Journal of Computer Science

Original Research Paper

Kenyan Sign Language Translation Using SSD MobileNet-v2 FPNlite Model

Henry Muchiri Muriithi and Geoffrey Kasembeli Wanjala

School of Computing and Engineering Sciences, Strathmore University, Nairobi, Kenya

Article history
Received: 17-08-2024
Revised: 12-10-2024
Accepted: 23-11-2024

Corresponding Author:
Henry Muchiri Muriithi
School of Computing and
Engineering Sciences,
Strathmore University, Nairobi,
Kenya
Email: hmuchiri@strathmore.edu

Abstract: Speech impairment is a disability that impacts an individual's ability to communicate effectively through speech and hearing. Those affected often rely on alternative forms of communication, such as sign language. While sign language has become increasingly widespread in recent years, a significant challenge persists for non-sign language users, particularly in Kenya, where effective communication with sign language users remains a barrier. As such the disability creates inequality with affected people not being able to have equal opportunities. To bridge this gap and assist in achieving UN sustainable development goal number 10, which strives to reduce inequality technology, has been adopted to bridge this gap. Recent advancements in deep learning and computer vision have led to significant progress in motion and gesture recognition leveraging these cutting-edge techniques. However, not much work has been done specifically in translating Kenyan sign language. Many of the solutions have focused on sign languages from other developed countries. The focus of this study, therefore, is to create a vision-based application, which offers Kenyan sign language translation to text thus aiding communication between signers and non-signers. The model, developed using SSD MobileNet V2 FPNlite, achieved an accuracy of 85% after 20,000 training steps over 40 epochs.

Keywords: Kenyan Sign Language Recognition, Kenyan Sign Language, Dataset Sign Language Recognition, SSD MobileNet-v2 FPNlite

Introduction

Communication forms the foundation for understanding and sharing ideas. When conveyed effectively, these ideas promote understanding, meaningful interactions, and a sense of normalcy in daily life. Holmer *et al*. (2017). The deaf community faces several communication challenges that hinder their interaction with the wider population. Hearing individuals often face challenges in understanding the language of deaf and mute people, as they communicate using sign language. Mastering sign language requires proper training, not only for deaf individuals but also for anyone who wishes to engage and interact with them meaningfully. Overcoming these communication barriers involves time, financial resources, and accommodations, among other factors, all of which are essential to achieving the goal of bridging the gap between the hearing and deaf communities (Kusters and Hou, 2020). Various techniques have been applied worldwide to support people living with hearing difficulties. Human Sign. Language interpreters and teachers have been employed to bridge this gap, but their efforts remain insufficient to serve the entire population and fully facilitate communication.

Technology has introduced modern sign language approaches, including sensor-based systems that use gloves worn on the hands, as well as Kinect-based techniques that track hand trajectories and shapes to interpret sign language gestures Bulugu (2021). Other techniques involve vision-based methods that use machine learning and computer vision algorithms to detect and recognize sign language notations. Hybrid methods combine both approaches to improve the accuracy and effectiveness of sign language recognition (Papastratis *et al*., 2021). Recent research has increasingly focused on the implementation of computer vision-based solutions, owing to the framework's adaptability and its ability to incorporate facial expressions, body movements, and lip reading (Sultan *et al*., 2022). These studies have made significant progress in replacing device-based methods with vision-based techniques. These techniques leverage Artificial Intelligence (AI) and deep learning for Sign Language Recognition (SLR), focusing on translating hand gestures and body positions into meaningful sign language interpretations (Sultan *et al*., 2022). Below are some of the key studies in this area.

