IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.DOI

Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues

MUHAMMAD AL-QURISHI, (Member, IEEE), THARIQ KHALID, AND RIAD SOUISSI.
Research Department, Research and Innovation Division, Elm Company, Riyadh, KSA

Corresponding author: Muhammad Al-Qurishi (e-mail: mualqurishi@elm.sa).

This work was supported by the Research Department in Elm Company

ABSTRACT People with hearing impairments are found worldwide; therefore, the development of effective local level sign language recognition (SLR) tools is essential. We conducted a comprehensive review of automated sign language recognition based on machine/deep learning methods and techniques published between 2014 and 2021 and concluded that the current methods require conceptual classification to interpret all available data correctly. Thus, we turned our attention to elements that are common to almost all sign language recognition methodologies. This paper discusses their relative strengths and weaknesses, and we propose a general framework for researchers. This study also indicates that input modalities bear great significance in this field; it appears that recognition based on a combination of data sources, including vision-based and sensor-based channels, is superior to a unimodal analysis. In addition, recent advances have allowed researchers to move from simple recognition of sign language characters and words towards the capacity to translate continuous sign language communication with minimal delay. Many of the presented models are relatively effective for a range of tasks, but none currently possess the necessary generalization potential for commercial deployment. However, the pace of research is encouraging, and further progress is expected if specific difficulties are resolved.

INDEX TERMS Sign Language , Deep learning, Continuous Model, Machine learning, Pose Estimation

I. INTRODUCTION

FOR millions of people, sign language communication is the primary means of interacting with the world, and it is not difficult to imagine the potential applications involving effective sign language recognition (SLR) tools [1], [2]. For example, we could translate broadcasts that include sign language, create devices that react to sign language commands, or even design advanced systems to assist impaired people in conducting routine jobs. In particular, deep neural networks (DNNs) have emerged as a potentially groundbreaking asset for researchers, and the full impact of their application to the problem of SLR will likely be felt in the near future [3], [4]. SLR is a field dedicated to the automated interpretation of hand gestures and other signs used in communications between people with a speech or hearing impairment. Because hardware and software components have evolved to the point where developing advanced systems with real-time translation capacities appear to be within reach, a large number of exciting and innovative solutions have been proposed and tested in recent years [5]–[9]with the objective of building fully functional systems that can understand sign language and respond to commands given in this format. However, before any truly practical applications can be considered, it is imperative to perfect the interpretation algorithms to the point where false positives are rare [6], [10]–[13].

Owing to the numerous challenges inherent in this task, at this stage, it is not yet possible to design SLR tools that approach 100% accuracy on a large vocabulary [14], [15]. Thus, it is very important to continue developing new methods and evaluate their relative merits, gradually arriving at increasingly reliable solutions. While most researchers agree that deep learning models are the most suitable approach, the optimal network architecture remains a point of contention, with several competing designs achieving promising results. Detailed experimental evaluations are the only way to identify the best performing algorithms and refine these further using discoveries from other research teams when applicable. As most countries use their own variations of sign language, much of the research is conducted locally with persons skilled in using regional signs. With this in mind,

it is not surprising that a large number of scientific papers are targeting SLR problems and that the performance level of the proposed solutions is rapidly increasing from year to year [16], [17].

In the current literature, the various SLR solutions can essentially be divided into two major groups, depending on the primary data collection method. One group of methods relies on external sensors to gather insights regarding the actions of the signer, for example, through data gloves worn by the signer. Starner et al. [18] provided early example of a system based on wearable sensors, while many other authors have exploited this concept since then. However, there are practical considerations regarding sensor-based techniques, and therefore a majority of recent research has been directed toward vision-based methods, which rely on images, video, and depth data to determine the semantic content of hand signs. For example, Chen et al. [19] pioneered a hand gesture recognition method based on skin color, while many alternative techniques have since been proposed, some of which are based on filtering principles [20].

In particular, the commercial launch of the Microsoft Kinect device has unlocked a completely new level of insight [21]–[23] ,and researchers are still exploring how to leverage the power of depth vision to develop more accurate SLR tools. In terms of the type of neural network most suitable for SLR purposes, the convolutional neural network (CNN) model [24] was one of the first to gain major attention [25]–[28]. In addition to CNNs, other architectures such as hidden Markov models (HMMs) [19] and recurrent neural networks (RNNs) are frequently applied [29]. The support vector machine (SVM) model is frequently used for this purpose as well [30], [31],while random forest (RF) and K-nearest neighbor (k-NN) are sometimes chosen for the classification task [29], [32]. We summarize our work contributions in this paper as follows:

1) Comprehensive review and taxonomy of automated sign language recognition (ASLR) literature: We conducted a comprehensive review of automated sign language recognition using machine/deep learning methods and techniques published between 2014 and 2021. We concluded that several SLR methods currently in existence require some conceptual classification to make sense of all available data. Thus, we focus on elements that are common to almost all sign language recognition methodologies and discuss their relative strengths and weaknesses regarding specific SLR tasks and functionalities as part of this study.

2) Establishment of a general framework for creating SLR models: We propose a general framework based on the challenges and limitations we have identified in the literature. At this point, the value of machine learning/deep learning (ML/DL) methodologies for sign language recognition is beyond question, although discussions regarding the most promising directions of research continue. There is consensus that deeper models hold more promise for the eventual development of real-life SLR applications than traditional machine learning approaches, but at present, even the most sophisticated models fall considerably short of the necessary reliability.

3) Benchmark datasets and performance: An analysis of the benchmark datasets and performance used in the literature is conducted. The quality of available sign language datasets is essential for ensuring that SLR tools built and tested with them return relevant predictions. However, the availability of high-quality datasets of this kind is limited, and in some cases barely sufficient for serious testing. Some of the datasets mentioned in literature include the Corpus VGT consisting of over 140 hours of video input and including approximately 100 classes, PHOENIX14T dataset with video recordings of 9 different signers using more than 1000 unique signs, PHOENIX-Weather2014T with vocabulary related to weather, and ASLG-PC12 which includes various English-language versions of signs. Datasets are usually split into training, validation, and testing portions, so the models can be evaluated with the same type of input that was used to optimize them. However, due to different datasets used in different studies, direct comparison of the results across studies is not possible.

4) Identifying open Issues and challenges: After analyzing and discussing the existing methodologies, we draw some conclusions with respect to their limitations, open issues, and potential challenges. Differences between regional variations of sign language alphabets and vocabularies greatly complicate cross-border collaboration, especially considering the scarcity of high-quality datasets for languages with smaller numbers of speakers. This also makes it very difficult to develop and test more advanced applications, which require much larger training vocabularies. Most of the proposed methods are conceptually sound, yet they lack the level of accuracy and reliability that would be desired for a final solution. These problems are exacerbated in the continuous SLR sub-field, where semantic content is far more complex and thus more difficult to capture through statistical analysis.

The remainder of this paper is organized as follows. In Section II, we provide a brief background regarding some of the basic concepts discussed in this paper, such as deep learning, machine learning. Section III presents the review method used in this study. Machine learning and deep learning methods to design sign language recognition models are discussed in detail in Section IV along with the proposed framework. Types of models and languages related to the recognition process are discussed in Section V and Section VI, respectively. The related studies and surveys have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

been discussed in Section VII. Section VIII introduces the benchmark SLR datasets used for ML/DL and provides a comparative analysis of the ML/DL methods performance for sign language recognition. Section IX discusses open issues, challenges, and opportunities for future research. Finally, the conclusions of our study are presented in Section X.

II. BACKGROUND

In recent years, there have been ongoing efforts to develop automated methods for the completion of numerous linguistic tasks using advanced algorithms that can 'learn' based on past experience [33]. Sign language recognition (SLR) is an area where automation can provide tangible benefits and improve the quality of life for a significant number of people who rely on sign language to communicate on a daily basis [34]. The successful introduction of such capabilities would allow for the creation of a wide array of specialized services, but it is paramount that automated SLR tools are sufficiently accurate to avoid creating confusing or dysfunctional responses. In this section, we provide a brief background regarding some important approaches that have been utilized for automated SLR.

A. MACHINE LEARNING (ML)

The machine learning concept encompasses a number of stochastic procedures that can be used to predict the value of a certain parameter based on similar examples that the algorithm was previously exposed to. A simple example, illustrated by Algorithm 1, shows how a general formalization of the learning process takes place. There are many different methodologies that belong to this group; some of the best-known methods include naïve Bayes, random forest, K-nearest neighbor, logistic regression, and the support vector machine [33], [35]. All of these methods undergo a training phase, which can be either supervised (using labeled input data) or unsupervised (without labeled data) , and use input features to establish connections among variables and acquire predictive power. However, owing to their simplicity, such methods have limitations when there is a need to capture nuanced semantic hints, as is the case with most linguistic tasks. On the other hand, they can often provide the foundation for the development of more powerful analytic tools and serve as a measuring stick to evaluate progress.

Algorithm 1: LEARNING PROCESS

Input: x, is a d dimensional vector of features
Output: y, is the output decision

1: Target function $f : X \Rightarrow Y$ the ideal formula (Unknown)
2: Data: $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ training examples
3: Hypothesis $g : X \Rightarrow Y$ formula to be used
4: Learning algorithm $g \approx f$ final hypothesis

Machine learning techniques are used to aid in sign language recognition and have achieved some degree of success. Some of the earliest studies in this field were based on data input from wearable sensors, which provide a very direct translation of a user's movements. The data can be filtered using techniques such as SVM to provide a reasonably accurate recognition of the intended sign. Some of the aforementioned machine learning methods are used primarily to analyze static content (i.e., individual signs isolated in time and space), while in some cases, there have been attempts to interpret continuous segments of sign language speech, necessitating the use of dynamic models such as dynamic time warping or relevance vector machines. In general, basic stochastic models are better suited for simple SLR tasks, which is why they were extensively used in the early stages of research. These statistical models typically require less computing power than more complex architectures, although this depends on the number of analyzed features as well as the size of the dataset. As more complex ASLR applications naturally require the inclusion of additional variables and sometimes additional modalities, the simplicity of basic models remains attractive. Thus, simpler machine learning methods remain valuable tools and often serve as comparison benchmarks that can be used to evaluate the properties of newly proposed methods.

B. DEEP LEARNING

Recently, basic machine learning approaches have been largely replaced with deeper architectures that employ several layers and pass information in vector format between layers, gradually refining the estimation until positive recognition is achieved. Such algorithms are usually described as "deep learning" systems or deep neural networks, and they operate on principles similar to the machine learning strategies described above, although with far greater complexity. Based on the structure of the network, two architectures are commonly used for a number of different tasks: recurrent neural networks (RNNs) that include at least one recurrent layer, and convolutional neural networks (CNNs) that include at least one convolutional layer. Depending on the number and type of layers, these networks can exhibit different properties and are generally suitable for different types of tasks, while the training phase decisively impacts the performance of the algorithm. The general rule is that larger and more specific datasets allow for more robust network training, and therefore the quality of the training set is an important factor. Additional fine-tuning of a model can usually be achieved by changing some of the relevant hyper-parameters that define the training procedure [36].

The majority of research involving the automation of SLR tasks is currently based on methods that rely on a combination of images and depth data, which generate a tremendous amount of information that often requires analysis in real time (or at least taking the temporal dimension into account). With larger and more diverse datasets, simple machine learning methods tend to underperform, which is why many of the more sophisticated models are based either on RNN or

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

CNN design. Deep networks can be trained using multimodal input (e.g., skeletal data combined with depth images from Microsoft Kinect), and in some applications, they can achieve a recognition accuracy of over 98% under optimal conditions. The advantages of deep learning were demonstrated by Konstantinidis et al. [37]], who successfully used data from disparate sources to identify sign language words in isolated form, although their model displayed uneven performance depending on the dataset used. More demanding SLR tasks, such as interpretation of continuous speech or real-time translation, require even more sophisticated models, which in some cases require an increased number of layers (depth). While deep models appear to be a safe choice for the role of empowering automated SLR applications in the future, it remains to be seen whether the current architectures will survive in their present form or will evolve into new models that can 'understand' the semantic aspects of sign communication more astutely. Possible models that could be more widely used in the future include deep belief networks with a very large number of layers, as well as networks based on autoencoders.

III. REVIEW METHOD

In this study, we summarize and organize scientific data about the subject of Sign Language Recognition (SLR) for the benefit of the entire research community. In order to assist anyone interested in the fundamental knowledge in this field, we complemented the basic facts about each study with an impartial assessment of its quality and potential for positive contributions. We attempt to answer the following main research questions:

Question 1 – Which studies have been conducted addressing automated Sign Language Recognition, and what are the available datasets?

Question 2– What techniques in Automatic Sign Language Recognition for various languages are applied to date?

Question 3 – Which challenges remain unsolved in this scientific field?

One of the ultimate objectives of this paper is to lay the groundwork for future inquiries about SLR and clarify any ambiguous elements that might confuse some researchers. We accomplished this in three phases – preparatory, execution and presentation, with each stage including several steps. These steps included 1) selecting the most relevant research questions, 2) setting fundamental rules for the evaluation procedure, 3) formalizing the selection threshold, 4) assessing the quality of the work's premises and results, 4) looking into the methodological setup of the experiments, and 5) extracting any bits of information that contribute to answering the central questions.

A. REVIEW PROTOCOL

We followed a defined procedure during the literature review, allowing for a more objective evaluation of the paper content. This procedure consisted of numerous tasks, starting with selecting relevant variables, isolating the authors' strategic approaches, analyzing the methods and techniques used to obtain the results, sorting out quantitative output, and defining the principles for generalization and summarization.

B. INCLUSION AND EXCLUSION CRITERIA

During the collection of scientific works, a set of well-defined parameters were used to decide which works to include. Since the subject of this paper is SLR, only papers from this field were taken into consideration. The period covered is between 2014 and 2021, as shown in Fig 1, as the idea was to provide a systematization of contemporary research. In the following Table 1 , we provide a complete set of rules for selecting the research papers in a succinct format.

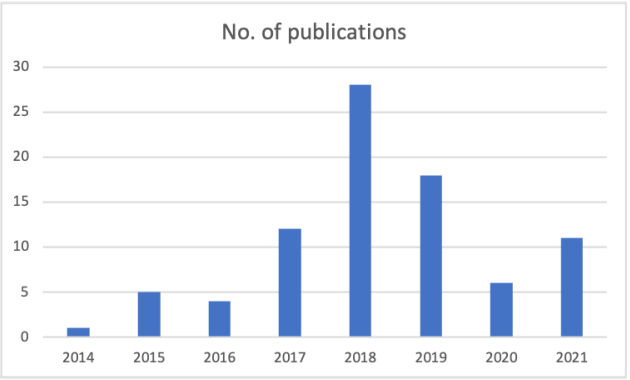


FIGURE 1: Number of publications on Sign Language Recognition by year

C. SEARCH STRATEGY

Finding the most relevant research material required an arduous process combing publicly available sources using a combination of automated tools and human workforce. Specific keywords drove the automated segment, which are displayed in Table 2.

This collection of studies is continually expanded through addition of individual papers that match the same level of relevance as those found by the algorithm. We included all of the most significant online repositories of scientific content in the search, from Google Scholar, MDPI, Springer, Elsevier, and IEEE explore to ACM and ArXiv. The proportion of papers from each source is shown in Fig 2.

The overall objective at this stage was to discover as many works that address the topic of SLR as possible. After completing this stage, we carefully analyzed the entire corpus of collected material using the forward/back technique. This allowed for a more detailed understanding of each paper, with the ability to track all references and follow the significant lines of research. In this way, it was possible to ensure that no foundational studies are missing from the study and that the final collection of SLR papers is truly representative of the most successful research directions.We then processed the collection based on the Mendeley method, which made it possible to easily identify and remove identical items

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 1: Inclusion and exclusion criteria for SLR studies.

Task	Criteria
Included Articles	• SLR papers in any native languages • SLR papers addressing alphabets, words and sentences • SLR papers with Depth cameras • SLR papers with open datasets • Published from the year 2014 onwards
Excluded Articles	• Not relevant to research questions • Incomplete or inaccessible text • Published before 2014 • Written in language other than English • Duplication

TABLE 2: Keywords for Searching Stage.

Group 1	Group 2
Sign language	Deep learning
Sign language recognition	Machine learning
Sign language translation	Pose estimation
Automated sign language	Hand gesture
	Transformer

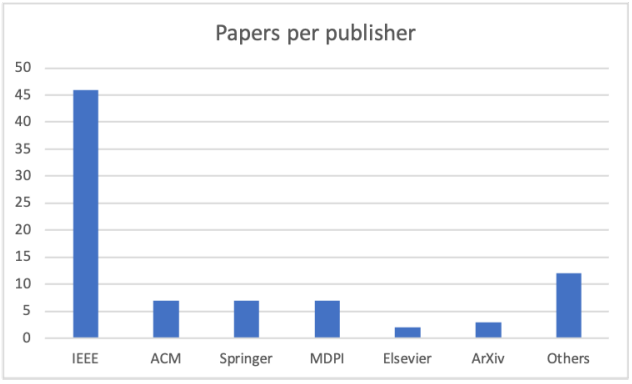


FIGURE 2: Number of publications on Sign Language Recognition by Publisher

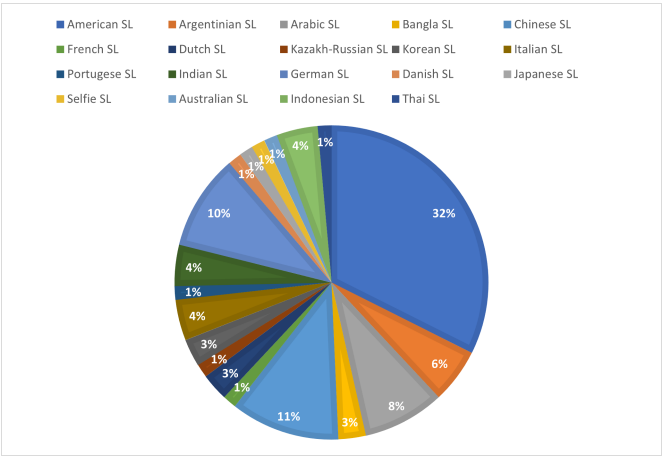


FIGURE 3: Number of publications on Sign Language Recognition by Language

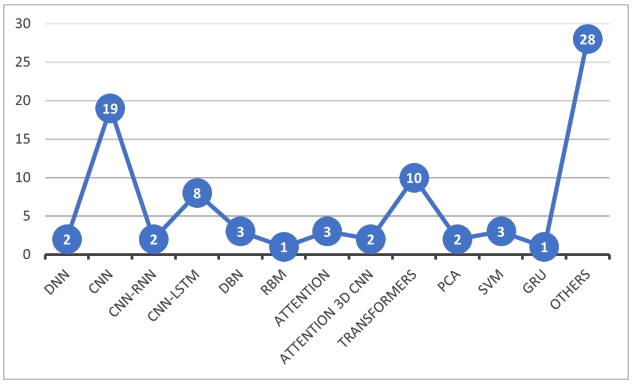


FIGURE 4: Number of publications on Sign Language Recognition by Architecture

from the list, making the content more readily searchable. We noted several trends in this part of the process, which included a breakdown of collected works based on the local variation of sign language they refer to. A majority of works in the collection (more than 30%) were related to the American variation, but French, Argentinian, Arabic, and many other SL variations are also represented, as seen in Fig. 3.

Another factor that was used to differentiate between papers is the type of architecture of the proposed solution. Full overview is available in Fig. 4.

D. STUDY SELECTION PROCESS

During the initial search, we found 196 different papers, although 11 of them were duplicates that we immediately disqualified. All original papers were reviewed using the principles outlined in Table5 and information available on the first page of the paper. In this manner, we removed all works not connected to the research field, collected from unreliable sources, or with other weaknesses. Examination of this kind identified 47 entries that did not meet the inclusion criteria; 138 core and relevant studies remained.

We examined the full text of the studies next, and removed any that failed to directly address SLR or to support their hypothesis with high-quality data removed as well. Next, we took the quality of all quotes and correct naming of sources into account, and performed online checks to ascertain the authorship of source papers. In the last phase, we performed a qualitative evaluation to determine which studies deserved to be reported on. The entire selection procedure cut down the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

number of included works to 84, but their level of scientific value and importance regarding the main research questions leaves nothing to be desired.

