



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Sign Language Detection Through Action Recognition System Using MP Holistic

Annapoorna B R^{*1}, Abhishek S^{*2}, Basavaraj Madiwalar^{*3}, Pavankumar V T^{*4}, Kishan N S^{*5}

Assistant Professor, Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India¹

Student, Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India²³⁴⁵

ABSTRACT: For people who are hard of hearing, sign language is an essential communication tool. Developing an accurate and real-time sign language recognition system is essential to facilitate effective communication between individuals with hearing impairments and the wider community. In this project, we propose a Sign Language Recognition System using Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) known for its ability to capture sequential patterns. The system aims to accurately interpret and translate sign language gestures into corresponding text or speech, enabling seamless communication between individuals who use sign language and those who do not.

The approach involves capturing and pre-processing sign language video data to extract relevant features. The pre-processed data is then fed into an LSTM network, which learns the temporal dependencies and patterns inherent in sign language gestures. The trained model is capable of recognizing a diverse set of sign language gestures, including both static and dynamic signs. To enhance the system's performance, we explore techniques such as data augmentation and transfer learning.

KEYWORDS: Sign Language Recognition, Recurrent neural Network, Long Short-Term Memory

I. INTRODUCTION

Sign language is a rich and intricate form of non-verbal communication used primarily by individuals who are deaf or hard of hearing to convey thoughts, ideas, and emotions. Unlike spoken languages, sign languages rely on visual and gestural components, making use of handshapes, facial expressions, and body movements to communicate complex messages. The global deaf community is diverse, with different countries and regions often having their unique sign languages, further emphasizing the richness and complexity of this mode of communication. The origins of sign languages can be traced back to Deaf communities that developed their methods of communication over centuries. These languages are not merely simplified versions of spoken languages; they are complete and natural languages with their grammatical structures and vocabularies. For example, American Sign Language (ASL), British Sign Language (BSL), and French Sign Language (LSF) are distinct sign languages with their grammar and vocabulary, despite the presence of spoken English and French in these respective countries.

This project holds significant implications for the deaf and hard-of-hearing community, offering a technological solution to bridge communication gaps. The proposed system, once implemented, has the potential to enhance accessibility, foster inclusivity, and empower individuals with hearing impairments to engage more fully in various aspects of life.

As we delve into the realms of advanced technology and machine learning, the fusion of Sign Language Detection and LSTM-based Action Recognition represents a promising avenue for creating impactful assistive tools. This research not only contributes to the field of assistive technology but also underscores the broader significance of leveraging technology to create a more inclusive and connected world for everyone.

II. REVIEW OF LITERATURE

As a part of this survey review, almost 50 papers were downloaded to present a systematic technical analysis of Sign Action Detection from various digital libraries, which include IEEE Xplore and many more. After studying

the paper title, abstract, introduction, experiment, and future scope, the 16 most suitable papers for the review have been identified,

In this part of the article, analysis and planning work is done. The literature review discusses previous attempts made in the field of research and development of Sign Action Recognition and the usage of various technologies. The discussion on the literature survey includes significant contributions made by different scholars in this discipline.

This paper presents a real-time computer vision-based Bengali Sign Language (BdSL) recognition system that detects the probable hand from captured images. The system detects the hand in each frame using cascaded classifiers based on Haar-like features. The hand sign is extracted based on Hue and Saturation values corresponding to human skin color. After normalization, the hand sign is converted to a binary image and classified by comparing it with pre-trained binary images using the K-Nearest Neighbours (KNN) Classifier. The system can recognize 6 Bengali Vowels and 30 Bengali Consonants. The system is trained using 3600 training images, with each signer performing 10 signs for each corresponding Bengali alphabet. The system achieved recognition accuracy of 98.17% for Vowels and 94.75% for Consonants. The study focuses on the extraction and preprocessing of hand signs using skin color and binary features. The cropped image is scaled, RGB images converted to the HSV color system, and segmented using a human skin color-based approach. The image is then reduced to remove background, dilation, erosion, and smoothing techniques. A training module is generated using the extracted binary hand sign using KNN classifiers. [1].

