Dynamic Sign Language Recognition combining Dynamic Images and Convolutional Neural Networks

Abstract

In this paper, we present a deep learning-based method for Sign Language Recognition (SLR). Our approach represents multimodal information (RGB-D) through dynamic images to describe the hand location and movement. Also, we extract a representative frame that describes the handshape, which is the input to two multi-stream CNN models. Next, we apply a later fusion stage for the final classification. Experimental results prove the feasibility of our approach.

1. Introduction

Sign Language (*SL*) is a visual-gestural language that uses hand-shape variations, body movement, and even facial expression. The *SL* structure consists of primary (hand configuration, articulation, and movement) and secondary (orientation of the palm of hands, facial expressions) parameters that are combined sequentially or simultaneously (Brito, 1995). Recently, the success of automatic *SLR* systems has opened up a new way to convert sign gestures into text/speech (Kumar et al., 2017), contributing to improving the quality of life of hearing-impaired people. Nevertheless, the *SLR* is still a very challenging task due to the complexity of exploiting the information from primary and secondary parameters. Initial approaches were based on hand-crafted techniques, *e.g.* Yang et al. (2009) used Dynamic Time Warping technique (DTW); similarly, Ronchetti et al. (2016b) used Hidden Markov Models (HMMs). Recently, deep learning approaches are used to learn high-level features and process dynamic signs (Neto et al., 2018; Konstantinidis et al., 2018a; Zhang et al., 2017; Konstantinidis et al., 2018b). On the other hand, low-cost depth devices such as Microsoft Kinect (Zhang, 2012) provides multimodal information for a sign (RGB-D and skeleton data) that improves the recognition rate. Alternatively, some authors propose the generation of texture color maps to represent the 3D skeleton trajectory (Kumar et al., 2018; Wang et al., 2016); other methods encode video sequences into movement maps (Fernando et al., 2017; Bilen et al., 2017). In this paper, our goal is extract all the information of a sign (primary parameters), we combine CNN models with texture/movement maps (dynamic images) to achieve a robust *SLR* system.

2. Method Overview

We follow a method based on the dynamic image generation using the RGB–D and skeleton data captured by a Kinect device to summarize a video sequence in single flow images describing the posture, hand configuration, and motion of a sign. Figure 1 illustrates our proposed *SLR* method. First, we use the skeleton data to extract image patches belonging to the hand movement area. Then, these patches are used to generate five dynamic images. For RGB-D data, we use the Rank Pooling method (Bilen et al., 2016; Fernando et al., 2017) to generate dynamic images from color (DC) and depth (DD) videos. For skeleton data, we extend the Skeleton Optical Spectra (SOS) method (Hou et al., 2018) to generate different spectral channels to encode the skeleton joints. Finally, we generate the DXY, DYZ, and DXZ dynamic images that represent the movement of the joints projected on the three orthogonal Cartesian planes.

Due to the short time duration of a sign, the hand shape is missing because there are frames with high blur. Therefore, we extract the most representative frame with the handshape. First, we compute the distances between each consecutive joint coordinate of the hand and compute a vector of accelerations A_h. Finally, we divide A_h into M segments; for each segment, we calculate its corresponding standard deviation (*SD*). Then, we select the segment with the minimum value of *SD* and extract the hand area from the color (CH), and depth (DH) frames with the less relative degree of focus, using the energy of Laplacian as the measure algorithm (Murali et al., 1992).

Besides, we propose two multi-stream CNN models called 3S–RGBD–CNN and 3S–SKL–CNN. In the first model, DC, DD, and DH are the inputs, whereas, in the second model, DH, CH, and the concatenation of DXY, DYZ, and DXZ are the inputs. In our CNN models, we consider the first four convolutional blocks of the pre-trained *imagenet–vgg–f* model (Chatfield et al., 2014) using its filters as initial parameters. Next, the output of the blocks *Conv 4* of each stream in both models are stacked and used as input to a convolution layer. Then, we add three fully–connected layers. Finally, we apply a later fusion to calculate the average value of the two score vectors to obtain the final classification score s^{avg}, where the highest score represents the recognized sign class.

