Multi-Modal Multi-Channel American Sign Language Recognition

Elahe Vahdani[a,*], Longlong Jing[a,*], Matt Huenerfauth[b], Yingli Tian[a,c,**]

[a] *The Graduate Center, City University of New York*
[b] *The Rochester Institute of Technology*
[c] *The City College of New York, City University of New York*

Abstract

In this paper, we propose a machine learning-based multi-stream framework to recognize American Sign Language (ASL) manual signs and non-manual gestures (face and head movements) in real-time from RGB-D videos. Our approach is based on 3D Convolutional Neural Networks (3DCNN) by fusing multimodal features including hand gestures, facial expressions, and body poses from multiple channels (RGB, depth, motion, and skeleton joints). To learn the overall temporal dynamics in a video, a proxy video is generated by selecting a subset of frames for each video which are then used to train the proposed 3DCNN model. We collected a new ASL dataset, ASL-100-RGBD, which contains *42* RGB-D videos captured by a Microsoft Kinect V2 camera. Each video consists of *100* ASL manual signs, along with RGB channel, depth maps, skeleton joints, face features, and HD face. The dataset is fully annotated for each semantic region (i.e. the time duration of each sign that the human signer performs). Our proposed method achieves *92.88%* accuracy for recognizing *100* ASL sign glosses in our newly collected ASL-100-RGBD dataset. The effectiveness of our framework for recognizing hand gestures from RGB-D videos is further demonstrated on a large-scale dataset, Chalearn IsoGD, achieving the state-of-the-art results.

Keywords: American Sign Language Recognition, Hand Gesture Recognition, RGB-D Video Analysis, Multimodality, 3D Convolutional Neural Networks, Proxy Video

1. Introduction

American Sign Language (ASL) is a natural language conveyed through movements and poses of the hands, body, head, eyes, and face [1]. There are more than one hundred sign languages worldwide, and ASL is used throughout the U.S. and Canada, as well as other regions of the world, including West Africa and Southeast Asia. Within the U.S.A., about 28 million people today are Deaf or Hard-of-Hearing (DHH) [2]. There are approximately

[*] Equal contribution

[**] Corresponding author

Email addresses: evahdani@gradcenter.cuny.edu (Elahe Vahdani), ljing@gradcenter.cuny.edu (Longlong Jing), matt.huenerfauth@rit.edu (Matt Huenerfauth), ytian@ccny.cuny.edu (Yingli Tian)

$500,000$ people who use ASL as a primary language [3], and since there are significant linguistic differences between English and ASL, it is possible to be fluent in one language but not in the other. Most ASL signs consist of the hands moving, pausing, and changing orientation in space. Facial expressions in ASL are most commonly utilized to convey information about entire sentences or phrases, and are referred to as "syntactic facial expressions", as discussed in [4]. Individual ASL signs consist of a sequence of several phonological segments, which include:

- An essential parameter of a sign is the configuration of the hand, i.e., the degree to which each of the finger joints is bent, commonly referred to as the "handshape." In ASL, there are approximately 86 handshapes, which are widely used [5], and the hand may transit between handshapes during the production of a single sign.
- During an ASL sign, the signer's hands will occupy specific locations and perform movements through space. Some signs are performed by a single hand, but most are performed using both of the signer's hands, which move through the area in front of their head and torso. During two-handed signs, the two hands may have symmetrical movements, or the signer's dominant hand (e.g., the right hand of a right-handed person) will have more significant changes than the non-dominant hand.
- The orientation of the palm of the hand in 3D space is also a meaningful aspect of an ASL sign, and this parameter may differentiate pairs of otherwise identical signs.
- Some signs co-occur with specific "non-manual signals," which are generally facial expressions characterized by specific mouth gestures, eyebrow movement, head tilt/turn, or head movements (e.g., forward-backward relative to the torso).

Sign language recognition can be categorized to isolated or continuous recognition. Isolated sign language recognition focuses on recognizing isolated signs through movements of the hands and quick facial expression changes. In continuous sign language recognition, the temporal boundaries of individual signs are not provided and the transition movements between two consecutive signs is hard to detect. While some researchers, e.g., [6], have investigated the identification of facial expressions that extend across multiple signs to indicate grammatical information, in this paper, we describe our work on recognizing isolated signs. The category of facial expressions, which is specifically relevant to the task of recognizing individual signs, is referred to as "lexical facial expressions," which are considered as a part of the production of an isolated ASL sign (see examples in Fig. 1). Such facial expressions are, therefore, essential for the task of sign recognition. For instance, signs with negative semantic polarity, e.g., NONE or NEVER, tend to occur with a negative facial expression consisting of a slight head shake and nose wrinkle. Besides, specific ASL signs almost always happen in a context in which a particular ASL syntactic facial expression occurs. For instance, some question signs, e.g., WHO or WHAT, tend to co-occur with a syntactic facial expression (brows furrowed, head tilted forward), which indicates that an entire sentence is a WH Question. Thus, such a facial expression may be useful evidence to consider when building a recognition system for such signs.

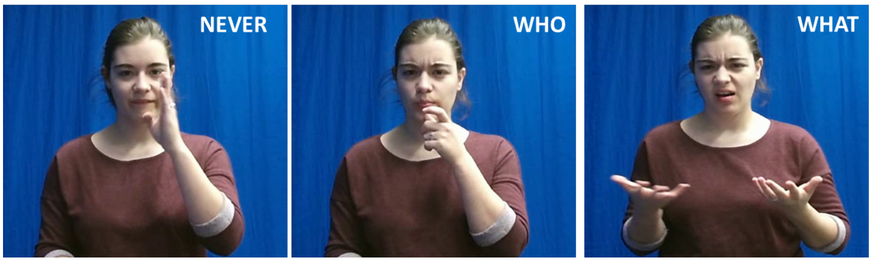


Figure 1: Example images of lexical facial expressions along with hand gestures for signs: NEVER, WHO, and WHAT. For NEVER, the signer shakes her head side-to-side slightly, which is a Negative facial expression in ASL. For WHO and WHAT, the signer is furrowing the brows and slightly tilting moving the head forward, which is a WH Question facial expression in ASL.

1.1. Motivations

In addition to the many members of the Deaf community who may prefer to communicate in ASL, many individuals seek to learn the language. Due to a variety of educational factors and childhood language exposure, researchers have measured lower levels of English literacy among many deaf adults in the U.S. [7]. Studies have shown that deaf children raised in homes with exposure to ASL have better literacy as adults, but it can be challenging for parents, teachers, and other adults in the life of a deaf child to rapidly gain fluency in ASL. The study of ASL as a foreign language in universities has significantly increased by 16.4% from 2006 to 2009, which ranked ASL as the 4th most studied language at colleges [8]. Thus, many individuals would benefit from a flexible way to practice their ASL signing skills.

Our research investigates technologies for recognizing signs performed in color and depth videos, as discussed in [9]. The focus of our research is to develop a real-time system that can automatically identify ASL signs, comprising manual and non-manual gestures, from RGB-D videos. This is aligned with our broader goal to design assistive technologies to support ASL education by providing ASL students immediate feedback about the fluency of their signing performances. While the development of user-interfaces for educational software was described in our prior work [9], this article instead focuses on the development and evaluation of our ASL recognition technologies, which underlie our educational tool. Beyond this specific application, automatic recognition of ASL signs from videos could enable new communication and accessibility technologies for people who are DHH. These tools may allow users to input information into computing systems by performing sign language or serve as a foundation for future research on machine translation technologies for sign languages.

1.2. Challenges

Sign language recognition shares properties with video action recognition but it has specific challenges caused by its unique characteristics. One challenge is visual complexity; for instance, slight difference in one hand's phonemes can generate another sign or be undefined. Also, for some pair of signs, hand gestures look identical, and we can only discriminate them by paying attention to the difference in facial expressions. In some cases, a hand gesture can impose multiple meanings depending on the number of repetitions. The other challenge is occlusion, i.e., hand-hand occlusion or hand-face occlusion where hands or face are partially visible in some moments of signing. To address these challenges, we design a multi-modal network to combine features from multiple modalities such as hand gestures, facial expressions, and body poses to better distinguish signs as some of the signs are only identifiable by simultaneous articulations of manual and non-manual sources. Furthermore, our network leverages information from multiple channels including RGB, depth, motion, and skeleton joints to better capture subtle movements of hands and facial expression for fine-grain analysis. Another challenge is the variation of signs performed by different signers such as pose or duration variations, pausing between signs or letters, wearing colored gloves or long sleeves shirts. Also, variation in the environment setup such as illumination, background, or distance from the camera can make the problem harder. To tackle this challenge, we have collected a new ASL dataset, ASL-100-RGBD, where 100 ASL signs have been collected and performed by *15* individual signers. To ensure a subject-independent evaluation, no same signer appears in both training and testing sets.

1.3. Scope of Contributions

As discussed in Section 2.1, most prior ASL recognition studies typically focus on isolated hand gestures without considering facial expressions and body poses or they only use RGB videos. In this paper, we propose a 3D multi-stream framework to recognize a set of grammatically important ASL signs from RGB-D videos in real-time. The proposed method operates by fusing multimodal features, including hand gestures, facial expressions, and body poses from multi-channel (RGB, depth, motion, and skeleton joints). To the best of our knowledge, we believe this is the first work that combines multi-channel videos (RGB and depth) with the fusion of multi-modal features for ASL recognition. Furthermore, most datasets are either do not have "depth" data or they are in other sign languages (not American) or they are designed for continuous sign language recognition (not isolated). To the best of our knowledge, ASL-100-RGBD is the only American sign language dataset collected for isolated signs that includes RGB and depth data (RGBD). The main contributions of the proposed framework can be summarized as follows:

- We propose a 3D multi-stream framework using 3D convolutional neural networks for ASL recognition in RGB-D videos by fusing multi-modal features such as hand gestures, facial expressions, and body poses in multiple-channels including RGB, depth, motion, and skeleton joints.
- We propose a temporal augmentation strategy to help the proposed 3D multi-stream network capture the long-term spatiotemporal information within video clips and augment the training data to handle the videos of relatively small datasets.
- We have created a new ASL dataset, ASL-100-RGBD, including multiple modalities (facial movements, hand gestures, and body pose) and multiple channels (RGB, depth, skeleton joints, and HD face) by collaborating with ASL linguistic researchers [10]. This dataset contains annotations of the time duration when the human in the video performs each ASL sign. The dataset is available to the research community.

- We further evaluate the proposed framework to recognize hand gestures on the Chalearn LAP IsoGD dataset [11], which consists of 249 gesture classes in $47,933$ RGB-D videos. Our framework achieves the state-of-the-art results using fewer channels (5 channels instead of 12 in previous work).

2. Related Work

2.1. RGB-D based ASL Recognition

Sign language (SL) recognition has been studied for three decades since the first attempt to recognize Japanese SL by Tamura and Kawasaki in 1988 [12]. The existing SL recognition research can be classified as sensor-based methods, including data gloves and body trackers to capture and track the hand and body motions [13, 14, 15, 16], and non-intrusive camera-based methods by applying computer vision technologies [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]. While most studies analyze the manual gestures, some methods exploit the linguistic information conveyed by the face and head of the signers, such as [42, 6, 43, 44]. More details about SL recognition can be found in these survey papers [45, 46, 47, 48, 49, 50, 51, 52, 53]. The availability of cost-effective RGB-D cameras in recent years, such as Microsoft Kinect V2 [54], Intel Realsense [55], Orbbec Astra [56], has facilitated capturing high-resolution RGB videos, depth maps, and tracking skeleton joints in real-time. Compared to traditional 2D RGB images, RGB-D images provide photometric and geometric information, motivating the research on ASL recognition using RGB and depth information [57, 58, 59, 60, 17, 36, 61, 62, 63, 64, 65, 66]. In this article, we briefly summarize ASL recognition methods using RGB-D images or videos.

