# Recent Advances of Deep Learning for Sign Language Recognition

1st Lihong Zheng
*School of Computing and Mathematics*
*Charles Sturt University*
Wagga Wagga, Australia
lzheng@csu.edu.au

2nd Bin Liang
*School of Computing and Mathematics*
*Charles Sturt University*
Wagga Wagga, Australia
bliang@csu.edu.au

3rd Ailian Jiang
*Dept. of Computer Science and Technology*
*Taiyuan University of Technology*
Taiyuan, China
ailianjiang@126.com

***Abstract*—To assist the social interaction of deaf and hearing impaired people, efficient interactive communication tools is expected. With the growing research interest in action and gesture recognition in the recent years, many successful applications for sign language recognition comprise new types of sensors including low-cost depth camera and advanced machine learning technologies. In this paper, we present a complete overview of deep learning based methodologies for sign language recognition. We discuss various types of such approaches designed for the recognition from viewpoints of available modalities provided by depth sensors, feature extraction and classification. In addition, we summarise the currently available datasets of sign language, including gestures of finger spelling and vocabulary words, which can be used as an assessing tool for those people who are learning sign languages. We then discuss the main current research works with particular interest on how they treat the different types of data, discussing their main features and identify opportunities and challenges for future research.**

***Index Terms*—deep learning, classification, feature extraction, sign language recognition, depth image**

## I. Introduction

Sign language is the primary language used by people with impaired hearing and speech. People use sign language gestures as a means of non-verbal communication to express their thoughts and emotions. But non-signers find it extremely difficult to understand, hence trained sign language interpreters are needed during medical and legal appointments, educational and training sessions. Over the past five years, there has been an increasing demand for interpreting services. Other means, such as video remote human interpreting using high-speed internet connections, have been introduced. They will thus provide an easy to use sign language interpreting service, which can be used, but has major limitations.

Auslan is the primary language used by the Australian Deaf Community, with an estimated 30,000 users in Australia. Of these 9700 are also speech-impaired, and, it is this group, which needs professional translation services. A less well-documented, but potentially large group, comprises people who have perfect language functioning but through injury or disease, are unable to speak. Sign language translation to text or speech (through using speech synthesisers) would give them everyday communication. Professional Auslan interpreters can help them in such situations. However, given a recognised shortage of Auslan interpreters, particularly in regional and rural areas, an urgent need for developing efficient interactive communication tools for signers with help of computers is to be met. With the wishes of increasing the ability of deaf and hearing impaired people to access services in their local community without engaging the use of a professional sign language interpreter. Such computer-aid systems can satisfy the unmet demand for professional interpreting services and significantly improve the quality of life. It will also be useful for sign language training and for people who are unable to speak through sickness or injury but have unimpaired manual dexterity. We believe that it will benefit deaf and hearing impaired people by offering them a flexible interpreting alternative when face-to-face interpreting is not available.

The purpose of this paper is to provide a complete overview of deep learning based methodologies for sign language recognition. We discuss various types of such approaches designed for the recognition from viewpoints of available modalities provided by depth sensors, feature extraction and classification. In addition, we also summarise the current available datasets of sign language, including gestures of finger spelling and vocabulary words. Thus this paper will investigate the technical challenges of real-time recognition of sign languages.

The paper is organised as follows. Section II gives a brief introduction of sign language recognition. Section III summarizes the various sensors. Section IV lists the current datasets of sign language. The current research works have been discussed in Section V. Finally section VI is conclusion and future work.

## II. Framework of Sign Language Recognition

Sign Language Recognition (SLR) has been studied in the decades since Human Computer Interaction (HCI) became commonly available. Gesture recognition forms the basis in translating sign languages. So, gesture recognition plays a critical role in sign language recognition.

The general framework of vision based SLR, as shown in Figure 1, consists of pre-processing, sign gesture representation, feature extraction, and classification. Pre-processing usually includes noise removal and hands

978-1-5386-2839-3/17/$31.00 ©2017 IEEE

localization. Sign gesture representation covers to identify a good way to present the key information. Feature extraction is to pull out condensed and discriminative descriptors to the following classifier for a better recognition purpose. Currently, the challenging problems faced in SLR are how to extract the most discriminating features from images or videos and which classifier is the best one to produce accurate results.

