International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Feb-24)

https://dx.doi.org/10.12785/ijcds/150152

Real Time System Based Deep Learning for Recognizing Algerian Sign Language

Kheldoun Ahmed[1], **Kouar Imene**[1] **and Kouar El Bachir**[1]

[1]*Department of Mathematics and Computer Science, University of Medea, Medea 26000, Algeria*

Received 4 Sep. 2023, Revised 6 Jan. 2024, Accepted 21 Jan. 2024, Published 1 Feb. 2024

Abstract: Sign language plays a crucial role in facilitating communication and interaction for the deaf community. However, the recognition of sign language poses unique challenges, especially in the context of Algerian Sign Language (ALGSL), where limited research has been conducted. Using recent advances in the field of deep learning, we present a novel ALGSL recoginition system using hand cropping and hand landmarks from successive video frames. Also, we propose a new key frame selection method to find a sufficient number of successive frames for the recognition decision, in order to cope with a near real-time system, where tradeoff between accuracy and response time is crucial to avoid delayed sign recognition. Our system is based on Autoencoder architecture enhanced by attention mechanism. The Autoencoder architecture combines both convolutional neural networks (CNN) for capturing spatial information and long-short-term memory (LSTM) for capturing temporal information. The proposed architecture is evaluated on our new ALGSL dataset and achieved an accuracy of 98,99%. Additionally, we test our architecture on different publicly datasets and shows outstanding results. Finally, we test the recognition of ALGSL gestures of our system for videos captured through a webcam.

Keywords: Algerian Sign Language, Sign language recognition, Deep learning, Convolutional neural networks, Long-short-term memory, attention, Mediapipe

1. INTRODUCTION

Sign language plays a crucial role in facilitating communication and interaction for the deaf community [1]. However, the recognition of sign language poses unique challenges, especially in the context of Algerian Sign Language (ALGSL) [2][3], where limited research has been conducted. This gap in knowledge presents an opportunity to delve into uncharted territory and contribute to the development of effective ALGSL recognition systems.

Sign languages exhibit significant differences from spoken languages regarding their lexicons and linguistic grammars. Consequently, hearing individuals face substantial difficulties communicating through sign languages without proper training, creating a communication gap between the deaf and hearing communities. To address this issue, recent attention has focused on the potential solutions offered by technologies and applications of sign language recognition [4]–[6]. However, sign language recognition represents a challenging research field, incorporating diverse techniques such as sensor-based approach [7], [8] and vision-based approach [9]–[12]. Sensor-based approach usually use different specialized equipment. Vision-based approach is based only on standard cameras and rely on image-processing techniques to interpret gestures. The vision-based approach is more natural and easier to use than the sensor-based approach [13].

The motivation behind our research stems from two critical aspects. Firstly, we recognize the pressing need to address the challenges faced by the deaf community, whose language and communication needs often receive inadequate attention. By focusing our efforts on ALGSL recognition, we aim to bridge this gap and provide a technological solution that enhances the communication and accessibility of deaf individuals. Secondly, we identified the absence of an available dataset for ALGSL, which presented a significant hurdle in our research. To overcome this challenge, we took a proactive approach and visited several deaf schools across Algeria, immersing ourselves in the language and culture of the deaf community. This involved actively engaging with deaf individuals, sign language interpreters, and native signers to capture a wide range of ALGSL signs, gestures, and expressions. By actively involving the deaf people and gathering a diverse dataset, we ensured that our research reflects the linguistic and regional variations within ALGSL.

Our main objective is to develop a real-time system that utilizes deep learning techniques [14]–[16], specifically an Autoencoder architecture enhanced by attention mechanism,

E-mail address: ahmedkheldoun@yahoo.fr, kouar.imene@gmail.com, kouar.elbachir@gmail.com

to interpret and recognize ALGSL gestures accurately. Real-time recognition, with response times of less than 2 seconds, was a significant challenge. Balancing the trade-off between accuracy and response time, we aimed to achieve reliable and efficient ALGSL recognition in real-world scenarios. Additionally, we focused on ensuring the robustness and scalability of our recognition system. This included incorporating a wide range of words and gestures into our dataset, allowing for the recognition of any ALGSL signer, not just those who were part of the training process. By minimizing overfitting and optimizing the performance of our model, we aimed to create a system that is capable of accurately interpreting ALGSL gestures with high generalizability.

The main contributions in this work are as follows:

- Built a new ALGSL video-based sign dataset, which comprises 89 distinct signs that were repeated five times by 10 signers.
- Proposed a key frame selection method in order to find a sufficient number of successive frames for the recoginition decision.
- Proposed an Autoencoder architecture enhanced by attention mechanism, which selectively focuses on important features.
- Built a real-time system that can recognize ALGSL gestures for videos captured through a webcam.

Throughout our research, we collaborated closely with the deaf community, valuing their expertise and incorporating their feedback to ensure the accuracy and cultural authenticity of the collected data. By actively involving them in the research process, we strived to create a recognition system that aligns with their needs and preferences. By addressing the challenges of ALGSL data collection, incorporating deep learning techniques, and engaging with the deaf community, our research endeavors seek to contribute to the advancement of ALGSL recognition systems. Our ultimate goal is to make a tangible impact on the lives of the deaf community in Algeria, empowering individuals who rely on sign language as their primary mode of communication and ensuring their inclusion and participation in various aspects of social, educational, and professional interactions.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 gives un overview about algerian sign language. In section 4, we show the methodology for developing our system. Section 5, presents the new video-based sign dataset for ALGSL. In section 6, we present our key frame selection mechanism. Section 7 presents the proposed Autoencoder architecture. In section 8, we present the obtained experimental results. Finally, Section 9 concludes and gives some further research directions.

