Brazilian Sign Language Recognition Using Kinect

José Elías Yauri Vidalón and José Mario De Martino(✉)

School of Electrical and Computer Engineering,
University of Campinas, Campinas, SP, Brazil
{elias,martino}@dca.fee.unicamp.br

Abstract. The simultaneous-sequential nature of sign language production, which employs hand gestures and body motions combined with facial expressions, still challenges sign language recognition algorithms. This paper presents a method to recognize Brazilian Sign Language (Libras) using Kinect. Skeleton information is used to segment sign gestures from a continuous stream, while depth information is used to provide distinctive features. The method was assessed in a new data-set of 107 medical signs selected from common dialogues in health-care centers. The dynamic time warping–nearest neighbor (DTW-kNN) classifier using the leave-one-out cross-validation strategy reported outstanding results.

Keywords: Sign language · Isolated sign language recognition · Brazilian Sign Language · Libras · Dynamic time warping · k–Nearest Neighbor

1 Introduction

In daily life, deaf and hearing impaired people use sign language as a communication system [1]. Sign language combines hand gestures, body postures, and facial expressions to convey meaning. The richness of sign language lexicon allows, as any other language, the expression of concepts, ideas, feelings, mood, or thoughts. Contrary to popular belief, sign language is not a universal language. There are many different sign languages around the world, for instance, the American Sign Language (ASL) in United States, British Sign Language (BSL) in England, Brazilian Sign Language (Libras) in Brazil. Furthermore, different countries that have the same spoken language may have their own sign language, e.g., although United States and England share the English as common oral language, ASL differs from BSL.

Despite sign language capabilities to communicate messages, there is a strong barrier between deaf and hearing people. This language barrier arises because deaf people usually do not master spoken and written language and only few hearing people can communicate using sign language. Aiming to reduce this language barrier, research efforts have been conducted in sign language recognition (SLR) [2–4]. Automatic SLR systems translate sign language into text and

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 391–402, 2016.
DOI: 10.1007/978-3-319-48881-3_27

can improve the interaction between deaf and hearing people. Critical situations where the communication is decisive, such as the access to emergency health services, may greatly benefit from automatic sign language technologies.

Currently, powered by new sensing technologies, new promising SLR approaches are being developed. The advent of depth cameras [5], also known as RGB-D cameras, has been an important milestone in the computer vision community because they can provide multimodal data, such as RGB or color images, depth range images, body skeleton, and user silhouettes, that can help to overcome the traditional restrictions of illumination changes and cluttered background of SLR systems based on traditional imaging systems.

Despite great progress in the last years, the building of robust and reliable SLR systems is still in its infancy. The high variability both in appearance and motion of signs, the signer dependence, the size of the vocabulary, the signing environment and imaging conditions still challenge any SLR algorithms.

This paper presents a method to recognize isolated signs of the Brazilian Sign Language (Libras) using Microsoft Kinect. First, motion analysis of the body skeleton allows for both segmenting signs from a continuous stream and categorizing them as either one-handed or two-handed. Next, the histogram of direction cosines (HDC) [6] are computed from the depth images of the segmented sign. To evaluate performance of the solution, a data-set of 107 medical signs were recorded. Our approach, based on dynamic time warping nearest neighbor classification strategy, reached an accuracy over 98.69 % on the data-set.

The remainder of this paper is organized as follows. Section 2 presents the related work in SRL. Section 3 explains the proposed method. Section 4 details the experimental results. Finally, Sect. 5 exposes conclusions and future work.

2 Sign Language Recognition

Sign language is a visual-spatial language that uses hands, body, head, and facial expressions to convey meaning [1,7]. In sign language, the meaningful unit is the sign. To be analyzed, a sign can be decomposed into manual and non-manual parameters. The manual parameters relate to the shape, location, movement, and orientation of the hands, while the facial expressions and body postures are the non-manual parameters. The manual component of the signs usually carries the most of the meaning, however the presence of non-manual components may change or modulate the meaning.

Automatic sign language recognition (SLR) aims to recognize and translate sign language into text [2]. To face the challenge, SLR methods focus either on recognizing isolated signs or recognizing continuous sentences. Methods for isolated sign recognition usually assume that the boundaries of signs are easy to estimate, so they are most focused on the recognition task. On the other hand, because the boundaries of signs in sentences are unclear, methods for continuous sign recognition are more complex because they have to estimate the start and end frames of signs before performing recognition tasks. Although isolated sign recognition is more simpler than the continuous case, it provides an important learning stage before going to develop continuous sign recognizers.