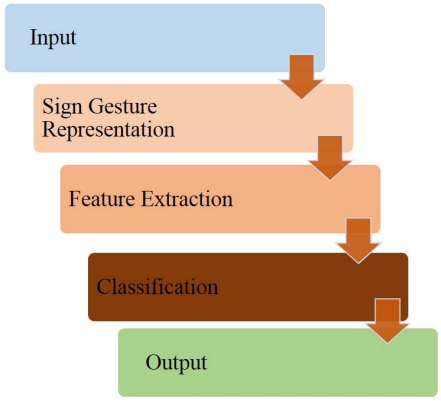


Fig. 1: General Framework of Sign Language Recognition.

The widely used 2D features include SIFT[1], HOG[2], HOF[3], STIP[3], and kernel descriptors [4]. Those features demonstrate good performance only for the case of single and clear object recognition. Additionally, 3D/4D space of time (HON4D) [5] and Random Occupancy Pattern (ROP)[6] features are proposed for the purpose of dealing with occlusions. Moreover, to address the issue of noise and occlusions in the depth maps, Space-Time Occupancy Patterns (STOP) [7] was proposed to characterize the 4D space-time patterns of human gestures to leverage the spatial and temporal contextual information while allowing for intra-class variations. [8] proposed motion history templates based on multiscale, pyramid motion history templates while considering multiple temporal scales and multiple levels of spatial grids. Commonly used classifiers have been seen as template matching, dictionary learning , bag of visual words [9, 10], Support Vector Machines (SVMs) [11, 12], and Hidden Markov Models (HMMs)[13].

## III. Sensor Technologies

Two common types of sign gesture recognition system are designed based on data acquisition: touched based or untouched/vision based. In the first category, optical, magnetic, or acoustic sensing devices were attached to hands or human body to report their positions. As shown in Figure 2, examples include data gloves [14], digital camera [15], accelerometer [16]. These sensing technologies vary along several dimensions, including accuracy, resolution, latency, the range of motion, user comfort, and cost. However, they typically require the user to wear a cumbersome device and carry a load of cables connecting the device to a computer. This hinders the ease and naturalness of the users interaction with the computer. Furthermore, extensive calibration is required when using these devices. This category adds an additional burden on users.

In contrast, untouched or vision based gesture recognition has become popular since Kinect first launched in 2010. Especially, Kinect [17], Leap motion controller [18] or Google Tango[19] provide depth maps. Several advantages of depth images include that it is insensitive to illumination changes, and the huge colour and texture variability induced by clothing, hair, skin and background could be reduced. There has been a growing amount of research on human gesture recognition by using depth information. Google's new Project Tango [19] is a limited-run experimental phone handed out to some developers recently. It has a Kinect-like vision and a revolutionary chipset, including a really sophisticated 3D scanner in a phone. This will make depth map based approaches much more convenient to signers.

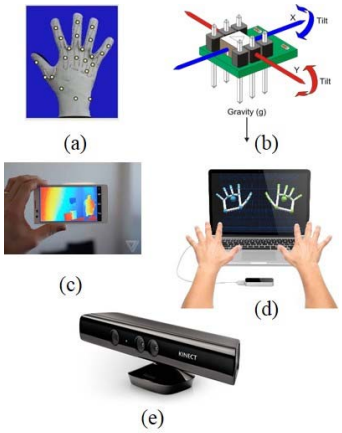


Fig. 2: Various Devices used in SLR. (a) Data glove[14], (b) Accelerometer[16], (c) Google Tango[19], (d) Leap motion controller[18], (e) Kinect[17].

Apparently sensor-based systems have the pitfall which require color markers or special gloves to acquire the necessary input. Vision based systems assume that users stand in front of a camera and begin to deliver their message using gestures.