IV. AUTOMATED SIGN LANGUAGE RECOGNITION FRAMEWORK

Most automated SLR research is concerned with similar problems, namely the need to interpret hand and body movements associated with sign language characters in a clear and unambiguous manner. Because the main objectives are similar, the studies in this area also share similar methodology, even if their procedures may not be identical. Figure 5 presents the general model shared among the majority of researchers in this area. The input layer of the solution consists of an input device based on SLR data collection methods, as shown in Figure 6, and includes a visual display to present hand signs. The second layer is the pre-processing layer that performs gesture data filtering and can decode a sign into the required data format. In some cases, there are additional steps, such as sample normalization or merging information contained in successive frames of a video. The first procedure performed by the system after receiving sign data is feature extraction. All proposed methods have to provide solutions for the two most important tasks: extraction of relevant features, and classification of entries to determine the most likely sign being presented.

There are many different types of features that can be used as the primary source of information, such as visual features, hand movement features, 3D skeletal features, and facial features, among others . The selection of features to be included for algorithm training is one of the most important factors that determine the success of the SLR method. The data are typically processed and transformed into a vector format before being input to the modeling layer, and multiple channels may be fused together to analyze their joint contribution to sign recognition.

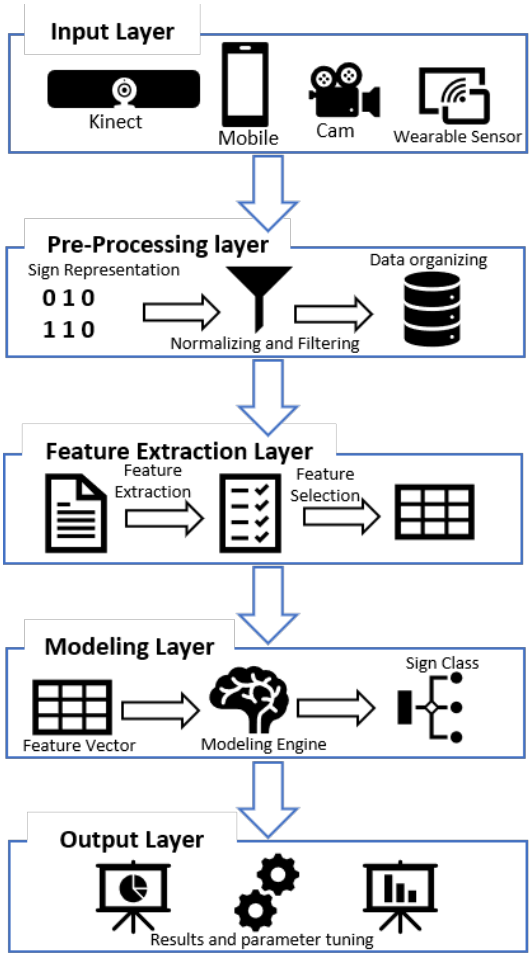


FIGURE 5: Automated Sign Language Recognition Framework

A. DATA COLLECTION

The interactive computing domain has evolved extensively in recent times. Consequently, a need for efficient human–computer interaction techniques has arisen. Sign language recognition is among the methods that can support further development of this domain. Sign language recognition enables the transfer of well-known gestures to a receiver. Techniques used to collect sign language recognition data can be hardware-based, vision-based, or hybrid.

1) Hardware-based

Hardware-based approaches are designed to circumvent computer vision problems during sign language recognition. These challenges may develop when recognizing signs from a video, for example. In many cases, hardware-based approaches use devices or wearable sensors. Wearable devices used in sign language recognition often use sensors attached to the user or implement a glove-based approach. These devices (whether sensors, gloves, or rings) can convert sign language into text or speech. With respect to wearable sensors and devices, the authors in [38]–[41] describe how they capture depth and intensity images obtained from a Microsoft Kinect sensor and a SOFTKINECT sensor. A similar category of observations features direct measurement methods that involve the use of sensors attached to the hands or body, as well as motion capture systems [42]. Huan et al. [43] observed that sensor-based approaches are never natural because burdensome instruments must be worn. Instead, they propose a novel approach, Real-Sense, which can detect and track hand locations naturally.

In recent years, the vast popularity of device-based approaches has resulted in renewed interest toward developing human gesture and action recognition methods. Among the device-based methods, Kinect is more commonly applied than the Leap motion controller (LMC) or Google Tango [13], [38]–[40], [43]–[50]. Wang et al. [51] identified Leap Motion as an excellent product that uses computer vision to achieve a useful interactive function. The significance of LMC is reinforced by the fact that learning and practicing sign language is not common in society, as discussed in [52].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

Some other methods rely on specially designed gloves for input such as [53]–[55], while a range of other technological devices such as accelerometers [56] and depth recording devices [57]. Some of the most basic sensor configurations include coloration of the fingers on gloves, as in [58], [59], allowing for easy and inexpensive motion tracking.

Gloves equipped with digital capture capacities were introduced by [60] and utilized to deduct hand signs of the Arabic sign language variation with a reduced number of sensors. While the cost of creating and using special equipment of this kind is considerable, it is still many times cheaper than purchasing some of the high-tech products available in the market.Authors in [61] chose motion controller as the primary input device, which allowed them to track objects in three dimensions with extreme precision at a 120 fps rate. The controller they used was developed for the purpose of tracking hand motion, so the researchers were able to follow many key points on the hands from one frame to the next. The same device was used by [62] to differentiate between 50 unique isolated hand signs, with absolute precision attained.

2) Vision-based

In recent years, research on sign language recognition systems has focused more on vision-based methods because they provide little to no restraints on users, unlike sensor-based approaches. In vision-based techniques, depth and pose estimation data are collected from users. A discussion regarding depth data and pose estimation can be found in section V. some of the recent SLR studies rely on input in the visual format. For example, depth information and RBG are some of the formats that can be commonly encountered in this field as demonstrated by [17]. Previous research by Rioux-Maldague et al. [44] indicates that use of depth data has increased because of the increased number of 3D sensors available in the market. A Microsoft Kinect sensor was used in their experiment, which has an image resolution of 640 × 480 and uses a traditional intensity camera to obtain depth images. Recent publications have also obtained depth data using vision-based approaches [40], [63], [64]. Depth data can be in the form of video sequences [65]–[70] or images [40], [71]–[74] obtained using a normal camera or a mobile device. Oyedotun et al. [74] used hand gesture grayscale images measuring 248 × 256 pixels. According to Zheng et al. [17], the use of depth data is advantageous to maintain privacy and to streamline the human body extraction process. Furthermore, depth data are invariant to changes in illumination, hair, clothing, skin, and background [17].

Aside from depth data, pose estimation has been used to facilitate vision-based techniques. Rioux-Maldague et al. [44]used a combination of regular intensity images and depth images to group different hand poses. They tracked the hands using functions that are publicly available in the OpenNI+NITE framework. While using pose estimation, computationally heavy heat maps for 2D joint locations were generated, and a 3D hand pose was inferred based on inverse kinematics and depth channels. Koller et al. [63] further described the state-of-the-art aspect of hand shape recognition, where the configuration of a hand pose is determined by the positions and angles of the joints. Currently, many experiments use these joint positions and angles because they can be estimated based on depth images and pixel-wise hand segmentation. Other experiments, such as those by Zimmermann et al. [75] use a hand pose estimation system combined with a classifier trained to recognize hand gestures.

While vision-based methods are non-invasive, they are constrained by the inadequate performance of conventional cameras. Another challenge is that uncomplicated hand features can cause ambiguities, while advanced features require extra processing time [39].

3) Hybrid

In some instances, hybrid approaches have been used to collect sign language recognition data. Hybrid methods exhibit similar or better performance compared to other methods with respect to proportional automatic speech or handwriting recognition. In hybrid approaches, vision-based cameras together with other types of sensors, such as infrared depth sensors, are combined to acquire multi-mode information regarding the shapes of the hands [76]. This approach requires calibration between the hardware and vision-based modalities, which can be particularly challenging. The fact that this method does not require retraining means that it is faster and can be used to examine the impact of deep learning techniques. Koller et al. [77] conducted an experiment and opted for the cleaner hybrid method, otherwise referred to as automatic speech recognition (ASR), to examine the direct impact of this type of data on a CNN.

Using still photos or continuous recordings in RGB format has the advantage of good resolution, but depth imaging does a better job at determining how far an item might be located from a fixed point. There are certain algorithms that use both types of visual data in combination [72]. Thermal imaging is also an intriguing possibility, even if it is used more rarely than the previous two formats. IR heat sensors can leverage the emitting of radio waves or reflection of light rays to construct an image as well. This type of information has been used with success for tasks such as facial recognition or body contouring, but has not yet found its way into SLR studies [78]. Skeletal data can also be used as a source of input, mostly in the form of hand joint position during SLR gestures. Another type of input is derived from motion capture, where information changes are tracked from one image to another. Models of this kind usually define the optical sequence as a vector describing the movement of pixels in series of still images, while so called scene sequence can be tracked in video materials referring to the motion of three-dimensional objects within the scene, relative to the distance from the camera lens [79].

While all of the input devices can be effective in the right scenario, their performance significantly fluctuates depending on the context. Still, more advanced input sources such as depth sensors and Real Sense/Kinect recording systems

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

can create three-dimensional representations which carry far more information than simple two-dimensional images from a fixed angle [78], [80].

B. SLR DATA PRE-PROCESSING AND FEATURE EXTRACTION FOR DEEP NEURAL NETWORKS

SLR data pre-processing plays a critical role in sign language recognition engineering. As such, data processing may involve sign representation, normalization and filtering, data formatting and organization, feature extraction, and feature selection.

1) Sign Representation

Sign language is a type of visual language that utilizes grammatically structured manual and non-manual sign representations to facilitate the communication process. These representations may range from the hand shape to the orientation of the palm, finger or hand movement and location, as well as head tilting, mouthing, and other aspects of facial expression. Tang et al. [39] used eight representative frames organized in a time sequence. Their representations showed the movement of two hands that began by moving closer to each other before moving apart. In [40],all gestures used in an experiment were represented by the hand of the signer. A hand segmentation phase was also used to represent the shape of the hand sign. Similarly, Koller et al. [63] represented 60 hand shape classes using a double state, while the garbage class was represented by a single state. Another experiment by Zhou et al. [81] evaluated only right-handed signers. In this case, the right hand was used to represent the dominant hand, while the left hand was the submissive hand. Hossen et al. [45] focused on the Bengali Sign Language, having 51 letters that were represented in the experiment using 38 signs, which were developed by combining related sound alphabets into single signs.

In the Bahasa Indonesia Language, one word is represented by at most five signs, as discussed in [69]. This means that every word and affix has an independent signed Indonesian (SIBI) representation and is represented by one sign that is consistently performed. Another experiment by Huang et al. [43], used 66 input units and 26 output units to represent 26 signs.

Past experiments have also attempted to compare body and hand features. In [15], it was observed that body features make up a somewhat better representation compared to hand features for sign language recognition. In essence, using body features improved the recognition of sign language by 2.27% [70]. These observations can be attributed to the fact that body joints are more dependable and robust than hand joints.

2) Normalization and filtering

In machine learning and deep learning, normalization refers to all actions and procedures aimed at standardizing the input based on a set of predefined rules with the ultimate objective to improve performance of the AI tool. This procedure is typically performed during the data pre-processing stage, and may include various statistical operations or media processing tasks. The exact type of normalization procedure that is optimal for the current implementation depends on the format of the input (i.e. text, image, or video), the level of variability within the sample, the type of machine learning architecture, the purpose of the automation tool, etc. Due to its impact on performance, normalization is commonly included in most contemporary Sign Language Recognition methodologies and its contributions are empirically verified [59], [82]. As SLR studies use a lot of different input modalities and pursue a range of different objectives, it is logical that the scope of normalization techniques found in this field is quite broad. Most of the techniques are visual in nature, and involve changes of images to fit them into a standard format that can be readily interpreted by the algorithm. This is frequently done by altering the data on the level of pixels, since this is how information is encoded in the machine learning models during the feature extraction and network training stages.

Some of the simplest examples of normalization methods used in SLR are image resizing and re-shaping, as demonstrated in Kratimenos et al [83] and several other works [59], [84]. Garurel et al [85] also normalize the size of each frame to fit feature map dimensions, using mean values and standard deviations obtained during training to find the most optimal size. Cropping is another frequently used method that can improve the quality of visual data and make sign recognition more reliable by removing sources of possible confusion for the algorithm. Input images are typically cropped in such a way to eliminate all regions except those depicting hands and face, which are crucial for sign language communication. In [86], cropped images are normalized based on the average length of the neck, thus negating the impact of the distance from camera for every image. In [87], a benchmark signer is selected and input from other signers is standardized based on positions of key joints. Contour extraction is used to this end as well, for example in [88], with the main focus on the areas corresponding to hands, with background removed from the image. For SLR methods that rely primarily on video for raw input, frame downsampling is frequently used to standardize the quality of various clips and reduce computational demands.

In [44], normalization and filtering processes were applied. The intensity histogram of an image was equalized and all pixels were normalized to the [0,1] interval. Gabor filters were then applied to the processed images using four different scales and orientations. An attempt was made to apply bar filters to the depth and intensity images to obtain the primary contours of the hands. Gabor filters were also used in an experiment by Li et al. [76] to obtain hand features that could be used for classification. While using Gabor filters, images were normalized to a size of 96×96 pixels. In another experiment in [40], principal component analysis (PCA) filter convolutions learned from input images were used. As part of the preprocessing, Koller et al. [63] applied a per-pixel mean normalization to images and used pre-trained convolutional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

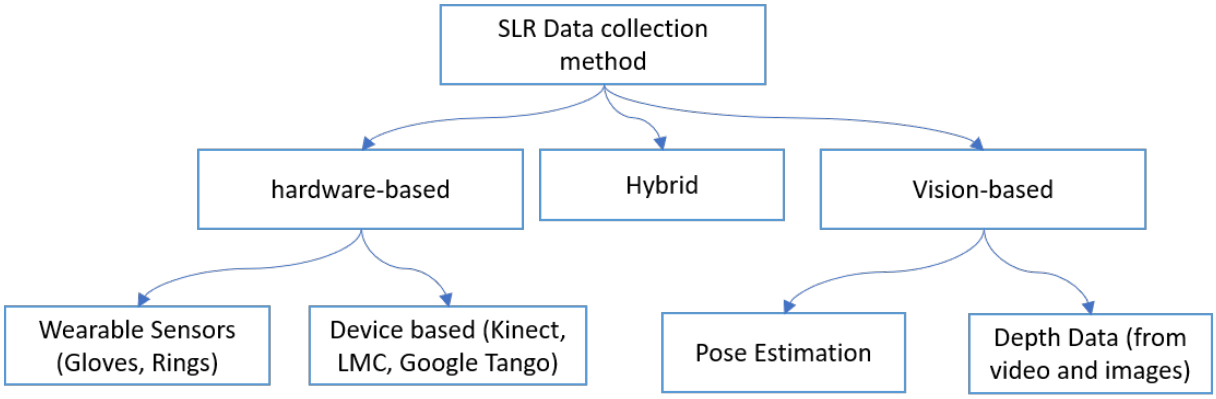


FIGURE 6: Primary data collection methods employed for SLR

filters located in the lower layers of their CNN model. Zhou et al. [81] did not conduct any normalization process in their experiment because the extracted features occurred naturally in the range of [-1,1]. Another experiment by Yang et al. [65] set a threshold to filter the minor skin-color area, an approach that enhanced the robustness of the system by using the second layer of their CNN model as a filter.

Other examples of normalization can be observed in the experiments by [67], [70], [71].Balayn et al. [67] normalized Japanese sign language (JSL) motion sentences and used them as inputs and outputs for Seq2Seq models. Konstantinidis et al. [70] normalized hand positions, which were used as inputs for the classifier together with cropped hand regions. In their attempt to examine Chinese Sign Language, [71] obtained a total of 1,260 images of basic signs in Chinese, which were normalized to 256 × 256 optimized background samples. Their model used 16 filters in the first convolutional layer. The filters had a width and height of 7 and a channel width of 3. Similarly, Koller et al. [77] applied a global mean normalization process to images before fine-tuning their CNN model.

Experiments to format and organize data in various ways have been reported. Tang et al. [39] organized the hidden layers of their models using various planes within which all units shared similar weights. In another experiment by Jiang et al. [71], the data were divided into training and test sets, with the training set containing 80% of the total images and the test set containing the remaining 20%. In a different experiment that used a Kinect sign language dataset, Huang et al. [41] formatted and organized their data into 25 vocabularies that were extensively used in daily life. Each word was played by nine signers, and each signer repeated each word three times. Using this approach, each word was organized into 27 samples, yielding a total of 25 × 27 samples. Eighteen samples were selected for training, and the remaining samples were used for testing.

Many studies from this field also include filtering and data augmentation steps, which have the purpose of improving the quality of input and consequently boosting the accuracy of the model. Random sampling or discarding of frames is one of the most straightforward techniques found in literature, where approximately 20% of input is eliminated. In [89], this technique is complemented by random changes of brightness, saturation, and other image parameters. Some of the data augmentation methods used in [90] include Gaussian Noise, Just Counter, and Future Prediction. The PoseLTSM tool also employs some operations aimed at augmenting the input images, with rotation of the hands around fixed points in the wrists as one of the most original ideas. As with normalization, the choice of filtering and data augmentation techniques is directly related to the properties of the model and the type of input, so it must be made with full understanding of each individual implementation and its objectives.

3) Feature extraction

Feature extraction is a crucial step in all of SLR models, since it impacts how the models are trained and consequently how quickly they can become effective at distinguishing between different signs/words. In all cases, features are derived from raw data, and they refer to positions of body parts (key points in hands and face) relevant for sign language communication. Features are calculated based on statistical operations, and assigned weights proportional to their discriminatory value [90]. In effect, features are expressed as vectors in the latent space and allow the neural model to learn the probabilities of their association with particular classes.

Several different feature engineering schemes are discussed, and in some cases a special tool was used for their extraction. The final number of features as well as weight distribution between them is typically optimized based on their impact on accuracy and scalability of the model [40], [81]. Various authors [38], [39] conducted feature extraction processes in their sign language recognition experiments. Wu et al. [38] carried out high-level feature extraction by fixing the architecture of the network as [NX, N2, 1000, 1000, 1000, 1000, NTC), where NX represents the dimension of the observation domain and N2 represents the number of hidden nodes. In another experiment, Rioux-Maldague et al. [44] presented a novel feature extraction method for recognizing hand poses using depth and intensity images. Images were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

de-interlaced by maintaining every other line in an image to resize them from 128 × 64 to 64 × 64. Each resulting 64 × 64 image was unrolled as a 1 × 4096 intensity vector. Tang et al. [39] extracted hand features by considering the two hands as a whole, making the recognition process much more accurate. A similar experiment in [40] used PCANet for feature extraction to solve the challenges associated with processing different image modalities. Li et al. [42] exemplified the process of feature extraction by transferring sensor signals from both hands into feature vectors. Such an approach bypasses the approach of reconstructing the precise shape of the hand, its orientation, and position.

Likewise, Camgoz et al. [64] used 2D CNNs to conduct spatial feature extraction. The 2D convolution layers obtained feature maps by using weights to convolve images. Additionally, observations from [21] also demonstrated that various stages of convolution and subsampling can be used to extract spatial-temporal features. Based on these principles, Huang et al. [41] extracted hand-crafted features from a video containing sign language and used the features to train a Gaussian mixture model-hidden Markov model (GMM-HMM). Unlike Huang et al. [41], who oversaw the feature extraction process manually, features such as finger length, finger width, and angle of the finger were input directly to the DNN in a separate study [43]. Instead of using 2D CNNs, some experiments have used 3D-CNNs owing to their capability to consider spatial and temporal relationships. For instance, the authors in [11] used a ResNet model rooted in a 3D CNN model to generate a representation of each video clip considered. Within the same domain, the authors in [45] developed a neural network that uses a stack of layers to extract features. In [71], a convolution layer was used to extract various features of the input. The authors in [72] used a trained CNN as the feature extractor for an SVM.

Another experiment by Konstantinidis et al. [37] extracted a mixture of video and skeletal features from video sequences. The video features were the image and optical flow, while the skeletal features were the body, hand, and face. The VGG-16 network pre-trained on ImageNet was used to extract video features, whereas FlowNet2 was used for the optical flow images. A similar experiment by Konstantinidis et al. [70] used a mixture of the ImageNet VCG-19 network and $\text{conv}4_4$ for feature extraction. The key features extracted during the experiment included 18 body and 21 hand joints in 2D. Rao et al. [68] conducted human-like feature extraction and recognition. These features are those used by human interpreters to recall signs accurately.

