Using Machine Learning to Transcribe Visual Arabic Sign Language

Anas Al –Tirawi, Fatma Al – Amoudi, Rahaf Al – Shabrawi, Judi Al – Baghajati

atirawi@dah.edu.sa, faalamoudi@dah.edu.sa
raaalshabrawi@dah.edu.sa @dah.edu.sa,
jhalbaghajati@dah.edu.sa
[1] Dar Al Hekma University, Jeddah 23535-2440, Saudi Arabia

Abstract. In the era of technological advancement, clear communication is crucial, but using online resources can be challenging for those with hearing impairments. While assistance programs exist for English speakers, many regions, including the Middle East, lack such resources. Our goal is to develop software that translates hand gestures seen in videos into Arabic text through computer vision. People with hearing disabilities can now successfully pursue opportunities in their education and careers, interact with society, and obtain information efficiently.

Keywords: Arabic Sign Language, Machine Learning, Computer Vision, Hearing Disabilities, Image processing

1 Introduction

Despite technological advancements, many solutions have been developed to aid hearing-impaired individuals who speak. However, there is a notable absence of equivalent resources for the Arabic language. This difference demonstrates the necessity of adopting new approaches to overcome communication hurdles encountered by deaf people living in the Middle East.

In this context, researchers have been motivated by the transformative potential of technology in eliminating barriers and promoting inclusiveness. We will utilize machine learning and computer vision tools to develop a software program that facilitates communication between deaf individuals using Arabic Sign Language (ArSL) through text. This innovative solution has the potential to enable hearing-impaired individuals to interact with digital content, actively participate in their communities, and pursue educational and employment opportunities on an equal footing with their peers. This research's aim is to develop a machine-learning system capable of accurately analyzing hand movements in video clips of ArSL and instantly translating them into Arabic texts. The main objectives of this research are as follows:1. Gather and prepare a comprehensive dataset of ArSL videos with Arabic translations. 2. Develop and deploy an effective, efficient machine-learning model for hand gesture recognition and translation. 3. Evaluate the performance of the developed system through extensive testing and validation, focusing on accuracy, speed, and ease of use. The rest of the paper is structured as follows: Section two summarizes earlier research and literature on sign language recognition and translation. Additionally, Section Three outlines the methodology for data collection and pre-processing, along with the design and implementation of the machine learning model. Furthermore, Section Four presents an experimental framework and evaluation results. Finally, in section five we discuss the results and recommendations of the study with a summary of its contributions.

2 Literature Review

The application of machine learning methods to the discipline of ArSL recognition has resulted in the implementation of various techniques that have led to tremendous improvements in this field. From the use of CNNs, transfer learning, and varying deep architectures to the development of numerous models, multiple studies have gone to great lengths to refine recognition accuracy. Suliman [1] developed an LSTM and CNN-based approach that successfully extracts features while maximizing accuracy. This approach that uses a combination of CNN and sequence modeling achieves 95.9% accuracy when Ardebil signers are given the medium to interact. This demonstrates that this approach is effective. This is followed by Saleh [2] who employed deep learning and the transfer learning of convolutional neural networks to successfully reach a level of recognition accuracy close to 100% in identifying 32 different gestures signifying Arabic sign language. The study points to the possibility that models can be transferred from the classroom to the field leading to insights that can substantially improve the recognition performance. According to Alzohairi [3], the concept of image-based recognition was tackled. This notion was especially connected to visual descriptors such as Histograms of Oriented Gradients (HOG) which were vital in recognizing the Arabic Script Language (ArSL) letters. It furthermore illustrates the effectiveness of the feature extraction approaches that boost the recognition precision. In their work, Duwairi [4] proposed an innovative strategy based on transfer learning and state-of-the-art deep learning models achieving an accuracy of recognition of Arabic alphabets in sign language b altogether 97%. This study achieves the goal of utilizing existing stacked deep network architectures by accurately recognizing proofs. In the latest edition, Buttar et al. [5] designed a live gesture recognition system that is based on a combination of skeleton models and YOLOv6 to detect static gestures. That reproduces both recognition effectiveness and enables real-time translation which makes it a good choice for communication between these individuals and with other people. At the same time, the researchers Tharwat, Ahmed, and Bouallegue [6] stated that it was possible to develop an Arabic Alphabet Sign Language Recognition System following with the achieved accuracy of 99.5% that was accomplished using the K-Nearest Neighbor (KNN) classifier. In this system, the deaf community is treated of learning Arabic sign language letters and read Quranic surahs. AArSLRS includes some of the major blocks such as, data processing, preprocessing, feature extraction, and classification features, which can be mixed and matched to assist in identifying the Arabic sign language. Moreover, Almasre & Al-Nuaim [7] were able to encode and identify ArSL words, which achieved high accuracy levels for different word classes up to the rung of 97.059%. Studies conducted by the researchers therefore emphasize the fact that DSL is imperative and ML algorithms should be used for precise ArSL classification.