## IV. Common Datasets of SLR

There are about seven main large data sets of sign language. They are American Sign Language Lexicon Video Dataset [24], MSR Gesture3D [6], Auslan data set [25], LTI-Gesture Database[20], RWTH German Fingerspelling Database [21], DEVISIGN Chinese Sign Language dataset[22], and Indian Sign Language dataset[23].

The **American Sign Language Lexicon Video Dataset** (ASLLVD) [24] consists of videos of more than 3,000 ASL signs in citation form, each produced by 1-6 native ASL signers, for a total of almost 9,800 tokens. This dataset includes multiple synchronized videos showing the signing from different angles. For compound signs,

the dataset includes annotations for each morpheme. To facilitate computer vision-based sign language recognition, the dataset also includes numeric ID labels for sign variants, video sequences in uncompressed-raw format, and camera calibration sequences. A total of 2,284 monomorphemic lexical signs were collected. For some signs, there is more than one variant, resulting in a total number of distinct sign variants that is greater: 2,793. For 621 of those sign variants, they have examples from a single signer; for 989 of them, they have examples from 2 signers, etc., and for 141 of those sign variants, they have examples from all 6 of native signers. For 175 of the signs, they have more than 6 tokens. Compressed versions of those sequences and annotations are publicly available from the project website[42].

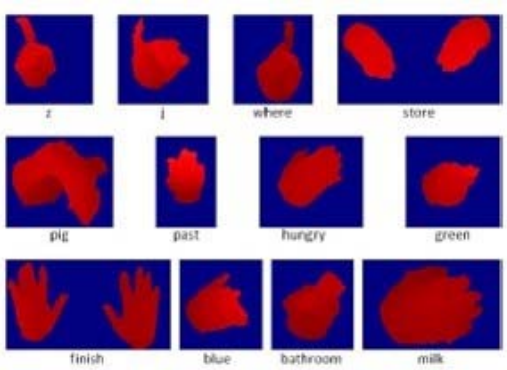


Fig. 3: Gesture Samples of MSR3D dataset.

The **MSR Gesture3D dataset**[6] is a dynamic hand gesture data set of depth sequences captured by a depth camera. It contains 12 dynamic hand gestures defined by the American Sign Language (ASL): ”z”, ”j”, ”where”, ”store”, ”pig”, ”past”, ”hungry”, ”green”, ”finish”, ”blue”, ”bathroom”, and ”milk”. There are 10 subjects, each one performing each gesture 2 or 3 times. This data set presents more self-occlusions than other data set. Notice that all of the gestures in this experiment are dynamic gestures, which means both the shape and movement of the hands are important for the semantics of the gesture. The resolutions of the depth maps are various. Some example frames of the gestures are shown in Figure 3.

**Auslan Signbank** [25] is a language resources site for Auslan (Australian Sign Language). It has about 7797 sign words and 26 finger spellings. Auslan differs to the sign language used within other countries. Auslan is a visual language, with no oral form. It uses hand shapes and movements, facial expressions and body expressions to express a visual means of communication. Each sign is made up of 5 five main parts; handshape, orientation, location, movement and facial expression. Figure 4 presents the sign words summary.The Auslan Signbank allows people to use key word to search the corresponding video of the sign language, which is played with an explanation of the word. It has included also fingerspelling and number signs.

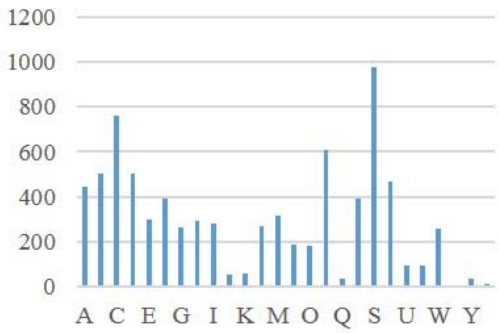


Fig. 4: Statistics of Auslan Signbank words.

**LTI-Gesture Database** [20] was created at the Chair of Technical Computer Science at the RWTH Aachen and is not freely available. It contains 14 dynamic gestures. The resolution of each video sequence is 106 x 96 grey-scale pixel. The videos were recorded with an infrared camera inside a car. In total, 364 sequences were recorded, of which 84 are used for testing and 280 are used for training. Below Figure 5 shows some examples of the different gestures.