There have been a few experiments that seek to avoid or simplify the feature extraction process. For instance, Yang et al. [65] used a CNN owing to its capability to avoid complicated feature extraction processes. Therefore, it allowed direct image input into the sign language recognition system.

4) Feature selection

Feature selection is a crucial step in the design of practically all machine-learning-based sign language recognition model. Basically, it involves a reduction of data to a limited number of relevant statistical parameters, which are then fed into the machine learning network as input [91]. The idea is to include only those features that significantly contribute to the ability of the algorithm to distinguish between different classes, effectively limiting the number of computations necessary to obtain an accurate prediction. Thus, the exact number of selected features may vary from one model to the next, depending on the type of algorithm used, structure and volume of raw data, and the main tasks that the machine learning classifier will be expected to complete [92].

Researchers use many different methodologies to rank features based on their relevance and select those that deserve to be included. Broadly speaking, there are two major types of feature selected techniques – supervised and unsupervised [91]. In terms of the principles used to rank the features, we can talk about Filter methods (such as variance threshold, correlation coefficient, or Chi-square test) which capture some of the native properties of each feature, and Wrapper methods (i.e. forward feature selection or backward feature elimination), which measure how a proposed set of features works with a particular algorithm [92]. There are also Embedded (LASSO regularization or random forest importance) and Hybrid approaches, which combine some of the main strengths of both Filter and Wrapper methods. With so many possibilities for feature selection, researchers need to take the specifics of their project into account and use the scheme that best suits the classifier, the key tasks, and the data [93].

Some experiments that conduct feature selection include those in [39], [81]. In [39], a deep neural network was used, reducing the need to manually select certain features. The deep neural network autonomously detects and obtains useful features . Another example of the feature selection process was presented in [81], where 215 distinct test sentences were selected to represent conventional conversations in sign language. Another experimental work by Konstantinidis et al. [70] selected only 12 out of the total of 18 features extracted from body skeleton joints. The selection was based on the fact that the signers in a sign language dataset are usually in a sitting position , and the skeleton joints of their legs are usually not visible. Apart from CNN, some experiments also used PCA to facilitate the feature selection process. The use of PCA is guided by the fact that PCA is a conventional dimensionality reduction approach that can be useful when processing image data, which typically involves high-dimensional space. For instance, the authors in [76] used PCA to conduct feature selection and dimensional reduction. A different experiment by Huang et al. [43] illustrated the use of a DNN (deep learning or feature learning) in the generation and selection of features. In essence, a DNN has the capability to autonomously analyze and generate features from raw data.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

C. SIGN LANGUAGE MODELING AND RECOGNITION

Sign language modeling focuses on developing an articulate model from the phonetic to the semantic level for language representation. The modeling process covers various aspects, ranging from the use of the signing space to the synchronization of manual and non-manual features such as eye gaze and facial expressions. On the other hand, sign language recognition entails pattern matching, computer vision, linguistics, and other aspects of natural language processing [94]. The objective of sign language recognition is to establish different methods and algorithms that can recognize already developed signs and perceive their meaning. The techniques for sign language modeling and recognition discussed in this section include classic , deep learning, SLR continuous models, and SLR isolated models.

1) Machine Learning

Machine learning refers to the science of using computers to complete a task without having to program them explicitly. In many cases, machine learning algorithms are usually provided with general guidelines that characterize the model along with the necessary data. The data usually contain information to allow the model to complete a given task. This means that a machine learning algorithm can achieve its task when the model is adjusted based on the associated data. Examples of machine learning algorithms include SVM, PCA, and LDA, among many others.

Support Vector Machines

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for classification problems involving two groups. Feeding new sets of labeled training data into an SVM model can result in groupings of new examples. Past experiments have used SVM for this and many other purposes. Nguyen et al. [72] applied multiclass SVMs to learn extracted data . The validation accuracy of the CNN-SVM model was lower than that of the HOG-LBP-SVM model. However, the CNN-SVM model had a better chance of avoiding overfitting . The demand for real-time performance was evaluated in [76] to compare the most popular classifiers, which combine softmax and linear SVM. SVM and softmax performed better than other advanced classifiers in terms of accuracy. Additionally, it was observed that an SVM classifier featuring a linear kernel required more training time but performed better than the softmax-based classifier. Similarly, an experiment by [43] attempted to compare the performance of DNN and SVM by using the same dataset. The results indicated that DNN had a better recognition rate than SVM. Similarly, the authors in [46] identified SVM as a suitable classifier for real-time sign language recognition. While exploring American Sign Language, Chong et al. [52] used SVM and a deep neural network as a sign language recognizer. The outcome of the experiment showed that the rate of sign language recognition for 26 letters was 80.30% when using SVM and 93.81% when using DNN. It was also observed that the recognition rates for a grouping of 26 letters and 10 digits were slightly lower at 72.79% for SVM and 88.79% for DNN. The performance of the SVM was inferior to that of the DNN in sign language recognition. Similarly, Huang et al. [49] applied SVM to their approach for recognizing a large-vocabulary sign language. The SVM scheme used in the experiment facilitated the process of mean pooling over clipped features to produce a fixed dimension vector as the video feature representation. Huang et al. [49] trained an SVM for classification based on video features. Despite the use of SVM, the authors noted that their machine learning approach disregards temporal information during the mean-pooling process. The effectiveness of the SVM in a hybrid system was also evaluated in [95]. The experiment examined the classification accuracy of a HOG+SVM system. The hybrid system included a HOG feature extractor that produced 64-dimensional features and an SVM classifier that was fed canonical handshapes. Improvements in accuracy with the HOG+SVM system were between 14.18% and 18.33% compared to SVM alone.

Principal Component Analysis

PCA is used in computer vision to reduce dimensionality or to extract features. Many recent experiments have used PCA in sign language recognition as a dimensionality reduction mechanism. PCA can best be described as an orthogonal linear transformation that converts the original data into a new coordinate system having a reduced number of dimensions. In [40], a fingerspelling recognition system based on PCA was proposed. The convolutional layer of the proposed PCANet system features PCA. Another investigation focused on training a CNN on 1 million hand images using PCA [63]. Koller et al. [63] utilized 1024-dimensional feature maps and applied PCA to reduce the dimensionality to 200. Another experiment by [67] used PCA to select data streams exhibiting a high variance represented by approximately 492 dimensions. The use of PCA on Kinect data has also helped to reduce cases of overfitting. In a different experiment, [51] used PCA to expand a matrix into a 210-dimensional vector. These dimensional vectors are useful in the creation of an enhanced scheme for the mel frequency cepstral coefficient (MFCC), which is useful for sign language recognition.

Some experiments have compared their proposed approaches to a hybrid version of PCA. In [96], the proposed method was compared to other methods, including SAE+PCA. The outcome of the comparison indicated that SAE+PCA performed better than the proposed method, and achieved 99.05% accuracy. Other experiments have also shown interest in a variation of PCA, referred to as recursive principal component analysis (RPCA) for feature extraction. While exploring the features of SLR systems, [97] reported that using RPCA achieved a classification rate of 98%.

2) Hidden Markov Model HMM

This method relies on statistical operations that can reveal trends from the complex interaction of motions within a space-time continuum. It was first applied to the field of SLR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

by [98] in 1996, while [99] used in 1997 to classify isolated hand gestures based on visual input, achieving solid performance with the most optimal parameters. Variations such as dual HMM [100] or factorial HMM [101] were suggested at approximately the same time, seeking to build on the promising performance of the base model. Those studies confirmed that the model requires a lot of data during the training stage in order to arrive at sound statistical projections. Soon after, Wilson & Bobick [102] proposed a parameter-based improvement of this method, while authors in [103] proposed using parallel computing within this paradigm. The same principle was developed further by [104] to solve language-based problems. This approach was demonstrably more cost-efficient than any of the earlier HMM implementations and capable of reaching accuracy in excess of 94% for static signs and over 84% for dynamic signs in continuous speech by using 80% of the sample for training and 20% to test the model. Another class of models from this group is called input & output HMM, and was first developed by [105] to deal with material that is less homogenous. The same concept can be applied with success to track positions of hands during sign language communication, as demonstrated by [106], with accuracy of output of more than 70% when distinguishing between 16 signs based exclusively on hand movement.

Further development of the input/output HMM model was achieved in 2009 by [107], who introduced a cut-off point a thus managed to push the accuracy over 90%, albeit only when the total number of signs to be recognized was smaller than 20. An alternative was proposed by [108] in 2003, who called their method Left & Right HMM but were unable to significantly improve SLR performance over earlier version. A combination of HMM with GMM models can be useful for hand sign recognition even when the available data is scarce, as shown by [109], although reliability of the system decreases in this case. Hidden Markov Models were also used by [110] to analyze data collected with the help of multiple video cameras. While those methods have certain benefits, their application to the field of SLR requires additional work. In recent years, some researchers tried to use HMM alongside other methodologies in order to obtain better results. One such attempt was done by [111] in 2011, where this method was deployed together with PCA to determine key features of hand signs. On the other hand, authors in [112] added HMM to an RNN model tasked with tracking contours of hands during sign language communication, but were successful only when working with a limited number of already known signs. Yang et al [113] developed a variation of HMM that was aimed at shortening the calculation time, but this method requires certain conditions to be met, for example the length of each gesture must be limited. In the work of Belgacem et al [114], CRF method and HMM were used in combination to process training samples with scarce distribution, but with a lot of possible options the discrimination process is still very demanding.

Many continuous processing tasks experience terrestrial alignment challenges, which can often be resolved using hidden Markov models. In [63], an EM-based algorithm was incorporated into HMMs to facilitate weak supervision and overcome the challenges associated with video processing. Zhou et al. [81] used HMM techniques to develop a model framework that makes continuous sign language recognition possible. The use of HMM allows the resulting system to scale up to a larger vocabulary, allows modeling of signs and of transitions between signs, and decoding and training are possible even with new deep learning algorithms. In another experiment, the authors in [41] evaluated the Gaussian mixture model-hidden Markov model (GMM-HMM) as a baseline method. Trajectory and hand-shape features were extracted and used to train the GMM-HMM for recognition. An average accuracy rate of 90.8% was achieved when using trajectory as well as hand-shape features. A similar experiment by [49] also used GMM-HMM to facilitate temporal pattern recognition (automatic speech recognition as well as sign language recognition). Alternatively, combining HMM and BLSTM-NN yielded an accuracy of 97.85% for single-hand signs and 94.55% for double-hand signs [115].

Another experiment by Cui et al. [3] examined the role of HMMs in continuous sign language recognition. HMMs are among the most popular temporal models for sign language recognition. However, the framework developed in Cui et al.'s study performed better than HMMs. Their framework used recurrent neural networks in the sequence learning module.

3) Deep Learning Techniques

Deep learning is an incipient field of machine learning that focuses on learning representations of data [38]. However, the ability of deep learning techniques to capture semantics contained within data is limited by the complexity of the models and the underlying details of the input to the system [37], [38]. Advances in the field of deep learning have strong implications and applications for sign language interpretation using neural networks. Key deep learning techniques that have been applied in recent experiments include backpropagation, convolutional neural networks, recurrent neural networks, recurrent convolutional neural networks, attention-based approaches, deep belief networks, PCANets, SubUNets, logistic regression, transfer learning, and hybrid deep architecture.

Backpropagation

Backpropagation is a supervised learning algorithm used to train feedforward neural networks. The basic equations describing the learning process are given by (1) and (2) . This classic multilayer perceptron (MLP) technique was used by Rioux-Maldague et al. [44] to train a translation layer. The output layer was trained using normal backpropagation to interpret the activations of various restricted Boltzman machines (RBMs) into a 24-dimension softmax vector for every 24 letters. Training was conducted based on 200 epochs of backpropagation and used both weight decay and early stopping. A fine backpropagation phase was also conducted

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

using the entire network but at a much lower learning rate. In addition, Wu et al. [38] adopted the standard backpropagation method to adjust the weight of each modality

$$\theta^{t+1} = \theta^t - \alpha \frac{\partial E}{\partial \theta} \tag{1}$$

$$E = \frac{1}{2N} \sum_{i=1}^{N} (y_i - y_i\hat{})^2 \tag{2}$$

where theta is a weight θ, alpha α is the learning rate.

Deep Belief Network

Some sign language experiments have used a deep belief network (DBN) to classify learning representations. DBNs are comparable to multilayer perceptrons (MLPs), but they have many additional layers in their structure. The extra layers in DBNs provide enhanced learning potential, even though these layers are usually difficult to train. However, recent work has facilitated DBN training. For instance, Rioux-Maldague et al. [44] used a DBN consisting of three restricted Boltzmann machines (RBMs) and a single extra translation layer. Tang et al. [39] used DBNs to implement hand posture recognition. Based on the recognition results, the DBN attained a high recognition accuracy of 98.12%, which was better than the baseline HOG+SVM approach. Similarly, Huang et al. [43] established a deep neural network that can recognize various signs based on Real-Sense. The technique uses 3D coordinates of finger joints because the model can learn key recognition features from the raw data. The average rate of recognition of this DNN based on Real-Sense was 98.9%, while that of a DNN based on Kinect was 97.8%. An additional experiment that used the deep belief network was conducted in [96]. An American Sign Language dataset was used to examine the structure of a deep belief network and its performance in gesture recognition. The experiment compared DBN with other classic methods for recognizing gestures (a convolutional neural network and a stacked denoise auto encoder), and the results demonstrated a much higher performance by the designed DBN.

Convolutional Neural Network (CNN)

A Convolutional Neural Network receives an input image, assigns significance to different aspects of the image, and differentiates one image from another. Figure 7 shows the basic CNN architecture mode for sign language recognition. CNNs require a much lower level of pre-processing compared to other deep learning algorithms [63]. While these networks perform strongly in many tasks [65], they require large amounts of labeled training data [67], [71]. Hand shape recognition, a process influenced by the pose of the subject, has a remarkably high rate of intra-class ambiguity, which results in an added burden to acquire training data. In many cases, only a few specific labeled datasets exist in the gesture and sign language recognition field. As such, CNN has been used because it can be trained easily. In [63], a CNN was embedded within an iterative expectation-maximization (EM) algorithm, which allowed the CNN to be trained using a very large number of model images. The CNN achieved a recognition accuracy of 62.8% on over 3000 hand shape images that were labeled manually.

Some experiments focus on American Sign Language, such as in [72]. The methodology applied an end-to-end CNN architecture to a training dataset for comparison purposes. Additionally, CNN and SVM were combined to act as a feature descriptor, producing acceptable accuracy. Another related experiment by Li et al. [76] used CNN to process images of a large size. The CNN shares the weights of the images, thereby significantly reducing the number of parameters that need to be learned. This also reduces the risk of overfitting. CNN also finds invariant features that are particularly useful during image processing. By combining CNN with various PCA layers, [76] developed a hierarchical model that proved useful for recognizing fingerspelling in American Sign Language. Similarly, the authors in [73] developed a CNN focused on grouping fingerspelling images using a mixture of image intensity and depth data. The CNN was evaluated by applying it to American Sign Language with respect to fingerspelling recognition, and the developed CNN performed better than CNNs evaluated in previous studies. Specifically, the CNN achieved a precision of 82% and a recall of 80%.

Similar observations concerning American Sign Language were noted in an experiment by Taskiran et al. [116], where a CNN structure was used to extract and classify features obtained from the American Sign Language. The CNN model had the following features: an input layer, a pooling layer, two 2D convolutional layers, two dense layers, and a flattening layer. The resulting system achieved high accuracy, even when evaluating letters that had shared gestures. Daroya et al. [117] used a CNN model to examine the performance of a framework they proposed. The experiment applied Alexnet (an effective CNN model) and altered a few parameters to adapt it to their dataset consisting of 28 × 28 pixel images. In another trial, Shahriar et al. [118] attempted to recognize American Sign Language using a real-time approach. Images used in the experiment were categorized using CNN and deep learning. A CNN was used to obtain features from the images, while the deep learning method was used to train a classifier to identify sign language. Specifically, the CNN model was trained to produce a 4096-dimensional feature vector for the following classes: face, A, palm, and v. Similar to [119], the authors in [118] also used AlexNet, a built-in neural network consisting of 25 layers that was pre-trained extensively. The output of the image features indicated that the CNN model, together with the deep learning method, managed to classify input images with a high level of accuracy. Similarly, Cayamcela et al. [120] used a CNN model to translate American Sign Language from a real-time perspective. The CNN model was trained using a dataset consisting of several instances obtained from the American Sign Language alphabet. The CNN obtains features from every pixel and develops precise predictions based on a translator. In general, the CNN model achieved higher accuracy than its comparable counterparts.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

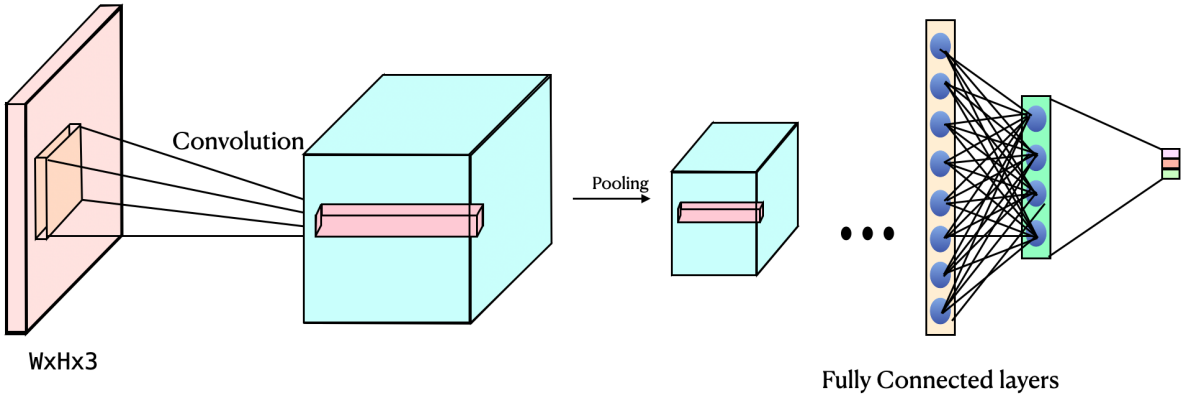


FIGURE 7: Basic CNN Model used in Sign Language Recognition

In [65], a CNN was integrated into a novel video-based recognition method and used to obtain upper-body images precisely from videos. In the experiment, the CNN model was trained for recognition, thereby simplifying the feature extraction process. The CNN circumvented the complicated feature extraction process by allowing direct image input. In addition, a Chinese Sign Language recognition method using this CNN model achieved a high accuracy of 99%. Similarly, [45] used deep convolutional neural networks (DCNNs) to develop a new method that can facilitate Bengali Sign Language recognition. Hossen et al. [45] used a network consisting of a convolution layer, a ReLU layer, a max-pooling layer, a fully connected layer, a dropout layer, and a softmax layer , which achieved an accuracy of 84.68%. This accuracy is remarkably high, considering that a very small dataset was used to train and test their network.

A similar experiment focusing on the Chinese Sign Language used a CNN consisting of six layers to facilitate fingerspelling recognition [71]. The deep learning approach consisted of various components such as dropout, maximum pooling, and batch normalization. The CNN achieved an overall accuracy of 88.10 +/- 1.48%, and a maximum accuracy of 90.87%, which was higher than other established approaches. Recent experiments have focused on the Arabic Sign Language. Shahin et al. [121] introduced a system that could recognize Arabic sign language using a vision-based approach. In the design of the system, deep learning methodologies relying on CNN were used to attain a high level of accuracy without the need for sensors . The results of the experiment were very promising for the application of CNN in the recognition of Arabic sign language. In addition, Yasir et al. [122] used the CNN approach to train a dataset obtained from the Bangla Sign Language. The use of this data classification technique was guided by the fact that CNNs require little pre-processing when compared to other image classification algorithms. The resulting model had a validation accuracy of 94.88%.