According to the sensor being used for capturing sign language, SLR methods can be categorized into wearable sensor based or vision based. Wearable sensors combine data-gloves and body markers to track the hands and the body motion. On the other hand, vision based methods use cameras that mimic the human vision for imaging the scene. Although color cameras allow approaches to detect and track the hands, body, and facial characteristics, they are sensible to changes in illumination and background conditions. The advent of depth sensors offers new forms of deal with images [5]. Geometric information contained in depth images has become an essential tool to improve approaches only based on color images, as well as a new source for new discriminative features [8].

Recent SLR methods seek to take advantage of the multimodal data provided by depth sensors. Usually, color and depth images are used to extract shape features of the hands, body, and facial expressions, while skeletons are used to provide motion features of body parts. To perform recognition, machine learning approaches are mostly used [9]. In this context, signs can be modeled either as time-series or as a single feature vector that assumes that all signs have the same length. Dynamic time warping (DTW), hidden Markov models (HMM), and lately conditional random field (CRF) algorithms are suitable for the former, while support vector machine (SVM), random forest (RF), neural network (NN), and deep learning methods are applicable for the latter.

In the following paragraphs, we highlight some recent proposed approaches that use the Kinect depth sensor for SRL. Nonetheless, a extensive thorough review of SLR methods can be found in [2–4].

Escobedo-Cardenas and Camara-Chavez [10] used SIFT features extracted from intensity and depth images in a bag-of-words combined with the upper body skeleton positions to recognize 20 signs of the Italian Sign Language (ISL). They assumed that all signs have N key-frames (N = 10) for removing their temporal variation. Performance evaluation reported 88.39 % of average recognition using SVM classifier.

Pigou et al. [11] presented a method for feature learning based on 2D convolutional neural network (CNN). The CNN processes N key-frames (N=32) from both intensity and depth images for feature extraction. Performance evaluation in 20 sign of ISL achieved 95.68 % of average recognition using ANN classifier.

Conly et al. [12] proposed a method to retrieve the most similar signs to a given one. Based on the movement trajectories of the hands, DTW computes similarities between signs and returns a list of the top-k most close signs. Performance evaluation in a data-set of 1113 ASL signs shows that in 62.00 % of cases the sign being queried is found in the top 20 list.

Hanjie et al. [13] used HOG and body skeleton features to recognize Chinese Sign Language (CSL). To reduce the HMM recognition time, a low-rank approximation of feature vectors furnishes both the key-frames of signs and the presumable number of hidden states of the model. Hence, the HMM speeds-up to three times the recognition time. Performance evaluation on data-sets of

370 and 1000 CSL signs reported 94.00 % and 84.00 % of average recognition, respectively.

The recognition of Brazilian Sign Language (Libras) was also addressed by some researches. Anjo et al. [14] had a 100 % of success in recognized 10 static poses of the manual alphabet using Kinect and ANN. Souza and Pizzolato [15] used Kinect to recognize both finger-spelled words and isolated signs using SVM and CRF, respectively. Later, Moreira et al. [16] used a fingertip detector and tracker sensor to recognize 26 letters of the manual alphabet. They achieved 61.53 % of average recognition using ANN. Recently, Bastos et al. [17] used HOG and Zernike moments to recognize 40 predefined static signs. They reported 96.77 % of average recognition using ANN.

Despite the progress, automatic recognition of sign language is still in its infancy and Kinect has not been fully explored to develop applications that might benefit to deaf and hearing impaired people.

This paper presents a method to recognize isolated signs of the Brazilian Sign Language using Kinect. Instead of dealing with multiple data sources, we propose to use the only depth image to extract discriminative features. In our approach, the signer performs signs continuously following the *stop–motion* strategy (i.e., the hands are down and stopped before and after a sign is performed). Accordingly, signs are segmented in time by a simple motion analysis of the hands. Motion analysis also allows for labeling the signs as either one-handed or two-handed to reduce the searching space during the classification stage. Finally, signs are modeled as time-series which are classified using the dynamic time warping–nearest neighbor (DTW-kNN) algorithm.

