A New Approach for Arabic Sign Language Recognition *(ArSLR)*

Adham Mohamed Farouk
Scientific Computing Department,
Faculty of Computers and Information,
Sciences, Ain Shams University, Cairo-Egypt
20201700101@cis.asu.edu.eg

Abdelrahman Mohamed Zenhom
Scientific Computing Department,
Faculty of Computers and Information,
Sciences, Ain Shams University, Cairo-Egypt
20201700445@cis.asu.edu.eg

Eman Mahmoud Abdelaleem
Scientific Computing Department,
Faculty of Computers and Information,
Sciences, Ain Shams University, Cairo-Egypt
20201700166@cis.asu.edu.eg

Sara Ahmed Fadel
Scientific Computing Department,
Faculty of Computers and Information
Sciences, Ain Shams University, Cairo-Egypt
20201700329@cis.asu.edu.eg

Khloud Khaled Attia Elsayed
Scientific Computing Department,
Faculty of Computers and Information
Sciences, Ain Shams University, Cairo-Egypt
20201700238@cis.asu.edu.eg

Mohamed Saeed Mohammed
Scientific Computing Department,
Faculty of Computers and Information
Sciences, Ain Shams University, Cairo-Egypt
20201700904@cis.asu.edu.eg

Roula Hani Hassan
Scientific Computing Department,
Faculty of Computers and Information
Sciences, Ain Shams University, Cairo-Egypt
Roula.hani@cis.asu.edu.eg

Howida A. Shedeed
Scientific Computing Department,
Faculty of Computers and Information
Sciences, Ain Shams University, Cairo-Egypt
dr_howida@cis.asu.edu.eg

2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES) | 979-8-3503-7851-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/NILES63360.2024.10753193

Abstract—This paper presents a new approach for Arabic sign language (ArSL) recognition, employing advanced feature extraction and machine learning techniques. Using Media-Pipe, a robust framework for real-time hand and pose detection, key features are extracted from static and dynamic sign language videos. For static sign language recognition, we integrated TensorFlow Lite, which achieves an impressive accuracy of 99.1% with a CNN model and 99.5% with a Support Vector Machine (SVM) model. For dynamic sign language, we implemented a Long Short-Term Memory (LSTM) model within TensorFlow, which attained an accuracy of 93%. Enhanced performance reached 98% with a hybrid CNN-LSTM model. Our approach addresses the limitations of existing methods by eliminating the need for specialized equipment and ensuring high performance on mobile devices. Another main contribution of this research is the developing of our own dataset for static and dynamic models for Arabic Sign language. The developed static dataset is for 27 Alphabet classes (1000 images per class) and 1000 images for each word of 10 words. For the Dynamic dataset, 60 videos were captured for each class. To automate the process of data collection and feature extraction, an automated function was developed to capture 100 images within 3 seconds. Also, a script utilizing OpenCV and Media-Pipe was implemented to capture and process video frames and extracts key points essential for model training.

Keywords:
Sign Language (SL), Arabic Sign Language (ArSL), Arabic Sign Language Recognition (ArSLR), LSTM model, CNN model.

I. INTRODUCTION

Sign language is used by more than 70 million people worldwide to enable the speech-impaired community to interact with each other's. According to the World Federation of the Deaf, there are more than 300 sign languages used by 70 million deaf people worldwide. Many individuals with speech and hearing impairments cannot read or write in their native languages. Sign language (SL) is the primary mode of communication for deaf individuals, relying predominantly on gestures to convey meaning. SL employs various body components such as fingertips, wrists, arms, heads, bodies, and facial expressions. As a systematic combination of hand gestures and facial expressions, SL facilitates daily communication between the speech-impaired community and their surroundings.

Hearing impairment encompasses partial or total hearing loss in one or both ears, with disability levels ranging from mild to profound. People with hearing or speech impairments primarily depend on sign language (SL) for everyday communication. Among different gestures, SL is the most formal, featuring a wide array of signs, each with a specific meaning. A functional sign language recognition (SLR) system can enable deaf individuals to communicate with non-signers without the need for an interpreter.