Previous experiments have focused on the Indian Sign Language. Rao et al. [123] observed that it is exceedingly difficult to classify complex head and hand movements owing to their ever-changing shapes. Because of this, the use of CNN was proposed to recognize gestures in the Indian sign language. They trained the CNN using three varying sample sizes, each consisting of different sets of subjects and viewing angles. Different CNN architectures were designed and evaluated, from which much better recognition accuracy was achieved. Specifically, Rao et al. achieved a recognition rate of 92.88% when using CNN. Another test in this domain was conducted by Sajanraj et al. [124], who developed a real-time system to convert Indian Sign Language into text. A deep learning method (CNN) was introduced to classify the sign language. The accuracy of the resulting system was 99.56%. Additionally, the authors in [15] used a VGG-19 network to recognize sign language from video sequences. VGG-19 is a type of CNN that has been trained using more than 1 million images obtained from the ImageNet database. Generally, the VGG-19 network is 19 layers deep and can categorize images into 1,000 object categories. Konstantinidis et al. [70] used VGG-19 because it has learned rich feature representations for different image ranges. Within the same scope, the research contribution by Koller et al. [77] demonstrated a scheme that can be used to train a CNN in a supervised manner. The experiment took the outputs of the CNN classifier and incorporated them with an HMM approach, thereby allowing iterative learning of video data. Through this approach, a significant improvement was reported in the classification performance of the deep learning technique. Huang et al. [41] proposed a 3D CNN approach designed to automatically obtain discriminative spatial-temporal features from raw video streams. In their CNN architecture, for every type of visual source, nine frames measuring 64×48 were considered and centered on the existing frame as input. The 3D CNN achieved an accuracy of 88.5% when implemented on a gray channel, and 94.2% when implemented on multi-channels.

Some experiments have applied deep learning to classify RGB images. For instance, in [117], a CNN was used as an approach that can classify RGB images of static hand poses

(representing a letter) associated with sign language. This method is based on DenseNet. In essence, the approach could classify sign language images in real time and performed well with an accuracy of 90.3%. Similarly, Rastgoo et al. [125] used a CNN model called a faster region-based convolutional neural network (Faster-RCNN) to detect hands in an input image. The purpose was to examine how a generative deep model can be used to obtain data from modeled data distribution probabilities and whether it can enhance the recognition performance of up-to-the-minute alternatives for recognizing sign language. The CNN detected input images as either original, cropped, or noisy cropped images [125]. CNNs have also been used in attempts to resolve the challenging task of gesture and sign language recognition in a constant video stream. For instance, Pigou et al. [126] used a deep learning approach and temporal convolutions to address this problem. The CNN model featured certain improvements that made it easier to conduct the classification process. The use of temporal convolutions was important for coping with the spatiotemporal nature of the data. Upon evaluation, the CNN model achieved a top-10 frame-wise accuracy of 73.3% when trained on the Corpus NGT and 55.7% on the Corpus VGT.

A recent experiment by Gunawan et al. [47] modified a CNN model and used the outcome to recognize sign language. The modified CNN is referred to as the i3d inception model and is based on the inception v1 model. The architecture of this model was used because of its capability to enhance the outcomes of previous experiments that used the ResNet-50 models, Two-Stream Fusion + IDT, and the C3D Ensemble. The i3d inception model was composed of 67 convolutional layers, including the input and output layers. In addition, the model contained nine inception modules. The outcome of the experiment indicated that the i3d inception model achieved fair training accuracy but an extremely low rate of validation. Correspondingly, Soodtoetong et al. [48] used a 3D-CNN to assess its efficiency in sign language recognition. The 3D-CNN model was used to determine the predictive gestures. The results of the experiment demonstrated that the 3D-CNN algorithm could identify gesture motions accurately, with the highest recognition rate being 92.24%.

Another experiment by Nakjai et al. [127] used a CNN model as part of the base model of YOLO. YOLO was used in the experiment to detect objects in real time, and CNN was its support model. The Darknet-19 architecture was used, which consisted of 19 convolution layers as well as 5 max-pooling layers with varying numbers of filters and filter sizes. Furthermore , Papadimitriou et al. [95] used a CNN variant to introduce a hybrid, vision-based, two-stage system that could effectively extract the shape of the hand. The convolution operation was changed in the CNN to enhance the learning capacity of the model. The alteration focused on the convolution scheme, leading to nonlinear behavior of the network output. The AlexNet architecture was used as part of the CNN. Additionally, the developed model followed the normal CNN layer pipeline, which involves convolution, pooling, and corresponding activation functions. The classification accuracy of the proposed method was tested and outperformed existing alternatives.

Recurrent Neural Network (RNN)

RNN is an influential model used to facilitate sequential data modeling. This approach has been used extensively and has been proven successful in a variety of important tasks, such as speech recognition, natural language processing, video recognition, and language translation. Figure 8 shows the basic RNN Encoder-Decoder architecture used for sign language recognition. Fang et al. [128] used a bidirectional RNN and long short-term memory (LSTM) in their experiment to facilitate universal and non-intrusive word and sentence-level translation of sign language. The outcome of the experiment indicated that the RNN model could successfully capture the important features of American Sign Language words.

A feature of RNNs that has been applied in some experiments is LSTM. For instance, Kavarthapu et al. [129] applied a bidirectional LSTM as the encoder and a second LSTM within the embedding layer as the decoder. The use of bidirectional LSTM in sign language recognition is significant because it allows the collection of information in an abstract manner. A standard LSTM was used to minimize the loss function. The results demonstrated that the bidirectional LSTM performed very well. Its performance could be attributed to the aptitude of the bidirectional LSTM. Correspondingly, Rakun et al. [130] attempted to use LSTM to recognize Indonesian Sign Language. LSTM was used in the experiment because the model can use full sequences as input and does not depend on pre-clustered per-frame data. The outcome of the experiment indicated that the 2-layer LSTM model achieved the best performance among the models compared and was 95.4% accurate in classifying root words. However, the LSTM model achieved a much lower accuracy of 77% when used on inflectional words, which can be attributed to the challenges involved in identifying prefixes and suffixes. The architecture used in [131] featured an RNN consisting of LSTM cells. In the architecture, the feature vector from every frame was provided as the input at every time step. The output layer was composed of a softmax classifier. LSTM was used to guarantee the real-time translation of sign language. The resulting model could translate continuous sign language videos into comprehensive sentences in English and was regarded as being highly effective in facilitating communication through sign language.

A few recent experiments have also used LSTM to recognize Indonesian sign language gestures. In [132], 2-layer LSTM neural networks were used to identify Sistem Isyarat Bahasa Indonesia (SIBI) gestures. The neural network achieved very high accuracy rates of 91.74% for prefix, 98.94% for root, and 97.71% for suffix datasets [132]. Attempts to address the challenges associated with sign language translation have led to increased use of hierarchical deep recurrent fusion (HRF) networks. Guo et al. [50] developed a hierarchical recurrent architecture to encrypt visual semantics with varying visual granularities. The HRF

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

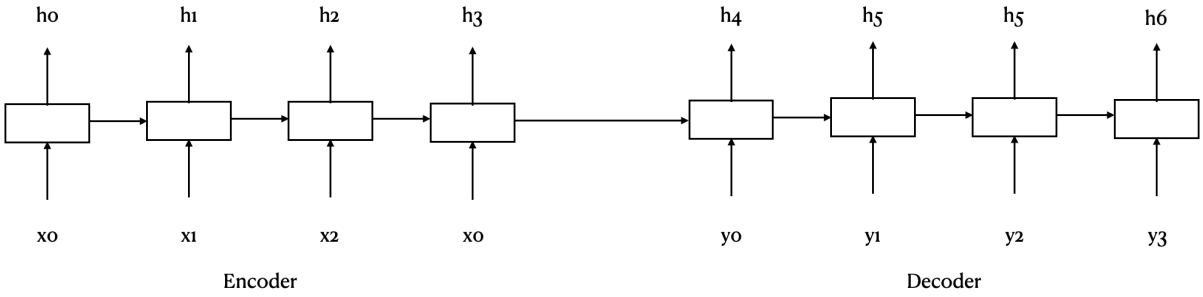


FIGURE 8: Basic RNN Encoder-Decoder model used in Sign Language Recognition

decodes a sentence by using complementary RGB visemes as well as skeleton signemes. The steps used were as follows. The HRF encoded the entire visual content after translating the video into various neural languages. Next, Guo et al. explored the use of adaptive clip summarization (ACS) to delve into sign action patterns in sign language translation. They also proposed an adaptive temporal segmentation scheme that differed from past models that obtain key frames or clips over a fixed time interval. In the next step, a hierarchical adaptive temporal encoding network was developed that condensed the time span. In addition to HRF, LSTM was selected as the basic RNN unit. The top layer of LSTM was responsible for learning the persistent characteristics of the original features. The medium layer was responsible for learning the recurrent features of compact visemes or signemes. The bottom layer transformed the visual information into textual semantics. As mentioned earlier, the main idea of the suggested model was to learn the descriptors of sub-visual words, such as visemes and signemes. Detailed experiments indicated that the HRF framework, working on the basis of LSTM, was highly effective.

Recurrent Convolutional Neural Networks (RCNNs)

Cui et al. [133] introduced a recurrent convolutional neural network to map video segments to glosses. They used an RCNN to extract features and facilitate sequence learning. By developing their architecture using RCNN, the performance could be equated to state-of-the-art models without having to introduce additional information. In this sense, the RCNN assisted in the process of continuous sign language recognition.

PCANet

Another type of deep learning technique used in sign language experiments is PCANet. Although this deep learning method has been proposed only recently, it is highly effective. As evident in [40], PCANet is very successful in solving many problems associated with object recognition and can be used to learn features obtained from intensity and depth images. Fingerspelling was recognized using two PCANet models to cover every color and depth input present in images. Empirically, the use of a two-stage PCANet is sufficient to achieve acceptable performance. As a result, developing a deeper architecture may not necessarily enhance the performance of this deep learning technique. Additionally, Aly et al. [13] used the PCANet deep learning architecture to recognize the alphabet in American Sign Language. Unlike [40], Aly et al. [13] proposed two approaches that could be used to train the PCANet models: the single PCANet and user-specific PCANet feature models. The single PCANet was trained using samples obtained from all users. In contrast , the user-specific PCANet was used to train various PCANet models, where individual models learned certain features from individual users. The extracted features were then identified using a linear SVM classifier. Inspired by the many achievements of the PCANet deep learning architecture, the model was used to autonomously learn depth features from segmented regions of the hand.

SubUNets

A few other experiments have used SubUNets to facilitate sequence-to-sequence tasks. In [64], the authors used SubUNets, which is a new deep learning architecture that produces a series of outputs from video. Unlike the other video-to-text methods, the approach mimics the contextual subunits of a task while simultaneously training the network for the key task. When dealing with the challenges of sign language recognition, SubUNets detect and identify individual signs in a certain video and generate a text translation. SubUNet features three tiers of neural networks. The first tier includes CNNs, which take images as inputs and are responsible for extracting spatial features. The second tier uses bidirectional LSTM (BLSTM) layers, which model the spatial features obtained from the CNNs [64]. The final tier includes a connectionist temporal classification (CTC) loss layer, which allows training the networks using videos of different lengths and label sequences. After being trained on the Deep Hand 1 million hands dataset, SubUNets achieved a Top-1 accuracy of 80.3% and a Top-5 accuracy of 93.9%.

Hybrid Deep Architectures

In many instances, the use of a single deep learning technique is challenging. As a result, some experiments have combined deep learning techniques. For instance, [39] noted that the process of training DBNs was difficult to parallelize across different computers. They evaluated this issue by using CNNs for comparison purposes. The recognition results indicated that CNN achieved a high recognition accuracy rate of 94.17%, although this was lower than the accuracy of the hybrid DBN approach.

Wang et al. [66] proposed a hybrid deep architecture to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

address the continuous sign language translation (CSLT) problem. The hybrid model featured the combination of a temporal convolution (TCOV) module, a bidirectional gated recurrent unit (BGRU) module, and a fusion layer (FL) module. In the model, TCOV is responsible for capturing short-term temporal transitions, whereas BGRU preserves the long-term context transitions that occur across temporal dimensions. The FL then links (fuses) the embedded features in both the TCOV and BGRU outputs to learn their corresponding relationships. Experimental results demonstrated that this hybrid deep architecture improved accuracy by 6.1% in terms of the word error rate (WER) compared to single deep learning techniques.

A CNN has also been used in combination with a bidirectional recurrent neural network (Bi-RNN). Combining these techniques, the authors in [69] used a 3D CNN to obtain features from every video frame and a Bi-RNN to generate unique features from the sequential behavior present in individual video frames. On average, the hybrid approach exhibited a higher average word error rate and a similar character error rate when compared to the Lipnet model. Comparably, Cui et al. [3] combined a deep CNN with a Bi-LSTM to extract features. The CNN model proved useful in learning spatiotemporal representations from the input of video streams. Then, Bi-LSTMs were used to learn more complicated dynamics. Bi-LSTMs iterate LSTM computations by calculating both forward and backward hidden sequences. The authors employed Bi-LSTMs because unidirectional RNNs are limited in the sense that they can only calculate hidden steps based on past time steps.

The authors in [49] employed attention-based 3D-CNNs to facilitate the recognition of large vocabularies in sign language. The attention-based framework has two primary advantages. First, the model can learn spatio-temporal features based on raw video input without having previous knowledge. Second, attention mechanisms assist in selecting clues. In this case, attention-based 3D-CNNs were assessed using Chinese Sign Language data and the ChaLearn14 benchmark. The outcome demonstrated the higher accuracy of the approach compared to other advanced algorithms. In [115], transfer learning was used to tune an ASLR model to detect Indian sign language. Transfer learning helped in the learning of new classes even in situations when new training sets were limited in size.

While focusing on American Sign Language, Oyedotun et al. [74] applied a mixture of deep learning-based networks to recognize hand gestures collected from a public database. The techniques applied were CNN and a stacked denoising autoencoder (SDAE). The recognition rate of CNN was 91.33%, while that of SDAE was 92.83% when evaluated on test data that were not part of the training data. Another experiment by Bantupalli et al. [134] examined American Sign Language using a mixture of CNN and RNN. In this case, Inception, a CNN model, was used to identify spatial features from a video stream designated for sign language recognition. Next, the experiment used LSTM and an RNN model to obtain temporal features from sequences of videos using two approaches: outputs were generated from softmax and from the pooling layer of the CNN. Despite the success of the experiment, the authors suggested that the use of capsule networks rather than Inception may have yielded better results in sign language recognition.

4) Transformer-Based Approach

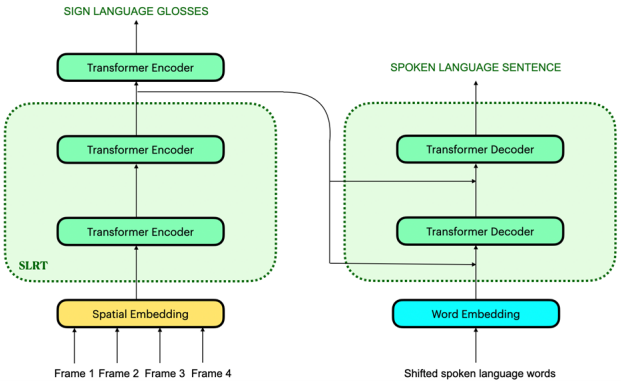
A range of different methodological approaches to sign language recognition can be found in the reviewed literature, but there are some basic principles shared by nearly all of them. In particular, the studies are focused on attention-based neural models with transformer architecture [135]. In this computing paradigm, encoder and decoder stacks are used to train the model for the classification of sign language samples as you can see in the diagram in Fig. 9. This approach has been proven successful with other types of tasks, and offers some unique advantages over earlier models. In this case, the models are expected to capture the relationship between temporal and spatial cues, and deduce the intended sign based on them. A tokenization procedure is performed to break down the input and output into frames/key points and word embeddings [136].

One of the unique limitations of transformer models is that they lack positional information for the inspected sequences, necessitating the introduction of the temporal ordering step. Therefore, feature extraction is another necessary element of all transformer-based neural models, where the most relevant features derived from input tokens are selected and later used for model training [85], [136]. Some of the features delineate between signs (inter-cue features), while others are useful to differentiate the particular gloss from similar ones (intra-cue features) [89], [137].

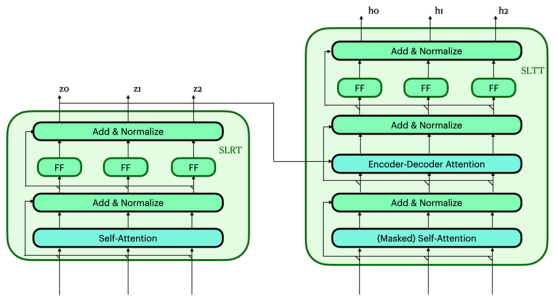
In one hybrid model, a separate neural network of the CNN type is used to extract the features from video input, greatly improving the efficiency of the process. The classification step is typically performed by a Bi-directional Long Short Term Memory (Bi-LSTM) module or an encoder-decoder stack, comprising several successive layers in both cases. The exact depth of the model and the number of deployed attention heads varies depending on the intended use of the model and other factors, and can be optimized for best performance based on the empirical evaluations. For example, some studies propose using only two layers in transformer models (as opposed to standard six used for Natural Language Processing), while others introduce a linear projection layer and a softmax attention layer on top of the stack [83], [86]. A normalization procedure is used to improve the efficiency of model training, which is driven by maximizing conditional probabilities and minimizing cross-entropy loss, while a validation procedure fine-tunes the model for the particular purpose.

Networks of this type were tested in several roles, including for isolated [138] and continuous SLR [139], as well as translation of sign language into spoken language. Video footage and skeletal data were used as input modalities,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access



(a) Transformer-based SLR Main Architecture



(b) A detailed overview of a single layered Sign Language Transformer

FIGURE 9: Transformer-based Model used in Sign Language Recognition [136]

but this methodology could conceivably also be used with different modalities [88]. Versatility of the deep learning approach with transformer architecture is very welcome in this challenging field, since the output can be specialized through the selection of training dataset and features, as well as training hyper parameters. Several interesting ideas were presented in the reviewed literature that could additionally refine the ability of encoder models to understand sign language, for example gloss-level supervision or the use of specialized pose estimation tools. With those improvements, some of the long-standing difficulties in the SLR field could finally be permanently resolved [140].

All studies from this group include an experimental evaluation of the proposed deep learning model, typically comparing its results with those obtained with alternative SLR approaches. The methods based on transformer architecture tend to outperform simple sequence to sequence models and other benchmarks by a significant margin on most datasets. The most optimal version of the algorithm can typically correct predictions in up to 85% of cases for tasks such as pose estimation, around 70-75% for isolated SLR, and up to 45% for the more demanding translation task. In some cases, the gains over competing methods were small, but in certain instances the improvements were quite dramatic. In addition to the task, several other factors such as the size of vocabulary, the size of the training dataset, and exact configuration of the network etc., could affect the quality of output [141]. While insights gained from those tests are extremely valuable, at this point it's hard to draw any firm conclusions about the most optimal setup that would guarantee high performance regardless of factors, such as the identity of sign performer, local variation of sign language, and environmental influences. From the data collected so far, it appears that deep neural networks of transformer type have a role in this scientific field, but it remains to be seen exactly what that role should be and how it can be leveraged for expanding the range of possible SLR applications [82].

While the methods based on transformer architecture bring tangible improvements over earlier deep learning SLR systems, their accuracy is still not near the level where they could be used in everyday practice without issues. Low accuracy is especially apparent with more complex tasks, and it tends to decrease as the complexity of analyzed sign language samples grows [139]. It's possible that gaps in performance are due to training samples and selected features rather than the fundamental data processing approach, but this postulation needs to be ascertained by more comprehensive testing and the possible inclusion of additional input modalities and localized sign language variations [89]. Based on the presented results, universal autonomous tools capable of continuous SLR that is signer-independent and language-independent remains a distant goal. Evaluation of the proposed encoder models suggests that a slightly different architecture might be optimal for SLR than for linguistic tasks, so it would be very interesting to see innovative attempts to redefine transformer models and develop them with the explicit purpose of interpreting sign language [142].

D. USING HAND GESTURE FOR SLR

Because of the importance of hand gestures for SLR, and given the significant amount of scientific research that has been carried out in this field, we limit our review to the most important points while mentioning the most important studies. Gesture interpretation has been a subject of scientific research for several decades; consequently, numerous reviews of this field have been conducted at various points in time. One of the earliest reviews was performed by Gavrila [143], who considered several 2D and 3D models for the analysis of human motion. Moeslund and Granum [144] provided a comprehensive recapitulation of two decades of research involving gesture tracking and recognition, while Ribeiro and Gonzaga [145] focused primarily on real-time approaches available at the time. Some of the more recent publications include an updated review of opportunities and challenges in this field undertaken by Rautaray and Agrawal [146]. Kumar et al. [147] discussed various feature extraction techniques, while Mohandes et al. [148] presented a survey of sensor-based and direct measurement methods for sign language

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

recognition. Because this field has experienced significant progress and undergone many reviews over the past two decades, we provide a brief overview of the current state of research in the field of hand gesture and sign language recognition by automated systems.