Speech recognition is one of the areas of machine learning that has recently had very good results, especially in the field of sign language interpretation. There is where a concern by the Ahli [8] who put forward a machine learning model built specifically to transfer sign language videos automatically into text format. According to the Ben Ali's model, the system involves video sequences including spatial and time characteristics, acquired from an Aromatic boatneck crop top and Low-rise distressed straight-leg jeans with shoe shopping video. The system uses different machine learning algorithms for training and prediction.

Generally, these articles represent the fact that the algorithms based on machine learning boast high efficacy in the area of ArSL recognition foundations and promote efficiency in communications with impaired speech and hearing people.

3 Methodology

3.1 Machine Learning and Computer Vision Approaches

Recent advancements in computer vision and machine learning techniques have catalyzed a surge in interest in Arabic Sign Language (ArSL) recognition. Despite the Arabic Language ranks among the most widely spoken languages globally, Arabic Sign Language (ArSL) remains relatively nascent in its development. Because Arabic-speaking countries have a wide variety of regional dialects, there are numerous ArSLs, each with distinctive qualities of its own. One of the important obstacles to ArSL recognition is the “diglossia” phenomenon, where various dialects are spoken in place of written languages. However, most ArSLs share the same alphabet and limited vocabulary. To address this issue, researchers have suggested using multi-modality ArSL databases that concentrate on joining both manual and non-manual gestures. These databases provide a valuable resource for training and evaluating ArSL recognition systems. These databases add a wide variety of gestures and expressions, which aid in building strong models that can handle the complexities of different ArSLs.

Using Long Short-Term Memory (LSTM) networks is one noteworthy development in this field. Recurrent neural networks (RNNs), such as LSTM networks, are especially well-suited for ArSL recognition tasks because they are excellent at recognizing temporal dependencies in sequential data. Accurate ArSL recognition systems have been made possible by the fusion of computer vision and machine learning techniques, which has also created opportunities for real-time translation and interpretation applications. By improving their access to information and communication channels, these developments have the potential to greatly empower the deaf and hard-of-hearing community. The following sections provide more detail on the particular approaches we used in our study, including the datasets we used, the model architectures we created, and the evaluation metrics we used to gauge how well our ArSL transcription system performed.

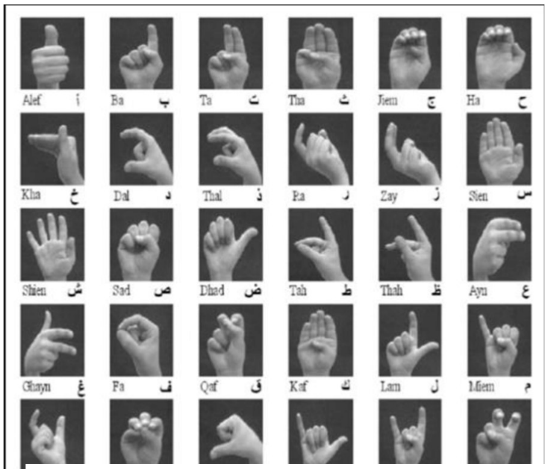


Figure 1: Arabic Sign Language Alphabets

3.2 Dataset Selection and Preprocessing:

3.2.1 Dataset Selection:

For our research on Arabic Sign Language (ArSL) transcription, we utilized one primary dataset:

1. **KFUMP Arabic Sign Language (KArSL-502):**
   KArSL is the largest video dataset for Word-Level Arabic Sign Language (ArSL). The dataset consists of 502 isolated sign words which are collected by Microsoft Kinect V2. Signs are performed by three professional signers and repeated 50 times each.

A comprehensive and representative training set for our ArSL transcription system was the driving force behind our decision to use KArSL-502. Our goal is to represent the depth and breadth of ArSL gestures, including both common signs and regional differences.

3.2.2 Preprocessing:

The primary objective we set during the preprocessing stage was to optimize the dataset to guarantee its quality and appropriateness for the next stage of model training. We approached this phase with meticulous attention to detail because we understood how important high-quality data is to the success of machine learning efforts. We sought to improve the overall dependability and effectiveness of the datasets by methodically converting the data into a usable form for the model. Our aim was accomplished by utilizing an extensive range of preprocessing methods that were customized to the unique features of the information. The videos were cropped to 256 x 256 and each were separated into individual frames for faster feature extraction. To guarantee uniform data distribution across features and samples, normalization and standardization techniques were implemented. Our goals were to lower biases and improve model stability by maintaining equal length for all the videos. This was done by padding all the videos to be the same length as the longest video in the group, which increased the consistency of the data.