Some early work of SL recognition based on RGB-D cameras only focused on a small number of signs from static images [57, 60, 67]. Pugeault and Bowden proposed a multi-class random forest classification method to recognize 24 static ASL fingerspelling alphabet letters by ignoring the letters j and z (as they involve motion) and combining appearance and depth information of handshapes captured by a Kinect camera [57]. Keskin *et al.* [67] recognized 24 static handshapes of the ASL alphabet, based on scale-invariant features extracted from depth images, fed to a Randomized Decision Forest for classification at the pixel level, where the final recognition label was voted based on a majority. Ren et al. proposed a modified Finger-Earth Mover's Distance metric to recognize static handshapes for 10 digits captured using a Kinect camera [60].

While these systems only used static RGB and depth images, some studies employed the RGB-D videos for ASL recognition. Zafrulla *et al.* developed a hidden Markov model (HMM) to recognize 19 ASL signs collected by Kinect camera and compared the performance with that from colored-glove and accelerometer sensors [58]. For the Kinect data, they compared the system performance between the signer seated and standing and found that higher accuracy resulted when the users were standing. Yang developed a hierarchical conditional random field method to recognize 24 manual ASL signs (seven one-handed and 17 two-handed) from the handshape and motion in RGB-D videos [63]. Lang *et al.* [68] presented a HMM framework to recognize 25 signs of German Sign Language using depth-camera specific features. Mehrotra *et al.* [69] employed a support vector machine (SVM) classifier to recognize 37 signs of Indian Sign Language based on 3D skeleton points captured using a Kinect camera. Almeida et al. [62] also employed an SVM classifier to recognize 34 signs of Brazilian Sign Language using handshape, movement, and the position captured by a Kinect. Jiang *et al.* proposed recognizing 34 signs of Chinese Sign Language based on the color images and the skeleton joints captured by a Kinect camera [61]. Recently, Kumar *et al.* [70] combined a Kinect camera with a Leap Motion sensor to recognize 50 signs of India Sign Language.

As discussed above, SL consists of hand gestures, facial expressions, and body poses. However, most existing methods have only focused on hand gestures without considering facial expressions and body poses. A few attempted to analyze hands and face [44, 19, 6, 71, 27, 43], but they only use RGB videos. To the best of our knowledge, we believe this is the first work that combines multi-channel RGB-D videos (RGB and depth) with the fusion of multi-modal features (hand, face, and body) for ASL recognition.

2.2. Machine Learning-based Action and Hand Gesture Recognition

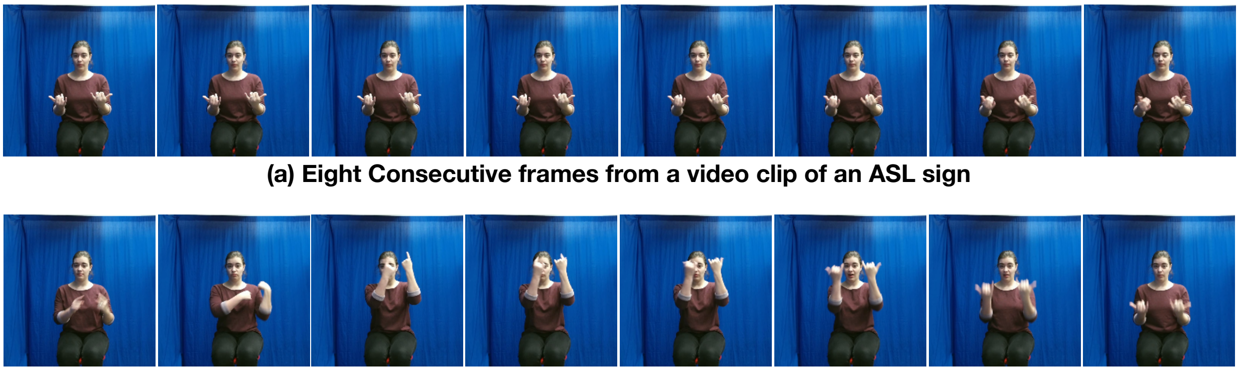
In addition to prior research on sign-recognition technologies, there has been significant research in action and hand gesture recognition, which is relevant to consider [72, 73, 74, 75, 76, 77, 78, 79, 80, 81]. Since the work of AlexNet [82] which makes use of the powerful computation ability of GPUs, deep neural networks (DNNs) have enjoyed a renaissance in various areas of computer vision, such as image classification [83, 84], object detection [85, 86], image description [87, 88], and others. Many efforts have been made to extend CNNs from image to video domain [89], which is more challenging because of the large volume of video data; therefore, processing video data in the limited GPU memory is not tractable. An intuitive way to extend image-based CNN structures to the video domain is to perform the fine-tuning and classification process on each frame independently. Then, conduct a later fusion, such as average scoring, to predict the action class of the video [90]. To incorporate temporal information in the video, [91] introduced a two-stream framework. One stream was based on RGB images, and the other, on stacked optical flows. Although it proposed an innovative way to learn temporal information using a CNN structure, in essence, it was still image-based, since the third dimension of stacked optical flows collapsed immediately after the first convolutional layer.

To model the sequential information of extracted features from different segments of a video, [87] and [92] proposed to input features into Recurrent Neural Network (RNN) structures, and they achieved good results for action recognition. The former emphasized pooling strategies and how to fuse different features, while the latter focused on how to train an end-to-end DNN structure that integrates CNNs with RNNs. These networks mainly use CNN to extract spatial features, then RNN is applied to extract the temporal information of the spatial features. 3DCNN was recently proposed to learn the Spatio-temporal features with 3D convolution operations [93],[94],[95],[96], and [97] has been widely used in video analysis tasks such as video caption and action detection. 3DCNN is usually trained with fixed-length clips (usually 16 frames [94],[97],) and later fusion is performed to obtain the final category of the entire video. The R(2+1)D network [98] separates spatial and temporal learning by using a 2D convolution for spatial features and a 1D convolution for temporal features. This separation allows the model to learn spatial and temporal features effectively and is computationally more efficient than 3D convolutions. Hara *et al.* [94] proposed the 3D-ResNet by replacing all the 2D kernels in 2D-ResNet with 3D convolution operations. With its advantage of avoiding gradient vanishing and explosion, the 3D-ResNet outperforms many complex networks.

ASL recognition shares properties with video action recognition; therefore, many networks for video action recognition have been applied to this task. Pigou *et al.* proposed temporal residual networks for gesture and sign language recognition [27] and temporal convolutions on top of the features extracted by 2DCNN for gesture recognition [22]. Huang *et al.* proposed a Hierarchical Attention Network with Latent Space (LS-HAN), which eliminates the pre-processing of the temporal segmentation [24]. Pu *et al.* proposed to employ a 3D residual convolutional network (3D-ResNet) to extract then visual features. The features are then fed to a stacked dilated convolution network with connectionist temporal classification to map the visual features into text sentence [25]. Camgoz *et al.* attempted to generate spoken language translations from sign language video [26]. Camgoz *et al.* proposed SubUNets for simultaneous hand shape and continuous sign language recognition [29]. Cui *et al.* proposed a weakly-supervised framework to train the network for continuous sign language recognition with videos only having the ordered gloss labels [28]. Zhou *et al.* proposed STMC network [99] to represent spatial cues with a 2DCNN (VGG [100]) and temporal cues with the bidirectional Long-Short Term Memory (BLSTM) [101]. Jiang *et al.* proposed SAM-SLR [102] to exploit whole body skeleton features for sign language in both RGB and RGB-D channels. Moryossef *et al.* also evaluated representations based on skeleton poses for sign language recognition [103]. Hu *et al.* designed a hand-model-aware framework for sign language with hand meshes and poses as the intermediate representation [104]. Zhang *et al.* proposed a global feature descriptor for time series modeling and a local feature extractor to model hands for sign language recognition [37]. Bohavcek *et al.* proposed a transformer model for word-level sign language recognition and introduced a robust pose normalization scheme to model hand poses [105]. Han *et al.* adopted a deep R(2+1)D network and argued that decomposing 3D convolution filters into separate spatial and temporal convolutions is beneficial for sign language recognition [106]. Bilge *et al.* proposed a zero-shot sign language recognition to train the models with the seen sign classes and recognize the instances of unseen sign classes [107]. In prior work, our research team proposed a 3D-FCRNN for ASL recognition by combining the 3DCNN and a fully connected RNN [36].

2.3. Public Camera-based ASL Datasets

As discussed in Section 2.1, technology to recognize ASL signs from videos could enable new educational tools or assistive technologies for people who are DHH, and there has been significant prior research on sign language recognition. However, a limiting factor for much of this research is the scarcity of video recordings of sign language that have been annotated with time interval labels of the sign glosses. For ASL, there have been some annotated



(a) Eight Consecutive frames from a video clip of an ASL sign

(b) Randomly sampled eight frames from the video clip of the same ASL sign

Figure 2: Generating representative proxy video by our proposed random temporal augmentation. (a) Eight consecutive frames from a video clip of an ASL sign. (b) Randomly sampled eight frames from the video clip of the same ASL sign. With the same number of frames, the proxy video captures more temporal dynamics of the ASL sign.

video-based datasets [108] or collections of motion capture recordings of humans wearing special sensors [109]. Most publicly available datasets, e.g. [110, 71], contain general ASL vocabularies from RGB videos and a few with RGB-D channels. Table 1 demonstrates the properties of some well-known sign language datasets.

2D Camera-based ASL databases: The American Sign Language Linguistic Research Project (ASLLRP) dataset contains video clips of signing from the front and side and includes a close-up view of the face [108], with annotations for 19 short narratives (1,002 utterances) and 885 additional elicited utterances from four Deaf native ASL signers. It includes annotations such as the start and endpoints of each sign and a unique gloss label for each sign. The start and endpoints of a range of non-manual behaviors are also labeled with respect to the linguistic information that they convey (serving to mark, e.g., different sentence types, topics, negation, etc.). Instances of non-manual behaviors include raised/lowered eyebrows, head position and periodic head movements, mouth gestures, and other expressions of the face. Dreuw *et al.* [111] produced several subsets from the ASLLRP dataset as benchmark databases for automatic recognition of isolated and continuous sign language. The American Sign Language Lexicon Video Dataset (ASLLVD) [112] is a large dataset of videos of isolated signs. It contains video sequences of about 3,000 distinct signs, each produced by 1 to 6 native ASL signers recorded by four cameras under three views (front, side, and face region). The annotations are provided, including start/end frames and class labels of every sign (i.e., gloss-based identification) plus locations of hands and face at every frame. The RVL-SLLL ASL Database [113] consists of three sets of ASL videos with distinct motion patterns, distinct handshapes, and structured sentences, respectively. These videos were captured from 14 native ASL signers (184 videos per signer) under different lighting conditions. For annotation, the videos with distinct motion patterns or distinct handshapes are saved as separate clips. However, there are no detailed annotations for the videos of structured sentences which limits the usefulness of the database. There

Table 1: The summary of sign language datasets of isolated signing.

Dataset	Sign Language	Signers	Vocabulary	Clips	Modalities
BosphorusSign22k [122]	Turkish	6	744	22,542	RGB+D
AUTSL [123]	Turkish	43	226	38,336	RGB+D
CSL (SLR500) [124]	Chinese	50	500	125,000	RGB+D
Polytropon [125]	Greek	1	2,703	3,517	RGB+D
ITI-GSL [126]	Greek	7	310	40,785	RGB+D
Signum [127]	German	25	455	11,375	RGB
BOBSL [128]	British	39	2,281	1,940	RGB
ASLLVD [112]	American	6	2,742	9,000	RGB
MS-ASL [117]	American	222	1,000	25,513	RGB
ASL-LEX [115]	American	69	1,000	-	RGB
ASL-LEX 2.0 [114]	American	-	2723	-	RGB
WLASL [116]	American	119	2,000	21,000	RGB
ASL-100-RGBD (ours)	American	22	100	4,150	RGB+D

are some other ASL datasets with only RGB channels such as ASL-LEX 2.0 [114], ASL-LEX [115], WLASL [116], and MS-ASL [117] for isolated sign language recognition and RWTH-BOSTON-104 [118], [119], RWTH-BOSTON-400 [120] and CopyCat [121] datasets for continuous sign language recognition.