3 Proposed Method

Figure 1 illustrates the proposed framework for sign language recognition. In short, the framework uses both the depth image and the skeleton data provided by Kinect [18]. To segment signs from a continuous stream, the system detects stop–motion patterns based on the skeleton information and also determines the hand dominance (i.e., one-handed or two-handed). Once identified the start and end of a sign, the histogram of direction cosines [6] features is computed for the depth images. During training, feature descriptors together with the hand dominance labels are stored in a database of sign models. During testing, the unknown sign is recognized via dynamic time warping–nearest neighbor classifier.

Next, we describe the main stages of the framework:

3.1 Sign Segmentation

To be more closer to real-life situations where a speaker produces sequence of words, our system allows the subjects to sign constantly following a stop–motion scheme. Our stop–motion scheme establishes that the hands are down and stopped before and after a sign is performed, so patters of “silence” (stop) and “activity” (motion) are easily detected in the continuous stream of signs.

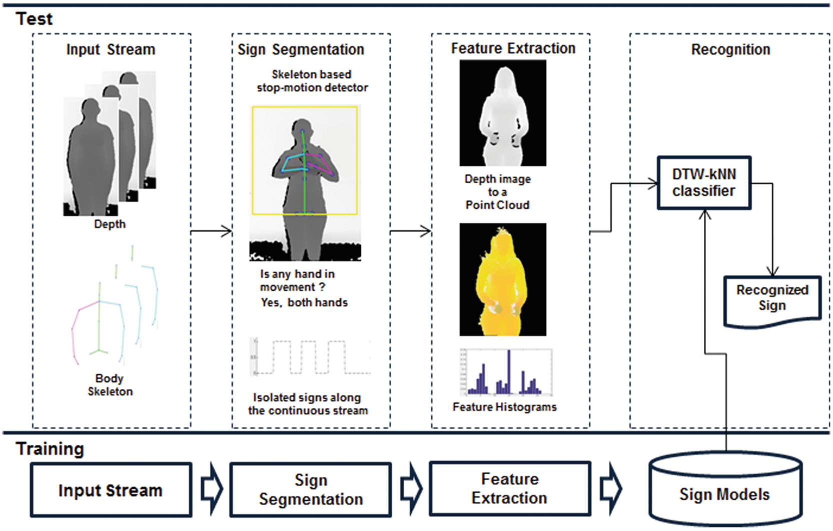


Fig. 1. Framework for sign language recognition.

A stop state indicates a non-sign segment, whereas a motion state indicates a sign segment.

In order to detect stop–motion segments, the system uses the 3D skeleton data provided by Kinect. The user usually shrinks the arms when he/she is signing, therefore the angle ω between the forearm and arm in the 3D space gives a clue whether a stop state or a moving state is happening in the stream. For each frame over time t, angles are evaluated according to:

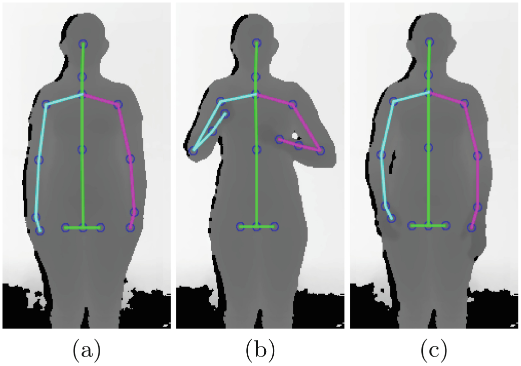
$$S(t) = \begin{cases} 1, & \omega < Th_{angle} \\ 0, & otherwise \end{cases} \quad (1)$$

resulting in a sequence $S(t)$ of 0s and 1s, for each arm. Usually, a transition from zero to one indicates the start of the sign, whereas the transition from one to zero means the end of the sign. Measurement of 1s in $S(t)$ also allows to determine the dominant hand. Figure 2 presents examples of stop and motion frames of a sign.

3.2 Depth Image Preprocessing

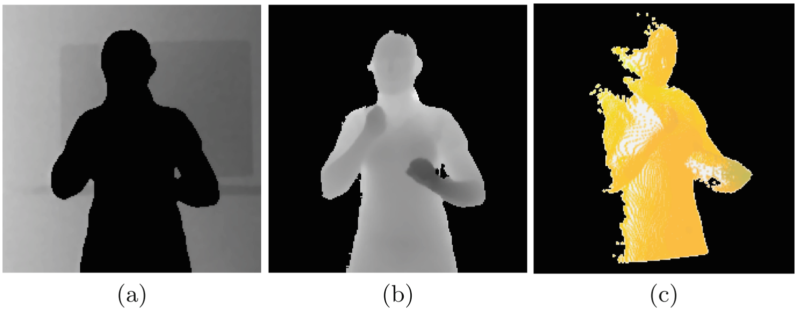
Once identified the boundaries of signs, their depth images are processed as follows:

Step 1: Defining the region of interest. Since the subject usually occupies a small region of the image and the signing occurs in the upper region of the body, the system defines a region of interest (ROI) in the first depth frame.



