Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 1, 2025, 95-105.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A Persian Continuous Sign Language Dataset

Razieh Rastgoo*

Electrical and Computer Engineering Faculty, Semnan University, Semnan, Iran

Article Info

Article History:
Received 05 September 2024
Revised 11 November 2024
Accepted 12 January 2025

DOI:10.22044/jadm.2025.14973.2600

Keywords:
Sign Language Recognition (SLR), Sign Language Production (SLP), Dataset, Generative Adversarial Network (GAN), Persian.

**Corresponding author: rrastgoo@semnan.ac.ir (R. Rastgoo).*

Abstract

Sign language (SL) is the primary mode of communication within the deaf community. Recent advances in deep learning have led to the development of various applications and technologies aimed at facilitating bidirectional communication between the Deaf and hearing communities. However, challenges remain in the availability of suitable datasets for deep learning-based models. Only a few public large-scale annotated datasets are available for sign sentences, and none exist for Persian Sign Language sentences. To address this gap, we have collected a large-scale dataset comprising 10,000 sign sentence videos corresponding to 100 Persian sign sentences. This dataset includes comprehensive annotations such as the bounding box of the detected hand, class labels, hand pose parameters, and heatmaps. A notable feature of the proposed dataset is that it contains isolated signs corresponding to the sign sentences within the dataset. To analyze the complexity of the proposed dataset, we present extensive experiments and discuss the results. More concretely, the results of the models in key sub-domains relevant to Sign Language Recognition (SLR), including hand detection, pose estimation, real-time tracking, and gesture recognition, have been included and analyzed. Moreover, the results of seven deep learning-based models on the proposed datasets have been discussed. Finally, the results of Sign Language Production (SLP) using deep generative models have been presented. We report the experimental results of these models from these sub-areas, showcasing their performance on the proposed dataset.

1. Introduction

By 2050, approximately 2.5 billion people will experience some degree of hearing loss, with at least 700 million requiring hearing rehabilitation. In light of this, it is crucial to study Sign Languages (SL), the primary mode of communication among Deaf individuals [1]. Like other natural languages, SLs adhere to linguistic rules [2]. However, SLs lack standardized written forms. Additionally, most recent communication technologies are designed for spoken or written languages, not SLs. Consequently, many members of the hearing community are unfamiliar with SLs, leading to significant communication barriers [3-14]. Sign Languages, like natural languages, operate on three levels: letter, word, and sentence. Sign letters are static signs consisting solely of poses without any movement of body parts. Sign words and sign sentences, on the other hand, involve dynamic poses combined with body movements. A sign word represents a single word, whereas a sign sentence comprises a sequence of words forming a sentence. Despite both having sequential characteristics, sign sentences are inherently more complex. Sign words and sign sentences are also referred to as isolated and continuous signs, respectively [6]. To address the communication barriers faced by the Deaf community and develop practical communication applications, it is essential to consider all three levels of SL.

However, this paper focuses specifically on sign sentences (continuous signs).

Considering recent advances in deep learning, different applications and technologies have been developed to facilitate the bidirectional communication between the Deaf and the hearing communities [7]. However, the challenges regarding the deep learning-based models, such as the necessity of large amount of data as well as the hardware equipment for model training are still remaining. Regarding the data challenge, there are some large-scale datasets for the sign letter and sign word levels [7]. However, only few public large-scale annotated datasets suitable for deep learning-based models are available for sign sentence [15-18]. Since there is not any dataset for Persian Sign sentences, we collected a large-scale dataset, including 10'000 sign sentences videos corresponding to 100 Persian sign sentences. The main contributions of this work can be listed as follows:

- A **dataset**, including 10'000 videos of 100 Persian sign sentences using 10 Deaf and hearing people has been collected. One important characteristic of the proposed dataset is that it contains isolated signs corresponding to the sign sentences within the dataset.
- The proposed dataset contains a set of **annotations**, including the bounding box of the detected hand, the class label, the hand pose parameters, and the heatmap.
- The extensive **experiments** have been presented on the proposed dataset and the results have been discussed. Also, the results of the sign video generation using the Generative Adversarial Network (GAN) are discussed.

The remainder of this paper is as follows: recent advances in SL as well as the sign sentence datasets are reviewed in section 2 and 3, respectively. Description of the proposed dataset as well as the statics of this dataset is presented in section 4. Results of the model are analyzed and evaluated in section 5. Finally, we discuss the advantages as well as the limitation and conclude the work with the insight to the future work in section 6 and 7, respectively.