RGB-D Camera-based ASL and Gesture Databases: Recently, several RGB-D databases have been collected for hand gesture and SL recognition [59, 23, 110]. Here we only briefly summarize RGB-D databases for ASL. The "Spelling-It-Out" dataset consists of 24 static handshapes of ASL fingerspelling alphabet, ignoring the letters "j" and "z" as they involve motion. Four signers repeat 500 samples for each letter in front of a Kinect camera [57]. The NTU dataset consists of 10 static hand gestures for digits 1 to 10 and was collected from 10 subjects by a Kinect camera. Each subject performs 10 different poses with variations in hand orientation, scale, articulation for the same gesture, and there is a color image and the corresponding depth map for each one [60]. The Chalearn LAP IsoGD dataset [11] is a large-scale hand gesture RGB-D dataset, which is derived from Chalearn Gesture dataset (CGD 2011) [129]. This dataset consists of $47,933$ RGB-D video clips fallen into 249 classes of hand gestures including mudras (Hindu/ Buddhist hand gestures), Chinese numbers, and diving signals. Although it is not about ASL recognition, it can be used to learn RGB-D features from different environment settings. Using the learned features as a pretrained model, the fine-tuned ASL recognition models are more robust to handle different backgrounds and scales (e.g. distance variations between Kinect camera and the signer). There are other sign language datasets with RGBD channels for isolated signs in Greek (ITI-GSL isol. [126], Polytropon [125]), Turkish (BosphorusSign [130], BosphorusSign22k [122], AUTSL [123]), and Chinese (SLR500 [124]) languages. How2Sign [131] and ASL-Homework-RGBD [132] are new ASL datasets with RGBD channels for continuous sign language recognition.

To support our research, we have collected and annotated a new RGB-D ASL dataset, ASL-100-RGBD, described in Section 4, with the following properties:

- *100* ASL signs have been collected and performed by *15* individual signers (often with multiple recordings from each signer).
- The ASL-100-RGBD dataset has been captured by a Kinect V2 camera and contains multiple channels including RGB, depth, skeleton joints, and HD face.
- Each video consists of 100 ASL signs shown in Fig. 4. The temporal boundary of each sign is annotated by ASL linguists, who labeled each span with one of 100 text labels.
- The *100* ASL signs have been strategically selected to support sign recognition educational tools with the detailed vocabulary composition described in Section 4. Many of these signs are characterized by both hand gestures and changes in facial expressions.

3. The Proposed Method for ASL Recognition

The pipeline of our proposed method is illustrated in Fig. 3. There are two main components in the framework: random temporal augmentation to generate proxy videos (which represent the overall temporal dynamics of the video clip of an ASL sign) and 3DCNN to recognize the class label of the sign.

3.1. Random Temporal Augmentation for Proxy Video Generation

The performance of the deep neural network greatly depends on the amount of the training data. Large-scale training data and different data augmentation techniques usually are needed for deep networks to avoid over-fitting. During training, different kinds of data augmentation techniques, such as random resizing and random cropping of images, are already widely applied in 3DCNN training. In order to capture the overall temporal dynamics, we apply a random temporal augmentation, to generate a proxy video for each sign video clip channel, by selecting a subset of frames, which has proved to be very effective for our proposed framework.

Videos are often redundant in the temporal dimension, and some consecutive frames are very similar without observable difference, as shown in Fig. 2 (a) which displays 8 consecutive frames in a video clip of an ASL sign while the proxy video in 2 (b) displays the 8 frames selected from the same video clip by random temporal augmentation. With the same number of frames, the proxy video provides more temporal dynamics. Thus, proxy videos are generated to represent the overall temporal dynamics for each ASL sign. To generate proxy videos, we uniformly divide the span of frames into T intervals and randomly sample one frame from every interval. If the total number of frames is less than T, it is padded with the last frame to the length of T. These proxy videos make it feasible to train a deep neural network on the dataset. The process of proxy video generation by random sampling is formulated in Eq. (1) below:

$$S_i = random(\lfloor N/T \rfloor) + \lfloor N/T \rfloor * i, \tag{1}$$

where N is the total number of frames in a signing video, T is the number of sampled frames, S_i is the i-th sampled frame, and $random(N/T)$ generates one random number in range $\lfloor 0, N/T \rfloor$ for every $i \in [0, T-1]$.

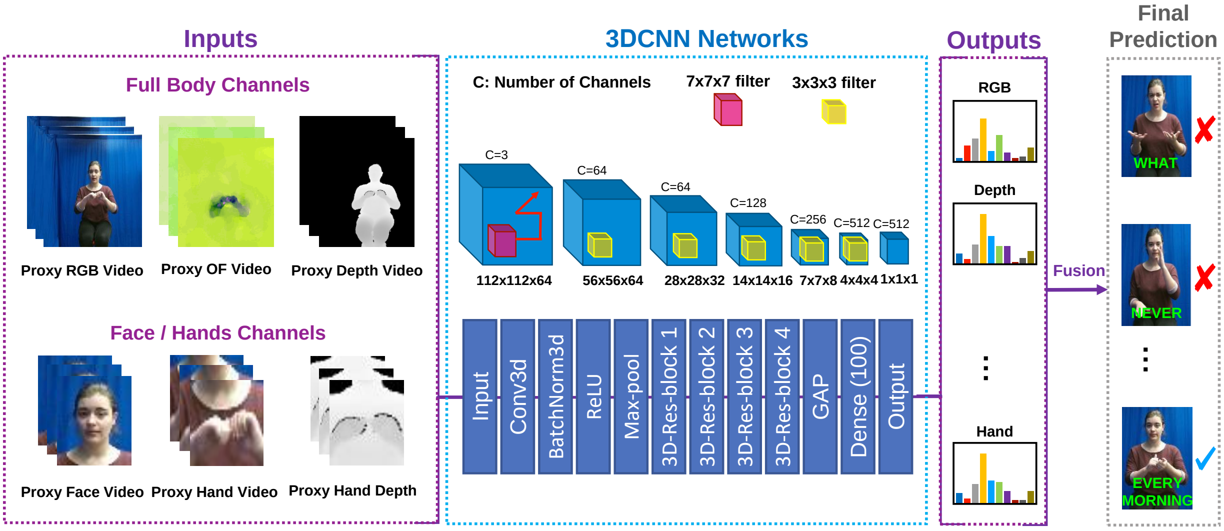


Figure 3: The pipeline of the proposed multi-channel multi-modal 3DCNN framework for ASL recognition. There are multiple channels such as RGB, Depth, and Optical flow, and multiple modalities including hand gestures, facial expressions and body poses. Hands and face regions are cropped to better model hand gestures and the facial expression changes. The whole framework consists of two main components: proxy video generation and 3DCNN modeling. First, proxy videos are generated for each ASL sign by selecting a subset of frames spanning the whole video clip of each ASL sign, to represent the overall temporal dynamics. Then the generated proxy videos of RGB, Depth, Optical flow, RGB of hands, and RGB of the face are fed into the multi-stream 3DCNN component. The predictions of these networks are weighted to obtain the final results of ASL recognition. The detailed architecture of our network is shown in Table 2.

3.2. 3D Convolutional Neural Network

3DCNN was first proposed for video action recognition [95] and was improved in C3D [97] by using a similar architecture to VGG [100]. It obtained state-of-the-art performance for several video recognition tasks. The difference between the 2DCNN and 3DCNN operation is that 3DCNN has an extra-temporal dimension, capturing the spatial and temporal information between video frames more effectively. After the emergence of C3D, many 3DCNN models were proposed for video action recognition [133],[93],[96]. The 3D-ResNet is the 3D version of ResNet, which introduced identical mapping to avoid gradient vanishing and explosion, making the training of very deep convolutional neural networks feasible. The size of the convolution kernel in 3D-ResNet is $w \times h \times t$ (w is the width of the kernel, h is the height of the kernel, and t is the temporal dimension of the kernel), while it is $w \times h$ in 2D-ResNet. In this paper, 3D-ResNet is chosen as the base network for ASL recognition.

The detailed architecture of our network is shown in Table 2. In the 3DResNet, there are five convolution blocks, where the first one consists of one convolution layer, one batch normalization layer, one ReLU layer, followed by one max-pooling layer. The next four convolution blocks are 3D residual blocks with skip connections. The number of kernels in the five convolution blocks are $\{64, 64, 128, 256, 512\}$. The Global Average Pooling (GAP) is followed after the fifth convolution block to produce a 512-dimensional feature vector. Then one fully connected layer and Softmax function are applied to produce the final prediction.

Table 2: The detailed architecture of our network. C is the number of classes which is 100 for ASL-100-RGBD dataset. GAP is Global Average Pooling.

Layer	Channels	Height	Width	Temporal
Input	3	112	112	64
Conv3d	64	56	56	64
BatchNorm3d	64	56	56	64
ReLU	64	56	56	64
Max-pool	64	28	28	32
3D-Res block	64	28	28	32
3D-Res block	128	14	14	16
3D-Res block	256	7	7	8
3D-Res block	512	4	4	4
GAP	512	1	1	1
FC	C	-	-	-

All the networks are optimized with cross-entropy loss with Stochastic Gradient Descent (SGD) optimizer. The cross-entropy loss function is formulated below. N is the number of samples in each mini-batch and C is the number of classes; $C = 100$ for ASL-100-RGBD. y_i is the ground-truth label for sample i and $\hat{y_i}$ is the prediction (output of the network). y_i and $\hat{y_i}$ are both C-dimensional vectors. y_i^c is 1 if video i belongs to class c, for $1 \leq c \leq C$, otherwise, it equals to 0. $\hat{y_i}$ is a probability vector where $\hat{y_i^c}$ is the predicted probability that video i belongs to class c.

$$L = -\frac{1}{N}(\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \cdot log(\hat{y_i^c})). \tag{2}$$

A hybrid framework comprising two 3DCNN networks is designed to recognize three main components of signing videos, such as hand gesture, facial expression, and body pose. The first 3DCNN (Body Network) captures the full-body movements by receiving multi-channel proxy videos generated from RGB, depth, and optical flow. The second 3DCNN (Hand-Face network) is designed to capture the coordinates of hands and face with the inputs of multi-channel proxy videos generated from the cropped regions of the left hand, right hand, and face. Only RGB and depth channels of hand regions are used in the Hand-Face network because optical flow cannot accurately track the quick and large motions of hands. Also, only the RGB channel of face region is employed since facial expressions generally change much less in-depth. The prediction results of the networks are weighted to obtain the final prediction of each ASL sign.