The proposed system consists of several stages for processing images accurately. The first stage involves real-time image capture using a webcam, which is then converted to a compatible image for processing. The system then performs greyscale conversion and morphological operations, such as erosion. The camera is set up to make the hand visible and visible. The system captures video input objects using a web camera, with 5 frames per trigger. The system camera initially captures the video every second, and a snapshot is taken for further processing

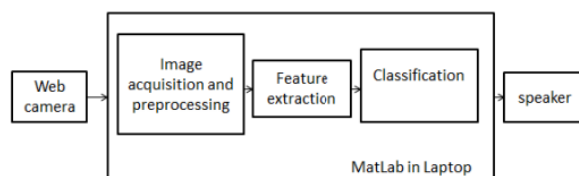


FIG: 1 FRAMEWORK OF THE HAND GESTURE RECOGNITION SYSTEM

Skin segmentation is performed to identify the hand gesture, converting the image to the YCbCr domain for better segmentation. The system then compares the detected binary image with the area of the detected binary result, subjecting it to morphological closing operation to obtain a well-defined segmented gesture image. Hand tracking is also performed, determining the centroid point of the user's motion and allowing for recognition. The feature extraction is done using a canny multistage detection algorithm with a wide range of edges in an attempt to detect the hand gesture. [2]

The paper discusses the differences between RNN and LSTM, focusing on their differences and how LSTM refines RNN and overcomes the vanishing gradient problem. RNNs rely on previous inputs, while LSTM blocks provide a solution by replacing RNN units with memory blocks. Each block has a memory of the previous network status and can flexibly update hidden states and forget previous hidden states. The LSTM architecture is introduced to solve the problem of SLR, where RNNs are trained for tasks with long-term delay between inputs and outputs. The LSTM blocks contain one self-connected memory cell and three multiplicative units: the input gate, the output gate, and the forget gate. The LSTM architecture is designed to be more efficient and effective in tasks with long-term delays between inputs and outputs. [3].

Research has focused on representing actions using body silhouettes without attempting to segment or match individual body parts to a model. This has been used in early work to train Hidden Markov Models (HMMs) of actions, which are applied in the recognition phase to find the model that best matches the observed symbol sequence.

Feature representations for action recognition use 2D and 3D features like stick figures, silhouettes, or volumes to represent information about the position and movement of different body parts. Most approaches use an explicit model of the human body and aim to optimize the match between model projections and an observed image frame while maintaining a correspondence of joints between frames. However, 2D-time representations are view-dependent, and 3D pose estimation requires more complex models. [4]

This work proposed a Hand Gesture Recognition System using Microsoft Kinect for Xbox, built on the Candescant NUI project and using the Open NI framework for data extraction from the Kinect sensor. The system includes three main processes: Hand Detection, Finger Identification, and Gesture Recognition. Hand detection involves separating hands from the background using depth information, obtaining two clusters of hand pixels, determining convex hulls of hands, and detecting hand contours. Finger identification involves common pixels being collected, and finger names are determined according to their relative distances [5].

This paper proposes a novel CNN-based selfie sign language recognition system to achieve higher recognition rates. The model is constructed with an input layer, four convolutional layers, five rectified linear units (ReLU), two stochastic pooling layers, one dense and one SoftMax output layer. The model uses stochastic pooling for feature representation and classification, with two layers of pooling initiated to avoid substantial information loss. The input video frames are pre-processed by resizing them to $128 \times 128 \times 3$ to increase the computational capability of the high-performance computing (HPC) on which the program is implemented. The CNN uses a tanh activation function with an additive bias, and the output of a convolutional layer is generally denoted with the standard equation. The maximum value of a feature is obtained using the pooling technique, which reduces data variance. The architecture is implemented with a stochastic pooling technique by calculating the maximum value of a feature over a region. The results are compared with other traditional state-of-the-art techniques such as Mahalanobis distance classifier, Adaboost, ANN, and Deep ANN [6].

