Expert Systems With Applications 164 (2021) 113794

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Sign Language Recognition: A Deep Survey

Razieh Rastgoo [a], Kourosh Kiani [a,*], Sergio Escalera [b]

[a] Electrical and Computer Engineering Department, Semnan University, Semnan, 3513119111, Iran
[b] Department of Mathematics and Informatics, Universitat de Barcelona and Computer Vision Center, 585, 08007 Barcelona, Spain

ARTICLE INFO

Keywords:
Sign language recognition
Pose estimation
Deep learning
Computer Vision
Face recognition
Application

ABSTRACT

Sign language, as a different form of the communication language, is important to large groups of people in society. There are different signs in each sign language with variability in hand shape, motion profile, and position of the hand, face, and body parts contributing to each sign. So, visual sign language recognition is a complex research area in computer vision. Many models have been proposed by different researchers with significant improvement by deep learning approaches in recent years. In this survey, we review the vision-based proposed models of sign language recognition using deep learning approaches from the last five years. While the overall trend of the proposed models indicates a significant improvement in recognition accuracy in sign language recognition, there are some challenges yet that need to be solved. We present a taxonomy to categorize the proposed models for isolated and continuous sign language recognition, discussing applications, datasets, hybrid models, complexity, and future lines of research in the field.

1. Introduction

According to the World Federation of the Deaf, there are over 300 sign languages around the world that 70 million deaf people are using them (Murray, 2018). Sign language recognition would help break down the barriers for sign language users in society. Most of the communication technologies have been developed to support spoken or written language (which excludes sign language). While communication technologies and tools such as Imo (Pagebites, 2018) and WhatsApp (Acton & Koum, 2009) have become an important part of our life, deaf people have many problems for using these technologies. Daily communication of the deaf community with the hearing majority community can be facilitated using these technologies. As a result, sign language, as a structural form of the hand gestures involving visual motions and signs, is used as a communication system to help the deaf and speech-impaired community for daily interaction. Sign language involves the usage of different parts of the body, such as fingers, hand, arm, head, body, and facial expression (Cheok, Omar, & Jaward, 2017). There are five main parameters in sign language, which are hand-shape, palm orientation, movement, location, and expression/non-manual signals. To have an accurate sign word, all of these five parameters must be performed correctly. Many applications benefit from sign language recognition advantages such as translation systems, interpreting services, video remote human interpreting, human–computer interaction (Deng, Yang, Zhang, Tan, Chang, & Wang, 2017; Supancic, Rogez, Yang, Shotton, & Ramana, 2018), online hand tracking of human communication in desktop environments (Tagliasacchi, Schröder, Tkach, Bouaziz, Botsch, & Pauly, 2015), real-time multi-person recognition systems (Cao, Simon, Wei, & Sheikh, 2017), games, virtual reality environments, robot controls, and natural language communications (Wadhawan & Kumar, 2020). Besides, since the Red Green Blue Depth (RGBD) cameras, with high capability to produce depth map, have become cheap in recent years, they are widely used to decrease the cost of hand pose recognition systems. Furthermore, most of the well-known and big technology companies, like Google, Microsoft, and Facebook, contributed to some projects in Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR) technology, as new interactive personal computers. This trend has widely broadened the applications of hand sign and pose estimation. Moreover, some research works have been proposed in Human–Computer Interaction (HCI) focusing on hands movement controlling (Wadhawan & Kumar, 2020). So, the development of automatic hand sign language translation systems is necessary to satisfy many applications requirements to promote equal communication opportunity to improve public welfare.

With the advent of deep learning in recent years, many research efforts have been conducted to sign language recognition (Asadi-Aghbolaghi, Clapés, Bellantonio, Jair Escalante, Ponce-López, Baró, Guyon, Kasaei, & Escalera, 2017; Deng et al., 2017; Ferreira, Cardoso, & Rebelo, 2019; Guo, Wang, & Chen, 2017; Lim, Tan, Lee, & Tan, 2019; Oberweger, Wohlhart, & Lepetit, 2015; Rastgoo, Kiani, & Escalera, 2018; Wadhawan & Kumar, 2020; Zheng, Liang, & Jiang, 2017; Zimmerman, Lanier, Blanchard, Bryson, & Harvill, 1987). One

* Corresponding author.
E-mail addresses: rrastgoo@semnan.ac.ir (R. Rastgoo), Kourosh.kiani@semnan.ac.ir (K. Kiani), sescalera@ub.edu (S. Escalera).

https://doi.org/10.1016/j.eswa.2020.113794
Received 20 February 2020; Received in revised form 10 July 2020; Accepted 23 July 2020
Available online 6 August 2020

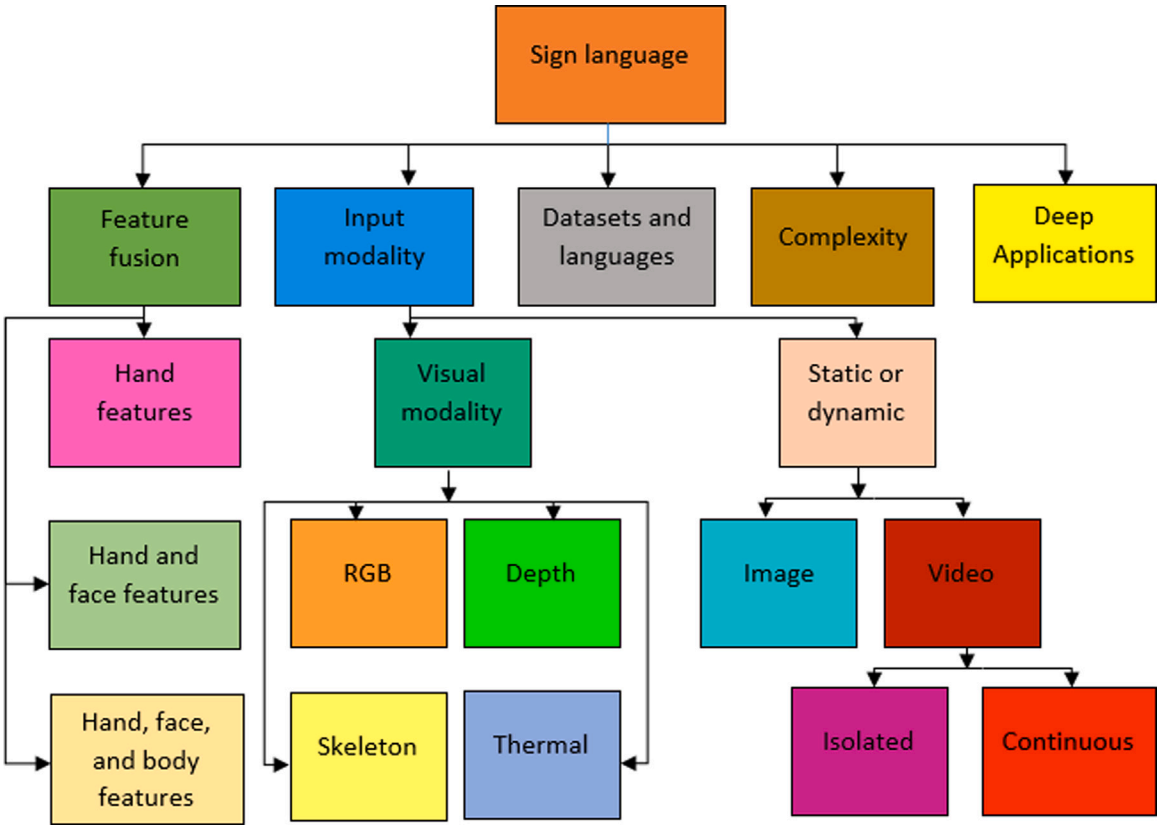

Fig. 1. Taxonomy of deep models for sign language recognition.

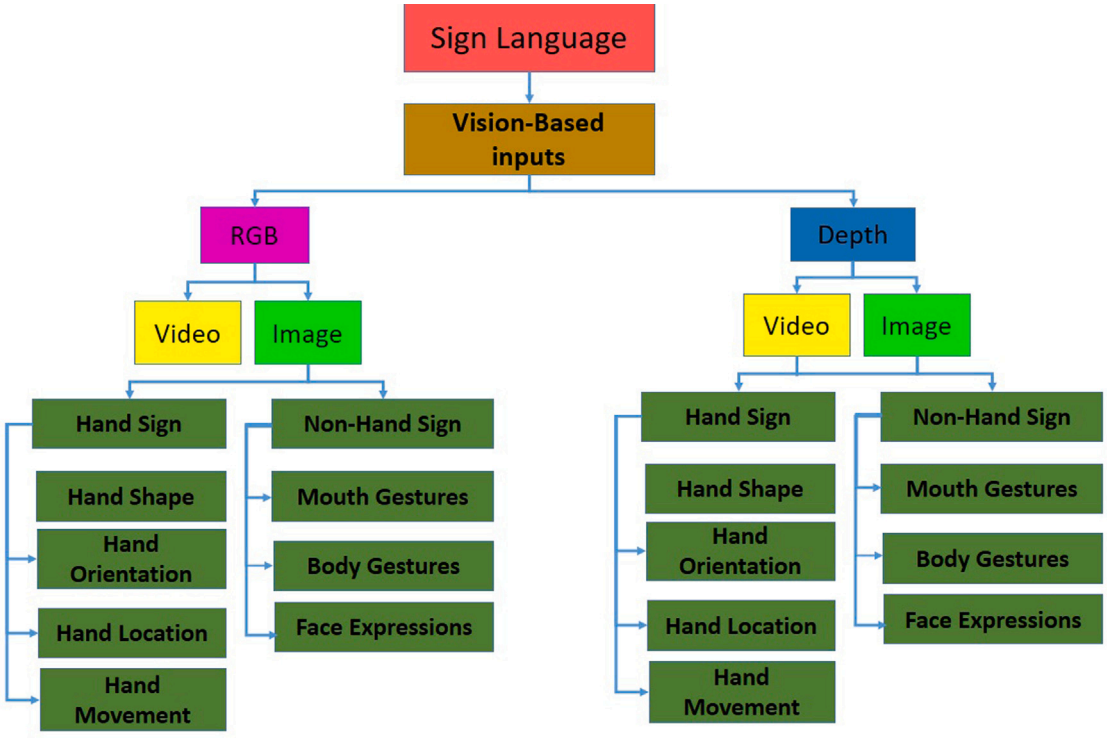

Fig. 2. Vision-based sign language models classification.

of the early efforts on hand and gesture recognition is dated back to 1987, where a hand gesture interface is proposed by Zimmerman et al. (1987) to estimate the hand position and orientation using the magnetic flux sensors of a glove. However, there are different challenges to address in visual sign language recognition (e.g. inter and intra subject variability, illumination conditions, partial occlusions, different points

of view and resolutions, background artifacts) that makes it difficult to define a universal automatic model for automatic sign language recognition. In addition to the present list of challenges, there are some other critical challenges in sign language recognition. One of the main challenges is developing a sign language recognition system to translate the sing words or sentences into the text or voice to facilitate the communication between deaf people and the hearing majority in a real-world situation. The current works in sign language recognition use datasets including videos of only one sign or images of only one character. We need to develop systems that can be applied in a real chat or conversation between a deaf person and a hearing one. To do this, the proposed systems need to be efficient and intelligent enough to split the input videos, including some characters, words, or sentences, into separate characters, words, or sentences. Another main challenge in this area is the high difference between sign languages of different countries. We need to have a multi-lingual system capable of translating different sign languages into text or voice. Due to the importance of sign language recognition in the deaf and speaking disabled community, we present a comprehensive review of the recent sign language recognition works in computer vision using deep learning, identifying future lines of research. In this work, we perform a comprehensive review of recent works for sign language recognition, defining taxonomy to group existing works and providing with an associated discussion on their pros and cons.

The remainder of this paper is organized as follows. Section 2 includes a brief review of Deep Learning algorithms. Section 3 presents a taxonomy of the sign language recognition area. Hand sign language, face sign language, and human sign language literature are reviewed in Sections 4, 5, and 6, respectively. Section 7 presents the recent models in continuous sign language recognition. Recent hybrid models are included in Section 8. Finally, We discuss the main challenges and conclude the work on Section 9.

2. Why deep learning?

In this section, we present a brief introduction to Deep Learning. Over recent years, deep learning methods outperformed previous state-of-the-art machine learning techniques in different areas, especially in Computer Vision and Natural Language Processing (Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018). Some of the most significant deep learning models used in computer vision problems are Convolutional Neural Network (CNN) (Wu, 2019), Deep Boltzmann Machine (RBM) (Fischer & Igel, 2012), Deep Belief Network (DBN) (Hinton, 2007), Auto Encoder (AE) (Grosse, 2017), Variational Auto Encoder (VAE) (Doersch, 2016), Generative Adversarial Network (GAN) (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, & Bengio, 2014), and Recursive Neural Network (RNN) including Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) (Wang, 2016). One of the main goals of deep models is to avoid the need to build/extract features. Deep learning allows computational models of multiple processing layers to learn and represent data with multiple levels of abstraction to imitate the human brain mechanism and implicitly capture the complex structures of large-scale data. The first effort on human brain simulation is dated back to 1943, where McCulloch and Pitts (1943) tried to understand how the brain could produce highly complex patterns by using interconnected basic cells, called neurons. The trend of major contributions continued so that DBN was one of the prominent breakthroughs in Deep Learning introduced by Hinton, Osindero, and Teh (2006). Deep learning contains a wealthy group of methods, including neural networks, hierarchical probabilistic models, and a variety of unsupervised and supervised feature learning algorithms. One of the important factors that contributed to the huge boost of deep learning is the advent of large-scale, high-quality, and publicly available labeled datasets along with the capability of parallel GPU computing. Other factors, such as the alleviation of the vanishing gradient, proposing some new regularization techniques (such as dropout, batch normalization, and data augmentation), and development of some strong frameworks like TensorFlow (TensorFlow, 2020), Theano (Frederic, Lamblin, Pascanu, et al., 2012), and MXNET (MXNET, 2020), had a significant role in Deep Learning advancement. In this survey, we focus only on Deep Learning-based models for sign language recognition in computer vision.

3. Taxonomy

In this section, we present a taxonomy that summarizes the main concepts related to deep learning in sign language recognition. In the rest of this section, we explain different feature fusions for sign language recognition, input modality, datasets, and applications of this area. Figs. 1 and 2 show the proposed taxonomy that we describe in this section.

3.1. Feature fusion

To improve the recognition accuracy of sign language, different features can be fused that we organize them in three categories, which are: using only the hand pose features, using the hand and face pose features, using the hand, face, and body pose features. Details of these categories are explained in the following sub-sections.

3.1.1. Hand pose features

Using the hand pose features has been more considered in recent years with the advent of the accurate depth sensors (Chen, Wanga, Guoa, & Zhanga, 2020; Dibra, Wolf, Oztireli, & Gross, 2017; Doosti, 2019; Wang, Chen, Liu, Qian, Lin, & Ma, 2018). In this category, only hand features are used to sign language recognition. After hand detection from input data, the hand features are extracted using different deep learning architectures such as CNN, RBM, RNN, GAN, and so on (Cao et al., 2017; Cheok et al., 2017; Deng et al., 2017; Dibra et al., 2017; Escobedo-Cardenas & Camara-Chavez, 2020; Gomez-Donoso, Orts-Escolano, & Cazorla, 2019; Guo et al., 2017; Li, Xue, Wang, Ge, Ren, & Rodriguez, 2019; Oberweger et al., 2015; Rastgoo et al., 2018; Rastgoo, Kiani, & Escalera, 2020a, 2020b; Supancic et al., 2018; Tagliasacchi et al., 2015; Wang et al., 2018; Zheng et al., 2017). So, having an efficient hand detection and feature extraction model is challenging in this category. While CNN has an impressive capability to cope with the still images, it does not efficiently cover the sequence information. So, CNN is combined with another deep learning model, such as RNN, LSTM, and GRU, to benefit from the capability of these models in sequence feature extraction from visual data.

3.1.2. Hand and face pose features

Since the face pose includes some important grammatical and prosodic features, using these features as complementary features with the hand features could improve the sign language recognition accuracy. This category focuses on the proposed models using the combined features of hand and face pose. Few models have been proposed in this category due to some challenges in tracking human faces from videos such as the head tilting and side-to-side movements. Furthermore, the proposed models have to be able to track and recognize the facial features in natural settings, not in a constrained environment. So it is essential to have a model with accurate recognition that could be applied in the real world, not in some complicated laboratory conditions. Challenges related to extreme head movements from side to side, frequent self-occlusions of the face by the signer's hands and hair also have to be considered in this area. Because of these complexities, some of the models work only on a part of the face, such as eye or leap, to decrease these complexities (Koller, Ney, & Bowden, 2015).

Table 1
Sign language datasets including video samples. F: face, H: hand, h: head, W: word, S: sentence, C: Country, CN: Class Number, SubN: Subject Number, SampN: Sample Number, LL: Language Level (word or sentence), A: Annotation.

Y	Dataset	C	CN	SubN	SampN	LL	A
2011	Boston ASL LVD (Thangali, Nash, Sclaroff, & Neidle, 2011)	USA	3300	6	9800	W	H
2012	DGS Kinect 40 (Cooper, Ong, Pugeault, & Bowden, 2012)	Germany	40	15	3000	W	–
2012	RWTH-PHOENIX-Weather (Forster, Schmidt, Hoyoux, Koller, Zelle, Piater, & Ney, 2012)	Germany	1200	9	45760	S	F, H
2012	GSL 20 (Adaloglou, Chatzis, Papastratis, Stergioulas, Papadopoulos, Zacharopoulou, Xydopoulos, Atzakas, Papazachariou, & Daras, 2019)	Greek	20	6	840	W	–
2013	PSL Kinect 30 (Oszust & Wysocki, 2013)	Poland	30	1	300	W	–
2013	PSL ToF 84 (Oszust & Wysocki, 2013)	Poland	84	1	1680	W	–
2014	DEVISIGN-G (Chai, Guang, Lin, Xu, Tang, Chen, & Zhou, 2013)	China	36	8	432	W	–
2014	DEVISIGN-D (Chai et al., 2013)	China	500	8	6000	W	–
2014	DEVISIGN-L (Chai et al., 2013)	China	2000	8	24000	W	–
2015	SIGNUM (Koller, Forster, & Hermann, 2015)	Germany	450	25	33210	S	–
2016	MSR (Chen, Zhang, Hou, Jiang, Liu, & Yang, 2017)	USA	12	10	336	W	–
2016	LSA64 (Ronchetti, Quiroga, Estrebou, L, & Rosete, 2016)	Argentina	64	10	3200	W	H, h
2016	TVC-hand gesture (Kim, Ban, & Lee, 2017)	Korea	10	1	650	–	–
2018	PHOENIX14T (Camgoz, Hadfield, Koller, Ney, & Bowden, 2018)	Germany	1066	9	67781	S	–
2020	RKS-PERSIANSIGN (Rastgoo et al., 2020a)	Iran	100	10	10000	W	H

3.1.3. Hand, face, and body pose features

Some models fuse the features of hand, face, and the other parts of the human body to benefit from the fused features and improve the recognition accuracy. Using these fused features, sign language recognition models could be improved to be more robust to occlusions, severe deformations, and appearance variations (Kocabas, Karagoz, & Akbas, 2018; Newell, Yang, & Deng, 2016; Wei, Ramakrishna, Kanade, & Sheikh, 2016). In this category, the proposed models benefit from the body features, as these features could improve the recognition accuracy in the complex situations of the hand or face occlusions.

3.2. Input modality

Generally, vision-based and glove-based approaches are the two main categories considered for sign language recognition models. While the vision-based models use the captured video data of the signers for different signs, the glove-based models employ some mechanical or optical sensors attached to a glove in order to use the electrical signals for hand pose detection. Vision-based models provide more natural and real systems based on the information that human can sense from the surroundings (Zheng et al., 2017). Focusing on vision-based models, we explain the details of the input modalities in this subsection.

3.2.1. Visual modality

We present the details of visual modality from two perspectives, as follows:

- **Input data modality**: There are some visual input data modalities in sign language recognition area in recent years. RGB and depth input data are two common types of input data used in sign language recognition models. While RGB images or videos include the high-resolution contents, depth inputs have the accurate information related to the distance between the image plane and the corresponding object in the image. Some of the models use the advantages of two input modalities, simultaneously. We also survey the models in this category in the next section. Thermal modality is another modality that it is not as common as RGB and depth modalities. Infrared (IR) thermal sensors have the capability of imaging scenes and objects based on the IR light reflectance or radiation emittance. Although many works have used the benefits of thermal information for face recognition and human body detection, few hand sign recognition models have focused on thermal information learning (Kim et al., 2017). Another type of the input modality is the skeleton, as an encoded form of the joint sequences, that has been provided in some gesture and hand datasets. Flow information, as the motion features of every pixel in a sequence of the video frames, is another input modality that have been widely used by some researchers in order to combine with the CNN features. Two types of flow information, Optical Flow (OF) and Scene Flow (SF), have been used in some models. While OF, as a displacement vector of pixel positions, is usually used for RGB image sequences, SF, as a dense or semi-dense 3D motion field of a scene with respect to the camera, is applied to depth frame sequences.
- **Devices**: There are different devices to record the input data modalities and camera is the most common one of them. Different cameras support different formats and qualities of input data. Microsoft Kinect, as a strong device, is widely used due to its ability to provide high-quality depth and RGB video streams, simultaneously. Another device is the flex sensors planted inside the gloves to acquire data of the palms and fingers movement. Leap Motion Controller (LMC) system is another device to detect and track hands, fingers, and finger-like objects. The point is that we need to delineate what kind of applications are targeted because selection of the suitable device heavily depends on the application. For example, while Kinect camera can work well for many real applications, it highly depends on some environmental conditions. The same go with the other devices. Using these depth sensors and cameras such as Kinect and RealSense, the proposed models can employ 3D information to decrease the ambiguity of 2D information recorded from traditional devices (Marı n Jimeneza, Romero-Ramireza, Munoz-Salinasa, & Medina-Carnicer, 2018; Lifshitz, Fetaya, & Ullman, 2016).