2. Related Work

There is a significant amount of research reported in sign language recognition. The authors in [17] have proposed an approach combining inception model (CNN) and RNN for recognising isolated hand gestures. They achieved an accuracy of 95.2% over the Argentinian Sign Language dataset (LSA64) [18]. However, this work is not suitable for real time recognition as they pre-process all the frames contained in a video. Also, for each frame they apply the background removal which significantly increases the processing time. In [19], the authors propose the Multiple extraction and Multiple prediction (MEMP) network which consists of three consecutive 3DCNN and convolutional LSTM layers for gesture recognition. The model achieved an accuracy of 99.06% for LSA64 dataset but using a complex model is time-consuming for prediction. In [20], the authors have proposed an approach that used three-stream CNN. In the first step, the input video was used to generate three types of motion templates. In the second step used pre-trained CNNs to extract features. Finally, they use fusion for classification. In [21], the authors have proposed a system which combines 3DCNN and ConvLSTM for dynamic gesture recognition. The system has been evaluated using several datasets. They achieved an accuracy of 98.5% over LSA64 dataset. In [22], a system based spatial-temporal graph convolutional network (ST-GCN) has been proposed for sign language recognition. Also, the system used the OpenPose library [23] to estimate the skeletons of individuals in videos. The system has been evaluated and achieved an accuracy of 61.04% over American sign language (ASLLVD-20)[24]. Recently, in [25], a graph-based 3DCNN (3DGCN) architecture is proposed for sign language recognition. Furthermore, the authors used the attention mechanism to enhance the spatial representation of the gesture. The system was evaluated in several datasets and obtained promising results. The proposed system achieved an accuracy of 94.84% (resp. 68.75%) over LSA64 (resp. ASLLVD-20) dataset.

3. Algerian Sign Language

The ALGSL is a gestural language used by deaf individuals in Algeria as a means of communication through signs. The Algerian Sign Language is officially recognized by the law of May 8, 2002, as the first language of the deaf community in Algeria, making it the only country in the Arab world and in Africa to officially recognize sign language [2].

The ALGSL can be traced back to its origin from French Sign Language (LSF). ALGSL is composed of a total of 42 signs for alphabet [3], which include 35 static and 6 dynamic signs. It is noteworthy that these signs are produced using a single hand gesture (refer to Figure 1). Moreover, each static sign is identified by two distinct features: configuration and orientation.

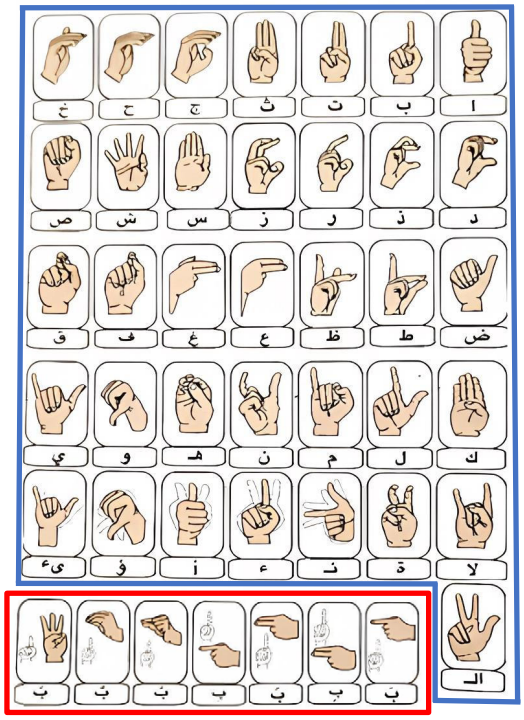


Figure 1. Algerian Sign Language Alphabet (the ones with red border are the dynamic signs)

As an example, the sign for the letter "ب" in Algerian Sign Language is defined by two parameters [3] :

- Configuration: making a fist with the index finger extended.
- Orientation: the palm is facing upwards (or with the wrist in a downwards position).

4. Methodology

In proposals to solve the problem of sign language recognition, we build a system that can interpret sign language in real-time. In Figure 2, we present the different phases for developing our system. The first phase consists of the creation of new dataset for Algerian sign language. This dataset undergoes careful optimization to ensure its suitability for real-time processing. The second phase is data preprocessing. It involves preparing and cleaning the raw data to make it suitable for a machine or deep learning model. In this phase, we propose a new key frame extraction method to find a sufficient number of successive frames and we use Mediapipe library [26], which is the most efficient human landmarks estimator to detect and extract hand landmarks. The third phase consists, on one hand, in data augmentation using rotation and horizontal flip. On the other hand, in data spliting to generate data training and validation. Finally, the last step is to train and evaluate the proposed deep learning model. The trained model is used for sign language recognition.