The authors introduce a novel transformer-based architecture for Sign Language Translation that jointly learns Continuous Sign Language Recognition and Translation while being trainable in an end-to-end manner. This joint approach solves two co-dependant sequence-to-sequence learning problems, leading to significant performance gains. The authors evaluate the recognition and translation performances of their approaches on the challenging RWTHPHOENIX-Weather-2014T (PHOENIX14T) dataset. They report state-of-the-art sign language recognition and translation results achieved by their Sign Language Transformers, outperforming both sign video to spoken language and gloss to spoken language translation models, in some cases more than doubling the performance. The authors also share new baseline translation results using transformer networks for several other text-to-text sign language translation tasks. The authors emphasize the importance of distinguishing between sign and spoken languages, as the mapping between speech and sign is complex and there is no simple word-to-sign mapping [7].

This paper proposes using Inflated 3D Convolutional Neural Networks for large-scale signer-independent sign language recognition (SLR). The method relies only on RGB video data and does not require other modalities such as depth, making it beneficial for many applications where depth data is not available. The authors show that transferring spatiotemporal features from a large-scale action recognition dataset is highly valuable to the training for SLR. The method is evaluated on the ChaLearn249 Isolated Gesture Recognition dataset and outperforms other state-of-the-art RGB-based methods. The authors present a simple yet effective approach for the recognition of isolated sign language gestures on RGB-only data, outperforming all current state-of-the-art methods. They also show the benefit of using pre-trained weights from action recognition tasks in SLR, which has not been applied to comparable tasks of recognition of well-structured hand gestures [8].

The study proposes a method for hand cropping and normalization using an open-source real-time human pose estimation framework called openpose. Openpose is a deep learning-based framework that detects 2D key points of each individual in an image, improving machine understanding of human activity. It takes an RGB image as input and returns a list of coordinates for all human body key points. The method estimates the hand orientation to one of nine basic directions: perpendicular to the frame's plane, vertical pointing up, diagonal pointing to the top right, horizontal pointing to the right, diagonal pointing to the bottom right, vertical pointing down, diagonal pointing to the bottom left, horizontal pointing to the left, and diagonal pointing to the top left. The cropped hand region is resized to 112×112 pixels and the hand direction is estimated using the horizontal and vertical distances between the wrist and elbow joints. The cropped region is centered on the wrist joint, with an appropriate value of 40 pixels. If only the horizontal distance is negligible, the hand axis is nearly vertical, with the vertical coordinates of the wrist and elbow joints used to indicate

the direction of the axis. If the wrist joint is horizontally less than the elbow joint, the direction is up, and the cropped region is shifted down to avoid inaccuracies [9].

This paper presents a real-time continuous gesture recognition system for sign language using a DataGlove TM. The system solves the critical problem of end-point detection in a stream of gesture input and performs statistical analysis based on four parameters in a gesture: posture, position, orientation, and motion. The prototype system has a lexicon of 250 vocabularies in Taiwanese Sign Language (TWL) and uses hidden Markov models (HMMs) for 51 fundamental postures, 6 orientations, and 8 motion primitives. The system can continuously recognize a sentence of gestures based on these vocabularies in real-time, with an average recognition rate of 80.4%. Previous work has focused on gesture-to-speech interfaces, multilayer perceptron models, and simplified methods for recognizing letters and numbers in ASL. The system described in this paper is an extension of the previous one, incorporating position, orientation, and motion models to enhance performance. The goal is to recognize large sets of vocabulary in sign language by recognizing constructive postures and context information [10].