3.2.2. Static or dynamic

Two forms of input data, static or dynamic, are used in sign language recognition models to extract the necessary features. Many deep-based models have been proposed to use the still or sequence inputs in recent years. While dynamic inputs include the sequential information that could be useful to improve the sign language recognition accuracy, there are still some challenges for using this data such as computation complexity of input sequences. Furthermore, dynamic inputs can be split into isolated dynamic inputs and continuous dynamic inputs. While the isolated dynamic inputs are used at word level, the continuous dynamic inputs are employed at sentence level. Additional challenges in continuous dynamic inputs include tokenization of the sentences into separate words, detection of start and end of a sentence, and managing the abbreviations and synonyms in the sentence. In the next sections, we survey the sign language recognition models that have used these inputs.

3.3. Datasets and different sign languages

We list the most relevant datasets, including videos, for sign language recognition area in Table 1, respectively. For each dataset in this

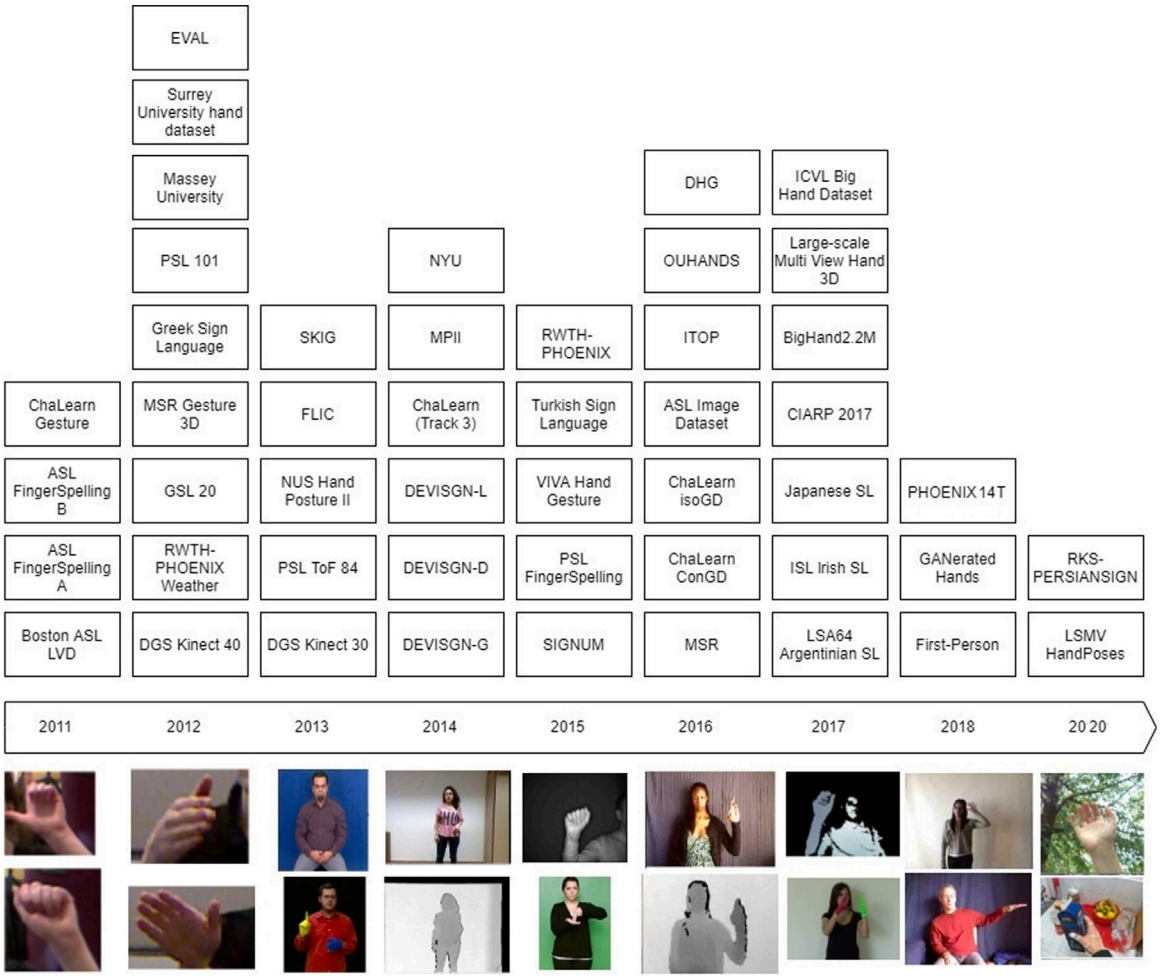

Fig. 3. Sign language Datasets in one glance.

Table 2
Sign language datasets including images.

Year	Dataset	C	CN	Sub	Samp.
2011	ASL Fingerspelling A (Pugeault & Bowden, 2011)	USA	24	5	131000
2011	ASL Fingerspelling B (Pugeault & Bowden, 2011)	USA	24	9	–
2016	LSA16 handshapes (Ronchetti, Quiroga, Estrebou, & Lanzarini, 2016)	Argentina	16	10	800
2015	PSL Fingerspelling ToF (Kapuscinski, Oszust, Wysocki, & Warchol, 2015)	Poland	16	3	960

table, we specify eight fields including Year (Y), Dataset name, Country (C), Class Number (CN), Subject Number (SubN), Sample Number (SampN), Language Level (LL), and Annotation (A). These datasets have different environments, qualities, constraints, and complexities.

While the sign language recognition models use different languages in the input data, American Sign Language (ASL) has attracted more attention due to more popularity and usage. The other languages, such as India, Germany, Netherlands, Greece, Poland, Argentina, Turkey, China, have also been used in the proposed datasets (Adaloglou et al., 2019; Andriluka, Pishchulin, Gehler, & Bernt, 2014; Baró, Gonzàlez, Fabian, Bautista, Oliu, Escalante, Guyon, & Escalera, 2015; Chai et al., 2013; Chen et al., 2017, 2017; Cooper et al., 2012; Escalera, Gonzàlez, Baró, Reyes, Lopés, Guyon, Athitsos, & Escalante, 2013; Forster et al., 2012; Ganapathi, Plagemann, Koller, & Thrun, 2012; Haque, Peng, Luo, Alahi, Yeung, & Fei-Fei, 2016; Kapuscinski et al., 2015; Koller, Forster et al., 2015; Liu & Shao, 2013; Matilainen, Sangi, Holappa, & Silven, 2016; Molchanov, Gupta, Kim, & Kautz, 2015; Oszust & Wysocki, 2013; Pugeault & Bowden, 2011; Ronchetti, Quiroga, Estrebou, & Lanzarini, 2016; Ronchetti, Quiroga, Estrebou, Lanzarini et al., 2016; Sapp & Taskar, 2013; Smedt, Wannous, & Vandeborre, 2016; Thangali et al., 2011; Tompson, Stein, Lecun, & Perlin, 2014; Wan, Zhao, Zhou, Guyon, Escalera, & Li, 2016; Yuan, Ye, Stenger, Jain, & Kim, 2017). The trend of sign and gesture datasets can be found in Fig. 3. Some of the most popular sign datasets, including image modality, have been listed in Tables 2 and 3. Also, some of the gesture datasets and also some samples of sign and gesture datasets have been presented in Table 3, Fig. 4, and Fig. 5, respectively.

As one can see in Tables 1–3, it should be better to increase the numbers of sign categories to have a more realistic method generalization for real applications. Furthermore, We have to note that most of these datasets are for sign classification, not detection/spotting. Only a few datasets, such as Montalbano II dataset (Escalera et al., 2013), include a detection task.

3.4. Task complexity

Sign language, as a visual language for the deaf community, defines some grammatical rules to concatenate the movements of the hand, face, and body parts. While there are different parameters in hand sign such as movement, shape, orientation, and place of articulation,

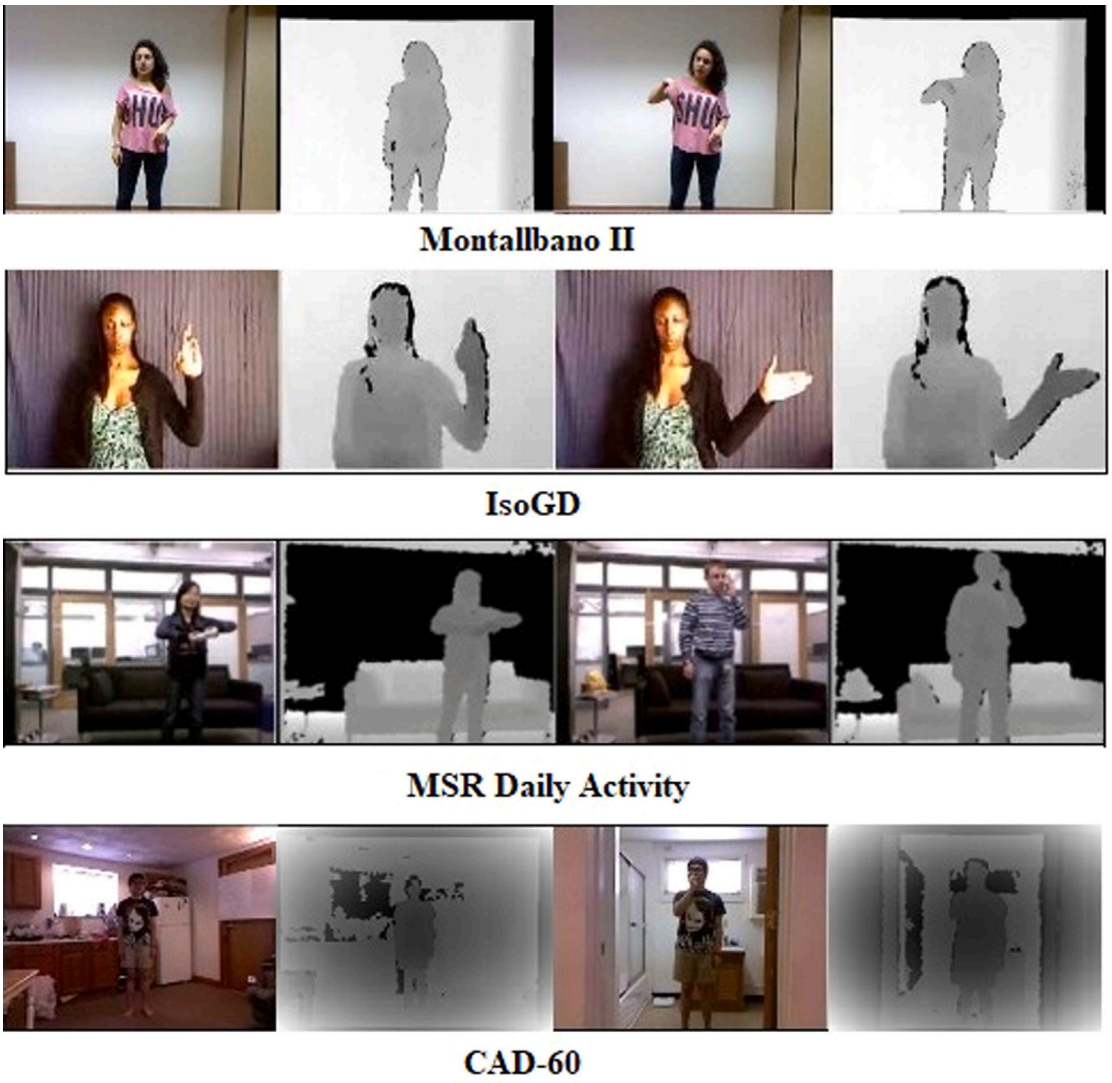

Fig. 4. Some samples of gesture recognition datasets.

Fig. 5. Some samples of sign language datasets.

face sign uses the eye gaze, eyebrows, mouth, and head orientation parameters. In addition to these parameters, sign language benefits from multi-channel information of the hand shape, motions, body pose, and even facial gestures. Using these parameters, the level complexity of signs depends on the sign modality. There are three levels of signs, which are a static sign, dynamic sign, and continuous dynamic sign. While the input modality of a static sign is an image, a video modality is employed in dynamic and continuous dynamic signs. However, a video

Table 3
Some of the gesture recognition datasets.

Year	Dataset	Modality	Class num.
2011	ChaLearn Gesture (Escalera et al., 2013)	RGB, Depth	15
2012	MSR-Gesture3D (Chen et al., 2017)	RGB, Depth	12
2014	ChaLearn (Track 3) (Baró et al., 2015)	RGB, Depth	20
2015	VIVA Hand Gesture (Molchanov et al., 2015)	RGB	19
2016	ChaLearn conGD (Wan et al., 2016)	RGB, Depth	249
2016	ChaLearn isoGD (Wan et al., 2016)	RGB, Depth	249

sample in dynamic sign language recognition includes only one sign, while a video sample containing multiple signs is used in continuous dynamic sign language recognition. In this way, there is a rising trend in parameter complexity from static sign to dynamic and continuous dynamic signs.

3.5. Deep applications

Hand detection, hand tracking, hand pose estimation, hand gesture recognition, and hand pose recovery are some of the most important sub-areas of hand sign language recognition that are widely used in human–computer interaction applications (Newell et al., 2016). Nowadays that computers have become an important part of our lives, they can be used to facilitate the communications of deaf and hearing impaired people. For example, the interpreting services, such as remote video human interpreting using high-speed internet connections, can be used but there are some major limitations for these services that need to be resolved before usage. These applications are important not only from an engineering point of view but also for the impact on society. Further applications, such as automatic indexing of signed videos, human behavior understanding, online human communication using hand tracking, multi-person pose estimation, and the other human–computer interaction applications, related to sign language area, can also be considered to use (Cao et al., 2017; Deng et al., 2017; Marı n Jimeneza et al., 2018; Supancic et al., 2018; Tagliasacchi et al., 2015).

4. Hand sign language recognition

In this category, only hand features are used to sign language recognition. There are some important sub-areas for hand sign language recognition which are: hand detection, hand pose estimation, real-time hand tracking, hand gesture recognition, and hand pose recovery. In this section, we present the proposed models of last four years in these sub-areas with an associated discussion on their pros and cons. Tables 4–7, show the details of these models. In these tables, we categorize the presented models based on the same datasets, same evaluation metrics, same features, and same input modalities used in the models. In these tables, we use some abbreviations for goal field which are: HP: Hand Pose, HT: Hand Tracking, HD: Hand Detection, HSR: Hand Sign Recognition, HG: Hand Gesture, RHR: Real-time Hand Recognition.

4.1. Hand detection

Hand detection currently has a momentous role in sign language recognition area. However, many researches have been conducted to improve hand detection models, this task still includes many challenges in computation time and detection accuracy aspects (Le, Jaw, Lin, Liu, & Huang, 2018). Some of the prominent object detection methods have been fine-tuned using transfer learning approach in order to use their strong capabilities in hand detection area. One of the commendable object detection methods is Region-based Convolutional Neural Network (R-CNN) model provided by Girshick, Donahue, Darrell, and Malik (2015). R-CNN uses the region proposals to detect the objects in the input images. It is slow because it performs a CNN forward pass for each object proposal in the model input, without sharing computation. To improve the R-CNN performance, Fast-RCNN was suggested to efficiently classify object proposals using deep convolutional networks (Girshick, 2015). In Fast-RCNN, the image is fed as input to the CNN in order to provide the convolutional feature map. The region proposals are detected in feature maps and forwarded to the next step. To avoid using the selective search method used in the RCNN and Fast-RCNN to detect the region proposals, Shaoqing Ren et al. proposed a CNN to learn the region proposals in the model (Ren, He, Girshick, & Sun, 2015). The Faster-RCNN model has been fine-tuned and used by some of the proposed models for hand detection applications (Bambach, Lee, Crandall, & Yu, 2015; Yang, Li, Fermuller, & Aloimonos, 2015).

The region-based methods, RCNN, Fast-RCNN, and Faster-RCNN, do not consider the whole image and work on the regions of the input image to localize the objects. To solve this problem, another model, namely You Only look Once (YOLO), has been proposed to object detection using just one CNN to predict the bounding boxes and the class probabilities for these boxes. YOLO splits the input image into a grid that predefined numbers of bounding boxes are considered in each grid. The class probability of each bounding box is estimated using CNN, and the detected object will be located at the bounding boxes with maximum class probabilities. The limitation of YOLO is that it struggles with small objects in the input image (Redmon, Divvala, Girshick, & Farhadi, 2016). To improve the detection accuracy of YOLO, another object detection model, namely Single Shot Multi-box Detector (SSD), has been proposed using adding the feature maps from different layers on top of the YOLO model. This makes the SSD more accurate and faster (Liu, Anguelov, Erhan, Szegedy, Reed, Fu, & Berg, 2016). In this section, we review the proposed models for hand detection using deep learning approaches.

4.1.1. RGB-based hand detection methods

Simon et al. proposed a real-time convolutional-based method to hand detection using a multi-view camera system from RGB still images. They used a keypoint detector to provide noisy labels. After that, a multi-view geometry approach has been used to convert and re-project the 2D detected keypoints into the 3D view. They generated some labeled data in this way and used them to train the proposed model for hand keypoints detection. The model has been trained only on their own dataset and reported the results. In the best case, they achieved average error of 3.65 for Tip Touch that is equal to having an improvement with a 1.66 margin in comparison with state-of-the-art. However, their model needs to be robust enough to work with fewer cameras and in less controlled environments (e.g., with multiple cell phones) (Simon et al., 2017).

Yan et al. provided a multi-scale CNN for unconstrained hand detection in still images. They used a generic region proposal algorithm, followed by multi-scale information fusion from the VGG16 model to hand detection scheme. They integrated the features from multiple layers of a CNN model to have a multi-scale representation of hand objects. The evaluation results on Oxford Hand Detection Dataset and VIVA Hand Detection Challenge showed that the model achieved a detection accuracy of 49.6% and 92.8% on these datasets with a relative state-of-the-art improvement of 1.6% and 2.1% (Yan, Xia, Smith, Lu, & Zhang, 2017).

4.1.2. Depth-based hand detection

Neverova et al. provided a CNN-based model for hand detection and segmentation using both of the unlabeled and synthetic data. They used the structural information not into the model architecture but into the training objective and benefited from the advantages of very fast test-time processing and the ability to parallelize. While the evaluation results on the synthetic data showed the improvement of the segmentation and detection accuracy obtaining a detection accuracy of 82.0%, they need to adopt the model to real data (Neverova et al., 2014).

Table 4
Deep sign language recognition models, categorized based on the datasets used for evaluation.