In This research Arabic sign language application is developed to help Arabian deaf and mute individuals communicate with each other and with the surrounding environment. It enables the community to understand their needs or translate their sign language. Therefore, we decided to develop a simple application to convert speech to sign

language and vice versa to ease our communication with them. The application can also translate text-to-sign and sign-to-text. Text is translated to sign language using stored sign videos to represent the movements of the translated words. Also "sign to text" translates sign language to text using a mobile camera to capture sign language images, which a model then interprets and converts to text.

The user will use this app to type a text, and the system will translate it to sign language. Conversely, someone else can capture a picture, video, or even real-time sign language, and the system will convert it to text to understand their needs and facilitate communication without needing to learn sign language. By creating this app, we aim to help both the deaf and mute community and those who want to communicate with them, thus overcoming communication barriers and avoiding embarrassment for both parties. Figure 1 shows the Arabic Sign language images for the Alphabets.

The developed Android application in this research utilizes computer vision for real-time sign language recognition. This research aspires to create a valuable tool for bridging communication gaps between the deaf and mute communities and those who want to communicate with them. The research application focuses on helping disabled Arabic speakers by providing an Arabic sign language translator.

In this research work, we utilize the Long Short-Term Memory (LSTM) network model for Arabic Sign Language Recognition. By leveraging the capabilities of LSTM models, this research work aims to enhance the accuracy and efficiency of ArSL recognition, bridging communication gaps and fostering greater inclusivity for the deaf community within Arabic-speaking regions.

LSTM networks are particularly well-suited for sequence prediction tasks due to their ability to capture long-range dependencies. Despite the advancements, several challenges remain in developing robust ArSL Recognition systems using LSTM:

Data Scarcity: A significant challenge is the limited availability of large, annotated ArSL datasets. This hampers the training of deep learning models, which require extensive data to generalize well.

Variability in Signing Styles: Differences in signing styles among individuals can affect recognition accuracy. Addressing this variability is crucial for developing generalized models.

Complexity of ArSL Grammar: The grammatical structure of ArSL, which may include non-manual signals (facial expressions, body posture), adds to the complexity of recognition tasks.

The paper's organization is as follows: Section 2 provides information about related work. Section 3 presents the system architecture of the recognition system. Section 4 presents the results. Finally, section 5 discusses the conclusion and future work.

I. Related Work

The field of sign language recognition, particularly Arabic Sign Language Recognition (ArSLR) using Long Short-Term Memory (LSTM) networks, has seen significant advancements. This section reviews the existing literature and research efforts related to ArSLR and the application of LSTM in sign language recognition systems.

Sign language recognition has been an active area of research for several decades. Early efforts focused on static gesture recognition using image processing techniques. More recent approaches have leveraged advancements in machine learning and deep learning to handle dynamic gestures and continuous sign language sequences.

Fig. 1. ArSL images for Alphabets

Research specific to Arabic Sign Language (ArSL) has been comparatively limited but growing. Various approaches have been explored, including the image-based or Sensor-based approach.

Image-Based Recognition: Early studies used image processing techniques to recognize static signs. In [1] Authors utilized image segmentation and feature extraction to classify ArSL alphabets.

In [2] The authors developed a robust deep learning model that is trained on the ArSL2018 dataset to convert images of the ArSL alphabets into Arabic alphabets. Accuracy reached 99.4 using the ResNet-18 model for recognizing alphabets only.

Sensor-Based Approaches: Some studies employed sensor data, such as accelerometers and gyroscopes, to capture motion data for ArSL. In [5] Authors developed wearable sensors to detect hand movements for ArSL recognition. In [6] and [7] authors use Leap Motion Sensors to capture hand signs. The classification was done using NN models with accuracy reached to 98% and 88% respectively.

