Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20240302022
Volume 3, Issue2 (May-August 2024), PP: 148-153.
www.indjcst.com
ISSN No: 2583-5300

Sign Tone: Two Way Sign Language Recognition and Multilingual Interpreter System Using Deep Learning

S. Surya[1], Saranya S[2], Swetha R[3], Hemashree G[4]
[1]*Assistant Professor, Department of Information Technology, Er. Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India.*
[2,3,4] *Department of Information Technology, Er. Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India.*

To Cite this Article: *S. Surya[1], Saranya S[2], Swetha R[3], Hemashree G[4], "Sign Tone: Two Way Sign Language Recognition and Multilingual Interpreter System Using Deep Learning", Indian Journal of Computer Science and Technology, Volume 03, Issue 02 (May-August 2024), PP: 148-153.*

Abstract: *In this paper, a comparative experimental assessment of computer vision-based methods for sign language reclog- nation is conducted. By implementing the most recent deep neural network methods in this field, a thorough evaluation on multiple publicly available datasets is performed. The aim of the present study is to provide insights on sign language recognition, focusing on mapping non-segmented video streams to glosses. For this task, two new sequence training criteria, known from the fields of speech and scene text recognition, are introduced. Furthermore, a plethora of pretraining schemes is thoroughly discussed. Finally, a new RGB+D dataset for the Greek sign language is created. To the best of our knowledge, this is the first sign language dataset where three annotation levels are provided (individual gloss, sentence and spoken language) for the same set of video captures.*

Key Words: *Sign Language Recognition, Greek Sign Language, Deep Neural Networks, Stimulated CTC, Conditional Entropy CTC.*

I.INTRODUCTION

Spoken languages make use of the "vocal - auditory" channel, as they are articulated with the mouth and perceived with the ear. All writing systems also derive from, or are representations of, spoken languages. Sign languages (SLs) are different as they make use of the "corporal - visual" channel, produced with the body and perceived with the eyes. SLs are not international and they are widely used by the communities of the Deaf. They are natural languages since they are developed spontaneously wherever the Deaf have the opportunity to congregate and communicate mutually [1]. SLs are not derived from spoken languages; they have their own independent vocabularies and their own grammatical structures [1]. The signs used by the Deaf, actually have internal structure in the same way as spoken words. Just as hundreds of thousands of English words are produced using a small number of different sounds, the signs of SLs are produced using a finite number of gestural features. Thus, signs are not holistic gestures but are rather analyzable, as a combination of linguistically significant features. Similarly, to spoken languages, SLs are composed of the following indivisible features:

- Manual features, i.e. hand shape, position, movement, orientation of the palm or fingers, and
- Non-manual features, namely eye gaze, head-nods/ shakes, shoulder orientations, various kinds of facial expression as mouthing and mouth gestures.

Combinations of the above-mentioned features represent a gloss, which is the fundamental building block of a SL and represents the closest meaning of a sign [2]. SLs, similar to the spoken ones, include an inventory of flexible grammatical rules that govern both manual and non-manual features [3]. Both of them, are simultaneously (and often with loose temporal structure) used by signers, in order to construct sentences in a SL. Depending on the context, a specific feature may be the most critical factor towards interpreting a gloss. It can modify the meaning of a verb, provide spatial/temporal reference and discriminate between objects and people. Due to the intrinsic difficulty of the Deaf community to interact with the rest of the society (according to [4], around 500,000 people use the American SL to communicate in the USA), the development of robust tools for automatic SL recognition would greatly alleviate this communication gap. As stated in [5], there is an increased demand for interdisciplinary collaboration including the Deaf community and for the creation of representative public video datasets.

