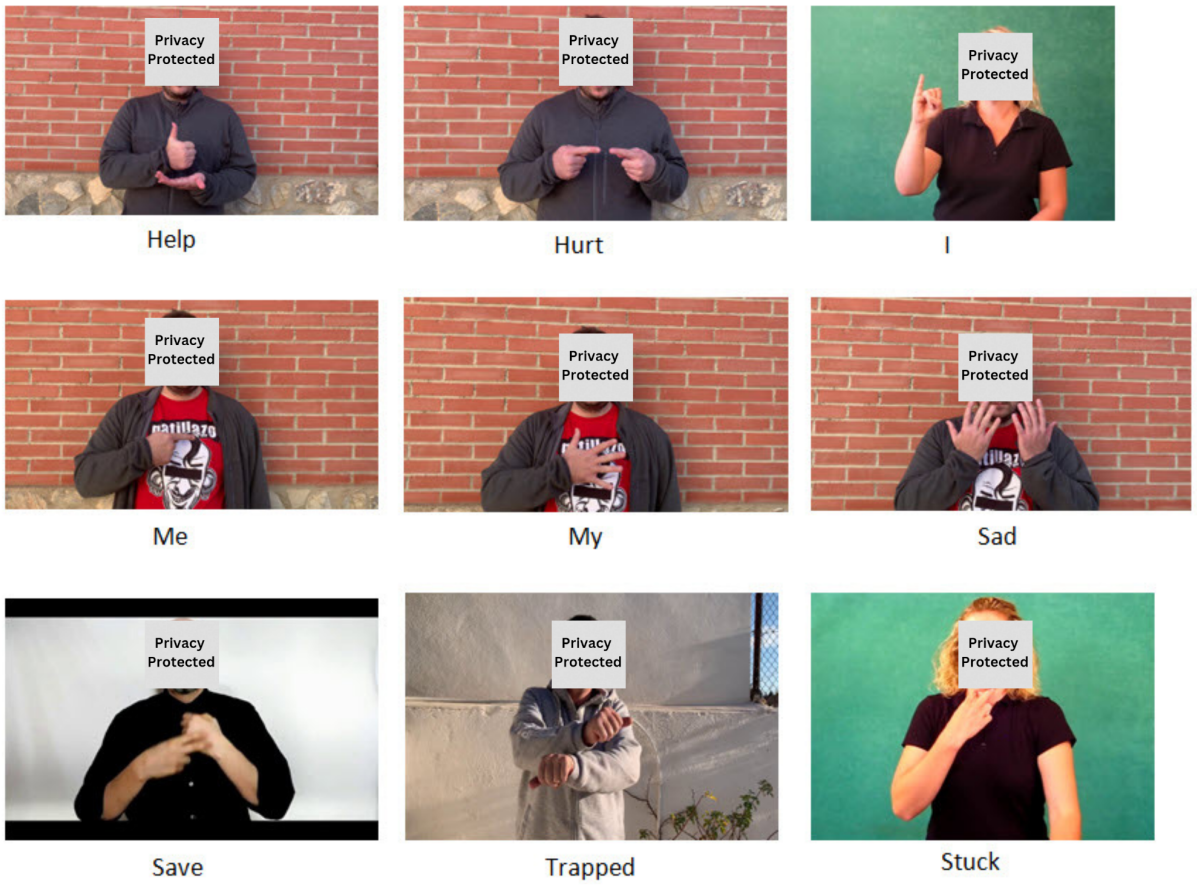

Fig. 2. ASL signs .