Recent research has focused on deep learning models to improve recognition accuracy. In [8] Authors developed CNNs which have been employed to extract spatial features, while RNNs (Recurrent Neural Network), particularly LSTMs, have been utilized to capture temporal dependencies in sign sequences. The work demonstrated the use of CNN-LSTM models for recognizing ArSL words from video sequences. In [9] CNN network was used to classify the Arabic Sign language. The model is tested on ArASL2018 dataset..

In [10] Authors applied a combination of CNN and LSTM networks to capture both spatial and temporal features from video frames, achieving state-of-the-art results in continuous Arabic Sign Language (ArSL) recognition. CNN and LSTM classifiers achieved accuracies of 94.40% and 82.70%, respectively.

Bidirectional LSTMs have been explored to capture context from both past and future frames, enhancing the model's ability to understand the overall gesture sequence. In [11] Authors demonstrated the effectiveness of Bi_LSTM networks in improving recognition accuracy for sign language videos.

II. SYSTEM ARCHITECTURE

The proposed system architecture for detecting sign language involves several stages; data acquisition, features extraction and finally the classification stage. The process handles input from three diverse sources: real-time video streams, static images, and pre-recorded videos. The following sections describe each component, the flow of data through the system, and the specific challenges and solutions encountered during the data collection stage.

A. Input Sources

1. Real-time Video: This involves capturing frames in real-time from a live video feed.
2. Static Images: This involves processing a single input image.
3. Pre-recorded Videos**:** This involves extracting frames from a pre-recorded video file.

B. Data Collection Stage

The data collection stage for this research encountered several significant challenges:

1. Limited Public Datasets:
   - Scarcity: Publicly available datasets, especially those containing dynamic data, are limited.
   - Dependence on Private Datasets: Most current research utilizes private datasets, which are not accessible for public use.
2. Unusable Public Datasets:
   - Size: Available public datasets often contain too few images, typically around 500 images.
   - Quality: Issues such as inconsistent image sizes, color variations, and data corruption are prevalent in these datasets.
3. Requirements:
   - Quantity Requirements: Many images per class are needed, with a minimum threshold of 1,000.
   - Dynamic Data: For dynamic datasets, 60 frames per video are required for each movement.

To address these challenges, we developed several solutions:

- **Automated Image Capture Function:** An automated function was developed to capture 100 images within 3 seconds.
- **Automated Data Collection Script:** A script utilizing OpenCV and Media-Pipe was implemented to automate the data collection process. This script captures and processes video frames and extracts key points essential for model training.

Then the total collected dataset size: 27,000+ images for the Arabic Alphabet (27 classes x 1000 images) and 1000 images for each word of 10 words.

The classified Arabic words in our dataset are:
(انت، الساعة كام، اين المكان، سعيد، حزين، شكرا، الحمد لله، إلى اللقاء، مساء النور)

This comprehensive approach to data collection ensures that our system can handle the various input sources effectively while overcoming the limitations of existing datasets. The integration of automated image capture and data collection scripts enhances the efficiency and accuracy of the dataset, supporting robust model training and improving the overall performance of the sign language recognition system. System architecture steps are as shown in Figure 2.