Zahid *et al*. (2023) developed a ResNet convolutional neural network model for recognizing Urdu Sign Language from live video. The model accurately identified sign language symbols and displayed them in written form in real time. Özdemir *et al*. (2023) developed an American sign language recognition model from digital videos. The model applied a Detection Transformer (DETR) object recognition vision transformer technique in combination with the ResNet152 and Feature Pyramid Network (FPN) deep learning model. The dataset used contained 12 video fragments of nine classes with the labels: “Love”, ‘Good”, “You”, “Meet”, “Yes”, “No”, “Please”, “Name”, " andMy". The downside was that the approach was computationally more complex and had more parameters than the previous methods. Sreemathy *et al*. (2023) proposed an Indian sign language recognition model. The model was developed by combining YOLOv4 for the detection and classification of the hand signs and SVM with Mediapipe as the feature extractor. The model was trained and developed using a self-developed dataset containing 676 images from 80 classes from the day-to-day vocabulary. This work focused on both static and continuous hand gesture recognition.

Alyami *et al*. (2024) presented a framework for isolated Arabic Sign Language (ArSL) recognition using hand and face key points. The framework employed a media pipe pose estimator to extract the key points of sign gestures in the video stream. They believe that pose-based approaches for sign language recognition would provide lightweight and fast models that can be adopted in real-time applications. The study applied Long-Term Short Memory (LSTM), Temporal Convolution Networks (TCN), and Transformer-based models for sign language recognition. The proposed models were evaluated using two different sign languages, the Arabic and Argentinian sign languages.

Özdemir *et al*. (2023) proposed an isolated SLR framework based on Spatial-Temporal Graph Convolutional Networks (ST-GCNs) and Multi-Cue Long Short-Term Memory (MC-LSTMs) to exploit multi-cue (e.g., body, hands and face) information for sign language recognizing. The developed model was evaluated on two Turkish sign language benchmark datasets, BosphorusSign22k and AUTSL.

Gueuwou *et al*. (2023) developed an end-to-end machine translation of African sign languages. The dataset applied consisted of videos in six African sign languages: Ghanaian Sign Language, Nigerian Sign Language, Kenyan Sign Language, Zambian Sign Language, Zimbabwean Sign Language and South African Sign Language. The videos consisted of Bible verses extracted from the Jehovah’s Witnesses sign language website. The evaluation of the developed model using BLEU scores reported extremely low performance. They attribute this performance to the length of the training data which was four times longer than existing benchmark datasets.

The related works section highlights great strides and achievements made in vision-based sign language recognition models. Many of the models achieved competitive recognition abilities and were found to be based on various deep learning technologies. A review of the related work also revealed that a majority of the scholarly works focused primarily on the translation of sign languages from developed countries and very little on sign language from developing countries, especially in the African continent (Gueuwou *et al*., 2023). This is a point of concern because according to the World Health Organization (2021), 80% of people with hearing impairment live in middle and low-income countries. Different nations have their own sign languages such as American Sign Language, Indian Sign Language, British Sign Language, and Japanese Sign Language (Zahid *et al*., 2023; Sreemathy *et al*., 2023). Each language has its unique characteristics, such as the use of one or both hands and facial expressions to convey emotions. With over 300 sign languages worldwide, there is currently no universal sign language (Alyami *et al*., 2024).

This study aims to address the identified gap by performing sign language recognition specifically for Kenyan Sign Language (KSL) using the SSD MobileNet-v2 FPNlite deep learning model. The model employs the Single Shot Detector (SSD) architecture with MobileNet-v2 as the backbone and Feature Pyramid Network lite (FPNlite) as the feature extractor. This approach leverages the benefits of both SSD and MobileNet-v2 for object detection while ensuring low computational complexity (Kumar *et al*., 2023). This represents the first contribution of the paper.

In Kenya, it is estimated that around 150,000 people have hearing impairments. This is a significant number, underscoring the critical importance of addressing the issue at hand.

Additionally, existing Kenyan sign language datasets only contain either few KSL vocabulary and in the case of Gueuwou *et al*. (2023) vocabulary retrieved from bible verses that might not be applicable in day-to-day conversations. This study therefore extends existing datasets by providing additional KSL vocabulary including commonly applied conversational phrases. The application of this enhanced dataset makes developed recognition models more useful and relevant in supporting the hearing impaired in regular conversations. The data set is publicly available at Kaggle. This is the second contribution of this study.