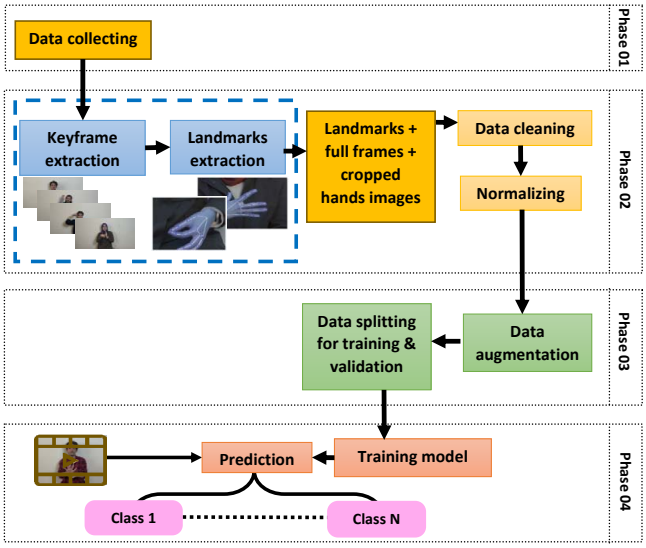


Figure 2. General diagram of the proposed approach

5. Algerian Sign Language Dataset (ALGSL89)

The development of an Algerian Sign Language dataset, aimed at facilitating the creation of a dictionary and training an automatic sign recognition system. This dataset contains 4885 videos showcasing ten non-expert subjects and one deaf individual each perform five repetitions of 89 distinct sign types. The selected signs, comprising both verbs and nouns, represent the most commonly used elements of the ALGSL lexicon and are classified into nine thematic categories: time, colors, places, justice-related terms, medical terminology, months, interrogatives, family, general communication, and specific verbs. Figure 3 provides several examples of these signs.

Figure 3. Snapshots of six key-frames random distinct signs extracted from the ALGSL89 dataset

Table I provides a detailed list of the signs used in our ALGSL89 dataset. It showing the ID, name, and hand(s)

used in each sign of the dataset. The "H" column specifies whether the sign was performed with the right hand ("R"), left hand ("L"), or both hands ("B").

TABLE I. List of the signs used in ALGSL89 dataset, along with the ID and hand(s) used for each sign.

ID	Name	H	ID	Name	H	ID	Name	H	ID	Name	H
00	Semaine	B	24	Aéroport	B	48	Juillet	R	72	Prénom	B
01	Dimanche	R	25	Banque	B	49	Aout	R	73	Nom de famille	B
02	Lundi	R	26	Restaurant	B	50	Septembre	B	74	Oui	R
03	Mardi	R	27	Hôtel	R	51	Octobre	B	75	Non	R
04	Mercredi	R	28	Rue	B	52	Novembre	B	76	Bonjour	R
05	Jeudi	B	29	École	B	53	Décembre	B	77	Merci	R
06	Vendredi	R	30	Université	R	54	Mère	R	78	Derien	R
07	Samedi	R	31	Loi	B	55	Père	R	79	Question	R
08	Année	B	32	Avocat	B	56	Frère	R	80	Message	B
09	Heur	B	33	Juge	B	57	Sœur	R	81	Toilette	R
10	Minute	B	34	Liberté	B	58	Fille	R	82	Quand	B
11	Couleur	R	35	Témoin	B	59	Garçon	R	83	Comment	B
12	Blanc	R	36	Héritage	B	60	Entrer	B	84	Ou	R
13	Noir	R	37	Malade	B	61	Sortir	B	85	Combien	B
14	Rouge	R	38	Vaccination	R	62	Aimer	R	86	Hier	R
15	Bleu	R	39	Médecin	R	63	Acheter	B	87	Demain	R
16	Jaune	R	40	Médicament	R	64	Manger	R	88	Rendez-vous	B
17	Vert	R	41	Premier secours	B	65	Chercher	B			
18	Rose	R	42	Janvier	R	66	Demande	B			
19	Hôpital	B	43	Février	R	67	Écrire	B			
20	Police	B	44	Mars	B	68	Appeler	B			
21	Tribunal	B	45	Avril	B	69	Perdre	B			
22	Mosquée	B	45	Mai	R	70	Trouver	R			
23	Pharmacie	R	47	Juin	R	71	Traduire	B			

The ALGSL89 dataset was captured in eleven different sessions, primarily within the indoor setting of our university classrooms. Each session focused on a specific set of signs, with most signs being captured in a single session. However, some signs required additional sessions for refinement and correction. More repetitions than required were captured during each session to account for potential errors. We utilized indoor lighting for most of the recordings and supplemented it with an artificial white projector to ensure consistent lighting across all sessions, regardless of the different recording times. All recordings were captured using a OnePlus 8 smartphone, which records videos at 4K 30fps with a 16:9 ratio. A tripod was set 1.6 meters away from the wall and at a height of 1.4 meters to maintain consistency across all recordings (more details illustrated can be seen in Figure 4).

Due to unforeseen circumstances, we had to adjust the recording location for deaf Subject, who was unable to participate in the same session as the others. We arranged a session at the deaf school in Algeria, where the individual was located. Despite this change, we strove to maintain the same recording conditions to preserve the dataset's quality and integrity.

The majority of the subjects wore dark-colored clothing and performed the signs against a white wall background to maintain consistency across all videos. To enhance the database's diversity and realism, we imposed minimal constraints on the subjects while performing the signs, ensuring a natural execution and a more realistic representation of real-world signing conditions. We believe that incorporating non-expert subjects added a greater level of diversity to the dataset.