Throughout the preprocessing pipeline, particular attention was paid to preserving data consistency and integrity. This included splitting the data appropriately for testing and validation, carefully extracting features, and timing gestures. Following rigorous preprocessing guidelines created a strong basis for later model training and assessment.

3.2.3 Feature Extraction

MediaPipe framework by Google delivers an easy solution for hand and face detection by providing landmark detection and drawing. The hand landmark detection utility allows the tracking and identification of hand gestures and key points.

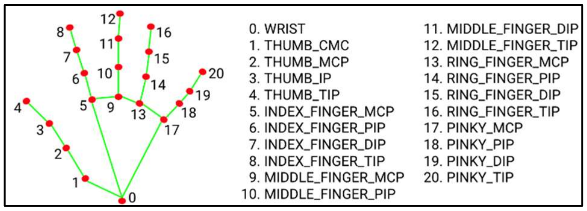


Figure 2: Hand landmarks

As seen in figure 2, there are 21 landmarks detected in each hand which results in 42 key points for both hands. Each key point is assigned a number for faster feature extraction.

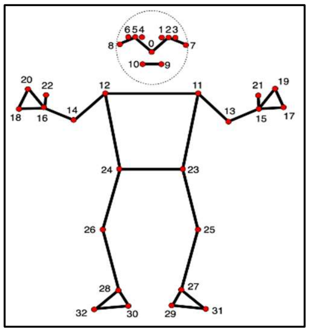


Figure 3: Pose landmarks

Figure 2 features pose landmarks detected by MediaPipe's pose estimation. With 33 landmarks, the X, Y, and Z dimensions are computed to predict the hidden parts of the body. Face gestures are also a part of the signs in sign language, hence capturing the facial features is a necessity. MediaPipe helps to capture the main landmarks with 468 key points.

Each key point has 3 dimensions, x, y, and z. The total number of face key points is (468 × 3) = 1404 key points. Pose key points are calculated in a slightly different manner, as the visibility of a frame's pixel is based on how hidden the pixel is by another body part. Thus, the total number of key points is indicated by a number of key points × (3 dimensions + visibility) which is respectively (33 × (3 + 1)) = 132 key points. The hands key points are 21 in each hand and in three dimensions, so the total number of key points is 42 × 3 = 126 key points extracted from hands. The total number of key points extracted in each frame is the following:

1404 + 132 + 126 = 1662 key points that indicate a gesture.

In each video frame, features were extracted and stored in a separate NumPy array (.npy file) for the labeling process and model preparation. The number of signs used in this model is 150 signs, each having a variable number of frames and repeated around 50 times.

3.2.4 Data labeling

To label the data, the features were loaded from the stored NumPy files and padded for consistency, then they were stored in a NumPy array with the following dimensions (6353, 116, 1662) where 6353 is the number of sequences, 116 is the length of the longest sequence, and 1662 is a number of key points extracted in each sequence.

The labels were stored in a separate NumPy array with two dimensions (6353, 150) which are the number of sequences, and the number of signs detected.

3.3 Model Architecture and Training:

Within the "Model Architecture and Training" section, the main emphasis is on outlining the neural network architecture design and optimization procedure for Arabic Sign Language (ArSL) transcription. To optimize model performance, this phase entails carefully choosing and configuring the right models and fine-tuning training protocols. Understanding the computational frameworks and strategies used to efficiently transcribe ArSL gestures is provided by describing the architectural decisions and training methodologies used.