(a) (b) (c)

Fig. 2. Skeleton based stop–motion detection. (a) Stop, (b) Motion, and (c) Stop frames.



(a) (b) (c)

Fig. 3. Depth image preprocessing: (a) After Cropping with a ROI, (b) After depth thresholding, and (c) After mapping to a point cloud.

The ROI is specified around the Head, Spine-Base, Shoulder-Left, and Shoulder-Right joints of the 2D skeleton given by Kinect. The skeleton keeps the aspect ratio of the user, so the ROI is robust against changes in both the user size and location.

Using the ROI, all the depth images of a sign are cropped. An illustration of the cropping result is shown in Fig. 3a.

Step 2: Removing the background. To segment the body of the user, we perform segmentation along the depth axis. Depth values beyond a threshold are zoomed in the image. The system uses the threshold

$$Th_{depth} = Head_{depth} + \varDelta \tag{2}$$

the depth value of the Head position plus an additional depth value $\varDelta$. An illustration of the segmentation result is shown in Fig. 3b.

Step 3: Mapping to a point cloud: Segmented depth images are mapped to point clouds using the intrinsic camera parameters of Kinect. A point cloud is a set of spatially organized points along the X, Y, Z coordinates of the camera. An illustration of the point cloud of a depth image is shown in Fig. 3c.

3.3 Feature Extraction

As a feature descriptor, we use the histogram of direction cosines (HDC). The HDC was successfully used to classify static hand postures of ASL [6], however in this work we extend it to classify isolated signs.

Direction cosines are the cosine angles between a vector and the Cartesian axes, and a HDC histogram accumulates direction angles in the same way as to histogram of oriented gradients [19].

For a vector $\boldsymbol{v} = a\hat{i} + b\hat{j} + c\hat{k}$ in the 3D Cartesian coordinates, the direction cosines are:

$$\begin{aligned} p = \cos\alpha &= \frac{a}{a^2+b^2+c^2} \\ q = \cos\beta &= \frac{b}{a^2+b^2+c^2} \\ r = \cos\theta &= \frac{c}{a^2+b^2+c^2} \end{aligned} \tag{3}$$

where $p^2 + q^2 + r^2 = 1$. The angles α, β, and θ can be obtained by inverting the function.

Geometrically, direction cosines characterize a vector using its orientation relative to the Cartesian axes. For a set of vectors, direction cosines portrays the surface encompassed by the vectors.

To increase the distinctiveness of the original HDC, we propose a slight modification in the weighted vote on each bin of the histogram. Steps to compute HDC from a point cloud PC are:

1. Determine the central point p_c of PC.
2. Generate the directional vectors for all points p_i of PC.

$$\boldsymbol{v}_{p_i} = \{p_c - p_i | \forall p_i \in PC\} \tag{4}$$

3. For each $\boldsymbol{v}_{p_i}$ estimates its direction cosines (according Eq. 3) to obtain the orientation angles α, β, and θ as well as the magnitude $\|\boldsymbol{v}_{p_i}\|$.
4. Calculate the histogram of cumulative magnitudes for each coordinate axis. Each histogram encompasses 9-bins from 0 to 180°. Each vector $\boldsymbol{v}_{p_i}$ casts an orientation-based vote in which its magnitude is weighted and distributed to three histograms. The closer the vector is to an axis, the greater the weight is to the respective histogram.
5. The final feature vector FV consists of the concatenation of the three cumulative histograms. Finally, FV is normalized scaling to unit length.

$$FV = \{h_x, h_y, h_z\} \tag{5}$$

For a depth image, the HDC gives a 27-dimensional feature vector, i.e., 3 histograms × 9 bins = 27.