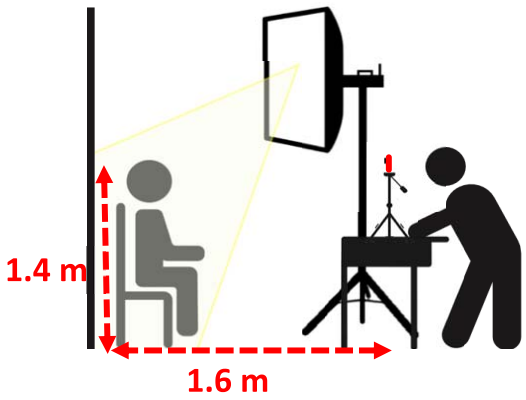


Figure 4. The configuration of the recording setup

6. Key Frame Selection Mechanism

After successfully creating the ALGSL89 dataset, the next crucial step in our methodology is key frame extraction mechanism, which is an important requirement for our sign language recognition system. Frame extraction consists in the conversion of continuous video motion into a series of separate frames, primed for individual processing and analysis. It's essential to recognize that key frame extraction method can considerably impact the overall effectiveness of a deep learning model trained on video data. In fact, key frame extraction is very important in sign video processing and analysis since it greatly reduces computing time. Therefore, the delay in sign recognition, which is the necessary time taken by the system in order to provide the output text from an input video, may be well-reduced. In [27], we have proposed five various key frame extraction methods. For each method, we have presented its strengths and weaknesses. Thorough an extensive evaluations, we have compared the different proposed key frame extraction in terms of duration of video processing, number of frames processed, selected frames, and the type of output. In the following, we present only the best one. In this method, for each frame, a laplacian variance was used as a focus measure operator in order to consider only non blurred image. Then, the frame preprocessing combines both hand cropping and frame resizing, as well as extracting hand landmarks using MediaPipe Hands (21 keypoints for each hand). The process involves reading each frame of the video, converting it to RGB, and processing it with MediaPipe Hands. If hand landmarks are detected, they are used to create a bounding box around the hand, which is then scaled up and cropped. If two hands are detected, the cropped hand images are resized, stacked horizontally, and normalized. If only one hand is detected, the cropped hand image is resized and normalized. For videos with no detected hand landmarks, a zero-filled array is added. The

detailed steps of this method are presented in Figure 5.

7. Autoencoder model

In Figure 6, we present our proposed architecture that combines both convolutional neural networks (CNNs) [15][28] and long-short-term memory (LSTM) networks [16] [29] along with an attention mechanism [30]. This forms an Autoencoder model for the task of sign language recognition. The Autoencoder is a type of neural network used for learning efficient coding of input data. It's composed of an encoder, a bottleneck, and a decoder.

- **Encoder** : The encoder part of our model consists of two branches. The first branch processes the video frames input using a series of 2D convolutional layers, max-pooling layers, and dense layers. It applies a 2D convolutional operation in a time-distributed manner to capture spatial information. The second branch processes the hand landmarks input through a fully connected layer. Both branches extract relevant features from their respective inputs.
- **Bottleneck** : The encoded features from the two branches are concatenated and passed through an attention mechanism, which selectively focuses on important parts of the combined feature representation.
- **Decoder** : The decoder part of our model utilizes an LSTM (Long Short-Term Memory) network to decode the encoded features. The LSTM layer takes the attention-weighted features as input and processes them in a recurrent manner, capturing sequential dependencies. The output of the LSTM layer is then fed into a series of dense layers with time-distributed connections. The final output is a sequence of probabilities representing the classification of sign language gestures.

8. Experimental results and Discussion

The proposed system in this work was implemented using Python. We use some deep learning libraries such as Tensorflow and OpenCV. The evaluation results was conducted on Intel i5-3570, 4.00 GHz and has 16.0 GB of RAM. Firstly, an extensive tests was conducted to fine-tune some parameters such as the number of extracted frames and the crop scale. After, the evaluation of the proposed system is conducted on the ALGSL89 dataset and the other publicly datasets.

A. Parameter tuning

1) Number of extracted frames

The number of frames selected is determined based on a technique that utilizes a variable called the frame step. The frame step is calculated by dividing the Total number of frames per video (TNF) by the number of frames to be selected ($NF2Sel$) (Equation (1)). This method ensures that the selected frames are distinct from each other and effectively cover the entire video.

$$Frame\ Step = TNF \ / \ NF2Sel \tag{1}$$

By using this frame extraction technique, we can guarantee that the chosen frames capture diverse moments throughout the video sequence, providing a comprehensive representation of the gestures performed. This leading to improved recognition accuracy and robustness in our system. Table II shows the results obtained according on the variation in the number of frames to be extracted. It is shown that the tradeoff between accuracy and time of preprocess a single video is better when the number of extracted frames equal to 8.

2) Crop scale

Now, we turn our attention to the second important parameter, namely crop scale. This parameter determines the size of the hand image that is extracted during the preprocessing stage. The choice of crop scale has a significant impact on the accuracy of our model. By adjusting the zoom level, we can control the level of detail captured in the hand images, thereby influencing the model's ability to accurately recognize gestures. Careful tuning of this parameter is crucial to strike the right balance between capturing sufficient hand details and avoiding excessive noise or loss of important information. Table III provides an overview of the performance metrics for each crop scale value. We can observe that a higher crop scale value of 1.5 yields a slightly higher accuracy compared to a crop scale value of 2. This suggests that a larger zoom level contributes to improved gesture recognition performance by capturing more intricate hand details.