Dataset	Year	Ref.	Goal	Model	Modality	Results
Own dataset	2014	(Neverova, Wolf, Taylor, & Nebout, 2014)	HP	CNN	2D, Depth	82.0 (Acc.)
	2014	(Tompson et al., 2014)	HP	CNN	2D, Depth	33.0 (error)
	2015	(Kang, Tripathi, & Nguyen, 2015)	HD	CNN	Depth	99.0
	2015	(Tang, Lu, Wang, Huang, & Li, 2015)	HP	CNN, DNN	2D, Depth, RGB	0.899 ms (Ave. time)
	2016	(Han, Chen, Li, & Chang, 2016)	HG	CNN	2D, RGB	93.80
	2017	(Simon, Joo, Matthews, & Sheikh, 2017)	HD	CNN	RGB	4.15 mm
	2018	(Rao, Syamala, Kishore1, & Sastry, 2018)	HSR	CNN	2D, RGB	92.88
	2018	(Ye, Tian, Huenerfauth, & Liu, 2018)	HSR	CNN	3D, RGB	69.2
	2019	(Gomez-Donoso et al., 2019)	HP	CNN	static, RGB, Depth	5 mm
	2020	(Wadhawan & Kumar, 2020)	HSR	CNN	static, RGB	99.72
NYU	2015	(Oberweger et al., 2015)	HP	CNN	3D, Depth	20 mm
	2017	(Deng et al., 2017)	HP	CNN	3D, Depth	17 mm
	2017	(Guo et al., 2017)	HP	CNN	2D, Depth	66.00
	2017	(Fang & Lei, 2017)	HP	CNN, AE	2D, Depth	17 mm
	2017	(Yuan et al., 2017)	HP	CNN	2D, Depth	21.4 mm
	2017	(Madadi, Escalera1, Baro, & Gonzalez, 2017)	HP	CNN	2D, Depth	15.6 mm
	2018	(Chen et al., 2020)	HP	CNN	2D, Depth	11.811 mm
	2018	(Rastgoo et al., 2018)	HSR	RBM	2D, RGB, Depth	90.01
	2016	(Sinha, Choi, & Ramani, 2016)	HP	CNN	static, Depth	9 mm
	2017	(Ge, Liang, Yuan, & Thalmann, 2017)	HP	CNN	static, Depth	9 mm
	2017	(Dibra et al., 2017)	HP	CNN	static, Depth	9 mm
	2018	(Ge, Liang, Yuan, & Thalmann, 2018)	HP	CNN	static, Depth	9.46 mm
	2018	(Moon, Chang, & Lee, 2018)	HP	CNN	static, Depth	8.42 mm
	2018	(Baek, Kim, & Kim, 2018)	HP	GAN	static, Depth	14.1 mm
	2018	(Kazakos, Nikou, & Kakadiaris, 2018)	HP	CNN	static, RGB, Depth	9 mm
	2018	(Spurr, Song, Park, & Hilliges, 2018)	HP	VAE	static, RGB, Depth	10.5 mm
	2020	(Rastgoo et al., 2020a)	HSR	SSD, 2DCNN, 3DCNN, LSTM	dynamic, RGB	4.64 mm
ICVL	2015	(Oberweger et al., 2015)	HP	CNN	3D, Depth	10 mm
	2017	(Deng et al., 2017)	HP	CNN	3D, Depth	11 mm
	2017	(Guo et al., 2017)	HP	CNN	2D, Depth	7.8 mm
	2017	(Fang & Lei, 2017)	HP	CNN, AE	2D, Depth	9 mm
	2017	(Yuan et al., 2017)	HP	CNN	2D, Depth	12.3 (error)
	2017	(Dibra et al., 2017)	HP	CNN	static, Depth	8 mm
	2018	(Chen et al., 2020)	HP	CNN	2D, Depth	6.793 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	static, Depth	8 mm
	2018	(Moon et al., 2018)	HP	CNN	static, Depth	6.28 mm
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	8.5 mm
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	19.5 mm
MSRA	2016	(Oberweger, Riegler, Wohlhart, & Lepetit, 2016)	HP	CNN	2D, 3D, Depth	5.58 mm (error)
	2017	(Guo et al., 2017)	HP	CNN	2D, Depth	9.80 mm
	2017	(Madadi et al., 2017)	HP	CNN	2D, Depth	18 mm
	2017	(Yuan et al., 2017)	HP	CNN	2D, Depth	21.3 (error)
	2017	(Ge et al., 2017)	HP	CNN	static, Depth	6 mm
	2018	(Chen et al., 2020)	HP	CNN	2D, Depth	8.649 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	static, Depth	8 mm
	2018	(Moon et al., 2018)	HP	CNN	static, Depth	7.49 mm
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	12.5 mm
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	10 mm
FLIC	2014	(Toshev & Szegedy, 2014)	HP	DNN	2D, RGB	96.0
	2016	(Newell et al., 2016)	HP	CNN	2D, RGB	99.0 (Elbow)
	2016	(Wei et al., 2016)	HP	CNN	2D, RGB	97.59
LSP	2014	(Toshev & Szegedy, 2014)	HP	DNN	2D, RGB	78.0
	2016	(Wei et al., 2016)	HP	CNN	2D, RGB	84.32
isoGD	2016	(Duan, Zhou, Wany, Guo, & Li, 2016)	HG	CNN	2D, Depth, RGB	67.19
	2017	(Wang, Li, Liu, Gao, Tang, & Ogunbona, 2017)	HG	CNN	2D, Depth	55.57
	2020	(Rastgoo et al., 2020b)	HSR	SSD, CNN, LSTM	2D, RGB	86.32
MPII	2016	(Newell et al., 2016)	HP	CNN	2D, RGB	90.90 (total)
	2016	(Wei et al., 2016)	HP	CNN	2D, RGB	87.95
ITOP	2016	(Haque et al., 2016)	HP	CNN	2D, Depth	80.50
	2017	(Guo et al., 2017)	HP	CNN	2D, Depth	84.90
	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	3D, Depth	97.5 (AUC)
RGBD-HuDaAct	2016	(Duan et al., 2016)	HG	CNN	2D, Depth, RGB	96.74
STB	2017	(Zimmermann & Brox, 2017)	HP	CNN	3D, RGB	94.0 (AUC)
	2018	(Mueller, Bernard, Sotnychenko, Mehta, Sridhar, Casas, & Theobalt, 2018)	HT	CNN	static, RGB	96.5 (AUC)
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	98.3(AUC)
	2019	(Li et al., 2019)	HP	CNN	static, RGB	8.34 (err)
	2019	(Gomez-Donoso et al., 2019)	HP	CNN	static, RGB, Depth	5 mm
EVAL	2016	(Haque et al., 2016)	HP	CNN	2D, Depth	74.10
Dexter	2017	(Zimmermann & Brox, 2017)	HP	CNN	3D, RGB	49.0 (AUC)

(*continued on next page*)

Table 4 (*continued*).

Dataset	Year	Ref.	Goal	Model	Modality	Results
Real video samples	2015	(Tagliasacchi et al., 2015)	HT	CNN	2D, Depth	8 mm (MSE)
	2019	(Ferreira et al., 2019)	HS	CNN	static, RGB, Depth	3.17, 92.61
RWTH-PHOENIX-Weather	2015	(Koller, Ney et al., 2015)	HSR	CNN	2D, RGB	55.70 (Precision)
BigHand2.2M	2017	(Yuan et al., 2017)	HP	CNN	2D, Depth	17.1 (error)
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	13.7 mm
Human3.6M	2018	(Wang et al., 2018)	HP	CNN	2D, Depth	62.8 mm
UBC3V	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	3D, Depth	88.2 (AUC)
Massey 2012	2018	(Rastgoo et al., 2018)	HR	RBM	2D, RGB, Depth	99.31
SL Surrey	2018	(Rastgoo et al., 2018)	HR	RBM	2D, RGB, Depth	97.56
ASL Fingerspelling A	2018	(Rastgoo et al., 2018)	HR	RBM	2D, RGB, Depth	98.13
OUHANDS,	2018	(Dadashzadeh, Tavakoli Targhi, & Tahmasbi, 2018)	HG	CNN	2D, Depth	86.46
Egohands	2017	(Dibia, 2017)	HT	CNN	static, RGB	96.86 (mAP)
Dexter	2018	(Mueller et al., 2018)	HT	CNN	static, RGB	64.0 (AUC)
EgoDexter	2018	(Mueller et al., 2018)	HT	CNN	static, RGB	54.0 (AUC)
RHD	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	84.9(AUC)
B2RGB-SH	2019	(Li et al., 2019)	HP	CNN	static, RGB	7.18 (err)
DHG-14/28 Dataset	2019	(Chen, Zhao, Peng, Yuan, & Metaxas, 2019)	HG	CNN	dynamic, RGB	91.9
SHREC'17 Track Dataset	2019	(Chen et al., 2019)	HG	CNN	dynamic, RGB	94.4
RWTH-BOSTON-50	2019	(Lim et al., 2019)	HS	CNN	dynamic, RGB	89.33
ASLLVD	2019	(Lim et al., 2019)	HS	CNN	dynamic, RGB	31.50
EgoGesture	2019	(Kopuklu, Gunduz, Kose, & Rigoll, 2019)	HG	CNN	dynamic, RGB	94.03
NVIDIA benchmarks	2019	(Kopuklu et al., 2019)	HG	CNN	dynamic, RGB	83.83
isoGD	2020	(Elboushaki, Hannane, Afdel, & Koutti, 2020)	HG	CNN	dynamic, RGB, Depth	72.53
SKIG	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	99.72
NATOPS	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	95.87
SBU	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	97.51
RKS-PERSIANSIGN	2020	(Rastgoo et al., 2020a)	HSR	SSD, 2DCNN, 3DCNN, LSTM	dynamic, RGB	99.80

4.1.3. Discussion

Most of the deep-based models in four recent years have used a CNN or a combination of a CNN with another deep learning approach for hand detection from different types of input data, such as image, video, skeleton, flow features, and so on. While CNN has a potent capability to feature extraction from still images, RNN models, such as LSTM and GRU, have more powerful capability to use the sequence information of the input videos than the CNN. However, some of the proposed models not only use the combination of these approaches, CNN and RNN, to benefit from the capabilities of both methods but also use the fused features of multi-modal inputs.

Although hand detection has been extensively studied in recent years and many models have been proposed, there are still a lot of challenges that need to be tackled. Due to heavy occlusions in hand keypoints, even accurate manual keypoint annotations are too difficult to do. So, providing an accurate model to learn the hand keypoints is challenging in this area. While hand detection is often the first step in many tasks, such as action recognition and sign language recognition, it is a very difficult task due to high variations of the hand shapes and gestures. Heavy occlusion, low resolution, varying illumination conditions, different hand gestures, and complex interactions between hands and objects or other hands are some of the most important challenges in this area. Another important issue is the situation that a hand may hold objects, appear at different scales with closed or open palms, have different articulations of the fingers, or hold other hands. We think that using the features of the other human body parts, such as face and body, fusing different types of input data, such as image, skeleton, flow information, text, and so on, and also using the other deep learning model fused with the traditional approaches could help to improve the detection accuracy. Also benefiting from the face and human pose datasets and combining data from different domains with the hand detection datasets could be considered for more improvement. There are a lot of impressive models for face and person detection that fusing the features of them by hand features may improve the detection accuracy (Gattupalli, Ghaderi, & Athitsos, 2016; Haque et al., 2016; Huang, Zhou, Li, & Li, 2015; Marı n Jimeneza et al., 2018; Koller, Ney et al., 2015; Newell et al., 2016; Rao et al., 2018; Wang et al., 2018; Wei et al., 2016). Furthermore, using the new hardware capabilities and developments with efficient model implementations can provide the real hand detection required in real applications.

4.2. Hand pose estimation

Hand pose estimation has matured rapidly due to the introduction of depth sensors in recent years (Doosti, 2019; Supancic et al., 2018). Here, we provide a complete analysis of the state-of-the-art methods for hand pose estimation. A good review of recent progress for hand pose estimation methods from the depth sensor can be found in Barsoum (2016), Doosti (2019), Li et al. (2019).

4.2.1. RGB-based hand pose models

Zimmermann and Brox proposed a deep network to learn the 3D articulation priors and keypoints in the input RGB images. Their model includes three main steps for localization, cropping, and estimation of the hand that CNN is used in all steps. Furthermore, they introduce a large scale 3D hand pose dataset based on the synthetic hand models to train the model. They evaluated the model on two datasets, Stereo Hand Pose Tracking Benchmark and Dexter, and reported the Area Under the ROC Curve (AUC) of 94.0 and 49.0 on these datasets with 10.9 and 5 relative state-of-the-art improvement. While the performance of the model is even competitive to approaches with input depth maps, it needs to be learned by an annotated large-scale dataset with real-world images and diverse pose statistics (Zimmermann & Brox, 2017). Li et al. designed an end-to-end approach to estimate 3D hand pose from stereo cameras. They developed a framework using 2D keypoint regressor to estimate the sparse disparity of the hand joints. Furthermore, they proposed a large scale synthetic dataset with stereo RGB images and full 3D hand pose annotations to efficiently learn the hand model. A reference tracking algorithm along with a evaluation protocol has been suggested for model evaluation. Benefiting from stereo cameras, evaluation results show the superiority of the model performance in comparison with the state-of-the-art models for hand pose estimation obtaining an estimation error of 8.34 on STB dataset with 1.16 relative improvement (Li et al., 2019). Gomez-Donoso et al. presented a pipeline, including two CNNs for accurate real-time tridimensional hand pose estimation using a single RGB frame. While the first CNN is employed for hand detection, the second one accurately estimates the tridimensional positions of the hand pose joints. Furthermore, they contributed on a large-scale dataset including the images of hands and the corresponding 3D joint annotations. Evaluation results on the

Table 5
Deep sign language recognition models, categorized based on the evaluation metric.

Evaluationmetric	Year	Ref.	Goal	Model	Modality	Dataset	Results
Accuracy	2014	(Neverova et al., 2014)	HP	CNN	2D, Depth	proposed dataset	82.0
	2014	(Toshev & Szegedy, 2014)	HP	DNN	2D, RGB	FLIC, LSP	96.0 , 78.0
	2015	(Kang et al., 2015)	HD	CNN	Depth	proposed dataset	99.0
	2016	(Han et al., 2016)	HG	CNN	2D, RGB	proposed dataset	93.80
	2016	(Duan et al., 2016)	HG	CNN	2D, Depth, RGB	Chalearn IsoGD, RGBD-HuDaAct	67.19, 96.74
	2016	(Newell et al., 2016)	HP	CNN	2D, RGB	FLIC, MPII	99.0 (Elbow), 90.90 (total)
	2016	(Wei et al., 2016)	HP	CNN	2D, RGB	MPII, LSP, FLIC	87.95, 84.32, 97.59
	2016	(Haque et al., 2016)	HP	CNN	2D, Depth	EVAL, ITOP	74.10, 80.50
	2017	(Wang et al., 2017)	HG	CNN	2D, Depth	ChaLearn	55.57
	2018	(Rao et al., 2018)	HSR	CNN	2D, RGB	own dataset	92.88
	2018	(Ye et al., 2018)	HSR	CNN	3D, RGB	own dataset	69.2
	2018	(Rastgoo et al., 2018)	HR	RBM	2D, RGB, Depth	Massey 2012, SL Surrey, NYU, ASL Fingerspelling A	99.31, 97.56, 90.01, 98.13
	2018	(Dadashzadeh et al., 2018)	HG	CNN	2D, Depth	OUHANDS,	86.46
	2020	(Wadhawan & Kumar, 2020)	HSR	CNN	static, RGB	own dataset	99.72
	2019	(Chen et al., 2019)	HG	CNN	dynamic, RGB	DHG-14/28 Dataset	91.9
	2019	(Chen et al., 2019)	HG	CNN	dynamic, RGB	SHREC'17 Track Dataset	94.4
	2019	(Ferreira et al., 2019)	HS	CNN	static, RGB, Depth	real-time frames	93.17, 92.61
	2019	(Lim et al., 2019)	HS	CNN	dynamic, RGB	RWTH-BOSTON-50	89.33
	2019	(Lim et al., 2019)	HS	CNN	dynamic, RGB	ASLLVD	31.50
	2019	(Kopuklu et al., 2019)	HG	CNN	dynamic, RGB	EgoGesture	94.03
	2019	(Kopuklu et al., 2019)	HG	CNN	dynamic, RGB	NVIDIA benchmarks	83.83
	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	isoGD	72.53
	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	SKIG	99.72
	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	NATOPS	95.87
	2020	(Elboushaki et al., 2020)	HG	CNN	dynamic, RGB, Depth	SBU	97.51
mAP	2017	(Dibia, 2017)	HT	CNN	static, RGB	Egohands	96.86
Error	2014	(Tompson et al., 2014)	HP	CNN	2D, Depth	proposed dataset	0.33
	2017	(Yuan et al., 2017)	HP	CNN	2D, Depth	ICVL, NYU, MSRC, BigHand2.2M	12.3, 21.4, 21.3, 17.1
	2019	(Li et al., 2019)	HP	CNN	static, RGB	STB	8.34
	2019	(Li et al., 2019)	HP	CNN	static, RGB	B2RGB-SH	7.18
	2015	(Tagliasacchi et al., 2015)	HT	CNN	2D, Depth	real video samples	8 mm
	2015	(Oberweger et al., 2015)	HP	CNN	3D, Depth	ICVL, NYU	10 mm, 20 mm
	2016	(Oberweger et al., 2016)	HP	CNN	2D, 3D, Depth	MSRA	5.58 mm
	2017	(Deng et al., 2017)	HP	CNN	3D, Depth	ICVL, NYU	11 mm, 17 mm
	2017	(Guo et al., 2017)	HP	CNN	2D, Depth	ICVL, NYU, MSRA, ITOP	7.8 mm, 34.00, 9.80 mm, 15.10
	2017	(Simon et al., 2017)	HD	CNN	RGB	own dataset	4.15 mm
	2017	(Fang & Lei, 2017)	HP	CNN, AE	2D, Depth	NYU, ICVL	17 mm, 9 mm
	2017	(Madadi et al., 2017)	HP	CNN	2D, Depth	NYU, MSRA	15.6 mm, 18 mm
	2018	(Chen et al., 2020)	HP	CNN	2D, Depth	ICVL, NYU, MSRA	6.793 mm, 11.811 mm, 8.649 mm
	2018	(Wang et al., 2018)	HP	CNN	2D, Depth	Human3.6M	62.8 mm
	2016	(Sinha et al., 2016)	HP	CNN	static, Depth	Dexter1	16.35 mm
	2016	(Sinha et al., 2016)	HP	CNN	static, Depth	NYU	9 mm
	2017	(Ge et al., 2017)	HP	CNN	static, Depth	NYU	9 mm
	2017	(Dibra et al., 2017)	HP	CNN	static, Depth	NYU	9 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	static, Depth	NYU	9.46 mm
	2018	(Moon et al., 2018)	HP	CNN	static, Depth	NYU	8.42 mm
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	NYU	14.1 mm
	2018	(Kazakos et al., 2018)	HP	CNN	static, RGB, Depth	NYU	9 mm
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	NYU	10.5 mm
	2017	(Dibra et al., 2017)	HP	CNN	static, Depth	ICVL	8 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	static, Depth	ICVL	8 mm
	2018	(Moon et al., 2018)	HP	CNN	static, Depth	ICVL	6.28 mm
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	ICVL	8.5 mm
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	ICVL	19.5 mm
	2017	(Ge et al., 2017)	HP	CNN	static, Depth	MSRA	6 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	static, Depth	MSRA	8 mm
	2018	(Moon et al., 2018)	HP	CNN	static, Depth	MSRA	7.49 mm
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	MSRA	12.5 mm
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	MSRA	10 mm
	2019	(Gomez-Donoso et al., 2019)	HP	CNN	static, RGB, Depth	STB	5 mm
	2019	(Gomez-Donoso et al., 2019)	HP	CNN	static, RGB, Depth	own dataset	5 mm
	2018	(Baek et al., 2018)	HP	GAN	static, Depth	Big Hand 2.2M	13.7 mm
Run time	2015	(Tang et al., 2015)	HP	CNN, DNN	2D, Depth, RGB	own dataset	0.899 ms
Precision	2015	(Koller, Ney et al., 2015)	HSR	CNN	2D, RGB	RWTH-PHOENIX-Weather	55.70
AUC	2017	(Zimmermann & Brox, 2017)	HP	CNN	3D, RGB	SPT and Dexter	94.0, 49.0
	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	3D, Depth	UBC3V, ITOP	88.2, 97.5
	2018	(Mueller et al., 2018)	HT	CNN	static, RGB	STB	96.5
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	STB	98.3
	2018	(Mueller et al., 2018)	HT	CNN	static, RGB	Dexter	64.0
	2018	(Mueller et al., 2018)	HT	CNN	static, RGB	EgoDexter	54.0
	2018	(Spurr et al., 2018)	HP	VAE	static, RGB, Depth	RHD	84.9

Table 6
Deep sign language recognition models, categorized based on the feature types.

