Unraveling a Decade: A Comprehensive Survey on Isolated Sign Language Recognition

Noha Sarhan Simone Frintrop
Department of Informatics, Universität Hamburg
Hamburg, Germany
{noha.sarhan, simone.frintrop}@uni-hamburg.de

Abstract

Sign language plays a crucial role as a distinct and vital mode of communication for diverse groups of people in society. Each sign language encompasses a wide array of signs, each characterized by unique local and global articulations, e.g. hand shape, motion profile, and the arrangement of the hands, face, and body. Consequently, the domain of visual Sign Language Recognition (SLR) presents a complex and challenging research area within the field of computer vision, even with state-of-the-art models. This survey paper provides a comprehensive overview of Isolated Sign Language Recognition (ISLR), covering various aspects including input modality, modelled sign language parameters, fusion methods, and transfer learning, all of which have an impact on the performance of SLR methods. In addition, we present an overview of publicly available benchmark datasets for ISLR as well as analyze the state-of-the-art results achieved on these datasets. By examining these different aspects along with benchmarking strategies, we provide insights into the advancements, challenges, and potential directions in ISLR research.

1. Introduction

Sign language (SL) is a visual language that is used by deaf and hard-of-hearing people to communicate. It is composed of hand gestures, facial expressions, and body movements. It is estimated that there are over 70 million sign language users worldwide. Sign languages are not universal, each language has its own linguistic rules and grammatical structures, as well as having a unique vocabulary that does not necessarily have a one-to-one correspondence to spoken language. Sign languages employ multiple complementary channels to convey information [47], which can be grouped under two main categories: manual and non-manual features [3, 4].

The manual parameters refer to the hand motion, shape, orientation and place of articulation. They play a fundamental role in conveying the core lexical and grammatical information in sign language. Accurately capturing and interpreting these manual parameters is crucial for recognizing and understanding sign language gestures, as they convey the bulk of the linguistic content. Non-manual parameters encompass facial expressions, head movements, body posture, and other elements that accompany and enhance the meaning of the sign being gestured. These parameters provide valuable context, emotional nuances, and grammatical information.

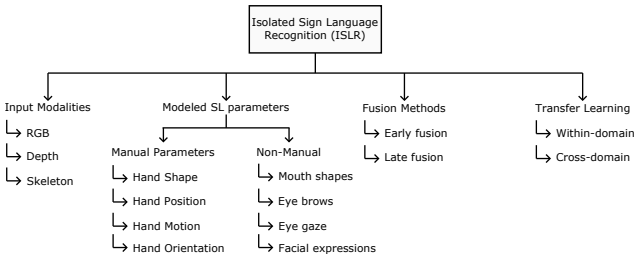

Figure 1. Taxonomy and different aspects underpinning this survey paper.

Sign language recognition (SLR) is the task of automatically recognizing the sign language signs from videos or images. SLR has a wide variety of applications, such as real-time translation, sign language education, and sign language-based human-computer-interaction. SLR can be subdivided into three different branches: 1) finger-spelling, a relatively simple task that involves a fixed set of characters for that language, usually made of static hand gestures that are presented in still images [35, 9]; 2) Isolated SLR (ISLR), a more challenging task which involves the recognition of individual signs that are performed in isolation in short video [37, 20, 19, 30]; and 3) Continuous SLR (CSLR), is the recognition of a sequence of signs that are performed in a continuous manner, thus requires the extra overhead of identifying individual signs in the sequence [6, 25, 24]. In

this paper we focus on ISLR.

In the past decade, SLR has witnessed a surge, more specifically, with the advents of deep learning and availability of public datasets [44, 1, 5]. However, the task remains challenging due to problems posed by background clutter, partial occlusion, view-point, lighting changes, execution rate and biometric variations. These challenges remain even with current deep learning approaches[33]. In addition, ISLR systems must be robust to noise and occlusion.

This survey provides a comprehensive overview of the recent advances in deep learning-based ISLR. To gather this data, we extensively review top conferences, journals, and recent ISLR challenges. We first discuss the different input modalities that have been used for ISLR, including RGB videos, depth maps, and skeletal data, or a combination thereof. We then survey the different types of sign language parameters that have been modeled by deep learning methods, including manual parameters such as hand shape and motion, and non-manual parameters such as facial expressions and head pose. Next, we review the different fusion methods that have been employed to combine information from multiple sources or modalities. Finally, we discuss the use of transfer learning techniques to improve the performance of ISLR methods. Figure 1 highlights the taxonomy underpinning this survey paper.

The choice of input modality for ISLR can have a significant impact on the performance of the recognition system. RGB videos are the most commonly used input modality for ISLR, as they provide a rich representation of the sign language gesture. However, RGB videos can be challenging to process due to the high dimensionality of the data and the presence of noise and occlusions. Depth maps and skeletal data can provide complementary information to RGB videos, and can help to improve the robustness of ISLR systems.

The performance of ISLR systems is also dependent on the type of sign language parameters that are modeled. Manual parameters such as hand shape and motion are the most commonly modeled parameters, as they are essential for conveying the meaning of sign language gestures. However, non-manual parameters such as facial expressions and head pose can also provide useful information for ISLR. For example, facial expressions can be used to convey emotions, and head pose can be used to indicate the direction of focus.