B. Evaluation of the Autoencoder

We first examine the training and validation accuracy on ALGSL89 dataset. Figure 7(a) illustrates the training and validation accuracy curves, which demonstrate the model's ability to learn and generalize from the training data while maintaining accuracy on unseen validation data. The Autoencoder achieves a good accuracy which is approximately 100%. Next, we explore the training and validation loss in Figure 7(b). The loss curves provide valuable insights into the convergence and stability of the model during the training process. Lower loss values indicate better optimization and improved generalization capabilities of the model.

Then, we evaluate our Autoencoder on the publicly dataset LSA64 [18]. LSA64 is a dataset specifically curated for Argentinian Sign Language (LSA), comprising $3,200$ videos of 10 non-expert subjects performing 64 different sign types. The dataset covers a wide range of commonly

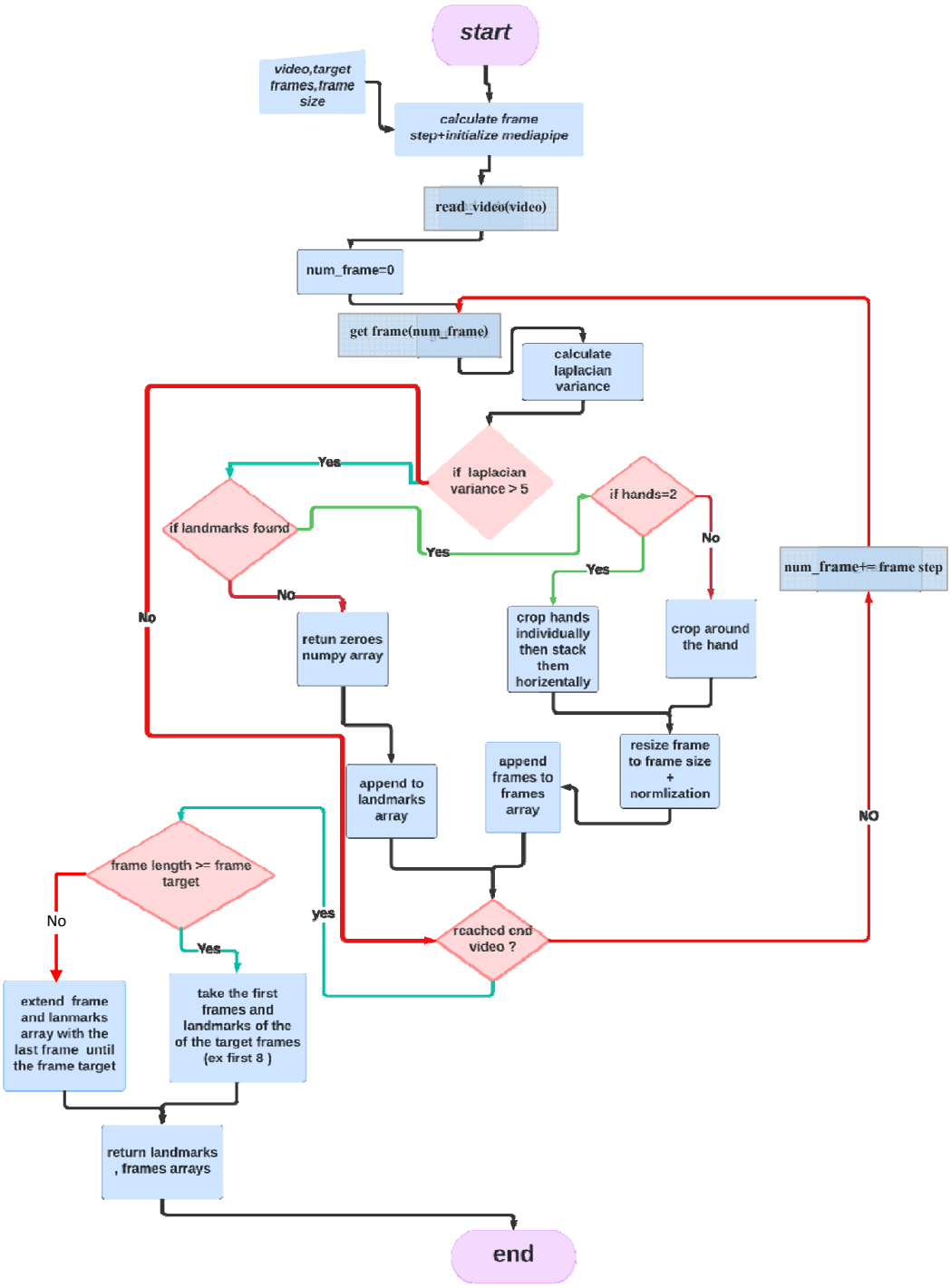


Figure 5. Flowchart of Key frame extraction

used signs in LSA, including verbs and nouns. The recordings were conducted in both outdoor and indoor environments, providing variation in lighting conditions. Subjects wore fluorescent gloves to facilitate hand segmentation in the images. LSA64 serves as a valuable resource for research and development in automatic sign recognition and understanding, contributing to advancements in sign language technology.

Table IV presents a comparaison of various classifiers and their corresponding features and validation accuracy. While our Autoencoder model achieved a good accuracy

TABLE II. Comparison results according to the number of extracted frames.