Materials and Methods

The primary objective of this study is to develop a machine-learning model that can translate Kenyan Sign Language notations and symbols into English-readable text. This process involves capturing the notations of a deaf individual through web cameras, images, or video inputs and then using computer vision algorithms to convert them into English text, thereby facilitating communication. This represents the first contribution of this study.

Data Collection

The study employed both primary and secondary data sources. Secondary data was obtained from the Kenya Sign Language (KSL) dataset, available on Kaggle. However, this dataset was limited to a small set of KSL notations, which restricted its scope. To address this limitation, we collected primary data to broaden the dataset.

Primary data collection involved capturing images of various KSL notations using a camera. These notations were based on the Kenya National Association of the Deaf booklet to ensure accuracy. Each sign was performed and photographed multiple times under different lighting conditions and variations, creating a diverse dataset that would enhance the model's training.

For this proof-of-concept study, we focused on the following notations: Church, enough, friend, hello, I am married, I love you, love, me, mosque, no, seat, temple, thank you so much, and yes. These notations were chosen because they represent common phrases widely used within the Kenyan deaf community. They encompass basic expressions suitable for various contexts, including home, school, work, and social gatherings, making them highly relevant for daily interactions.

Table (1) summarizes the dataset used in this study, categorized by data source. Overall, the combined dataset contains 11,330 images across all classes, ensuring a robust foundation for training and evaluating the model's performance.

Figure (1) illustrates the integrated words and phrases sourced from both the primary dataset and the Kaggle site. By combining the diverse vocabulary from the two sources, we aimed to enhance the model's ability to recognize and interpret various Kenya Sign Language notations effectively. This integration ensures that the model is trained on a richer and more varied linguistic context, ultimately improving its performance in real-time sign language recognition tasks.

Data Annotation

In order to prepare the dataset for the machine learning task, the data was preprocessed by first labeling the images with the help of labeling open-source image package. Secondly, the images were augmented by flipping, rotation, and zooming techniques. Lastly, the image sizes were standardized into 320 by 320 pixels.

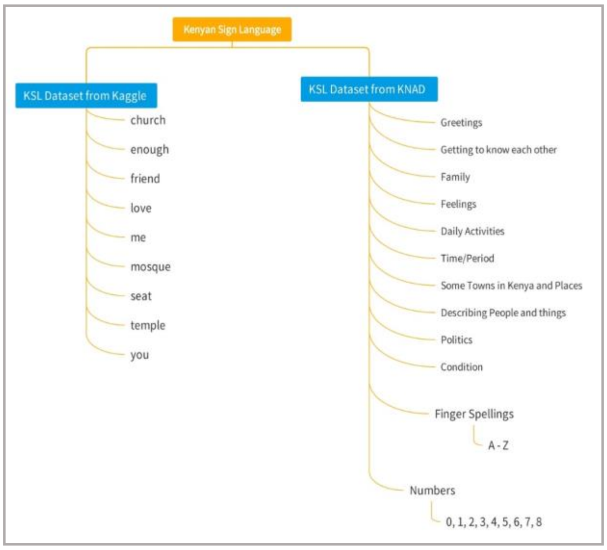

Fig. 1: Class compositions of the integrated Kenya Sign Language dataset

Table 1: Dataset composition summary

Data source category	Number of classes	Number of images
Primary DataSet	80	2400
Kaggle SITE	9	8930
Total		11,330

Model Development and Training

The machine learning model was implemented using tensor flow version 2. The SSD MobileNet V2 FPNLite 320×320 variant was selected for its optimal balance of shorter training time and high accuracy compared to other variants. SSD, similar to Faster-RCNN, is employed for object detection and classification in input images. It integrates the YOLO regression approach with a set of default bounding boxes to detect features at different scales. The model was trained and tested using the dataset provided in Table (1), with 80% of the dataset located for training and 20% for testing.

Choice of Model Architecture

The model architecture consists of three main components:

1. SSD Framework: The SSD framework is employed for its ability to perform object detection in a single pass, utilizing multiple feature maps at different scales to enhance detection accuracy
2. MobileNet-v2 Backbone: MobileNet-v2 serves as the backbone network, providing a lightweight and efficient feature extraction mechanism. Its

architecture includes linear bottlenecks and inverted residuals, which contribute to improved performance without excessive computational demands.