Goal	Year	Ref.	Model	Modality	Dataset	Results
HP	2014	(Neverova et al., 2014)	CNN	2D, Depth	proposed dataset	82 (Acc.)
	2014	(Toshev & Szegedy, 2014)	DNN	2D, RGB	FLIC, LSP	96, 78
	2016	(Newell et al., 2016)	CNN	2D, RGB	FLIC, MPII	99 (Elbow), 90.90 (total)
	2016	(Wei et al., 2016)	CNN	2D, RGB	MPII, LSP, FLIC	87.95, 84.32, 97.59
	2016	(Haque et al., 2016)	CNN	2D, Depth	EVAL, ITOP	74.10, 80.50
	2014	(Tompson et al., 2014)	CNN	2D, Depth	proposed dataset	33 (error)
	2017	(Yuan et al., 2017)	CNN	2D, Depth	ICVL, NYU, MSRC, BigHand2.2M	12.3, 21.4, 21.3, 17.1 (error)
	2015	(Oberweger et al., 2015)	CNN	3D, Depth	ICVL, NYU	10 mm, 20 mm
	2016	(Oberweger et al., 2016)	CNN	2D, 3D, Depth	MSRA	5.58 mm (error)
	2017	(Deng et al., 2017)	CNN	3D, Depth	ICVL, NYU	11 mm, 17 mm
	2017	(Guo et al., 2017)	CNN	2D, Depth	ICVL, NYU, MSRA, ITOP	7.8 mm, 66.00, 9.80 mm, 84.90
	2017	(Fang & Lei, 2017)	CNN, AE	2D, Depth	NYU, ICVL	17 mm, 9 mm
	2017	(Madadi et al., 2017)	CNN	2D, Depth	NYU, MSRA	15.6 mm, 18 mm
	2018	(Chen et al., 2020)	CNN	2D, Depth	ICVL, NYU, MSRA	6.793 mm, 11.811 mm, 8.649 mm
	2018	(Wang et al., 2018)	CNN	2D, Depth	Human3.6M	62.8 mm
	2015	(Tang et al., 2015)	CNN, DNN	2D, Depth, RGB	own dataset	0.899 ms(Ave. time)
	2017	(Zimmermann & Brox, 2017)	CNN	3D, RGB	Stereo Hand Pose Tracking Benchmark and Dexter	94, 49.0 (AUC)
	2018	(Marı n Jimeneza et al., 2018)	CNN	3D, Depth	UBC3V, ITOP	88.2, 97.5 (AUC)
	2016	(Sinha et al., 2016)	CNN	static, Depth	Dexter1	16.35 mm
	2016	(Sinha et al., 2016)	CNN	static, Depth	NYU	9 mm
	2017	(Ge et al., 2017)	CNN	static, Depth	MSRA	6 mm
	2017	(Ge et al., 2017)	NN	static, Depth	NYU	9 mm
	2017	(Dibra et al., 2017)	CNN	static, Depth	ICVL	8 mm
	2017	(Dibra et al., 2017)	CNN	static, Depth	NYU	9 mm
	2018	(Ge, Liang et al., 2018)	CNN	static, Depth	NYU	9.46 mm
	2018	(Ge, Liang et al., 2018)	CNN	static, Depth	ICVL	8 mm
	2018	(Ge, Liang et al., 2018)	CNN	static, Depth	MSRA	8 mm
	2018	(Moon et al., 2018)	CNN	static, Depth	ICVL	6.28 mm
	2018	(Moon et al., 2018)	CNN	static, Depth	NYU	8.42 mm
	2018	(Moon et al., 2018)	CNN	static, Depth	MSRA	7.49 mm
	2018	(Baek et al., 2018)	GAN	static, Depth	ICVL	8.5 mm
	2018	(Baek et al., 2018)	GAN	static, Depth	MSRA	12.5 mm
	2018	(Baek et al., 2018)	GAN	static, Depth	NYU	14.1 mm
	2018	(Baek et al., 2018)	GAN	static, Depth	Big Hand 2.2M	13.7 mm
	2018	(Kazakos et al., 2018)	CNN	static, RGB, Depth	NYU	9 mm
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	STB	98.3(AUC)
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	RHD	84.9(AUC)
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	ICVL	19.5 mm
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	NYU	10.5 mm
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	MSRA	10 mm
	2019	(Gomez-Donoso et al., 2019)	CNN	static, RGB, Depth	own dataset	5 mm
	2019	(Gomez-Donoso et al., 2019)	CNN	static, RGB, Depth	STB	5 mm
	2019	(Li et al., 2019)	CNN	static, RGB	STB	8.34 (err)
	2019	(Li et al., 2019)	CNN	static, RGB	B2RGB-SH	7.18 (err)
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	STB	98.3(AUC)
	2018	(Spurr et al., 2018)	VAE	static, RGB, Depth	RHD	84.9(AUC)
	2019	(Li et al., 2019)	CNN	static, RGB	STB	8.34 (err)
	2019	(Li et al., 2019)	CNN	static, RGB	B2RGB-SH	7.18 (err)
RHR	2018	(Rastgoo et al., 2018)	RBM	2D, RGB, Depth	Massey 2012, SL Surrey, NYU, ASL Fingerspelling A	99.31, 97.56, 90.01, 98.13
HG	2016	(Han et al., 2016)	CNN	2D, RGB	proposed dataset	93.80
	2016	(Duan et al., 2016)	CNN	2D, Depth, RGB	Chalearn IsoGD, RGBD-HuDaAct	67.19, 96.74
	2017	(Wang et al., 2017)	CNN	2D, Depth	ChaLearn	55.57
	2018	(Dadashzadeh et al., 2018)	CNN	2D, Depth	OUHANDS,	86.46
	2019	(Kopuklu et al., 2019)	CNN	dynamic, RGB	EgoGesture	94.03
	2019	(Kopuklu et al., 2019)	CNN	dynamic, RGB	NVIDIA benchmarks	83.83
	2020	(Elboushaki et al., 2020)	CNN	dynamic, RGB, Depth	isoGD	72.53
	2020	(Elboushaki et al., 2020)	CNN	dynamic, RGB, Depth	SKIG	99.72
	2020	(Elboushaki et al., 2020)	CNN	dynamic, RGB, Depth	NATOPS	95.87
	2020	(Elboushaki et al., 2020)	CNN	dynamic, RGB, Depth	SBU	97.51
	2019	(Chen et al., 2019)	CNN	dynamic, RGB	DHG-14/28 Dataset	91.9
	2019	(Chen et al., 2019)	CNN	V	SHREC'17 Track Dataset	94.4
HSR	2018	(Rao et al., 2018)	CNN	2D, RGB	own dataset	92.88
	2018	(Ye et al., 2018)	CNN	3D, RGB	own dataset	69.2
	2015	(Koller, Ney et al., 2015)	CNN	2D, RGB	RWTH-PHOENIX-Weather	55.70 (Precision)
	2020	(Wadhawan & Kumar, 2020)	CNN	static, RGB	own dataset	99.72
	2019	(Ferreira et al., 2019)	CNN	static, RGB, Depth	real-time frames	93.17, 92.61
	2019	(Lim et al., 2019)	CNN	dynamic, RGB	RWTH-BOSTON-50	89.33
	2019	(Lim et al., 2019)	CNN	dynamic, RGB	ASLLVD	31.50
HD	2015	(Kang et al., 2015)	CNN	Depth	proposed dataset	99
	2017	(Simon et al., 2017)	CNN	RGB	own dataset	4.15 mm

(continued on next page)

Table 6 (*continued*).

Goal	Year	Ref.	Model	Modality	Dataset	Results
HT	2015	(Tagliasacchi et al., 2015)	CNN	dynamic, Depth	real video samples	8 mm (MSE)
	2017	(Dibia, 2017)	CNN	static, RGB	Egohands	96.86 (mAP)
	2018	(Mueller et al., 2018)	CNN	static, RGB	STB	96.5 (AUC)
	2018	(Mueller et al., 2018)	CNN	static, RGB	Dexter	64 (AUC)
	2018	(Mueller et al., 2018)	CNN	static, RGB	EgoDexter	54 (AUC)

Stereo Hand Pose Tracking Benchmark show that the model outperformed state-of-the-art alternatives in hand pose estimation obtaining a Percentage of Correct Keypoints (PCK) of 99.85% with 9.85% relative improvement (Gomez-Donoso et al., 2019). Spurr et al. designed a model to learn a statistical hand model represented by a trained cross-modal latent space via a VAE. They derive the variational lower bound that permits training of a single latent space using multiple modalities. In this space, similar input poses are embedded close to each other independent of the input modality. While the proposed model is trained on the RGB input images, it is tested on different combinations of modalities where the aim is to estimate 3D hand poses as output. In parallel, the VAE framework permits to generate samples consistently in each modality. Reported results using AUC metric show 17.4 relative improvement in model performance in comparison with the state-of-the-art models in hand pose estimation on the RHD dataset (Spurr et al., 2018).

4.2.2. Depth-based hand pose models

Oberweger et al. have proposed several CNN architectures to predict the 3D hand joint locations from a depth map. They used a learned prior model as a bottleneck layer, with fewer neurons than the last layer, to improve the hand pose estimation accuracy. Furthermore, a refinement stage has been provided using spatial pooling and sub-sampling of multiple input regions centered on the initial joints estimation. This model achieved the estimation error of 10 mm and 20 mm on ICVL and NYU datasets obtaining 1 mm and 3 mm relative improvement in comparison with state-of-the-art models. Performance analysis of this model show that the location prediction of the joints is constrained by the learned hand model. This means that if the missing regions of prediction get too large, the accuracy gets worse (Oberweger et al., 2015). A 3D neural network architecture has been provided by Deng et al. to 3D hand pose estimation from a single depth image. They converted an input depth map to a 3D volumetric representation and fed it into a 3D CNN without any ground truth reference point for network initialization. Some synthetic depth images have been rendered from existing real image datasets to increase the training data. They achieved an estimation accuracy of 74% and 96% with 9% and 3% relative improvement in comparison with state-of-the-art alternatives on NYU and ICVL datasets (Deng et al., 2017). Guo et al. suggested a tree-structured Region Ensemble Network (REN) for hand pose estimation using a regression-based method. Partitioning the last convolution outputs of the CNN into several grid regions, the results from fully-connected (FC) regressors on each region are fused and input to another FC layer for final estimation. Different training strategies have been used to improve the performance of the joints localization. The results of the model evaluation on three public datasets, ICVL, NYU, and ITOP datasets showed that their model achieved the state-of-the-art results for hand pose estimation with 1.05 mm, 3.7 mm, and 4.4% of relative improvement (Guo et al., 2017). Oberweger et al. proposed an efficient and accurate method to put a label for each frame of a hand depth video and estimate the 3D locations of the joints. They sample some frames of each video by some users, namely reference frames. Users provide the initial 2D estimations of the visible joints. After that, some spatial, temporal, and appearance constraints are applied to these 2D joints in order to estimate the full 3D poses of the hand over the complete sequence. While the evaluation results on MSRA dataset showed that the model outperformed state-of-the-art model with 4.38 mm relative improvement, their model was too complicated to use (Oberweger et al., 2016). Another deep learning approach, including a CNN with an embedded forward kinematics based layer for intermediate representation in the network, has been proposed by Zhou et al. to hand pose estimation. The proposed model has used prior geometric knowledge in the learning process. While the model has some problems with inaccurate joint annotations and small viewpoint changes of the ICVL dataset, it achieves an estimation accuracy of 16.9 mm on NYU dataset. Result of this model is comparable with state-of-the-art alternatives on NYU with an approximately 1 mm relative distance (Zhou, Wan, Zhang, Xue, & Wei, 2016). Ge et al. proposed a multi-view CNN-based model to project the query depth image onto three orthogonal planes and regress the 2D heat-maps of each plan in order to estimate the hand joint positions. The final 3D hand pose estimation with learned pose priors is achieved using the multi-view heat-maps. Experimental results on NYU, ICVL, and MSRA datasets show that the proposed method outperforms state-of-the-art alternatives achieving a relative improvement of 1.1 mm, 0.5 mm, and 0.5 mm on these datasets, respectively (Ge, Ren, & Yuan, 2018). Fang and Lei suggested a CNN model with an embedding denoising auto-encoder in the bottom layer of the network to hand pose estimation. used an auto-encoder as the nonlinear dimension converter in the model to learn the prior knowledge and consolidate this embedding layer monotonically into the CNN to improve hand pose estimation accuracy. They reported a relative state-of-the-art improvement of approximately 1 mm and 2 mm on ICVL and NYU datasets (Fang & Lei, 2017). Chen et al. suggested a Pose guided structured Region Ensemble Network (Pose-REN) to enhance the estimation accuracy of hand pose from a depth image. They partitioned the CNN feature maps into some regions and fused the estimated joints of these regions based on a tree-structured fully connections. A refined estimation and iterative cascaded method have been applied to estimate the final hand pose. Evaluation results showed that the model obtained state-of-the-art with a relative improvement of 7.04%, 0.88 mm, and 2 mm on ICVL, NYU, and MSRA datasets, respectively (Chen et al., 2020). Yuan et al. proposed a tracking system using the magnetic sensors and inverse kinematics to automatically acquire the hand joints annotations from a depth map. They also provided a dataset using six magnetic sensors, five on each fingernail and one on the back of the palm, to record the 6D measurements of the joints. After that, inverse kinematics with 31 degrees of freedom (dof) and kinematic constraints have been applied to estimate the final joint locations. This model achieved a relative state-of-the-art improvement of 2.6 mm, 1.2 mm, and 29.2 mm in estimation error on ICVL, NYU, and Bighand datasets, respectively (Yuan et al., 2017). Supancic et al. provided an analysis of the state-of-the-art methods focusing on hand pose estimation from a single depth frame. They defined an evaluation metric and a simple nearest-neighbor baseline to compare the recognition accuracy of different models. Comparison results of some existence models on NYU, ICL, and EGO datasets have been presented and analyzed extensively using the proposed metric. Evaluation results confirm the effectiveness of the proposed metric for analyzing different methods and datasets (Supancic et al., 2018). Moon et al. designed a model for mapping 3D hand and human pose estimation parameters into a voxel-to-voxel space to estimate the likelihood for each keypoint per each voxel. This model benefits from a 3D CNN for real-time estimation of hand keypoints. Evaluation results on the three publicly available 3D hand and human pose estimation datasets, used in the HANDS 2017 challenge, show that the model outperforms the other models in hand and human pose estimation. This model achieved a

Table 7
Deep sign language recognition models, categorized based on the input modality.

Modality	Year	Ref.	Goal	Model	Dataset	Results
2D, Depth	2014	(Neverova et al., 2014)	HP	CNN	own dataset	82 (Acc.)
	2014	(Neverova et al., 2014)	HP	CNN	own dataset	82 (Acc.)
	2014	(Tompson et al., 2014)	HP	CNN	own dataset	33 (error)
	2015	(Kang et al., 2015)	HD	CNN	own dataset	99
	2017	(Guo et al., 2017)	HP	CNN	NYU	66
	2017	(Fang & Lei, 2017)	HP	CNN, AE	NYU	17 mm
	2017	(Yuan et al., 2017)	HP	CNN	NYU	21.4 (error)
	2017	(Madadi et al., 2017)	HP	CNN	NYU	15.6 mm
	2018	(Chen et al., 2020)	HP	CNN	NYU	11.811 mm
	2016	(Sinha et al., 2016)	HP	CNN	NYU	9 mm
	2017	(Ge et al., 2017)	HP	CNN	NYU	9 mm
	2017	(Dibra et al., 2017)	HP	CNN	NYU	9 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	NYU	9.46 mm
	2018	(Moon et al., 2018)	HP	CNN	NYU	8.42 mm
	2018	(Baek et al., 2018)	HP	GAN	NYU	14.1 mm
	2017	(Guo et al., 2017)	HP	CNN	ICVL	7.8 mm
	2017	(Fang & Lei, 2017)	HP	CNN, AE	ICVL	9 mm
	2017	(Yuan et al., 2017)	HP	CNN	ICVL	12.3 (error)
	2017	(Dibra et al., 2017)	HP	CNN	ICVL	8 mm
	2018	(Chen et al., 2020)	HP	CNN	ICVL	6.793 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	ICVL	8 mm
	2018	(Moon et al., 2018)	HP	CNN	ICVL	6.28 mm
2D, Depth	2018	(Baek et al., 2018)	HP	GAN	ICVL	8.5 mm
	2017	(Guo et al., 2017)	HP	CNN	MSRA	9.80 mm
	2017	(Madadi et al., 2017)	HP	CNN	MSRA	18 mm
	2017	(Yuan et al., 2017)	HP	CNN	MSRA	21.3 (error)
	2017	(Ge et al., 2017)	HP	CNN	MSRA	6 mm
	2018	(Chen et al., 2020)	HP	CNN	MSRA	8.649 mm
	2018	(Ge, Liang et al., 2018)	HP	CNN	MSRA	8 mm
	2018	(Moon et al., 2018)	HP	CNN	MSRA	7.49 mm
	2018	(Baek et al., 2018)	HP	GAN	MSRA	12.5 mm
	2017	(Wang et al., 2017)	HG	CNN	ChaLearn IsoGD	55.57
	2016	(Haque et al., 2016)	HP	CNN	ITOP	80.50
	2017	(Guo et al., 2017)	HP	CNN	ITOP	84.90
	2016	(Haque et al., 2016)	HP	CNN	EVAL	74.10
	2015	(Tagliasacchi et al., 2015)	HT	CNN	real video samples	8 mm (MSE)
	2017	(Yuan et al., 2017)	HP	CNN	BigHand2.2M	17.1 (error)
	2018	(Baek et al., 2018)	HP	GAN	Big Hand 2.2M	13.7 mm
	2018	(Wang et al., 2018)	HP	CNN	Human3.6M	62.8 mm
	2018	(Dadashzadeh et al., 2018)	HG	CNN	OUHANDS,	86.46
3D, Depth	2015	(Oberweger et al., 2015)	HP	CNN	NYU	20 mm
	2017	(Deng et al., 2017)	HP	CNN	NYU	17 mm
	2015	(Oberweger et al., 2015)	HP	CNN	ICVL	10 mm
	2017	(Deng et al., 2017)	HP	CNN	ICVL	11 mm
	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	ITOP	97.5 (AUC)
	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	UBC3V	88.2 (AUC)
2D, RGB	2016	(Han et al., 2016)	HG	CNN	own dataset	93.80
	2017	(Simon et al., 2017)	HD	CNN	own dataset	4.15 mm
	2018	(Rao et al., 2018)	HSR	CNN	own dataset	92.88
	2020	(Wadhawan & Kumar, 2020)	HSR	CNN	own dataset	99.72
	2014	(Toshev & Szegedy, 2014)	HP	DNN	FLIC	96
	2016	(Newell et al., 2016)	HP	CNN	FLIC	99.0 (Elbow)
	2016	(Wei et al., 2016)	HP	CNN	FLIC	97.59
	2014	(Toshev & Szegedy, 2014)	HP	DNN	LSP	0.78
	2016	(Wei et al., 2016)	HP	CNN	LSP	84.32
	2016	(Newell et al., 2016)	HP	CNN	MPII	90.90 (total)
	2016	(Wei et al., 2016)	HP	CNN	MPII	87.95
	2018	(Mueller et al., 2018)	HT	CNN	STB	96.5 (AUC)
	2019	(Li et al., 2019)	HP	CNN	STB	8.34 (err)
	2015	(Koller, Ney et al., 2015)	HSR	CNN	RWTH-PHOENIX-Weather	55.70 (Precision)
	2017	(Dibia, 2017)	HT	CNN	Egohands	96.86 (mAP)
	2018	(Mueller et al., 2018)	HT	CNN	Dexter	64 (AUC)
	2018	(Mueller et al., 2018)	HT	CNN	EgoDexter	54 (AUC)
	2019	(Li et al., 2019)	HP	CNN	B2RGB-SH	7.18 (err)
3D, dynamic, RGB	2018	(Ye et al., 2018)	HSR	CNN	own dataset	69.2
	2017	(Zimmermann & Brox, 2017)	HP	CNN	STB	94 (AUC)
	2017	(Zimmermann & Brox, 2017)	HP	CNN	Dexter	49.0 (AUC)
	2019	(Lim et al., 2019)	HS	CNN	RWTH-BOSTON-50	89.33
	2019	(Lim et al., 2019)	HS	CNN	ASLLVD	31.50
	2019	(Chen et al., 2019)	HG	CNN	DHG-14/28 Dataset	91.9
	2019	(Chen et al., 2019)	HG	CNN	SHREC'17 Track Dataset	94.4

(*continued on next page*)

Table 7 (*continued*).

Modality	Year	Ref.	Goal	Model	Dataset	Results
2D, RGB, Depth	2018	(Rastgoo et al., 2018)	HR	RBM	Massey 2012	99.31
	2018	(Rastgoo et al., 2018)	HR	RBM	SL Surrey	97.56
	2018	(Rastgoo et al., 2018)	HR	RBM	ASL Fingerspelling A	98.13
	2019	(Gomez-Donoso et al., 2019)	HP	CNN	own dataset	5 mm
	2018	(Rastgoo et al., 2018)	RHR	RBM	NYU	90.01
	2018	(Kazakos et al., 2018)	HP	CNN	NYU	9 mm
	2018	(Spurr et al., 2018)	HP	VAE	NYU	10.5 mm
	2018	(Spurr et al., 2018)	HP	VAE	ICVL	19.5 mm
	2018	(Spurr et al., 2018)	HP	VAE	MSRA	10 mm
	2016	(Duan et al., 2016)	HG	CNN	Chalearn IsoGD	67.19
	2016	(Duan et al., 2016)	HG	CNN	RGBD-HuDaAct	96.74
	2015	(Tang et al., 2015)	HP	CNN, DNN	own dataset	0.899 ms (Ave. time)
	2018	(Spurr et al., 2018)	HP	VAE	RHD	84.9 (AUC)
	2019	(Ferreira et al., 2019)	HS	CNN	real-time frames	93.17, 92.61
	2018	(Spurr et al., 2018)	HP	VAE	STB	98.3(AUC)
	2019	(Gomez-Donoso et al., 2019)	HP	CNN	STB	5 mm
dynamic, RGB, Depth	2020	(Elboushaki et al., 2020)	HG	CNN	isoGD	72.53
	2020	(Elboushaki et al., 2020)	HG	CNN	SKIG	99.72
	2020	(Elboushaki et al., 2020)	HG	CNN	NATOPS	95.87
	2020	(Elboushaki et al., 2020)	HG	CNN	SBU	97.51
	2019	(Kopuklu et al., 2019)	HG	CNN	EgoGesture	93.75, 94.03
	2019	(Kopuklu et al., 2019)	HG	CNN	NVIDIA benchmarks	78.63, 83.83
2D, 3D, Depth	2016	(Oberweger et al., 2016)	HP	CNN	MSRA	5.58 mm (Ave. err.)

relative improvement of 0.51 mm, 3.39 mm, and 1.16 mm on the ICVL, NYU, and MSRA datasets, respectively (Moon et al., 2018). Dibra et al. used a simple CNN which is pre-trained only on synthetic depth images generated from a single 3D hand model. They fine-tuned the model on the unlabeled depth images from a real user's hand. Experimental results on two public datasets showed that the model performance was comparable with state-of-the-art methods in hand pose estimation obtaining an error estimation of 8 mm and 9 mm on ICVL and NYU datasets (Dibra et al., 2017). Baek et al. applied a GAN for hand pose estimation by making a one to one relation between depth disparity maps and 3D hand pose models. This model refines the initial skeleton estimations for further accuracy improvement. Evaluation results on ICVL, MSRA, NYU, and Big Hand 2.2M demonstrate that the proposed model is on par with state-of-the-art models in hand pose estimation, achieving an estimation error of 8.5 mm, 12.5 mm, 14.1 mm, and 13.7 mm on these datasets, respectively (Baek et al., 2018). Ge et al. proposed a 3DCNN for real-time hand pose estimation from single depth images. A 3D volumetric representation of the hand depth image is fed to a 3DCNN to get the 3D spatial structure of the input image. A 3D data augmentation is performed on the training data to increase the robustness of the input images to variations in hand sizes and global orientations. Results on MSRA show that the proposed model outperforms state-of-the-art methods in hand pose estimation with a relative improvement of 3 mm. Furthermore, the evaluation results on the proportion of joints within in different error thresholds on NYU dataset confirm the relative state-of-the-art improvement of 10% (Ge et al., 2017). Sinha et al. proposed a CNN-based model for real-time 3D hand pose estimation using depth data. They provided a hierarchical pipeline for hand pose estimation using the combination of global pose orientation and finger articulations in a principled way. They used an efficient matrix completion method for joint angle parameters estimation using the initialized pose matrix. Experimental results on Dexter1 confirm a relative state-of-the-art improvement of 3.25 mm. Indeed, this model achieved a relative state-of-the-art improvement of approximately 4% in estimation accuracy on NYU dataset (Sinha et al., 2016).