Sign Language Recognition (SLR) can be defined as the task of inferring glosses performed by a signer from video captures. Even though there is a significant amount of work in the field of SLR, a lack of a complete experimental study is profound. Moreover, most publications do not report results in all available datasets or share their code. Thus, experimental results in the field of SL are rarely reproducible and lacking interpretation. Apart from the inherent difficulties related to human motion analysis

(e.g. differences in the appearance of the subjects, the human silhouette features, the execution of the same actions, the presence of occlusions, etc.) [6], automatic SLR exhibits the following key additional challenges:

- The Deaf often employ a grammatical device known as "role-shifting" when narrating an event or a story with more that one characters [7]. Therefore, exact position in surrounding space and context have a large impact on the interpretation of SL. For example, personal pronouns (e.g. "he", "she", etc.) do not exist. Instead, the signer points directly to any involved referent or, when reproducing the contents of a conversation, pronouns are modeled by twisting his/her shoulders or gaze. Additionally, the Deaf leverage the space in front of them (signing space) in order to localize people or places [8]. The latter is referred as placement in most SLs. By placing a person or a city somewhere in his/her signing space, the singer can refer to a person by pointing in its assigned space or show where a place is located, relative to the placed city.
- Many glosses are only distinguishable by their constituent non-manual features and they are typically difficult to be accurately detected, since even very slight human move- ments may impose different grammatical or semantic interpretations depending on the context [9].
- The execution speed of a given gloss may indicate a different meaning or the particular signer's attitude. For instance, signers would not use two glosses to express "run quickly", but they would simply speed up the execution of the involved signs [9].
- Signers often discard a gloss sub-feature, depending on previously performed and proceeding glosses. Hence, different instances of the exact same gloss, originating even from the same signer, can be observed.
- For most SLs so far, very few formal standardization ac- tivities have been implemented, to the extent that signers of the same country exhibit distinguishable differences during the execution of a given gloss [10].

Historically, before the advent of deep learning methods, the focus was on identifying isolated glosses and gesture spotting. Developed methods were often making use of hand crafted techniques [11], [12]. For spatial representation of the different sub-gloss components, they usually used handcrafted features and/or fusion of multiple modalities. Temporal model- ing was achieved by classical sequence learning models, such as Hidden Markov Model (HMM) [13], [14], [15] and hidden conditional random fields [16]. The rise of deep networks was met with a significant boost in performance for many video- related tasks, like human action recognition [17], [18], gesture recognition, [19], [20], motion capturing [21], [22], etc. SLR is a task closely related to computer vision. This is the reason that most approaches tackling SLR have adjusted to this direction. In this paper, SLR using Deep Neural Network (DNN) methods is investigated. The main contributions of this work are summarized as follows:

- A comprehensive, holistic and in-depth analysis of multi- ple literature DNN-based SLR methods is performed, in order to provide meaningful and detailed insights to the task at hand.
- Two new sequence learning training criteria are proposed, known from the fields of speech and scene text recogni- tion.
- A new pretraining scheme is discussed, where transfer learning is compared to initial pseudo-alignments.
- A new publicly available large-scale RGB+D Greek Sign Language (GSL) dataset is introduced, containing real- life conversations that may occur in different public services. This dataset is particularly suitable for DNN- based approaches that typically require large quantities of expert annotated data

II.RELATED WORK

The various automatic SLR tasks, depending on the model- ing's level of detail and the subsequent recognition step, can be roughly divided in (Fig. 1):

- Isolated SLR: Methods of this category target to address the task of video segment classification (where the seg- ment boundaries are provided), based on the fundamental assumption that a single gloss is present [23], [24], [20].
- Sign detection in continuous streams: The aim of these approaches is to detect a set of predefined glosses in a continuous video stream [13], [25], [26].
- Continuous SLR (CSLR): These methods aim at rec- ognizing the sequence of glosses that are present in a continuous/non-segmented video sequence [27], [28], [29]. This category of approaches exhibits characteristics that are most suitable for the needs of real-life SLR applications [5]; hence, it has gained increased research attention and will be further discussed in the remainder of this section.