3. Feature Pyramid Network Lite (FPNlite): FPNlite is integrated to facilitate the extraction of features at multiple scales, allowing the model to effectively detect objects of varying sizes

Due to its multi-scale target feature extraction approach, SSD offers faster detection speeds than Faster-RCNN and higher detection accuracy than YOLO-v2 (Kumar *et al*., 2023). A supervised learning method was applied to train the model and experimental hyperparameter tuning was performed to optimize the model's performance and achieve the best results.

Model Deployment and Validation

The model was deployed on a web application using a front-end React framework and a Flask backend, ensuring public accessibility. The web-based user interface is designed to be user-friendly, allowing anyone to translate Kenyan Sign Language notations into clear English text for effective communication. The backend handles user login functionality and processes the model's logic engine.

Results

Model Training Results

To optimize the model's performance, several parameters were fine-tuned, including the training steps and the number of epochs. The model underwent multiple rounds of fine-tuning to achieve the desired detection accuracy for real-time textual inference of notations from webcam video input. The results of each hyperparameter tuning exercise are presented in Table (2). The results indicate a steady increase in the model's accuracy, rising from 58% at 500 steps to an optimal performance of 85% after 20,000 steps of training. Correspondingly, the model's accuracy improved from 75-85%.

Figure (2) presents the loss values recorded throughout the training phase of the model. As depicted, the loss value decreases progressively as the training steps increase, ultimately reaching an acceptable minimum at 20,000 steps.

Table 2: Hyper-parameter tuning results

Hyperparameter	Results
Training steps	
500	Maximum average precision of 0.58 (58%)
10,000	Maximum average precision of 0.6 (60%)
20,000	Maximum average precision of 0.85 (85%)
Number of epochs	
20	Accuracy of 75%
30	Accuracy of 75 %
40	Accuracy of 85%

To assess the performance of the developed model, the study utilized key evaluation metrics, specifically maximum Average Precision (mAP) and average recall. These metrics are essential for understanding how well the model can identify and classify objects within images and videos. As illustrated in Fig. (3), the model achieved an average precision score of 0.78, which translates to 78%. This score indicates a strong ability to correctly identify relevant instances among all predicted instances, reflecting a high level of accuracy in its classifications.