4.2.3. Multi-modal hand pose estimation models

Tang et al. proposed a real-time hand pose estimation model using the combination of traditional and deep models. While they used morphological transform from two modalities for hand detection, a Deep Neural Networks (DBNs) has been applied to hand pose estimation in real-time. They evaluated the model on some provided videos and claimed that the proposed model is fast in all stages achieving the average recognition time of 0.899 ms per each video input (Tang et al., 2015). Kazakos et al. designed a CNN-based model using the fusion of RGB and Depth information in a double-stream architecture for hand pose estimation. While the RGB and depth images are fed into two separate CNNs for feature extraction, the intermediate layers of the CNNs are fused for final hand pose estimation. Evaluation results demonstrate that while the depth of the network is crucial for hand pose estimation, the double stream nets performs very similarly with the net trained only with depth images. The proposed model obtained a comparable estimation error with state-of-the-art methods on NYU Hand pose dataset (Kazakos et al., 2018).

4.2.4. Discussion

With the advent of deep learning for hand sign language recognition in recent years, having a large amount of data has been proven to be a crucial part of the model learning. Using depth cameras has facilitated the creation of large-scale datasets with automatic annotations of key-point locations for data using some magnetic sensors attached to the hand. While these magnetic sensors provide accurate annotations for depth inputs, they are not effective for RGB inputs because they deform the hand appearance in the RGB inputs. So, most of the proposed models in hand pose estimation area are based on the depth inputs and few models have been suggested to RGB inputs. A broad benchmark evaluation has shown that deep models appear particularly well-suited for pose estimation (Supancic et al., 2018). The impressive capability of CNN to work with the still images is not enough for video inputs and it needs to be combined with another deep approach for video inputs in order to more cover the sequence information. Based on the performance analysis of different CNN structures with regard to hand shape, joint visibility, viewpoint, and articulation distributions, 3D hand pose estimation, using deep learning approaches, in the isolated scenes have a lower mean error in comparison with the cluttered and complex scenes. Furthermore, it is possible to integrate the forward kinematic process of an articulated hand model into the deep learning framework for accurate hand pose estimation (Supancic et al., 2018). Using the prior knowledge in geometric hand model in the learning process also has led to significant results (Zhou et al., 2016).

While hand pose estimation roughly has been solved for scenes with isolated hands, the proposed models still struggle to analyze cluttered scenes where hands may be interacting with the other objects and surfaces. Furthermore, self-occlusion (between fingers), close similarity between fingers, dexterity of the hands, speed of the pose and high

dimension of the hand kinematic parameters are so challenging. Also, when segmentation is hard due to active hands or clutter, many existing models fail to work properly. Using the multi-modal inputs, such as image, video, skeleton, flow features, text, and so on, along with the real and diverse training sets can be more considered because the input data is as important as the choice of model architecture. The current datasets for hand pose estimation are not suitable enough to apply in real communication of sign language because they are highly restricted not only in environmental conditions and background complexities but also in numbers of the signs in each video. In other words, we need a dataset that includes the daily phrases or sentences similar to the real world, not in a constrained environment with the predefined slots. Articulated hand pose estimation is still an open problem and under intensive research from both academia and industry. We think that using the complementary features, such as face and body, more real and rich datasets, benefited from the face and human datasets, along with the new hardware and wearable devices capabilities could improve the hand pose estimation accuracy in the future.

4.3. Real-time hand tracking

Hands are the most important object in the inputs of the sign language recognition models. Tracking the detected hands are one of the substantial challenges for video inputs due to high occlusions of hand fingers and joints. In this sub-section, we review the deep-based suggested models for hand tracking in the last four years.

4.3.1. RGB-based real-time hand tracking

Dibia proposed a repository to real-time hand detection using SSD model. He used the Tensorflow library for implementation of the model. A multi-stage process, including assembling dataset, cleaning, splitting into training/test partitions and generating an inference graph, has been used to train the model. The evaluation results on Egohands dataset showed that the model had a comparable performance with the state-of-the-art models in this area with a detection accuracy of 98.86% (Dibia, 2017). Mueller et al. designed a real-time 3D hand tracking model using the combination of CNN with a kinematic 3D hand model from Monocular RGB input. This model is robust to occlusions and varying camera viewpoints. A novel geometrically consistent GAN image-to-image translation network is used for synthetic generation of training data. In other words, they proposed a model that translates synthetic images to real images with the same statistical distribution as real-world hand images. Furthermore, they contributed to a new RGB dataset with annotated 3D hand joint positions. Evaluation results on the Stereo and Dexter datasets show a relative state-of-the-art accuracy improvement of 1.7% and 15% (Mueller et al., 2018).

4.3.2. Depth-based real-time hand tracking

Kang et al. proposed a real-time finger-spelling recognition model using CNN from the depth map. They have recorded some depth videos in multiple subjects and used different learning configurations to classify the alphabets and numbers. The evaluation results on their dataset have led to the accuracy of 0.99% (Kang et al., 2015).

4.3.3. Multi-modal real-time hand tracking

Kopuklu et al. presented a hierarchical and two-stream CNN model to operate online efficiently by using sliding window approach for real-time hand gesture detection and classification from two modalities, RGB and Depth videos. The proposed model includes two CNNs for detection and classification. Evaluation results on EgoGesture and NVIDIA Dynamic Hand Gesture Datasets show the relative state-of-the-art accuracy improvement of 4.45% and 4.53% (Kopuklu et al., 2019).

4.3.4. Discussion

Real-time hand detection is now an attractive area in the research community as a next step to provide a complete system for online human communication in desktop environments. Using the deep learning approaches along with the recent improvements of GPUs have led to impressive improvement in accuracy and speed for real-time hand tracking area. Some of the models use not only the CNN capabilities, as a deep learning model, but also the advantages of a depth sensor for input data providing (Kang et al., 2015). Furthermore, the advent of fast CNN-based models such as Faster-RCNN (Ren et al., 2015), YOLO (Redmon et al., 2016), and SSD (Liu et al., 2016) make deep-based models as the attractive candidates for real-time hand detection and tracking applications. Accurate and real-time hand tracking is a challenging problem in computer vision area due to highly articulated human hands. Fast processing in an uncontrolled environment considering the rapid hand motions is a too difficult requirement of the real-time hand tracking models. Since it is difficult to satisfy this requirement, some of the hand tracking models apply some restrictions on the user or the environment to facilitate the processing. For example, some models consider only a uniform or static background, avoid the rapid hand motions, or assume the hand as a skin-colored object. These limitations may not be applicable to real-world systems. We think that using the faster hardware to process the input data, a pre-trained model for hand detection and also hand localization trained on a large amount of data, an effective and fast refinement approach to correct the false hand detection, and also benefiting from multi-modal inputs could improve the hand tracking accuracy models.

4.4. Hand gesture recognition

Hand gesture recognition area, as an engrossing area in computer vision, can provide fundamental information for HCI applications. There are some undesirable challenges related to high hands occlusions and background complexities in this area. A good discussion on these challenges can be found in Escalera, Athitsos, and Guyon (2016). Here, in this subsection, we present deep-based models for hand gesture recognition in four recent years.

4.4.1. RGB-based hand gesture models

A CNN-based model along with a simple Gaussian skin color model and background subtraction has been provided by Han et al. to hand gesture recognition from visual data. The proposed skin model controls the light affection on skin color and filters out non-skin colors of an image. They evaluated the model on own dataset and reported the recognition accuracy of 93.80% from a camera image (Han et al., 2016). Dadashzadeh et al. proposed a two-stage deep model for hand gesture segmentation and recognition. A CNN model has been used in two stages for precise pixel-level semantic segmentation as well as the final hand gesture recognition. A combination of fully convolutional deep residual neural network and atrous spatial pyramid pooling has been used for segmentation step. Evaluation results on OUHANDS dataset show a relative state-of-the-art recognition accuracy improvement of 1.6% for static hand gestures (Dadashzadeh et al., 2018). Elboushaki et al. proposed a multi-dimensional feature learning model, namely MultiD-CNN, for gesture recognition from RGB-D videos. This model benefits from spatiotemporal features of two modalities using the combination of 3DCNN and LSTM. Temporal information of the model input is encoded into a motion representation employing a two-stream architecture based on 2D-ResNets. The proposed model extracts deep features from motion representation by investigating different fusion strategies at different levels. Results of this model on SKIG, NATOPS, and SBU datasets demonstrate a relative state-of-the-art recognition accuracy improvement of 0.19%, 8.52%, and 4.11% (Elboushaki et al., 2020). Chen et al. presented a Dynamic Graph-Based Spatial–Temporal Attention (DG-STA) model for hand gesture

recognition. A fully-connected graph, including a self-attention mechanism for automatically learning the node and edge features in both spatial and temporal domains, is constructed from a hand skeleton. A novel spatial–temporal mask is applied to significantly reduction of the computational cost. Evaluation results on DHG-14/28 and SHREC'17 confirm the superior performance of this model in comparison with the state-of-the-art models in hand gesture recognition with 0.9% and 3% relative accuracy improvement (Chen et al., 2019). Canuto dos Santos et al. developed a deep-based model using two ResNet models and a soft-attention ensemble layer for dynamic gesture recognition. A condensing technique, namely star RGB, is proposed to summarize an input RGB video into only one RGB image. This image is passed to the rest of the model including two ResNets, a soft-attention ensemble, and a fully connected layer for final classification. Experimental results on Montalbano and GRIT datasets show a relative state-of-the-art accuracy improvement of 0.78% and 6.68% (Canuto-dos Santos, Leonid-Aching-Samatelo, & Frizera-Vassallo, 2020).

4.4.2. Depth-based hand gesture models

Wang et al. proposed a CNN model for gesture recognition and evaluated on the Large-scale Isolated Gesture Recognition at the ChaLearn Looking at People (LAP) challenge 2016 and verified the effectiveness of the proposed method. In their model, three representations of depth sequences have been constructed from a sequence of depth maps using bidirectional rank pooling to provide the spatio-temporal information. Using these representations, they fine-tuned the CNN model trained on the image data for classification of depth sequences without considering large parameters to learn. This model achieved a relative state-of-the-art accuracy improvement of 16.34% on IsoGD dataset (Wang et al., 2017).

4.4.3. Skeleton-based hand gesture recognition

Devineau et al. provided a deep CNN model for hand gesture recognition using only hand-skeletal data. They used the parallel convolutions to train the sequences of hand-skeletal positions of the joints in different time resolutions. The evaluation results on the DHG dataset showed that the model achieved the state-of-the-art results by 3% relative improvement. Their model demonstrated that the parallel processing of sequences using CNNs can be competitive with neural architectures that use some cells specifically designed for sequences, such as GRU and LSTM cells. The biggest drawback of the model is that it only works on complete sequences of input data. Furthermore, due to the weight sharing between all channels, the overall model performance is decreased. So, they need to have a trade-off between model accuracy and its total parameters count, respectively (Devineau, Xi, Moutarde, & Yang, 2018).

4.4.4. Multi-modal hand gesture models

Duan et al. provided a convolutional Two-Stream Consensus Voting Network (2SCVN) to explicitly model the short-term and long-term structure of the RGB sequences. To decrease the complexity of the background, a 3D Depth-Saliency CNN stream (3DDSN) has been used in parallel to provide the motion features. These two networks, 2SCVN and 3DDSN, have been fused in one framework to improve the recognition accuracy. The evaluation results of the multi-modal model on Chalearn IsoGD benchmark and RGBD-HuDaAct dataset showed that the model outperformed the state-of-the-art models on these datasets with a relative improvement of 4.47% and 0.61% (Duan et al., 2016). Molchanov et al. proposed a recurrent 3D CNN-based model for dynamic hand gesture recognition from multi-modal inputs. Four input modalities, including RGB, depth, OF, and stereo-IR sensors data, have been fused to boost the recognition accuracy of the model. They achieved the state-of-the-art results on SKIG and ChaLearn datasets with a relative accuracy improvement of 0.9% and 1% (Molchanov, Yang, Gupta, Kim, Tyree, & Kautz, 2016). Wu et al. provided a two-stream CNN-based model for hand gesture recognition and identification from depth map and OF input information. Their model has the capability of multiple gestures generalization for only one person or multiple person generalization for only one gesture. The most interesting contribution of this model is the ability of unseen gestures verification and identification. Results on the MSR Action3D dataset show a relative state-of-the-art gesture recognition improvement of 18.91% (Wu, Chen, Ishwar, & Konrad, 2016). Rastgoo et al. have proposed a hand sign recognition model using RBM from visual data. Two modalities, RGB and depth, have been considered in the model input in three forms: original image, cropped image, and noisy cropped image. In the first step, the hand of each crop is detected using a CNN. After that, for each modality, three forms of an input image have been input to RBMs. The outputs of the RBMs for two modalities are fused in another RBM in order to recognize the output sign label. The proposed multi-modal model is trained on four publicly available datasets, Massey University Gesture Dataset 2012, Fingerspelling Dataset from the University of Surrey's Center for Vision, Speech and Signal Processing, NYU, and ASL Fingerspelling A. Results showed that the model achieved state-of-the-art with a relative accuracy improvement of 27.31%, 28.56%, 2.9%, and 11.13%, respectively (Rastgoo et al., 2018).

4.4.5. Discussion

Hand gestures and gesticulations are a common form of human communication. Vision-based gesture recognition has faced up much attention from both the academic and the industrial communities due to its prominent applications in HCI and sign language area. Many deep-based models have been proposed over the last few years. A CNN-based model along with a simple Gaussian skin color model and background subtraction (Han et al., 2016), a two-stage deep model for hand gesture segmentation and recognition using CNN (Dadashzadeh et al., 2018), a CNN model using only hand-skeletal data (Devineau et al., 2018), a stacked RBMs model (Rastgoo et al., 2018), and a recurrent 3D CNN-based model (Molchanov et al., 2016) are just some of the proposed models for hand gesture recognition using deep learning approaches. Similar to other sub-area of sign language recognition, there are much more gesture recognition models for depth modality than the RGB one due to advancement and availability of depth-sensors such as Kinect. To benefit from the complementary advantages of all input modalities, some models used the multi-modal inputs to fuse the structural information from the depth channel, high-resolution pixel information of RGB inputs, motion features, and also skeletal input data (Duan et al., 2016; Molchanov et al., 2016; Rastgoo et al., 2018). There are still some limitations regarding hand gesture recognition area due to hand gesture variations, illuminations, background complexity, and large diversity in how people perform gestures. Another challenge is related to a real-time gesture recognition area. Since human feedback time to show the gestures in real-world can vary in different gestures, this can present the challenge of detecting and classifying gestures immediately upon or before their completion to provide rapid feedback. Furthermore, using the new hardware capabilities and also recognizing the unseen gestures could be more considered, especially for real-time applications. We think that more adaptive selection of the optimal hyper-parameters of the model can improve the recognition speed of real-time models. Joint features of face gesture, body gesture, face pose, and body pose, as complementary features, can also be useful to improve the hand gesture recognition accuracy. Furthermore, an effective spatio-temporal data augmentation method to deform the input volumes of hand gestures in order to facilitate the final recognition can be helpful. Since data is a central part of the deep learning models, using different forms of input data, such as image or video, OF, skeleton, text, and so on, can provide more accurate hand gesture recognition models.

4.5. Hand pose recovery

Hand pose recovery attracted special attention in recent years due to the availability of low-cost depth cameras such as Microsoft Kinect. Although hand pose estimation area had remarkable improvements using deep learning approaches, 3D hand pose estimation and recovery coped with some challenges. In this sub-section, we present deep-based models for hand pose recovery in four recent years.

4.5.1. Depth-based hand pose recovery

Tompson et al. provided a four-part model including a randomized decision forest classifier for image segmentation, a robust method for labeled dataset generation, a CNN for feature extraction, and an inverse kinematics step for real-time pose recovery. They used the intermediate heat-map features to extract the accurate and reliable 3D pose information at interactive frame-rates using inverse kinematics. While the evaluation results on own dataset showed that the model achieved the recognition error of 33.0 for hand pose recovery, this model can track two hands only if they are not interacting (Tompson et al., 2014). Madadi et al. proposed a hierarchical tree-like structured CNN for hand pose recovery using end-to-end training. They fused the branches in predefined subsets of hand joints to learn the higher order dependencies among joints in the final pose. Furthermore, some appearance and physical constraints of hand motion and deformation have been defined in a loss function. Evaluation results on NYU dataset showed the proposed model outperformed state-of-the-art models in hand pose recovery with a relative improvement of 1.3 mm (Madadi et al., 2017).

4.5.2. Discussion

Hand pose recovery area has been tremendously studied in recent years. Different deep-based models have been proposed by different researchers. Most of the models include CNN-based models that are highly data-dependent models and have powerful capabilities in coping with a highly nonlinear output space. Fusing CNN outputs with another deep approach had also considerable attention among the proposed models.

Although the availability of affordable depth cameras have allowed researchers to use non-invasive, precise, robust to illumination and color changes approaches to hand pose recovery and led to significant advances in this area, there are still several open challenges to tackle. Fingers self-occlusions, hand-body occlusions, low resolution/noisy depth images, and the inherent complexity of modeling hand motion due to its highly articulated nature are some of these challenges. The current datasets mainly provide front-face hand deformations, which are not appropriate to compare state-of-the-art approaches against hard cases with high occlusions. Furthermore, a little attention has been paid to embed temporal motion information in hand pose recovery problems. We think that we need a system for efficient hand pose recovery in non-controlled settings involving self-occlusions of hands. While the proposed models must be robust against highly-variable hand poses, also they need to be able to recover occluded joints both efficiently and accurately using a powerful refinement approach. Also, using the novel hardware capabilities for input data recording along with the different fusions of input data types and features could improve the hand pose recovery accuracy. In addition, due to having the impressive progress in human pose recovery in recent years, using these complementary features can be more considered to propose a more accurate hand pose recovery model.

5. Sign language recognition using hand and face

Humans mainly look at the face during sign language communication. So, movement of different parts of the face plays a significant role and constitutes natural patterns with large variability. In this section, we review deep sign language recognition models using facial features.

5.1. RGB-based models for sign language recognition using hand and face

Koller et al. suggested a combined model including a CNN and an HMM for weakly supervised learning of mouth shapes in the input frames without explicit frame labels. Experimental results on RWTH-PHOENIX-Weather corpus showed a relative state-of-the-art recognition accuracy improvement of 8% (Koller, Ney et al., 2015). Rao et al. provided a CNN-based model using head and hand movements along with their constantly changing shape features. Different CNN architectures have been applied and selected the best one. They proposed an Indian dataset including sign language videos for 200 signs in 5 different viewing angles under various background environments. Experimental results on this dataset showed a recognition accuracy of 92.88% (Rao et al., 2018).

5.2. Discussion

Facial signs in sign language happen simultaneously with head pose changes and hand signs. So, the proposed models try to learn the shape features of hand, face, and head in order to fuse these features and improve the recognition accuracy. Because of the impressive capabilities of CNN to feature extraction from input images, it has been used in most of the models with the input image. For video input, CNN is not as effective as image input and so it usually is combined with another approach, such as RNN, to more cover the sequence information. In addition, using the 3D shape features of the face and also combining the deep learning approaches with traditional methods have been used in the proposed models.

Due to the fast motion of the head and face in different signs and also face occlusions by hands during the signing, tracking the facial features is challenging. Hence, the communication of the deaf persons usually are considered only as a set of hand movements and do not pay attention to the facial and body features. However, a model has been proposed using the fused features of hand and head (Rao et al., 2018), this model has been limited to the constant background and predefined environmental conditions that cannot be generalized for real-world applications. Furthermore, this model has been evaluated only on the private dataset, not public datasets. We think that fusing the features of hand with the face, head, and body features in an unlimited environment can improve the recognition accuracy. Also, using the other forms of input data, such as flow information, skeleton, thermal, text, and so on, can be more considered in order to benefit from the fusion features of these inputs. Due to the limitation of the large and diverse dataset including both of the hand and face annotations, using the human body pose datasets can be helpful here (Andriluka et al., 2014; Ionescu, Papava, Olaru, & Sminchisescu, 2014; Sapp & Taskar, 2013; Varol, Romero, Martin, Mahmood, Black, Laptev, & Schmid, 2017). Furthermore, using the symmetrical face appearance in order to consider just half of the input face and decrease the parameters and complexity of the model can also be effective for recognition accuracy improvement.

6. Sign language recognition using hand, face, and human body

Human body pose estimation is one of the fundamental tools in the areas related to understanding people behavior, especially action and sign language recognition. Using the body features can improve the recognition accuracy in high occlusions and severe deformation situations. In this section, we review deep human pose estimation models proposed for sing language recognition. Furthermore, we present some of the accurate human pose estimation models in recent four years that could be applied in sign language recognition area.