A majority of sign language characters and words can be expressed with simple hand gestures, which makes correct recognition of hand shapes a very practical feature for automated systems. However, the process of recognizing hand gestures involves many difficulties, which may be related to different hand sizes and shapes among signers, as well as different skin shades. In addition, various individuals may use unique styles to display certain elements when signing. Such difficulties can be resolved through the use of advanced analytic techniques aimed at identifying patterns independent of the signer's identity or the physical properties of their hands [149].

Because deep learning networks have the capability to identify latent connections among many different variables, they can be effectively used to analyze hand gestures in ASLR. Depending on the regional variations of sign languages, both one-handed and two-handed gestures can be used to express certain words or phrases, with single-handed signs usually assigned basic meanings, such as letters or numbers. Thus, hand gesture analysis alone has the potential to correctly recognize simple linguistic content from still images or videos, as well as other sources. In other applications, hand gesture analysis may be complemented by other techniques, such as tracking head movements [150]. Given that hand motion is the central building block of all sign language communication systems, this aspect of SLR is unlikely to lose relevance despite increased focus on full-body tracking and continuous sign language interpretation. However, pure hand gesture analysis techniques are likely to be combined with other methods to obtain the best results. This is exemplified by the increasing number of hybrid models that consider many different elements of a signer's behavior [151], [152].

E. USING POSE ESTIMATION FOR SLR

Because body configuration plays an important role in sign language recognition, pose estimation techniques are among the core tools in this area. The basic idea is to determine the exact pose of the entire body based on the positions of certain fixed points that can be ascertained through measurement. While this can be accomplished in various ways, deep learning algorithms have proven effective in this task given sufficient training using well-chosen samples. This is especially true when high-quality input is provided, preferably from more than one source/modality [153].

1) Pose estimation based on RGB images

A method using a convolutional neural network was suggested by [154] for determining pose of the human body by analyzing pictures of variable size, thus comparing how individual body parts are spatially organized. The final prediction required the operations of pooling and up-sampling to be repeated in several iterations. When this model was experimentally tested with two different datasets, it accomplished excellent results that outperformed the baselines for approximately 1.7-2.4%.

Another model using the same type of neural network was presented by [155], and it leveraged mutually dependent variables to predict the body position. In addition to a CNN network, this method also includes already prepared knowledge maps, and it employs a procedure that doesn't have to create a graphical representation in order to produce accurate output. This was confirmed by empirical testing where the so called Convolutional Pose Machine method outperformed all alternatives by 9% on the MPII set, 6% on the LSP dataset, and 3% on FLIC.

A model with cascading architecture was constructed by [156] in 2014, using DNN as the basic tool for estimating the positions of joints on the body and their mutual relations. This model frames the problem as a matter of regression, which turned out to be a very suitable paradigm, as evidenced by the performance of the model which was above the marks set by earlier solutions by 2% and 17% on two commonly used datasets. In the work of [157], a new dataset for SLR research was presented along with a benchmark for body position predictions for the comparison of techniques for pose estimation based on deep learning. Interestingly, they studied the possibilities for transfer learning and found evidence that this phenomenon applies in the field of SLR [158].

A similar model for pose estimation that uses RGB photos and deep learning was suggested by [159], starting from the linear SMPL model. In this work, three-dimensional representations of body joints were deployed as intermediaries and a regression of parameters was performed. The model relies on autoencoders to act as the link between the regressed SMPL and a convolutional neural network, guaranteeing that any structural imperfections would be corrected. Those improvements yielded a tangible performance boost in comparison with the basic SMPL on Surreal and Human 3.6M sets.

Sign language communication relies on more than one channel, and in addition to hand movements, facial expressions and body postures are commonly used to express meaning. While much work has been performed in the area of automatic recognition of hand gestures, the literature concerning the analysis of body positions is not as abundant. Tompson et al. [160] attempted to address this problem by analyzing the relationships among various body parts using a CNN. On the other hand, Yang and Ramanan [30] organized the data in tree-like fashion and used an SVM as the classifier.

Another notable study in this area was conducted by Chen and Yuille [161], who used a graphical model to represent the spatial configuration of body joints. This approach was improved by Pfister [162], who extracted temporal information from successive images to improve the capability of the system to interpret body positions, while Toshev et al. [156] provided an alternative method for body joint position eval-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

uation. Although there are many competing principles and ideas, the latter two approaches were chosen by the authors of this study as the starting points for conducting experiments on a new dataset, with the objective of establishing the most optimal methodology for body posture detection.

One more recent solution based on a convolutional network for analyzing graphs was proposed by [163], where a human body is presented in three dimensions by multiple points and links between them. This model deploys an attention mechanism to discriminate between data and contextualize this schematic representation. According to the results of experimental testing, this model can bring some modest gains in the range between 0.7% and 3.4% compared to alternative methods on various SLR datasets.

2) Pose estimation based on depth imaging

A model formulated by [164] combines components of convolutional and recurring neural networks with a self-correcting feature that can improve previous predictions. This model builds a 3D vector space constructed from local data and extracts partial body poses from it while accounting for the noise. The authors tested the model on an original dataset and found that it is indeed superior to any of the existing alternatives. Depth imaging also has a central role in the solution suggested by [165], which is named Depth Ranking Pose Estimation for 3D images. In this concept, CNN network is used for deciding between candidate pairs in the initial phase, followed by another step in which 3D pose estimations are made, thus combining depth data with two-dimensional images to great effect. This model was evaluated using the standard Human 3.6M dataset, yielding a significant accuracy improvement on the scale of over 6 mm over any competing 3D pose estimation methods.

A model named DDP or Deep Depth Pose was proposed by [166], where body positions were approximated by the construction of maps based on depth information. Such maps were created in advance and contained many different body positions along with all relevant joints. This approach was practically proven to be effective, outperforming the benchmarks by more than 11%.

3) Analysis of various pose estimation models

Since body position estimation has an important role in many different fields of research including SLR, there have been many attempts to formulate a successful model based on deep learning networks of convolutional and recurrent types. Introducing three-dimensional imaging and construction of depth maps has greatly increased the recognition capacity of such models. Some of the techniques aimed at achieving further gains in terms of accuracy comprise cascade or tree-like structures, imposing of certain limits, etc. From experimental evaluations, it is clear that recent models are far more powerful and reliable than their predecessors, but even the most complex solutions are still far from universal applicability [82].

Research directed towards better interpretation of body positions remains a central topic for scientific research. In particular, researchers are working to ensure that exact positions of each joint can be determined even when ambient noise is present in the images or parts of the body are blocked from view. There has been a lot of progress with 3D mapping of the body positions, but a part of the complication arises from the fact that multiple 3D positions can correspond to a single 2D pose. An additional complication is caused by the difficulty of labeling 3D joint images, necessitating the use of technologically advanced input devices. On the other hand, effective regression of 3D information requires highly precise mapping of spatial relations between key body points. Many existing models track multiple aspects including precise location of every joint in three dimension, from various angles and with regard to specific body shapes. Such models represent the foundation for future SLR research that new methods can build upon. Technological advancement of capturing devices also contributes to improved pose recognition and shape prediction abilities of new systems. Fusion of different types of data (i.e. thermal imaging or hybrid data) with vision-based indicators can make the systems more reliable under real life conditions and thus represents a promising direction of research.

Using sensor technology, the positions of the key points (i.e., limbs and joints) are transferred directly, whereas in image-based methods, those positions are inferred based on 2D images. Because of this essential difference, the methods chosen for completion of this task must reflect the type of input as well as the intended outcome [167]. Deducing the pose is instrumental in correcting the interpretation of the content of sign language communication [152]. This aspect is particularly important for continuous SLR, where individual signs are displayed in a non-stop stream, and changes in body position can be indicative of the intended meaning of the entire expression.

Many factors can affect the performance of pose estimation algorithms, from the choice of features that can include both 2D and 3D data points, to the classifier depth and architecture. Many of the latest tools for pose estimation achieve relatively high accuracy, but they are too prone to false recognition to consider them suitable for immediate practical application on a mass level [168]. In recent studies, there has been a tendency to use advanced technology, including the Microsoft Kinect device, to detect body poses based on multiple parameters, which is clearly a direction that will be exploited further in the near future as better sensors and tracking devices become available [151], [156]. Finally, high-quality resources necessary for evaluating new methods are emerging. The existence of readily available large SLR datasets stimulates more meticulous testing and brings us a step closer to the commercialization stage for this technology.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

V. SLR MODEL TYPES

There are two types of models related to the recognition process of sign languages: the isolated model and the continuous model. In the following sections we will show the work that has been done in this aspect.

A. SLR CONTINUOUS MODELS

As part of sign language recognition and modeling, some experiments have used continuous models. For example, Wu et al. [38] proposed a new bimodal dynamic network suitable for continuous recognition of gestures. The model relied on the positions of the 3D joints, as well as audio utterances of the gesture tokens. Koller et al. [63] demonstrated the use of an EM-based algorithm for continuous sign language recognition. The EM-based algorithm was designed to address the temporal alignment problem associated with continuous video processing tasks. Similarly, Li et al. [42] proposed a framework that addresses some of the scalability challenges associated with continuous sign language recognition. Another experiment by Camgoz et al. [64] developed an end-to-end system designed for continuous sign language alignment and recognition. The model is based on explicit subunit modeling. Similarly, Wang et al. [66] suggested a connectionist temporal fusion method having the capability to translate continuous visual languages in videos into textual language sentences.

Additional studies on continuous SLR models have been conducted by Rao et al. [68]. A system was developed and evaluated at various times using continuous Indian Sign Language sentences developed from 282 words. Similarly, Koller et al. [77] used a database consisting of continuous signing in German Sign Language. In [46], animations were processed continuously. However, this approach proved to be extremely challenging because the animations were difficult to work with after processing. While exploring the challenges of continuous translation, Pigou et al. [126] observed that deep residual networks can be used to learn patterns in continuous videos containing gestures and signs. The use of deep residual networks can minimize the need for preprocessing. In [17], a model was developed that can enhance existing sign language recognition methods by between 15% and 38% relatively, and by 13.3% absolutely. Cui et al. [133] also suggested a weakly supervised approach that could recognize sign language continuously with the help of deep neural networks. This approach achieved an outcome comparable to state-of-the-art approaches .

B. SLR ISOLATED MODELS

Until recently, a majority of sign language recognition experiments have been carried out on isolated sign samples. These models examine a sequence of images or signals based on hand movements obtained from sensor gloves [97]. Sensor gloves often represent a complete sign. For instance, Koller et al. [63] used a dataset that featured isolated signs from Danish and New Zealand sign languages. Another experiment by [37] proposed an isolated SLR system designed to extract discriminative aspects from videos, where each signed video corresponded to one word. After evaluating the challenges of continuous translation, Escudeiro et al. [46] resorted to an isolated approach. In essence, every gesture was created separately, making it easier to use animations with ease. Different observations by Fang et al. [128] suggested the use of a hierarchical model reliant on deep recurrent neural networks. The model successfully combined the isolated low-level American Sign Language characteristics into an organized high-level representation that could be used for translation.

Recent developments in sign language experiments have also suggested that the use of regions of interest (ROIs) to isolate hand gestures and sign language features can enhance the accuracy of recognition [134]. In [131], the authors used an isolated gloss recognition system to facilitate real-time sign language translation. The isolated gloss recognition system included video pre-processing as well as a time-series neural network module. Another experiment by Latif et al. [169] also considered video segments based on an estimated "gloss-level." While making their observations, Cui et al. [3] set their receptive field to the estimated length of an isolated sign. A recent study by Huang et al. [49] focused on a basic isolated sign language recognition task. The use of an attention-based 3D-CNN was proposed to recognize a large vocabulary. The model was advantageous because of it took advantage of the spatio-temporal feature learning capabilities of the 3D-CNN. Papadimitriou et al. [95] used the American Sign Language Lexicon Video Dataset, which consists of video sequences of isolated American Sign Language signs.

C. DELIBERATIONS ABOUT CONTINUOUS AND ISOLATED SLR

SLR comprises two distinct modes – isolated and continuous, each of which requires a different approaches and is associated with very specific challenges. In particular, one key distinction is that direct supervision is much more essential for continuous SLR. In isolated SLR, all the relevant content is concentrated in a limited area of a single image, but in continuous SLR it is necessary to carefully align the sections of the video in chronological order and ensure that each sentence is properly tagged. That's one example of the complexities associated with continuous sign language recognition, which is far more demanding in terms of computing efficiency. This must be taken into account during the evaluation of methodologies, as well as the feature selection process. If sequential labeling is done correctly and the most predictive features are selected, the resulting model has a higher chance of being accurate with continuous video analysis.

Over the last several years, smart applications of deep learning systems have removed many obstacles in this field as well as many other related automation tasks, but real breakthroughs that could lead to broad application by general population are still ahead. The attention mechanism is an intriguing element that works well with different types of data, and can be used to describe complex interactions in

space and time (for example, Graph Neural Networks applications). Further research will show whether this approach is the most optimal for resolving the issues complicating continuous SLR at the moment.

VI. SIGN LANGUAGE RECOGNITION BASED ON LOCALIZATION

Many basic concepts surround the use of sign language. First, sign languages are never international. Most, but not all, nations use different sign languages. Sign language is popular in American, British, Arabic, and Chinese settings, among many others. Table 3 provides an overview of various studies undertaken using different sign languages. For instance, American Sign Language (ASL), the most popular localization, includes independent grammar rules that are not a visual form of English. Application of this localization was evident in the experiment by Rioux-Maldague et al. [44], where the authors applied their proposed technique and classified ASL based on grammatical rules. Another experiment by Tang et al. [39] considered 36 hand postures obtained from American Sign Language to facilitate posture training and recognition. However, there are other systems that derive non-ASL signs and use them in English order. Such examples include experiments that have focused on Italian Sign Language. In [38], a dataset consisting of 20 Italian cultural or anthropological signs was used to evaluate a novel bimodal dynamic network designed to recognize gestures. The Italian dataset consisted of 393 labeled sequences and a total of 7,754 gestures.

Arabic Sign Language is also considered the preferred communication approach among many deaf people. In [40], depth and intensity images in the Arabic language were used to develop a system that can recognize associated signs . The proposed system was tested using a dataset obtained from three different users, resulting in an accuracy of 99.5%. The authors in [121], [169] also used an Arabic Sign Language dataset. In some cases, sign language experiments have focused on Chinese. In [81], vocabulary was adopted from 510 distinct words obtained from Chinese Sign Language. Among these words, 353 were single-sign words, while the remaining were multi-sign words. Yang et al. [65] also showed interest in Chinese Sign Language and used the instructional video We Learn Sign Language to meet the objective of their experiment. Another experiment by Jiang et al. [71] used Chinese Sign Language to facilitate the fingerspelling process. Furthermore, the authors in [49]–[51], [97] used Chinese Sign Language in their experiments.

A few other experiments evaluated Argentine Sign Language. An example would be [37], where an initial dataset was obtained from 10 subjects speaking Argentine Sign Language. Konstantinidis et al. [70] also used Argentine Sign Language featuring 10 subjects to explore hand and body skeletal recognition. Rather than focusing on a single language, some experiments use a mixture of sign languages. For example, Koller et al. [63] employed a mixture of Danish and New Zealand sign languages in their effort to examine how to train a CNN on 1 million hand images. The sign languages were obtained from two representative videos based on publicly available lexicons. The Danish data did not have any motion blur, while the New Zealand version had some motion blur. In another experiment, Camgoz et al. [64] focused on Danish, New Zealand, and German sign languages to evaluate the role of SubUNets in sign language recognition.

VII. RELATED STUDIES

The importance of hand gestures and sign recognition is indisputable; well-designed technologies of this type have the potential to impact millions of lives in a positive manner. This is reflected by the amount of new research in this area [143], [144], [146], [170], [171], which is growing rapidly as new technological platforms become available. By compiling a comprehensive list of SLR methods that are currently being discussed in research circles [6], [16], [172], we aim to provide the foundation for future researchers who are searching for references and inspiration. We discuss two major groups of solutions: vision-based (including both static and dynamic) methods and sensor-based SLR methods. Regarding the first group, various segmentation and feature extraction techniques are reviewed in [6], along with examples of successful neural classifiers. The latter group is discussed primarily in the context of a specific sensor device that enables data capture, while data processing is explained only briefly.

In terms of performance evaluation, the two metrics that are highly relevant to all of the discussed studies are classification accuracy and sample size. Classification accuracy is the percentage of correctly recognized signs and can be within the 0–100% range, with a higher percentage indicating better recognition results. Some of the methods in this field reached very high accuracy levels above 98% [173]–[178], but it is important to understand the conditions under which an algorithm can be expected to perform to its full potential as well as the scope of its possible applications. Most importantly, reviewing these studies can direct any interested researchers toward the most relevant research that may contain further information regarding a topic of interest. The sample size represents a combination of the total number of gestures that were displayed during the experimental evaluation and the number of classes to which a particular sign could belong [179], [180]. A larger sample size indicates more reliable results, and is always preferable. The sample size used for model training was listed in cases where it was specified in the original study.

The majority of reviews in this field lack sufficient space to provide in-depth discussions of all methods, and instead provide only a general snapshot [145], [181]–[184]. Additionally, some aspects of the sign recognition algorithms were not significantly discussed; for example, data preprocessing methods have been omitted because of the uneven presence of such information in the studies that were reviewed. In addition, methods that rely on sensors or customized input devices were not given proper consideration. Individual ap-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 3: Categorization of Sign Language Studies

Reference(s)	Sign Language
[13], [39], [44], [52], [72]–[74], [76], [95], [96], [116]–[120], [125], [128], [131], [134]	American Sign Language
[38]	Italian Sign Language
[40], [121], [169]	Arabic Sign Language
[49]–[51], [65], [71], [81], [97]	Chinese Sign Language
[70]	Argentine Sign Language
[63], [64]	Danish and New Zealand Sign Language
[45], [122]	Bengali Sign Language
[3], [66], [77], [95], [133]	German Sign Language
[67]	Japanese Sign Language
[68], [115], [123], [124]	Indian Sign Language
[69], [130], [132]	Indonesian Sign Language
[46]	Portuguese Sign Language
[126]	Dutch Sign Language
[127]	Thai Sign Language
[158]	Korean Sign Language

plications of SLR technology were also presented in a very succinct form, despite their relevance for the large number of users. This topic definitely deserves more attention in order to create innovation space that would allow for addressing numerous practical and philosophical issues that were not adequately covered in the previous period [**?**], [82]. The completeness of a literature survey is also relative, as new studies are rapidly published such that the solutions listed in any survey will eventually become obsolete. Hence, the value of this scientific resource will gradually decrease, and it will have to be replaced with a more updated version at some point.

VIII. BENCHMARKS

As we have seen in the previous sections, advanced learning algorithms have been used in the context of SLR with various degrees of success. As new deep learning architectures are being devised, some could bring improvements to studies of sign language interpretation and push them a step closer to practical application. Improved accuracy of basic sign recognition would open the doors for more advanced linguistic operations, including translation into spoken language, and prediction of the following signs. Since there are many regional variations of sign language, it is preferable to develop methods with universal potential. Deep learning networks can be trained on specific language corpora, which illustrates why this approach is so promising. It is hoped that the challenging issue of continuous SLR could be decisively solved with further perfection of currently researched methods based on artificial intelligence and deep learning. In this part, we will try to compare previous work with each other from several aspects such as datasets, features, and performance analysis.

A. DATASETS

This section presents some of the most important and available datasets containing hand gestures that can be used for the evaluation of SLR tools. The emphasis is on ensuring large enough dictionaries to facilitate more robust testing and more sophisticated applications. Currently, some high-quality sets can be used for this purpose, depending on the chosen geographic variation of sign language. For UK version of sign language, researchers have multiple datasets at their disposal, including RWTH-Boston-1, RWTH-Boston-50, and RWTH-Boston-400, ranging in number of different signs from 10 to 400.

High quality data corpus is also available for German sign language, with DGS Kinect-40, SIGNUM, and RWTH-PHOENIX-Weather as the most prominent examples. Those sets contain between 35 and 1225 unique signs, have a large number of authentic sentences by up to 9 skilled signers, and are labeled with the first and last frame of each sign clearly defined in terms of facial and hand features. ASLLVD proposed by Thangali et al. in 2011 is the most significant resource for studying American Sign Language, and contains over 30 thousand signs performed by 6 different persons. This is also a labeled set with designated frames marking the beginning and end of every sentence.