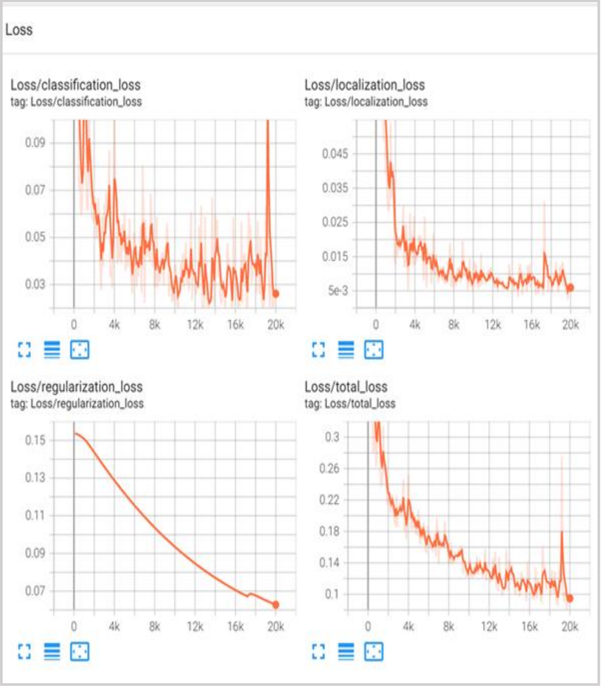

Fig. 2: Training and validation loss graph

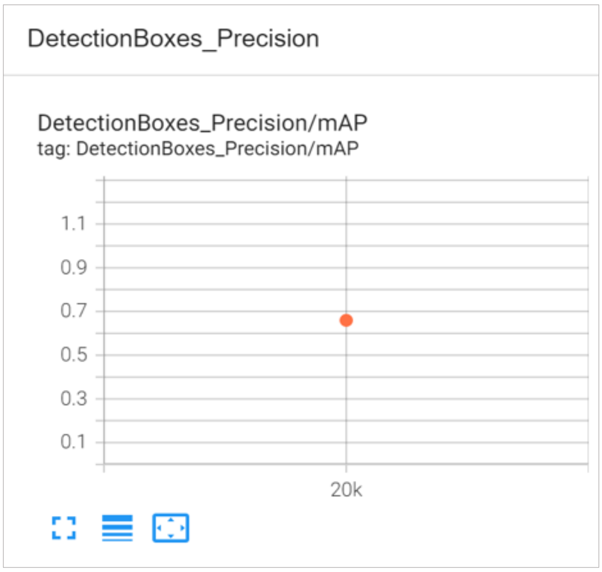

Fig. 3: Reported precision scores

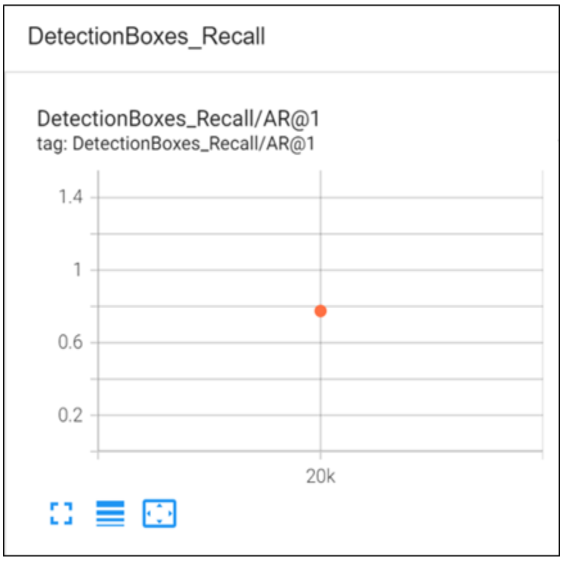

Fig. 4: Recall scores

Meanwhile, Fig. (4) presents a recall score of 0.85, equating to 85%. This metric measures the model's ability to correctly identify all relevant instances within the dataset, signifying that a substantial majority of the actual notations were successfully detected.

Discussion

The experiment on Kenyan Sign Language (KSL) translation using the SSD MobileNet-v2 FPNlite model yielded promising results, highlighting the impact of fine-tuning and optimization on model performance. The fine-tuning results demonstrate a progressive enhancement in detection accuracy. This improvement underscores the effectiveness of the training process in refining the model's capabilities.

This significant reduction in loss indicates that the model's predictions became increasingly accurate over time. A lower loss value signifies that the model made very few inaccurate guesses, demonstrating its improved ability to generalize and accurately classify the input data. Consequently, this finding suggests that the training process effectively enhanced the model's performance, leading to reliable outcomes in the context of real-time sign language recognition

The reported precision scores demonstrate the model's strong ability to accurately classify relevant instances from its predictions, emphasizing its reliability in distinguishing Kenyan Sign Language notations. The high recall scores further highlight the model's effectiveness in identifying a significant portion of the actual notations within the dataset, highlighting its robustness in real-time detection scenarios. These exceptional precision and recall scores not only surpass the typical thresholds for object detection models but also underscore the model's superior capability to detect and accurately translate sign language notations.

Existing sign language interpretation machine learning models have demonstrated varying performance metrics, often influenced by their underlying architectures and training methodologies. For instance, traditional models such as Hidden Markov Models (HMMs) have achieved moderate accuracy levels, typically around 60-70%, but struggle with capturing the complexities of continuous signing. More recent deep learning approaches, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have improved performance, achieving accuracy rates of approximately 75-80% (Khan *et al*., 2021).

