Expert Systems with Applications 41 (2014) 7259–7271

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Review

Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors

Sílvia Grasiella Moreira Almeida [a,b], Frederico Gadelha Guimarães [c,*], Jaime Arturo Ramírez [c]

[a] Graduate Program in Electrical Engineering, Federal University of Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, Minas Gerais, Brazil
[b] Federal Institute of Minas Gerais, Rua Pandiá Calógeras 898, Ouro Preto, Minas Gerais, Brazil
[c] Department of Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

ARTICLE INFO

Article history:
Available online 28 May 2014

Keywords:
Brazilian Sign Language Recognition
RGB-D sensors
Feature extraction

ABSTRACT

In contrast to speech recognition, whose speech features have been extensively explored in the research literature, feature extraction in Sign Language Recognition (SLR) is still a very challenging problem. In this paper we present a methodology for feature extraction in Brazilian Sign Language (BSL, or LIBRAS in Portuguese) that explores the phonological structure of the language and relies on RGB-D sensor for obtaining intensity, position and depth data. From the RGB-D images we obtain seven vision-based features. Each feature is related to one, two or three structural elements in BSL. We investigate this relation between extracted features and structural elements based on shape, movement and position of the hands. Finally we employ Support Vector Machines (SVM) to classify signs based on these features and linguistic elements. The experiments show that the attributes of these elements can be successfully recognized in terms of the features obtained from the RGB-D images, with accuracy results individually above 80% on average. The proposed feature extraction methodology and the decomposition of the signs into their phonological structure is a promising method to help expert systems designed for SLR.

1. Introduction

Languages are complex systems of communication that humans use to express themselves, to manipulate objects and ideas and to foster cooperation and social bonds. The estimated number of languages in the world amount to a few thousands, the majority of which are vocal languages, also termed oral or spoken languages. Nevertheless, natural languages can also be signed ones, which is the most natural modality for deaf people. Both vocal and sign languages are composed by a limited set of building blocks called phonemes; in sign languages a limited set of shapes, orientations, locations and movements of the hands are combined to make up the words or morphemes of the language. The grammatical rules of the language link those morphemes into phrases and units of discourse. Many countries in the world have an official sign language, some examples are American Sign Language (ASL), French Sign Language (FSL), Italian Sign Language (ISL), Polish Sign Language (PSL), Japanese Sign Language (JSL), among others. In Brazil, the official sign language is the Brazilian Sign Language (BSL),[1] which was developed in the 19th century as a combination of the ancient Brazilian Sign Language and the French Sign Language. BSL is even different from the Portuguese Sign Language, even though Brazil and Portugal have the same vocal language, the Portuguese. Nowadays, BSL is recognized as an official language in Brazil with an estimated number of 5 million speakers, which is greater than the number of speakers of many vocal languages in the world.

Sign languages are visual-space languages, sharing spatial and visual elements such as shape of the hands, location and orientation of the signs, and the movements of the hands and the body together to convey ideas. All sign languages are characterized by these building blocks, differing basically in variations of these spatial and visual elements. In general, sign languages share a common phonological structure, with five elements[2]: (i) articulation point; (ii) Configuration of the hands; (iii) type of movements of the hands; (iv) orientation and (v) facial and body expressions. Each sign in each language is composed by a combination of these building blocks. These elements represent an important aspect of the language and can be exploited in automatic expert systems aimed at Sign Language Recognition (SLR). In contrast to speech recognition, whose speech features have been extensively explored in the

* Corresponding author. Tel.: +55 31 3409 3419.
E-mail addresses: silvia.almeida@ifmg.edu.br (S.G. Moreira Almeida), frederico-guimaraes@ufmg.br (F.G. Guimarães), jramirez@ufmg.br (J. Arturo Ramírez).

[1] In Portuguese, BSL is known as LIBRAS.

[2] Complete phonological, semantics, grammar and spelling structures defined in BSL, can be found in Quadros and Karnopp (2004).

http://dx.doi.org/10.1016/j.eswa.2014.05.024
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

research literature, feature extraction in SLR is still a very challenging problem, still lacking effective features for recognition that are independent from signers. Feature extraction in SLR requires different methods and approaches than in speech recognition. Most of the research literature focuses on specialized techniques of computer vision for extracting features but it is possible to find other approaches relying on gloves, sensors on fingers and hands and, more recently, RGB-D sensors, which capture not only color information but also depth infrared images. In a wider context, research on SLR has a potential impact on human–computer interaction, since both modalities of natural languages (vocal and signed) could in principle be used as an interface to interact with computers.