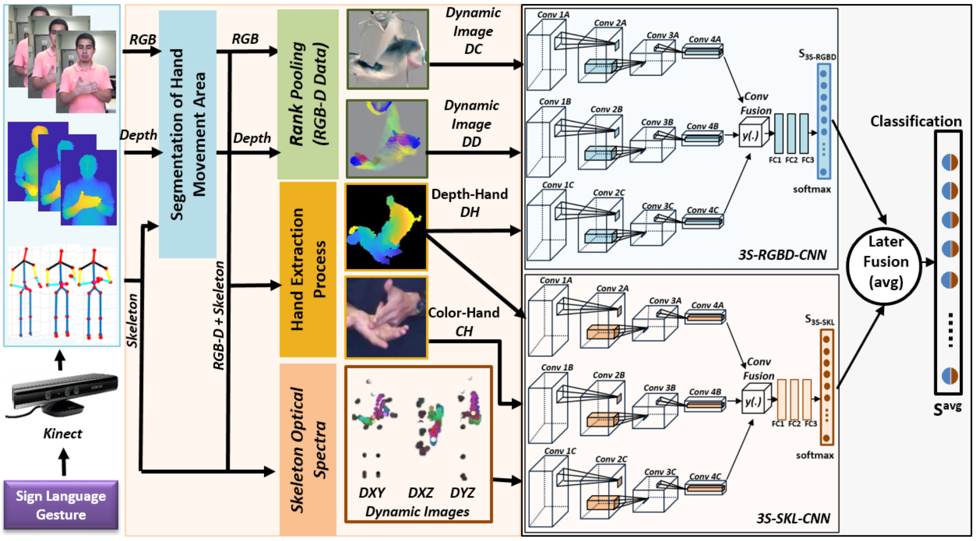

Figure 1. Overview of the proposed Sign Language Recognition method.

3. Experiments on The LSA64 dataset

First, we use the LSA64 dataset (Ronchetti et al., 2016a), which consists of 64 classes. The dataset does not provide depth and skeleton data, whereas it presents several samples balanced by class with high similarity. Therefore, we apply the *OpenPose* library to estimate body points (Cao et al., 2017). We conduct our experiments following the same experimental protocol defined by the authors. Table 1 shows the results. We note that the 3S–SKL–CNN model overcomes several proposed methods (using only the skeleton and shape of the hand) we can discriminate the majority of signs on the LSA64 dataset (99.82 %). However, we still need local information to improve results, *i.e.* the posture movement of RGB-D data (3S–RGBD–CNN). The later fusion improves the final recognition stabilizing the predictions of the CNNs (99.91%). We observe this in the standard deviation, which decreases to 0.33.

Table 1. Comparative results on the LSA64 Dataset.

Method	Accuracy (mean ± std)
ProbSOM (Ronchetti, 2018)	91.70
3DCNN (Neto et al., 2018)	93.90 ± 1.40
ALL (sequence agnostic) (Ronchetti et al., 2016b)	97.44 ± 0.59
ALL-HMM (Ronchetti et al., 2016b)	95.92 ± 0.95
Deep Network (Konstantinidis et al., 2018b)	98.09 ± 0.59
skeleton + LSTMs (Konstantinidis et al., 2018a)	99.84 ± 0.19
3S–RGBD–CNN	96.92 ± 0.56
3S–SKL–CNN	99.82 ± 0.48
Later Fusion (3S–RGBD + 3S–SKL)	99.91 ± 0.33

4. Experiments on The LIBRAS dataset

We also propose a public Brazilian Sign Language dataset (LIBRAS) composed of 56 highly similar classes based on minimal pairs (Sandler, 2012). The dataset was performed by five subjects, generating 600 samples per subject. The dataset contains complete RGB-D and skeleton data captured by a Kinect device V1. For experiments, we use cross-validation with five folds using different subjects for training, validation, and testing. Table 2 shows the results. We observe that the hand-crafted and P-CNN methods achieve low performance due to the complexity of the dataset. Again, we note that the 3S–SKL–CNN model (74.25%) overcomes the results of the 3S–RGBD–CNN model (72.44%); however, the 3DCNN-LSTM outperform it with 74.27%. Nonetheless, the later fusion achieves 75.21% of accuracy with a low standard deviation (2.97).