6.1. RGB-based human pose estimation models

Newell et al. proposed a CNN-based model, namely stacked hourglass, to human pose estimation using different scales of the input image in order to monitor the spatial relationships associated with different parts of the human body. They used the successive steps of pooling and up-sampling to provide the prediction. Evaluation results on FLIC and MPII datasets showed that the model achieved state-of-the-art results for human pose estimation with an average relative improvement of 1.7% and 2.4% (Newell et al., 2016). Wei et al. proposed a CNN-based model, namely Convolutional Pose Machine (CPM), for articulated pose estimation using long-term dependencies among the variables of the model. The model includes a sequential architecture composed of convolutional networks that directly operate on belief maps from previous stages. A refined estimation process has been used for part locations without the need for explicit graphical model-style inference. They achieved state-of-the-art performance on MPII, LSP, and FLIC datasets with a relative accuracy improvement of 9%, 6.11%, and 3%, respectively (Wei et al., 2016). Toshev and Szegedy designed a DNN-based cascade model to estimate the joints by defining a joints regression problem. The proposed model is capable of not only capturing the full context of each body joint but also estimating the location of each joint. Evaluation results on FLIC and LSP datasets showed that the model achieved state-of-the-art results in human pose estimation with an approximately relative accuracy improvement of 17% and 2% (Toshev & Szegedy, 2014). Gattupalli et al. proposed a dataset, namely SLR, and provided a baseline for human pose estimation to evaluate the performance of two deep-based hand pose estimation methods on the proposed dataset. Furthermore, they analyzed the impact of transfer learning to the pose estimation and confirmed the improvement of the estimation accuracy using transfer learning (Gattupalli et al., 2016). Madadi et al. proposed a deep model for human pose recovery using Skinned Multi-Person Linear Model (SMPL) and deep neural networks in a still RGB image. 3D joints, as an intermediate representation, have been used to regress the SMPL parameters. In addition, a denoising autoencoder connects the CNN to SMPLR model for structural errors recovery. Evaluation results on SURREAL and Human3.6M datasets showed a relative improvement over SMPL-based state-of-the-art models about 4 mm and 12 mm (Madadi, Bertiche, & Escalera, 2020). Bin et al. suggested a model, namely Pose Graph Convolutional Network (PGCN), to capture the structural connections between 3D human body keypoints. An attention mechanism is employed to focus on the most crucial structural information and refine both short-range and long-range correlations between 3D human keypoints. Evaluation results on single-person and multi-person estimation datasets confirm the superiority of the model achieving a relative state-of-the-art estimation accuracy improvement of 1.3%, 0.7%, and 3.4% on MPII, LSP, and COCO datasets, respectively (Bin, Chen, Wei, Chen, Gao, & Sang, 2020).

6.2. Depth-based human pose models

Haque et al. provided an end-to-end 3D human pose estimation model using a CNN and recurrent network architecture with a top-down error feedback mechanism to self-correct the previous pose estimations. The model inputs the local regions into a learned viewpoint invariant feature space to selectively estimate partial poses in the presence of noise and occlusion. Evaluation results on own dataset confirm the performance improvement in human pose estimation (Haque et al., 2016). Wang et al. provided a two-stage Depth Ranking Pose 3D estimation (DRPose3D) model using a Pairwise Ranking Convolutional Neural Network (PRCNN) and a 3D Pose Network(DPNet). The model extracts depth rankings of human joints from input images and estimates the 3D poses from both depth rankings and 2D human joint locations. Evaluation results on the Human3.6M benchmark showed that the proposed model outperformed state-of-the-art models for 3D pose estimation area with a relative improvement of 6.2 mm (Wang et al., 2018). MarJimenez et al. provided a deep-based model, namely Deep Depth Pose (DDP), to 3D human pose estimation from depth maps. A depth map including a person and a set of predefined 3D prototype poses are input to the DDP model to estimate the 3D position of the body joints. The proposed model outperformed state-of-the-art models on ITOP and UBC3V datasets with a relative improvement of 11.3% and 13.1% (Marı n Jimeneza et al., 2018).

6.3. Multi-modal human sign language recognition

Huang et al. proposed a deep sign recognition model using 3D CNN from multi-modal inputs. Three input modalities, including RGB, depth, and skeleton data, have been used as multi-channel video streams, to boost the recognition accuracy. They validated the model on own dataset and reported the effectiveness of the model obtaining a recognition accuracy 94.2% (Huang et al., 2015).

6.4. Discussion

Human pose estimation is an important area related to a variety of applications, especially sign language recognition area. A lot of human pose estimation models have been proposed in recent years using deep learning approaches, especially CNN and RNN. Using the 3D information of depth maps has led to significant improvement in this area. The proposed models tried to improve the estimation accuracy of human pose using different approaches such as cascading, tree-structure, 3D estimation, constraint definition, and so on. Although the experimental results of different models approved that the estimation accuracy has more improved in recent years, much more work is necessary to solve the challenges.

Human pose estimation in unconstrained conditions is an intensive research area in computer vision area. Using body features can help to improve the recognition accuracy of human pose estimation but estimation of different joints and parts of the human body remains challenging under partial occlusions and noisy situations. Localization of different body parts and joints is the main core of human pose estimation. While 3D information of the human body has been used by a lot of models in recent years, this information could be challenging where several 3D poses could be projected to the same 2D joints. Furthermore, annotating of 3D joints is very hard and needs the sophisticated tracking devices. Another challenge is regarding 3D pose regression that it is not possible without accurate modeling the correlation of joints. However, some of the proposed models for human pose and shape estimation have considered few constraints including pose angles, shape priors, and 3D joint localization, they are quite accurate in unconstrained and real environments. We think that we can benefit from these models in sign language recognition area to improve the accuracy of the sign models. Furthermore, using the new hardware for 3D joints labeling of the human body is crucial to use the 3D information for more accurate human pose and shape estimation. Also, providing an accurate model for joint correlations and using the other input forms of data, such as thermal or synthetic data, as an auxiliary data, can be helpful for accuracy improvement of human pose estimation. In addition, there are some large human pose datasets including the accurate annotations and the least constraints (Ionescu et al., 2014; Varol et al., 2017) that can be used in sign language recognition area to improve the recognition accuracy and use in real applications.

7. Continuous dynamic sign language recognition

In recent years, with the availability of large datasets, such as RWTHPHOENIX-Weather-2014 (Forster, Schmidt, Koller, Bellgardt, & Ney, 2014), some researchers have been attracted to continuous dynamic sign language recognition (Cihan Camgöz, Hadfield, Koller, & Bowden, 2017; Cui, Liu, & Zhang, 2019; Koller, Zargaran, Ney, &

Bowden, 2016; Mocialov, Turner, Lohan, & Hastie, 2017; Pu, Zhou, & Li, 2018; Wei, Zhou, Pu, & Li, 2019). While isolated sign language recognition uses an image or video including only one sing in the input model, continuous dynamic sign language recognition concerns about multiple signs per video input. One of the main challenges in continuous dynamic sign language recognition is video segmentation to make multiple videos including only one sign per each video segment. In this section, we present the recent works in continuous sign language recognition and related areas. Table 8 shows the details of these models.

7.1. RGB-based continuous dynamic sign language recognition

Pu et al. proposed a CNN-based model for continuous dynamic sign language recognition from RGB video input. They combined 3D residual convolutional network (3D-ResNet) with a stacked dilated CNN and a Connectionist Temporal Classification (CTC) for visual feature extraction and making a mapping between the sequential features and the text sentence. They used an iterative optimization strategy to overcome the problem of less contribution between CTC and CNN parameters. After generating an initial label for a video clip using CTC, they fine-tune CNN to refine the generated label. Evaluation results on RWTH-PHOENIX-Weather dataset demonstrate the superiority of the model obtaining a relative state-of-the-art word error rate improvement of 1.4% (Pu et al., 2018). Mocialov et al. employed the combination of a heuristic approach with the stacked LSTMs for video segmentation using the epenthesis identification and automatic classification of the segmented videos, respectively. They performed an analysis on the sign numbers and efficiency of different features for final recognition. Evaluation results reported only on different sign classes without comparing with state-of-the-art. In the best case, this model achieved a recognition accuracy of 95% on the NGT dataset, including 40 sign classes (Mocialov et al., 2017). Wei et al. defined the continuous sign language recognition as a grammatical-rule-based classification problem using the combination of 3D convolutional residual network and bidirectional LSTM. They used two modules, word-independent classifiers (WIC) module and n-gram classifier (NGC) module, to split a sentence into a sequence of consecutive words. The confidence scores provided by these modules are used to concatenate the features of the words in a sentence. Evaluation results on a CSL SPLIT I dataset demonstrated a relative state-of-the-art precision improvement of 2% (Wei et al., 2019).

7.2. Depth-based continuous dynamic sign language recognition

Camgoz et al. presented a deep-based and end-to-end framework for continuous dynamic sign language recognition. They employed the SubUNets approach to improve the learning procedure of the intermediate representations. This model uses the multi-channel information such as hand shape, motions, body pose and facial gestures, from input data to generate a sequence of outputs from a given video. Evaluation results on One-Million Hands demonstrated that this model obtained a comparable sign recognition rate to previous research works on this dataset achieving a word error rate of 40.7 (Cihan Camgöz et al., 2017).

7.3. Multi-modal continuous dynamic sign language recognition

Cui et al. developed a continuous sign language recognition framework using the combination of CNN and Bi-LSTM. They used an iterative optimization process to obtain the representative features from CNN. Model performance is iteratively improved using the training and tuning procedures of the recognition model. Besides, this model benefits from the multi-modal fusion of RGB images and OF information. Experimental results on two public benchmarks, RWTH-PHOENIX-Weather 2014 and SIGNUM, confirm outperforming of state-of-the-art by a relative improvement of more than 15% (Cui et al., 2019).

7.4. RGB-based continuous dynamic Human pose estimation

Ye et al. suggested a 3D Recurrent Convolutional Neural Network (3DRCNN) for gesture recognition and joint localization using their temporal boundaries within continuous videos and fusing multi-modal features. Dividing the original videos into some short video clips, the extracted features of the relations among these video clips have considered as the temporal information. Sliding window approach has been used to merge the temporal information of the Consecutive clips with the same semantic meaning. To evaluate the method, they proposed a dataset including some words and sentences videos and reported the effectiveness of the proposed model on this dataset achieving an estimation accuracy of 69.2% (Ye et al., 2018).

7.5. Discussion

Sign language recognition includes two main categories, which are isolated sign language recognition and continuous sign language recognition. The supervision information is a key difference between the two categories. While isolated sign language recognition is similar to the action recognition area, the continuous sign language recognition concerns about not only the recognition task but also the accurate alignment between the input video segments and the corresponding sentence-level labels. Generally, continuous sign language recognition is more challenging than isolated sign language recognition. Indeed, isolated sign language recognition can be considered as a subset of continuous sign language recognition. Two factors play a key role in the performance evaluation of continuous sign language recognition, which include feature extraction from frame sequences of the input video and alignment between the features of each video segment and the corresponding sign label. Obtaining more descriptive and discriminative features from the video frames could result in a better performance for a continuous sign language recognition system. While recent models in continuous sign language recognition have a rising trend in model performance relying on deep learning capabilities in computer vision and NLP, there is still much room for performance improvement in this area. Considering the attention mechanism, using multiple input modalities to benefit from multi-channel information, learning structured spatio-temporal patterns (such as Graph Neural Networks models), and employing the prior knowledge on sign language are only some of the possible future directions in this area.

8. Hybrid models using deep and traditional methods

In this section, we present the recent works in sign language recognition and related areas that use the combination of deep-based descriptors and classic classifiers for training. Table 9 shows the details of these models. While some works use traditional descriptors and classic classifiers for training (Cheron, Laptev, & Schmid, 2015; Escobedo-Cardenas & Camara-Chavez, 2015), these works are not under the scope of this survey.

8.1. RGB-based Hybrid hand sign language recognition

Rastgoo et al. proposed a deep-based model to hand sign language recognition using SSD, CNN, LSTM benefiting from hand pose features. They improved hand detection accuracy of SSD model using five online sign dictionaries. Furthermore, they employed some hand-crafted features and combined with the extracted features from CNN model. Evaluation results on RKS-PERSIANSIGN and isoGD datasets confirm the superiority of the proposed model achieving state-of-the-art on isoGD dataset with 4.25% relative margin (Rastgoo et al., 2020b). Koller et al. used the embedding of CNN into a HMM for continuous sign language recognition. They used the CNN outputs as true Bayesian posteriors and train the model end-to-end as a hybrid CNN-HMM. Evaluation results on RWTHPHOENIX-Weather 2014 Multi-signer dataset showed that this model decreased the error rates from 51.6/50.2 to 38.3/38.8 on dev/test in comparison with state-of-the-art models (Koller, Zargaran et al., 2016).

Table 8
Deep continuous dynamic sign language recognition models; CDSLR: Continuous Dynamic Sign Language Recognition, CDHPE: Continuous Dynamic Human Pose Estimation.

Modality	Year	Ref.	Goal	Model	Dataset	Results
RGB video	2018	(Pu et al., 2018)	CDSLR	3DCNN	RWTH-PHOENIX-Weather 2014	37.3 (error rate)
RGB video	2017	(Mocialov et al., 2017)	CDSLR	heuristic approach and LSTM	NGT	80.70 (accuracy)
RGB video	2019	(Wei et al., 2019)	CDSLR	3DCNN and Bi-LSTM	Chinese dataset	94.90
Depth video	2017	(Cihan Camgöz et al., 2017)	CDSLR	CNN	One-Million Hands	40.8 (error rate)
RGB video, OF	2019	(Cui et al., 2019)	CDSLR	3DCNN, Bi-LSTM	RWTH-PHOENIX-Weather 2014, SIGNUM	22.86, 2.80 (error rate)
RGB video	2018	(Ye et al., 2018)	CDHPE	3DRCNN	own dataset	69.2

Table 9
Hybrid sign language recognition models; SLR: Sign Language Recognition, GR: Gesture Recognition, PE: Pose Estimation.

Modality	Year	Ref.	Goal	Model	Dataset	Results
RGB video	2020	(Rastgoo et al., 2020b)	SLR	SSD, CNN, LSTM, hand-crafted features	RKS-PERSIANSIGN, isoGD	98.42, 86.32 accuracy
RGB video	2016	(Koller, Zargaran et al., 2016)	SLR	CNN, HMM	RWTHPHOENIX-Weather 2014 Multi-signer dataset	38.3, 38.8 (error rate)
RGB image	2018	(Chen, Ting, Wu, & Fu, 2018)	GR	CNN, SVM	own dataset	49.88
RGB video	2020	(Escobedo-Cardenas & Camara-Chavez, 2020)	GR	CNN, HCM	UTD-MHAD, isoGD, UFOP-LIBRAS	94.81, 67.36, 64.33 (accuracy)
RGB, Depth image	2016	(Ma1, Chen, & Wu, 2016)	GR	CNN, SVM	ASL	96.1 (accuracy)
Depth image	2018	(Chen, Ting, Wu, & Fu, 2018)	PE	CNN, SPM	ICVL, NYU, NTU	8.64, 15.90, 12.81 (error)

8.2. RGB-based Hybrid gesture recognition

Chen et al. proposed a hybrid model to hand gesture recognition, including a CNN for automatic feature extraction and SVM for final classification, from the input of raw EMG image. Experimental results on own dataset confirmed the higher accuracy with 2.5% and 9.7% margins in comparison with the cases that only CNN or traditional methods was used (Chen, Tong et al., 2018).

8.3. Multi-modal hybrid gesture recognition

Cardenas and Chavez proposed a hybrid model to hand gesture recognition using the combination of CNN and Histogram of Cumulative Magnitudes (HCM). They used three input modalities including RGB, Depth, and Skeleton. A skeleton estimation method along with a sampling method is employed to include a fixed number of keyframes from the input video. They fuse the extracted spatio-temporal features and fed them into a linear Support Vector Machine (SVM) classifier for final recognition. Evaluation results on UTD-MHAD, ChaLearn LAP IsoGD, and UFOP-LIBRAS, confirmed the effectiveness of this model achieving the accuracy of 94.81%, 67.36%, and 64.33%, respectively. While this model performed comparably with state-of-the-art methods on isoGD and UFOP-LIBRAS datasets, it outperformed state-of-the-art on UTD-MHAD with relative improvement of 0.16% (Escobedo-Cardenas & Camara-Chavez, 2020). Ma et al. deployed a CNN-based model for hand gesture recognition from two modalities, RGB and Depth images. The gesture region is extracted using a depth image-based segmentation method. After feature extraction using a CNN, final recognition is done using the SVM method. Experimental results on own dataset confirm that the proposed model, benefiting from the combination of CNN and SVM, achieved a recognition accuracy of 96.1% (Ma1 et al., 2016).

8.4. Depth-based hybrid hand pose estimation

Chen et al. proposed a vision-based framework for 3D hand pose estimation using the combination of a Spherical Part Model (SPM) and a deep CNN. In this framework, prior knowledge of the human hand is used for accurately hand pose estimation from a depth map. Using the hand-centric coordinate system, SPM employed to obtain skeletal configurations from the most stable joints and use spherical representation. Results on NYU and NTU datasets demonstrated a relative estimation improvement of 0.063 mm and 3.358 mm in comparison with state-of-the-art methods (Chen, Ting et al., 2018).

8.5. Discussion

Recently, deep learning-based models attracted more attention in the research community in comparison with the traditional computer vision techniques. However, this is not mean that traditional computer vision techniques have got obsolete. Some problems may benefit from having a trade off between both the powerful capabilities of deep learning (in particular in those cases of having large amounts of data) and the specific problem-tailored design of handcrafted features. To this end, using the capabilities of both categories, as a hybrid model, has considered by some researchers in recent years (Chen, Tong et al., 2018, 2018; Escobedo-Cardenas & Camara-Chavez, 2020; Koller, Zargaran et al., 2016; Ma1 et al., 2016; Rastgoo et al., 2020b). A combination of a deep-based and automatic feature extractor, such as CNN, with a classical classifier method, such as Support Vector Machine (SVM), has widely used in the hybrid models. Furthermore, employing some hand-crafted features along with the CNN-based features is another approach in the hybrid models (Chen, Ting et al., 2018; Rastgoo et al., 2020b). Combining prior knowledge with deep-based features can help to develop systems with lower complexity and maybe more accurate in some special domains.

9. Discussion and conclusion

Sign language and different forms of sign-based communication are prominent to large groups in society. With the advent of deep learning approaches, sign language recognition area have faced up a significant accuracy improvement in recent years. In this survey, we reviewed the proposed models of sign language recognition area using deep learning in recent four years based on a proposed taxonomy. Many models have been proposed by researchers in recent years. Most of the models have used the CNN model for feature extraction from input image due to the impressive capabilities of CNN for this goal. In the case of video input, RNN, LSTM, and GRU have been used in most of the models to cover the sequence information. Also, some models have combined two or more approaches in order to boost the recognition accuracy. Moreover, different types of input data, such as RGB, depth, thermal, skeleton, flow information have been used in the models. Tables 4–9 show the proposed models details for sign language recognition in recent four years. Furthermore, the pros and cons of these models are presented in the Table 10. Here, we present the challenges based on our taxonomy:

- **Feature fusion**: While many models have been proposed by different researchers, there are still several challenges in this area that need to be solved. Feature fusion can be applied in input data

Table 10
Summary of deep sign language recognition models.

Year	Ref.	Feature fusion	Input modality	Dataset	Description.
2014	(Neverova et al., 2014)	Hand	static, Depth	own dataset	**Pros**: This model significantly improved hand gesture recognition accuracy using unlabeled real-world samples. **Cons**: Need to adapt with real data.
2014	(Tompson et al., 2014)	Hand	static, Depth	own dataset	**Pros**: Acceptable generalization performance in coping with hand shape changes. **Cons**: The recognition accuracy of the hand pose model has not been evaluated. Furthermore, the model is not robust against hand occlusion.
2014	(Toshev & Szegedy, 2014)	Hand	static, RGB	FLIC, LSP	**Pros**: High precision pose estimation using a simple but yet powerful formulation. **Cons**: Need to propose a new architecture which could be potentially better tailored towards localization problems and especially in pose estimation in particular because they used a generic model which was originally designed for classification tasks and applied it for hand joints localization.
2015	(Kang et al., 2015)	Hand	static, Depth	own dataset	**Pros**: High recognition performance for observed signers in real-time. **Cons**: Need to include more data from different subjects to improve the results.
2015	(Oberweger et al., 2015)	Hand	static, Depth	ICVL, NYU	**Pros**: Accurate and fast. **Cons**: Location prediction of the joint is constrained by the learned hand model.
2015	(Tang et al., 2015)	Hand	static, Depth,	own dataset	**Pros**: Robust to occlusion and RGB insensitive to movement, scaling and rotation. **Cons**: training process of DBN is difficult to parallelize.
2015	(Tagliasacchi et al., 2015)	Hand	dynamic, Depth	real video	**Pros**: Novel, robust and accurate hand tracking algorithm. **Cons**: This model track only hand and suffers from low accuracy in two-hand tracking.
2015	(Koller, Ney et al., 2015)	Face	dynamic, RGB	RWTH-PHOENIX	**Pros**: Accurate and robust modeling of mouth shapes. **Cons**: Need to be more generalize with more data because the recognition accuracy of the model depends on the shuffling of the training samples.
2015	(Huang et al., 2015)	Body	dynamic, RGB,	Own dataset	**Pros**: Boosting the recognition accuracy of the model using multi-channels of video streams. **Cons**: The evaluation results of the model highly depends on own dataset. Need to be more general by evaluating on some public datasets.
2016	(Oberweger et al., 2016)	HP	Depth, Skeleton dynamic, Depth	MSRA	**Pros**: Accurate 3D hand pose annotation and estimation. **Cons**: Estimation accuracy of the model highly depends on the annotation accuracy of input data. The model is semi-automated and need some human users to annotate the visible joints in frames. In addition, the model is highly complex.
2016	(Han et al., 2016)	Hand	static, RGB	own dataset	**Pros**: Some simple pre-processing methods applied on input data that improved the recognition performance of the model. **Cons**: Need to increase the gesture labels and input data for more generalization of the model.
2016	(Duan et al., 2016)	Hand	static, Depth, RGB	Chalearn IsoGD, RGBD-HuDaAct	**Pros**: Using the spatial and temporal information complementary to improve the recognition accuracy and reduce the estimation variance. **Cons**: While two input modality are used in the model, the temporal information of only one modality, RGB, are employed.
2016	(Newell et al., 2016)	Hand	static, RGB	FLIC, MPII	**Pros**: Accurate and robust against heavy occlusion and multiple people in close proximity. **Cons**: No robust to some complicated poses.
2016	(Wei et al., 2016)	Human	static, RGB	MPII, LSP, FLIC	**Pros**: Robust to create an effective communication between joints to accurately pose estimation. **Cons**: The model is not accurate in handling multiple people in close proximity.
2016	(Haque et al., 2016)	Hand	static, Depth	EVAL, ITOP	**Pros**: Accurate pose estimation on alternate viewpoints and partially robust to noise and occlusion. **Cons**: Not accurate in joints localization.
2016	(Molchanov et al., 2016)	Hand	dynamic, RGB, Depth, OF, stereo-IR sensors	SKIG, ChaLearn	**Pros**: Accuracy improvement of hand gesture model using effective modality fusion. **Cons**: Temporal information between all clips of each video can be used in an efficient way to improve hand gesture recognition accuracy.
2016	(Wu et al., 2016)	Hand	dynamic, Depth, OF	MSR Action 3D	**Pros**: Proposing an accurate two-stream CNN model for gesture verification and identification. **Cons**: Generalization of the model is poor for unseen gestures.
2017	(Deng et al., 2017)	Hand	static, Depth	ICVL, NYU	**Pros**: Integrating both local 3D feature and global context without any further post-processing. **Cons**: Estimation accuracy of hand pose model highly depends on data augmentation.