This paper discusses the development of an automated sign language recognition (SLR) system, which uses a vision-based isolated hand gesture detection and recognition model. The model is trained using a large dataset and the best algorithm, with a convex hull for feature extraction and KNN for classification. The model achieved 65% accuracy, reducing the communication gap among speech and hearing-impaired individuals. The complexity of the model varies due to the different semantics and hand gesturing of sign languages, making it challenging to develop a universal SLR model. [11].

The proposed architecture consists of a feature extraction module using a deep CNN, temporal fusion layers, and a sequence learning module using RNNs with bidirectional long short-term memory (Bi-LSTM) architecture. The architecture uses an end-to-end recognition system to generate alignment proposals between video segments and gestural labels. The feature extraction module is trained and fine-tuned iteratively to handle a large number of gestural segments with supervisory labels. The system uses a CNN followed by temporal convolutional and pooling layers to learn spatiotemporal representations from input video streams. Bi-LSTMs are employed to learn complex dynamics by mapping sequences of spatiotemporal representation to sequences of ordered labels. The output categorical probabilities of M gloss labels at time k are computed through a SoftMax classifier, which takes the concatenation of hidden states of Bi-LSTM as the input. The architecture uses an iterative optimization scheme to effectively train the deep architecture [12].

DeepSLR is a method for recognizing hand gestures using two armbands attached to forearms. These armbands consist of an IMU sensor and sEMG sensors, which capture the acceleration and angular velocity of arm movements and muscle activities reflecting finger motions.

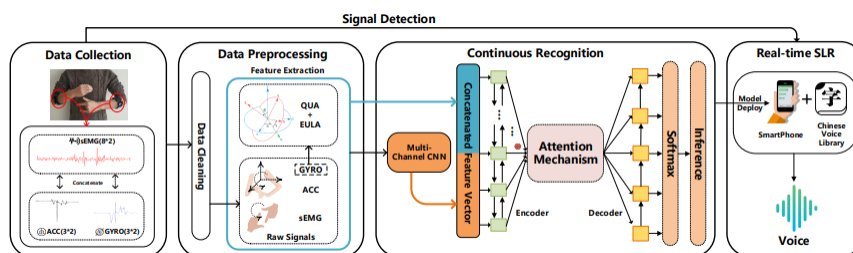


Fig 2. System overview of DeepSLR.

Data collection involves data cleaning and feature extraction to normalize real-time signals and reduce noises. Continuous recognition is achieved using an attention-based encoder-decoder model with a multi-channel CNN, which improves system scalability for different users. The model uses a beam search with a grammar-based language model to infer the most probable word sequence from the probability matrixes. The trained model is deployed on a smartphone, and signals from the armbands are sent to the smartphone using Bluetooth in real time. The output sentence can be displayed in voice using a Chinese voice library [13].

This work explains the components and working principles of an interactive system that detects human hand movement and controls the soft hand to obtain corresponding actions through collected motion data. The system is

divided into three parts, with the first part detecting the action of the human hand and passing it back to the host. The inertial sensor attached to the hand and torso collects the position of the human body in space and transmits the data back to the host for processing. The second part processes and judges the data returned by the acquisition device and issues the control signal to the soft hand part. The software in the host performs this function, converting the spatial position of the human hand feature point into the bending angle information of each finger and the angle information between the fingers. The third part is mainly based on a kind of soft hand, which receives the control signal from the host and controls the air valve to make the soft hand perform the corresponding action. The soft hand has 9 degrees of freedom and can make most of the movements of the human hand. The data processing algorithm aims to convert coordinate data obtained from the Novotron software into status information for each finger. The software solves the information obtained by the sensor according to the internal model as the position coordinate of the finger joint, the end, and the angle of rotation according to the coordinate axis.