Table 2. Comparative results on the UFOP-LIBRAS Dataset.

Method	Accuracy (mean ± std)
SC-CHM (hand-crafted) (Escobedo & Camara, 2016)	63.30 ± 2.90
P-CNN (CNN + SVM) (Chéron et al., 2015)	68.14 ± 1.32
3DCNN–LSTM (Zhang et al., 2017)	74.27 ± 3.30
3S–RGBD–CNN	72.44 ± 3.35
3S–SKL–CNN	74.25 ± 3.28
Later Fusion (3S–RGBD + 3S–SKL)	75.21 ± 2.97

5. Conclusion

In this paper, we proposed a method for Sign Language Recognition. Most of the works in the literature use recurrent models or 3DCNN architectures; whereas, we use dynamic images to encode the movement and location of the hand with two Multi-stream CNN models. The experimental results prove the feasibility of our approach achieving satisfactory results on the LSA64 dataset outperforming state-of-the-art methods. On the LIBRAS dataset, the compared methods achieve less than 80% of accuracy, due to the high similarity between different classes.

References

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., and Gould, S. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3034–3042, 2016.

Bilen, H., Fernando, B., Gavves, E., and Vedaldi, A. Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Brito, L. F. *Por uma gramática de línguas de sinais*. Tempo Brasileiro, 1995.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields, 2017.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, pp. 1–12, 2014.

Chéron, G., Laptev, I., and Schmid, C. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3218–3226, 2015.

Escobedo, E. and Camara, G. A new approach for dynamic gesture recognition using skeleton trajectory representation and histograms of cumulative magnitudes. In *Graphics, Patterns and Images (SIBGRAPI), 2016 29th SIBGRAPI Conference on*, pp. 209–216. IEEE, 2016.

Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., and Tuytelaars, T. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2017.

Hou, Y., Li, Z., Wang, P., and Li, W. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):807–811, 2018.

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

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

Kumar, E. K., Kishore, P., Sastry, A., Kumar, M. T. K., and Kumar, D. A. Training cnns for 3-d sign language recognition with color texture coded joint angular displacement maps. *IEEE Signal Processing Letters*, 25(5):645–649, 2018.

Kumar, P., Gauba, H., Roy, P. P., and Dogra, D. P. Coupled hmm-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86:1–8, 2017.

Murali, S., Choi, T.-S., and Nikzad, A. Focusing techniques. *Applications in Optical Science and Engineering. International Society for Optics and Photonics*, 1992.

Neto, G. M. R., Junior, G. B., de Almeida, J. D. S., and de Paiva, A. C. Sign language recognition based on 3d convolutional neural networks. In *International Conference Image Analysis and Recognition*, pp. 399–407. Springer, 2018.

Ronchetti, F. Reconocimiento de gestos dinámicos y su aplicación al lenguaje de señas. In *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).*, 2018.

Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., and Rosete, A. Lsa64: A dataset of argentinian sign language. *XX II Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016a.

Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., and Rosete, A. Sign languague recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language. In *Ibero-American Conference on Artificial Intelligence*, pp. 338–349. Springer, 2016b.

Sandler, W. The phonological organization of sign languages. *Language and linguistics compass*, 6(3):162–182, 2012.

Wang, P., Li, Z., Hou, Y., and Li, W. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 102–106. ACM, 2016.

Yang, R., Sarkar, S., and Loeding, B. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):462–477, 2009.

Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., and Bennamoun, M. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3120–3128, 2017.

Zhang, Z. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, 2012.