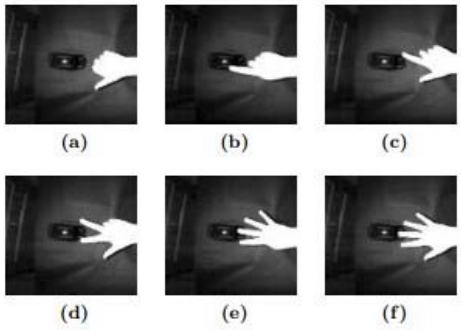


Fig. 5: Some examples of the LTI-Gesture database showing different gestures of numbers. (a) one thumb, (b) one finger, (c) two, (d) three, (e) four, (f) five

The 280 sequences dedicated for training a continuous gesture recognition system were split into a train and test set. These two sets were then used to test a single gesture recognition system.

**RWTH German Fingerspelling Database**[21] contains 35 gestures representing the letters of the alphabet, German umlauts, and the numbers from one to five. The dataset comprises 20 different signers, who did two recordings each for every gesture. Most of the gestures are static except for the ones for the letters $J$, $Z$, Ä, Ö and Ü, which are dynamic. In order to keep this experiment simple, they ran the experiments on the subset restricted to 30 static gestures. The database contains recordings by two different cameras, but they used only one camera. The short videos sequences have a resolution of 320×240 pixels. They grabbed the middle frame from each video sequence and used those color images and gesture class labels as training data. This dataset has

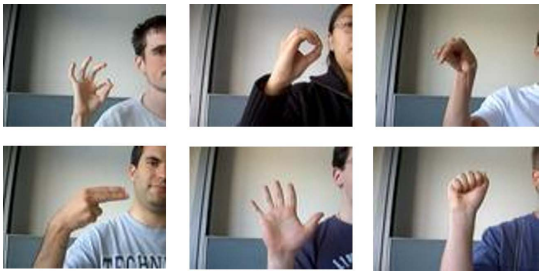

Fig. 6: Some examples of the RWTH German Fingerspelling database.

1160 images. Figure 6 shows some examples.

**DEVISIGN**[22] is a Chinese Sign Language dataset, which provides the worldwide researchers of SLR (Sign Language Recognition) community a large vocabulary Chinese sign language dataset for training and evaluating their algorithms. It has a subset (DEVISIGN-G) of DEVISIGN that is composed of 26 letters and 10 numbers performed by 8 different subjects. Among them, the signs are recorded twice for 4 subjects (2 males and 2 females) and once for the other 4 subjects (2 males and 2 females). Two examples are shown in Figure 7.

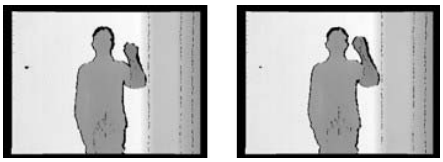

Fig. 7: Two examples of the signs in DEVISIGN-G dataset.

**Indian Sign Language dataset** [23] contains about 7500 sign gestures. These are captured simultaneously by Leap Motion and Kinect sensors. The dataset consists of 50 dynamic sign gestures that are performed by 10 different signers, 8 of the signers were male and rest were females candidates. Every sign input was repeated 15 times by each signer which makes 750 different gestures associated to a single signer, thus a total of 7500-word samples have been recorded. The dataset consists of both single and double hand dynamic sign gestures. Out of these 50 dynamic sign gestures, 28 words were performed by single hand (right hand only) and rest of the signs (i.e. 22) were performed using both hands. The dataset is available to download[43].

There are also some other sign language datasets such as British [26], Danish sign language [27], New Zealand[28], Arabic ones [29] and RWTH-PHOENIX-Weather2012 [30] and RWTH-PHOENIX-Weather Multi-signer 2014[31]. More details please refer to those reference papers cited.

## V. Recent Advances of Deep Learning for Sign Language Recognition

There are many researchers working in this active area. In this paper, we are reviewing the most recent papers published from 2014 to 2017 that apply deep learning for SLR. Table 1 summarizes the key information in term of features, models, accuracy and datasets. We discuss them in more details as follows.