A. Continuous Sign Language Recognition

By definition, CSLR is a task very similar to the one of continuous human action recognition, where a sequence of glosses (instead of actions) needs to be identified in a contin- uous stream of video data. However, glosses typically exhibit a significantly shorter duration than actions (i.e. they may only involve a very small number of frames), while transitions among them are often very subtle for their temporal boundaries to be efficiently recognized. Additionally, glosses may only involve very detailed and fine-grained human movements (e.g. finger signs or facial expressions), while human actions usually refer to more concrete and extensive human body actions. The latter facts highlight the particular challenges that are present in the CSLR field [3]. Due to the lack of gloss-level annotations, CSLR is regularly casted as a weakly supervised learning problem. The majority of CSLR architectures usually consists of a feature extractor, followed by a temporal modeling mechanism [30], [31]. The feature extractor is used to compute feature representations from individual input frames (using 2D CNNs) or sets of neighboring frames (using 3D CNNs). On the other hand, a critical aspect of the temporal modeling scheme enables the modeling of the SL unit feature representations (i.e., gloss- level, sentence-level). With respect to temporal modeling, sequence learning can be achieved using HMMs, Connec- tionist Temporal Classification (CTC) [32] or Dynamic Time.

Warping (DTW) [33] techniques. From the aforementioned categories, CTC has in general shown superior performance and the majority of works in CSLR has established CTC as the main sequence training criterion (for instance, HMMs may fail to efficiently model complex dynamic variations, due to expressiveness limitations [28]). However, CTC has the tendency to produce

overconfident peak distributions that are prone to overfitting [34]. Moreover, CTC introduces limited contribution towards optimizing the feature extractor [35]. For these reasons, some recent approaches have adopted an iterative training optimization methodology. The latter essentially comprises a two-step process. In particular, a set of temporally-aligned pseudo-labels are initially estimated and used to guide the training of the feature extraction module. In the beginning, the pseudo-labels can be either estimated by statistical approaches [3] or extracted from a shallower model [28].

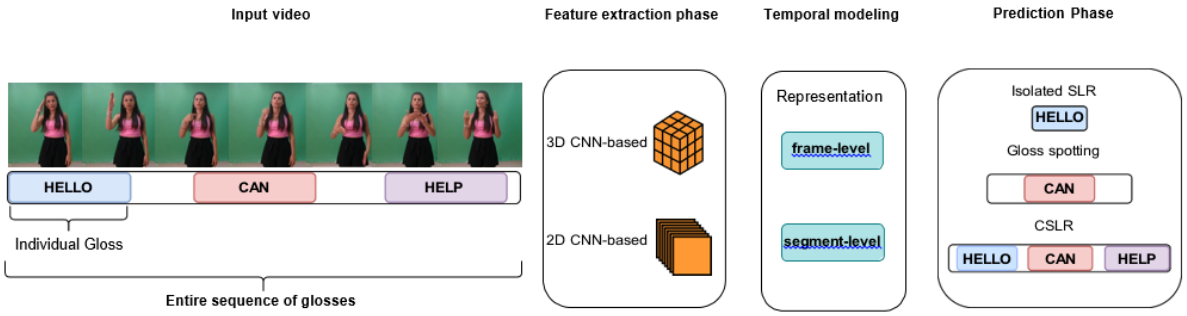

Figure 1 An overview of SLR categories

After training the model in an isolated setup, the trained feature extractor is utilized for the continuous SLR setup. This process may be performed in an iterative way, similarly to the Expectation Maximization (EM) algorithm [36]. Finally, CTC imposes a conditional independence constraint, where output predictions are independent, given the entire input sequence.

B. 2D CNN-based CSLR approaches

One of the firstly deployed architectures in CSLR is based on [37], where a CNN-HMM network is proposed. Google-LeNet serves as the backbone architecture, fed with cropped hand regions and trained in an iterative manner. The same network architecture is deployed in a CSLR prediction setup [38], where the CNN is trained using glosses as targets instead of hand shapes. Later on, in [39], the same authors extend their previous work by incorporating a Long Short- Term Memory unit (LSTM) [40] on top of the aforementioned network. In a more recent work [27], the authors present a three-stream CNN-LSTM-HMM network, using full frame, cropped dominant hand and signer's mouth region modalities. These models, since they employ HMM for sequence learning, have to make strong initial assumptions in order to overcome HMM's expressive limitations.