In this paper we present a methodology for feature extraction in Brazilian Sign Language that explores the phonological structure of the language and relies on an RGB-D sensor for obtaining data. Our goal is to extract features that would allow, in future work, the creation of an expert system capable of automatically translating signs from BSL to vocal Portuguese. To avoid retraining the system to every new sign and to overcome the scalability challenge related to the high number of different signs in the language, we relate these features to the phonological structure of the BSL. The use of an RGB-D sensor is very convenient in this application, since it is a low cost sensor with relatively easy programming. Therefore, we propose herein an innovative methodology in which the computational system should be guided by the linguistic system in the feature extraction process.

From the RGB-D images we obtain seven vision-based features. Each feature is related to one, two or three structural elements in BSL. We investigate this relation between extracted features and structural elements based on shape, movement and position of the hands. We consider that exploiting this relation is an important contribution of our approach when compared to other work on SLR. Since each sign is composed by these elements, the main idea is that establishing these feature-structure relations will improve the pattern recognition system. Moreover, news signs can be easily included in the implemented recognition system using their structural elements description, instead of having to train the whole system for each specific sign (or word or morpheme). Currently, BSL has a universe of around 10,000 signs according to Capovilla, Raphael, and Maurício (2012a, 2012b), thus it would be impractical to develop an SLR system for recognizing each individual sign. In principle, the ideas advanced here could be applied to other systems in SLR elsewhere, just requiring adaptations to the specificities of the sign language of interest.

In our results, to illustrate the proposed methodology, we have selected a set of 34 signs from Capovilla et al. (2012a, 2012b) as shown in Fig. 1. These signs represent 34 distinct morphemes in BSL. They were not chosen at random, but with the advise of a BSL expert, in such a way that these signs represent a wide range of structural elements of the language, and therefore are very representative of the universe of signs in BSL.

These 34 signs are recorded in a database using RGB-D sensors. The following information about each sign is recorded: (i) intensity image; (ii) depth image; (iii) skeleton image and (iv) positions of the body. The authors will make this dataset available in public domain for other researchers. After feature extraction, we employ Support Vector Machines (SVM) for pattern recognition and classification. We used two different kernels for classification and results between 70% and 97% in accuracy were obtained for each relation "extracted feature/structural element in BSL".

The remainder of the paper is organized as follows: first, in Section 2, we present an overview of related work. In Sections 3–5, the structure of BSL, our methodology and our system classification, respectively, are described. Section 6 shows the experiments and results. Finally, in Section 7 we present our conclusions.

2. Related work

Sign Language Recognition (SLR) is an important part of the larger application field of Hand Gestures Recognition (HGR) in Human–Computer Interaction. Complete works describing applications and techniques in HGR can be seen in Watson and College (1993), Chakraborty, Sarawgi, Mehrotra, Agarwal, and Pradhan (2008), Suarez and Murphy (2012), Chen, Wei, and Ferryman (2013), Palacios, Sagüés, Montijano, and Llorente (2013), Zhang, Yang, and Tian (2013) and Ren, Yuan, Meng, and Zhang, 2013 over the last two decades. But when SLR is the subject, some specific issues are recurrent, regardless of the methodology used. Designing solutions that achieve good performance despite all the constraints and difficulties imposed by the complexity of the problem is the real challenge.

Issues related to sign composition, interaction, relationships between hands, different classes of signs, lexicon complexity and non-standard sign translation and other sign language paradigms are discussed in Bossard, Braffort, and Jardino (2003) and Caridakis, Karpouzis, Drosopoulos, and Kollias (2012). Also, in Bossard et al. (2003), authors propose an SLR system design with the following system architecture in three levels: (i) sign recognition, (ii) sign selection and (iii) detection of relationships between signs. Each level itself can involve a very complex system design.