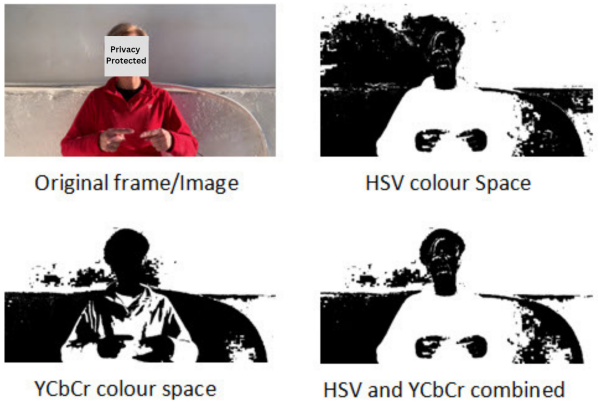

Fig. 3. Skin tone detection

images.The second Conv2D layer also applies a convolution operation with a kernel size of 3x3 and a dilation rate of 2. The GlobalMaxPooling2D layer then reduces the dimensionality of the output from the Conv2D layers. The Dense layers then classify the output from the GlobalMaxPooling2D layer.

The feature extraction for the models LSTM, GRU and DCNN was done using InceptionV3 Architecture. We have also used InceptionV3 for classification so that we can compare it with other models.

We have also experimented with models by adding drop out layers.

IV. Experimentation and Results

The tables 2 and 3 show the results of various models used for classification of sign language. We can see that the best result was achieved using DCNN model. We were able to achieve 95% accuracy in validation data and 98% accuracy in test data. The models have also been tested by adding droput layers. We saw a slight increase in accuracy for test data when using LSTM. The other models performed better without dropout layers.Dropout layer is a regulaeization technique commonly used in neural networks to prevent overfitting [18].

When the predictions of LSTM and DCNN model was combined to form and ensemble, it gave the highest accuracy for sample test data. The ensemble model gave an accuracy