In contrast, our model based on SSD MobileNet-v2 FPNlite has surpassed these benchmarks, achieving an average precision score of 78% and a recall score of 85%, demonstrating a robust ability to accurately detect and interpret sign language notations in real-time. This improvement underscores the effectiveness of the SSD MobileNet-v2 FPNlite model in balancing accuracy and computational efficiency, making it particularly suitable for applications in dynamic environments.

After successful training, the model was deployed as a web application. Figure (5) illustrates a sample operation of the model, showing how it highlights detected hand gestures within a bounding box. This highlighting corresponds to the gesture labels defined during the image data pre-processing stage. The model not only identifies the gesture but also displays its corresponding textual meaning above the box, along with a confidence level shown as a percentage. This feature allows users to easily understand the notation by reading the text on the screen. The web application is designed for online communication, making it especially useful for virtual meetings or collaborative activities. Users can seamlessly interact with individuals who use sign language, promoting clearer and more inclusive communication.

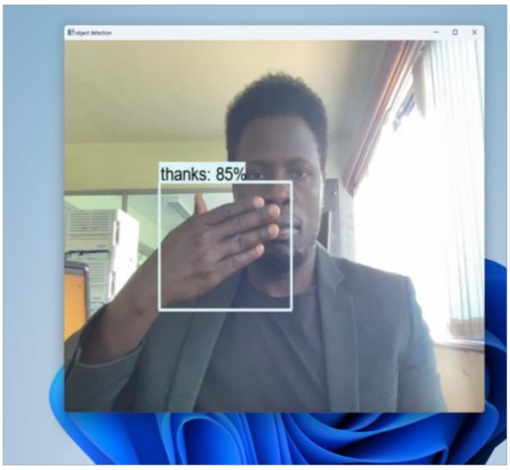

Fig. 5: Illustration of the model's operation

The model demonstrates impressive capabilities, including the ability to detect hand gestures from distances of up to 4 meters from the camera, ensuring effective communication even in larger spaces. Additionally, it maintains its performance in lower lighting conditions, showcasing its robustness and adaptability. Overall, these features enhance the practicality of the application in diverse environments and scenarios, further supporting its utility in real-time sign language recognition.

Conclusion

The primary objective of this research was to develop a Kenyan Sign Language translation model capable of converting sign language notations into readable English text. This facilitates communication between signers, non-signers, and deaf individuals, promoting better interaction and understanding. The notations covered in the study included finger spellings, numbers, and basic daily expressions such as greetings, family terms, emotions, daily activities, time periods, and interactive phrases. The study followed a sequential set of carefully planned milestones to achieve the research objectives.

The study aimed to review existing computer vision algorithms and models used for sign language recognition. It examined both hardware-based and software-based algorithms employed in sign language translation. The review highlighted the need for a model to translate Kenyan Sign Language and emphasized the importance of a software-based solution due to its affordability compared to hardware-based options like glove-based systems.

With this gap identified, the study set out to design a computer vision model for recognizing Kenyan Sign Language. The model was developed using transfer learning, following a prototyping-based methodology and a quantitative research design. The model's development parameters included an image size of 320×320 in JPEG format, 20,000 training steps, and an average learning rate of 0.054. The model achieved an optimal accuracy of 85% after 20,000 training steps. Additionally, it successfully detected hand gestures from a distance of up to 4 meters and under poor lighting conditions.

The model's strong performance underscores its potential for real-world application in translating Kenyan Sign Language. This would provide significant benefits to the hearing-impaired community in Kenya, who currently depend heavily on human translators to facilitate communication in various settings such as education, healthcare, and social interactions. By automating sign language recognition and translation, this technology could bridge communication gaps, offering greater independence and accessibility for the deaf and hard of hearing. Additionally, the model could be integrated into mobile apps or online platforms, enabling users to communicate more easily without needing a translator, especially in remote areas or during virtual meetings where translators may not always be available. This advancement represents a meaningful step toward enhancing inclusivity and reducing communication barriers in Kenya.

Acknowledgment

I would like to express my sincere gratitude to my advisor, Dr. Henry Muchiri, for his invaluable guidance and support throughout my research journey. His expertise and encouragement have been instrumental in shaping my work.