In [30], the authors introduce an end-to-end system in CSLR without iterative training. It consists of two streams, one responsible for processing the full frame sequences and one for processing only the signer's cropped dominant hand. In [31], the authors employ a 2D CNN-LSTM architecture and in parallel with the LSTMs, a weakly supervised gloss detection regularization network, consisting of stacked tempo- ral 1D convolutions. The same authors in [28] extend their previous work by proposing a module composed of a series of temporal 1D CNNs followed by max pooling, between the feature extractor and the LSTM, while fully embracing the iterative optimization procedure. In [2], a hybrid 2D-3D CNN architecture [41] is developed. Features are extracted in a structured manner, where temporal dependencies are modeled by two LSTMs, without pretraining or using an iterative procedure. This approach however, yields the best results only in continuous SL datasets where a plethora of training data is available.

C. 3D CNN-based CSLR approaches

One of the first works that employs 3D-CNNs in SLR is introduced in [42]. The authors present a multi-modal approach for the task of isolated SLR, using spatio-temporal Convolutional 3D networks (C3D) [43], known from the re- search field of action recognition. Multi-modal representations are lately fused and fed to a Support Vector Machine (SVM) [44] classifier.
The C3D architecture has also been utilized in CSLR by [45]. The developed two-stream 3D CNN processes both full frame and cropped hand RGB images. The full network, named LS-HAN, consists of the proposed 3D CNN network, along with a hierarchical attention network, capable of latent space-based recognition modeling. In a later work [46], the authors propose the I3D [47] architecture in SLR. The model is deployed on an isolated SLR setup, with pretrained weights on action recognition datasets. The signer's body bounding box is served as input. For the evaluated dataset it yielded state-of-the-art results. In [35], the authors adopted and enhanced the original I3D model with a gated Recurrent Neural Network (RNN). Their aim is to accommodate features from different time scales.I3D has also been used as a baseline model in [48] on a large-scale isolated SLR dataset and achieved great recognition accuracy. In another work [49], the authors introduce the 3D-ResNet architecture to extract fea- tures. Furthermore, they substitute LSTM with stacked dilated temporal convolutions and CTC for sequence alignment and decoding. With this approach, they manage to have very large receptive fields while reducing time and space complexity, compared to LSTM. Finally, in [50], Pu et al. propose a framework that also consists of a 3D-ResNet backbone. The features are provided in both an attentional encoder-decoder network [51] and a CTC decoder for sequence learning. Both decoded outputs are jointly trained while the soft-DTW [52] is utilized to align them.

III.PUBLICLY AVAILABLE DATASETS

Existing SLR datasets can be characterized as isolated or continuous, taking into account whether annotation are provided at the gloss (fine-grained) or the sentence (coarse- grained) levels. Additionally, they can be divided into Signer Dependent (SD) and Signer Independent (SI) ones, based on the defined evaluation scheme. In particular, in the SI datasets a signer cannot be

present in both the training and the test set. In Table I, the following most widely known public SLR datasets, along with their main characteristics, are illustrated:

- The Signum SI and the Signum subset [53] include laboratory capturings of the German Sign Language. They are both created under strict laboratory settings with the most frequent everyday glosses.
- The Chinese Sign Language (CSL) SD, the CSL SI and the CSL isol. datasets [45] are also recorded in a predefined laboratory environment with Chinese SL words that are widely used in daily conversations.
- The Phoenix SD [54], the Phoenix SI [54] and the Phoenix-T [55] datasets comprise videos of German SL, originating from the weather forecast domain.
- The American Sign Language (ASL) [46] dataset con- tains videos of various real-life settings. The collected videos exhibit large variations in background, image quality, lighting and positioning of the signers.

A. The GSL dataset

Dataset description: In order to boost scientific research in the deep learning era, large-scale public datasets need to be created. In this respect and with a particular focus on the case of the GSL recognition, a corresponding public dataset has been created in this work. In particular, a set of seven native GSL signers are involved in the capturings. The considered application includes cases of Deaf people interacting with different public services, namely police departments, hospitals and citizen service centers. For each application case, 5 individual and commonly met scenarios (of increasing duration and vocabulary complexity) are defined. The average length of each scenario is twenty sentences with the mean individual sentence length amounting to 4.23 glosses. Subsequently, each signer was asked to perform the pre-defined dialogues in GSL five consecutive times. In all cases, the simulation considers a Deaf person communicating with a single public service employee, while all interactions are performed in GSL (the involved signer performed the sequence of glosses of both agents in the discussion).