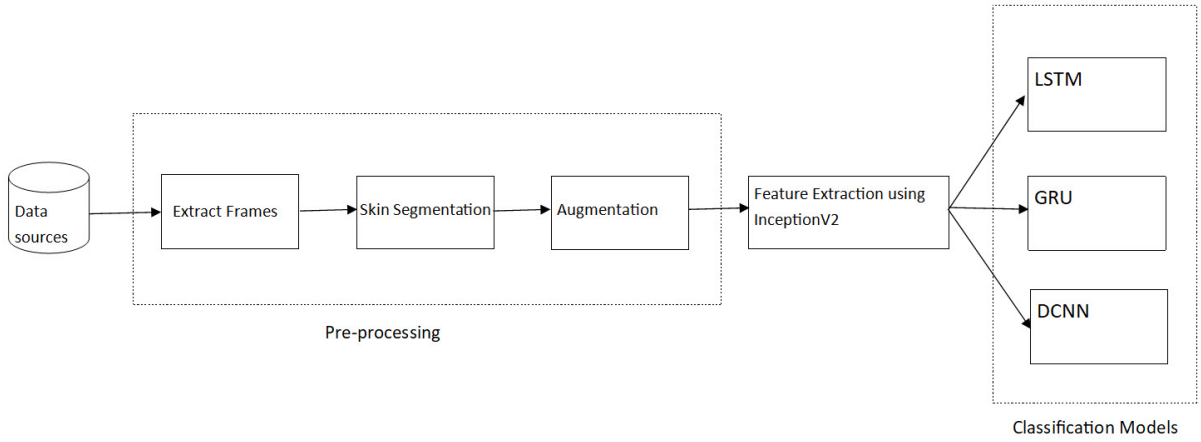

Fig. 4. Flow Diagram

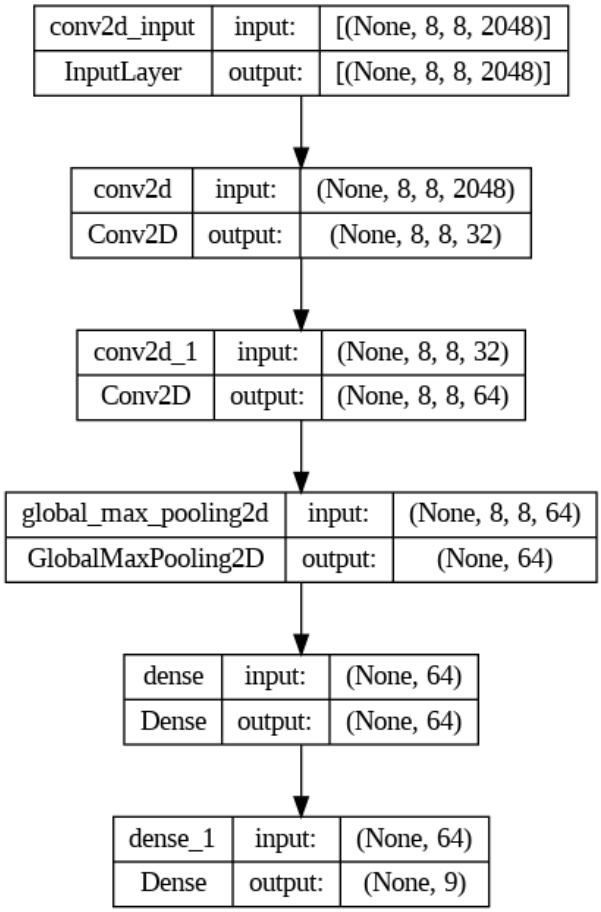

Fig. 5. DCNN model plot

of 98.7% on sample test data. To make the ensemble first we make predictions using the LSTM and DCNN models on the sample test data - LSTM predictions and DCNN predictions, respectively.

Then, we combine the predictions of the two models using averaging to create the ensemble predictions. This is done by adding the two sets of predictions and dividing by 2 to get the average.

TABLE II
RESULTS WITHOUT DROPOUT LAYERS

Model	Epochs	Number of parameters	Train	Val	Test
LSTM	63	134,483,209	93	88	88
GRU	91	100,863,753	91	90	87
CNN	20	5,608,489	99	89	89
DCNN	20	613,097	99	95	98
InceptionV3	20	23,910,185	93	88	90

TABLE III
RESULTS WITH DROPOUT LAYERS

Model	Epochs	Number of parameters	Train	Val	Test
LSTM	124	134,483,209	66	88	91
GRU	100	100,863,753	57	73	71
CNN	100	5,608,489	99	91	90
DCNN	100	613,097	98	89	87

V. CHALLENGES ENCOUNTED

- Limited American Sign Language open dataset result, although both WSASL and MASASL both holding over 2000-word count but each word only content numbers of video sample. As a result, we decided to create our own video dataset and collected sign languages from over 29 volunteered signers, the process was time consuming and long.
- The finalised dataset is small due to limited dataset sources and that leads to decease data image variety and limited the augmentation output quality.
- The overall frame extraction process did not perform practically well when the content is captured throughout noisy background or clothing is overly similar to the signer skin tone. Although colour threshold is an excellent approach here by isolating the clothing colour from the background, but this technique is not suitable in video data.

VI. Conclusion

In this paper we have compared five different deep learning recognition model to identify 9 distinct American Sign Language in call for help specific purpose and These are LSTM, GRU, CNN, DCNN and InceptionV3. We found that the DCNN model can the highest level of accuracy which was 95% on validation data and 98% on test data and when ensemble was created using DCNN and LSTM we were able to achieve an accuracy of 98.7% on sample test data.

Sign Language Recognition in computer vision is still a relatively new and yet on-going developing area, more ethical regulations and methodologies are expected to emerge for the following 5 years. Hence this project aims to help to bridge the current social media community support gap for Sign Language based recognition model and allows better inclusivity for all users.

References

[1] A.Petrosyan, "Number of internet and social media users worldwide of April 2023," Statista , 22 May 2023. [Online]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide/. [Accessed 12 July 2023].

[2] "Countries that use ASL 2023," World Population Review , [Online]. Available: https://worldpopulationreview.com/country-rankings/countries-that-use-asl. [Accessed 2023 07 02].