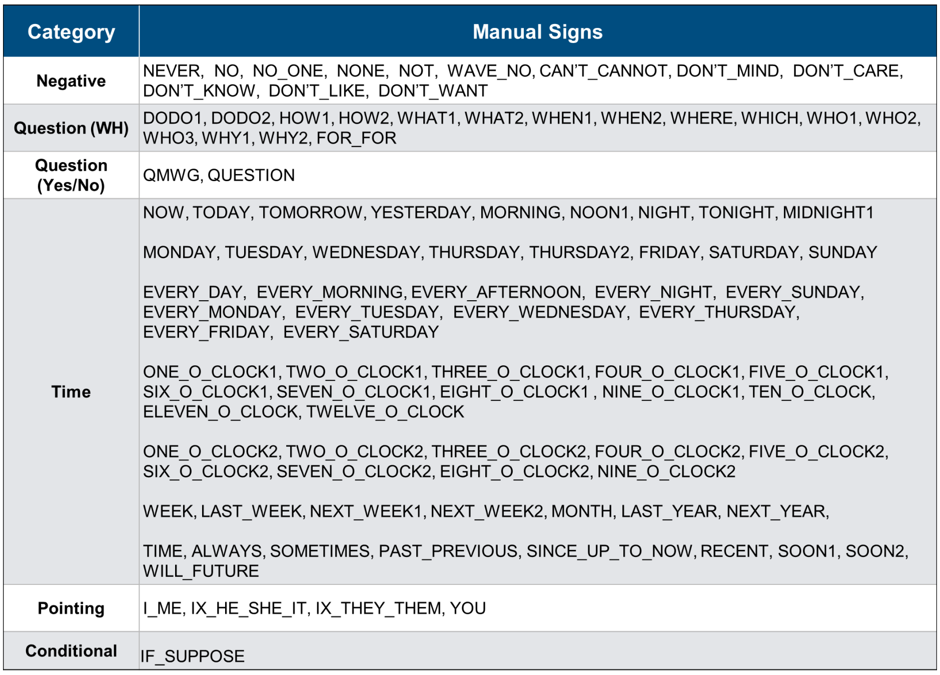
The optical flow images are calculated by stacking the x-component, the y-component, and the magnitude of the flow. Each value in the image is then rescaled to *0* and *255*. This practice has yielded good performance in other studies [87, 92]. As observed in the experimental results, the performance can be improved by fusing all the features generated by RGB, optical flow, and depth images. This indicates that different channels provide complementary information for ASL recognition through training deep neural networks.

4. ASL Dataset: "ASL-100-RGBD"

As mentioned in Section 2.3, we collected a new dataset from native ASL signers (individuals who have been using the language since very early childhood) in collaboration with ASL computational linguistic researchers. Each signer performed a list of *100* ASL signs (See the full list of ASL signs in Fig. 4) by using a Kinect V2 camera. Participants responded affirmatively to the following screening question: Did you use ASL at home growing up or attending a school as a very young child where you used ASL? Participants were provided with a slide-show presentation that asked them to perform a sequence of *100* individual ASL signs, without lowering their hands between signs. Since this new dataset includes *100* signs with RGB and depth data, we refer to it as the "ASL-100-RGBD" dataset.

During the recording session, a native ASL signer met the participant and conducted the session. Prior research in ASL computational linguistics has emphasized the importance of having only native signers present when recording ASL videos so that the signer does not produce English-influenced signing [109]. The dataset comprises *100* ASL signs, produced by *22* fluent signers, each often contributing multiple recordings. The participants, 15 men and 7 women, ranged in age from 20 to 51, with a median age of 23. Each recorded video consists of the 100 ASL signs, and the start-time and end-time of each of the signs have been annotated. Several signers missed few ASL signs in some videos during the recording. Typically two to three videos were recorded from each signer, which produced a total collection of 42 videos (each video contains about 100 signs) and *4*, *150* samples of ASL signs. To facilitate this collection process, we have developed a recording system based on Kinect 2.0 RGB-D camera to capture multiple modalities (facial expressions, hand gestures, and body poses) from multiple channels (RGB video and depth video) for ASL recognition. The recordings also include skeleton (25 joints for every video frame) and HD face (1,347 points) channels. The video resolution is *1920* x *1080* pixels for the RGB channel and *512* x *424* pixels for the depth channel, respectively.

The *100* ASL signs in this collection were selected strategically to support the research on sign recognition for ASL educational applications. The signs were chosen based on the vocabulary that is traditionally included in introductory ASL courses. Specifically, as discussed in [9], our recognition system must identify a subset of ASL signs that relate to a list of errors often made by students who are learning ASL. Our proposed educational tool [9] would receive as input a video of a student who is performing ASL sentences, and the system would automatically identify whether the student's performance may include one of several dozen errors, which are common among students learning ASL. As part of this system's operation, we require a sign-recognition component that can identify if a video of a person includes any of these *100* signs and the period in which the sign occurs. When one of these 100 key signs are identified, the system will consider other properties of the signer's movements, including hand shapes, timing, and repetitions [9], to determine whether the signer may have made a mistake in their signing.



Category	Manual Signs
Negative	NEVER, NO, NO_ONE, NONE, NOT, WAVE_NO, CAN'T_CANNOT, DON'T_MIND, DON'T_CARE, DON'T_KNOW, DON'T_LIKE, DON'T_WANT
Question (WH)	DODO1, DODO2, HOW1, HOW2, WHAT1, WHAT2, WHEN1, WHEN2, WHERE, WHICH, WHO1, WHO2, WHO3, WHY1, WHY2, FOR_FOR
Question (Yes/No)	QMWG, QUESTION
Time	NOW, TODAY, TOMORROW, YESTERDAY, MORNING, NOON1, NIGHT, TONIGHT, MIDNIGHT1 MONDAY, TUESDAY, WEDNESDAY, THURSDAY, THURSDAY2, FRIDAY, SATURDAY, SUNDAY EVERY_DAY, EVERY_MORNING, EVERY_AFTERNOON, EVERY_NIGHT, EVERY_SUNDAY, EVERY_MONDAY, EVERY_TUESDAY, EVERY_WEDNESDAY, EVERY_THURSDAY, EVERY_FRIDAY, EVERY_SATURDAY ONE_O_CLOCK1, TWO_O_CLOCK1, THREE_O_CLOCK1, FOUR_O_CLOCK1, FIVE_O_CLOCK1, SIX_O_CLOCK1, SEVEN_O_CLOCK1, EIGHT_O_CLOCK1 , NINE_O_CLOCK1, TEN_O_CLOCK, ELEVEN_O_CLOCK, TWELVE_O_CLOCK ONE_O_CLOCK2, TWO_O_CLOCK2, THREE_O_CLOCK2, FOUR_O_CLOCK2, FIVE_O_CLOCK2, SIX_O_CLOCK2, SEVEN_O_CLOCK2, EIGHT_O_CLOCK2, NINE_O_CLOCK2 WEEK, LAST_WEEK, NEXT_WEEK1, NEXT_WEEK2, MONTH, LAST_YEAR, NEXT_YEAR, TIME, ALWAYS, SOMETIMES, PAST_PREVIOUS, SINCE_UP_TO_NOW, RECENT, SOON1, SOON2, WILL_FUTURE
Pointing	I_ME, IX_HE_SHE_IT, IX_THEY_THEM, YOU
Conditional	IF_SUPPOSE

Figure 4: The full list of the 100 ASL signs in our “ASL-100-RGBD” dataset under 6 semantic categories. These ASL signs are strategically selected to support the technology and educational tools for sign language recognition. Many of these signs are characterized by both hand gestures and facial expression changes.

For instance, the *100* signs include words related to questions (e.g., WHO, WHAT), time-phrases (e.g., TODAY, YESTERDAY), negation (e.g., NOT, NEVER), and other categories that relate to key grammar rules of ASL. A full listing of the words included in this dataset is shown in Fig. 4. Note that there is no one-to-one mapping between English words and ASL signs, and some ASL signs have variations in their appearance, e.g., due to geographic/regional differences or other factors. For this reason, some words in Fig. 4 appear with integers after their name, e.g., THURSDAY and THURSDAY2, to reflect more than one variation in how the ASL sign may be produced. For instance, THURSDAY indicates a sign produced by the signer's dominant hand in the ”H” alphabet-letter handshape, with gentle circling in space. On the other hand, THURSDAY2 indicates a sign produced with the signer's dominant hand quickly switching from the alphabet-letter handshape of ”T” to ”H” while held in space in front of the torso. Both are commonly used ASL signs for the concept of ”Thursday” with two different representations.

As shown in Fig. 4, the words are grouped into 6 semantic categories (Negative, WH Questions, Yes/No Questions, Time, Pointing, and Conditional), suggesting that particular facial expressions are likely to co-occur with these words when used in ASL sentences. For instance, time-related phrases that appear at the beginning of ASL sentences tend to co-occur with a specific facial expression (head tilted back slightly and to the side, with eyebrows raised). Additional details about how detecting words in these various categories would be useful in the context of educational software appear in [9].

After the videos were collected from participants, the videos were analyzed by a team of ASL linguists, who produced time-coded annotations for each video. The linguists used a coding scheme in which an English identifier label was used to correspond to each of the ASL signs used in the videos, in a consistent manner across the videos. For example, all of the time spans in the videos when the human performed the ASL sign “NOT” were labeled with the English string ”NOT” in our linguistic annotation.

The ASL-100-RGBD dataset is available via the Databrary platform (Huenerfauth, 2020). A sample video [1] that visualizes the face and body-tracking information in this dataset is available. Fig. 5 demonstrates several frames of each channel of an ASL sign from our dataset including RGB, skeleton joints (25 joints for every frame), depth map, basic face features (5 main face components), and HD Face (1,347 points). The dataset [2] is available to the research community.

5. Experiments and Discussions

In this section, extensive experiments are conducted to evaluate the proposed approach on the newly collected “ASL-100-RGBD” dataset and Chalearn LAP IsoGD dataset [11].

5.1. Implementation Details

Same 3D-ResNet architecture is employed for all experiments. Different channels and modalities are fed to the network as input. The input channels are RGB, Depth, RGBflow (i.e. Optical flow of RGB images), and Depthflow (i.e. Optical flow of depth images) and the modalities are hands, face, and full body. The fusion of different channels and modalities are studied and compared.

Our proposed models are trained in PyTorch on four Titan X GPUs. To avoid overfitting, the pretrained models from Kinetics or Chalearn datasets are used and then random cropping and random rotation are applied to augment the data. The original resolution of RGB videos is 1920×1080 pixels. In order to meet the limitation of the computer memory, in our experiment, the center area of 800×800 pixels (where the signer is located) is resized to 134×134 as the input. In every iteration of the training, 112×112 image patches are randomly cropped from the 134×134 input images for data augmentation. During the testing, only the center patch of size 112×112 (from the 134×134 input image) is used for the prediction (no data augmentation is needed during testing). Random rotation (with a degree randomly selected in a range of $[-10, 10]$) is applied on the cropped patch to further augment the dataset. The models are then fine-tuned for *50* epochs with an initial learning rate of $\lambda = 3 \times 10^{-3}$, reduced by a factor of *10* after every *25* epochs.

To apply the pretrained 3D-ResNet models on 3 bands in RGB image format to one channel depth images or optical flow images, the depth images are simply converted to 3 bands as RGB image format. For the optical flow images, the pretrained 3D-ResNet models take the x-component, the y-component, and the magnitude of flow as the R, G, and B bands in the RGB format.

[1] A sample video is available http://media-lab.ccny.cuny.edu/wordpress/datecode/.