Studies of Polish Sign Language variation can use one of the 3 high-quality data sets, including PSL Kinect 30, PSL ToF 84, and PSL 101. Those datasets contain only isolated words (totaling between 30 and 101 signs) and have the limitation of being performed by only one person. Sign corpus IITA-ROBITA ISL is available for Indian researchers, and it was developed collaboratively between 2010 and 2017 by several research teams. Unfortunately, the entire set is performed by a single signer and contains only 23 signs. From all the aforementioned datasets, two stand out for their universal usability – ASLLVD and RWTH-PHOENIX-Weather. Those publically available sign language sets are suitable for interpretation of sign language in conditions that most resemble the real world, which is why they are often used as benchmarks in SLR studies for determining the effectiveness of proposed computing techniques.

Access to specialized datasets is currently one of the limiting factors in SLR research, which is why almost all researchers focus on this issue. The problem is exacerbated by the fact that separate datasets are required for each regional variation of sign language and for each different type of linguistic task. In some studies, the authors constructed new datasets from scratch by making video recordings of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

sign language users and obtaining other measurements, while in other cases, well-known local datasets were used instead. A typical dataset contains multiple repetitions of the same sign by several signers, with the objective of facilitating signer-independent recognition capacity after training. Some datasets presented in the literature are considerably larger than others, and this aspect should be taken into consideration when assessing the reliability of results.

Our examination of the datasets in all the reviewed research papers was conducted based on firmly defined criteria as we can see in Tables 4- 12, and relied on the discussions in the literature . Given that all papers are primarily interested in decoding sign language elements of various complexity levels, the databases used have many common features and can effectively be classified based on these features. The criteria were selected with the idea of providing a framework for direct comparison between studies, although in some cases, certain categories may not be applicable or some data may not have been reported by the authors. In this manner, our overview attempts to illustrate both the commonalities and differences among datasets upon which the conclusions of each study were based. Owing to space constraints and the need for clarity, training, testing, and evaluation datasets were typically merged together, so in some of the studies, the actual structure of a particular dataset may be more complex than is apparent from the size listed in the table. A more focused examination of each particular example is recommended for those interested in the practical use of any SLR dataset belonging to this group.

TABLE 4: Datasets Arabic Sign Language Recognition

Reference(s)	#Subjects	#Classes	#items	Arch
[185]	10	3	180	NA
[186]	30	30	900	NA
[187]	21	150	150	NA
[40]	3	28	1400	PCANet, SVM
[169]	40	32	54049	NA
[121]	40	32	54049	Resnet18

A quick glance at Tables 4– 12 is sufficient to note the large degree of diversity among the datasets with respect to the parameters used. This is a natural result of the fact that sign language studies employ a variety of methodological concepts and may explore mutually unrelated aspects of sign language recognition. It is important to understand the distinction between isolated and continuous SLR and the types of datasets suitable to each approach – for example, alphanumerical characters or words are typically used for recognition of isolated language elements, while sentences or even longer segments of speech are necessary for continuous SLR experiments. The datasets also greatly differ in terms of size and complexity, which is important to consider when attempting to evaluate the generalization potential of a given model. However, even the largest datasets are far from exhaustive and are typically limited by available resources and practical concerns.

One encouraging trend is that additional datasets documenting many different regional variations of sign language are becoming available. This is important because SLR research is universally relevant, so building automated tools capable of recognizing local versions of hand signs should be a priority. Multi-modal datasets are also becoming more common, which is a positive development signaling the next stage of SLR research and opening additional possibilities for innovative ideas. On the negative side, most datasets were created using a very small number of signers and feature a small number of classes, which brings into question their representative value. Consequently, the accuracy of any automated tools that rely on those datasets could be compromised when faced with slightly different presentations of sign language gestures. In any AI-related research field, the availability of high-quality datasets for model training and testing is a crucial factor that can affect the pace of progress. As a relatively new area of interest, SLR research initially suffered from this problem, but the studies reviewed offer evidence that the situation is steadily improving in this respect. There are several widely used datasets that can be considered ‘standard’ and can be used whenever broad compatibility of the experimental results is desired. On the other hand, new datasets focused on local sign language systems are quickly emerging and could potentially be reused to fuel additional research in the same geographic location. Despite the optimistic outlook, it is necessary to recognize that currently available datasets differ greatly in terms of quality, size, and structure, potentially necessitating the compilation of new datasets to support specific directions of research.

B. PERFORMANCE EVALUATION

A vast majority of research papers are concerned with accurate recognition of sign language material, and the primary metrics they use attempt to measure this capability. Consequently, some studies use common percentage-based accuracy indicators, such as precision and recall, as well as their combination, known as the F1 score. Depending on the stage of an experiment, some authors differentiate between training accuracy, testing accuracy, and validation accuracy. The training time necessary to accomplish reasonable accuracy is another factor that was tracked in some studies, and it was most commonly expressed in epochs. Processing time and input video length were less frequently considered to be sufficiently relevant to warrant direct measurement, but could be expressed in seconds and/or the number of frames. Tables 13 and 14 provide overviews of two performance benchmarks used in ASLR studies.

Almost every research paper reviewed includes a quantitative evaluation of the proposed sign recognition algorithm. Testing varies greatly in scope and complexity, with particular tests being administered depending on the objectives of the study. In general, the tests were designed to estimate how effectively the algorithm could differentiate between sign language words or sentences, often in comparison with several benchmark methods. Because of the diverse nature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 5: Datasets for American Sign Language

Reference(s)	#Subjects	#Classes	#items	Architecture
[188]	NA	36	2524	NA
[189]	6	3300	9800	NA
[157]	6+2	NA	808+479	NA
[190]	3 (1 M, 2 F)	15+20	NA	NA
[44]	5	24	60000	Deep Belief Network
[39]	6+2	36	288 videos	Deep Neural Network
[77]	7	40	2137 sentences	CNN + HMM
[76]	5	24	120000 images	Sparse AutoEncoder, PCA
[41]	9	25	657	3D CNN
[73]	5	26	60000 images	Convolutional Neural Network
[13]	5	24	60000 images	PCANet with SVM
[95]	6,20	24,3000	3000, 4416 image frames	Altered convolution operation,CNN
[120]	NA	26	78000	CNN
[74]	NA	24	2040 gestures	CNN and stacked denoising autoencoder
[128]	11	56, 100	7306 samples	hierarchical bidirectional DRNN
[116]	NA	36	900 images	NA
[117]	NA	24	100000 images	DenseNet
[119]	NA	24	34627 images	Capsule based Deep Neural Network
[134]	NA	24	62400	CNN-LSTM
[52]	12	26 letters, 10 digits	NA	DNN
[72]	20	5	2425 images	CNN-SVM

TABLE 6: Datasets for Sign Language Recognition for EU Countries Languages

Reference(s)	#Subjects	#Classes	#items	Architecture
[37]	10 + NA	64 + 50	3200 + 1297	NA
[70]	10	6450	32001535 videos	CNN, OpenPose, Stacked LSTM
[70]	10	64	3200 videos	CNN-Stacked LSTM
[47]	10	10	500 videos	3D CNN
[63]	6, 8, 2009	60	1,134,319 frames	CNN with embedded EM algorithm
[64]	23	60	1.2 million hand images	CNN+BLSTM
[191]	100	40	11 + 200 per class	NA
[126]	78, 53, 21	100, 100, 249	55224, 12599 video-gloss pairs, 22535 video files	Deep Residual Network, Bi-LSTM
[192]	18	60	5 hours of video content	NA
[3]	9	455	65,22711,874	Deep CNN and Bidirectional RNNs
[75]	20	35	43,986 images	Segmentation + Pose Estimation
[133]	NA	9	5,672 sentences	Connectionist Temporal Classification
[66]	40	10	6841 videos	TCOV, BGRU, Fusion Layer
[3]	91	455	6841 sentences 2340 sentences	Deep CNN+BiLSTM
[38]	NA	20	13,858	Deep Belief Network
[193]	3	10	2000 videos	NA
[46]	NA	57	NA	SVM, Dynamic Time Warping

TABLE 7: East Asian Countries Sign Language Recognition Datasets

Reference(s)	#Subjects	#Classes	#items	Architecture
[45]	3	45	54,000	NA
[194]	NA	10	1075	CNN
[45]	NA	37	1147 images	DCNN
[195]	NA	90	9000	NA
[43]	3	26	78	NA
[50]	50	179	5000 videos	3D CNN
[97]	3-50	20-3000	100-16000 sentences	NA
[49]	50	500	125000	Attention based 3D CNN
[65]	NA	40	NA	CNN
[71]	NA	30	1260 samples	CNN
[49]	50	50020	125000+14000 instances	Attention based 3D CNN
[115]	15	20	30000 images	Deep Neural Networks
[124]	NA	26, 9	NA	CNN
[68]	10	NA	282 words	Deep Neural Network
[130]	2	163	1630 word sequences	LSTM
[69]	10	30	3,006 videos 30 sentences	3D CNN-BiRNN
[48]	NA	5	100 images	3D CNN
[132]	NA	NA	NA	2 layer LSTM
[67]	1	195	812 sentences	Deep LSTM
[158]	14	419+105	14,672	NA
[158]	14	524	14672 videos	OpenPose feature extraction, GRU
[127]	12	25	30000 images	YOLO based CNN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 8: Other Sign Language Recognition Datasets

Reference(s)	#Subjects	#Classes	#items	Architecture
[196]	NA	20	8000-10000	NA
[197]	28	14/28	2800	NA
[51]	NA	NA	NA	NA
[42]	NA	NA	NA	NA
[129]	NA	8	1600 gestures	Encoder-Decoder with LSTM
[17]	NA	NA	NA	NA
[123]	5	200	72000 images	CNN
[96]	NA	24	60000 images	NA
[118]	NA	NA	NA	CNN
[131]	NA	NA	NA	Attention based Sequence model
[196]	NA	20	8000-10000	NA
[197]	28	14/28	2800	NA
[51]	NA	NA	NA	NA
[42]	NA	NA	NA	NA
[129]	NA	8	1600 gestures	Encoder-Decoder with LSTM
[17]	NA	NA	NA	NA
[123]	5	200	72000 images	CNN
[96]	NA	24	60000 images	NA
[118]	NA	NA	NA	CNN
[131]	NA	NA	NA	Attention based Sequence model

TABLE 9: Datasets for Sign Language Based on Alphabetic Linguistic Content

Reference(s)	#Input Modality
[196]	RGB video + depth info
[188]	2D Photo
[45]	2D photo
[63]	RGB
[186]	Image
[45]	RGB
[73]	RGBD
[13]	RGBD, Kinect
[97]	RGB, Kinect, Gloves
[95]	RGB Video
[120]	RGB
[40]	RGB, Depth
[75]	RGB
[74]	RGB
[17]	NA
[71]	RGB
[119]	RGB
[52]	3D models
[115]	RGB
[124]	RGB
[96]	RGB
[125]	RGB, Depth RGB
[169]	RGB
[121]	RGB
[72]	RGB
[127]	RGB

TABLE 10: Datasets for Sign Language Based on Hand Gesture Linguistic Content

Reference(s)	#Input Modality
[197]	2D and 3D skeletal representations, depth data
[64]	RGB
[126]	RGB, RGB-D Kinect
[44]	intensity camera , Kinect
[39]	RGB image, Kinect
[76]	RGBD
[65]	RGB Video
[3]	RGB
[129]	6D IMU data
[116]	RGB
[117]	RGB
[134]	RGB videos
[48]	RGB, Kinect

of the tests, the results can generally be compared across different studies with only some reservations; in general, many methods performed reasonably well and recognized more than 90% of the displayed signs. In some cases, the reported effectiveness was above 97%, but this usually involved less complex tasks and often could not be maintained over multiple datasets. For continuous SLR tasks, recognition rates above 80% can be considered very strong, especially when they are consistent over multiple datasets.

A notable trend found was that almost all algorithms exhibited mixed performance from one sign to another, and in general, only a handful of confusing signs were typically responsible for a large portion of false recognitions. These frequent mistakes often persisted regardless of the classifier or training procedure, and were caused either by the similarity between hand gestures or other systemic factors. This finding implies that certain difficulties in the structure and form of sign language, rather than methodological deficiencies, are impeding the construction of more effective tools and serves as a reminder that currently available SLR algorithms are still prone to error and must be constantly compared with human-created estimations to avoid miscommunication.

In general, the performance of the suggested models is typically evaluated in terms of capacity for correct execution of the primary task, i.e., sign language recognition or translation. Average accuracy for the entire dataset is given as the main indicator of model performance, with a higher percentage indicative of a more accurate system. In some cases, top-1, top-5, and top-10 accuracy were calculated, expressing the model's ability to identify 'most likely' candidates rather than one correct answer. A BLUE score was used to assess the quantitative output of translation models with values between 0 and 100 as depicted in Table 15, while qualitative analysis was based on comparison with ground

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 11: Datasets for Sign Language Based on Linguistic Content in Words and Sentences

Reference(s)	#Input Modality
[37]	RGB video
[185]	Kinect
[189]	Video
[157]	RGB image extracted from video
[191]	Video, Kinect
[190]	Video
[187]	RGB video, depth video, 3D skeletal data, facial features
[41]	RGB video, Kinect, 3D skeletal data
[195]	Kinect, RGB image, skeletal data
[50]	RGB video
[49]	RGB, Kinect, Skeleton point data
[128]	Infrared
[133]	RGB
[66]	RGB
[3]	RGB
[37]	RGB Video
[49]	RGB, depth, skeleton
[193]	Video
[68]	NA
[130]	RGB, Kinect
[69]	RGB Video
[70]	RGB Video
[47]	RGB Video
[67]	RGB, Kinect
[158]	RGB from two angles, Video

TABLE 12: Datasets for Sign Language Based on Other Linguistic Content

Reference(s)	#Input Modality
[37]	RGB video
[185]	Kinect
[189]	Video
[157]	RGB image extracted from video
[191]	Video, Kinect
[190]	Video
[187]	RGB video, depth video, 3D skeletal data, facial features
[41]	RGB video, Kinect, 3D skeletal data
[195]	Kinect, RGB image, skeletal data
[50]	RGB video
[49]	RGB, Kinect, Skeleton point data
[128]	Infrared
[133]	RGB
[66]	RGB
[3]	RGB
[37]	RGB Video
[49]	RGB, depth, skeleton
[193]	Video
[68]	NA
[130]	RGB, Kinect
[69]	RGB Video
[70]	RGB Video
[47]	RGB Video
[67]	RGB, Kinect
[158]	RGB from two angles, Video

TABLE 13: State-of-the-art Sign Language Recognition Accuracy Results

Reference(s)	Accuracy in %
Konstantinidis et al. [37]	99.84
Aujeszky and Eid [185]	92
Sun et al. [196]	97.3
Magar and Parajuli [188]	96
Hossen et al. [45]	84.68
Mao et al. [195]	94.7
Xue et al. [189]	98
Devineau et al. [197]	91.28
Mocialov et al. [191]	95
Sabyrov et al. [193]	73
Alzohairi et al. [186]	63.56
Lim et al. [190]	89.33
Elpeltagy et al. [187]	55.57
Ko et al. [158]	93.28
Wu and Shao [38]	70.1
Rioux-Maldague and Giguere [44]	77
Tang et al. [39]	98.12
Koller et al. [77]	55.7
Li et al. [76]	99.1
Huang et al. [41]	94.2
Huang et al. [43]	98.9
Ameen and Vadera [73]	80.34
Guo et al. [50]	92.9
Aly et al. [13]	88.7
Joy et al. [115]	97
Ko et al. [158]	93.28
Papadimitriou and Potamianos [95]	99.56
Cayamcela and Lim [120]	99.39
Huang et al. [49]	88.7
Cui et al. [3]	91.93
Aly et al. [40]	99.5
Koller et al. [63]	62.8
Li et al. [42]	87.4
Oyedotun and Khashman [74]	92.83

truth as interpreted by human operators. A combination of accuracy and training sample size is used to construct the learning curve, which demonstrates how the performance changes as the volume of training sample increases. Word error rate (WER) analysis is conducted in some studies, as shown in Table 16, to determine which glosses are most confused with each other.

To be effective, the neural classifier must first be trained on data resembling the samples it will encounter during testing and/or practical use. The training data usually involve a basic group of sign language characters, words, or sentences presented in a format that the system was designed to decipher, which is typically annotated by human observers. After training is conducted, the model can be used to deduce the sign language elements in the same format with varying degrees of accuracy. In some studies, several classifiers were tested on the same tasks to evaluate their relative strengths and weaknesses, while in others the focus was on discovering the most suitable combinations of features. The capability of the neural model is generally limited to the signs learned from the training set, but some generalization regarding different people displaying the same sign can be achieved. Therefore, the optimization of training parameters is one of the most important elements of SLR research and can have tremendous impact on the utility value of the proposed solutions. More advanced systems aim to develop real-time translation capacity and to interpret more complex segments of continuous sign language speech. Such applications are vastly more complex than simple recognition of alphabetic characters or isolated words, and they frequently have to analyze multiple signs together to understand the meaning behind a given sequence. In response, researchers have to deploy hybrid architectures and sophisticated sequence-to-sequence models intended to capture semantic nuances and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 14: State-of-the-art Sign Language Recognition Accuracy Results

Reference(s)	Accuracy in %
Fang et al. [128]	94.5
Pigou et al. [126]	73.3
Kavarthapu and Mitra [129]	97.7
Zimmermann and Brox [75]	81.7
Yang and Zhu [65]	99
Konstantinidis et al. [37]	70
Islam et al. [194]	95
Taskiran et al. [116]	98.05
Daroya et al. [117]	90.3
Jalal et al. [119]	99.74
Bantupalli and Xie [134]	91
Chong and Lee [52]	83.78
Hossen et al. [45]	84.68
Balayn et al. [67]	53
Rao et al. [123]	92.88
Ma et al. [96]	83.72
Sajanraj and Beena [124]	99.56
Rastgoo et al. [125]	99.31
Shahriar et al. [118]	94.7
Rakun et al. [130]	77
Rao and Kishore [68]	90
Konstantinidis et al. [70]	98.09
Gunawan et al. [47]	100
Soodtoetong and Gedkhaw [48]	92.24
Cui et al. [3]	91.93
Huang et al. [49]	88.7
Shahin and Almotairi [121]	99.48
Jiang and Zhang [71]	88.1
Nguyen and Do [72]	98.36
Adimas et al. [132]	98.81
Nakjai et al. [127]	87.31
Mathieu et al. [138]	92.92
Amit et al. [142]	98.4
Bansal et al. [84]	71.9
Roy et al. [88]	77.75
Meng et al. [87]	98.08
Gajurel et al. [85]	46.96
Mathieu et al. [86]	74.7
Agelos et al. [83]	94.77

avoid confusing similar signs.

IX. OPEN ISSUES AND FUTURE DIRECTIONS

After reviewing numerous studies related to SLR, the most obvious weakness is the fragmented nature of research in this field. Many research teams have achieved promising results using a wide variety of approaches, but there is little overlap among these studies, and the joint utilization of multiple effective tools is slow to emerge. The lack of a general consensus regarding the most valuable features and the optimal neural network architecture may be an obstacle to achieving better practical results. Recognition of continuous sign language speech remains a considerable challenge, and even the best automated systems struggle with linguistic nuances that can be expressed in sign language sentences. This may be partially a consequence of the fact that most available datasets contain only limited vocabularies and simple sentences, while training models for advanced linguistic tasks requires far more extensive libraries containing diverse examples.

Understanding sign language communication remains a formidable challenge for automated systems. On closer observation, the reasons for the continued inability of machines to consistently and accurately interpret sign language sequences are not as mysterious as they appear to be at first glance. Any natural language features a complex interplay of many rules and relationships, which are difficult to summarize in a mathematical format that can be programmed into computers. This explains why the current generation of sign language recognition (SLR) tools fares quite well with alphabetic characters and simple words and phrases, yet struggles to handle continuous streams of communications such as conversations and narratives.

These shortcomings will almost certainly be remedied in the future, as this field is regarded as socially relevant and consequently receives significant attention from some of the world's most accomplished research teams. It may be argued that the upcoming period is critical for overcoming some of the obstacles that stand in the way of more rapid progress. Some of the main areas where it would be reasonable to expect significant changes over the next few years include the following.