In many cases, it is beneficial to fuse information from multiple sources or modalities to improve the performance of ISLR systems. Fusion methods can be classified into three categories: early fusion, late fusion, and hybrid fusion. Early fusion methods combine the data from multiple sources at the feature level, while late fusion methods combine the data at the decision level. Hybrid fusion methods combine both early and late fusion approaches.

Figure 2. Two samples taken from ChaLearn LAP IsoGD [50] to visualize the RGB input data.

In addition, this survey explores the utilization of transfer learning techniques in ISLR. We investigate how pre-trained models or knowledge from related tasks can be leveraged to improve recognition performance, reduce the need for large annotated datasets, and accelerate the training process. We discuss various transfer learning approaches and analyze their efficacy in the context of ISLR

Benchmarking and evaluation play a critical role in advancing the state-of-the-art in ISLR. To visualise the data, a sample RGB from ChaLearn LAP IsoGD [50] is shown in Figure 2. Therefore, we present a dedicated section that provides an overview of publicly available benchmark datasets for ISLR. We discuss prominent datasets, such as AUTSL [44], WLASL [26], and BosphorusSign22k [5], and highlight their characteristics, including the number of classes, sample size, number of signers. Additionally, we present state-of-the-art results achieved on these datasets, showcasing the progress and shortcomings in ISLR.

This survey stands apart from other survey papers [33, 49] by providing:

- Insightful categorization and analysis of ISLR methods based on different input modalities, SL parameters, fusion techniques, and transfer learning; highlighting the pros and cons of each aspect.

- Comprehensive coverage of the most commonly used benchmark datasets, along with deep learning based methods developed in the last ten years, thereby providing readers with a complete overview of recent research results and state-of-the-art methods.

- Discussion of the challenges of vision-based ISLR; analysis of the limitations of available methods and discussion of potential research directions.

2. Insights into State-of-the-Art

In this section we present an overview of the number of studies on ISLR covering four aspects: input modality, modelled sign language parameters, fusion methods, and transfer learning.

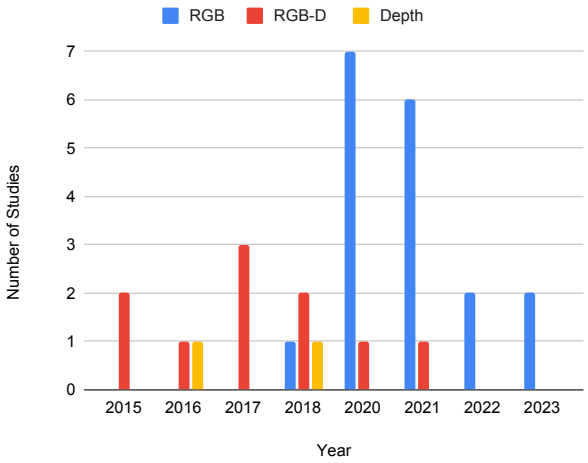

Figure 3. Number of published deep-learning based ISLR studies w.r.t. type of input modality in the past decade. Since skeletal data is extracted from either RGB or Depth data, we refrain from including in this plot as a separate modality, but reflect it in the corresponding modality from which it was extracted.

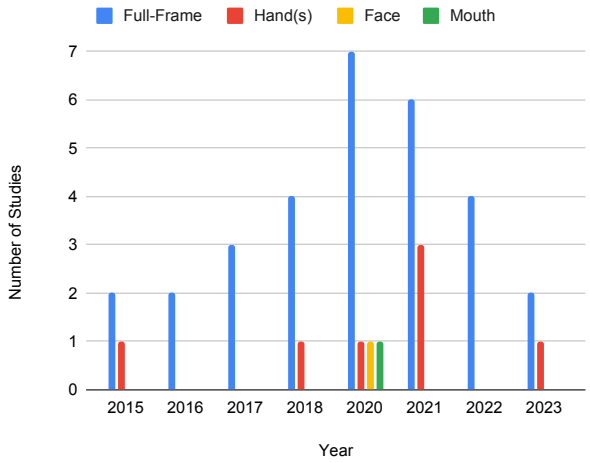

Figure 4. Number of published deep-learning based ISLR studies w.r.t. modelled sign language parameters in the past decade.

2.1. Different Input Modalities

SLR has witnessed a paradigm shift after 2005, when intrusive methods of acquisition (e.g. sensor gloves, colored gloves, etc) have been less used, and instead non-intrusive vision-based methods became more and more common (e.g. RGB, Depth) . To date, the most common input modality used in research studies is RGB video [20, 32, 18, 11, 45, 37]. RGB video provides a rich representation of the hand shapes, movements, and body postures used in sign language. However, RGB video can be sensitive to noise, occlusion, and background clutter.

Another paradigm shift took place in 2010 with the recent development of cost-effective RGB-D sensors (e.g. Microsoft Kinect and Asus Xtion), there has been growing interest in using depth data for ISLR since. This is largely because the extra dimension (depth) is insensitive to illumination changes.In addition, depth data can provide more accurate information about the 3D structure of the hand and body, which can be useful for distinguishing between similar-looking signs. Consequently, several methods based on RGB-D data have been proposed and the approach has proven to be a promising direction for SLR [20, 54, 52]. Interestingly, two studies in the last decade have relied solely on the use of depth data [53] and [51].