2. Sign language background

Sign language processing can be divided into two main tasks: Sign Language Recognition (SLR) and Sign Language Production (SLP). SLR focuses on translating sign language into spoken or written language, while SLP performs the reverse task, generating sign language from spoken or written input. Together, these processes aim to create a bidirectional communication system that can be used in real-world applications. However, both SLR and SLP face significant challenges. One issue is that most people are unfamiliar with sign language, and there is no universally accepted standard for it. Sign languages differ across regions, and signs can be highly complex due to variations in hand shape, orientation, movement, location, and non-manual signals like facial expressions. These complexities result in large variations within the same sign class (intra-class variability) and subtle differences between different sign classes (inter-class variability), making it difficult to create systems that can robustly recognize a wide range of signs. In addition, SLP systems must generate photorealistic sign sequences that accurately convey the meaning of spoken or written input, which is challenging because of the grammatical and structural differences between spoken and sign languages. The translation between these two types of languages involves more than just a word-for-word mapping; the order of signs and tokens can differ significantly from that of spoken language, and this must be accounted for in any system designed for sign language production. Another obstacle in this field is the misconception that deaf individuals are comfortable reading spoken language, leading to an underemphasis on translating written or spoken language into sign language. In reality, many deaf individuals may not be proficient in reading or writing the spoken language of their country, as the written forms of most spoken languages differ greatly from sign languages. Furthermore, many sign languages lack a formal written system, adding another layer of complexity to this translation task. Applications of SLR, in particular, are widespread in areas such as robotics, human-computer interaction, education, and virtual reality. However, the focus tends to be on recognizing sign language rather than generating it, partly because of the assumption that deaf individuals can rely on written communication. Developing robust SLP systems that can generate accurate and realistic signs from text or voice inputs in diverse real-world scenarios remains a crucial and underexplored area. Anyway, both of these tasks, SLR and SLP, need to be carefully studied in order to make a bi-directional translation system for Deaf community. In-line with our previous works in SLR/SLP [2, 3, 5-9, 11-14, 19-22], in this paper, we present a Persian Sign Sentence dataset for providing a test-bed in continuous SLR/SLP.

3. Current sign sentence datasets

Providing a large-scale annotated dataset is an important step in accelerating progress in the field of sign language processing, including both of SLR and SLP tasks [6]. Most of the existing datasets have been developed for the sign letter and sign words [23-30]. These datasets are not enough for real-world applications of sign language processing that involve natural conversational with complete sentences. However, only few public large-scale annotated datasets are available for processing sign sentences using deep learning-based models [31-38] (See Table 1). As this table shows, these datasets include only four types of sign languages, Chinese [31], German [32-34], American [35-37], and British [39]. More specifically, SIGNUM [34] and the BSL Corpus [38] include the RGB videos recorded in controlled environment. Furthermore, the RWTH-Phoenix-2014 [17] and the extended version of it, RWTH-Phoenix-2014T [15], are widely used in Neural Machine Translation (NMT) [33] and production works [41-42]. These datasets contain the data of the weather forecast from a TV broadcast performed by 9 signers. In addition, the Public DGS Corpus [43] and the Video-Based CSL (Chinese Sign Language) [31] datasets contain the RGB videos along with the corresponding 2D and 3D poses estimated using the OpenPose [44]. The RWTH-BOSTON-104 [17] and the NCSLGR [35] datasets consist of the videos from the American Sign Language (ASL). In addition, the recently released dataset, How2Sign [38], is a multimodal and multi-view continuous ASL dataset, containing a parallel corpus of more than 80 hours of sign language videos and a set of corresponding modalities including speech, English transcripts, and depth. With the aim of increasing the variety of the datasets for sign language processing and also facilitating the communications of Persian Deaf people in the word, we propose a Persian dataset, including 10’000 annotated videos of 100 Persian sign sentences. Currently, our dataset contains only one modality, the RGB videos.

Table 1. Current datasets in continuous sign language.

Name	Language	#Signer	Include isolated signs	#vocab/token	Gloss	Pose
Video-Based CSL [30]	CSL	50		178	N	Y
SIGNUM [33]	DGS	25	×	450	Y	N
RWTH-Phoenix-2014T [31]	DGS	9	×	3k	Y	N
Public DGS Corpus [41]	DGS	327	×	-	Y	Y
BSL Corpus [37]	BSL	249	×	5k	Y	N
RWTH-BOSTON-104 [35]	ASL	3	×	104	Y	N
NCSLGR [34]	ASL	4	×	1.8k	Y	N
How2Sign [36]	ASL	11	×	16k	Y	Y
ASLLRP [11]	ASL	4	×	16.6k	Y	N
PSLS (Ours)	PSL	10	√	490	N	Y

Figure 1. Some samples of the proposed continuous signs: (a) I am happy with my family, (b) I like coffee, (c) I am a student, (d) I watch news every day.

4. Proposed dataset

In this section, we present the details of the proposed dataset, including an overview of the

proposed dataset, dataset statistics, and the proposed demo. Some samples of the continuous signs have been shown in Figure 1. Based on the signer requests, the identification of these signers have been hidden in this figure.

4.1. Dataset set-up

To record the videos in the proposed dataset, a camera with medium resolution (1920x1080 / Full HD) has been employed. Aiming to use the recorded videos in deep learning models, the frames of these videos have been resized into 224x224x3 using the Python libraries [43].

4.2. Dataset overview

The proposed dataset consists of 10,000 RGB videos featuring 100 Persian sign language sentences, each performed by 10 signers. The dataset includes a balanced mix of 5 female (three Deaf and two pre-trained hearing) and 5 male participants (three Deaf and two pre-trained hearing). Each sentence is repeated 10 times, resulting in a total of 10,000 RGB video samples. To enhance the dataset's realism and applicability to real-world scenarios, various backgrounds were incorporated during data collection. The video resolution is set to 224×224 pixels with three color channels (RGB). The RGB frames were manually annotated using the LabelImg software [44], including bounding boxes for detected hands, class labels, hand pose parameters, and heatmaps. This dataset is designed to serve as a benchmark for developing sign language recognition systems and facilitating communication for the hearing-impaired. The dataset will be made publicly available in the near future.