I am also deeply thankful to my fellow researchers and friends for their camaraderie and support. Their willingness to share ideas and resources has made this journey both enriching and enjoyable.

Special thanks to Strathmore University School of Computing and Engineering Sciences for providing the necessary resources and funding that made this research possible.

Funding Information

The authors have not received any financial support or funding to report.

Author’s Contributions

Henry Muchiri Muriithi: Analysis of existing literature, Manuscript writing and review.

Geoffrey Kasembeli Wanjala: Analysis of existing literature, data collection, Model development, validation of the study, analysis, interpretation of results

Ethics

This study upholds the principles of honesty and integrity and strictly avoids any instances of plagiarism, data fabrication, or falsification. Prior to commencing the research, the Strathmore University Ethics Review Board was consulted to ensure that ethical standards were maintained throughout the study's implementation. All previous authors' works have been properly cited to acknowledge their contributions. Additionally, a research permit has been obtained from the relevant authorities, along with a document affirming the originality of this study.

References

Alyami, S., Luqman, H., & Hammoudeh, M. (2024). Isolated Arabic Sign Language Recognition Using a Transformer-Based Model and Landmark Keypoints. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(1), 1–19. https://doi.org/10.1145/3584984

Bulugu, I. (2021). Sign language recognition using Kinect sensor based on color stream and skeleton points. *Tanzania Journal of Science*, *47*(2), 769–778. https://doi.org/10.4314/tjs.v47i2.32

Gueuwou, S., Takyi, K., Müller, M., Nyarko, M. S., Adade, R., & Gyening, R.-M. O. M. (2023). AfriSign: Machine Translation for African Sign Languages. *4th Workshop on African Natural Language Processing*.

Khan, A., Hussain, M., & Saeed, R. (2021). A Review on Sign Language Recognition Systems. *2nd International Conference on Applied Electromagnetics, Signal Processing, & Communication (AESPC)*, 1–5. https://doi.org/10.1109/AESPC52704.2021.9708526

Kumar, S., Kumar, R., & Saad. (2023). Real-Time Detection of Road-Based Objects Using SSD MobileNet-v2 FPNlite with a New Benchmark Dataset. *4th International Conference on Computing, Mathematics and Engineering Technologies (ICoMET)*, 1–5. https://doi.org/10.1109/icomet57998.2023.10099364

Kusters, A., & Hou, L. (2020). Linguistic Ethnography and Sign Language Studies. *Sign Language Studies*, *20*(4), 561–571. https://doi.org/10.1353/sls.2020.0018

Özdemir, O., Baytaş, İ. M., & Akarun, L. (2023). Multi-Cue Temporal Modeling for Skeleton-Based Sign Language Recognition. *Frontiers in Neuroscience*, *17*, 1148191. https://doi.org/10.3389/fnins.2023.1148191

Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2021). Artificial Intelligence Technologies for Sign Language. *Sensors*, *21*(17), 5843. https://doi.org/10.3390/s21175843

Sreemathy, R., Turuk, M., Chaudhary, S., Lavate, K., Ushire, A., & Khurana, S. (2023). Continuous Word Level Sign Language Recognition Using an Expert System Based on Machine Learning. *International Journal of Cognitive Computing in Engineering*, *4*, 170–178. https://doi.org/10.1016/j.ijcce.2023.04.002

Sultan, A., Makram, W., Kayed, M., & Ali, A. A. (2022). Sign Language Identification and Recognition: A Comparative Study. *Open Computer Science*, *12*(1), 191–210. https://doi.org/10.1515/comp-2022-0240

World Health Organization. (2021). *World Report on Hearing*. ISBN-10: 978-92-4-002048-1.

Zahid, H., Syed, S. A., Rashid, M., Hussain, S., Umer, A., Waheed, A., Nasim, S., Zareei, M., & Mansoor, N. (2023). A Computer Vision-Based System for Recognition and Classification of Urdu Sign Language Dataset for Differently Abled People Using Artificial Intelligence. *Mobile Information Systems*, *2023*, 1–17. https://doi.org/10.1155/2023/1060135