Overall, the resulting dataset includes 10,295 sentence in- stances, 40,785 gloss instances, 310 unique glosses (vocabu- lary size) and 331 unique sentences. For the definition of the dialogues in the identified application cases, the particularities of the GSL and the corresponding annotation guidelines, GSL linguistic experts are involved.

The proposed Greek SLR dataset contains: a) temporal gloss annotations, b) sentence annotations, and c) translated annotations to the Modern Greek language. All the referenced annotations are performed in the same set of video captures. This is in contrast to other SL datasets that either contain only a small subset of isolated signs, or no translation to the spoken language. Thus, our dataset can serve as a benchmark for multiple SL tasks: isolated SLR, CSLR, and SL translation. This enables method evaluation from isolated to continuous SLR, or even SL translation, on the same videos.

Fig. 2: Example Keyframes of the introduced GSL dataset

2) GSL Evaluation Sets:

Regarding the evaluation settings, the dataset includes the following setups: a) the continuous GSL SD, b) the continuous GSL SI, and c) the GSL isol. In GSL SD, roughly 80% of the videos are used for training, corresponding to 8,189 instances. The rest 1,063 (10%) are kept for validation and 1,043 (10%) for testing. The selected test gloss sequences are not used in the training set, while all the individual glosses exist in the training set. In GSL SI, the recordings of one signer are left out for validation and testing (588 and 881 instances, respectively), which is approximately 14% of the total data. The rest 8821 instances are utilized for training. A similar strategy is followed in GSL isol., where the validation set consists of 2,290 gloss instances, the test set 3,500, while the remaining 34,995 are used for training.

3) Linguistic Analysis and Annotation Process:

As already mentioned, the provided annotations are both at individual gloss and sentence level. Native signers annotated and labelled individual glosses, as well as whole sentences. Sign linguists and SL professional interpreters consistently validated the annotation of the individual glosses. A great effort was devoted in determining individual glosses following the “one form one meaning” principle (i.e. a distinctive set of signs), taking into consideration the linguistic structure of the GSL and not its translation to the spoken standard modern Greek. We addressed and provided a solution for the following issues: a) compound words, b) synonyms, c) regional or stylistic variants of the same meaning, and d) agreement verbs

IV.SLR APPROACHES

In order to gain a better insight on the behavior of the various automatic SLR approaches, the best performing and the most widely adopted methods of the literature are dis- cussed in this section. The selected approaches cover all different categories of methods that have been proposed so far. The quantitative comparative evaluation of the latter will facilitate towards providing valuable insights for each SLR methodology.

A. Sub U Nets

Camgoz et. al [30] introduce a DNN-based approach for solving the simultaneous alignment and recognition problems, typically referred to as "sequence-to-sequence" learning. In particular, the overall problem is decomposed of a series of specialized systems, termed SubUNets. Each SubUNet processes the frames of the video independently. Their model follows a 2D CNN-LSTM architecture, replacing HMM with LSTM-CTC. The overall goal is to model the spatio-temporal relationships among these SubUNets to solve the task at hand. More specifically, SubUNets allow to inject domain-specific expert knowledge into the system regarding suitable intermedi- ate representations. Additionally, they also allow to implicitly perform transfer learning between different interrelated tasks.

B. GoogLeNet + TConvs

In contrast to other 2D CNN-based methods that employ HMMs, Cui et. al [28] propose a model that includes an extra temporal module (TConvs), after the feature extractor (GoogLeNet). The TConvs module consists of two 1D CNN layers and two max pooling layers. It is designed to capture the fine-grained dependencies, which exist inside a gloss (intra- gloss dependencies) between consecutive frames, into compact per-window feature vectors. The intermediate segment repre- sentations approximate the average duration of a gloss. Finally, bidirectional RNNs are applied in order to capture the context information between gloss segments. The total architecture is trained iteratively, in order to exploit the expressive capability of DNN models with limited data.