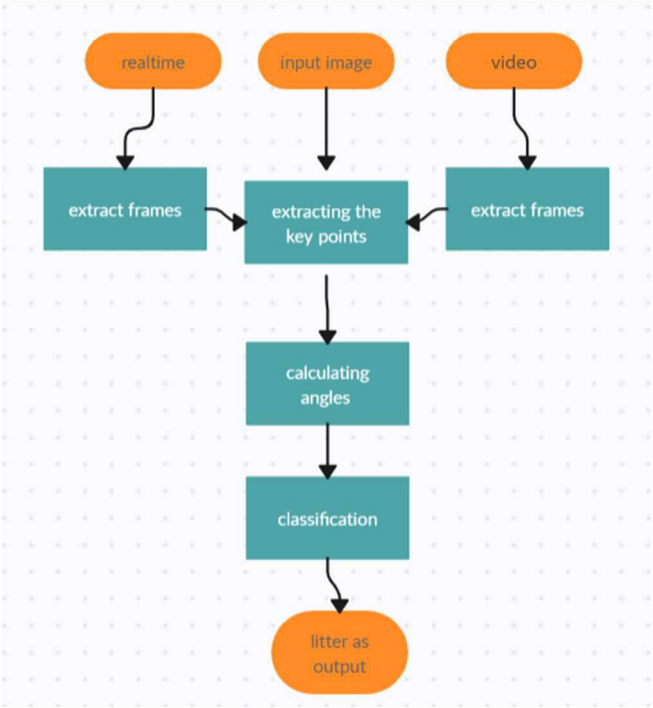

Fig 2. Proposed system Architecture.

C. Collecting Dataset

The issue with the dataset was that Arabic sign language datasets available online are exceedingly rare, and most research papers on this topic created their own datasets without publishing them online. To overcome the limitation issues for the available public dataset addressed in the previous section, we opted to create our own dataset, following the example set by researchers, to provide the model with high-quality data resembling real-life images. We collected images by photographing the hands gestures of the authors and their family, and friends. This task was challenging as we aimed for at least 1000 images per class. To expedite the process, we developed a function capable of capturing 100 images in less than 3 seconds. For the dynamic component, four actions were included, each with 60 videos.

Each video will consist of 30 frames, and each frame contains 258 landmarks.

D. Feature Extraction

The feature extraction stage is a crucial component of our system for sign language detection, involving several detailed techniques to ensure accurate and comprehensive data representation. We utilize Media-Pipe to extract key hand features from real-time frames. Media-Pipe is adept at detecting and tracking 21 key points for each hand, including essential points such as finger joints. This granular detection enables precise identification of hand positions and movements, which are fundamental for interpreting sign language gestures

Fig 3. Sample from the developed dataset, illustrating the extracted key features.

In addition to detecting key points, the system calculates angles between specific hand key points, providing additional context about hand posture. This step is crucial for distinguishing between similar gestures that may differ only in the orientation or angle of the fingers and hand, thereby enhancing the accuracy of the feature representation. To ensure the correct interpretation of features, our system identifies and separates key points for the right and left hands. This distinction is important as it prevents misinterpretation of mirrored gestures and ensures that the features extracted reflect the actual hand movements accurately.

All extracted features and calculated angles are compiled into a single feature vector. This vector serves as a comprehensive representation of the hand gesture, encompassing both positional and angular information. The feature vector is then used as input for the subsequent classification stage. To improve the model's generalization capabilities, we apply data augmentation techniques. These include reflection and translation operations along the x-axis. By increasing the diversity of the training data, these augmentations help the model perform better on unseen data, ultimately leading to more robust and reliable gesture recognition. Figure 3 illustrates a sample from the developed dataset, explaining the extracted key features.

E. Classification Model

The classification stage of our system employs a hybrid model that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This hybrid approach is designed to handle the complex nature of sign language, which involves both spatial and temporal dimensions. CNNs are used to extract spatial features from the input data. They excel at capturing static aspects of hand gestures, such as finger positions and hand shapes. CNNs recognize patterns in image-like data, which is essential for understanding the visual components of sign language. LSTMs are designed to handle sequential data and capture temporal dependencies. They learn and remember long-term dependencies in the sequence of hand movements, which is crucial for recognizing dynamic gestures that unfold over time. By combining CNNs and LSTMs, the model leverages the strengths of both architectures. CNN handles spatial aspects, while the LSTM captures the temporal evolution of gestures. This combination is particularly effective for sign language, which involves complex sequences of movements. Figure 4 illustrates the proposed architecture of the CNN Model.