4.3. Dataset statistics

In this section, we present a statistical overview of the dataset, highlighting key aspects that enhance its applicability for sign language recognition tasks:

Sign Sentences: Based on an analysis of the most frequently used signs in daily communication, we created 100 sentences using the top 100 most common signs.

Samples per Class: To meet the requirements of deep learning models, each sign sentence is represented by 100 sample recordings. This ensures a sufficient volume of data for the models to effectively learn high-level features from the inputs.

Background Complexity: To simulate realistic recognition environments, the dataset includes a variety of backgrounds with differing levels of complexity, enhancing the model's robustness in real-world applications.

Subject Diversity: To ensure the dataset is person-independent and suitable for generalized recognition, it includes subjects with varying physical configurations and characteristics.

Hand Pose Data: Each sample contains 3D coordinates for 21 keypoints per hand, providing detailed hand pose information essential for accurate gesture recognition.

Hand Bounding Box: The dataset includes annotations for the bounding boxes of detected hands in each video frame. These annotations were generated using the LabelImg software [43].

Heatmap Annotations: Heatmap annotations for video frames, derived from the Convolutional Neural Network (CNN)-based model described in [20], are also part of the dataset.

Table 1 and Figure 1 provide an overview of the dataset and offer sample illustrations for reference.

4.4. Demo

With the aim of making the sign language as a popular language for hearing-impaired people, an API as well as a GitHub repository have been prepared to show the details of the proposed dataset. In this way, a Flutter-based dictionary is proposed to search the sign sentences and show the corresponding videos. This application can be used in SLR and authentication systems to solve communication barriers. Our dataset and the corresponding demo will be publicly available after the paper publication.

5. Experimental results

This section presents the performance of fifteen models across Sign Language Recognition (SLR) and related fields, using the proposed dataset, as detailed in Table 2. SLR techniques primarily utilize body and hand features. Key sub-domains relevant to SLR include hand detection, pose estimation, real-time tracking, and gesture recognition. We report the experimental results of several models from these sub-areas, showcasing their performance on the dataset (See Figure 2). The details of these models, along with their respective outcomes, are summarized in Table 2. Additionally, we evaluate the synthesis of sign language videos using the dataset, providing insights into the results. Finally, we present state-of-the-art evaluation outcomes specifically for SLR. Three evaluation metrics are used to report the numerical results:

Intersection over Union (IoU): IoU is calculated as the ratio of the area of overlap between the

predicted region and the ground truth region to the area of their union. It is expressed as:

$$\text{IoU}=\frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (1)$$

where Area of Overlap is the area where the predicted region and the ground truth region intersect. Area of Union is the total area covered by both the predicted and ground truth regions combined.

Average pixel errors in the location of the keypoints: It is defined as the average of the errors in the pixel of the keypoints locations.

Recognition/Detection accuracy: Recognition/ Detection accuracy is the proportion of true positive detections (correctly identified hands or hand parts) to the total number of ground truth instances (all hands or hand parts in the dataset). It is typically expressed as:

$$\text{Detection Accuracy}=\frac{\text{Number of Correct Detections}}{\text{Total Number of Instances}} \quad (2)$$

5.1. Hand detection

Considering the significant role of the hand detection in SLR area, different models have been suggested to enhance the detection accuracy of these models. Although, there are still many challenges in computation time and detection accuracy aspects [45]. Using the transfer learning approach, some of the well-known object detection models have been fine-tuned to use for hand detection. For instance, the region-based methods, such as Region-based Convolutional Neural Network (RCNN), Fast Region-based Convolutional Neural Network (Fast-RCNN), and Faster Region-based Convolutional Neural Network (Faster-RCNN), as well as the You Only look Once (YOLO) and Single Shot Multi-Box Detector (SSD) models are some of the CNN-based models used for hand detection. Moreover, a real-time and multi-view convolutional-based method has been proposed by Simon et al. for hand detection from RGB still images. In this way, a keypoint detector is used to make the noisy labels. Moreover, a multi-view geometry approach is employed to convert and reproject the 2D detected keypoints into the 3D view. Results on own dataset show that the proposed model obtains the average error of 3.65 compared to state-of-the-art model [46]. However, the results on the proposed dataset in this paper show the average error of 11.15 (Figure 2).

5.2. Hand pose estimation

In recent years, hand pose estimation has made significant advancements due to the advent of depth sensors, impacting various applications [47, 48]. Zimmermann and Brox introduced a three-step CNN-based model designed to learn 3D articulation priors and keypoints from RGB images. To train the model, a large-scale synthetic hand pose dataset was developed. Performance evaluation on the Stereo Hand Pose Tracking Benchmark and Dexter datasets showed notable improvements. Specifically, the model achieved an Area Under the ROC Curve (AUC) of 94.0 and 49.0, respectively, with relative state-of-the-art improvements of 10.9 and 5 points [46]. The results of this model on the proposed dataset are depicted in Figure 3 for further analysis.