Another input modality that has been explored for ISLR is skeleton data [20, 2]. Skeleton data represents the positions of the joints in the body. Skeleton data can be extracted from RGB video or depth data using pose estimation algorithms, e.g. OpenPose [7]. Skeleton data is a compact representation of sign gestures, which makes it well-suited for use in mobile devices and other resource-constrained systems. SLR models working on pose data have one or two orders of magnitude fewer parameters than those that process the video directly. However, skeleton data does not provide information about the hand shapes or the 3D structure of the hand and body.

In Figure 3, we plot the number of ISLR studies that are rely on different modalities in the past decade. While depth data, RGB data, or a combination thereof has been commonly used, in the past 3 years, there is a trend of relying more on just RGB data. This is due to the limitation is that depth data is often less available than RGB video. This is particularly important when deep learning methods models pre-trained on depth-data do not exist. In addition, several sign language video data lack depth, e.g. TV broadcasts [1] and YouTube videos [30]. Accordingly, since 2020, there has been more research aiming to rely on only RGB data, evident by the work in [39, 37], and the recent ChaLearn Looking at People Challenge on ISLR in CVPR 2021 [43], which had an RGB-only track [48, 20, 18, 45, 11]. Moreover, Sarhan *et al.* [38] proposed generating pseudo depth data to mitigate this problem, while still retaining the benefits of depth data.

2.2. Modelled Sign Language Parameters

In this section, we investigate the sign language parameters and features that are extracted based on the input data. Both manual and non-manual parameters are important for the recognition of sign language. Manual parameters are essential for identifying the individual signs that are being made, while non-manual parameters are used to convey additional meaning, such as emphasis, emotion, or sarcasm. Therefore, research efforts in SLR focus on devel-

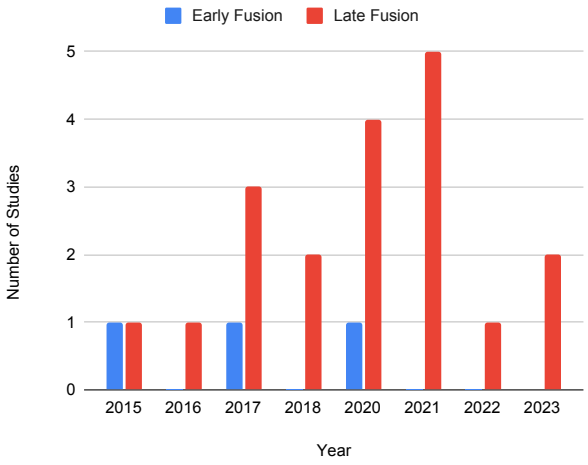

Figure 5. Number of published deep-learning based ISLR studies w.r.t. fusion method in the past decade.

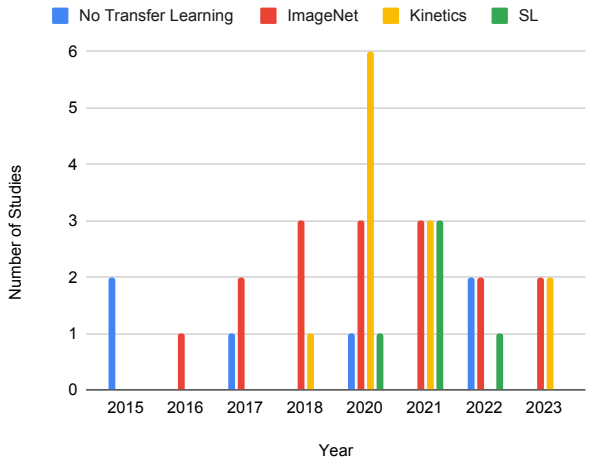

Figure 6. Number of published deep-learning based ISLR studies w.r.t. dataset used for transfer learning in the past decade. SL indicates some sign language dataset.

oping techniques that effectively capture and analyze both manual and non-manual parameters to ensure comprehensive and robust recognition of SL.

While earlier, feature extraction methods relied on manually extracting these parameters using image processing techniques, with deep learning, it has become more common to use global feature representations that are based on full-frame inputs [44, 38, 20]. However, in attempt to increase SLR accuracy, and capture the fine-grained features, there are still some methods that aim to highlight areas that focus on certain parameters. This can be done via image crops [51], e.g. hand crops, face crops, or mouth crops, or by employing some form of attention mechanism to focus the processing on relevant areas as done by [39, 37].

In Figure 4, we show the number of studies that specifically model certain parameters: full-frame, hands, face, and mouth. We observe that in the early years of using deep learning techniques for ISLR, the use of global full-frame feature was dominated, as opposed to feature extraction. However, it was quickly seen that it was not enough, and more studies started modelling other parameters, especially the hands, being the essential part to cover the manual features. In the past two years, the use of full-frames started to diminish. Using full-frame videos blew up the number of parameters used for the models. The use of other low order data, e.g. skeletal data, started to gain traction, as they resulted in lighter models, that do not require pre-training.