In Parton (2006), the author presents an overview about how techniques in different areas of Artificial Intelligence deal with SLR and translations. Designing SLR systems is a multidisciplinary task that can involve areas such as robotics, virtual reality, computer vision, neural networks, Virtual Reality Modeling Language (VRML), three-dimensional (3D) animation, natural language processing and intelligent computer-aided design. In Loeding, Sarkar, Parashar, and Karshmer (2004) and Ong and Ranganath (2005) we can find discussions about the progress and future of automated systems for SLR. Most recently, authors in Futane, Dharaskar, and Thakare (2012), Ong, Cooper, Pugeault, and Bowden (2012) and Cooper, Ong, Pugeault, and Bowden (2012) present a comparative study of different approaches in Sign Language Recognition. Out of possible approaches described in their work, such as glove based techniques, vision-based techniques and analysis of drawing gestures, we believe that vision-based techniques are the most natural way of constructing a human–computer interface and tackling the related challenges, in this case (i) segmentation of moving hands, (ii) tracking and analysis of hand motion and (iii) recognition itself.

The first issue to address when designing systems for SLR is related to the interaction between the user doing the signs and the computational interface. This interaction can employ glove-based systems or vision-based systems. In the first case, extensive overviews about gloves equipment and sensors in many applications in HGR can be found in Dipietro, Sabatini, and Dario (2008) and Parvini et al. (2009). In these two papers, the authors discuss many applications with data gloves. In Starner and Pentland (1995), Starner, Weaver, and Pentland (1998), Ong et al. (2012) and Cooper et al. (2012), the authors use colored gloves as markers. In Brashear, Starner, Lukowicz, and Junker (2003) multiple sensors are used in a system for mobile sign recognition and in Zhang et al. (2011) multiple sensors are used in a framework for Sign Language Recognition. In both glove-based systems and vision-based systems, sensors and markers are used to provide information about velocity, direction, position, orientation and angles of each hand. Glove-based systems can provide better precision and efficacy in detecting and tracking hands. However, they might be more expensive and require additional equipment for the user.

On the other hand, vision-based systems can be much cheaper and more comfortable for users. Therefore, vision-based systems have received increased attention from researchers and