C. I3D+BLSTM

Inflated 3D ConvNet (I3D) [47] was originally developed for the task of human action recognition. Compared to 2D CNNs, 3D CNNs are able to directly learn spatiotemporal features from frame sequences. As such, its application has demonstrated outstanding performance on isolated SLR [46], [48]. In particular, the I3D architecture is an extended version of GoogLeNet, which contains several 3D convolutional layers followed by 3D max-pooling layers. The key insight of this architecture is the endowing of the 2D sub-modules (filters and pooling kernels) with an additional temporal dimension. The time dimension depends mostly on frame rate. For this reason, the stride and pooling size in are designed to be asymmetric to the spatial dimensions. This methodology makes feasible to learn spatio-temporal features from videos, while it leverages efficient known architecture designs and parameters.

V.EXPERIMENTAL EVALUATION

In order to provide a fair evaluation, we re-implemented the selected approaches and evaluated them on multiple large-scale datasets, in both isolated and continuous SLR. Re- implementations are based on the original authors' guidelines and any modifications are explicitly referenced. For the con- tinuous setup, the criteria CTC, En CTC, and EnStimCTC are evaluated in all architectures. For a fair comparison between different models, we opt to use the RGB full frame modality, since it is the common modality between selected datasets and it is more suitable for real-life applications. In addition, we conduct experiments on GSL SI and SD datasets using the depth modality (Table VII). We omit the iterative optimization process, instead we pretrain each model on the respective dataset's isolated version if present. Otherwise, extracted pseudo-alignments from other models (i.e. Phoenix) are used for isolated pretraining (imple- mentations and experimental results are publicly available to enforce reproducibility in SLR3). On the CSL SI dataset all methods, except for 3D- ResNet+BLSTM, have comparable recognition performance. They achieve high recognition accuracy due to the large size of the dataset and the small size of the vocabulary. I3D+BLSTM seems to benefit the most when trained with EnStimCTC, with 3.73% absolute WER reduction. GoogLeNet+TConvs has the best performance with 2.41% WER, with an absolute reduction is 1.65% less compared to CTC training (Fig. 4). This method outperforms the current state-of-the-art method on CSL SI [2] by an absolute WER improvement of 1.39%. On CSL SD, all models perform considerably worse compared to CSL SI, which is a challenging CSLR task since the dataset has a small combination of unique sentences (100 sentences) and the test set contains different sentences, i.e., unseen sentences, from those on the training set, with 6% of the sentences on CSL SD used for testing. I3D+BLSTM trained with EnStimCTC loss has the best performance with a WER of 60.68% on the test set.The proposed GSL dataset contains nearly double the vo- cabulary and roughly three times the number of unique gloss sentences, with less training instances. More importantly, in GSL the isolated subset draws instances from the same distri- bution as the continuous one. Fig. 7 illustrates the alignments produced by I3D+BLSTM given different pretraining schemes. It can be stated that proximal transfer learning significantly outperforms training with pseudo-alignments in this setup, both quantitatively and qualitatively. Thus, by annotating the same videos one can create larger SL datasets and more efficient CSLR systems for new SL datasets.

Figure 3: Visual comparison of ground truth alignments with the predictions of the proposed training criteria. GoogLeNet+TConvs is used for evaluation on the GSL SD dataset.

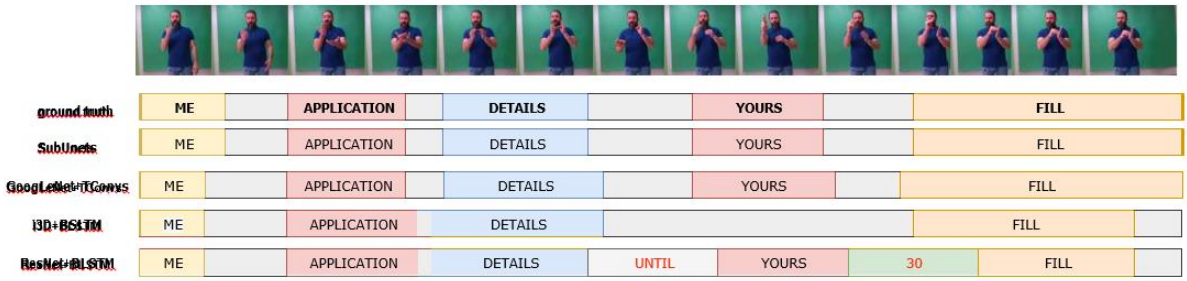

Fig. 6: Visual comparison of ground truth alignments with the predictions of each method. The methods are trained with EnStimCTC loss and are evaluated on the GSL SI dataset