[3] H. B. ,. O. A. ,. D. A. ,. A. B. ,. A. A. a. H. A. Amal Babour, "Intelligent gloves: An IT intervention for deaf-mute people," De Degruyter , vol. 32 , no. 1, 2023.

[4] Handtalk, "About us," Handtalk, [Online]. Available: https://www.handtalk.me/en/about/. [Accessed 06 07 2023].

[5] Sign Language Recognition Techniques- A Review M. Safeel, T. Sukumar, S. K. S, A. M. D, S. R and P. S. B, "Sign Language Recognition Techniques- A Review," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-9, doi: 10.1109/INOCON50539.2020.9298376.

[6] Sign Language Recognition Based on Gesture Recognition/Holistic Features Recognition: A Review of Techniques O. Asthana, P. Bhakuni, P. Srivastava, S. Singh and K. Jindal, "Sign Language Recognition Based on Gesture Recognition/Holistic Features Recognition: A Review of Techniques," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 713-718, doi: 10.1109/ICIPTM54933.2022.9754140.

[7] M. Qutaishat, H. Moussa, B. Takruri, H. Abed Al-Malik, "American sign language (ASL) recognition based on Hough transform and neural networks," Expert Systems with Applications, vol. 32, no. 1, pp. 24-37, 2007. ISSN 0957-4174. doi: 10.1016/j.eswa.2005.11.018.

[8] D. Dahmani, M. Cheref, S. Larabi, "Zero-sum game theory model for segmenting skin regions," Image and Vision Computing, vol. 99, pp. 103925, 2020. ISSN 0262-8856. doi: 10.1016/j.imavis.2020.103925.

[9] Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion Hao Tang, Hong Liu, Wei Xiao, Nicu Sebe,Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion,Neurocomputing,Volume 331,2019,Pages 424-433,ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.11.038

[10] V. S. Katocha, U. S. T. Shagun, "Indian Sign Language recognition system using SURF with SVM and CNN," ScienceDirect, Jul. 14, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000121. [Accessed Nov. 19, 2022].

[11] Deepsign: Sign Language Detection and Recognition Using Deep Learning Kothadiya, Deep Bhatt, Chintan Sapariya, Krenil Patel, Kevin Gil, Ana Corchado Rodríguez, Juan. (2022). Deepsign: Sign Language Detection and Recognition Using Deep Learning. Electronics. 11. 1780. 10.3390/electronics11111780.

[12] R. R. V. Adithya, "Hand gestures for emergency situations: A video dataset based on words from Indian sign language," 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920309100. [Accessed 2022 11 06]

[13] M. Oliveira, H. Chatbri, S. Little, Y. Ferstl, N. E. O'Connor and A. Sutherland, "Irish Sign Language Recognition Using Principal Component Analysis and Convolutional Neural Networks," 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, NSW, Australia, 2017, pp. 1-8, doi: 10.1109/DICTA.2017.8227451.

[14] D. a. R. C. a. Y. X. a. L. H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1459–1469.

[15] O. Koller, "MS-ASL American Sign Language Dataset," Microsoft, 15 June 2018. [Online]. Available: https://www.microsoft.com/en-us/research/project/ms-asl/. [Accessed 10 02 2023].

[16] X. Lei, H. Pan and X. Huang, "A Dilated CNN Model for Image Classification," in IEEE Access, vol. 7, pp. 124087-124095, 2019, doi: 10.1109/ACCESS.2019.2927169.

[17] B. Rekabdar and C. Mousas, "Dilated Convolutional Neural Network for Predicting Driver's Activity," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 2018, pp. 3245-3250, doi: 10.1109/ITSC.2018.8569818.

[18] Srivastava, Nitish Hinton, Geoffrey Krizhevsky, Alex Sutskever, Ilya Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15. 1929-1958.