A. TYPE OF INPUT

A majority of reviewed models make use of depth imaging, although some are focused on the RGB images with a higher amount of details to facilitate efficient SLR. Sequential information has been useful as well, most commonly for tracking objects and scenes, along with information about the skeleton (i.e. joint positions). Thermal imaging is less frequently used for SLR, but can bring additional gains when combined with some of the basic types of information such as images. On the level of signs, there is a distinction between static and dynamic signs, with the latter group having a subgroup used in continuous SLR. Based on the current trends, it can be assumed that continuous video and complex signs will become a key focus of study in the next period. It appears that all the preconditions are in place for this shift of focus to occur.

B. GLOBAL RESOURCE BASE

One of the overarching themes in SLR research is the chronic lack of high-quality input databases. Large and diverse datasets are available only for American Sign Language and a few other variations, such as Chinese or Indian, but researchers in smaller countries lack the samples needed for model training and testing. This is slowly changing as the volume of study into SLR continues to grow, and the accumulated resources are becoming sufficient to support the next wave of research. While the situation is certainly improving, it remains difficult to test more advanced applications that require large vocabularies to demonstrate the full capacities of existing or future methods. On the other hand, direct collaboration among research teams and more proactive sharing of available resources could alleviate current issues to a considerable degree and provide a blueprint for more impactful networking. Sign language recognition is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

TABLE 15: Bilingual Evaluation Understudy Score Comparison

Reference(s)	Dev.BLEU-4	Dev.BLEU-3	Dev.BLEU-2	Dev.BLEU-1	Test.BLEU-4	Test.BLEU-3	Test.BLEU-2	Test.BLEU-1
Jiang et al. [140]	10.91	14.01	19.53	31.97	10.35	13.49	19.11	31.86
Yin Kayo et al. [94]	-	-	-	3.2	-	-	-	22.17
Saunders et al. [90]	11.54	14.48	19.63	30.94	11.68	14.55	19.70	31.56
Camgoz et al. [136]	22.12	26.83	34.03	46.56	21.80	26.75	34.46	47.20
Kayo et al. [137]	24.68	29.81	37.60	50.31	25.40	30.58	38.36	50.63

TABLE 16: Word Error Rate Comparison

Reference(s)	Val.Del	Val.Ins	Val.WER	Test.Del	Test.Ins	Test.WER
Cui et al. [133]	13.7	7.3	39.4	12.2	7.5	38.7
Camgoz et al. [64]	20.6	3.2	43.9	19.8	3.2	43.1
Wang et al. [66]	12.8	5.2	37.9	11.9	5.6	37.8
Kumar et al. [131]	-	-	-	-	-	43.7
Ilias et al. [139]	-	-	23.7	-	-	23.4

a worldwide problem, and the only way to resolve it requires a truly global effort.

On the other hand, there are numerous regional variations of sign language, relying on unique combinations of hand and facial gestures to express meaning. For this reason, there is a clear need for high-quality datasets including all relevant input modalities. With regard to hand signs, there is currently a lack of adequate labeled sets that would enable testing of SLR tools under natural conditions, but this has been changing recently. Hopefully, improving datasets will eventually facilitate development of practically applicable SLR models. For this, it's necessary to label longer parts of sign language speech, not just individual elements as is mostly the case right now. Basically, new datasets need to reflect the diversity of communication with sign language so that newly developed methods could be a step closer to reality. Real communications are continuous and without artificial limits, and modern SLR tools must be able to handle long sequences of signs without problems. With use of deep learning networks entering a mature stage, this ambitious goal could be reached soon.

C. COMBINING DIFFERENT FEATURES

This issue has been addressed by many studies, but many difficulties still need to be worked out. It can be desirable to combine features when trying to describe multiple parts of the human body. This issue is typically complicated by the fact that data can be in different formats and include textual elements, images of different types, depth and skeleton data, etc. Fusing some of this data together can result in improved feature engineering and consequently a more accurate model. Three main areas of the body where such features are concentrated include hands, facial region, and the torso. Limiting the attention to hands only can result in imperfect models that fail to properly interpret some of the signs.

Specific areas relevant for success in this regard include detection of hand position, estimation of hand shapes and gestures, real time following of hand movement and similar tasks, and in many ways all of those tasks can present problems. For example, there is extremely high variability in the size and shapes of hands of different signers, while on the other hand different fingers can look very similar and sometimes block each other from view. Ambient conditions including the amount of available light also come into play. Those difficulties are magnified when input images are in low resolution of there are interfering objects, and when complicated gestures need to be analyzed. To alleviate some of those concerns, researchers resort to feature fusion and include facial characteristics into the mix.

On the other hand, rapid motion of the face and neck during sign language use present their own challenges, including partial blocking of key areas. The third group of features – those related to signer's body – can be added as well and bring an additional recognition improvement. Hence, versatile models capable of leveraging features from different parts of the body have an edge and present a better starting point for future research.

D. SEQUENCE SLR MODELS

While notable success has been attained in the realm of isolated SLR, where the algorithms merely have to recognize the single alphabetic sign or word, the same cannot be said for continuous SLR, which involves interpretation of longer segments of speech. Contextual relationships among signs have a strong impact on the meaning of the sentences, so this task cannot be reduced to the recognition of individual gestures.

Contemporary attempts to develop continuous SLR capacity have demonstrated only limited effectiveness and frequently make mistakes when sophisticated analysis of semantic details is required. This is obviously one of the hot topics of SLR research that will continue to be examined in many different ways, searching for a configuration that can overcome the difficulties preventing the emergence of highly effective tools. Based on current findings, we believe that research in this area will focus on deeper and more complex neural network models that employ additional layers and combine several types of layer compositions to gain additional processing power.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3110912, IEEE Access

E. IMPROVING RECOGNITION ACCURACY

To be used commercially and trusted by everyday users, any technology must be highly reliable (>99%) and highly consistent. This is not the case with the SLR applications available today, as they typically still report a small but persistent percentage of false positives and false negatives. The rate of incorrectly recognized items increases as the size of the vocabulary and the complexity of the tasks increases, which is why very few SLR tools are currently deployed in practice. Some of the proposed solutions are conceptually sound and appear suitable for further development, but they are often created by small teams that lack the resources to conduct large-scale testing and refine the training procedures. For the next phase, it is necessary to summon broader support and gather sufficient funds and resources to make high-level accuracy optimization possible. The systems will have to be tested under a wide variety of settings and deliver reasonably useful results even when the external conditions are less than ideal (for example, input images taken under poor lighting conditions).

F. IMPROVING THE EFFICIENCY OF SLR SOLUTIONS

In the past, the focus of scientific research has been on developing the fundamental capability to meaningfully connect observed hand and body gestures and fixed units of sign language. While this is understandable for an early stage of scientific examination, it is will necessary to increase attention on the usability dimension in the future. Some of the earliest SLR solutions required body-worn sensors and other bulky equipment, but newer systems are considerably less demanding and may include only a few miniature cameras. Interaction between the user and the system is another topic that will have to be considered more seriously in the future, with the idea of providing the user with a level of control over the software used by a system. It is equally important to create feedback mechanisms that allow for the instant discovery of common errors while ensuring that user opinions are valued. The previous period of discoveries has renewed interest in SLR research and has resulted in many competing theoretical postulations.

While deep learning networks are broadly accepted as the most suitable technology for tackling this difficult linguistic problem, there is much work needed before fully automated systems capable of understanding continuous streams of sign language communication can be created. In the upcoming decade, we can expect some of the already known solutions to mature to the point where their accuracy is nearly perfect, and it is possible that a major breakthrough will occur at any given point. As the SLR methods become more reliable, it is nearly a given that more creative and meaningful mainstream applications will emerge, delivering direct benefits to the entire population of hearing/speech-impaired people anywhere in the world.

X. CONCLUSION

There is little doubt that the current momentum in the field of sign language recognition will continue into the foreseeable future – the number of potential beneficiaries of such solutions is simply too great to ignore. Recent advances in this area have been largely fueled by the use of deep learning models, which are currently being perfected and will only become more widely accepted in the coming years. Over the past decade, many original and highly innovative suggestions have been used to build SLR tools by extracting features from sensor data or visual streams and feeding them into neural classifiers.

In this paper, we covered most of the currently known methods for SLR tasks based on deep neural architectures that were developed over the past several years, and divided them into clusters based on their chief traits. There is a multitude of options in this regard, as this area of research has been attracting a lot of attention lately. The most common design deploys a CNN network to derive discriminative features from raw data, since this type of network offers the best properties for this task. When information is collected in multimedia format, some of the architectures that can be used include Long Short Term Memory, Recurrent Neural Networks, and GRU. In many cases, multiple types of networks were combined in order to improve final performance. Those models are capable of processing information from different sources and in different formats, including still images, depth information, thermal scans, skeletal data and sequential information have all been used with success.

Some of the proposed models were demonstrated to be very effective, albeit on tasks with limited scope. Studies come from all parts of the world and address many different variations of sign language, which is very important toward ensuring global coverage. Despite some remaining issues, it is fair to conclude that the scientific community is making steady progress toward developing real-time, two-way translation systems that can eventually be deployed in the real world. Before this occurs, it will be necessary to achieve more consistent performance and eliminate some common confusion points (where most algorithms tend to misinterpret an intended sign).

Some hybrid models are emerging that combine the best characteristics of several types of neural networks, and solutions of this type may represent the most logical path forward with respect to advanced SLR applications. It is reasonable to expect breakthroughs in this field in the future, and many of the research studies may include key elements that will eventually become a part of the final solution to automated sign language recognition. Even at this stage, many SLR tools can be practically used to some extent, and can provide immediate relief to disabled people as well as point to future directions of research.

REFERENCES

[1] T. Kim, J. Keane, W. Wang, H. Tang, J. Riggle, G. Shakhnarovich, D. Brentari, and K. Livescu, "Lexicon-free fingerspelling recognition

from video: Data, models, and signer adaptation," *Computer Speech & Language*, vol. 46, pp. 209–232, 2017.

[2] M. A. Ahmed, B. B. Zaidan, A. A. Zaidan, M. M. Salih, and M. M. b. Lakulu, "A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017," *Sensors*, vol. 18, no. 7, p. 2208, 2018.

[3] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019.

[4] P. S. Santhalingam, P. Pathak, J. Košecká, H. Rangwala *et al.*, "Sign language recognition analysis using multimodal data," in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2019, pp. 203–210.

[5] A. C. Duarte, "Cross-modal neural sign language translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1650–1654.

[6] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019.

[7] Q. Xiao, Y. Zhao, and W. Huan, "Multi-sensor data fusion for sign language recognition based on dynamic bayesian network and convolutional neural network," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 335–15 352, 2019.

[8] E. K. Kumar, P. Kishore, M. T. K. Kumar, and D. A. Kumar, "3d sign language recognition with joint distance and angular coded color topographical descriptor on a 2–stream cnn," *Neurocomputing*, vol. 372, pp. 40–54, 2020.

[9] J. Wu and R. Jafari, "Wearable computers for sign language recognition," in *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Springer, 2017, pp. 379–401.

[10] J. Shang and J. Wu, "A robust sign language recognition system with multiple wi-fi devices," in *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*, ser. MobiArch '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 19–24. [Online]. Available: https://doi.org/10.1145/3097620.3097624

[11] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4165–4174.

[12] Q. Xiao, M. Qin, and Y. Yin, "Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people," *Neural Networks*, 2020.

[13] W. Aly, S. Aly, and S. Almotairi, "User-independent american sign language alphabet recognition based on depth image and pcanet features," *IEEE Access*, vol. 7, pp. 123 138–123 150, 2019.

[14] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 279–286.

[15] S. Wei, X. Chen, X. Yang, S. Cao, and X. Zhang, "A component-based vocabulary-extensible sign language gesture recognition framework," *Sensors*, vol. 16, no. 4, p. 556, 2016.

[16] N. B. Ibrahim, H. H. Zayed, and M. M. Selim, "Advances, challenges and opportunities in continuous sign language recognition," *Journal of Engineering and Applied Sciences*, vol. 15, no. 5, pp. 1205–1227, 2020.

[17] L. Zheng, B. Liang, and A. Jiang, "Recent advances of deep learning for sign language recognition," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2017, pp. 1–7.

[18] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[19] F.-S. Chen, C.-M. Fu, and C.-L. Huang, "Hand gesture recognition using a real-time tracking method and hidden markov models," *Image and vision computing*, vol. 21, no. 8, pp. 745–758, 2003.

[20] X. Wang, M. Xia, H. Cai, Y. Gao, and C. Cattani, "Hidden-markov-models-based dynamic hand gesture recognition," *Mathematical Problems in Engineering*, vol. 2012, 2012.

[21] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and kinect devices," in *2014 IEEE International conference on image processing (ICIP)*. IEEE, 2014, pp. 1565–1569.

[22] X. Lu, B. Qi, H. Qian, Y. Gao, J. Sun, and J. Liu, "Kinect-based human finger tracking method for natural haptic rendering," *Entertainment Computing*, vol. 33, p. 100335, 2020.

[23] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2sign: Towards sign language production using neural machine translation and generative adversarial networks," *International Journal of Computer Vision*, pp. 1–18, 2020.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] E. Escobedo, L. Ramirez, and G. Camara, "Dynamic sign language recognition based on convolutional neural networks and texture maps," in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2019, pp. 265–272.

[26] S. Hayani, M. Benaddy, O. El Meslouhi, and M. Kardouchi, "Arab sign language recognition with convolutional neural networks," in *2019 International Conference of Computer Science and Renewable Energies (ICCSRE)*. IEEE, 2019, pp. 1–4.

[27] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38 044–38 054, 2019.

[28] P. Witoonchart and P. Chongstitvatana, "Application of structured support vector machine backpropagation to a convolutional neural network for human pose estimation," *Neural Networks*, vol. 92, pp. 39 – 46, 2017, advances in Cognitive Engineering Using Neural Networks. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608017300321

[29] S. He, "Research of a sign language translation system based on deep learning," in *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Oct 2019, pp. 392–396.

[30] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR 2011*. IEEE, 2011, pp. 1385–1392.

[31] P. Q. Thang, N. T. Thuy, and H. T. Lam, "The svm, simpsvm and rvm on sign language recognition problem," in *2017 Seventh International Conference on Information Science and Technology (ICIST)*, April 2017, pp. 398–403.

[32] R. A. A. R. Agha, M. N. Sefer, and P. Fattah, "A comprehensive study on sign languages recognition systems using (svm, knn, cnn and ann)," in *Proceedings of the First International Conference on Data Science, E-learning and Information Systems*, 2018, pp. 1–6.

[33] E. Alpaydin, *Introduction to machine learning*, 2020.

[34] A. Wadhawan and P. Kumar, "Sign language recognition systems: A decade systematic literature review," *Archives of Computational Methods in Engineering*, pp. 1–29, 2019.

[35] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data*. AMLBook New York, NY, USA:, 2012, vol. 4.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "A deep learning approach for analyzing video and skeletal features in sign language recognition," in *2018 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2018, pp. 1–6.

[38] D. Wu and L. Shao, "Multimodal dynamic networks for gesture recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 945–948.

[39] A. Tang, K. Lu, Y. Wang, J. Huang, and H. Li, "A real-time hand posture recognition system using deep neural networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 2, pp. 1–23, 2015.

[40] S. Aly, B. Osman, W. Aly, and M. Saber, "Arabic sign language fingerspelling recognition from depth and intensity images," in *2016 12th International Computer Engineering Conference (ICENCO)*. IEEE, 2016, pp. 99–104.

[41] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3d convolutional neural networks," in *2015 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2015, pp. 1–6.

[42] K. Li, Z. Zhou, and C.-H. Lee, "Sign transition modeling and a scalable solution to continuous sign language recognition for real-world applications," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 8, no. 2, pp. 1–23, 2016.

[43] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using real-sense," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, 2015, pp. 166–170.

[44] L. Rioux-Maldague and P. Giguere, "Sign language fingerspelling classification from depth and color images using a deep belief network," in *2014 Canadian Conference on Computer and Robot Vision*. IEEE, 2014, pp. 92–97.

[45] M. Hossen, A. Govindaiah, S. Sultana, and A. Bhuiyan, "Bengali sign language recognition using deep convolutional neural network," in *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2018, pp. 369–373.

[46] P. Escudeiro, N. Escudeiro, R. Reis, J. Lopes, M. Norberto, A. B. Baltasar, M. Barbosa, and J. Bidarra, "Virtual sign–a real time bidirectional translator of portuguese sign language," *Procedia Computer Science*, vol. 67, pp. 252–262, 2015.

[47] H. Gunawan, N. Thiracitta, A. Nugroho *et al.*, "Sign language recognition using modified convolutional neural network model," in *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*. IEEE, 2018, pp. 1–5.

[48] N. Soodtoetong and E. Gedkhaw, "The efficiency of sign language recognition using 3d convolutional neural networks," in *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 2018, pp. 70–73.

[49] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3d-cnns for large-vocabulary sign language recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 9, pp. 2822–2832, 2018.

[50] D. Guo, W. Zhou, A. Li, H. Li, and M. Wang, "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation," *IEEE Transactions on Image Processing*, vol. 29, pp. 1575–1590, 2019.

[51] N. Wang, Z. Ma, Y. Tang, Y. Liu, Y. Li, and J. Niu, "An optimized scheme of mel frequency cepstral coefficient for multi-sensor sign language recognition," in *International Conference on Smart Computing and Communication*. Springer, 2016, pp. 224–235.

[52] T.-W. Chong and B.-G. Lee, "American sign language recognition using leap motion controller with machine learning approach," *Sensors*, vol. 18, no. 10, p. 3554, 2018.

[53] W. Jingqiu and Z. Ting, "An arm-based embedded gesture recognition system using a data glove," in *The 26th Chinese Control and Decision Conference (2014 CCDC)*. IEEE, 2014, pp. 1580–1584.

[54] A. Z. Shukor, M. F. Miskon, M. H. Jamaluddin, F. bin Ali, M. F. Asyraf, M. B. bin Bahar *et al.*, "A new data glove approach for malaysian sign language detection," *Procedia Computer Science*, vol. 76, pp. 60–67, 2015.

[55] N. B. Ibrahim, M. M. Selim, and H. H. Zayed, "An automatic arabic sign language recognition system (arslrs)," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 470–477, 2018.

[56] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 572–578.

[57] S. G. M. Almeida, F. G. Guimarães, and J. A. Ramírez, "Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7259–7271, 2014.

[58] B. Hisham and A. Hamouda, "Arabic sign language recognition using ada-boosting based on a leap motion controller," *International Journal of Information Technology*, vol. 13, no. 3, pp. 1221–1234, 2021.

[59] U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, and A. Abid, "Advances in machine translation for sign language: approaches, limitations, and challenges," *Neural Computing and Applications*, pp. 1–43, 2021.

[60] M. I. Sadek, M. N. Mikhael, and H. A. Mansour, "A new approach for designing a smart glove for arabic sign language recognition system based on the statistical analysis of the sign language," in *2017 34th National Radio Science Conference (NRSC)*. IEEE, 2017, pp. 380–388.

[61] N. Rossol, I. Cheng, and A. Basu, "A multisensor technique for gesture recognition through intelligent skeletal pose analysis," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 350–359, 2015.

[62] L. Quesada, G. López, and L. Guerrero, "Automatic recognition of the american sign language fingerspelling alphabet to assist people living with speech or hearing impairments," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 4, pp. 625–635, 2017.

[63] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3793–3802.

[64] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3075–3084.

[65] S. Yang and Q. Zhu, "Video-based chinese sign language recognition using convolutional neural network," in *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*. IEEE, 2017, pp. 929–934.

[66] S. Wang, D. Guo, W.-g. Zhou, Z.-J. Zha, and M. Wang, "Connectionist temporal fusion for sign language translation," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1483–1491.

[67] A. Balayn, H. Brock, and K. Nakadai, "Data-driven development of virtual sign language communication agents," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 370–377.

[68] G. A. Rao and P. Kishore, "Selfie sign language recognition with multiple features on adaboost multilabel multiclass classifier," *Journal of Engineering Science and Technology*, vol. 13, no. 8, pp. 2352–2368, 2018.

[69] M. C. Ariesta, F. Wiryana, A. Zahra *et al.*, "Sentence level indonesian sign language recognition using 3d convolutional neural network and bidirectional recurrent neural network," in *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*. IEEE, 2018, pp. 16–22.

[70] D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Sign language recognition based on hand and body skeletal data," in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE, 2018, pp. 1–4.

[71] X. Jiang and Y.-D. Zhang, "Chinese sign language fingerspelling via six-layer convolutional neural network with leaky rectified linear units for therapy and rehabilitation," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 9, pp. 2031–2090, 2019.