Rioux-Maldague et al.[32] applied their technique to American Sign Language fingerspelling classification using a Deep Belief Network (DBN). Depth and intensity images of hand pose were captured from a Microsoft KinectTMsensor. DBN has 3 Restricted Boltzmann Machines (RBM) (with size of 1500, 700 and 400 units) and one translation layer. They evaluated results on a multi-user data set with two scenarios: one with all known users and one with an unseen user. This model achieved 99% recall and precision on the first, and 77% recall and 79% precision on the second. Their method is also capable of real-time sign classification and is adaptive to any environment or lightning intensity.

Huang, et al.[33] provided the 3D coordinates of finger joints in real time. They built a deep neural network (DNN) based on Real-Sense to recognise different signs. A three hidden layers DBN with 500, 500, and 2000 hidden units in each hidden layer were set up separately in the model. For Real-Sense-DNN, there were 66 input units and 26 units as output representing the 26 signs. For Kinect-DNN, there were 3232 input units and 26 output units. The DNN took the 3D coordinates of finger joints as input directly without using any hand-crafted features since a deep model is capable of learning suitable features for recognition from raw data. In their experiment, to demonstrate the effectiveness of Real-Sense, they collected two datasets by Real-Sense and Kinect respectively, then tested DNNs based on each dataset for recognition. The accuracy results can reach about 97.8% and 98.9% on those two data sets.

Unlike existing methods for SLR using hand-crafted features to describe sign language motion and build classification models based on those features. Huang et al.[34] proposed a novel 3D convolutional neural network (CNN) which extracts discriminative spatial-temporal features from raw video stream automatically without any prior knowledge, avoiding designing features. The model consists of eight layers. After the input layer, the next four layers are convolution layers (C1) followed by sub-sampling (S1) and convolution (C2) which is further linked by sub-sampling (S2). This is followed by a 3rd convolution layer (C3) with no subsampling following. This is further followed by two fully-connected layers containing the output layer. Specifically, various sized kernels were used in different layers. To boost the performance, multi-channels of video streams, including colour information, depth clue, and body joint positions, were used as input to the 3D CNN in order to integrate colour, depth, and trajectory information.

Koller et al.[35] embedded pre-trained 22-layer CNN model within an iterative EM algorithm for a frame-based classifier on weakly labelled sequence data. This allows the CNN to be trained on a vast number of example images in the context of hand shape recognition. The iterative EM algorithm leverages the discriminative ability of the CNN to iteratively refine the frame level annotation and subsequent training of the CNN. By

embedding the classifier within an EM framework the CNN can easily be trained on 1 million hand images. The model has been evaluated on three different datasets over 3000 manually labelled hand shape images of 60 different classes.

Oyedotun and Khashman[36] applied deep learning to the problem of hand gesture recognition for the whole 24 hand gestures of the Thomas Moeslunds gesture database. They have tested 3 different sized convolutional neural networks and 3 stacked denoising autoencoders (SDAEs). CNNs are of different depth sizes of 2, 3, or 4 hidden layers while SDAEs are with 1 to 4 hidden layers. Convolution kernels of various and suitable sizes are used in the networks. These models are all capable of learning the complex hand gesture classification task with lower error rates.

Li et al.[37] designed a feature learning approach based on sparse autoencoder (SAE) and principle component analysis for recognizing finger-spelling or sign language, from RGB-D inputs. Auto-encoder is a type of feed-forward neural network, under the unsupervised setting, whose output is required to be equal to the input. There is a hidden layer between input and output. SAEPCA network merely learns the low-level edge features. The proposed model of feature learning is consisted of two components: First, features are learned respectively from the RGB and depth channels, using SAE with CNNs. Second, the learned features from both channels are concatenated and fed into a multiple layer PCA to get the final feature. Experimental results on American Sign Language (ASL) data set demonstrate that the proposed feature-learning model is significantly effective, which improves the recognition rate from 75% to 99.05% and outperforms the state-of-the-art.