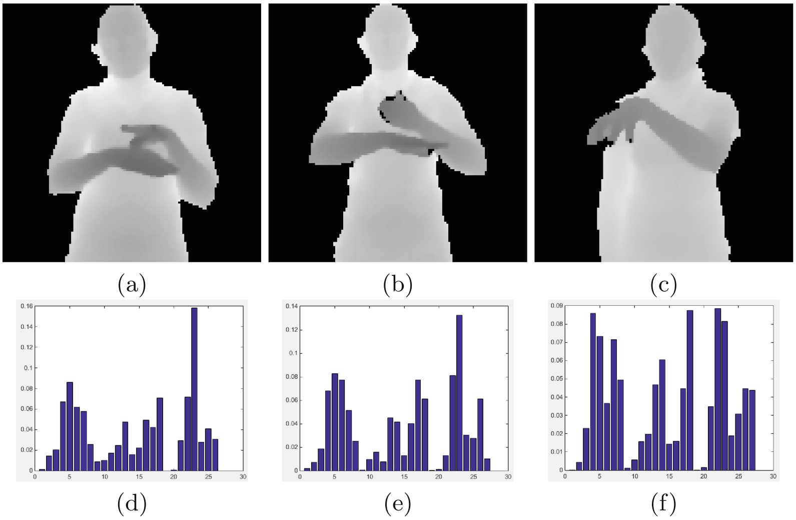


Fig. 4. Different depth images (a), (b), (c) and their HDC features (d), (e), (f), respectively.

Figure 4 illustrates the HDC features calculated from three different depth images. Visually, the two images have very similar postures, varying slightly in the hand shape, whereas the posture in the third image is quite different of them. In order to figure out the degree of closeness of the images, we measured the Euclidean distance between histogram H_i : $distance(\mathbf{H_1}, \mathbf{H_2}) = 0.0458, distance(\mathbf{H_1}, \mathbf{H_3}) = 0.0883$, and $distance(\mathbf{H_2}, \mathbf{H_3}) = 0.0851$. Accordingly, the two first images are similar, but slightly different from the third.

3.4 DTW-Based Recognition

Subjects sign in different ways –e.g., different impetus, speed, and style– so the same sign can change in time even for the same user. Therefore, the time-series model of signs have different lengths, which should be identified in order to recognize the signs.

Because its simplicity and flexibility to compare two different time-series which are similar but locally out of phase, we propose to use the dynamic time warping (DTW) technique. DTW allows for a non linear mapping of one time-series to another by minimizing the distance between them [20].

The proposed system recognizes signs in two steps:

1. Measuring the DTW distance between the queried time-series against the time-series stored in a pre-built database.
2. Classifying the queried time-series using the k-Nearest Neighbor algorithm based on DTW measurements.

4 Experimental Results

We have collected a new data-set using Kinect v2 [18]. The vocabulary consists of 107 medical signs of the Brazilian Sign Language (Libras)–65 one-handed and 42 two-handed signs. Signs were recorded in continuous streams following the stop–motion scheme. A deaf informant performed each sign 5 times. The distance between the sensor and user is between 1.5–2.0 m. An outline of the vocabulary is exposed in Table 1.

Table 1. Medical sign vocabulary in Libras

Sicken, Medical scheduling, Now, Severe, Needle, Allergy, Tomorrow, Tonsillitis, Year, Anxiety, Appendicitis, Heart attack, Well, Bronchitis, Head, Mumps, Surgery, Pill, Medical consultation, Contusion, Chronic, To heal, His, Delirium, Insanity, Dengue, Tooth, Depression, Brain stroke, Dehydration, Diabetes, Disease, To ache, Headache, Electrocardiogram, He, Address, Nursing, Sprain, Poisoning, Stable, Stethoscope, Stomach, I, Medical exam, Fever, Fracture, Future, Flu, Bleeding, Hepatitis, Hypertension, Today, Hospital, Age, Unstable, Respiratory infection, Injection, Intoxication, To go, Laceration, Injury, Knife injury, Gun injury, Slight, Hurt, Doctor, My, Die, Very/Too, To cannot, To want not, To have not, Nausea, Name, Yesterday, Hearing people, Past, Kidney stone, Chest, To can, A few, Need, Clinic history, Psychosis, Lung, Burs, To want, X-rays, Medical prescription, Remedy, Medical risk, Bad, Salmonella, Healthy, To feel, Your, Yes, Deaf people, To have, Dizziness, Cough, Vaccine, To come, You, Vomit

We used the following parameters in our experiment:

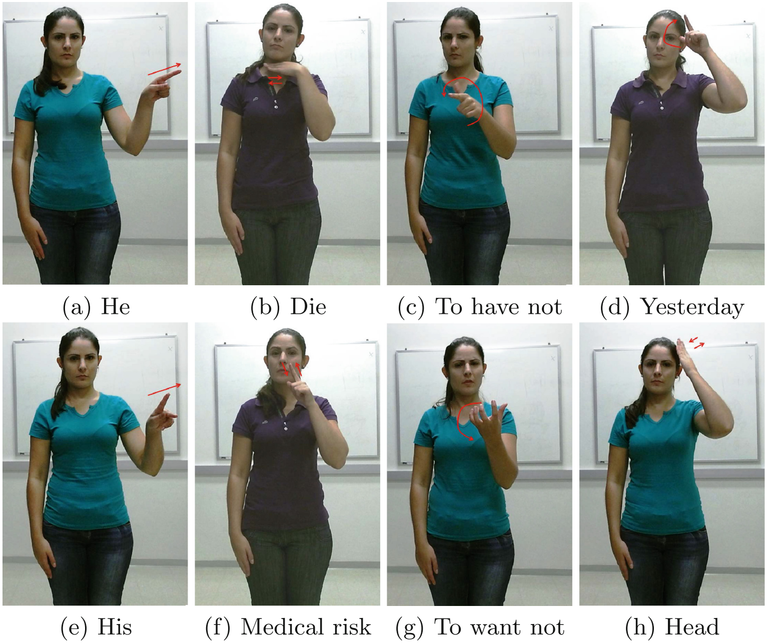
- Angle threshold $\omega = 130$ between the arm and forearm for temporal segmentation of signs.
- Region of interest (ROI) around the Head, Spine-base, Shoulder-left, and Shoulder-right for cropping depth images.
- Depth value $\varDelta = 200\,\text{mm}$ as a step value for background subtraction behind the user.
- Median mask 3×3 for filtering noise in depth images.
- Mapping depth images to point clouds (PC) for computing their histogram of direction cosines (HDC) features.
- DTW–Nearest Neighbor classifier.
- Leave-one-out cross-validation (LOOCV) for performance evaluations [9].

The data-set contains $\varSigma = 107 \times 5 = 535$ signs. To evaluate the classification performance we use the LOOCV because the signs contain few examples. In this way, there are $n = 535$ cases to be evaluated, so the case i is tested against the training set which consists of all cases except i.

Table 2. Classification performance result

Average accuracy	Average precision	Average recall
98.69	98.88	98.69

After experimenting, we achieved an average accuracy of 98.88 %. The result was promising, however, we perceived that the computational complexity of feature extraction is highly correlated with the size of the images. Such computational cost can be reduced by processing the depth image pixels at a given offset. An offset = 2 reduces the computation time up to a quarter, without affecting performance (average accuracy of 98.69 %). Nonetheless, offsets greater than 2 steps diminish the discriminative power of feature vectors due to the loss of fine details in the image (average accuracy lower than 90 %). Table 2 shows in details the classification performance of the DTW–nearest neighbor classifier.


(a) He (b) Die (c) To have not (d) Yesterday

(e) His (f) Medical risk (g) To want not (h) Head

Fig. 5. Screen-shot of some misclassified signs

The results evidence the discriminative power of depth based features and the feasibility to use for isolated sign recognition. However, it is worth remarking that several signs were misclassified. For instance, the system fails in differentiation the signs He and His, Die and Medical risk, To have not and To want not, and Yesterday and Head. Screen-shots of some misclassified signs are shown in Fig. 5.

It seems that signs that are misclassified are those that differ in one or two characteristics, usually named minimal pairs [1,7]. For instance, the signs He–His and To have not–To want not differ only in the hand configuration, whereas the signs Die–Medical risk and Yesterday–Head differ both in configuration and subtle movements of the hands. Moreover, these signs are hardly distinguishable since the arm poses are similar and only vary in either the hand shape or the hand movement beyond the wrist. A dedicated recognizer of the hands may help to detect and disambiguate signs that have slight differences between them.