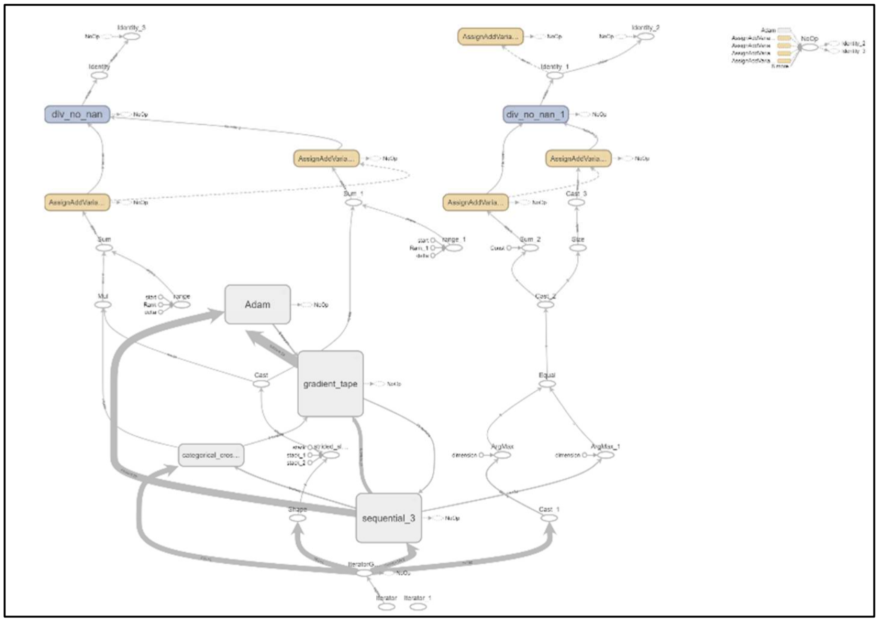
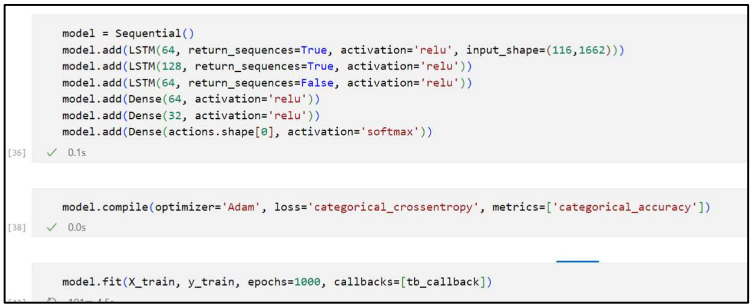


Figure 4: The design of the Recurrent Neural Network

Strict evaluation metrics are used to monitor model performance and direct iterative refinement during the training phase. This includes metrics that shed light on the model's capacity to faithfully translate ArSL gestures. Through ongoing assessment and optimization of the model based on these metrics, the objective is to create a stable and dependable transcription system that can efficiently overcome communication obstacles for people who use sign language. The model chosen for the transcription system was the LSTM (Long Short-Term Memory) model, which is a type of RNN (Recurrent Neural Network) that is aimed at dealing with long-term dependencies and sequential data, which is well suited for our use case. The layers used in our model are the following: three LSTM layers and three dense layers, each with a different number of hidden units, and an input shape of 116 maximum sequence length, and 1662 features.



```
    model = Sequential()
    model.add(LSTM(64, return_sequences=True, activation='relu', input_shape=(116,1662)))
    model.add(LSTM(128, return_sequences=True, activation='relu'))
    model.add(LSTM(64, return_sequences=False, activation='relu'))
    model.add(Dense(64, activation='relu'))
    model.add(Dense(32, activation='relu'))
    model.add(Dense(actions.shape[0], activation='softmax'))
[36] ✓ 0.1s

    model.compile(optimizer='Adam', loss='categorical_crossentropy', metrics=['categorical_accuracy'])
[38] ✓ 0.0s

    model.fit(X_train, y_train, epochs=1000, callbacks=[tb_callback])
```

Figure 5: LSTM Model layers

4 Results

This section contains the results of our research on transcription from Arabic Sign Language (ArSL), which we have carefully examined. By conducting thorough experiments and analysis, we have gained important insights into the effectiveness of the approaches we have suggested. In addition to quantitative metrics, qualitative analyses offer a more in-depth understanding of the strengths and weaknesses of our models. The models' ability to faithfully translate ArSL gestures is demonstrated by sample outputs and visualizations, which provide a concrete depiction of their performance. Preprocessing methods, training schedules, and model architectures all lend themselves to interesting comparison studies. Through a comparative analysis of our methods with current cutting-edge techniques, we shed light on the advantages and disadvantages of each approach, enabling future researchers to make well-informed decisions.

Our analyses provide insightful information about the variables affecting model performance as well as the usefulness of our conclusions. Through the analysis of qualitative and performance metrics, we can obtain a deeper understanding of the fundamental mechanisms that govern the accuracy of ArSL transcription. Furthermore, by improving the deaf and hard-of-hearing community's access to information and communication channels, our research has implications for practical applications in sign language recognition.