2.3. Fusion Methods

SLR is intrinsically multi-modal, given the various number of features/parameters that are used to represent a gesture. In order to improve the performance of ISLR systems, it is often beneficial to fuse information from multiple input modalities, as well as model different SL parameters. As a result, most recent SLR propose a multi-stream ensemble for each input type [18, 11, 20, 37], which are fused together. Fusion methods used in ISLR can be classified into categories: early fusion and late fusion. Early fusion methods combine the information from the different input modalities at an early stage of the processing pipeline. Late fusion methods combine the information form the different input modalities at a later stage of the process pipeline.

In Figure 5, we plot the number of deep learning studies that employ early fusion and those that employ late (score) fusion in the past decade. We observe that most SLR research uses late fusion (score fusion) approaches, where score probabilities of each stream is fused at the end to get one final prediction. This is probably because it has less model complexity and can achieve better run-time performance. Some methods naively average the score predictions of every stream [17, 37, 39, 20, 38], while others Gökçe *et al.* [17] use a weighted score fusion.

2.4. Transfer Learning

Transfer learning can be beneficial for ISLR because it can help overcome the problem of data scarcity. The large datasets that are used to train the initial model can provide the smaller dataset with a lot of useful information. Pretraining, a common strategy in computer vision, produces more generic feature representation and may alleviate overfitting for target tasks. For object recognition tasks, it is common to pre-train the backbone on ImageNet [12], or on Kinetics [8] for human action recognition tasks, or large web sources [14] for the downstream tasks. To date, an isolated sign language dataset that is as massive as ImageNet

or Kinetics does not yet exist. For instance, the large-scale ISLR dataset, AUTSL [44], has on average 169.6 video clips per class compared to 1200 images per class in ImageNet. Datasets for SLR have always been small due to the difficulty and expertise required for acquiring and annotating them. Even with datasets becoming larger [44, 31], they are still not large enough to train Deep CNNs from scratch. Therefore cross-domain transfer learning becomes inevitable.

In Figure 6, we observe that in the years 2016 until 2021, it was more common to rely on ImageNet. Researchers would model SLR videos as still images in order to rely on CNNs pre-trained on ImageNet, the strongest annotated dataset available at the time. Starting 2018, after the release of Kinetics dataset, researchers recent papers successfully utilized I3D CNNs pre-trained on large human action recognition datasets [36].

In the last 2 years, with the availability of larger ISLR, we start to see some within-domain transfer learning, were researchers rely on larger sign language datasets for pre-training, showing promising results as will be shown in Section 3.2.

3. Datasets and Benchmarking

In this section we present an overview of major, publicly available benchmark datasets for ISLR as well as analyze the state-of-the-art results achieved on these datasets. ISLR models are usually evaluated by one metric, accuracy. Benchmark datasets, challenges, and state-of-the-art models do not provide more metrics. Unfortunately, this is not so helpful to give further insights to the results and understanding the limitations of the proposed methods. Some datasets [21] report top-1 and top-5 instance accuracy, as well as top-1 and top-5 class accuracy. The benefits on including top-5 accuracy is that it accounts for ambiguity in the language, which could be resolved in context, just as is the case in spoken languages. Calculating per class accuracy allows to account for an unbalanced test set, and thereby better for reflecting performance than plain accuracy.

3.1. Benchmark Datasets

SLR stands as an active domain of research; however, a notable obstacle lies in the paucity of realistic large-scale sign language datasets. As a result, a majority of studies in the literature rely on training and evaluating their models with limited private or publicly accessible small-scale datasets [22, 29, 57, 56]. However, in order to train a deep learning based SLR model, the amount of training data is crucial. in recent years, larger datasets have been published [16, 43, 44, 26, 1], which contain a large vocabulary size, large number of samples, with many signers. These datasets help building practical SLR models. Although each of them has several challenges, video samples usually have a plain or simple background. This makes it difficult to develop models that can be used in daily life.

Below, we summarize the most important ISLR datasets. We refrain from including datasets that use intrusive methods, e.g. colored gloves, such as LSA64 [34], and smaller datasets that are not commonly used as benchmarks. All datasets mentioned below are signer-independent. Each signer appears only in either training, validation or test set. This is especially important because a powerful model would pick up particularities about individual persons, and recognition scores would be overly optimistic due to data leakage.

ChaLearn LAP IsoGD [50]: The ChaLearn LAP RGB-D Isolated Gesture Dataset (IsoGD) contains 47,933 RGB-D tow-modality video sequences manually labeled into 249 categories, of which 35,878 samples belong to the training set. Each RGB-D video represents one gesture instance, having 249 gesture labels performed by 21 different individuals. The IsoGD benchmark is one of the latest and largest RGB-D gesture recognition benchmarks and has a clear evaluation protocol, on which the 2016 ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge has been held.

Montalbano [16]: is a gesture dataset released by ChaLearn2014 Looking At People Challenge, which consists of 20 Italian gestures performed by 27 users. it contains 940 video sequences, each containing 10 to 20 gesture samples and around 14,000 samples in total (6,850 train, 3,454 validation, and 3,579 test samples). The videos are recorded with MS Kinect in 640×480 pixel resolutions and four types of data are provided: RGB, depth, user segmentation, and skeleton.