Number of extracted frames	Frame step	Time of preprocess single video	Validation Acc. (%)	Loss	Selected
4 Frames	7	1.68 sec	50.92 %	0.0031	
8 Frames	3	3.7 sec	98.99 %	0.00025	√
16 Frames	1	8.99 sec	99.30 %	0.00021	

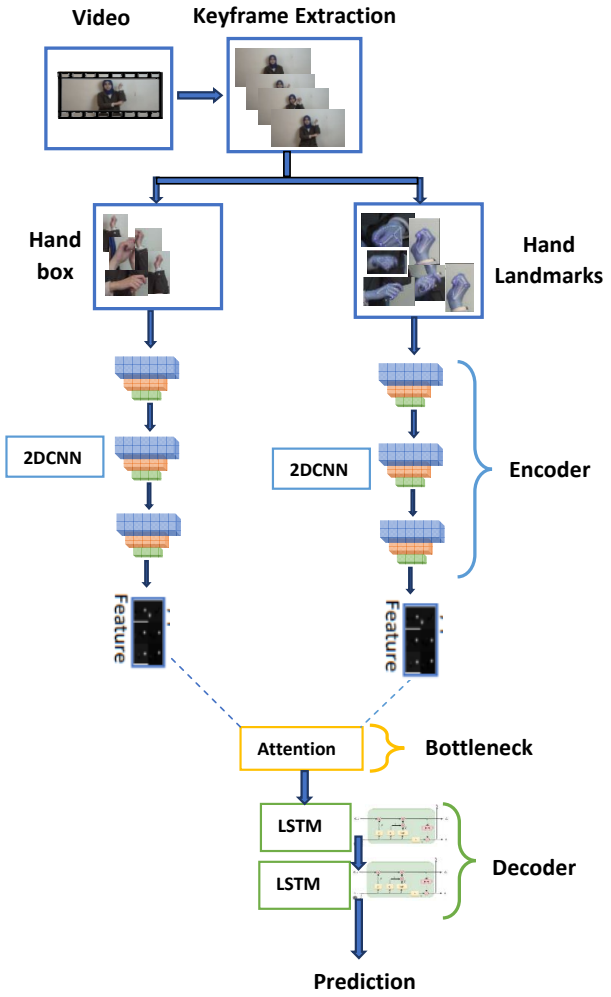
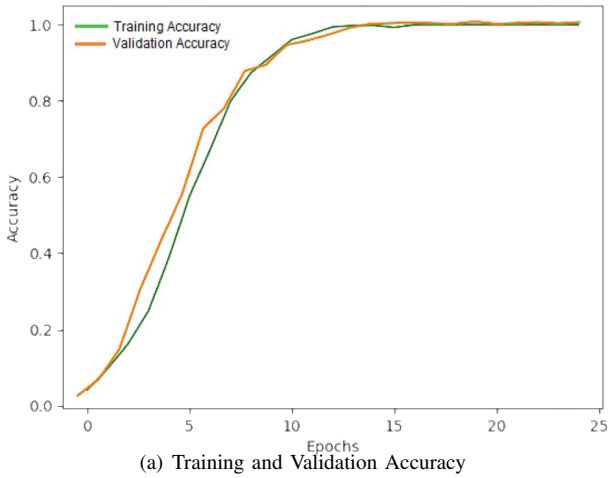


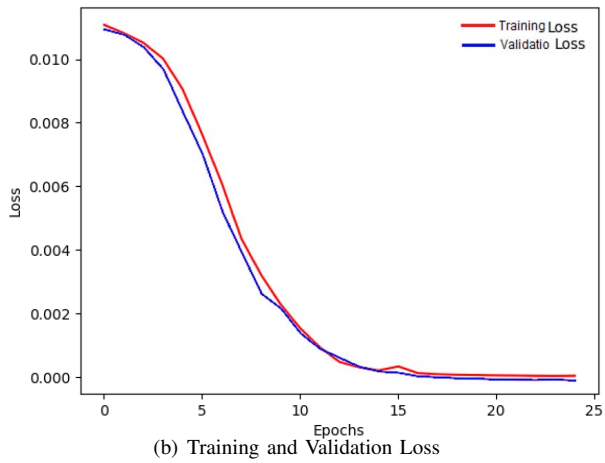
Figure 6. Architecture of the proposed Autoencoder model

TABLE III. Comparison results according to the crop scale.

Crop scale	Validation Acc. (%)	Loss	Selected
Crop scale = 1.5	98.99 %	0.00025	√
Crop scale = 2.0	97.61 %	0.00034	



(a) Training and Validation Accuracy



(b) Training and Validation Loss

Figure 7. Performance of the Autoencoder on ALGSL89 dataset.

(98.67%), it is important to note that direct comparison with other models may not be straightforward. Other classifiers in the Table IV might have achieved higher accuracies, but they could have faced challenges such as overfitting or required longer processing times to provide results.

After that, we evaluate our Autoencoder on the publicly dataset ASLLVD-20 which is a partial dataset of the entire dataset ASLLVD [24]. ASLLVD-20 comprising 20 selected sign gestures from American Sign Language (ASL). The dataset focuses on capturing skeletal data of sign language movements and provides a valuable resource for research and development in ASL recognition and understanding. Each sign gesture is represented by a small video segment, allowing for targeted analysis and exploration of ASL

TABLE IV. Performance Comparison of LSA64 dataset

Classifier	Feature	Validation Acc. (%)
CNN + RNN [97]	Hand segmentation	95.2%
MEMP (3DCNN + ConvLSTM) [99]	RGB frame	**99.06%**
Three-stream CNN [100]	Skeletal + Motion Template Fusion	97.81 %
3DCNN + ConvLSTM [101]	RGB frame	98.5 %
3DGCN [105]	Graphe of hand landmarks	94.84 %
Autoencoder (Ours)	RGB Cropped hand + hand landmarks	98.67 %

TABLE V. Performance Comparison of ASLLVD-20 dataset

Classifier	Feature	Validation Acc. (%)
ST-GCN[104]	skeleton + keypoints	61.04%
Basic 3DGCN [105]	skeleton + hand landmarks	62.5%
Enhanced 3DGCN [105]	skeleton + hand landmarks	**68.75** %
Autoencoder (Ours)	RGB Cropped hand + hand landmarks	66.36 %

gestures within a concise subset.