[2] The ASL-100-RGBD dataset is available via the Databrary platform http://doi.org/10.17910/b7.1062

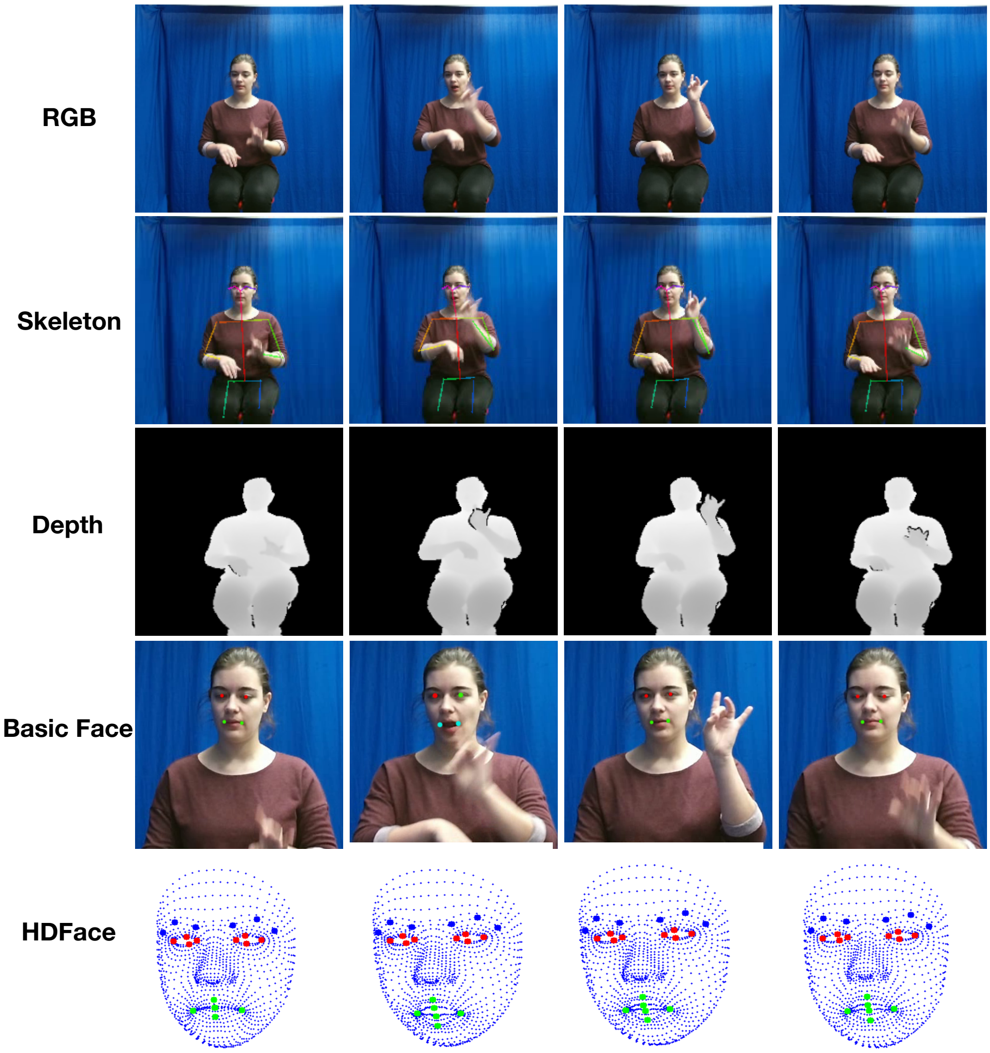


Figure 5: Four sample frames of an ASL sign from our dataset, in different channels including RGB, skeleton joints (25 joints for every frame), depth map, basic face features (5 main face components), and HD Face (1,347 points.)

5.2. Experiments on ASL-100-RGBD

To prepare the training and testing for evaluation of the proposed method on "ASL-100-RGBD" dataset, we first extracted the video clips for each ASL sign. We use $3,250$ ASL clips for training (*75%* of the data) and the remaining *25%* ASL clips for testing. To ensure a subject-independent evaluation, no same signer appears in both training and testing datasets. To augment the data, a new 16-frame proxy video is generated from each video by selecting different subset of frames for each epoch during the training phase. In testing, 16 frames are randomly sampled from the uniformly divided intervals of the entire video and fed to network to obtain the final prediction.

5.2.1. Effects of Data Augmentations

The training dataset which contains $3,250$ ASL video clips of 100 ASL manual signs is relatively small for 3DCNN training and could easily cause an over-fitting problem. To extract more representative temporal dynamics and avoid over-fitting, we applied a random temporal augmentation technique to generate proxy videos for each ASL clip (a new proxy video for each epoch). The ASL recognition results of using the proposed proxy video (16 frames per video) are compared with the traditional method (using the same number of consecutive frames). The network, 3DResNet-34, dose not converge when trained with 16 consecutive frames, while the network trained with proxy video obtained 68.4% on the testing dataset. This is likely due to the majority of movements being from hands in these videos and the consecutive frames could not effectively represent the temporal and spatial information. Therefore, the network could not classify the clips based on only 16 consecutive frames. We also evaluate the effect of random cropping (using a batch size of 112 × 112) and random rotation (with a random number of degrees in a range of $[-10, 10]$).

Table 3 lists the effects of different data augmentation techniques for recognizing 100 ASL signs on only RGB channel. With proxy videos, the 3DCNN model obtains 68.4% accuracy on the testing data for recognizing 100 ASL signs. By adding random cropping, the performance is improved by 4.4% and adding the random rotation further improved the performance to 75.9%. In the following experiments, proxy videos together with random cropping and random rotation are employed to augment the data.

Table 3: The comparison of the performance of different data augmentation methods on only RGB channel with 16 frames for recognizing 100 ASL signs. All the models are pretrained on Kinetics and finetuned on ASL-100-RGBD dataset. The best performance is achieved with random proxy videos, random cropping, and random rotation.

Augmentations	Fusions			
Random Proxy Video	✗	√	√	√
Random Crop	✗		√	√
Random Rotation	✗			√
Performance	Not converging	68.4%	72.8%	**75.9**%

5.2.2. Effects of Network Architectures

In this experiment, the ASL recognition results of different number of layers at 18, 34, 50, and 101 for 3DResNet are compared on full RGB, optical flow, and depth images. As shown in Table 4, the performance of 3DResNet-18, 3DResNet-50, and 3DResNet-101 achieve comparable results on RGB channel. However, the performance on optical flow and depth channels are much lower than that of RGB channel because the network has been pretrained on from Kinetics dataset which contains only RGB images. As shown in Table 4, 3DResNet-34 obtained the best performance for all RGB, optical flow, and depth channels. Hence, 3DResNet-34 is chosen for all the subsequent experiments.

Table 4: The effects of number of layers for 3DResNet with 16 frames on RGB, optical flow, and depth channels. All the models are pretrained on Kinetics and finetuned on ASL-100-RGBD dataset.

Network	RGB (%)	Optical Flow (%)	Depth (%)
3DResNet-18	73.2	61.9	65.0
3DResNet-34	**75.9**	**62.8**	**66.5**
3DResNet-50	72.3	55.4	62.0
3DResNet-101	72.5	55.0	61.5

5.2.3. Effects of Pretrained Models

To evaluate the effects of pretrained models, we fine-tune 3DResNet-34 with pretrained models from the Kinectics [134] and the Chalearn LAP IsoGD datasets [11], respectively. Kinetics dataset consists of RGB videos of diverse human actions which involve different parts of body while the Chalearn LAP IsoGD dataset contains both RGB and depth videos of various hand gestures including mudras (Hindu/ Buddhist hand gestures), Chinese numbers and diving signals, as shown in Fig. 6.

The results are shown in Table 5. The temporal duration is fixed to 16 and the channels are RGB, Depth, and RGBflow. The pretrained models from large datasets such as Kinetics or Chalearn can significantly boost the classification performance for all the modalities because the pretrained models provide prior knowledge as a good starting point for network optimization. In all channels, the performance using the pretrained models from the Chalearn dataset is better than pretrained models from Kinetics dataset. This is probably because all the videos in Chalearn dataset are focused on hand gestures and the network trained on this dataset can learn prior knowledge of hand gestures. The Kinetics dataset consists of general videos from YouTube and the network focuses on the prior knowledge of motions. Therefore, for each channel the pretrained model on the same channel of Chalearn dataset is used in the subsequent experiments.

5.2.4. Effects of Temporal Duration of Proxy Videos

We study the effects of temporal duration (i.e. number of frames used in proxy videos) by finetuning 3DResNet-34 on ASL-100-RGBD dataset with 16, 32, and 64 frames. Note that the same temporal duration is also used to train the corresponding pretrained model on the Chalearn dataset. Results are shown in Table 6. The performance of the network

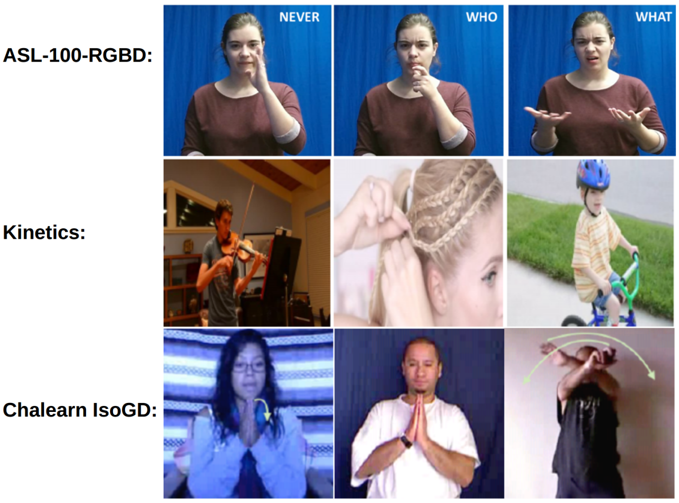


Figure 6: Example images of three datasets. ASL-100-RGBD: various ASL signs. Kinetics dataset: consisting of diverse human actions, involving different parts of body. Chalearn IsoGD: various hands gestures including mudras (Hindu/ Buddhist hand gestures) and diving signals.

Table 5: The comparison of the performance of recognizing 100 ASL signs on 3DResNet-34 trained from scratch and with different pretrained models.

Channels	Scratch (%)	Kinetics (%)	Chalearn (%)
RGB	59.0	75.9	**76.4**
Depth	52.5	66.5	**68.2**
RGB Flow	46.3	62.8	**66.8**

with 64 frames achieves the best performance. Therefore, 3D-ResNet-34 with 64 frames is used in all the following experiments.

Table 6: The comparison of the performance of networks with different temporal duration (i.e. number of frames used in proxy videos). All the models are pretrained on Chalearn dataset and finetuned on ASL-100-RGBD dataset by using the same temporal duration.

Channel	16 frames (%)	32 frames (%)	64 frames (%)
RGB	76.38	80.73	**87.83**
Depth	68.18	74.21	**81.93**
RGB Flow	66.79	71.74	**80.51**

5.2.5. Effects of Different Input Channels

In this section, we examine the fusion results of different input channels. The RGB channel provides global spatial and temporal appearance information. The depth channel provides the distance information, and the optical flow channel captures the motion information. The network is finetuned on the three input channels respectively. The geometric mean fusion is used to obtain the final predictions.

Table 7 shows the performance of ASL recognition on ASL-100-RGBD dataset for each input channel and different fusions. While RGB channel alone achieves *87.83%*, by fusing with optical flow, the performance is boosted up to *89.02%*. With the fusion of all the three channels (RGB, Optical flow, and Depth), the performance is further improved to *89.91%*. This indicates that depth and optical flow channels contain complementary information to RGB channel for ASL recognition.

Table 7: The performance comparison of networks with different input channels and their fusions. All the models are pretrained on Chalearn dataset and finetuned on ASL-100-RGBD dataset with 64 frames.

Channels	Fusions					
RGB	√			√	√	√
Depth		√		√		√
Optical Flow			√		√	√
Performance	87.83%	81.93%	80.51%	**89.91%**	89.02%	89.71

5.2.6. Effects of Different Modalities

We attain further insight into the learned features of the model for RGB channel. In Fig 7 we visualize some examples of the attention maps of the fifth convolution layer on our test dataset generated by the trained RGB 3DCNN model for ASL recognition. These attention maps are computed by averaging the magnitude of activations of convolution layer which reflect the attention of the network. The attention maps show that the model mostly focused on **hands** and **face** of the signer during the ASL recognition process.

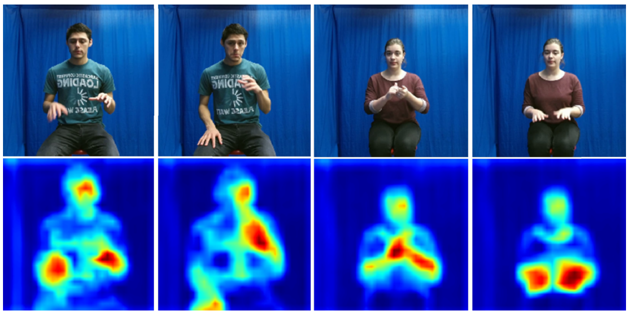


Figure 7: The example RGB images and their corresponding attention maps from the fifth convolution layer of the 3DResNet-34 on the test dataset of ASL-100-RGBD, showing that the hands and face have most of the attention.