[72] H. B. Nguyen and H. N. Do, "Deep learning for american sign language fingerspelling recognition system," in *2019 26th International Conference on Telecommunications (ICT)*. IEEE, 2019, pp. 314–318.

[73] S. Ameen and S. Vadera, "A convolutional neural network to classify american sign language fingerspelling from depth and colour images," *Expert Systems*, vol. 34, no. 3, p. e12197, 2017.

[74] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941–3951, 2017.

[75] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4903–4911.

[76] S.-Z. Li, B. Yu, W. Wu, S.-Z. Su, and R.-R. Ji, "Feature learning based on sae–pca network for human gesture recognition in rgbd images," *Neurocomputing*, vol. 151, pp. 565–573, 2015.

[77] O. Koller, H. Ney, and R. Bowden, "Deep learning of mouth shapes for sign language," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 85–91.

[78] S. Kim, Y. Ban, and S. Lee, "Tracking and classification of in-air hand gesture based on thermal guided joint filter," *Sensors*, vol. 17, no. 1, p. 166, 2017.

[79] T. Xu, D. An, Z. Wang, S. Jiang, C. Meng, Y. Zhang, Q. Wang, Z. Pan, and Y. Yue, "3d joints estimation of the human body in single-frame point cloud," *IEEE Access*, vol. 8, pp. 178 900–178 908, 2020.

[80] W. Wong, F. H. Juwono, and B. T. T. Khoo, "Multi-features capacitive hand gesture recognition sensor: a machine learning approach," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8441–8450, 2021.

[81] Y. Zhou, G. Jiang, and Y. Lin, "A novel finger and hand pose estimation technique for real-time hand gesture recognition," *Pattern Recognition*, vol. 49, pp. 102–114, 2016.

[82] R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou, "Sign language production: A review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3451–3461.

[83] A. Kratimenos, G. Pavlakos, and P. Maragos, "Independent sign language recognition with 3d body, hands, and face reconstruction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4270–4274.

[84] D. Bansal, P. Ravi, M. So, P. Agrawal, I. Chadha, G. Murugappan, and C. Duke, "Copycat: Using sign language recognition to help deaf children acquire language skills," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–10.

[85] K. Gajurel, C. Zhong, and G. Wang, "A fine-grained visual attention approach for fingerspelling recognition in the wild," *arXiv preprint arXiv:2105.07625*, 2021.

[86] M. De Coster, M. Van Herreweghe, and J. Dambre, "Sign language recognition with transformer networks," in *12th International Conference*

on Language Resources and Evaluation. European Language Resources Association (ELRA), 2020, pp. 6018–6024.

[87] L. Meng and R. Li, "An attention-enhanced multi-scale and dual sign language recognition network based on a graph convolution network," *Sensors*, vol. 21, no. 4, p. 1120, 2021.

[88] P. P. Roy, P. Kumar, and B.-G. Kim, "An efficient sign language recognition (slr) system using camshift tracker and hidden markov model (hmm)," *SN Computer Science*, vol. 2, no. 2, pp. 1–15, 2021.

[89] N. M. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. Xydopoulos, K. Antzakas, D. Papazachariou, and P. none Daras, "A comprehensive study on deep learning-based methods for sign language recognition," *IEEE Transactions on Multimedia*, 2021.

[90] B. Saunders, N. C. Camgoz, and R. Bowden, "Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks," *International Journal of Computer Vision*, pp. 1–23, 2021.

[91] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[92] B. Butcher and B. J. Smith, "Feature engineering and selection: A practical approach for predictive models: by max kuhn and kjell johnson. boca raton, fl: Chapman and hall crc press, 2019, isbn: 978-1-13-807922-9." 2020.

[93] A. J. Ferreira and M. A. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1794–1804, 2012.

[94] K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, and M. Alikhani, "Including signed languages in natural language processing," *arXiv preprint arXiv:2105.05222*, 2021.

[95] K. Papadimitriou and G. Potamianos, "Fingerspelled alphabet sign recognition in upper-body videos," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[96] M. Ma, X. Xu, J. Wu, and M. Guo, "Design and analyze the structure based on deep belief network for gesture recognition," in *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 2018, pp. 40–44.

[97] S. M. Kamal, Y. Chen, S. Li, X. Shi, and J. Zheng, "Technical approaches to chinese sign language processing: A review," *IEEE Access*, vol. 7, pp. 96 926–96 935, 2019.

[98] R.-H. Liang and M. Ouhyoung, "A sign language recognition system using hidden markov model and context sensitive search," in *Proceedings of the ACM symposium on virtual reality software and technology*, 1996, pp. 59–66.

[99] K. Grobel and M. Assan, "Isolated sign language recognition using hidden markov models," in *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, vol. 1. IEEE, 1997, pp. 162–167.

[100] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1997, pp. 994–999.

[101] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," *Machine learning*, vol. 29, no. 2, pp. 245–273, 1997.

[102] A. D. Wilson and A. F. Bobick, "Parametric hidden markov models for gesture recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 9, pp. 884–900, 1999.

[103] C. Vogler and D. Metaxas, "Parallel hidden markov models for american sign language recognition," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1. IEEE, 1999, pp. 116–122.

[104] E.-J. Ong, O. Koller, N. Pugeault, and R. Bowden, "Sign spotting using hierarchical sequential patterns with temporal intervals," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1923–1930.

[105] Y. Bengio and P. Frasconi, "Input-output hmms for sequence processing," *IEEE Transactions on Neural Networks*, vol. 7, no. 5, pp. 1231–1249, 1996.

[106] A. Just, O. Bernier, and S. Marcel, "Hmm and iohmm for the recognition of mono-and bi-manual 3d hand gestures," IDIAP, Tech. Rep., 2004.

[107] C. Keskin and L. Akarun, "Stars: Sign tracking and recognition system using input–output hmms," *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1086–1095, 2009.

[108] N. Liu and B. C. Lovell, "Gesture classification using hidden markov models and viterbi path counting," in *VIIth Digital image computing: techniques and Applications*, 2003.

[109] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden markov model-based continuous gesture recognition system for hand motion trajectory," in *2008 19th international conference on pattern recognition*. IEEE, 2008, pp. 1–4.

[110] J. Appenrodt, A. Al-Hamadi, and B. Michaelis, "Data gathering for gesture recognition systems based on single color-, stereo color-and thermal cameras," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 3, no. 1, pp. 37–50, 2010.

[111] M. M. Zaki and S. I. Shaheen, "Sign language recognition using a combination of new vision based features," *Pattern Recognition Letters*, vol. 32, no. 4, pp. 572–577, 2011.

[112] P. V. Barros, N. T. Júnior, J. M. Bisneto, B. J. Fernandes, B. L. Bezerra, and S. M. Fernandes, "An effective dynamic gesture recognition system based on the feature vector reduction for surf and lcs," in *International Conference on Artificial Neural Networks*. Springer, 2013, pp. 412–419.

[113] W. Yang, J. Tao, and Z. Ye, "Continuous sign language recognition using level building based on fast hidden markov model," *Pattern Recognition Letters*, vol. 78, pp. 28–35, 2016.

[114] S. Belgacem, C. Chatelain, and T. Paquet, "Gesture sequence recognition with one shot learned crf/hmm hybrid model," *Image and Vision Computing*, vol. 61, pp. 12–21, 2017.

[115] J. Joy, K. Balakrishnan, and M. Sreeraj, "Signquiz: A quiz based tool for learning fingerspelled signs in indian sign language using aslr," *IEEE Access*, vol. 7, pp. 28 363–28 371, 2019.

[116] M. Taskiran, M. Killioglu, and N. Kahraman, "A real-time system for recognition of american sign language by using deep learning," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2018, pp. 1–5.

[117] R. Daroya, D. Peralta, and P. Naval, "Alphabet sign language image classification using deep learning," in *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, 2018, pp. 0646–0650.

[118] S. Shahriar, A. Siddiquee, T. Islam, A. Ghosh, R. Chakraborty, A. I. Khan, C. Shahnaz, and S. A. Fattah, "Real-time american sign language recognition using skin segmentation and image category classification with convolutional neural network and deep learning," in *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, 2018, pp. 1168–1171.

[119] M. A. Jalal, R. Chen, R. K. Moore, and L. Mihaylova, "American sign language posture understanding with deep neural networks," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 573–579.

[120] M. E. M. Cayamcela and W. Lim, "Fine-tuning a pre-trained convolutional neural network model to translate american sign language in real-time," in *2019 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2019, pp. 100–104.

[121] A. Shahin and S. Almotairi, "Automated arabic sign language recognition system based on deep transfer learning," *International Journal of Computer Science and Network Security*, vol. 19, no. 10, pp. 144–152, 2019.

[122] F. Yasir, P. Prasad, A. Alsadoon, A. Elchouemi, and S. Sreedharan, "Bangla sign language recognition using convolutional neural network," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. IEEE, 2017, pp. 49–53.

[123] G. A. Rao, K. Syamala, P. Kishore, and A. Sastry, "Deep convolutional neural networks for sign language recognition," in *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)*. IEEE, 2018, pp. 194–197.

[124] T. Sajanraj and M. Beena, "Indian sign language numeral recognition using region of interest convolutional neural network," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2018, pp. 636–640.

[125] R. Rastgoo, K. Kiani, and S. Escalera, "Multi-modal deep hand sign language recognition in still images using restricted boltzmann machine," *Entropy*, vol. 20, no. 11, p. 809, 2018.

[126] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3086–3093.

[127] P. Nakjai, P. Maneerat, and T. Katanyukul, "Thai finger spelling localization and classification under complex background using a yolo-based deep learning," in *Proceedings of the 11th International Conference on Computer Modeling and Simulation*, 2019, pp. 230–233.

[128] B. Fang, J. Co, and M. Zhang, "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, 2017, pp. 1–13.

[129] D. C. Kavarthapu and K. Mitra, "Hand gesture sequence recognition using inertial motion units (imus)," in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2017, pp. 953–957.

[130] E. Rakun, A. M. Arymurthy, L. Y. Stefanus, A. F. Wicaksono, and I. W. W. Wisesa, "Recognition of sign language system for indonesian language using long short-term memory neural networks," *Advanced Science Letters*, vol. 24, no. 2, pp. 999–1004, 2018.

[131] S. S. Kumar, T. Wangyal, V. Saboo, and R. Srinath, "Time series neural networks for real time sign language translation," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 243–248.

[132] D. M. Adimas, E. Rakun, and D. Hardianto, "Recognizing indonesian sign language gestures using features generated by elliptical model tracking and angular projection," in *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*. IEEE, 2019, pp. 25–31.

[133] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7361–7369.

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

[135] K. Yin and J. Read, "Attention is all you sign: sign language translation with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshop on Sign Language Recognition, Translation and Production (SLRTP)*, vol. 23, 2020.

[136] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 023–10 033.

[137] K. Yin and J. Read, "Better sign language translation with stmc-transformer," *arXiv preprint arXiv:2004.00588*, 2020.

[138] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from rgb video using pose flow and self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3441–3450.

[139] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, no. 7, p. 2437, 2021.

[140] T. Jiang, N. C. Camgoz, and R. Bowden, "Skeletor: Skeletal transformers for robust body-pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3394–3402.

[141] N. C. Camgoz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden, "Content4all open research sign language translation datasets," *arXiv preprint arXiv:2105.02351*, 2021.

[142] A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgoz, R. Bowden, T. Jiang, A. Rios, M. Muller, and S. Ebling, "Evaluating the immediate applicability of pose estimation for sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3434–3440.

[143] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer vision and image understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[144] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer vision and image understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[145] H. L. Ribeiro and A. Gonzaga, "Hand image segmentation in video sequence by gmm: a comparative analysis," in *2006 19th Brazilian Symposium on Computer Graphics and Image Processing*. IEEE, 2006, pp. 357–364.

[146] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[147] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in *2014 Fourth international conference on advanced computing & communication technologies*. IEEE, 2014, pp. 5–12.

[148] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to arabic sign language recognition," *IEEE transactions on human-machine systems*, vol. 44, no. 4, pp. 551–557, 2014.

[149] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton based sign language recognition using whole-body keypoints," *arXiv preprint arXiv:2103.08833*, 2021.

[150] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for sign language recognition and translation," *IEEE Transactions on Multimedia*, 2021.

[151] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Systems with Applications*, p. 113794, 2020.

[152] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 009–13 016.

[153] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-time sign language detection using human pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 237–248.

[154] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[155] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[156] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[157] S. Gattupalli, A. Ghaderi, and V. Athitsos, "Evaluation of deep learning based pose estimation for sign language recognition," in *Proceedings of the 9th ACM international conference on pervasive technologies related to assistive environments*, 2016, pp. 1–7.

[158] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Applied Sciences*, vol. 9, no. 13, p. 2683, 2019.

[159] M. Madadi, H. Bertiche, and S. Escalera, "Smplr: Deep learning based smpl reverse for 3d human pose and shape recovery," *Pattern Recognition*, vol. 106, p. 107472, 2020.

[160] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, "Modeep: A deep learning framework using motion features for human pose estimation," in *Asian conference on computer vision*. Springer, 2014, pp. 302–315.

[161] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," *Advances in neural information processing systems*, vol. 27, pp. 1736–1744, 2014.

[162] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3063–3072.

[163] Y. Bin, Z.-M. Chen, X.-S. Wei, X. Chen, C. Gao, and N. Sang, "Structure-aware human pose estimation with graph convolutional networks," *Pattern Recognition*, vol. 106, p. 107410, 2020.

[164] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards viewpoint invariant 3d human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 160–177.

[165] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, "Drpose3d: depth ranking in 3d human pose estimation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 978–984.

[166] M. J. Marin-Jimenez, F. J. Romero-Ramirez, R. Munoz-Salinas, and R. Medina-Carnicer, "3d human pose estimation from depth maps using a deep combination of poses," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 627–639, 2018.

[167] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *arXiv preprint arXiv:2012.13392*, 2020.

[168] X. Ji, Q. Fang, J. Dong, Q. Shuai, W. Jiang, and X. Zhou, "A survey on monocular 3d human pose estimation," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 6, pp. 471–500, 2020.

[169] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, and R. AlKhalaf, "Arasl: Arabic alphabets sign language dataset," *Data in brief*, vol. 23, p. 103777, 2019.

[170] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern recognition*, vol. 36, no. 3, pp. 585–601, 2003.

[171] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 1, pp. 13–24, 2009.

[172] N. A. Ibraheem and R. Z. Khan, "Vision based gesture recognition using neural networks approaches: a review," *International Journal of human Computer Interaction (IJHCI)*, vol. 3, no. 1, pp. 1–14, 2012.

[173] Q. Wan, Y. Li, C. Li, and R. Pal, "Gesture recognition for smart home applications using portable radar sensors," in *2014 36th Annual Inter-*

national Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2014, pp. 6414–6417.

[174] R. Lockton, "Hand gesture recognition using computer vision," *4th Year Project Report*, pp. 1–69, 2002.

[175] A. Tharwat, T. Gaber, A. E. Hassanien, M. K. Shahin, and B. Refaat, "Sift-based arabic sign language recognition system," in *Afro-european conference for industrial advancement*. Springer, 2015, pp. 359–370.

[176] C. Manresa, J. Varona, R. Mas, and F. J. Perales, "Hand tracking and gesture recognition for human-computer interaction," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 5, no. 3, pp. 96–104, 2005.

[177] C. Manresa-Yee, J. Varona, R. Mas, and F. J. Perales, "Hand tracking and gesture recognition for human-computer interaction," in *Progress in Computer Vision and Image Analysis*. World Scientific, 2010, pp. 401–412.

[178] T.-Y. Pan, L.-Y. Lo, C.-W. Yeh, J.-W. Li, H.-T. Liu, and M.-C. Hu, "Real-time sign language recognition in complex background scene based on a hierarchical clustering classification method," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*. IEEE, 2016, pp. 64–67.

[179] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 462–477, 2009.

[180] H. Lahiani, M. Elleuch, and M. Kherallah, "Real time hand gesture recognition system for android devices," in *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE, 2015, pp. 591–596.

[181] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[182] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[183] G. Murthy and R. Jadon, "A review of vision based hand gestures recognition," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 405–410, 2009.

[184] A. Chaudhary, J. L. Raheja, K. Das, and S. Raheja, "Intelligent approaches to interact with machines using hand gesture recognition in natural way: a survey," *arXiv preprint arXiv:1303.2292*, 2013.

[185] T. Aujeszky and M. Eid, "A gesture recogintion architecture for arabic sign language communication system," *Multimedia Tools and Applications*, vol. 75, no. 14, pp. 8493–8511, 2016.

[186] R. Alzohairi, R. Alghonaim, W. Alshehri, S. Aloqeely, M. Alzaidan, and O. Bchir, "Image based arabic sign language recognition system," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 3, 2018.

[187] M. Elpeltagy, M. Abdelwahab, M. E. Hussein, A. Shoukry, A. Shoala, and M. Galal, "Multi-modality-based arabic sign language recognition," *IET Computer Vision*, vol. 12, no. 7, pp. 1031–1039, 2018.

[188] A. T. Magar and P. Parajuli, "American sign language recognition using convolution neural network for raspberry pi," EasyChair, Tech. Rep., 2020.

[189] Q. Xue, X. Li, D. Wang, and W. Zhang, "Deep forest-based monocular visual sign language recognition," *Applied Sciences*, vol. 9, no. 9, p. 1945, 2019.

[190] K. M. Lim, A. W. C. Tan, C. P. Lee, and S. C. Tan, "Isolated sign language recognition using convolutional neural network hand modelling and hand energy image," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19 917–19 944, 2019.

[191] B. Mocialov, G. Turner, K. Lohan, and H. Hastie, "Towards continuous sign language recognition with deep learning," in *Proc. of the Workshop on the Creating Meaning With Robot Assistants: The Gap Left by Smart Devices*, 2017.

[192] V. Belissen, "Sign language video analysis for automatic recognition and detection," 2019.

[193] A. Sabyrov, M. Mukushev, and V. Kimmelman, "Towards real-time sign language interpreting robot: Evaluation of non-manual components on recognition accuracy." in *CVPR Workshops*, 2019.

[194] S. Islam, S. S. S. Mousumi, A. S. A. Rabby, S. A. Hossain, and S. Abujar, "A potent model to recognize bangla sign language digits using convolutional neural network," *Procedia computer science*, vol. 143, pp. 611–618, 2018.

[195] C. Mao, S. Huang, X. Li, and Z. Ye, "Chinese sign language recognition with sequence to sequence learning," in *CCF Chinese Conference on Computer Vision*. Springer, 2017, pp. 180–191.

[196] P. Sun, F. Chen, G. Wang, J. Ren, and J. Dong, "A robust static sign language recognition system based on hand key points estimation," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2017, pp. 548–557.

[197] G. Devineau, F. Moutarde, W. Xi, and J. Yang, "Deep learning for hand gesture recognition on skeletal data," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 106–113.

PLACE PHOTO HERE

MUHAMMAD AL-QURISHI (M'16) received the Ph.D. degree from the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2017. He was a Postdoctoral Researcher with the Chair of Pervasive and Mobile Computing (CPMC), CCIS, KSU. He is one of the founding members of CPMC. He is currently a Data Scientist working with the Research and Innovation Department, ELM Company. He has published several articles in refereed journals (IEEE, ACM, Springer, and Wiley). He received an Innovation Award for a Mobile Cloud Serious Game from KSU 2013 and the Best Ph.D. Thesis Award from CCIS, KSU, in 2018. His research interests include Natural Language Processing and Understanding, big data analysis and mining, pervasive computing, and machine learning.

PLACE PHOTO HERE

THARIQ KHALID received the B.Tech degree in Computer Science and Engineering National Institute of Technology, Calicut in 2007. He has done extensive work in the field of Machine learning, Natural Language Processing and Computer Vision at Samsung Research Institute Bangalore, Verse Innovations and Target Corp. Thariq received SPOT award from Samsung for his work in Arabic language model optimization in the year 2014, Gladiator award from Target Corp for his contriibutions in the field of DL at Target Corp in 2018. He is currently working as a Data Scientist with the Research and Innovation department, Elm Company. His research interests include Computer Vision, Perception systems for autonomous driving and Deep Learning.

PLACE PHOTO HERE

RIAD SOUISSI received his MSc degree in the Field Of Study Telecommunication and Information Systems, Ecole Centrale Paris. He is the Director of research department at Elm Company. Currently, his main focus and interest is in different research areas include Deep Learning and Analytics, Creating products and solutions for difficult problems using new technologies (AI, ML, Blockchain, IoT, etc).