MS-ASL [21] is an American sign language dataset (ASL) containing a vocabulary size of 1,000, with 25,513 samples in total for training, validation and testing, respectively. It is collected from a public video sharing platform, i.e. YouTube, where many videos are performed by ASL students and teachers. The Top-100 and Top-200 most frequent words are chosen as its two subsets, referred to as MSASL100, MSASL200. Unfortunately, this dataset is no longer accessible, and has expired online.

AUTSL [44]: is one of the largest ISLR that was used in the ChaLearn Looking at People Challenge in 2021 [43]. It consists of 36,302 samples from 226 sign categories, performed by 43 signers. Variable backgrounds and multiple signers. The videos are filmed at different locations and from different viewpoints. All samples are provided as separate RGB and depth video files with a spatial resolution of 512×512 pixels and a temporal resolution of 30 frames per second (FPS). The training set contains 28,142 samples from 31 different signers, the validation set 4,418 samples from 6 different signers and the test set 3,742 samples from

Dataset	Year	Modalities	Language	Vocab	#Subjects	#Samples
ChaLearn LAP IsoGD [50]	2014	RGB, Depth	Multiple	249	21	47,933
Montalbano [16]	2014	RGB, Depth	Italian	20	27	14,000
MS-ASL [21]	2018	RGB	American	1,000	222	25,513
AUTSL [44]	2020	RGB, Depth	Turkish	226	43	38,336
WLASL2000 [26]	2020	RGB	American	2,000	119	21,097
LSE_Lex40 [13]	2020	RGB, Depth	Spanish	40	32	1,368
BosphorusSign22k [31]	2020	RGB, Depth, Skeleton	Turkish	744	6	22,542
BSL-1K [1]	2021	RGB	British	1064	40	273,000

Table 1. Statistics of publicly available ISLR benchmark datasets that are commonly used for evaluation with deep learning techniques in the past decade.

6 different signers. The samples have varying lengths, with a median of 61 frames.

WLASL [26]: Word-Level American Sign Language dataset is a large-scale ASL dataset. The videos were directly extracted from public Internet resources: educational sign language websites and ASL tutorial videos on YouTube This database is publicly available and distributed in 4 different subsets according to the number of included glosses it contains: WLASL100, WLASL300, WLASL1000 and WLASL2000. It consists of 2,000 signs performed by 119 signers and 21,083 samples. Each sign is performed by at least 3 different signers. The dataset consists of only RGB videos. It is collected from 20 different educational sign language websites that provide lookup functions for ASL signs and from ASL tutorial videos on YouTube. In the videos, signers are in a nearly-frontal view with plain background, generally wearing a black colored clothes.

LSE_Lex40 [13]: is a subset of LSE_UVIGO, a multi-source Spanish Sign Language database collected in several scenarios for ISLR and XSLR purposes. Recordings were simultaneously gathered with a high-speed Nikon D3400 and a Kinect v2. Deaf people, SL interpreters and SL students participated in the recordings under lab controlled conditions

BosphorusSign22k [31]: is another large-scale, isolated Turkish sign language dataset that contains 744 signs, 22,542 video samples in which signs belong to health and finance domains, and also cover frequently used signs in daily activities. The dataset contains 6 signers; 1 of them is reserved for testing. It is derived from BosphorusSign [5]. While the dataset is a valuable addition, it is not helpful for improving SLR tasks, where distinguishing between instances of similar sign classes with similar manual and non-manual features is essential, rather more useful for specific applications with Q&A based interaction (e.g. banking, hospital desk applications). This is due to the way the dataset is categorized (linguistically), sign glosses with the same meaning but a different set of morphemes, were considered to belong to the same class.

MultiSign-ISLR [30]: is a new sign language corpora, developed with the aim of generating a large corpus for ISLR to address the resource scarcity and create a multilingual dataset especially for pre-training purposes. We refrain from adding it to Table 1 as it is made up of both isolated, continuous and continuous isolated gestures. While the collected dataset is in RGB videos, the authors process it to extract video frames of pose points. This alleviates privacy constraints, and allows to create much lighter models.

3.2. State-of-the-art Results and Performance Benchmarks

Table 2 presents an extensive overview of state-of-the-art results attained on major benchmark ISLR datasets. For each method we highlight the four aforementioned aspects: input modality, modelled SL parameters, fusion method, and transfer learning, and the corresponding reported accuracy.

In recent research direction for ISLR, there has been a noticeable shift towards pose-based approaches. De Coster *et al.* [11] introduced pose flow, drawing inspiration from optical flow, to represent body movements based on pose keypoints. They utilized visual transformer networks to effectively capture spatial and temporal dependencies in human pose. Similarly, Li *et al.* [26] presented pose-based temporal graph convolution networks to model spatial and temporal dependencies in human pose. Other works that solely base on pose or skeletal data include [2] and [48]. These pose-based methodologies showcase the growing interest and potential of using pose information for enhancing ISLR systems.