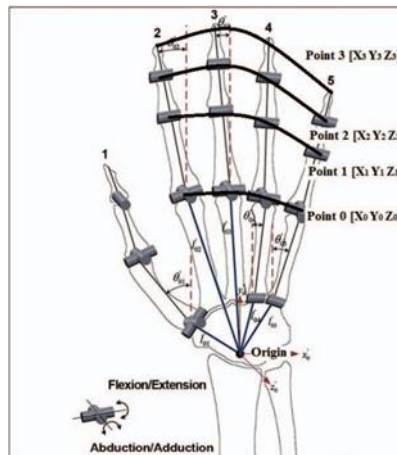


Fig 3. Point position on the finger

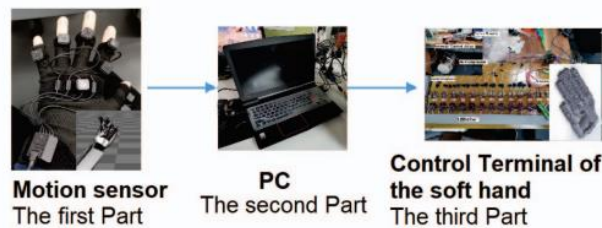


Fig 4. The overall structure of the system.

The coordinates of each node of the five fingers are obtained through calculations such as calculating the coordinate transformation matrix, obtaining the angle between the position feature vector of the middle finger and the plane normal vector of the palm, and determining the angle between the two [14].

This paper presents an efficient sign language translator device using a Convolutional Neural Network (CNN) and customized ROI segmentation. The system uses 5 sign gestures trained using a custom image dataset and implemented on a Raspberry Pi for portability. The ROI selection approach shows better outcomes than conventional approaches in terms of accuracy level and real-time detection from video streaming through a webcam. The method adopts a reverse engineering approach, where a bounding box is present on the display before classification starts, and the user moves the bounding box to the area where the sign is made by a hearing-impaired individual. To make predictions, the trained CNN model receives only the area contained by the bounding box. The main advantage of this process is that CNN does not need to learn a lot of features and detect the ROI. With only a small amount of data, it provides much accuracy and a faster detection rate. The proposed methodology is segmented into two parts: CNN training with training set images and the trained model being implemented in a webcam-connected Raspberry Pi. The convolutional neural

network configuration consists of three convolutional layers and two fully connected layers. The batch normalization technique is used with each convolutional layer to normalize the features of layers. After the completion of training, the model is tested with new test sets of images to come up with unbiased accuracy results. Hardware integration using a low-cost Raspberry Pi 3B is considered for this case, which resulted in a good combination while maintaining the same time and precision level previously tested on a computer. The main concern of this work was to reduce training time and increase accuracy rather than conventional approaches, so data preprocessing was very important for reducing overfitting and under-fitting [15].

This paper proposes a trainable deep learning network for isolated sign language recognition using accumulative video motion. The network comprises three networks: dynamic motion network (DMN), accumulative motion network (AMN), and sign recognition network (SRN). The DMN stream uses key postures to learn spatiotemporal information about signs, while the AMN stream generates an accumulative video motion frame. The extracted features are fed into the SRN for sign learning and classification. The proposed approach works well for isolated sign language recognition, particularly when it comes to static sign recognition.

ArSL, a unified language of several sign languages in Arabic countries, is mainly used in the Arab states of the Arab Gulf countries. The correlation between signs and spoken languages is complex and varies depending on the country more than the spoken language. A descriptive language that uses both manual and nonmanual gestures at the same time is sign language. Signs can be classified into static and dynamic signs. Static signs depend on hand movements for interpersonal communication, while dynamic signs involve manual and/or nonmanual motions of body parts. A video stream is required to represent signs in which the motion component is basic.

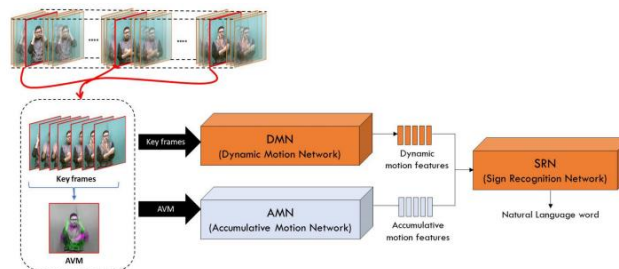


Fig 5. A framework of the sign recognition system.