In [29], the proposed method does not require any extra gloves or any visual marks. Local features from depth and intensity images are learned using unsupervised deep learning method called PCANet. The extracted features are then recognised using linear support vector machine c1assifier. The performance of the proposed method is evaluated on a dataset of real images captured from multi-users. Experiments are conducted using a combination of depth and intensity images and in addition, using depth and intensity images performed separately. The obtained results show that the performance of the proposed system improved by combining both depth and intensity information which give an average accuracy of 99.5%.

Koller et al.[38] firstly embedded a CNN into a HMM, while interpreting the outputs of the CNN in a Bayesian fashion. While treating the outputs of the CNN as true Bayesian posteriors and training the system as a hybrid CNN-HMM in an end-to-end fashion. The new model combines the strong discriminative abilities of CNNs with the sequence modelling capabilities of HMMs. The proposed model are able to improve over the state-of-the-art on three challenging benchmark continuous sign language recognition tasks by between 15% and 38% relative and up to 13.3% absolutely.

Kumar et al. [23] propose a novel multi-modal framework for SLR system using Leap Motion and Kinect sensor. Horizontal and vertical movement of fingers of sign gestures are captured. A combination of Hidden Markov Model (HMM) and Bidirec-tional Long Short-Term Memory Neural Network (BLSTM-NN) based sequential classifiers is used to boost-up the recognition performance. The framework has been tested on a dataset of 7500 Indian Sign Language (ISL) gestures comprised with 50 different sign-words.

Kim et al. [39] compared and found that better-performing models are segmental (semi-Markov) conditional random fields (CRF) using posteriors features of deep neural networks. In the signer-dependent setting, their recognisers achieve up to about 92% letter accuracy of fingerspelling of American Sign Language dataset. The multi-signer setting is much more challenging, but with neural network adaptation, they achieve up to 83% letter accuracies in this setting.

The developed convolutional network [40] is evaluated by applying it to the problem of fingerspelling recognition for American Sign Language. The model has two blocks of convolution and pooling layers, as well as a fully-connected feedforward neural network. Depth and 3-channel image data are used as the input. With the proposed architecture, the first block has two separate parts: One extracts the edges of RGB images; the other extracts the edges of the depth. The features are then combined in the second block. The evaluation shows that the developed convolutional network performs better than previous studies and has the precision of 82% and recall of 80%.

In summary, multi-modality can improve the recognition accuracy as they offer more information of sign gestures. The Deep Neural Network can learn good feature representations from raw images or video sequences straightway. DNN based approaches show better performance than the previous ones with hand-crafted based features.

## VI. Conclusion and Future Work

The automatic sign language recognition systems could have a major impact on the lives of many people who use the sign language to communicate. However, developing such systems that recognise signs from images is still a challenging task given the variation in size, position, and shapes, etc. Most popular works mainly rely on hand-crafted features and the state-of-the-art classifiers to recognise hand gestures resulted in the bottleneck performance. Deep neural networks show promising performance in many practical applications. These DNN based approaches can learn good feature representations from raw images or video sequences straightway, which is more robust and flexible. In addition, it is very handy to deal multi-modality data, for example, the RGB-D data, skeleton, finger points, etc. that can provide rich information of signers action. 3D model of the signer can be constructed through a cloud of points re-construction. The skeleton and depth data show benefits in keeping privacy; simplifying the human body extraction process; and being invariant to illumination changes, clothing, hair, skin, and background.

High-performance relays on the number of features to learn, patch-based learning for a convolutional filter, pooling method, classifier choosing, and so on. Furthermore,

unsupervised feature learning model is very useful in the situation where the labeled data is scarce but having lots of unlabeled data. Therefore, learning high-level features by deeper networks remains an open issue in future work, to see its potential in performance enhancement. Furthermore, for real-time consideration, GPU is an alternative choice as well as further improvement of the optimization algorithm.

The success of the recent advances in deep learning for SLR has made a unique contribution to the sign language research community. Hence, it will promote active deep neural networks based model to improve the accuracy and efficiency of sign language recognition system. Eventually, it will satisfy the unmet demand for professional interpreting services and significantly improve the quality of life of the deaf people.