Hence, we conduct experiments to analyze the effect of each modality (hand gestures, facial expression, and body poses) with the RGB channel. As shown in Fig. 3, the hand regions and the face regions are obtained from the RGB image based on the location guided

Table 8: The performance comparison of different modalities and their fusions. All the models are pretrained on Chalearn dataset and finetuned on ASL-100-RGBD dataset with 64 frames.

Channels	Fusions			
Body	√		√	√
Hand		√	√	√
Face				√
Performance	87.83%	80.9%	89.81%	**91.5**%

by skeleton joints. The performance of each modality and their fusions are summarized in Table 8.

In addition to the accuracy of ASL sign recognition, we further analyzed the accuracy of the six categories (see Fig. 4 for details) for each modality and their combinations in Table 9. For the categories that involve many facial expressions, such as **Question(Yes/No)** and **Negative**, the accuracy of hand modality is improved by more than 15% after fusion with face modality. For the **Conditional** category which utilizes more subtle facial expressions, the accuracy of hand modality is not improved after fusion with face modality.

Table 9: The performance (%) of different modalities and their fusions on six categories listed in Fig. 4: Conditional (Cond), Negative (Neg), Pointing (Point), Question (WH), Yes/No Question (Y/N) and Time. The last column is the accuracy (%) for ASL signs.

Modalities	Cond	Neg	Point	WH	Y/N	Time	Acc
Hand	90.0	78.1	68.4	84.3	68.4	81.4	80.9
Body	**100.0**	87.4	84.2	88.0	89.5	87.6	87.83
Body+Hand	90.9	86.6	**89.5**	88.7	**94.7**	90.2	89.81
Body+Hand+Face	90.9	**93.3**	84.2	**90.6**	84.2	**91.8**	**91.5**

5.2.7. Comparison of Different Fusion Methods

Various fusion methods have been used for video understanding tasks including average fusion, geometric mean fusion, jointly end-to-end training, and sparse fusion method. The average fusion method calculates the average of predictions as final prediction from predictions of multiple channels, and the weights for each channel can be adjusted based on the importance of each channel. The geometric mean fusion method calculates the geometric mean of predictions of all channels. These two fusion methods are widely used for video action recognition task due to their simplicity and effectiveness. The sparse fusion method is proposed to use a small neural network to learn how much each channel contributes to each class and the weighted score is used as the final prediction, and the jointly training fusion method trains all the networks together to jointly optimize them.

In this section, we study the effects of different fusion methods and report the performance of all the four fusion methods in Table 10. Among all these fusion methods, the geometric mean fusion method outperforms the other three fusion methods. Therefore, the geometric mean fusion method is employed for all the experiments in the paper.

Table 10: Results of different fusion methods on ASL-100-RGBD dataset by using five channels including RGB, RGB Flow, Depth, Cropped Hands, and Cropped Face.

Fusion Method	Accuracy (%)
Jointly Training	89.51%
Sparse Fusion	90.29%
Average Fusion	91.29%
Geometric Mean Fusion	**92.58%**

5.2.8. *Fusions of Different Channels and Modalities*

The fusion results of different input channels and modalities on ASL-100-RGBD dataset are shown in Table 11. The experiments are based on 3DResNet-34 with 64 frames, pre-trained on Chalearn dataset. Among all the models, fusion of **RGB+Depth+Hands RGB+ Face RGB** achieves the best performance with 92.88% accuracy. Adding RGBflow to this combination results in 92.48% accuracy which is comparable but not improved since the channels have redundant information.

Table 11: Performance of 3DResNet-34 with 64 frames with fusion of different channels and modalities.

Channels	Fusions			
RGB	√	√		√
Depth	√	√	√	√
RGBflow	√	√	√	
RGB of Hands	√	√	√	√
RGB of Face		√	√	√
Performance	91.19%	92.48%	92.48%	**92.88%**

5.3. *Experiments on Chalearn LAP IsoGD dataset*

5.3.1. *Effects of Network Architectures*

The 3D-ResNet is pretrained on Kinetics [134] for all the experiments in this section. To find the best network architecture for Chalearn dataset, the parameters of 3D-ResNet are studied on RGB videos. The results are shown in Table 12. By changing the number of layers to 18, 34, 50 while fixing the temporal duration to 32, ResNet-34 achieved the best accuracy.

We also evaluated the performance of ResNet-34 with different temporal duration of the proxy videos by using 16, 32, and 64 frames. Our results indicate that ResNet-34 with 64 frames has the best performance for Chalearn dataset, as shown in Table 13.

Table 12: Ablation study of number of layers of the network on RGB videos of Chalearn Dataset.

Network	Temporal Duration	Accuracy
ResNet-18	32	52.69%
ResNet-34	**32**	**56.28%**
ResNet-50	32	54.57%

Table 13: Ablation study of temporal duration of proxy videos on RGB channel of Chalearn Dataset.

Network	Temporal Duration	Accuracy
ResNet-34	16	45.00%
ResNet-34	32	56.28%
ResNet-34	**64**	**58.32%**

5.3.2. Effects of Different Channels and Modalities

We evaluate the effects of different channels including RGB, RGB flow, Depth, and Depth flow. Because the Chalearn dataset is designed for hand gesture recognition, we further analyze the effects of different hands (left and right), as well as the whole body. We develop a method to distinguish left and right hands in Chalearn Isolated Gesture dataset, and will release the coordinates of hands (distinguished between right and left hands) with the publication of this article. Since the Chalearn dataset is collected for recognizing hand gestures, here, the face channel is not employed.

We train 12 3D-ResNet-34 networks with 64 frames by using different combinations of channels and modalities respectively and show the results in Table 14. The accuracy of right hand is significantly higher than the left hand. The reason is that for most of the gestures in Chalearn dataset, the right hand is dominant and the left hand does not move much for many hand gestures.

Table 14: Performance of 3D-ResNet-34 with 64 frames on Chalearn Dataset for different channels and modalities.

Channel	Global Channel (%)	Left Hand (%)	Right Hand (%)
RGB	58.32	18.01	48.58
Depth	**63.16**	19.43	**54.15**
RGB Flow	60.26	**21.97**	48.79
Depth Flow	55.37	20.28	47.07

5.3.3. Effects of Fusions on Channels and Modalities

Here we analyze the effects of fusing different channels and modalities. The results are shown in Table 15. Using only RGB and depth channels, the accuracy is 67.58% which is improved to 69.97% by adding RGB flow. We observe that among all different triplets of channels, *Right Hand RGB + Depth + RGBflow* has the highest accuracy at 73.32%. By applying the geometric mean fusion on four channels *RGB+ RGBflow+ Right Hand RGB + Right Hand Depth*, our model achieves the accuracy about 75.88% which outperforms all previous work on Chalearn dataset. In the-state-of-the-art work of [135], the accuracy of average fusion is 71.93% for 7 channels and 70.37% for 12 channels, respectively.

Finally, the geometric mean fusion of all global channels (RGB, RGB flow, Depth, Depth flow) and Right Hand channels (Right Hand RGB, Right Hand RGB flow, Right Hand Depth, Right Hand Depth flow) resulted in 76.04% accuracy and the accuracy of 12 channels together resulted in 75.68%. This means that the 12 channels contain redundant information, and adding more channels does not necessarily improve the results.

Table 15: Performance of 3DResNet-34 with 64 frames for fusion of different channels and modalities on Chalearn dataset.

Channels	Fusions				
RGB	√	√		√	√
Depth	√	√	√		√
RGBflow		√	√	√	√
RGB of Right Hand			√	√	√
Depth of Right Hand				√	√
Performance	67.58%	69.97%	73.32%	75.53%	**75.88%**

Table 16: Comparison with the State-of-the-art Results on Chalearn IsoGD Dataset.

Framework	Accuracy on Test Set (%)
Our Results	**76.04**
MEMP (3DCNN + LSTM) [136]	**78.85**
MultiD-CNN [137]	72.53
MEMP (3DCNN) [136]	71.24
FOANet (Average Fusion) [135]	70.37
Lin et al. [138]	68.42
Chen et al. [139]	68.15
Duan et al. [140]	67.26
Miao et al. [141]	67.71
CAPF [142]	66.79
Zhou et al [143]	66.62
Wang et al. [144]	65.59
Zhang et al. [145]	60.47
Wang et al. [146]	59.21
Santos et al. [147]	52.18

5.3.4. Comparison with the-state-of-the-arts

Our framework achieves accuracy of 75.88% and 76.04% from the fusion of 5 and 8 channels, respectively, on Chalearn IsoGD dataset. Table 16 lists the state-of-the-art results from Chalearn IsoGD competition 2017. As shown in the table, our framework achieves comparable results to the-state-of-the-art methods.

MEMP [136] achieves a slightly higher results, 78.85%, by combining 3DCNNs with LSTMs. However, the performance of MEMP [136] drops 5% below our results when LSTMs are not employed. Rastgoo et al. in [148] improved the performance to 86.1% by exploiting additional information such as 3D hand keypoints. It is worth noting that FOANet [135] reported the accuracy of 82.07% by applying *Sparse Fusion* on the softmax scores of 12 channels (combinations of right hand, left hand, and whole body while each has 4 channels of RGB, Depth, RGBflow and Depthflow). The purpose of using sparse fusion is to learn which channels are important for each gesture. The accuracy of FOANet framework using average fusion is 70.37% which is around 6% lower than our results and nearly 12% lower than the accuracy of sparse fusion. While the authors of FOANet [135] had reported a 12% boost from using sparse fusion in their original experiments, our experiments do not reveal such a boost when implementing a system following the technical details provided in [135].

Table 17 lists the accuracy on individual channels of our network and FOANet [135]. In this table, the values inside the parenthesis represent the accuracy of FOANet. As shown in the table, in the Global channel, our framework outperforms FOANet in all the four channels by 10% to 25%. Also, for the RGB of Right Hand, we obtain a comparable accuracy (48%) as FOANet. However, FOANet is outperforming our results in the Right Hand for Depth, RGBflow, and Depthflow by nearly 10%. From our experiments, the performance of "Global" channels (whole body) in general is superior to the Local channels (Right/ Left Hand) because the Global channels include more information. By using the similar architecture, FOANet reported 64% accuracy from Depth of Right Hand and 38% from Depth of the entire frame. Instead, our framework achieves more consistent results. For example, in our framework the accuracy of Depth channel is higher than RGB and RGBflow for both Global and Right Hand, while the accuracy in FOANet for Depth and RGB are almost the same in the Global channel (around 40%) but very different in the Right Hand channel (17% difference.)

Table 17: The accuracy (%) of 12 channels on the test set of Chalearn IsoGD Dataset. Comparison between our framework and FOANet [135]. The bold numbers show the best results.

Channel	Global Channel (%)		Left Hand (%)		Right Hand (%)	
Method	Ours	FOANet	Ours	FOANet	Ours	FOANet
RGB	**58.32**	41.27	**18.01**	16.63	**48.58**	47.41
Depth	**63.16**	38.50	19.43	**24.06**	54.15	**64.44**
RGB Flow	**60.26**	50.96	21.97	**24.02**	48.79	**59.69**
Depth Flow	**55.37**	42.02	20.28	**22.71**	47.07	**58.79**

5.4. Efficiency Analysis

One major advantage of our proposed method is that it is efficient and runs in real-time. During the training phase, a small proxy clip sampled for each gesture clip is used to train the network. During testing, the prediction of each gesture clip is obtained by feeding its proxy video to the network in one pass. The performance and computation time of our proposed framework with 3DResNet-34 on different input channels on the Chalearn IsoGD testing set using a single NVIIDA PASCAL GPU are reported in Table 18. Our proposed framework runs 432 frames per second by using 6 channels input channels including RGB, RGB Flow, Depth, Cropped Left Hand, Cropped Right Hand, and Cropped Face which demonstrate the potential for real-time ASL recognition application. Table 19 reports the computational complexity of our model, 3D-ResNet34, with varying temporal durations, in terms of floating-point operations (FLOPs) on the RGB channel and whole-body modality of the ChaLearn IsoGD Dataset. As the table demonstrates, increasing the temporal duration improves accuracy but also leads to higher computational complexity.