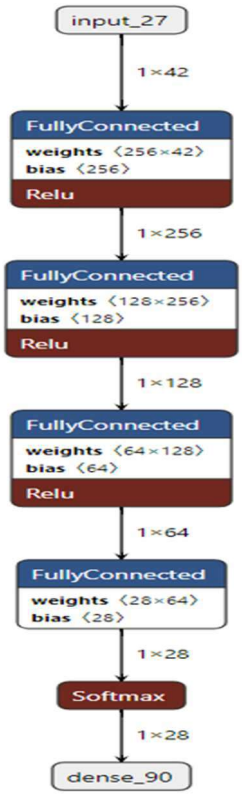

Fig 4. The proposed Neural network architecture

The hybrid CNN+LSTM model can process both static hand positions and the dynamic transitions between them, essential for accurately interpreting sign language. This approach yields better results for complex, time-dependent

gestures compared to using either CNN or LSTM alone. The model's architecture allows to handle a wide range of sign language gestures, making it versatile and adaptable to different types of signs.

The architecture of our CNN+LSTM model begins with feature extraction using Conv1D to identify local patterns from the input sequence. This is followed by dimensionality reduction via MaxPooling1D, which reduces the size of feature maps, focusing on the most prominent features and decreasing computational load. The LSTM component captures temporal dependencies and sequential patterns, crucial for understanding the flow of hand gestures. The output from the LSTM layer is then flattened to prepare it for dense layers, which learn and refine high-level feature representations for final classification. The final output layer generates probability distributions over possible hand gestures, enabling accurate classification. This hybrid model approach offers several advantages, including the effective capture of both spatial and temporal patterns, superior performance in sign language translation tasks, and the capability for real-time translation with high accuracy. The goal of this model is to achieve real-time sign language translation, facilitating seamless communication and enhancing accessibility for the deaf and hard-of-hearing.

IV Experimental Results

Experiments are done using different classifiers on our own private dataset with the proposed features vector. Table 1 illustrates the achieved accuracies using different models learned on our developed private dataset. As shown in the table, for static sign language recognition, we integrated TensorFlow Lite, which achieves an impressive accuracy of 99.1% with a CNN model and 99.5% with a Support Vector Machine (SVM). For dynamic sign language, we implemented a Long Short-Term Memory (LSTM) model within TensorFlow, which attained an accuracy of 93%. Enhanced performance reached to 98% with a hybrid CNN-LSTM model. Figure 5 illustrates the Overview of the Sign Language Detection and Translation System run. Figure 6 illustrates the overview of the system's steps to translate a gesture. Table 2 illustrates a comparison with recent research work in [3] that is close to the proposed work for classifying the static and dynamic Arabic signs. The results show that our proposed approach enhanced the performance of Arabic sign language recognition. Frame-based accuracies explain the higher performance with static and dynamic signs.

V Conclusion and Future Work

In This research Arabic sign language recognition Android application is developed to help Arabian deaf and mute individuals. The application converts text to sign language and vice versa.

In conclusion, our research demonstrates the superior efficacy of Media-Pipe combined with TensorFlow Lite and LSTM models for Arabic sign language recognition. The SVM model, paired with Media-Pipe, emerged as the optimal choice for static sign language, achieving a remarkable accuracy of 99.5%. For dynamic gestures, the hybrid CNN-LSTM model proved the most effective, reaching 98% accuracy. Our solution not only overcomes the challenges posed by the scarcity of Arabic sign language datasets but also ensures scalability and real-time performance on mobile platforms. Future work will include adding more dynamic signs for our dataset.

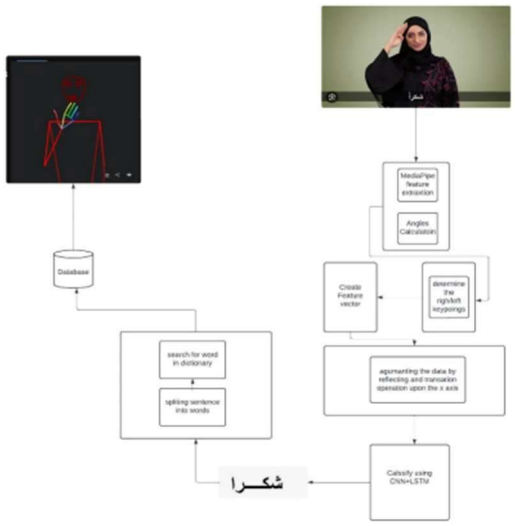

Fig 5. Overview of the Sign Language Detection and Translation System run for the word "شكرا"