## REFERENCES


[1] Lowe, D.G., Distinctive image features from scale-invariant keypoints. International journal of computer vision, 2004. 60(2): p. 91-110.

[2] Dalal, N. and B. Triggs. Histograms of oriented gradients for human detection. in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005. IEEE.

[3] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B., Learning realistic human actions from movies. in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. 2008. IEEE.

[4] Bo, L., Lai, K., Ren, X. and Fox, D., Object recognition with hierarchical kernel descriptors. in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. 2011. IEEE.

[5] Oreifej, O. and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[6] Wang, J., Liu, Z., Chorowski, J., Chen, Z. and Wu, Y., Robust 3d action recognition with random occupancy patterns, in Computer visionECCV 2012. 2012, Springer. p. 872-885.

[7] Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z. and Campos, M., Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. in Iberoamerican Congress on Pattern Recognition. 2012. Springer.

[8] Zheng, L. and B. Liang. Sign language recognition using depth images. in Control, Automation, Robotics and Vision (ICARCV), 2016 14th International Conference on. 2016. IEEE.

[9] Fei-Fei, L. and P. Perona. A bayesian hierarchical model for learning natural scene categories. in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005. IEEE.

[10] Lai, K., Bo, L., Ren, X. and Fox, D., A large-scale hierarchical multi-view rgb-d object dataset. in Robotics and Automation (ICRA), 2011 IEEE International Conference on. 2011. IEEE.

[11] Chang, C.-C. and C.-J. Lin, LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2011. 2(3): p. 27.

[12] Liang, B. and L. Zheng. 3D motion trail model based pyramid histograms of oriented gradient for action recognition. in Pattern Recognition (ICPR), 2014 22nd International Conference on. 2014. IEEE.

[13] Lv, F. and R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. Computer VisionECCV 2006, 2006: p. 359-372.

[14] Mohandes, M., M.A. Deriche, and S.O. Aliyu, Arabic sign language recognition using multi-sensor data fusion. 2017, Google Patents.

[15] Hongo, H., Ohya, M., Yasumoto, M., Niwa, Y. and Yamamoto, K., Focus of attention for face and hand gesture recognition using multiple cameras. in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). 2000.

[16] Zhang, X., et al., A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2011. 41(6): p. 1064-1076.

[17] Lai, K., J. Konrad, and P. Ishwar. A gesture-driven computer interface using Kinect. in 2012 IEEE Southwest Symposium on Image Analysis and Interpretation. 2012.

[18] Chuan, C.H., E. Regina, and C. Guardino. American Sign Language Recognition Using Leap Motion Sensor. in 2014 13th International Conference on Machine Learning and Applications. 2014.

[19] Google Tango, G., https://get.google.com/tango/.

[20] Ney, I., Dreuw, P., Seidl, T. and Keysers, D., Appearance-Based Gesture Recognition. 2005.

[21] RWTH German Fingerspelling database. //www-i6.informatik.rwth-aachen.de/ dreuw/fingerspelling.php.

[22] Chai, X., H. Wang, and X. Chen, The devisign large vocabulary of chinese sign language database and baseline evaluations. 2014, Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS.

[23] Kumar, P., Gauba, H., Roy, P. and Dogra, D., A multimodal framework for sensor based sign language recognition. Neurocomputing, 2017.

[24] Athitsos, V., et al. The American Sign Language Lexicon Video Dataset. in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2008.

[25] SignBank and Auslan, http://www.auslan.org.au/about/dictionary/.

[26] The British Sign Language (BSL) Corpus, http://www.bslcorpusproject.org/.

[27] Kristoffersen, J.H., et al., http://www.tegnsprog.dk/, 2008-2016.

[28] McKee, R.M.,D., Alexander,S. P. and Pivac, L., The Online Dictionary of New Zealand Sign Language. http://nzsl.vuw.ac.nz/, 2015.

[29] Aly, S.,Osman, B., Aly, W. and Saber, M., Arabic sign language fingerspelling recognition from depth and intensity images. in 2016 12th International Computer Engineering Conference (ICENCO). 2016.