Table 18: The speed analysis of the proposed network on the Chalearn IsoGD dataset. The channels are RGB, Depth, RGB Flow of whole body and the right hand.

# Channels	Accuracy (%)	FPS
4	75.53	650
5	75.88	537
6	**76.04**	432

Table 19: The computational complexity of 3D-ResNet34, with varying temporal durations, in terms of floating-point operations (FLOPs) on the RGB channel of the ChaLearn IsoGD Dataset.

Temporal Duration	Accuracy (%)	FLOPs(1×10^9)
16	45.0%	69.7
32	56.3%	133.8
64	58.3%	260.9

6. Conclusions

In this paper, we have proposed a 3DCNN-based multi-channel and multi-modal framework, which learns complementary information and embeds the temporal dynamics in videos to recognize ASL signs from RGB-D videos. To validate our proposed method, we collaborate with ASL experts to collect an ASL dataset of 100 manual signs including both hand gestures and facial expressions with full annotation on the sign labels and temporal boundaries (starting and ending points.) A Proxy video generation method is integrated with our framework to capture both spatial and temporal information of the entire gesture. The experimental results on our ASL-100-RGBD and Chalearn IsoGD datasets have demonstrated the effectiveness and efficiency of the proposed framework.

This technology for identifying the appearance of specific ASL signs has valuable applications for technologies that can benefit people who are DHH [29, 31, 30, 27, 43, 149, 150]. Our "ASL-100-RGBD" dataset together with the annotation is available to the research community to use this resource for training or evaluation of models for ASL recognition.

7. Acknowledgment

This material is based upon work supported by the National Science Foundation under award numbers IIS-1400802, IIS-1400810, IIS-1462280, and IIS-2041307.

References

[1] C. Valli, C. Lucas, K. J. Mulrooney, M. Villanueva, Linguistics of American Sign Language: An Introduction, Gallaudet University Press, 2011.

[2] American deaf and hard of hearing statistics, https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing.

[3] R. E. Mitchell, T. A. Young, B. Bachleda, M. A. Karchmer, How many people use asl in the united states? why estimates need updating, Sign Language Studies 6 (3) (2006) 306–335.

[4] K. Mulrooney, American Sign Language Demystified, Hard Stuff Made Easy, McGraw Hill, 2010.

[5] C. Neidle, A. Thangali, S. Sclaroff, Challenges in development of the american sign language lexicon video dataset (asllvd) corpus, in: Proceedings of the Language Resources and Evaluation Conference (LREC), 2012.

[6] D. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, C. Neidle, Recognition of nonmanual markers in asl using non-parametric adaptive 2d-3d face tracking, in: Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association, 2012.

[7] C. B. Traxler, The stanford achievement test: National norming and performance standards for deaf and hard-of-hearing students, Journal of deaf studies and deaf education 5 (4) (2000) 337–348.

[8] N. Furman, D. Goldberg, N. Lusin, Enrollments in languages other than english in united states institutions of higher education, fall 2010, Retrieved from http://www.mla.org/2009_enrollmentsurvey.

[9] M. Huenerfauth, E. Gale, B. Penly, S. Pillutla, M. Willard, D. Hariharan, Evaluation of language feedback methods for student videos of american sign language, ACM Transactions on Accessible Computing (TACCESS) 10 (1) (2017) 2.

[10] S. Hassan, L. Berke, E. Vahdani, L. Jing, Y. Tian, M. Huenerfauth, An isolated-signing rgbd dataset of 100 american sign language signs produced by fluent asl signers, in: In Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, 2020.

[11] J. Wan, S. Li, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition, in: Proceedings of CVPR 2008 Workshops, IEEE, 2016.

[12] S. Tamura, S. Kawasaki, Recognition of sign language motion images, Pattern Recognition 21 (4) (1988) 343–353.

[13] M. Kadous, Machine recognition of auslan signs using powergloves:towards large-lexicon recognition of sign language, in: Proceedings of the Workshop on the Integration of Gesture in Language and Speech, 1996, pp. 165–174.

[14] R.-H. Liang, M. Ouhyoung, A real-time continuous gesture recognition system for sign language, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 558–567.

[15] G. Fang, W. Gao, D. Zhao, Large-vocabulary continuoius sign language recognition based on transition-movement models, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 37 (1).

[16] W. Kong, S. Ranganath, Towards subject independent continues sign language recognition: A segment and merge approach, Pattern Recognition 47 (3) (2014) 1294–1308.

[17] C. Zhang, Y. Tian, M. Huenerfauth, Multi-modality american sign language recognition, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2016.

[18] T. Starner, J. Weaver, A. Pentland, Real-time american sign language recognition using desk and wearable computer based video, IEEE Pattern Analysis and Machine Intelligence 20 (12) (1998) 1371–1375.

[19] H. Yang, S. Sclaroff, S. Lee, Sign language spotting with a threshold model based on conditional random fields, IEEE Pattern Analysis and Machine Intelligence 31 (7) (2009) 1264–1277.

[20] R. Yang, S. Sarkar, B. Loeding, Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming, IEEE Pattern Analysis and Machine Intelligence 32 (3) (2010) 462–477.

[21] D. Kelly, J. McDonald, C. Markham, A person independent system for recognition of hand postures used in sign language, Pattern Recognition Letters 31 (11) (2010) 1359–1368.

[22] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, J. Dambre, Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video, International Journal of Computer Vision 126 (2-4) (2018) 430–439.

[23] L. Pigou, S. Dieleman, P.-J. Kindermans, B. Schrauwen, Sign language recognition using convolutional neural networks, in: Proceedings of European Conference on Computer Vision Workshops, 2014, pp. 572–578.

[24] J. Huang, W. Zhou, Q. Zhang, H. Li, W. Li, Video-based sign language recognition without temporal segmentation, arXiv preprint arXiv:1801.10111.

[25] J. Pu, W. Zhou, H. Li, Dilated convolutional network with iterative optimization for continuous sign language recognition, in: IJCAI, 2018, pp. 885–891.

[26] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, CVPR 2018 Proceedings.

[27] L. Pigou, M. Van Herreweghe, J. Dambre, Gesture and sign language recognition with temporal residual networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3086–3093.

[28] R. Cui, H. Liu, C. Zhang, Recurrent convolutional neural networks for continuous sign language recognition by staged optimization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[29] N. C. Camgöz, S. Hadfield, O. Koller, R. Bowden, Subunets: End-to-end hand shape and continuous sign language recognition., in: ICCV, Vol. 1, 2017.

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

[31] O. Koller, H. Ney, R. Bowden, Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3793–3802.

[32] Z. Liu, F. Huang, G. W. L. Tang, F. Y. B. Sze, J. Qin, X. Wang, Q. Xu, Real-time sign language recognition with guided deep convolutional neural networks, in: Proceedings of the 2016 Symposium on Spatial User Interaction, ACM, 2016, pp. 187–187.

[33] S. Gattupalli, A. Ghaderi, V. Athitsos, Evaluation of deep learning based pose estimation for sign language recognition, in: Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments, ACM, 2016, p. 12.

[34] O. Koller, S. Zargaran, H. Ney, R. Bowden, Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms, International Journal of Computer Vision 126 (12) (2018) 1311–1325.

[35] J. Charles, T. Pfister, M. Everingham, A. Zisserman, Automatic and efficient human pose estimation for sign language videos, International Journal of Computer Vision 110 (1) (2014) 70–90.

[36] Y. Ye, Y. Tian, M. Huenerfauth, Recognizing american sign language gestures from within continu-

ous videos, The 8th IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG) in conjunction with CVPR 2018.

[37] S. Zhang, Q. Zhang, Sign language recognition based on global-local attention, Journal of Visual Communication and Image Representation 80 (2021) 103280.

[38] K. Sadeddine, F. Z. Chelali, R. Djeradi, A. Djeradi, S. Benabderrahmane, Recognition of user-dependent and independent static hand gestures: Application to sign language, Journal of Visual Communication and Image Representation 79 (2021) 103193.

[39] J. Zheng, Y. Wang, C. Tan, S. Li, G. Wang, J. Xia, Y. Chen, S. Z. Li, Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23141–23150.

[40] L. Hu, L. Gao, Z. Liu, W. Feng, Continuous sign language recognition with correlation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2529–2539.

[41] L. Guo, W. Xue, Q. Guo, B. Liu, K. Zhang, T. Yuan, S. Chen, Distilling cross-temporal contexts for continuous sign language recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10771–10780.

[42] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. N. Metaxas, C. Neidle, Recognizing eyebrow and periodic head gestures using crfs for non-manual grammatical marker detection in asl, in: Proc. of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013.

[43] P. Kumar, P. P. Roy, D. P. Dogra, Independent bayesian classifier combination based sign language recognition using facial expression, Information Sciences 428 (2018) 30–48.

[44] U. von Agris, M. Knorr, K.-F. Kraiss, The significance of facial features for automatic sign language recognition, in: Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition, 2008.

[45] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, M. R. Morris, Sign language recognition, generation, and translation: An interdisciplinary perspective, in: In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19), 2019.

[46] S. Ong, S. C.and Ranganath, Automatic sign language analysis: A survey and the future beyond lexical meaning, IEEE Pattern Analysis and Machine Intelligence 27 (6) (2005) 873–891.

[47] A. Er-Rady, R. O. H. Thami, R. Faizi, H. Housni, Automatic sign language recognition: A survey, in: Proceedings of the 3rd International Conference on Advanced Technologies for Signal and Image Processing, 2017.

[48] R. Rastgoo, K. Kiani, S. Escalera, Sign language recognition: A deep survey, Expert Systems with Applications 164 (2021) 113794.

[49] M. C. Ariesta, F. Wiryana, G. P. Kusuma, et al., A survey of hand gesture recognition methods in sign language recognition., Pertanika Journal of Science & Technology 26 (4).

[50] O. Koller, Quantitative survey of the state of the art in sign language recognition, arXiv preprint arXiv:2008.09918.

[51] P. Barve, N. Mutha, A. Kulkarni, Y. Nigudkar, Y. Robert, Application of deep learning techniques on sign language recognition—a survey, Data Management, Analytics and Innovation (2021) 211–227.

[52] R. Minu, et al., A extensive survey on sign language recognition methods, in: 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2023, pp. 613–619.

[53] Z. Liang, H. Li, J. Chai, Sign language translation: A survey of approaches and techniques, Electronics 12 (12) (2023) 2678.

[54] Set up kinect for windows v2 or an xbox kinect sensor with kinect adapter for windows, https://support.xbox.com/en-US/xbox-on-windows/accessories/kinect-for-windows-v2-setup.

[55] Intel realsense technology: Observe the world in 3d, https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html.

[56] Orbbec astra, https://orbbec3d.com/product-astra/.

[57] N. Pugeault, R. Bowden, Spelling it out: Real-time asl fingerspelling recognition, in: Proc. of IEEE International Conference on Computer Vision Workshops, 2011, pp. 1114–1119.

[58] Z. Zafrulla, H. Brashear, T. Starner, P. Hamilton, H.and Presti, American sign language recognition with the kinect, in: In Proceedings of the International Conference on Multimodal Interfaces, 2011, pp. 279–286.

[59] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, M. Zhou, Sign language recognition and translation with kinect, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 2013.

[60] Z. Ren, J. Yuan, J. Meng, Z. Zhang, Robust part-based hand gesture recognition using kinect sensor, IEEE Trans. on Multimedia 15 (2013) 1110–1120.

[61] Y. Jiang, J. Tao, Y. Weiquan, W. Wang, Z. Ye, An isolated sign language recognition system using rgb-d sensor with sparse coding, in: Proceedings of IEEE 17th International Conference on Computational Science and Engineering, 2014.