5.3. Real-time hand tracking

Hand tracking, as a critical component in sign language recognition (SLR) models, presents a significant challenge due to frequent occlusions of the fingers and joints during signing. To address this, Dibia developed a CNN-based repository using the SSD model for real-time hand tracking. After data pre-processing, the model was trained and tested on the Egohands dataset, achieving an impressive detection accuracy of 98.86% [47]. The model was further evaluated using a proposed dataset, demonstrating a detection accuracy of 91.20%. In another approach, a hybrid model combining CNN and a kinematic 3D hand model was designed for real-time 3D hand tracking from monocular RGB input images. To enhance the dataset, a generative model was employed to synthesize training data, while an RGB dataset with annotated 3D hand joint positions was introduced. Results from the Stereo and Dexter datasets revealed improvements in accuracy by 1.7% and 15%, respectively, compared to previous state-of-the-art methods [48].

5.4. Hand gesture recognition

Hand gesture recognition provides the fundamental information for HCI applications. In this sub-section, some recent models for hand gesture recognition from the RGB and the skeletal data are analyzed.

5.4.1. RGB-based hand gesture models

Köpüklü et al. conducted an analysis of various CNN architectures, focusing on factors such as offline classification accuracy, parameter count, and computational complexity. Their experiments, performed on two well-known datasets, EgoGesture and NVIDIA Dynamic Hand Gesture, demonstrated that the ResNeXt-101 model achieved an impressive offline classification accuracy of 94.03% on the EgoGesture benchmark.

It also produced competitive results compared to state-of-the-art models on the NVIDIA benchmark [49]. In our evaluation of the proposed dataset using the same architecture, we achieved an offline classification accuracy of 90.25%, confirming its effectiveness in our scenario. Moreover, we present the results of the 3DCNN [2], I3D [3], a combination of CNN and Long Short-Term Memory (LSTM), and a combination of CNN and Gated Recurrent Unit (GRU), Transformer Encoder [5] with hand keypoints as input, Swin Transformer [40] plus LSTM, Vision Transformer (ViT) [40] plus LSTM, and Graph Convolutional Network (GCN) [13] plus LSTM in Table 3. As this table shows, Transformer Encoder with hand keypoints as input has the highest performance.

Table 2. Results of different tasks in sign language.

Ref.	Task	Input modality	Dataset	Result
[46]	Hand detection	RGB	Own dataset	Average error: 3.65
			Proposed	Average error: 11.15
[49]	Hand pose estimation	Depth	Stereo Hand Pose Tracking	AUC: 94.0
			Proposed	AUC: 55.0
[50]	Real-time hand tracking	RGB	Egohands	Accuracy: 98.86%
			Proposed	Accuracy: 91.20%
[51]	Real-time hand tracking	RGB	Stereo Hand Pose Tracking	AUC: 0.965
			Proposed	AUC: 0.52
[52]	Hand gesture recognition	RGB	NVIDIA benchmark	Accuracy: 94.03%
			Proposed	Accuracy: 90.25%
[53]	Hand gesture recognition	Skeleton	DHG	Accuracy: 84.35%
			Proposed	Accuracy: 80.10%
[54]	SLP	Skeleton	Proposed	WER: 39.20%
[55]	SLP	Skeleton	Proposed	WER: 37.10%
[56]	SLP	Skeleton	Proposed	WER: 36.20%
[57]	SLP	Skeleton	Proposed	WER: 35.25%
[60]	SLR	RGB	PHOENIX14	WER: 19.4%
[39]	SLP	Skeleton	PHOENIX14-T	BLEU-1: 29.74
			Proposed	BLEU-1: 27.10
[40]	SLP	RGB	PHOENIX14-T	BLEU-1: 32.41
			Proposed	BLEU-1: 30.05
[58]	SLP	Skeleton	PHOENIX14-T	BLEU-1: 34.94
			Proposed	BLEU-1: 30.15
[59]	SLP	RGB	PHOENIX14T	SSIM: 0.9338
			Proposed	SSIM: 0.9010

5.4.2. Skeleton-based hand gesture models

Devineau et al. proposed a CNN-based architecture for hand gesture recognition by leveraging hand skeleton data. The model utilizes parallel convolutions across multiple temporal resolutions to effectively learn the sequential patterns of skeletal hand movements. Their approach demonstrated a significant 3% improvement over the previous state-of-the-art on the DHG dataset [50]. Additionally, tests conducted on a new dataset showed a recognition accuracy of 89%, further validating the model's performance.