5 Analysis / Benchmarking

We carry out an extensive review of our research findings in Arabic Sign Language (ArSL) transcription, compared with recognized standards and literature. Our meticulous investigation includes an in-depth comparison of our methods with current cutting-edge practices, including assessments of model architectures, training schedules, and preprocessing methods. Utilizing this comparative analysis, we can reveal subtle insights regarding the relative benefits and drawbacks of each approach, thereby laying out a plan for further research and methodological improvements. We also explore the wider implications of our findings, including how they might progress the field of ArSL transcription and help the deaf and hard-of-hearing community become more inclusive and accessible. In the end, we hope to promote an inclusive society by accelerating the advancement of ArSL transcription techniques by identifying promising directions for future research and development. We also have critical conversations about the ethical implications and societal effects of our research, making sure that our efforts are in line with the values of social responsibility and inclusivity. By conducting such thorough analyses, we hope to further the scientific conversation about ArSL transcription while also fostering a more just and inclusive society for all people, irrespective of their communication requirements.

6 Conclusion

The proposed work presents the use of RNN-LSTM models enhanced with SelfMLP on a multi-modality video dataset for Arabic Sign Language recognition. A sign language user's everyday life is fraught with challenges when it comes to interacting with non-signers in a variety of settings, including shopping, healthcare, education, and transportation. Therefore, a system must be created to lessen the issues facing this community of the deaf and hard of hearing. This research has certain limitations, one of which is the small number of subjects and classes in the dataset used to classify signs. The suggested method will be used in the future to develop an Arabic Sign Language recognition system for a larger dataset that will include sign alphabets, numerals, words, and sentences from multiple signers, as well as different backgrounds, lighting conditions, and camera angles. This will enable us to identify sign videos in actual situations and train our model on bigger datasets. Even though our work has advanced this field significantly, there are still some issues with it. One limitation of our dataset is its relatively small number of subjects and classes. This underscores the need for future research endeavors to address this shortfall. In the future, we hope to create an ArSL recognition system that uses a wider range of datasets, including words, sentences, numerals, alphabets, and more from various signers in varying environments.

Our objective is to improve the efficacy and practicality of ArSL recognition systems through the expansion of our dataset and the training of our model on real-world scenarios. This will ultimately promote increased inclusivity and accessibility for the deaf and hard-of-hearing community.

References

1. W. Suliman, M. Deriche, H. Luqman and M. Mohandes, "Arabic Sign Language Recognition Using Deep Machine Learning," 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia, 2021, pp. 1-4, Doi: 10.1109/ISAECT53699.2021.9668405. keywords: {Gesture recognition; Machine learning; Assistive technologies; Feature extraction; Communications technology; Convolutional neural networks; Arabic Sign Language recognition; CNN; LSTM},

2. Saleh, Y., & Issa, G. F. (2020). Arabic Sign Language Recognition through Deep Neural Networks Fine-Tuning. *International Journal of Online and Biomedical Engineering (iJOE)*, *16*(05), pp. 71–83. https://doi.org/10.3991/ijoe.v16i05.13087

3. Reema Alzohairi, Raghad Alghonaim, Waad Alshehri and Shahad Aloqeely, “Image based Arabic Sign Language Recognition System” International Journal of Advanced Computer Science and Applications(ijacsa), 9(3), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090327

4. Duwairi, R. M., & Halloush, Z. A. (2022). Automatic recognition of Arabic alphabets sign language using deep learning. International Journal of Electrical and Computer Engineering (IJECE), 12(3), 2996–3004. DOI: 10.11591/ijece. v12i3.pp2996-3004.

5. Buttar, A.M.; Ahmad, U.; Gumaei, A.H.; Assiri, A.; Akbar, M.A.; Alkhamees, B.F. Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs. *Mathematics* **2023**, *11*, 3729. https://doi.org/10.3390/math11173729

6. Tharwat, G., Ahmed, A. M., & Bouallegue, B. (2021). Arabic Sign Language Recognition System for Alphabets Using Machine Learning Techniques. Department of Computers and Systems Engineering, Faculty of Engineering, Al-Azhar University, Cairo, Egypt; Department of Computer Engineering, King Khalid University, Abha, Saudi Arabia. Retrieved from https://downloads.hindawi.com/journals/jece/2021/2995851.pdf.

7. Almasre, M.A.; Al-Nuaim, H. Comparison of Four SVM Classifiers Used with Depth Sensors to Recognize Arabic Sign Language Words. *Computers* **2017**, *6*, 20. https://doi.org/10.3390/computers6020020

8. Ahli, Maitha Essa Mohammad, "Towards a Reliable Machine Learning-based Model Designed for Translating Sign Language Videos to Text" (2023). Thesis. Rochester Institute of Technology. Accessed from https://repository.rit.edu/theses/11490/