[62] S. G. M. Almeidaab, F. G. Guimarãesc, J. Ramírez, Feature extraction in brazilian sign language recognition based on phonological structure and using rgb-d sensors, Expert Systems with Applications 41 (16) (2014) 7259–7271.

[63] H.-D. Yang, Sign language recognition with the kinect sensor based on conditional random fields, Sensors 15 (2015) 135–147.

[64] P. Buehler, M. Everingham, D. P. Huttenlocher, A. Zisserman, Upper body detection and tracking in extended signing sequences, International journal of computer vision 95 (2) (2011) 180.

[65] N. Naz, H. Sajid, S. Ali, O. Hasan, M. K. Ehsan, Signgraph: An efficient and accurate pose-based graph convolution approach toward sign language recognition, IEEE Access 11 (2023) 19135–19147.

[66] A. A. SK, P. MVD, K. PVV, et al., Pose based multi view sign language recognition through deep feature embedding., International Journal of Intelligent Engineering & Systems 16 (3).

[67] C. Keskin, F. Kıraç, Y. Kara, L. Akarun, Hand pose estimation and hand shape classification using multi-layered randomized decision forests, in: In Proceedings of the European Conference on Computer Vision, 2012, pp. 852–863.

[68] S. Lang, M. Block, R. Rojas, Sign language recognition using kinect, in: In Proceedings of International Conference on Artificial Intelligence and Soft Computing, 2012, pp. 394–402.

[69] K. Mehrotra, A. Godbole, S. Belhe, Indian sign language recognition using kinect sensor, in: In Proceedings of the International Conference Image Analysis and Recognition, 2015, pp. 528–535.

[70] P. Kumar, H. Gauba, P. P. Roy, D. P. Dogra, A multimodal framework for sensor based sign language recognition, Neurocomputing 259 (2017) 21–38.

[71] O. Koller, J. Forster, H. Ney, Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers, Computer Vision and Image Understanding 141 (2015) 108–125.

[72] E. J. E. Cardenas, G. C. Chavez, Multimodal hand gesture recognition combining temporal and pose information based on cnn descriptors and histogram of cumulative magnitudes, Journal of Visual Communication and Image Representation 71 (2020) 102772.

[73] S. Ameur, A. B. Khalifa, M. S. Bouhlel, Chronological pattern indexing: An efficient feature extraction method for hand gesture recognition with leap motion, Journal of Visual Communication and Image Representation 70 (2020) 102842.

[74] L. Ding, Y. Wang, R. Laganière, D. Huang, S. Fu, A cnn model for real time hand pose estimation, Journal of Visual Communication and Image Representation 79 (2021) 103200.

[75] T. P. Moreira, D. Menotti, H. Pedrini, Video action recognition based on visual rhythm representation, Journal of Visual Communication and Image Representation 71 (2020) 102771.

[76] L. Jing, X. Yang, Y. Tian, Video you only look once: Overall temporal convolutions for action recognition, Journal of Visual Communication and Image Representation 52 (2018) 58–65.

[77] L. Song, G. Yu, J. Yuan, Z. Liu, Human pose estimation and its application to action recognition: A survey, Journal of Visual Communication and Image Representation (2021) 103055.

[78] H. Deng, J. Kong, M. Jiang, T. Liu, Diverse features fusion network for video-based action recognition,

Journal of Visual Communication and Image Representation 77 (2021) 103121.

[79] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, Y.-G. Jiang, Svformer: Semi-supervised video transformer for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18816–18826.

[80] F. Sato, R. Hachiuma, T. Sekii, Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6471–6480.

[81] I. R. Dave, C. Chen, M. Shah, Spact: Self-supervised privacy preservation for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20164–20173.

[82] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[83] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv:1310.1531.

[84] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, arXiv preprint arXiv:1409.4842.

[85] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE, 2014, pp. 580–587.

[86] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 346–361.

[87] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[88] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, arXiv preprint arXiv:1412.2306.

[89] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, T. Tuytelaars, Rank pooling for action recognition, IEEE transactions on Pattern Analysis and Machine Intelligence 39 (4) (2017) 773–787.

[90] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: CVPR, 2014.

[91] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[92] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.

[93] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. Mahdi Arzani, R. Yousefzadeh, L. Van Gool, Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification, ArXiv e-printsarXiv:1711.08200.

[94] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6546–6555.

[95] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (1) (2013) 221–231.

[96] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in: The IEEE International Conference on Computer Vision (ICCV), 2017.

[97] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[98] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

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

[100] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[101] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional lstm and other neural network architectures, Neural networks 18 (5-6) (2005) 602–610.

[102] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, Y. Fu, Skeleton aware multi-modal sign language recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3413–3423.

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

[104] H. Hu, W. Zhou, H. Li, Hand-model-aware sign language recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1558–1566.

[105] M. Boháček, M. Hrúz, Sign pose-based transformer for word-level sign language recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 182–191.

[106] X. Han, F. Lu, J. Yin, G. Tian, J. Liu, Sign language recognition based on r (2+ 1) d with spatial–temporal–channel attention, IEEE Transactions on Human-Machine Systems.

[107] Y. C. Bilge, R. G. Cinbis, N. Ikizler-Cinbis, Towards zero-shot sign language recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[108] C. Neidle, C. Vogler, A new web interface to facilitate access to corpora: Development of the asllrp data access interface (dai), in: Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC, 2012.

[109] P. Lu, M. Huenerfauth, Cuny american sign language motion-capture corpus: first release, in: Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, 2012.

[110] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, H. Ney, Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus, in: LREC, 2012, pp. 3785–3789.

[111] P. Dreuw, J. Forster, H. Ney, Tracking benchmark databases for video-based sign language recognition, in: Proc. ECCV International Workshop on Sign, Gesture, and Activity, 2010.

[112] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, A. Thangali, The asl lexicon video dataset, in: Proceedings of CVPR 2008 Workshop on Human Communicative Behaviour Analysis, IEEE, 2008.

[113] A. M. Martínez, R. B. Wilbur, R. Shay, A. C. Kak, The rvl-slll asl database, in: Proc. of IEEE International Conference Multimodal Interfaces, 2002.

[114] Z. S. Sehyr, N. Caselli, A. M. Cohen-Goldberg, K. Emmorey, The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language, The Journal of Deaf Studies and Deaf Education 26 (2) (2021) 263–277.

[115] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, K. Emmorey, Asl-lex: A lexical database of american sign language, Behavior research methods 49 (2) (2017) 784–801.

[116] D. Li, C. Rodriguez, X. Yu, H. Li, Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1459–1469.

[117] H. R. V. Joze, O. Koller, Ms-asl: A large-scale data set and benchmark for understanding american sign language, arXiv preprint arXiv:1812.01053.

[118] P. Dreuw, D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth, H. Ney, Spoken language processing techniques for sign language recognition and translation, Technology and Disability 20 (2) (2008) 121–133.

[119] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, H. Ney, Speech recognition techniques for a sign

language recognition system, hand 60 (2007) 80.

[120] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, H. Ney, Benchmark databases for video-based automatic sign language recognition, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008.

[121] H. Brashear, V. Henderson, K.-H. Park, H. Hamilton, S. Lee, T. Starner, American sign language recognition in game development for deaf children, in: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, 2006, pp. 79–86.

[122] O. Özdemir, A. A. Kındıroğlu, N. Cihan Camgoz, L. Akarun, BosphorusSign22k Sign Language Recognition Dataset, in: Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives, 2020.

[123] O. M. Sincan, H. Y. Keles, Autsl: A large scale multi-modal turkish sign language dataset and baseline methods, IEEE Access 8 (2020) 181340–181355.

[124] J. Zhang, W. Zhou, C. Xie, J. Pu, H. Li, Chinese sign language recognition with adaptive hmm, in: 2016 IEEE international conference on multimedia and expo (ICME), IEEE, 2016, pp. 1–6.

[125] E. Efthimiou, K. Vasilaki, S.-E. Fotinea, A. Vacalopoulou, T. Goulas, A.-L. Dimou, The polytropon parallel corpus, in: sign-lang@ LREC 2018, European Language Resources Association (ELRA), 2018, pp. 39–44.

[126] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, P. Daras, A comprehensive study on deep learning-based methods for sign language recognition, IEEE Transactions on Multimedia 24 (2021) 1750–1762.

[127] U. Von Agris, M. Knorr, K.-F. Kraiss, The significance of facial features for automatic sign language recognition, in: 2008 8th IEEE international conference on automatic face & gesture recognition, IEEE, 2008, pp. 1–6.

[128] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, et al., Bbc-oxford british sign language dataset, arXiv preprint arXiv:2111.03635.

[129] I. Guyon, V. Athitsos, P. Jangyodsuk, H. Escalante, The chalearn gesture dataset (cgd 2011), Machine Vision and Applications 25 (8) (2014) 1929–1951.

[130] N. C. Camgöz, A. A. Kındıroğlu, S. Karabüklü, M. Kelepir, A. S. Özsoy, L. Akarun, BosphorusSign: a Turkish sign language recognition corpus in health and finance domains, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 1383–1388.

[131] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, X. Giro-i Nieto, How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[132] S. Hassan, M. Seita, L. Berke, Y. Tian, E. Gale, S. Lee, M. Huenerfauth, Asl-homework-rgbd dataset: An annotated dataset of 45 fluent and non-fluent signers performing american sign language homeworks, in: In Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, 2022.

[133] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 4724–4733.

[134] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950.

[135] P. Narayana, J. R. Beveridge, B. A. Draper, Gesture recognition: Focus on the hands, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5235–5244.

[136] X. Zhang, X. Li, Dynamic gesture recognition based on memp network, Future Internet 11 (4) (2019) 91.

[137] A. Elboushaki, R. Hannane, K. Afdel, L. Koutti, 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 (2020) 112829.

[138] C. Lin, J. Wan, Y. Liang, S. Z. Li, Large-scale isolated gesture recognition using a refined fused model

based on masked res-c3d network and skeleton lstm, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 52–58.

[139] H. Chen, Y. Li, H. Fang, W. Xin, Z. Lu, Q. Miao, Multi-scale attention 3d convolutional network for multimodal gesture recognition, Sensors 22 (6) (2022) 2405.

[140] J. Duan, J. Wan, S. Zhou, X. Guo, S. Z. Li, A unified framework for multi-modal isolated gesture recognition, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (1s) (2018) 21.

[141] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, X. Cao, Z. Liu, X. Chai, Z. Liu, et al., Multimodal gesture recognition based on the resc3d network., in: ICCV Workshops, 2017, pp. 3047–3055.

[142] B. Zhou, P. Wang, J. Wan, Y. Liang, F. Wang, D. Zhang, Z. Lei, H. Li, R. Jin, Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20154–20163.

[143] B. Zhou, Y. Li, J. Wan, Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3563–3571.

[144] H. Wang, P. Wang, Z. Song, W. Li, Large-scale multimodal gesture recognition using heterogeneous networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3129–3137.

[145] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, M. Bennamoun, Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3120–3128.

[146] P. Wang, W. Li, Z. Gao, C. Tang, P. O. Ogunbona, Depth pooling based large-scale 3-d action recognition with convolutional neural networks, IEEE Transactions on Multimedia 20 (5) (2018) 1051–1061.

[147] C. C. dos Santos, J. L. A. Samatelo, R. F. Vassallo, Dynamic gesture recognition by using cnns and star rgb: A temporal information condensation, Neurocomputing 400 (2020) 238–254.

[148] R. Rastgoo, K. Kiani, S. Escalera, Real-time isolated hand sign language recognition using deep networks and svd, Journal of Ambient Intelligence and Humanized Computing 13 (1) (2022) 591–611.

[149] M. Palmeri, F. Vella, I. Infantino, S. Gaglio, Sign languages recognition based on neural network architecture, in: International Conference on Intelligent Interactive Multimedia Systems and Services, Springer, 2017, pp. 109–118.

[150] W. Liu, Y. Fan, Z. Li, Z. Zhang, Rgbd video based human hand trajectory tracking and gesture recognition system, Mathematical Problems in Engineering 2015.