Sincan and Keles [45] proposed an innovative approach leveraging RGB motion history images (MHI) to condensely summarize entire sign language videos into single frames. Their model effectively captures relevant spatial and motion patterns from these images, employing motion-based attention mechanisms to focus on pertinent spatial regions. Furthermore, they proposed a fusion model that combines RGB and RGB-MHI features, enhancing the representation of sign language gestures

In a multi-modal approach, Gökçe *et al.* [17] utilized

Dataset	Method	Modality	SL Params	Fusion	TL	Accuracy
ChaLearn LAP IsoGD	Sceneflow+CNN [52]	RGB, D	FF	Early,Score	ImageNet	36.27 %
	AMRL [53]	Depth	FF	-	ImageNet	39.23 %
	DDI+CNN [51]	Depth	FF+Hands	-	NTU RGB-D [41]	43.72 %
	Cooperative CNN [54]	RGB, D	FF	None	ImageNet	44.80 %
	xDETVP-TRIMPS [58]	RGB, D	FF	Score	UCF-101 [46]	45.02 %
	2SCVN-3DDSN [15]	RGB, D	FF	Score	ImageNet	49.17 %
	C3D [28]	RGB, D	FF	Score	ImageNet	49.20 %
	C3D+ConvLSTM [59]	RGB, D	FF	Score	Scratch	51.02 %
	I3D-SLR [36]	RGB	FF	Score	ImageNet,Kinetics	62.09%
	Attn-I3D (hybrid) [37]	RGB	FF	Score	ImageNet,Kinetics	65.02%
	TD-SLR [39]	RGB	FF,Hands	Score	ImageNet,Kinetics	70.91%
MS-ASL 1000	SignBERT [19]	Pose	Hands	Score	SL	57.06%
	Baseline-I3D [21]	RGB	FF	Score	ImageNet,Kinetics	57.69%
	BSL [1]	RGB,Pose	FF+Mouth	Score	Kinetics	61.55%
	SignBERT [19]	RGB	Hands	Score	SL	67.96%
AUTSL	Baseline [44]	RGB, D	FF	Score	ImageNet	62.02 %
	Baseline [44]	RGB	FF	Score	ImageNet	49.22 %
	S3D [48]	RGB	FF	None	Kinetics,SL	90.27 %
	VTN-PF [11]	RGB,PF	FF+Hands	-	ImageNet	92.92 %
	RGB-MHI [45]	RGB	FF	Weighted Score	Kinetics	93.53 %
	VLE-trans [18]	RGB	FF+Hands	Weighted Score	ImageNet	95.46 %
	MS-G3D [48]	RGB, Pose	FF	Weighted Score	Kinetics,SL	96.51 %
	TD-SLR [39]	RGB	FF	Score	ImageNet, Kinetics	97.93 %
	SAM-SLR [20]	RGB,Pose	FF	Score	Kinetics,SL	98.42 %
	SAM-SLR [20]	RGB,D	FF	Score	Kinetics,SL	98.53 %
WLASL 100	Pose-TGCN [26]	Pose	FF	-	-	55.43%
	SPOTER [2]	Pose	FF	-	-	63.18%
	I3D [26]	RGB	FF	-	ImageNet,Kinetics	65.89%
	TCK [27]	RGB	FF	-	Kinetics	77.52%
	SignBERT [19]	Pose	Hands	Score	SL	79.07%
	SignBERT [19]	RGB	FF+Hands	Score	SL	82.56%
WLASL 300	Pose-TGCN [26]	Pose	FF	-	-	38.32%
	SPOTER [2]	Pose	FF	-	-	43.78%
	I3D [26]	RGB	FF	-	ImageNet,Kinetics	56.14%
	TCK [27]	RGB	FF	-	Kinetics	68.56%
	SignBERT [19]	Pose	Hands	Score	SL	70.36%
	SignBERT [19]	RGB	FF+Hands	Score	SL	74.40%
WLASL 2000	Pose-TGCN [26]	Pose	FF	-	-	23.65%
	I3D [26]	RGB	FF	-	ImageNet,Kinetics	32.48%
	BSL [1]	RGB,Pose	FF+Mouth	Score	Kinetics	44.72%
	SignBERT [19]	Pose	Hands	Score	SL	45.17%
	SignBERT [19]	RGB	FF+Hands	Score	SL	52.08%
Bosphorus Sign2k	3D ResNet [31]	RGB	Full-Frame	Score	Kinetics	78.85%
	MC3-18 [17]	RGB	Full-Frame only	Weighted Score	Kinetics	86.91%
	RGB-MHI [45]	RGB	Full-Frame	Score	AUTSL, ImageNet	94.83 %
	MC3-18 [17]	RGB	FF+Hand+Face	Weighted Score	Kinetics	94.94%

Table 2. Performance comparison for different methods on commonly used RGB-D datasets in the past decade sorted by accuracy for each dataset. The column TL mentions the dataset used for transfer learning, where SL means some sign language dataset. D denotes Depth. FF denotes full-frame.

OpenPose to extract face and hand regions from sign language videos and used these modalities in conjunction with full-body images. Additionally, they separately cropped each hand and employed hand crops collectively in their analysis. This integration of various image regions significantly improved the understanding and recognition of sign language gestures.