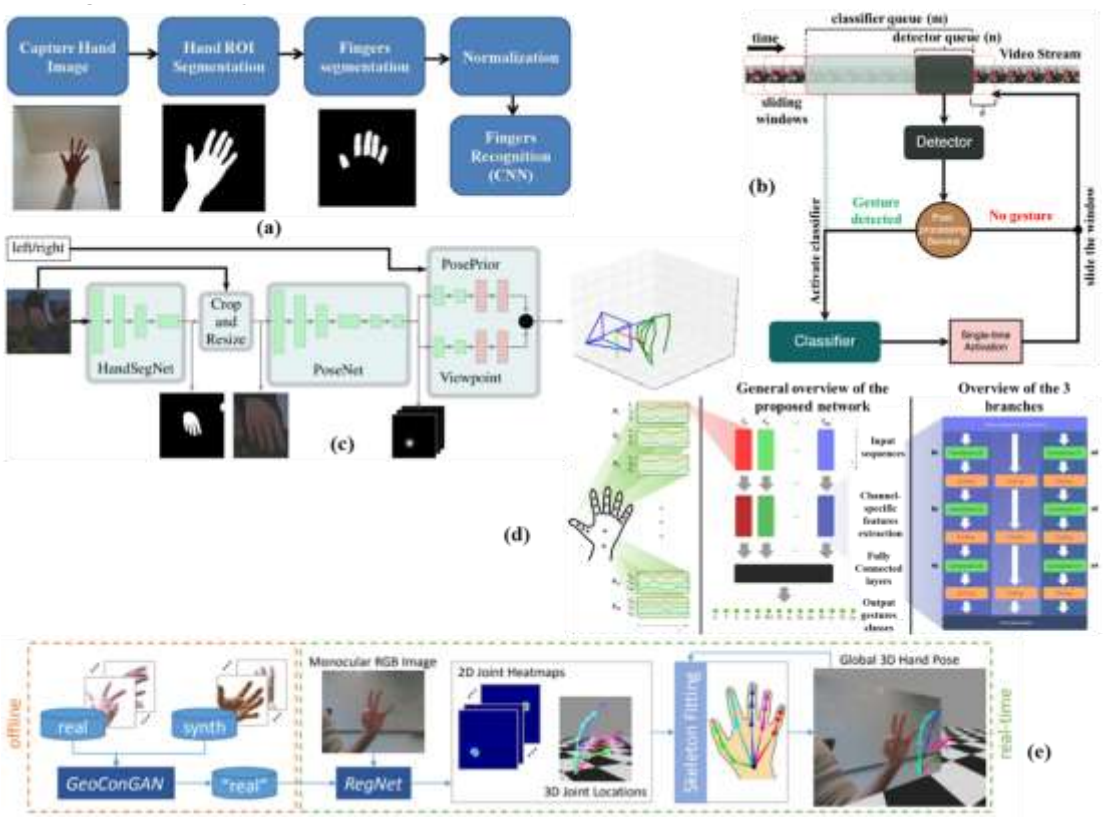

Figure 2. An overview of some models used in our experimental evaluation on the proposed dataset: (a) [42], (b) [49], (c) [46], (d) [50], (e) [48].

Table 3. Recognition accuracy of the seventh models on the proposed dataset.

Model	Recognition accuracy
3DCNN	0.62
I3D	0.65
Transformer Encoder	0.68
CNN+LSTM	0.64
CNN+GRU	0.63
Swin Transformer + LSTM	0.66
ViT + LSTM	0.65
GCN+LSTM	0.66

5.5. SLP

Here, we analyze the SLP from the skeletal data. To do this, the hand pose annotations in the proposed dataset are used to feed to some deep generative models: Generative Adversarial Network (GAN) [51], Conditional Generative Adversarial Network (cGAN) [52], Wasserstein Generative Adversarial Network (WGAN) [53], and Wasserstein Conditional Generative Adversarial Network (WCGAN) [54]. Relying on the capability of the deep generative models for the skeletal data distribution learning, these models can generate accurate skeleton-based videos. Figure 4 show a sample of the generated video frame by using these deep generative models. Moreover, Saunders et al. proposed a Transformer-based model for SLP using the continuous 3D multi-channel sign pose sequences. Relying on a counter decoding, this model tackle with variable length of continuous sequence generation by tracking the generation progress over time and predicting the end of sequence.

Figure 4. Results of hand pose estimation [46] corresponding to a video sequence from the proposed dataset.

Results on the PHOENIX14T dataset confirm the superiority of the model compared to state-of-the-art models [55]. In addition, Stoll et al. suggested a model, including some sub-processes to enhance the translation accuracy of the SLP task.

Figure 3. Results of hand detection [43] corresponding to some samples of the proposed dataset.

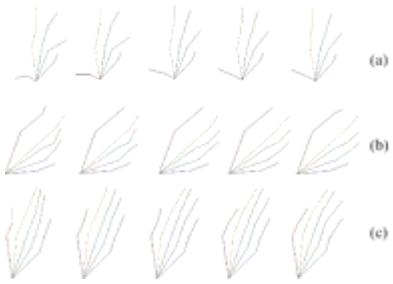

Figure 6. Visual results of the two SLP models in [56] and [57] on proposed dataset: (a) GT, (b) [56], (c) [57].

Figure 5. A sample of the generated video frame by using four deep generative models.

To this end, the spoken language sentences are translated into sign pose sequences by combining an NMT network with a Motion Graph. The obtained pose sequences are then employed to condition a generative model that generates photo realistic sign language video sequences. Evaluation results on the PHOENIX14T dataset show a BLEU-4 score of 16.34/15.26 on dev/test sets [56]. Hu et al. designed a Correlation Network (CorrNet) to employ the body trajectories across frames, aiming to identify signs. To this end, the correlation maps between the current frame and the adjacent frames are calculated to identify trajectories of all spatial patches. Using an identification module, the body trajectories within the correlation maps are gathered to obtain an overview of local temporal movements to recognize a sign. Results on four datasets, PHOENIX14, PHOENIX14-T, CSL-Daily, and CSL, show the accuracy improvement compared to state-of-the-art models in SLR [57]. Visual and numerical results of these models on the proposed dataset have been shown in Figsures 4-5 and Table 2, respectively.