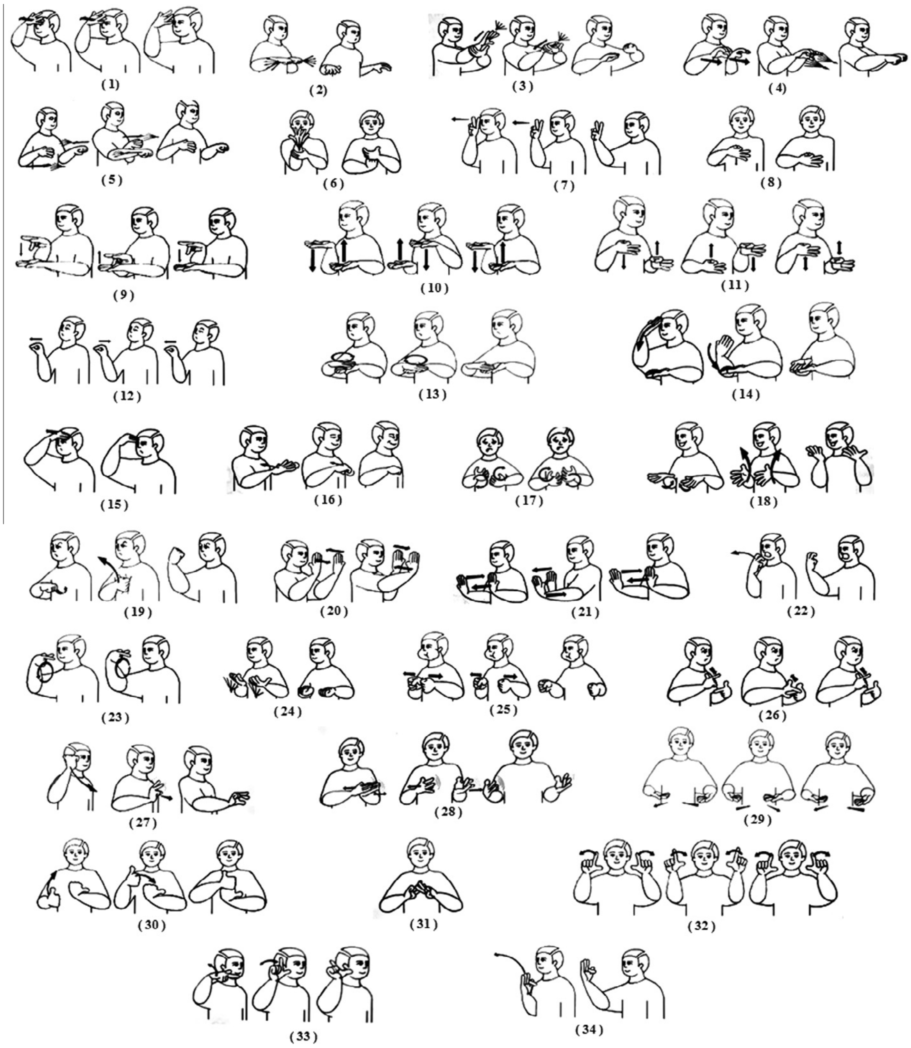


Fig. 1. The 34 signs used in our experiment. (1)"person", (2)"to spread", (3)"to copy", (4)"to catch", (5)"to gather", (6)"to disappear", (7)"to look", (8)"fair", (9)"truth", (10)"weight", (11)"justice", (12)"who", (13)"nothing", (14)"to believe", (15)"to forget", (16)"to love", (17)"to afflict", (18)"to commemorate", (19)"rancor", (20)"assembly meeting", (21)"to compare", (22)"to scream", (23)"to speak", (24)"to absorb", (25)"to fatten", (26)"to quarrel", (27)"perspicacious", (28)"to shine", (29)"maid", (30)"to replace", (31)"prison", (32)"television", (33)"yesterday", (34)"future".

developers. For these systems, general or special cameras capture images as data input. Papers describing system solutions using intensity (RGB), gray, and black and white images as their input can be found in Cui and Weng (2000), Zahedi, Keysers, and Ney (2005b), Zahedi, Keysers, Deselaers, and Ney (2005a), Haberdar and Albayrak (2005), Kawulok (2008), Dreuw, Stein, and Ney (2009), AL-Rousan, Assaleh, and Tala'a (2009) and Diraco, Leone, and Siciliano (2013). Most researchers in HGR are using RGB-D sensors as camera. With this sensor, we can obtain intensity (RGB), depth and skeleton (position) images, hence making it easier to detect and track hands. Authors in Liu and Fujimura (2004) provide a hand gesture recognition method using depth data obtained from a special camera. In recent papers, one of the most used RGB-D sensor in HGR and SLR in recent years is the Microsoft Kinect, mainly because of its low cost and the availability of a Developer Toolkit. More information about the sensor Kinect can be found in Cruz, Lucio, and Velho (2012) and Mankoff and Russo (2013). For HGR we can see the use of Kinect in the following work: (Chaaraoui, Padilla-López, Climent-Pérez, & Flórez-Revuelta, 2014; Chen et al., 2013; Dihl & Musse, in press; Frati & Prattichizzo, 2011; Li, 2012; Palacios et al., 2013; Ramey, González-Pacheco, & Salichs, 2011; Ramirez-Giraldo, Molina-Giraldo, Alvarez-Meza, Daza-Santacoloma, & Castellanos-Dominguez, 2012; Suarez & Murphy, 2012). In the specific case of SLR, Kinect has been used in the papers presented by Zafrulla, Brashear, Starner, Hamilton, and Presti (2011), Uebersax, Gall, Van den Bergh, and Van Gool (2011), Zaki and Shaheen (2011), Phadtare, Kushalnagar, and Cahill (2012), Boulares and Jemni (2012) and Oszust and Wysocki (2013). Authors in Trindade, Lobo, and Barreto (2012) propose a system design based on glove data and vision data together, joining both concepts.

Efforts to create public databases with signs in sign language have been done by some researchers. For instance, (Duduchi & Capovilla, 2006) shows an interface and a dataset for Brazilian Sign Language Recognition. Databases for American Sign Language Recognition are described in Yang, Sarkar, Loeding, and Karshmer (2006), Cooper and Bowden (2007) and Cooper, Pugeault, and Bowden (2011).

In terms of the pattern recognition and classification step of these systems, different methods have been used. In general, architectures based on Neural Networks (NN) are used, as in Huang and Huang (1998), Karami, Zanj, and Sarkaleh (2011), Sole and Tsoeu (2011), Karmokar, Alam, and Siddiquee (2012), Maraqa, Al-Zboun, Dhyabat, and Zitar (2012) and Ahmed (2012). SVM have also been used by Yang and Lee (2013). On the other hand, Hidden

Markov Models (HMM) have been used in many works for classification of the hands, as in Haberdar and Albayrak (2005), Haberdar and Albayrak (2006), Wang, Chen, Zhang, Wang, and Gao (2007), Yin, Starner, Hamilton, Essa, and Rehg (2009), AL-Rousan et al. (2009), Zafrulla et al. (2011), Zaki and Shaheen (2011) and Auephanwiriyakul, Phitakwinai, Suttapak, Chanda, and Theera-Umpon (2013). Authors in Aran, Burger, Caplier, and Akarun (2009) use HMM to detect manual and non-manual signs.