Table V presents a concise comparison of classifiers and their associated features. Notably, the Autoencoder-based approach achieves encouraging performance. It outperformes the ST-GCN[104] and Basic 3DGCN [105] architectures.

The performance of a classifier involves a trade-off between accuracy and other factors such as generalization, robustness, and real-time responsiveness. Our Autoencoder model strikes a balance in these aspects, offering a reasonable accuracy while being robust and responsive. It allows for efficient communication and interpretation of sign language gestures in real-world scenarios.

C. Developed System

Figure 8 shows the graphical interface of the proposed system. In this interface, users can use webcam to start recording gestures, and the results are displayed in both written and sound formats. Furthermore, users can upload pre-recorded videos, which can be analysed by the developed system and predicted the coresponding gestures.

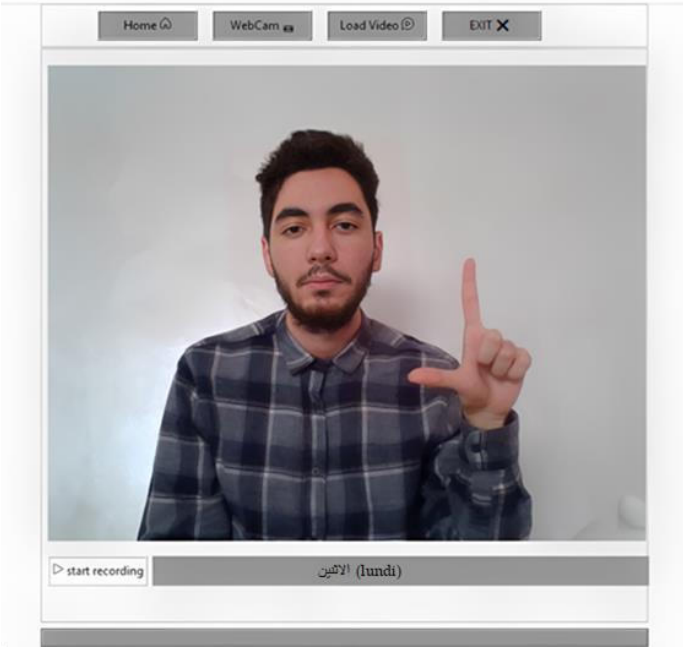


Figure 8. Graphical interface of the developed system

9. Conclusion and Future Works

This work has made significant progress in addressing the challenges of recognizing ALGSL and has contributed to the development of effective ALGSL recognition systems. We have achieved notable milestones and gained valuable insights. The major contributions of this work is, on one hand, the proposed of a new ALGSL video-based sign dataset which comprises 89 distinct signs. On the other hand, the successful development of a real-time application that accurately interprets ALGSL gestures. We specifically focused on frame extraction from videos, allowing us to extract significant frames, ensuring real-time processing and response. By employing deep learning techniques, particularly an Autoencoder architecture enhanced by attention mechanism, we achieved a balance between accuracy and response time, resulting in a recognition system with high reliability and efficient performance. Also, the proposed architecture is evaluated on two publicly datasets and compared to state-of-the-art architectures. Our Autoencoder shows outstanding performance and generalizes well on various datasets. There are several perspectives to consider for further research:

- To further improve the accuracy and generalizability of ALGSL recognition systems, expanding the dataset

is crucial. Collecting additional diverse samples, including a wider range of ALGSL signs, gestures, and expressions, would enhance the system's performance and adaptability.

- To provide a more comprehensive understanding of sign language communication, incorporating facial expression and emotion recognition as additional features would enrich the interpretation of ALGSL gestures. This can enhance the system's ability to capture nuanced expressions and improve overall accuracy.
- Building upon our work with isolated gestures, exploring continuous sign language recognition is essential. Developing models that can interpret complete sign language sentences and conversations would enable more natural and fluid communication for the deaf community, extending the system's usability and impact.

References

[1] R. Ruben, "Sign language: Its history and contribution to the understanding of the biological nature of language," *Acta otolaryngologica*, vol. 125, pp. 464–467, 2005.

[2] S. Lanesman, "Algerian jewish sign language: its emergence and survival," 2016.

[3] F. Nekkaa, "Détection automatique de la main : Application à la reconnaissance de la langue des signes arabe." Master thesis, Systèmes Distribués et Méthodes Formelles (SDMF), Université Abdelhamid Mehri-Constantine 2, 2014/2015.

[4] A. Eman, A. Reem, A. Aseel, A. Bushra, A. Hajer, A. Nahla, A. Areej, and A. Abdulwahab, "Arabic sign language recognition using convolutional neural network and mobilenet," *Arabian Journal for Science and Engineering*, vol. 48, pp. 2191–4281, 2023.