5 Conclusions and Future Work

This paper presented a method to recognize 107 medical signs of the Brazilian Sign Language (Libras) using Kinect. The method takes advantage of the geometric information contained in depth images to compute a high discriminative spatial-appearance feature. Classification experiments using DTW-kNN reported an striking result of 98.69 % in a single signer-dependent data-set. No tracking, no locations, and no region cropped of the hands were required. Furthermore, aiming to work in real-life scenarios, a skeleton based stop–motion detector was introduced to segment signs performed continuously. In order to improve the robustness of the proposed approach, skeleton and hand shape features will be added, the vocabulary will be extended with signs recorded from different users, and distance metric learning strategies will be explored in a future work.

Acknowledgment. The authors would like to thank the Brazilian agency CAPES for its financial support.

We also thank the anonymous reviewers for their valuable criticism and suggestion which have greatly contributed to the final version of the manuscript.

References

1. Valli, C.: Linguistics of American Sign Language: An Introduction. Gallaudet University Press, Washington, DC (2000)
2. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) Visual Analysis of Humans: Looking at People, pp. 539–562. Springer, London (2011)
3. Sahoo, A.K., Mishra, G.S., Ravulakollu, K.K.: Sign language recognition: state of the art. ARPN J. Eng. Appl. Sci. **9**, 116–134 (2014)
4. Classification, V.-B.S.L., Joudaki, S., bin Mohamad, D., Saba, T., Rehman, A., Al-Rodhaan, M., Al-Dhelaan, A.: Vision-based sign language classiffication: a directional review. IETE Tech. Rev. **31**, 383–391 (2014)
5. Lefloch, D., Nair, R., Lenzen, F., Schäfer, H., Streeter, L., Cree, M.J., Koch, R., Kolb, A.: Technical foundation and calibration methods for time-of-flight cameras. In: Grzegorzek, M., Theobalt, C., Koch, R., Kolb, A. (eds.) Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. LNCS, vol. 8200, pp. 3–24. Springer, Heidelberg (2013). doi:10.1007/978-3-642-44964-2_1
6. Escobedo Cardenas, E., Camara Chavez, G.: Finger spelling recognition from depth data using direction cosines and histogram of cumulative magnitudes. In: 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 173–179 (2015)
7. de Quadros, R.M., Karnopp, L.B.: Língua de Sinais Brasileira - Estudos Linguísticos. Artmed, Porto Alegre (2004)
8. Suarez, J., Murphy, R.: Hand gesture recognition with depth images: a review. In: RO-MAN 2012, pp. 411–417. IEEE (2012)

9. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge (2012)
10. Escobedo-Cardenas, E., Camara-Chavez, G.: A robust gesture recognition using hand local data and skeleton trajectory. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 1240–1244 (2015)
11. Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 572–578. Springer, Heidelberg (2015)
12. Conly, C., Zhang, Z., Athitsos, V.: An integrated RGB-D system for looking up the meaning of signs. In: Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2015), pp. 24: 1–24: 8. ACM, New York (2015)
13. Wang, H., Chai, X., Zhou, Y., Xilin, C.: Fast sign language recognition benefited from low rank approximation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–6. IEEE (2015)
14. Anjo, M.D.S., Pizzolato, E.B., Feuerstack, S.: A real-time system to recognize static gestures of Brazilian sign language (Libras) alphabet using kinect. In: Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems (IHC 2012), Porto Alegre, Brazil, pp. 259–268. Brazilian Computer Society (2012)
15. de Souza, C.R., Pizzolato, E.B.: Sign language recognition with support vector machines and hidden conditional random fields: going from finger spelling to natural articulated words. In: Perner, P. (ed.) MLDM 2013. LNCS, vol. 7988, pp. 84–98. Springer, Heidelberg (2013)
16. Matuck, G.R., Moreira, G.S.P., Saotome, O., da Cunha, A.M.: Recognizing the Brazilian signs language alphabet with neural networks over visual 3d data sensor. In: Bazzan, A.L.C., Pichara, K. (eds.) IBERAMIA 2014. LNCS, vol. 8864, pp. 637–648. Springer, Heidelberg (2014)
17. Bastos, I.L.O., Angelo, M.F., Loula, A.C.: Recognition of static gestures applied to Brazilian sign language (Libras). In: Proceedings of the 2015 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV 2015) (2015)
18. Microsoft Inc. Kinect for Windows SDK 2.0. (2014). https://developer.microsoft.com/en-us/windows/kinect/develop. Accessed 19 Aug 2016
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893 (2005)
20. Müller, M.: Information Retrieval for Music and Motion. Springer, Heidelberg (2007)