In this work we employ computer vision techniques and RGB-D sensor to segment and detect the movements of the hands. We extract features and relate them to the elements of the phonological structure of the language. The classifier system is trained to learn and detect this relation and these elements, in order to identify the sign (morpheme). In this way, scalability is straightforward since adding new signs to the system requires their description in terms of their structural elements, without needing to retrain and adjust the whole system anew.

3. Methodology for feature extraction

In this section, we describe each step in our methodology for feature extraction in SLR. First, we present, in Section 3.1, the phonological structure of BSL based on the work by Quadros and Karnopp (2004). Next, in Section 3.2, we discuss a video summarization technique employed in the methodology. In Section 3.3, we describe the database structure, and the hardware and software tools used to implement this approach and, in Section 3.4, we describe the method used to detect the region of interest (ROI) using intensity, depth and position images obtained by the RGB-D sensor.

3.1. Phonological structure in Brazilian Sign Language

We explore the phonological structure of BSL in our approach. For more details about BSL structure, see (Quadros & Karnopp, 2004). Four elements make up each sign in BSL: (i) configuration of the hands, (ii) articulation points, (iii) type of movement of the hands and (iv) orientation. However, the classification provided in Quadros and Karnopp (2004) is very extensive and detailed, and classifying each sign or morpheme exactly into values of these elements is a very complex task even for experts in BSL, let alone for a computer system. Therefore, we have to adopt a simplified set of these attributes and elements, which are summarized in Table 1. We describe each element below:

1. Configuration of the hands: This element of the phonological structure provides information about the shape of the hands, but there is no consensus about the amount of possible configurations available in Brazil for the BSL. We use as reference a study conducted in Quadros, Oliveira, and Miranda (2007), where the authors define 134 possible shapes for the hands and group them in 13 groups based on similarity. When treating this element, we have the following attributes and values: (i) group, from 1 to 13; (ii) axis alignment, either x, y, or z; (ii) variation of the configuration during the execution of the sign, yes or No.
2. Articulation points: it provides information about the location of the sign in the neutral space. The neutral space is the area ahead the body where the signs are done. We divide the neutral space into three areas, with one attribute corresponding to the (i) head, (ii) shoulder and (iii) body. The possible values for each attribute are either right, center or left.
3. Type of movement of the hands: provides information about the type of movement for the sign, in case the sign is dynamic. We adopt the following attributes for this element: (i) type, either up, down, right, left, inside or outside; (ii) frequency, either simple or repeated.
4. Orientation of the hand: provide information about the orientation of the palm of the hand in a sign. The following attributes and values are considered: (i) orientation, either up, down, inside, outside, or to the side; and (ii) change of orientation, yes or No.

Table 1
Elements of the phonological structure of BSL used for Sign Language Recognition. Each element consists of a set of attributes with possible values shown in the rightmost column.

Element	Attributes	Values
Configuration of the hands	group	$\{1,\ldots,13\}$
	axis alignment	$\{x,y,z\}$
	variation	$\{yes, no\}$
Articulation points	head	$\{right, center, left\}$
	shoulder	$\{right, center, left\}$
	body	$\{right, center, left\}$
Type of movement	type	$\{up, down, right, left, inside, outside\}$
	frequency	$\{simple, repeated\}$
Orientation	orientation	$\{up, down, inside, outside, totheside\}$
	variation	$\{yes, no\}$

3.2. Video summarization

Video summarization is an important topic in video classification and retrieval, where large videos should be compacted into more representative frames. In our case, we employ video summarization to reduce the number of frames in each sign, in order to avoid the processing of redundant frames that might be present in the recorded sign. Basic techniques for video summarization employ clustering techniques to group similar frames with respect to a given similarity metric.