[5] Q. Zhu, J. Li, F. Yuan, and Q. Gan, "Continuous sign language recognition via temporal super-resolution network," *Arabian Journal for Science and Engineering*, 2023.

[6] Y. Obi, K. S. Claudio, V. M. Budiman, S. Achmad, and A. Kurniawan, "Sign language recognition system for communicating to people with disabilities," *Procedia Computer Science*, vol. 216, pp. 13–20, 2022, 7th International Conference on Computer Science and Computational Intelligence 2022.

[7] M. S. Amin, M. T. Amin, M. Y. Latif, A. A. Jathol, N. Ahmed, and M. I. N. Tarar, "Alphabetical gesture recognition of american sign language using e-voice smart glove," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 2020, pp. 1–6.

[8] P. Lokhande, R. Prajapati, and S. Pansare, "Article: Data gloves for sign language recognition system," *IJCA Proceedings on National Conference on Emerging Trends in Advanced Communication Technologies*, vol. NCETACT 2015, no. 1, pp. 11–14, June 2015.

[9] M. Balaha, S. El-Kady, H. Balaha, M. Salama, E. Emad, M. Hassan, and M. Saafan, "A vision-based deep learning approach for independent-users arabic sign language interpretation," *Multimedia Tools and Applications*, vol. 82, pp. 1–20, 08 2022.

[10] W. Li, H. Pu, and R. Wang, "Sign language recognition based on computer vision," in *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2021, pp. 919–922.

[11] K. Bantupalli and Y. Xie, "American sign language recognition using deep learning and computer vision," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4896–4899.

[12] H. Gupta, A. Ramjiwal, and J. Jose, "Vision based approach to sign language recognition," *International Journal of Advances in Applied Sciences*, vol. 7, p. 156, 06 2018.

[13] S. Fakhfakh and Y. Ben Jemaa, "Gesture recognition system for isolated word sign language based on key-point trajectory matrix," vol. 22, 2019.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[15] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Q. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, 2021.

[16] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, 2021, 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.

[17] S. Masood, A. Srivastava, H. C. Thuwal, and M. Ahmad, "Real-time sign language gesture (word) recognition from video sequences using cnn and rnn," in *Intelligent Engineering Informatics*, V. Bhateja, C. A. Coello Coello, S. C. Satapathy, and P. K. Pattnaik, Eds. Springer Singapore, 2018, pp. 623–632.

[18] F. Ronchetti, F. M. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete, "Lsa64: An argentinian sign language dataset," 2016.

[19] X. Zhang and X. Li, "Dynamic gesture recognition based on MEMP network," *Future Internet*, vol. 11, no. 4, p. 91, 2019. [Online]. Available: https://doi.org/10.3390/fi11040091

[20] J. Imran and B. Raman, "Deep motion templates and extreme learning machine for sign language recognition," *The Visual Computer*, vol. 36, p. 1233–1246, 2020.

[21] D. Fathy and E. Elsayed, "Semantic deep learning to translate dynamic sign language," *International Journal of Intelligent Engineering and Systems*, vol. 14, p. 2021, 2021.

[22] C. C. de Amorim, D. Macêdo, and C. Zanchettin, "Spatial-temporal graph convolutional networks for sign language recognition," in *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*. Springer International Publishing, 2019, pp. 646–657.

[23] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 2018.

[24] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the American Sign Language lexicon video dataset (ASLLVD) corpus," in *Proceedings of the LREC2012 5th Workshop on the*

Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, O. Crasborn, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch, Eds. European Language Resources Association (ELRA), 2012, pp. 143–150.

[25] M. Al-Hammadi, M. Bencherif, M. Alsulaiman, G. Muhammad, M. Mekhtiche, W. Abdul, Y. Alohali, T. Alrayes, H. Mathkour, M. Faisal, M. Algabri, H. Altaheri, T. Alfaqih, and H. Ghaleb, "Spatial attention-based 3d graph convolutional neural network for sign language recognition," *Sensors*, vol. 22, p. 4558, 2022.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[27] K. Imene and K. el Bachir, "Deep-rsl: Deep learning for sign language recognition from a video sequence." Master thesis, Technologies and Web Systems Engineering, university of Yahia Fares, Medea, 2022/2023.

[28] J. Chang and S. Jin, "An efficient implementation of 2d convolution in cnn," *IEICE Electronics Express*, vol. 14, pp. 4299–4308, 2017.

[29] R. Siriak, I. Skarga-Bandurova, and Y. Boltov, "Deep convolutional network with long short-term memory layers for dynamic gesture recognition," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1, 2019, pp. 158–162.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, 2017.

Ahmed Kheldoun received the Ph.D. degree in computer science from the University of USTHB, Algeria, in 2018. He is currently an Associate Professor of computer sciences with the University of Medea, Algeria. His current research interests include Petri net theory and applications, software engineering, Business Processes, Distributed and reconfigurable systems, Web services, Artificial intelligence and Machine learning. Kheldoun received Engineer and Magister diplomat in computer science from high school of computer science (ESI), Algeria, in 2005 and 2008, respectively.

Imene Kouar received master's degree in Computer science from the university of Medea, Algeria in 2023, her current research interests include image processing, computer vision, machine learning, deep learning and web development.

El Bachir Kouar received master's degree in Computer science from the university of Medea, Algeria in 2023, his current research interests include image processing, computer vision, deep learning, web development and game design.