TABLE 1 ACCURACIES ACHIEVED USING DIFFERENT LEARNING MODELS

TABLE 2. COMPARISON WITH OTHER WORK IN [3]

	[3]	Our proposed Approach
Classification Models	3D CNN combined with a 2D point convolution network.	SVM + Media-Pipe for static Hybrid CNN-LSTM model is used for both static and dynamic gesture recognition.
Classified signs	28 static signs and 52 dynamic signs	28 static signs and 10 dynamic signs
Accuracy (static)	88.89%	99.5% with SVM + MediaPipe
Accuracy (Dynamic)	98.39%	98% with CNN-LSTM

Frame-based Accuracy Comparison (for static)	10 frames:85.8% 20 frames:91.7% 30 frames:94.1%	10 frames:97% 15 frames:99% 20 frames:99.5%

Frame-based Accuracy Comparison (for dynamic)	10 frames:85.8% 20 frames:89.1% 30 frames:94.1% 40 frames: 95.25% 100 frames:97.2%	10 frames:89% 20 frames:93.7% 30 frames:98% 40 frames:97.4% 60 frames:98 %

Fig 6. Overview of the system's steps to translate a gesture for the character "ك".

REFERENCES

[1] A. S. Elons, M. Ahmed and H. Shedid, "Facial expressions recognition for Arabic sign language translation", 9th International Conference on Computer Engineering & Systems (ICCES), Cairo, Egypt, 2014, pp. 330-335.

[2] Muhammad Al-Barham; Ahmad Abu Sa'Aleek; Mohammad Al-Odat; Ghada Hamad; Musa Al-Yaman, "Arabic sign language recognition using deep learning Models, 2022 13th International Conference on Information and Communication Systems (ICICS)

[3] M. Abdelali, A. Abdelali, H. El Majdoudi, and M.Ouhbi, "Arabic sign language recognition system using 2D Hands and body skeleton data," International Journal of Human-Computer Studies, vol. 171, p. 109068, 2023.

[4] M. Maraqa and F. Abu-Zaiter, "Arabic sign language recognition system for alphabets using machine learning techniques," Journal of Computer Science, vol. 17, no. 12, pp. 1223-1231, 2021.

[5] M. I. Al-Haj and M. I. Khamis, "Real-time Arabic sign language alphabets (ArSLA) recognition model using deep learning architecture," Sensors, vol. 22, no. 5, p. 1702, 2022.

[6] M. Mohandes, A. Aliyu, and M. Deriche, "Arabic sign language recognition using Leap Motion Controller", 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE).

[7] Elons, A.S., Ahmed, M., Shedid, H., Tolba, M.F., "Arabic sign language recognition using leap motion sensor", Proceedings of 2014 9th IEEE International Conference on Computer Engineering and Systems, ICCES 2014, pp. 368–373.

[8] T. M. Alawwad and M. Bchir, "Arabic sign language recognition using Faster R-CNN," Journal of King Saud University-Computer and Information Sciences, vol. 35, no. 7, pp. 1117-1124, 2023.

[9] Z. Al-Khatib, M. Maraqa, and S. Al-Hamad, "Design of Arabic sign language recognition model," International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 28-34, 2023.

[10] Talal H. Noor, Ayman Noor, Ahmed F. Alharbi, Ahmed Faisal, Rakan Alrashidi, Ahmed S. Alsaedi, Ghada Alharbi, Tawfeeq Alsanoosy, and Abdullah Alsaeedi, "Real-Time Arabic Sign Language Recognition Using a Hybrid Deep Learning Model", Sensors 2024, 24(11).

[11] T. M. Alawwad, M. Bchir, and M. Maher, "Arabic sign language recognition using a novel deep neural network architecture", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.