As for fusion techniques, Wang *et al.* [52] explored early fusion, combining extracted features from both depth and RGB modalities as a joint entity to create scene flow images. This strategy effectively leveraged complementary information from both modalities enhancing the representation of sign language gestures, thereby improving the overall recognition performance. The authors in [23] fused RGB and pose information, and model isolated SL videos using a a skeleton heatmap-based feature.

Regarding transfer learning, Wang *et al.* represented the data as scene flow images in 2D to benefit from pre-trained models on ImageNet.

Vázquez-Enríquez *et al.* [48] conducted experiments with pre-training models on different datasets. They showed that pre-training on a small dataset, such as WLASL200 or LSE_Lex40, and fine-tuning on a larger dataset like AUTSL did not significantly improve performance, though it led to faster convergence. However, pre-training on a large SLR dataset, specifically AUTSL, greatly benefited results when fine-tuned on smaller datasets, like LSE_Lex40.

4. Conclusion and Future Prospects

The domain of visual SLR presents a complex and challenging research area within the realm of computer vision, even with the use of state-of-the-art models. Through this comprehensive survey paper, we have provided a detailed overview of ISLR, delving into critical aspects such as input modality, modelled sign language parameters, fusion methods, and transfer learning, all of which significantly impact the performance of SLR methods.

In recent research, there is a noticeable trend towards skeleton-based methods, following the progress in human action recognition with spatial-temporal Graph Convolutional Networks (GCN) [55, 42]. While the field of ISLR has taken inspiration from these advancements [10, 48], it remains in its early stages, leaving ample room for exploration and innovation. Additionally, a new direction focuses on depth estimation methods, aiming to reduce reliance on specific acquisition methods. With the hope for light-weight skeleton-based models, and simple acquisition methods, the prospect of having accessible SLR on mobile phones seems promising.

Looking ahead, the adoption of within-domain transfer learning holds great promise for enhancing ISLR performance. Such transfer learning strategy, akin to those utilized in human action recognition, continue to evolve, offering exciting prospects for bridging the gap between different sign language recognition and improving overall recognition accuracy. In addition, techniques such as self-supervised learning and multilingual fine-tuning have proven effective in addressing low-resource data scenarios in natural language and speech processing domains. These techniques can be leveraged in ISLR, especially given that the datasets listed in Table 1 are considered low-resource. For instance, studies like those by Hu *et al.* [19] and Selvaraj *et al.* [40] delve into self-supervised training for ISLR, while NC *et al.* [30] provide a large, 10-language corpus that could serve as a pre-training dataset for multi-lingual fine-tuning.

Embracing these emerging techniques and further delving into the potential of self-supervised learning and multilingual fine-tuning can open new doors for future ISLR research. By incorporating these strategies, researchers can build more robust and accurate sign language recognition systems, enhancing communication and accessibility for the deaf and hard-of-hearing communities.

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, pages 35–53. Springer, 2020.

[2] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer or word-level sign language recognition. In *WACV*, pages 182–191, 2022.

[3] P Boyes Braem and RL Sutton-Spence. *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Press, 2001.

[4] Diane Brentari. *Sign language phonology*. Cambridge University Press, 2019.

[5] Necati Cihan Camgöz, Ahmet Alp Kındıroğlu, Serpil Karabüklü, Meltem Kelepir, Ayşe Sumru Özsoy, and Lale Akarun. Bosphorussign: A Turkish sign language recognition corpus in health and finance domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1383–1388, 2016.

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, pages 10023–10033, 2020.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017.

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[9] Djamila Dahmani and Slimane Larabi. User-independent system for sign language finger spelling recognition. *Jour-*

nal of Visual Communication and Image Representation, 25(5):1240–1250, 2014.

[10] Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. Spatial-temporal graph convolutional networks for sign language recognition. In *International Conference on Artificial Neural Networks*, pages 646–657. Springer, 2019.

[11] M. De Coster, M. Van Herreweghe, and J. Dambre. Isolated sign recognition from RGB video using pose flow and self-attention. In *CVPRW*, pages 3441–3450, 2021.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[13] Laura Docío-Fernández, José Luis Alba-Castro, Soledad Torres-Guijarro, Eduardo Rodríguez-Banga, Manuel Rey-Area, Ania Pérez-Pérez, Sonia Rico-Alonso, and Carmen García Mateo. LSE_UVIGO: A multi-source database for Spanish sign language recognition. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 45–52, 2020.

[14] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, pages 670–688. Springer, 2020.

[15] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Z. Li. A unified framework for multi-modal isolated gesture recognition. *TOMM*, 14(1s):1–16, 2018.

[16] Sergio Escalera, Xavier Baró, Jordi Gonzalez, Miguel A Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J Escalante, Jamie Shotton, and Isabelle Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *ECCVW*, pages 459–473. Springer, 2015.

[17] Çağrı Gökçe, Oğulcan Özdemir, Ahmet Alp Kındıroğlu, and Lale Akarun. Score-level multi cue fusion for sign language recognition. In *ECCVW*, pages 294–309. Springer, 2020.

[18] I. Gruber, Z. Krnoul, M. Hrúz, J. Kanis, and M. Bohacek. Mutual support of data modalities in the task of sign language recognition. In *CVPRW*, pages 3424–3433, 2021.