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, an in-depth analysis of the most characteristic DNN-based SLR model architectures was conducted. Through extensive experiments in three publicly available datasets, a comparative evaluation of the most representative SLR architectures was presented. Alongside with this evaluation, a new publicly available large-scale RGB+D dataset was introduced for the Greek SL, suitable for SLR benchmarking. Two CTC variations known from other application fields, EnCTC & StimCTC, were evaluated for CSLR and it was noticed that their combination tackled two important issues, the ambiguous boundaries of adjacent glosses and intra-gloss dependencies. Moreover, a pretraining scheme was provided, in which transfer learning from a proximal isolated dataset can be a good initialization for CSLR training. The main finding of this work was that while 3D CNN-based architectures were more effective in isolated SLR, 2D CNN-based models with an intermediate per gloss representation achieved superior results in the majority of the CSLR datasets. In particular, our implementation of GoogLeNet+TConvs, with the proposed pretraining scheme and EnStimCTC criterion, yielded state-of- the-art results on CSL SI. Concerning future work, efficient ways for integrating depth information that will guide the feature extraction training phase can be devised. Moreover, another promising direction is to investigate the incorporation of more sequence learn- ing modules, like attention-based approaches [65], in order to adequately model inter-gloss dependencies. Future SLR architectures may be enhanced by fusing highly semantic representations that correspond to the manual and non-manual features of SL, similar to humans. Finally, it would be of great importance for the Deaf-non Deaf communication to bridge the gap between SLR and SL translation. Advancements in this domain will drive research to SL translation as well as SL to SL translation, which have not yet been thoroughly studied.

REFERENCES

1. *W. Sandler and D. Lillo-Martin, Sign language and linguistic universals. Cambridge University Press, 2006.*
2. *Z. Yang, Z. Shi, X. Shen, and Y.-W. Tai, "Sf-net: Structured feature network for continuous sign language recognition," arXiv preprint arXiv:1908.01341, 2019.*
3. *O. Koller, J. Forster, and H. Ney, "Continuous sign language recogni- tion: Towards large vocabulary statistical recognition systems handling multiple signers," Computer Vision and Image Understanding, vol. 141, pp. 108–125, 2015.*
4. *R. E. Mitchell, T. A. Young, B. BACHELDA, and M. A. Karchmer, "How many people use asl in the united states? why estimates need updating," Sign Language Studies, vol. 6, no. 3, pp. 306–335, 2006.*
5. *D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudrealt, A. Braffort,*
6. *N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef et al., "Sign language recognition, generation, and translation: An interdisciplinary perspective," arXiv preprint arXiv:1908.08597, 2019.*
7. *G. T. Papadopoulos and P. Daras, "Human action recognition using 3d reconstruction data," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 8, pp. 1807–1823, 2016.*
8. *C. Padden, "Verbs and role-shifting in american sign language," in Proceedings of the fourth national symposium on sign language research and teaching, vol. 44. National Association of the Deaf Silver Spring, MD, 1986, p. 57.*
9. *K. Emmorey, "Space on hand: The exploitation of signing space to illustrate abstract thought." 2001.*
10. *H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in Visual Analysis of Humans. Springer, 2011, pp. 539–562.*
11. *F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete, "Lsa64: an argentinian sign language dataset," in XXII Congreso Ar- gentino de Ciencias de la Computacio´n (CACIC 2016)., 2016.*
12. *M. W. Kadous et al., "Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language," in Proceedings of the Workshop on the Integration of Gesture in Language and Speech, vol. 165, 1996.*