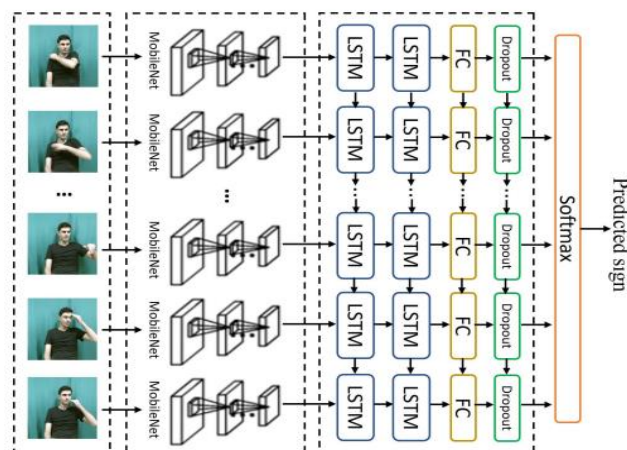


Fig 6. A framework of dynamic motion network model.

Sign language interpretation involves recognition and translation. Recognition involves identifying sign gestures in images or videos and returning their equivalent in a natural language. Isolated sign recognition systems accept a sign and output an equivalent word in a spoken language, while continuous sign language recognition systems identify a sequence of signs performed continuously and output a set of words in the form of sentences [16].

III. COMPARATIVE STUDY

Year	Ref. no	Comparative Analysis	
		Methodology Used	Key findings
2022	[13]	3DCNN, hand gesture recognition, Deep Learning.	A novel system for dynamic hand gesture recognition, using deep learning architectures for segmentation, local and global feature representations, and sequence feature globalization
2021	[9]	3D CNN	Inflated 3D Convolutional Neural Networks are proposed for large-scale signer-independent sign language recognition, relying on RGB video data and transferring spatiotemporal features from a large-scale action recognition dataset.
2020	[8]	K-Means Method	Addressed privacy concerns by applying federated learning techniques to sign language recognition. - Discussed trade-offs between privacy and accuracy.
2020	[7]	Deep Learning (CNNs and RNNs)	Explored the use of deep learning techniques, including CNNs and RNNs, for sign language recognition and translation. - Investigated the impact of different network architectures on accuracy.
2018	[6]	AdaBoost ANN, Deep ANN	Proposed a multimodal deep learning model combining video and sensor data for sign language recognition and translation. - Demonstrated improved accuracy compared to unimodal approaches.
2017	[5]	3D Pose Estimation, Convolutional Neural Network.	Employed 3D pose estimation to recognize signs and translate them to text. - Demonstrated improved accuracy compared to 2D methods.
2016	[4]	Hidden Markov Models (HMMs), Long Short-Term Memory	Focused on continuous sign language recognition for sentences and phrases. - Reviewed various methods, emphasizing context-aware translation.
2016	[3]	Hidden Markov model (HMM),	The Pentium PC -133 PC is used to implement the system. For statistical learning and vocabulary building, there were 196 sentences and 250 vocabulary items, respectively. The four models that follow all employ left-right hidden Markov models, and the feature vectors described in the preceding subsection are used as the HMMs' input.
2015	[2]	Artificial Neural Networks, MATLAB-based	The system carries out many phases of picture processing. For optimal outcomes, the system must correctly identify the acquired image. The ANN will assist the system in identifying the image without the need for laborious calculations or other complicated parts.
2014	[1]	Computer vision-based	The system detects the probable hand from the captured image. To identify the hand in each frame, the system employs cascaded classifiers based on Haar-like features. From the detected hand area, the system extracts the hand sign based on Hue and Saturation values corresponding to human skin color

IV. CONCLUSION

In conclusion, this survey highlights the significance of leveraging action recognition systems for sign language detection. The amalgamation of computer vision and machine learning techniques has paved the way for innovative approaches to bridging communication gaps for the deaf and hard of hearing. The diverse range of methodologies explored in this paper demonstrates the evolving landscape of sign language research. As technology continues to advance, the integration of these systems holds promise for enhanced accessibility and inclusivity. However, challenges such as real-time processing and model robustness persist. Further research and technological refinements are essential to fully unlock the potential of sign language detection through action recognition systems.