[19] Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. SignBERT: pre-training of hand-model-aware representation for sign language recognition. In *ICCV*, pages 11087–11096, 2021.

[20] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu. Skeleton aware multi-modal sign language recognition. In *CVPRW*, pages 3413–3423, 2021.

[21] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A large-scale data set and benchmark for understanding American sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[22] Tomasz Kapuscinski, Mariusz Oszust, Marian Wysocki, and Dawid Warchol. Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, 12(4):36, 2015.

[23] Ahmet Alp Kındıroglu, Oğulcan Özdemir, and Lale Akarun. Aligning accumulative representations for sign language recognition. *Machine Vision and Applications*, 34(1):12, 2023.

[24] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[25] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *BMVC*, 2016.

[26] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*, pages 1459–1469, 2020.

[27] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. Transferring cross-domain knowledge for video sign language recognition. In *CVPR*, pages 6205–6214, 2020.

[28] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2956–2964, 2017.

[29] Kian Ming Lim, Alan Wee Chiat Tan, Chin Poo Lee, and Shing Chiang Tan. Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, 78:19917–19944, 2019.

[30] Gokul NC, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh Khapra. Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. *NIPS*, 35:36202–36215, 2022.

[31] Oğulcan Özdemir, Ahmet Alp Kındıroğlu, Necati Cihan Camgöz, and Lale Akarun. BosphorusSign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283*, 2020.

[32] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *ECCVW 2014*, pages 572–578. Springer, 2015.

[33] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021.

[34] Franco Ronchetti, Facundo Quiroga, César Armando Estrebou, Laura Cristina Lanzarini, and Alejandro Rosete. Lsa64: An Argentinian sign language dataset. In *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016).*, 2016.

[35] Supawadee Saengsri, Vit Niennattrakul, and Chotirat Ann Ratanamahatana. TFRS: Thai finger-spelling sign language recognition system. In *2012 second international conference on digital information and communication technology and it's applications (DICTAP)*, pages 457–462. IEEE, 2012.

[36] Noha Sarhan and Simone Frintrop. Transfer learning for videos: from action recognition to sign language recognition. In *ICIP*, pages 1811–1815. IEEE, 2020.

[37] Noha Sarhan and Simone Frintrop. Sign, attend and tell: Spatial attention for sign language recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[38] Noha Sarhan, Jan M. Willruth, and Simone Frintrop. PseudoDepth-SLR: Generating depth data for sign language recognition. In *Computer Vision Systems: 14th International Conference, ICVS 2023*.

[39] Noha Sarhan, Christian Wilms, Vanessa Closius, Ulf Brefeld, and Simone Frintrop. Hands in focus: sign language recognition via top-down attention. In *ICIP*. IEEE, 2023.

[40] Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. *arXiv preprint arXiv:2110.05877*, 2021.

[41] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, pages 1010–1019, 2016.

[42] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019.

[43] Ozge Mercanoglu Sincan, Julio Junior, CS Jacques, Sergio Escalera, and Hacer Yalim Keles. Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In *CVPR*, pages 3472–3481, 2021.

[44] O. M. Sincan and H. Y. Keles. AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[45] O. M. Sincan and H. Y. Keles. Using motion history images with 3D convolutional networks in isolated sign language recognition. *IEEE Access*, 10:18608–18618, 2022.

[46] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[47] Clayton Valli and Ceil Lucas. *Linguistics of American sign language: An introduction*. Gallaudet University Press, 2000.

[48] M. Vazquez-Enriquez, J. L. Alba-Castro, L. Docío-Fernández, and E. Rodriguez-Banga. Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In *CVPRW*, pages 3462–3471, 2021.

[49] Ankita Wadhawan and Parteek Kumar. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28:785–813, 2021.

[50] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In *CVPRW*, pages 56–64, 2016.

[51] Pichao Wang, Wanqing Li, Zhimin Gao, Chang Tang, and Philip O Ogunbona. Depth pooling based large-scale 3-D action recognition with convolutional neural networks. *IEEE Transactions on Multimedia*, 20(5):1051–1061, 2018.

[52] Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, and Philip Ogunbona. Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In *CVPR*, pages 595–604, 2017.

[53] Pichao Wang, Wanqing Li, Song Liu, Zhimin Gao, Chang Tang, and Philip Ogunbona. Large-scale isolated gesture recognition using convolutional neural networks. In *ICPR*, pages 7–12. IEEE, 2016.

[54] Pichao Wang, Wanqing Li, Jun Wan, Philip Ogunbona, and Xinwang Liu. Cooperative training of deep aggregation networks for RGB-D action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[55] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[56] Quan Yang. Chinese sign language recognition based on video sequence appearance modeling. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542. IEEE, 2010.

[57] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 279–286, 2011.

[58] Guangming Zhu, Liang Zhang, Lin Mei, Jie Shao, Juan Song, and Peiyi Shen. Large-scale isolated gesture recognition using pyramidal 3D convolutional networks. In *ICPR*, pages 19–24. IEEE, 2016.

[59] Guangming Zhu, Liang Zhang, Peiyi Shen, and Juan Song. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access*, 5:4517–4524, 2017.