(*continued on next page*)

Table 10 (*continued*).

Year	Ref.	Feature fusion	Input modality	Dataset	Description.
2017	(Guo et al., 2017)	Hand	static, Depth	ICVL, NYU, MSRA, ITOP	**Pros**: Simple but accurate and fast model. **Cons**: No robust to occlusion.
2017	(Simon et al., 2017)	Hand	static, RGB	own dataset	**Pros**: Robust to accurately generate the annotations for keypoint detection. **Cons**: Model accuracy highly depends on using multiple cameras in controlled environments.
2017	(Zimmermann & Brox, 2017)	Hand	static, RGB	Stereo Hand Pose Tracking, Dexter	**Pros**: Approximately accurate to predict 3D hand poses from 2D keypoints. **Cons**: The performance seems mostly limited by the lack of an annotated large scale dataset with real-world images and diverse pose statistics.
2017	(Fang & Lei, 2017)	Hand	dynamic, Depth	NYU, ICVL	**Pros**: Accurate estimation of hand pose by exploiting dependencies between hand joints. **Cons**: No robust to occlusion and noisy inputs.
2017	(Yuan et al., 2017)	Hand	Static, Depth	ICVL, NYU, MSRC, BigHand	**Pros**: A suitable evaluation cross-benchmark for different models. **Cons**: Proposed tracking system highly depends on 6D magnetic sensors and inverse kinematics to obtain hand joints annotations.
2017	(Wang et al., 2017)	Hand	dynamic, Depth	MSRC, BigHand ChaLearn	**Pros**: Accuracy improvement of gesture recognition by using different presentations of input images. **Cons**: Recognition accuracy can be improved using more efficient way for features fusion in the model.
2017	(Madadi et al., 2017)	Hand	static, Depth	NYU, MSRA	**Pros**: High accurate capability to local pose recovery. **Cons**: Need to consider data complexity reduction of the model.
2017	(Dibra et al., 2017)	Hand	static, Depth	NYU, ICVL	**Pros**: Accurately estimation of 3D hand pose with the ability to refine on unlabeled depth images. **Cons**: Model has adopted to a single hand shape only.
2018	(Chen et al., 2020)	Hand	static, Depth	ICVL, NYU, MSRA	**Pros**: Accurately estimation of 3D hand pose. **Cons**: No robust and accurate when hands are interacting with other hands or objects.
2018	(Rao et al., 2018)	Face	dynamic, RGB	own dataset	**Pros**: Approximately accurate and robust against 5 various orientations. **Cons**: Recognition accuracy of the model can be improved using accurate and pre-trained CNN model instead of a shallow CNN.
2018	(Ye et al., 2018)	Face	dynamic, RGB	own dataset	**Pros**: Capability to learn the complementary information from multi-modal inputs via different fusions. **Cons**: Poor performance for sings including facial information.
2018	(Wang et al., 2018)	Hand	static, Depth	Human3.6M	**Pros**: Accurate capturing context and reasoning about pose in a holistic manner. **Cons**: Not accurate in joints localization.
2018	(Marı n Jimeneza et al., 2018)	Hand	static, Depth	UBC3V, ITOP	**Pros**: Accurate and robust to different viewpoints. **Cons**: Constrained on some especial types of poses.
2018	(Rastgoo et al., 2018)	Hand	static, RGB, Depth	Massey2012, ASL Surrey, NYU,ASL Fingerspelling A	**Pros**: Robust to noise and accurate by providing a generalization in instances of low amounts of annotated data. **Cons**: Need to decrease the complexity of the model by sharing the parameters.
2018	(Dadashzadeh et al., 2018)	Hand	static, Depth	OUHANDS	**Pros**: Accurate pixel-level semantic segmentation into hand region. **Cons**: Not accurate in recognition stage.
2018	(Devineau et al., 2018)	Hand	dynamic, Skeleton	DHG	**Pros**: Efficient recognition performance in time and accuracy. **Cons**: The model only works on complete sequences of input data.
2018	(Moon et al., 2018)	Hand	static, Depth	ICVL, MSRA, NYU	**Pros**: Estimation accuracy improvement of the model by converting 2D depth map into the 3D voxel representation. **Cons**: Model complexity is high due to doubling the number of channels of each feature map.
2019	(Chen et al., 2019)	Hand	dynamic, RGB	DHG-14/28, SHREC'17	**Pros**: They developed a general framework that can be used for other tasks aiming to learn spatial and temporal information from graph-based data. **Cons**: Need to generalize to additional datasets.
2019	(Gomez-Donoso et al., 2019)	Hand	static, RGB	own dataset, Stereo Hand Pose Tracking	**Pros**: Accurately prediction of 3D positions of hand joints. **Cons**: Suffering from a minor jittering when the results are rendered over time.
2019	(Lim et al., 2019)	Hand	dynamic, RGB	ASLLVD, RWTH-Boston-50	**Pros**: Accurate and compact sign language hand representation with a good discriminating power. **Cons**: No robust against different skin colors and hands occlusions.
2019	(Li et al., 2019)	Hand	static, RGB	STB, RSTB	**Pros**: Estimation accuracy improvement of 3D hand pose benefiting from stereo cameras capabilities. **Cons**: No robust to multiple hands or cases with hand/object interaction.

(*continued on next page*)

Table 10 (*continued*).

Year	Ref.	Feature fusion	Input modality	Dataset	Description.
2020	(Wadhawan & Kumar, 2020)	Hand	static, RGB	own dataset	**Pros**: Extensive evaluation results on 50 deep learning models using different optimizers. **Cons**: Need to fine-tune the recognition method using more real data.
2020	(Elboushaki et al., 2020)	Hand	dynamic, RGB, Depth	isoGD, SKIG, NATOPS, SBU	**Pros**: Recognition accuracy improvement by capturing the fine-grained motion details encoded in multiple adjacent frames of input video. **Cons**: Human gestures are highly related to different modalities.

Table 11
State-of-the-art models on the datasets corresponding to the sign language and related areas.

Dataset	Year	Ref.	Goal	Model	Modality	Results
NYU	2020	(Rastgoo et al., 2020a)	HSR	SSD, 2DCNN, 3DCNN, LSTM	Depth	4.64 mm
ICVL	2018	(Moon et al., 2018)	HP	CNN	Depth	6.28 mm
MSRA	2016	(Oberweger et al., 2016)	HP	CNN	Depth	5.58 mm (Ave. err.)
FLIC	2016	(Newell et al., 2016)	HP	CNN	RGB	99.0 (Elbow)
LSP	2016	(Wei et al., 2016)	HP	CNN	RGB	84.32
isoGD	2020	(Rastgoo et al., 2020b)	HSR	SSD, CNN, LSTM	RGB	86.32
MPII	2016	(Newell et al., 2016)	HP	CNN	RGB	90.90 (total)
ITOP	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	Depth	97.5 (AUC)
RGBD-HuDaAct	2016	(Duan et al., 2016)	HG	CNN	Depth, RGB	96.74
STB	2018	(Spurr et al., 2018)	HP	VAE	RGB, Depth	0.983(AUC)
EVAL	2016	(Haque et al., 2016)	HP	CNN	Depth	74.10
Dexter	2017	(Zimmermann & Brox, 2017)	HP	CNN	RGB	49.0 (AUC)
RWTH-PHOENIX-Weather 2012	2015	(Koller, Ney et al., 2015)	HSR	CNN	RGB	55.70 (Precision)
RWTH-PHOENIX-Weather 2014	2019	(Cui et al., 2019)	CDSLR	3DCNN, Bi-LSTM	RGB	22.86
BigHand2.2M	2018	(Baek et al., 2018)	HP	GAN	Depth	13.7 mm
Human3.6M	2018	(Wang et al., 2018)	HP	CNN	Depth	62.8 mm
NGT	2017	(Mocialov et al., 2017)	CDSLR	heuristic, LSTM	RGB	80.70 (accuracy)
UBC3V	2018	(Marı n Jimeneza et al., 2018)	HP	CNN	Depth	88.2 (AUC)
Massey 2012	2018	(Rastgoo et al., 2018)	HR	RBM	RGB, Depth	99.31
SL Surrey	2018	(Rastgoo et al., 2018)	HR	RBM	RGB, Depth	97.56
ASL Fingerspelling A	2018	(Rastgoo et al., 2018)	HR	RBM	RGB, Depth	98.13
OUHANDS,	2018	(Dadashzadeh et al., 2018)	HG	CNN	Depth	86.46
Egohands	2017	(Dibia, 2017)	HT	CNN	RGB	0.9686 (mAP)
Dexter	2018	(Mueller et al., 2018)	HT	CNN	RGB	0.64 (AUC)
EgoDexter	2018	(Mueller et al., 2018)	HT	CNN	RGB	0.54 (AUC)
RHD	2018	(Spurr et al., 2018)	HP	VAE	RGB, Depth	0.849(AUC)
B2RGB-SH	2019	(Li et al., 2019)	HP	CNN	RGB	7.18 (err)
DHG-14/28 Dataset	2019	(Chen et al., 2019)	HG	CNN	RGB	91.9
SHREC'17 Track Dataset	2019	(Chen et al., 2019)	HG	CNN	RGB	94.4
RWTH-BOSTON-50	2019	(Lim et al., 2019)	HS	CNN	RGB	89.33
ASLLVD	2019	(Lim et al., 2019)	HS	CNN	RGB	31.50
EgoGesture	2019	(Kopuklu et al., 2019)	HG	CNN	RGB	94.03
NVIDIA benchmarks	2019	(Kopuklu et al., 2019)	HG	CNN	RGB	83.83
SKIG	2020	(Elboushaki et al., 2020)	HG	CNN	RGB, Depth	99.72
NATOPS	2020	(Elboushaki et al., 2020)	HG	CNN	RGB, Depth	95.87
SBU	2020	(Elboushaki et al., 2020)	HG	CNN	RGB, Depth	97.51
First-Person	2020	(Rastgoo et al., 2020a)	HSR	SSD, 2DCNN, 3DCNN, LSTM	RGB	91.12
RKS-PERSIANSIGN	2020	(Rastgoo et al., 2020a)	HSR	SSD, 2DCNN, 3DCNN, LSTM	RGB	99.80

Table 12
A summary of the main characteristics of the reviewed models.

Query	Available choices	Most used
Sign languages	USA, Germany, Greek, Poland, China, Argentina, Korea, Iran	USA
Modalities	RGB, Depth, Skeleton Static (Image), Dynamic (Video)	Depth Static (Image)
Architectures	Static: CNN, RBM, GAN, AE Dynamic: RNN, LSTM, GRU, 3DCNN	CNN LSTM
Datasets with lowest performance	RWTH-PHOENIX-Weather, Human3.6M, isoGD	isoGD
Datasets with highest performance	FLIC, SKIG, Massey 2012, RKS-PERSIANSIGN	FLIC
Generative models	RBM, AE, VAE, GAN	VAE
Traditional classifiers	SVM, SPM, HCM	SVM
Traditional descriptors	Heuristics, HOG, HMM	HMM
Recognition modalities	Isolated, Continuous	Isolated
Evaluation metrics	Accuracy, Error rate, Precision, Recall, mAP, AUC	Accuracy
Features fusion	Hand, Face, Body	Hand
Feature types	HP, HG, HSR, HD, HT	HP
Input dimensions	2D, 3D	2D

or human body parts features. In input data fusion, different types of input data, such as RGB, depth, skeleton, flow information, text, and synthetic data, can be fused to have much more powerful features and improve the recognition accuracy. In the fusion of human body parts features, there are three parts, including hand, face, and body, that their features can be fused. However, there are many challenges to use the hand features for hand sign language recognition area, most of the models have relied on the hand features and tried to improve the sign language recognition accuracy just using the hand features. Hand sign language recognition area, including hand detection, hand pose estimation, hand gesture recognition, real-time hand tracking, and hand pose recovery, cops with a lot of challenges such as high variations of the hand shapes and gestures, self-occlusion (between fingers), close similarity between fingers, low resolution, varying illumination conditions, different hand gestures, and complex interactions between hands and objects or other hands. So, some of the models fused the hand features with the face features to decrease the effect of these challenges and improve the recognition accuracy. However, the facial features tracking is also challenging due to the fast motion of the head and face in different signs and also face occlusions by hands during the signing. In this regards, the human body features are also used as the complementary features along with the hand and face features to boost the recognition accuracy. So, sing language recognition area can benefit from the accurate models for human pose estimation that we reviewed some of them in this survey to improve the recognition accuracy.

- **Input modality**: While most of the models have used the advantages of depth modality, the other models have benefited from the high-resolution pixel information of the RGB modality for sign language recognition. Furthermore, some of the models have utilized the flow information, in the forms of OF and SF, skeleton modality, or synthetic data. While the thermal modality is not as common as RGB, depth, and the other modalities, using this modality with another familiar modality can be more considered to improve the feature quality. Regarding different types of signs, static, dynamic, and continuous dynamic, we think that the research community will move into designing for useful in practice sign language recognition models concentrating on learning unsegmented signs of long-term video streams. Current progress in deep learning for video understanding shows the feasibility of moving into this direction.
- **Datasets and different sign languages**: There are many datasets, with different data modalities and languages, for hand, face, and human body sign language recognition. While hand sing recognition area suffers from the inexistence of a large, diverse, and realistic dataset including accurate annotated data in the unconstrained environment, there are some accurate annotated datasets for human pose estimation, with the capability of applying in real-world applications, which could be used in sign language recognition area. Furthermore, we need the models with long untrimmed videos including some sign sentences in real communication, not just one sign, word, letter, or action in a video to have a more realistic model in order to apply in real-world communications. In other words, the sign language recognition systems, learned on real input data, have to be used in real communications with an unlimited environment. This is so complicated but we think that the research community will try to do it in the future especially based on deep learning approaches.
- **Task complexity**: Different deep-based models with different levels of complexity have been developed in recent years. Using various representations of input data along with different input modalities is led to propose different models with different levels of complexity. While static signs use the image modality in the model, dynamic and continuous dynamic signs tackle with the video input challenges. Furthermore, continuous dynamic signs, as the most complex signs, face a challenge in movement epenthesis to manage the transitions between signs in the input video. All of these leads to this fact that while there are many research works in static and dynamic sign language recognition (Deng et al., 2017; Gomez-Donoso et al., 2019; Guo et al., 2017; Li et al., 2019; Oberweger et al., 2015; Spurr et al., 2018; Zimmermann & Brox, 2017), few works have been developed for continuous dynamic sign language (Cihan Camgöz et al., 2017; Cui et al., 2019; Koller, Zargaran et al., 2016; Mocialov et al., 2017; Pu et al., 2018; Wei et al., 2019). Some of these models may require mechanisms to model spatio-temporal modeling of structures and patterns that localize signs, gestures, and poses. While these mechanisms increase the model complexity, they help to improve the model performance by providing discriminative features. Benefiting from deep learning capabilities for parallel computation, using more accurate fusion methods to integrate multiple input modalities, especially in continuous dynamic sign language, and employing the combination of fast traditional methods with deep-based models can help to decrease the complexity of the models in sign language recognition and related areas.
- **Applications**: In respect of applications, deep learning approaches have been successfully applied in many areas related to sign language recognition area such as machine translation, voice assistant, text assistant, and so on (Deng et al., 2017; John, Boyali, Mita, Imanishi, & Sanma, 2016; Mittal, 2018; Supancic et al., 2018). We expect that the sign language application area will be extended in future not only for deaf and speaking disable people but also for the other people of the society that rely on signing as a complementary language to verbal communication in daily interactions.
- **Hybrid models**: While there is a rising trend in employing deep learning-based models in the research community, this is not mean that traditional computer vision techniques have got obsolete. Some problems may benefit from having a trade off between both the powerful capabilities of deep learning (in particular in those cases of having large amounts of data) and the specific problem-tailored design of handcrafted features. This combination can help to develop more accurate systems in some special domains.
- **Other challenges**: One of the most important challenges in sign language recognition area is occlusion. Since each sign may consist of the whole or part of the hand, face, and body movements, occlusion of these parts during the signing can be led to the more complicated situations. Another important challenge is in real-time sign language recognition. We think that the research community will pay much more attention to deep learning models for real-time sign language recognition in the future. To have the more sophisticated and realistic models for applying in deaf and speaking community applications, we need to have a real-time translation system to connect these people to other people of the community. Some efforts have been done in this area and some models have been suggested but much more improvement is indispensable. Another challenge is multi-person sign language recognition that could be more considered in future models. Most of the proposed models just considered the individual sign for only one person that nevertheless takes into account the sign of other persons.

Finally, we presented an aggregated information about all of the works presented in this survey. Table 11 shows state-of-the-art models on different datasets in sign language recognition and related areas. As this table shows, trends of the proposed models on different datasets in sign language and related areas show that deep learning approaches successfully improved the model performance with a high margin. However, more endeavor is necessary for some challenging datasets such as isoGD, LSP, and EVAL. In most of the existing datasets, such

as NYU, ICVL, MSRA, ASL Fingerspelling A, RKS-PERSIANSIGN, the achieved performance by deep-based models are higher than the other challenging datasets. The proposed experimental results of different deep-based models confirm the effective role of using multi-modal and multi-channel information (Duan et al., 2016; Elboushaki et al., 2020; Rastgoo et al., 2018; Spurr et al., 2018). Furthermore, the proposed hybrid models successfully improved the model performance benefiting from the combination of some hand-crafted features with deep-based features (Chen et al., 2020; Escobedo-Cardenas & Camara-Chavez, 2020; Ma1 et al., 2016; Rastgoo et al., 2020b). These models benefit from having a trade off between both the powerful capabilities of deep learning (in particular in those cases of having large amounts of data) and the specific problem-tailored design of handcrafted features. Due to the undeniable power of CNN models for feature extraction from visual inputs, in most of the proposed deep-base models, CNN or a combination of CNN with other deep-based models is employed. Generative models, such as RBM and VAE, showed a comparable or better performance than other deep alternatives in coping with few data for sign language recognition (Rastgoo et al., 2018; Spurr et al., 2018). Since the dynamic modality is more challenging than the static one, most of the proposed models employed LSTM or 3DCNN for analyzing temporal dynamics. Table 12 shows a summary of the main characteristics relevant to sign language recognition regarding the reviewed models.

Next, we describe possible limitations of this survey and discuss some future recommendations for advancing the research in the field.

- **Limitations**: In this survey, we reviewed the vision-based proposed models of sign language recognition and related areas using deep learning approaches from the last five years. The main goal of this survey is to compactly summarize the vision-based sign language recognition models corresponding to the achieved results. This can facilitate the research way for other researchers in the field to access the latest developments, advantages, limitations, and future directions in sign language recognition. While we present the hybrid models in sign language recognition, we did not include the sensor-based models nor the traditional-based models. There are many sensor-based modalities considered for sign language recognition. Furthermore, other modalities achieved from other data collection devices can be also considered for possible usage. We presented a brief review in terms of the application domain of sign language recognition. Due to the importance of this domain in deaf and speak-impaired community, more details of this domain must be studied to open new windows for proposing some applications compatible with real world conditions. This may include considering in more detail the real needs in practice of this technology and their associated usability, privacy, generalization to different populations, and ethical dimensions.
- **Future directions**: While many models have been proposed for sign language recognition, more effort is required to provide more accurate and useful in practice models. We envision a multi-modal integration from the point of view of face, body, and hand visual cues with significantly enhance recognition performance of current models, providing the fine grain recognition analysis required in practice. We foresee that most of the challenges in sign language recognition area will be solved under the support of deep learning, faster hardware to process the input data, accurate multi-modal approaches, and new data covering the real variability and distribution of the problem at hand. While most of the presented models are in the scope of isolated sign language recognition, we expect the community to move in a near future into the direction of addressing the challenges of continuous sign language recognition, including continuous annotated datasets, tokenization, and long term multi-modal modeling of data, specially benefiting from the integration of vision and language models.

CRediT authorship contribution statement

Razieh Rastgoo: Methodology, Software, Validation, Data curation, Writing - original draft, Visualization. **Kourosh Kiani:** Conceptualization, Data curation, Writing - review & editing, Supervision, Project administration. **Sergio Escalera:** Conceptualization, Writing - review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya, ICREA under the ICREA Academia programme, and High Intelligent Solution (HIS) company in Iran.

Funding

This research received no external funding.

References

Acton, B., & Koum, J. (2009). WhatsApp. *Yahoo*, www.whatsapp.com.

Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G., Zacharopoulou, V., Xydopoulos, G., Atzakas, K., Papazachariou, D., & Daras, P. (2019). A comprehensive study on sign language recognition methods. *IEEE Transactions on Multimedia*.

Andriluka, M., Pishchulin, L., Gehler, P., & Bernt, S. (2014). 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*. Columbus, Ohio.

Asadi-Aghbolaghi, M., Clapés, A., Bellantonio, M., Jair Escalante, H., Ponce-López, V., Baró, X., Guyon, I., Kasaei, S., & Escalera, S. (2017). Deep learning for action and gesture recognition in image sequences: A survey. {*G*}*esture* {*R*}*ecognition*, 539–578.

Baek, S., Kim, K., & Kim, T.-K. (2018). Augmented skeleton space transfer for depth-based hand pose estimation. In *CVPR* (pp. 8330–8339). Salt Lake City, Utah, United States.

Bambach, S., Lee, S., Crandall, D., & Yu, C. (2015). Lending A hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*. Las Condes, Chile.

Baró, X., Gonzàlez, J., Fabian, J., Bautista, M., Oliu, M., Escalante, H., Guyon, I., & Escalera, S. (2015). ChaLearn Looking at People 2015 challenges: action spotting and cultural event recognition. In *CVPR 2015*. Boston, Massachusetts.

Barsoum, E. (2016). Articulated hand pose estimation review. arXiv:1604.06195.

Bin, Y., Chen, Z., Wei, X., Chen, X., Gao, C., & Sang, N. (2020). Structure-aware human pose estimation with graph convolutional networks. *Pattern Recognition*, *106*, Article 107411.

Camgoz, N., Hadfield, S., Koller, S., Ney, H., & Bowden, R. (2018). Neural sign language translation. In *CVPR* (pp. 7784–7793). Utah, United States.

Cao, Z., Simon, T., Wei, S., & Sheikh, Y. (2017). Real-time multi-person 2D pose estimation using part affinity fields. In *CVPR*. Hawaii, United States.

Chai, X., Guang, L., Lin, Y., Xu, Z., Tang, Y., Chen, X., & Zhou, M. (2013). Sign language recognition and translation with kinect.

Chen, T. Y., Ting, P. W., Wu, M. Y., & Fu, L. C. (2018). Learning a deep network with spherical part model for 3D hand pose estimation. *Pattern Recognition*, *80*, 1–20.

Chen, H., Tong, R., Chen, M., Fang, Y., & Liu, H. (2018). A hybrid CNN-SVM classifier for hand gesture recognition with surface EMG signals. In *2018 international conference on machine learning and cybernetics (ICMLC)* (pp. 619–624).

Chen, X., Wanga, G., Guoa, H., & Zhanga, C. (2020). Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2018.06.097.

Chen, C., Zhang, B., Hou, Z., Jiang, J., Liu, M., & Yang, Y. (2017). Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features. *Multimedia Tools and Applications*, *76*, 4651–4669.

Chen, Y., Zhao, L., Peng, X., Yuan, J., & Metaxas, D. . N. (2019). Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. In *BMVC, UK* (pp. 1–13).

Cheok, M., Omar, Z., & Jaward, M. (2017). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 1–23.

Cheron, G., Laptev, I., & Schmid, C. (2015). P-CNN: Pose-based CNN features for action recognition. In *IEEE International conference on computer vision (ICCV)*. Chile.

Cihan Camgöz, N., Hadfield, S., Koller, O., & Bowden, R. (2017). SubUNets: End-to-end hand shape and continuous sign language recognition. In *IEEE international conference on computer vision (ICCV) 2017*. Venice, Italy.

Cooper, H., Ong, W., Pugeault, N., & Bowden, R. (2012). Sign language recognition using sub-units. *Journal of Machine Learning Research 13*, 2205–2231.

Cui, R., Liu, H., & Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), 1880–1891.

Dadashzadeh, A., Tavakoli Targhi, A., & Tahmasbi, M. (2018). HGR-Net: A two-stage convolutional neural network for hand gesture segmentation and recognition. arXiv:1806.05653.

Deng, X., Yang, S., Zhang, Y., Tan, P., Chang, L., & Wang, H. (2017). Hand3D: Hand pose estimation using 3D neural network. arXiv:1704.02224.

Devineau, G., Xi, W., Moutarde, F., & Yang, J. (2018). Deep learning for hand gesture recognition on skeletal data. In *13th IEEE conference on automatic face and gesture recognition*. China.

Dibia, V. (2017). Handtrack: A library for prototyping real-time hand tracking interfaces using convolutional neural networks. *GitHub Repository*, https://github.com/victordibia/handtracking/tree/master/docs/handtrack.pdf.

Dibra, E., Wolf, T., Oztireli, C., & Gross, M. (2017). How to refine 3D hand pose estimation from unlabelled depth data?. In *International conference on 3D vision (3DV)*. Qingdao, China.

Doersch, C. (2016). Tutorial on variational autoencoders. arXiv:1606.05908.

Doosti, B. (2019). Hand pose estimation: A survey. arXiv:1903.01013.

Duan, J., Zhou, S., Wany, J., Guo, X., & Li, S. Ż. (2016). Multi-modality fusion based on consensus-voting and 3D convolution for isolated gesture recognition. arXiv:1611.06689.

Elboushaki, A., Hannane, R., Afdel, K., & Koutti, L. (2020). MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems With Applications*, *139*.

Escalera, S., Athitsos, V., & Guyon, I. (2016). Challenges in multi-modal gesture recognition. *Journal of Machine Learning Research*, *17*, 1–54.

Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopés, O., Guyon, I., Athitsos, V., & Escalante, H. (2013). Multi-modal gesture recognition challenge 2013: dataset and results. In *15th ACM conference, Sydney, Australia*. http://dx.doi.org/10.1145/2522848.2532595.

Escobedo-Cardenas, E., & Camara-Chavez, G. (2015). A robust gesture recognition using hand local data and skeleton trajectory. In *2015 IEEE international conference on image processing (ICIP), Quebec City, QC, 2015* (pp. 1240–1244).

Escobedo-Cardenas, E., & Camara-Chavez, G. (2020). Multi-modal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes. *Journal of Visual Communication and Image Representation*.

Fang, X., & Lei, X. (2017). Hand pose estimation on hybrid CNN-AE model. In *Proceedings of the 2017 IEEE, International conference on information and automation (ICIA), China*.

Ferreira, P., Cardoso, J., & Rebelo, A. (2019). On the role of multi-modal learning in the recognition of sign language. *Multimedia Tools and Applications*, *78*, 10035–10056.

Fischer, A., & Igel, C. (2012). An introduction to restricted Boltzmann machines. In *Proceedings of the 17th Iberoamerican congress on pattern recognition (CIARP 2012) LNCS 7441, uenos Aires, Argentina*. http://dx.doi.org/10.1007/978-3-642-33275-3_2.

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J., & Ney, H. (2012). RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *International conference on language resources and evaluation*. Istanbul, Turkey.

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., & Ney, H. (2014). Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *International conference on language resources and evaluation (LREC), harpa conference centre in Reykjavik (Iceland)*.

Frederic, B., Lamblin, P., Pascanu, R., et al. (2012). Theano: new features and speed improvements. In *NIPS Workshop, Canada*. http://deeplearning.net/sofware/theano/.

Ganapathi, V., Plagemann, C., Koller, D., & Thrun, S. Real-time human pose tracking from range data. In *ECCV* (pp. 738–751). Italy.

Gattupalli, S., Ghaderi, A., & Athitsos, V. (2016). Evaluation of deep learning based pose estimation for sign language recognition. arXiv:1602.09065.

Ge, L., Liang, H., Yuan, J., & Thalmann, D. (2017). 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *CVPR* (pp. 1991–2000). Hawaii, United States.

Ge, L., Liang, H., Yuan, J., & Thalmann, D. (2018). Robust 3D hand pose estimation in single depth images: from single-view CNN to multi-view CNNs. *IEEE Transactions on Image Processing*.

Ge, L., Ren, Z., & Yuan, J. (2018). Point-to-point regression pointnet for 3D hand pose estimation. In *ECCV* (pp. 1–17). Munich, Germany.

Girshick, R. (2015). Fast R-CNN. In *2015 IEEE international conference on computer vision (ICCV), Santiago, Chile*. http://dx.doi.org/10.1109/ICCV.2015.169.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(1), 142–158.

Gomez-Donoso, F., Orts-Escolano, S., & Cazorla, M. (2019). Accurate and efficient 3D hand pose regression for robot hand tele-operation using a monocular RGB camera. *Expert Systems With Applications*, *136*, 327–337.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NIPS*. Monteral, Canada.

Grosse, R. (2017). *CSC321 Lecture 20: Autoencoders*. Toronto University, http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/slides/lec20.pdf.

Guo, H., Wang, G., & Chen, X. (2017). Towards good practices for deep 3D hand pose estimation. arXiv:1707.07248.

Han, M., Chen, J., Li, L., & Chang, Y. (2016). Visual hand gesture recognition with convolution neural network. In *17th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)*. China.

Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., & Fei-Fei, L. (2016). Towards viewpoint invariant 3D human pose estimation. In *ECCV*. Amsterdam, Netherlands.

Hinton, G. (2007). Deep belief nets. In *NIPS*. Vancouver, B.C., Canada.

Hinton, G., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

Huang, J., Zhou, W., Li, H., & Li, W. (2015). Sign language recognition using 3D convolutional neural networKS. In *IEEE international conference on multimedia and expo (ICME)*. Turin, Italy.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: LArge scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Marı n Jimeneza, M., Romero-Ramireza, F., Munoz-Salinasa, R., & Medina-Carnicer, R. (2018). 3D Human pose estimation from depth maps using a deep combination of poses. *Journal of Visual Communication and Image Representation*, 627–639.

John, V., Boyali, A., Mita, S., Imanishi, M., & Sanma, N. (2016). Deep learning-based fast hand gesture recognition using representative frames. In *International conference on digital image computing: techniques and applications (DICTA)*. Australia.

Kang, B., Tripathi, S., & Nguyen, T. (2015). Real-time sign language finger-spelling recognition using convolutional neural networks from depth map. In *3rd IAPR Asian conference on pattern recognition (ACPR)*. Kuala Lumpur, Malaysia.

Kapuscinski, T., Oszust, M., Wysocki, M., & Warchol, D. (2015). Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, *12*(4).

Kazakos, E., Nikou, C., & Kakadiaris, I. (2018). On the fusion of rgb and depth information for hand pose estimation. In *25th IEEE international conference on image processing (ICIP)* (pp. 868–872). Athens, Greece.

Kim, S., Ban, Y., & Lee, S. (2017). Tracking and classification of in-air hand gesture based on thermal guided joint filter. *Sensors*.

Kocabas, M., Karagoz, S., & Akbas, E. (2018). MultiPoseNet: Fast multi-person pose estimation using pose residual network. In *CVPR*. Utah, United States.

Koller, O., Forster, J., & Hermann, N. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125.

Koller, O., Ney, H., & Bowden, R. (2015). Deep learning of mouth shapes for sign language. In *IEEE international conference on computer vision workshop (ICCVW), santiago, Chile*.

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2016). Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *BMVC, UK*.

Kopuklu, O., Gunduz, A., Kose, N., & Rigoll, G. (2019). Real-time hand gesture detection and classification using convolutional neural networks. arXiv:1901.10323.

Le, T., Jaw, D., Lin, I., Liu, H., & Huang, S. (2018). An efficient hand detection method based on convolutional neural network. In *7th IEEE international symposium on next-generation electronics*. Taipei, Taiwan.

Li, Y., Xue, Z., Wang, Y., Ge, L., Ren, Z., & Rodriguez, J. (2019). End-to-end 3D hand pose estimation from stereo cameras. In *BMVC*. UK.

Lifshitz, I., Fetaya, E., & Ullman, S. (2016). Human pose estimation using deep consensus voting. In *ECCV* (pp. 246–260).

Lim, K., Tan, A., Lee, C., & Tan, S. (2019). Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, *78*, 19917–19944.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., & Berg, A. (2016). SSD: Single shot multibox detector. In *ECCV* (pp. 21–37). Amsterdam, Netherlands.

Liu, L., & Shao, L. (2013). Learning discriminative representations from RGB-D video data. In *Proceedings of the twenty-third international joint conference on artificial intelligence (IJCAI)*. Beijing, China.

Ma1, M., Chen, Z., & Wu, J. (399–404). A recognition method of hand gesture with CNN-SVM model. In *International conference on bio-inspired computing: theories and applications* (pp. 399–404). Harbin, China.

Madadi, M., Bertiche, H., & Escalera, S. (2020). SMPLR: Deep SMPL reverse for 3D human pose and shape recovery. *Pattern Recognition*, *106*, http://dx.doi.org/10.1016/j.patcog.2020.107472.

Madadi, M., Escalera1, S., Baro, X., & Gonzalez, J. (2017). End-to-end global to local CNN learning for hand pose recovery in depth data. arXiv:1705.09606.

Matilainen, M., Sangi, P., Holappa, J., & Silven, O. (2016). OUHANDS Database for hand detection and pose recognition. In *International conference on image processing theory, tools and applications, Finland*. http://dx.doi.org/10.1109/IPTA.2016.7821025.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *5*, 115–133.
Mittal, V. (2018). Top 15 deep learning applications that will rule the world in 2018 and beyond. www.medium.com.
Mocialov, B., Turner, G., Lohan, K., & Hastie, H. (2017). Towards continuous sign language recognition with deep learning. semanticscholar. https://www.semanticscholar.org/paper/Towards-Continuous-Sign-Language-Recognition-with-Mocialov-Turner/f24c82e85906bc7325b296d37370febd65833fdd.
Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2015). Hand gesture recognition with 3D convolutional neural networks. In *IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. Boston, Massachusetts.
Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., & Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*. Las Vegas, NV, USA.
Moon, G., Chang, J., & Lee, K. (2018). V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In *CVPR*. Salt Lake City, Utah, United States.
Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., & Theobalt, C. (2018). Ganerated hands for realtime 3d hand tracking from monocular rgb. In *CVPR, Salt Lake City, Utah, United States* (pp. 1–11). http://dx.doi.org/10.1109/CVPR.2018.00013.
Murray, J. (2018). World Federation of the deaf. Rome, Italy. Retrieved from http://wfdeaf.org/our-work/. (Accessed 30 January 2020).
MXNET (2020). MXNET. Available online: Accessed date: Jun, 2020.
Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2014). Hand segmentation with structured convolutional learning. In Asian conference on computer vision (ACCV) 2014: Computer vision (pp. 687–702). Singapore.
Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision (ECCV)* (pp. 483–499).
Oberweger, M., Riegler, G., Wohlhart, P., & Lepetit, V. (2016). Efficiently creating 3D training data for fine hand pose estimation. In *CVPR*. Nevada, United States.
Oberweger, M., Wohlhart, P., & Lepetit, V. (2015). Hands deep in deep learning for hand pose estimation. In *Proceedings of 20th computer vision winter workshop (CVWW)* (pp. 21–30).
Oszust, M., & Wysocki, M. (2013). Polish sign language words recognition with Kinect. In *6th International conference on human system interactions (HSI)*. Sopot, Poland.
Pagebites, I. (2018). Imo. United States. http://www.imo.com.
Pu, J., Zhou, W., & Li, H. (2018). Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI18: Proceedings of the 27th international joint conference on artificial intelligence*. Stockholm.
Pugeault, N., & Bowden, R. (2011). Spelling it out: Real-Time ASL finger-spelling recognition. In *Proceedings of the 1st IEEE workshop on consumer depth cameras for computer vision, jointly with ICCV'2011*. Barcelona, Spain.
Rao, G., Syamala, K., Kishore1, P., & Sastry, A. (2018). Deep convolutional neural networks for sign language recognition. In *Conference on signal processing and communication engineering systems (SPACES)*. India.
Rastgoo, R., Kiani, K., & Escalera, S. (2018). Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. *Entropy*.
Rastgoo, R., Kiani, K., & Escalera, S. (2020a). Hand sign language recognition using multi-view hand skeleton. *Expert Systems With Applications*, *150*.
Rastgoo, R., Kiani, K., & Escalera, S. (2020b). Video-based isolated hand sign language recognition using a deep cascaded model. *Multimedia Tools and Applications*, http://dx.doi.org/10.1007/s11042-020-09048-5.
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. arXiv:1506.02640.
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*. Quebec, Canada.
Ronchetti, F., Quiroga, F., Estrebou, C., & Lanzarini, L. (2016). Handshape recognition for argentinian sign language using probsom. *Journal of Computer Science & Technology*, *16*(1).
Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., & Rosete, A. (2016). LSA64: An argentinian sign language dataset. In *Congreso Argentino de Ciencias de la Computación (CACIC 2016)*.
Canuto-dos Santos, C., Leonid-Aching-Samatelo, J., & Frizera-Vassallo, R. (2020). Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation. *Neurocomputing*, *400*, 238–254.
Sapp, B., & Taskar, B. (2013). MODEC: Multi-modal decomposable models for human pose estimation. In *CVPR*. Portland, Oregon.
Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multi-view bootstrapping. arXiv:1704.07809.
Sinha, A., Choi, C., & Ramani, K. (2016). DeepHand: Robust hand pose estimation by completing a matrix imputed with deep features. In *CVPR* (pp. 4150–4159). Las Vegas, NV, USA.
Smedt, Q., Wannous, H., & Vandeborre, J. (2016). Dynamic hand gesture recognition using skeleton-based features. In *CVPRW*. Las Vegas, Nevada, United States.
Spurr, A., Song, J., Park, S., & Hilliges, O. (2018). Cross-modal deep variational hand pose estimation. In *CVPR* (pp. 89–98). Salt Lake City, Utah, United States.
Supancic, J., Rogez, G., Yang, Y., Shotton, J., & Ramana, D. (2018). Depth-based hand pose estimation: methods, data, and challenges. *International Journal of Computer Vision*, 1180–1198.
Tagliasacchi, A., Schröder, M., Tkach, A., Bouaziz, S., Botsch, M., & Pauly, M. (2015). Robust articulated-ICP for real-time hand tracking. In *Eurographics symposium on geometry processing*.
Tang, A., Lu, K., Wang, Y., Huang, J., & Li, H. (2015). A real-time hand posture recognition system using deep neural networks. In *ACM transactions on intelligent systems and technology (TIST) - special section on visual understanding with RGB-D sensors*.
TensorFlow (2020). Tensorflow. Retrieved from Available online: Accessed date: Jun, 2020.
Thangali, A., Nash, J., Sclaroff, S., & Neidle, C. (2011). Exploiting phonological constraints for handshape inference in ASL video. In *CVPR*. USA.
Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, *33*, 1–10.
Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. arXiv:1312.4659.
Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M., Laptev, I., & Schmid, C. (2017). Learning from synthetic humans. In *CVPR*. Hawaii, United States.
Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Hindawi Computational Intelligence and Neuroscience*, 1–13. http://dx.doi.org/10.1155/2018/7068349.
Wadhawan, A., & Kumar, P. (2020). Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, 1–12. http://dx.doi.org/10.1007/s00521-019-04691-y.
Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., & Li, S. (2016). Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition.In *CVPRW 2016*. Nevada, United States.
Wang, T. (2016). *Recurrent neural network*. Machine Learning Group, University of Toronto, for CSC 2541, Sport Analytics, https://www.cs.toronto.edu/~tingwuwang/rnn_tutorial.pdf.
Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., & Ma, L. (2018). DRPose3D: Depth ranking in 3D human pose estimation. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18)* (pp. 978–984).
Wang, P., Li, W., Liu, S., Gao, Z., Tang, C., & Ogunbona, P. (2017). Large-scale isolated gesture recognition using convolutional neural networks. arXiv:1701.01814.
Wei, S., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*. Las Vegas, Nevada.
Wei, C., Zhou, W., Pu, J., & Li, H. (2019). Deep grammatical multi-classifier for continuous sign language recognition. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*. Singapore.
Wu, J. (2019). *Convolutional neural networks*. LAMDA Group, National Key Lab for Novel Software Technology Nanjing University, China, https://cs.nju.edu.cn/wujx/teaching/15_CNN.pdf.
Wu, J., Chen, J., Ishwar, P., & Konrad, J. (2016). Two-stream CNNs for gesture-based verification and identification: learning user style. In *Computer vision and pattern recognition (CVPR)*. Las Vegas, Nevada.
Yan, S., Xia, Y., Smith, J., Lu, W., & Zhang, B. (2017). Multi-scale convolutional neural networks for hand detection. *Applied Computational Intelligence and Soft Computing*, *2017*.
Yang, Y., Li, Y., Fermuller, C., & Aloimonos, Y. (2015). Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*.
Ye, Y., Tian, Y., Huenerfauth, M., & Liu, J. (2018). Recognizing American sign language gestures from within continuous videos. In *CVPR*. Utah, United States.
Yuan, S., Ye, Q., Stenger, B., Jain, S., & Kim, T.-K. (2017). Big hand 2.2M benchmark: Hand pose dataset and state of the art analysis. In *CVPR*. Honolulu, Hawaii, USA.
Zheng, L., Liang, B., & Jiang, A. (2017). Recent advances of deep learning for sign language recognition. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, NSW, Australia*. IEEE.
Zhou, X., Wan, Q., Zhang, W., Xue, X., & Wei, Y. (2016). Model-based deep hand pose estimation. In *IJCAI*.
Zimmerman, T., Lanier, J., Blanchard, C., Bryson, S., & Harvill, Y. (1987). A hand gesture interface device. In *87th Proceedings of the SIGCHI/GI conference on human factors in computing systems and graphics, toronto, Ontario, Canada* (pp. 189–192).
Zimmermann, C., & Brox, T. (2017). Learning to estimate 3D hand pose from single RGB images. In ICCV. Venice, Italy.