5.6. Stat-of-the-art evaluation on SLR

Here, we present the state-of-the-art evaluation results of SLR for continuous signs on the current datasets as well as the proposed dataset (See Table 2). The results presented in the tables and figures demonstrate the enhancement in model performance achieved through the deep learning models. These approaches have successfully increased the accuracy by a substantial margin. However, there is still room for further optimization and refinement, suggesting that additional improvements can be made to further boost the model's overall effectiveness.

6. Advantages and limitations

The proposed dataset is designed to support both Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR). A key advantage of this dataset is its inclusion of both continuous signs and their corresponding isolated signs, enhancing its utility for various applications. The dataset's flexibility can be further expanded by increasing the number of sign sentences and words in future updates. Additionally, introducing more signers and diverse backgrounds will allow for the development of more generalized models, thereby improving the dataset's applicability and accuracy across different scenarios. Addressing these enhancements will strengthen the dataset's value for a broader range of research and practical uses.

7. Conclusion and future work

In this paper, we presented a dataset for SLR, including the sign sentences as well as the corresponding isolated signs. This dataset contains the 10'000 videos of 100 Persian sign sentences using 10 Deaf and hearing people. In this way, some annotations are used for the video samples, including the bounding box of the detected hands, the class label, the hands pose parameters, and the heatmap. Furthermore, to analyze the complexity of the dataset in comparison with the other datasets, some experiments have been presented on the proposed dataset and the results have been discussed. To this end, the results of the models in SLR and some related area, such as hand detection, hand pose estimation, and hand gesture recognition, have been presented on the public datasets as well as the proposed dataset. Also, some experiments have been done on the generation capability of some deep generative models in sign video production. In overall, the proposed dataset can be effectively used in SLR as well as SLP applications. Considering the data requirements of deep learning models, the proposed dataset includes 100 samples per each class that is a positive characteristic of this dataset. However, the number of the sign sentences should be extended in order to use this dataset in real-world communications. Another characteristic of the proposed dataset that does not find in the current datasets, is the existence of both sign sentences as well as the corresponding sign words. In the future, we aim to extend this dataset to include more sign sentences. We also will include the gloss annotations in the dataset.

References

[1] World Health Organization, “Deafness and hearing loss,” https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed 17 June 2024.

[2] R. Rastgoo, K. Kiani, S. Escalera, M. Sabokrou M, “Multi-modal zero-shot dynamic hand gesture recognition,” *Expert Systems with Applications*, vol. 247, 123349, 2024.

[3] R. Rastgoo, K. Kiani, S. Escalera, “Word separation in continuous sign language using isolated signs and post-processing,” *Expert Systems with Applications*, vol. 249, 123695, 2024.

[4] R. Rastgoo, K. Kiani K, “Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning,” *Journal of Modeling in Engineering*, vol. 17, pp. 103-111, 2019.

[5] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, M. Sabokrou, “A survey on recent advances in Sign Language Production,” *Expert Systems with Applications*, vol. 243, 122846, 2024.

[6] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, M. Sabokrou, “All you need in sign language production,” *arXiv:2201.01609*, 2022.

[7] R. Rastgoo, K. Kiani, S. Escalera, “Sign language recognition: A deep survey,” *Expert Systems with Applications*, vol. 164, 113794, 2021.

[8] R. Rastgoo, K. Kiani, S. Escalera, “Hand pose aware multimodal isolated sign language recognition,” Multimedia Tools and Applications, vol. 80, pp. 127-163, 2021.

[9] R. Rastgoo, K. Kiani, S. Escalera, “Zs-slr: Zero-shot sign language recognition from rgb-d videos,” arXiv preprint arXiv:2108.10059, 2021.

[10] Z. Mohammadi, A. Akhavanpour, R. Rastgoo, M. Sabokrou, “Diverse hand gesture recognition dataset,” Multimedia Tools and Applications, vol. 83, pp. 50245-50267, 2024.

[11] R. Rastgoo, K. Kiani, S. Escalera, “A deep co-attentive hand-based video question answering framework using multi-view skeleton,” Multimedia Tools and Applications, vol. 82, pp. 1401-1429, 2023.

[12] R. Rastgoo, K. Kiani, S. Escalera, “ZS-GR: zero-shot gesture recognition from RGB-D videos,” Multimedia Tools and Applications, vol. 82, pp. 43781-43796, 2023.

[13] R. Rastgoo, K. Kiani, S. Escalera, “A Non-Anatomical Graph Structure for isolated hand gesture separation in continuous gesture sequences,” arXiv preprint arXiv:2207.07619, 2022.

[14] R. Rastgoo, K. Kiani, S. Escalera, “A transformer model for boundary detection in continuous sign language,” Multimedia Tools and Applications, 1-18, 2024.