REFERENCES

1. Rahaman, M. A., Jasim, M., Ali, M. H., & Hasanuzzaman, M. (2014, December). Real-time computer vision-based Bengali sign language recognition. In 2014 17th International Conference on Computer and Information Technology (ICCIT) (pp. 192-197). IEEE.
2. Pankajakshan, Priyanka C., and B. Thilagavathi. "Sign language recognition system." 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, 2015.
3. Rung-Huei Liang, Ming Ouhyoung A Real-time Continuous Gesture Recognition System for Sign Language 2016 Dept. of Information Management, Shih-Chien University, Taichih, Taipei 104, Taiwan, R.O.C. liang@scc1.scc.edu.tw
4. Tao Liu, Wengang Zhou, and Houqiang Li University of Science and Technology of China Department of Electronic Engineering and Information Science Hefei, Anhui, P.R. China SIGN LANGUAGE RECOGNITION WITH LONG SHORT-TERM MEMORY 978-1-4673-9961-6/16/\$31.00 ©2016 IEEE
5. M. Burić, M. Pobar, M. Ivašić Kos University of Rijeka/ Department of Informatics, Rijeka, Croatia an Overview of Action Recognition in Videos. MIPRO 2017, May 22- 26, 2017, Opatija, Croatia.
6. Rao, G. Anantha, et al. "Deep convolutional neural networks for sign language recognition." 2018 conference on signal processing and communication engineering systems (SPACES). IEEE, 2018.
7. Camgoz, Necati Cihan, et al. "Sign language transformers: Joint end-to-end sign language recognition and translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
8. Haddad, Mark, et al. "Instance-based learning for human action recognition." 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2020.
9. Sarhan, Noha, and Simone Frintrop. "Transfer learning for videos: from action recognition to sign language recognition." 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020.
10. Muneer al-Hammadi, Ghulam Huhammad Wadood, Mansour Alsulaiman, Mohammed a. Bencherif, Tareq s. Alrayes, Hassan Mathkour, and Mohamed Amine Mekhtiche Deep Learning-Based Approach for Sign Language Gesture Recognition with Efficient Hand Gesture Representation Digital Object Identifier 10.1109/ACCESS.2020.3032140.
11. Safeel, Mohammed, Tejas Sukumar, K. S. Shashank, M. D. Arman, R. Shashidhar, and S. B. Puneeth. "Sign language recognition techniques- a review." In 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1-9. IEEE, 2020.
12. Amrutha, K., & Prabu, P. (2021, February). ML-based sign language recognition system. In 2021 International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1-6). IEEE.
13. Shah, Farman, Muhammad Saqlain Shah, Waseem Akram, Awais Manzoor, Rasha Orban Mahmoud, and Daa Salama Abdelminaam. "Sign language recognition using multiple kernel learning: A case study of Pakistan sign language." Ieee Access 9 (2021): 67548-67558.
14. Youme, S. K., Chowdhury, T. A., Ahamed, H., Abid, M. S., Chowdhury, L., & Mohammed, N. (2021). Generalization of Bangla sign language recognition using angular loss functions. IEEE Access, 9, 165351-165365.
15. Ma, Hongxu, Qiang Wang, Xiang Ma, and Mohamed EM Salem. "A sign language interaction system based on pneumatic soft hand." In 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 568-573. IEEE 2020.
16. Luqman, Hamzah. "An Efficient Two-Stream Network for Isolated Sign Language Recognition Using Accumulative Video Motion." IEEE Access 10 (2022): 93785-93798.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details