[30] Forster, J., et al. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. in LREC. 2012.

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

[32] Rioux-Maldague, L. and P. Gigure, Sign Language Fingerspelling Classification from Depth and Color Images Using a Deep Belief Network. in 2014 Canadian Conference on Computer and Robot Vision. 2014.

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

[34] J. Huang, W. Zhou, H. Li and W. Li, Sign Language Recognition using 3D convolutional neural networks. in 2015 IEEE International Conference on Multimedia and Expo (ICME). 2015.

[35] Koller, O., H. Ney, and R. Bowden. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled. in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[36] Oyedotun, O.K. and A. Khashman, Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, 2016: p. 1-11.

[37] S. Li, B. Yu, W. Wu, S. Su and R.Ji, Feature learning based on SAEPCA network for human gesture recognition in RGBD images. Neurocomputing, 2015. 151: p. 565-573.

[38] O. Koller, O.Zargaran, N. Hermann and R. Bowden, Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. in Proceedings of the British Machine Vision Conference. 2016.

[39] Kim, T., et al., Lexicon-Free Fingerspelling Recognition from Video: Data, Models, and Signer Adaptation. Computer Speech & Language, 2017.

[40] Ameen, S. and S. Vadera, A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images. Expert Systems, 2017.

[41] Agris, U.v., M. Knorr, and K.F. Kraiss. The significance of facial features for automatic sign language recognition. in 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. 2008.

[42] http://www.bu.edu/asllrp/lexicon/

[43] https://sites.google.com/site/iitrcsepradeep7/

TABLE I: Performance Summary of Various Deep Learning Approaches for SLR

| **Authors** | **Modalities** | **Model** | **Precision** | **Dataset** |
|---|---|---|---|---|
| Rioux-Maldague et al. [32] 2014 | Static depth and intensity images | DBN (3 RBM layers + 1 translation layer) | 79% | 24 static signs of American SL |
| Huang et al. [33] 2015 | Finger position | DBN (3 hidden layers) | 98.9% | Real-Sense dataset (263 video clips) of 26 alphabets signs |
| Huang et al. [33] 2015 | RGB-D | DNN (3 hidden layers) | 97.8% | Real-Sense dataset (263 video clips) of 26 alphabets signs |
| Huang et al. [34]2015 | RGB, depth and trajectory | 3D CNN (3 Cov. layers, 2 SP layers, 2 FC layers) | 94.2% | 25 vocabularies |
| Li et al. [37] 2015 | RGB-D | sparse auto-encoder (SAE) +PCA | 99.05% | American SL fingerspelling |
| Koller et al. [35] 2015 | RGB images | CNN (22 layers) + EM algorithm | 90% | Danish sign language [27] and New Zealand sign language [28] |
| Koller et al. [35] 2015 | RGB images | CNN (22 layers) + EM algorithm | 55.9% | RWTH-PHOENIX-Weather2014 [30] |
| Oyedotun et al. [36] 2016 | Gray images | CNNs (2, 3 or 4 hidden layers) and SDAEs | 97.1% | 24 static signs of American SL |
| Aly et al. [29] 2016 | Intensity and depth | PCANet+SVM | 99.5% | Arabic SL fingerspelling |
| Koller et al. [38]2016 | RGB | Hybird CNN-HMM | 92.6% | SIGNUM [41]single signer |
| Koller et al. [38]2016 | RGB | Hybird CNN-HMM | 70% | RWTH-PHOENIX-Weather2012 [30] |
| Koller et al. [38]2016 | RGB | Hybird CNN-HMM | 62.2% | RWTH-PHOENIX-Weather Multi-signer 2014[31] |
| fAmeen [40] 2017 | image intensity and depth data | CNN | 82% | American SL fingerspelling |
| Kim et al. [39] 2017 | Video sequences | DNN+CRF | 92% | American SL fingerspelling |
| Kumar et al. [23] 2017 | RGB-D, finger and palm positions | HMM+ BLSTM-NN | 96.2% | Indian Sign Language (ISL) |