[15] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, “RWTH-phoenix-weather 2014t: Parallel corpus of sign language video, gloss and translation,” CVPR, Salt Lake City, UT, 2018.

[16] A. Duarte, Sh. Palaskar, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, J. GiroiNieto, “How2sign: A large-scale multimodal dataset for continuous American sign language,” *Sign Language Recognition, Translation, and Production workshop*., 2020.

[17] J. Forster, Ch. Schmidt, Th. Hoyoux, O. Koller, U. Zelle, J. Piater, H. Ney, “RWTH phoenix-weather: A large vocabulary sign language recognition and translation corpus,” LREC12, Istanbul, Turkey, pp. 3785–3789, 2012.

[18] C. Neidle, A. Opoku, D. Metaxas, “ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP),” *arXiv:2201.07899*, 2022.

[19] R. Rastgoo, K. Kiani, S. Escalera, Multi-Modal Deep Hand Sign Language Recognition in Still Images Using Restricted Boltzmann Machine. Entropy 20 (11):809.

[20] R. Rastgoo, K. Kiani, S. Escalera, “Hand sign language recognition using multi-view hand skeleton,” *Expert System with Applications*, vol. 150, 113336, 2020.

[21] R. Rastgoo, K. Kiani, S. Escalera, “Video-based isolated hand sign language recognition using a deep cascaded model,” *Multimedia Tools and Application*, vol. 79, pp. 22965-22987, 2020.

[22] R. Rastgoo, K. Kiani, S. Escalera, “Real-time isolated hand sign language recognition using deep networks and SVD,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 591–611, 2022.

[23] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, A. Thangali, “The American sign language lexicon video dataset,” *CVPRW*, pp. 1–8, 2018.

[24] N. Caselli, Z. Sevcikova, A. Cohen-Goldberg, K. Emmorey, “ASL-LEX: A lexical database for ASL,” *Behavior Research Methods*, vol. 49, pp. 784-801, 2016.

[25] H.R. Vaezi Joze, O. Koller, “MS-ASL: A largescale data set and benchmark for understanding American sign language,” *arXiv:1812.01053*, 2018.

[26] D. Li, C. Rodriguez, X. Yu, H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” *WACV*, pp. 1459–1469, 2020.

[27] A.M. Martınez, R.B. Wilbur, R. Shay, A.C. Kak, “Purdue rvl-slll asl database for automatic recognition of American sign language,” Fourth IEEE International

Conference on Multimodal Interfaces, pp. 167–172, 2002.

[28] S. Vintar, B. Jerko, M. Kulovec "Compiling the Slovene sign language corpus," *LREC*, vol. 5, pp. 159–162, 2012.

[29] S. Ghanbari Azar, H. Seyedarabi, "Trajectory-based recognition of dynamic Persian sign language using hidden Markov model," *Computer Speech & Language*, vol. 61, 101053, 2020.

[30] https://github.com/salar96/Iranian-Sign-Language-ISL-Dataset. Access Date: Dec., 2024.

[31] J. Huang, W. Zhou, O. Zhang, H. Li, W. Li, "Video-based sign language recognition without temporal segmentation," *AAAI*, 2018.

[32] O. Koller, J. Forster, H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108-125, 2015.

[33] N.H. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, "Neural sign language translation," *CVPR*, pp. 7784–7793, 2018.

[34] T. Hanke, M. Schulder, R. Konrad, E. Jahn, "Extending the public dgs corpus in size and depth," *Workshop on the Representation and Processing of Sign Languages*, pp. 75–82, 2020.

[35] U.V. Agris, K.F. Kraiss K.F, "Signum database: Video corpus for signer-independent continuous sign language recognition," *Workshop on Representation and Processing of Sign Languages*, pp. 243–246, 2010.

[36] C. Neidle, Ch. Vogler, "A new web interface to facilitate access to corpora: Development of the asllrp data access interface," *Proc. 5th Workshop on the Representation and Processing of Sign Languages*, 2012.

[37] M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, H. Ney, "Continuous sign language recognition-approaches from speech recognition and available data resources," *Workshop on Representation and Processing of Sign Languages*, 2006.

[38] A. Duarte, et al. "How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language," *CVPR*, 2021.

[39] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, K. Cormier K, "Building the British sign language corpus," Language *Documentation & Conservation*, vol. 7, pp. 136–154, 2013.

[40] B. Saunders, N.C. Camgoz, R. Bowden, "Progressive transformers for end-to-end sign language production," *ECCV*, 2020.

[41] B. Saunders, N.C. Camgoz, R. Bowden, "Adversarial training for multi-channel sign language production," *BMVC*, 2020.

[42] T. Hanke, M. Schulder, R. Konrad, E. Jahn E, "Extending the public dgs corpus in size and depth," *Workshop on the Representation and Processing of Sign Languages*, pp. 75–82, 2020.

[43]https://pillow.readthedocs.io/en/stable/reference/Image.html. Access Date: Dec., 2024.

[44] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh, "OpenPose: real-time multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172-186, 2021.

[45] T. Le, D. Jaw, I. Lin, H. Liu, S. Huang, "An efficient hand detection method based on convolutional neural network," *7th IEEE international symposium on next-generation electronics*, 2018.

[46] T. Simon, H. Joo, I. Matthews, Y. Sheikh, "Hand keypoint detection in single images using multi-view bootstrapping," *CVPR*, pp. 1145-1153, 2017.

[47] Sh. Wang, Sh. Wang, H. Kuang, F. Li, Z. Qian, M. Li, "A Survey of Deep Learning-based Hand Pose Estimation," EEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS), 2022.

[48] J. Supancic, G. Rogez, Y. Yang, J. Shotton, D. Ramana, "Depth-based hand pose estimation: methods, data, and challenges," *International Journal of Computer Vision*, vol. 126, pp. 1180–1198, 2018.

[49] C. Zimmermann, T. Brox, "Learning to estimate 3D hand pose from single RGB images," *ICCV*, Italy, 2017.

[50] V. Dibia, "Handtrack: A library for prototyping real-time hand tracking interfaces using convolutional neural networks," *GitHub*, https://github.com/victordibia/handtracking, 2017.

[51] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular RGB," *CVPR*, USA, 2018.

[52] O. Kopuklu, A. Gunduz, N. Kose, G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," *arXiv:1901.10323*, 2019.

[53] G. Devineau, W. Xi, F. Moutarde, J. Yang, "Deep learning for hand gesture recognition on skeletal data," *13th IEEE conference on automatic face and gesture recognition*, China, 2018.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative Adversarial Nets," *Advances in Neural*

Information Processing Systems (NIPS 2014), vol. 27, 2014.

[55] M. Mirza, S. Osindero, "Conditional Generative Adversarial Nets," *arXiv:1411.1784*, 2014.

[56] M. Arjovsky, S. Chintala, L. Bottou, "Wasserstein GAN," *arXiv:1701.07875v3*, 2017.

[57] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, Y. Zheng, "Recent Progress on Generative Adversarial Networks (GANs): A Survey," *IEEE Access*, vol. 7, pp. 36322–36333, 2019.

[58] B. Saunders, N.C. Camgoz, R. Bowden, "Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks," *International Journal of Computer Vision*, vol. 129, pp. 2113–2135, 2021.

[59] S. Stoll, N.C. Camgoz, S. Hadfield, R. Bowden, "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks," *Int J Comput Vis*, vol. 128, pp. 891–908, 2020.

[60] L. Hu, L. Gao, Z. Liu, W. Feng, "Continuous Sign Languag Recognition with Correlation Network," *CVPR*, 2023.

مجموعه داده زبان اشاره پیوسته فارسی

راضیه راستگو*

دانشکده برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۲۰۲۴/۰۹/۰۵؛ بازنگری ۲۰۲۴/۱۱/۱۱؛ پذیرش ۲۰۲۵/۰۱/۱۲

چکیده:

زبان اشاره، روش اصلی ارتباط در جامعه ناشنوایان است. پیشرفت‌های اخیر در یادگیری عمیق منجر به توسعه‌ی فناوری‌های مختلفی شده که هدفشان تسهیل ارتباط دوطرفه بین جوامع ناشنوا و شنوا است. با این حال، چالش‌هایی در راستای در دسترس بودن مجموعه داده‌های مناسب برای مدل‌های مبتنی بر یادگیری عمیق وجود دارد. تنها تعداد کمی مجموعه داده‌های عمومی و برچسب‌گذاری‌شده با مقیاس بزرگ برای جملات زبان اشاره وجود دارد و هیچ مجموعه داده‌ای برای جملات زبان اشاره فارسی در دسترس نیست. برای پر کردن این خلأ، ما یک مجموعه داده شامل ۱۰٬۰۰۰ ویدئو مربوط به ۱۰۰ جمله زبان اشاره فارسی جمع‌آوری کرده‌ایم. این مجموعه داده شامل برچسب‌گذاری جامع، از جمله مختصات دست‌های شناسایی‌شده، برچسب‌ها جملات، پارامترهای حالت دست و نقشه‌های حرارتی است. یکی از ویژگی‌های برجسته این مجموعه داده، وجود کلمات زبان اشاره مربوط به جملات موجود در مجموعه داده جمع‌آوری‌شده است. برای تحلیل پیچیدگی این مجموعه داده، آزمایش‌های گسترده‌ای ارائه شده و نتایج آن مورد بحث قرار گرفته‌اند. به طور مشخص‌تر، نتایج مدل‌ها در زیربخش‌های کلیدی مرتبط با شناسایی زبان اشاره، از جمله شناسایی دست، تخمین حالت، ردیابی در زمان واقعی و شناسایی حرکات، بررسی و تحلیل شده است. علاوه بر این، نتایج هفت مدل مبتنی بر یادگیری عمیق روی این مجموعه داده مورد بحث قرار گرفته‌اند. در نهایت، نتایج تولید زبان اشاره با استفاده از مدل‌های مولد عمیق ارائه شده است. ما نتایج آزمایش‌های این مدل‌ها را در این زیربخش‌ها گزارش کرده‌ایم که عملکرد آن‌ها را روی مجموعه داده پیشنهادی نشان می‌دهد.

کلمات کلیدی: شناخت زبان اشاره، تولید زبان اشاره، مجموعه داده، شبکه متخاصم مولد، فارسی.