We formulate the video summarization problem as a classic optimization problem known as the Maximum Diversity Problem (MDP), described in Freitas, Guimaraes, Pedrosa Silva, and Souza (2014). The MDP consists in finding a subset $M(|M| = m)$ from a set $N(|N| = n)$ of elements in such a way that the diversity of the m elements is maximized. Many relations of diversity can be used to define divergence values d_{ij} according to the practical application of the MDP. The problem is concisely described in Kuo, Glover, and Dhir (1993) and repeated in (1), where $x_i = 1$ means that element i is in the subset M:

$$\max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} x_i x_j, \ \textit{subject to} \ \sum_{i=1}^{n} x_i = m \tag{1}$$

where $x_i \in \{0,1\} \ \forall \ i = 1,\ldots,n$

The values d_{ij} in the diversity matrix $[\mathbf{D}]$ are calculated by a problem-specific metric, thus we have to define a metric of diversity between frames in videos. In this case, we can consider two aspects: temporal distance and color differences. In a temporal space, we are looking for the most distant frames in time. In a color space, we want to maximize diversity due to changes in color. The proposed diversity matrix is given by

$$[\mathbf{D}] = [\mathbf{C}] + [\mathbf{T}] \tag{2}$$

where $[\mathbf{C}]$ is a matrix with the color difference between frames i and j and $[\mathbf{T}]$ is a matrix with the time label of each frame.

The MDP is an NP-hard problem that can be solved with either exact methods or heuristics. In order to get a fast and efficient solution to the problem, we employ the Memetic Self-Adaptive Evolution Strategies (MSES) described in Freitas et al. (2014). It is an

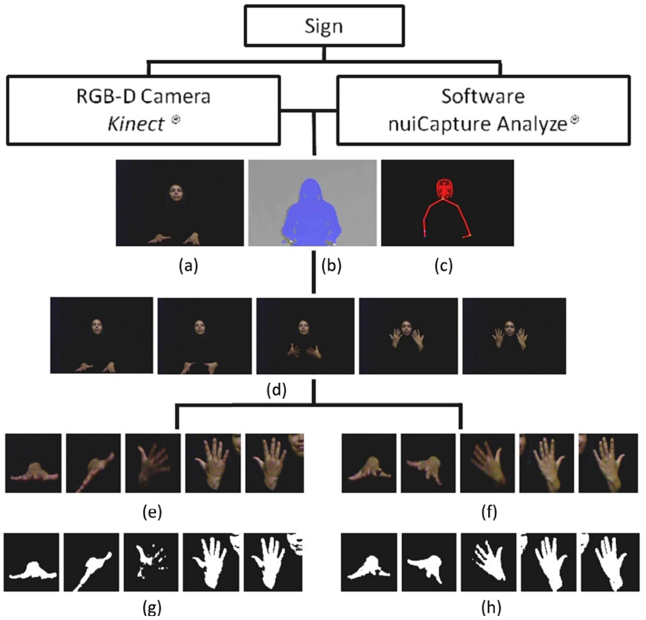


Fig. 2. Detection of region of interest – right and left hands for the sign "to commemorate". Video recorded using software nuiCaptureAnalyze working with Kinect sensor: (a) Intensity, (b) Depth and (c) Position. (d) Videos selected by video summarization using Maximum Diversity Problem to solve it. (e), (f) Region of interest in RGB detected to right and left hands, respectively. (g), (h) Region of interest detected in black and white using skin color detection.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

evolutionary algorithm with some local search heuristics designed for the MDP.

In this paper, we select a fixed number of representative frames for each sign. In Fig. 2(d), we show the $m = 5$ selected frames for the sign "Comemorar"[3] as an example.

3.3. Database and tools

The availability of databases in HGR and SLR for use by the scientific community in computer vision research is still a challenge. Some efforts are presented by Chunli, Wen, and Jiyong (2002), Dias, de Souza, and Pistori (2006) and Cooper and Bowden (2007), regarding data for SLR. Some researchers identify this need for standardized data, promoting collective actions to build large data in HGR using RGB-D sensors, see for instance (Escalera, González, Guyon, & Moeslund, 2013). The first difficulty, in the case of SLR, is reaching a consensus on which signs are representative of a given sign language and if a set of selected signs from a given sign language is useful and representative for testing and evaluating systems designed for other sign languages. Other issues are related to the resistance of the deaf community itself and regional variations as discussed in Schemer (2003), Johnston (2003) and Van Cleve (2003).

Given these difficulties, for the present work, we decided to create our own database, selecting 34 signs in BSL according to morphemes described in Capovilla et al. (2012a, 2012b), each sign being representative of many other signs due to similar characteristics. Given that there is no database available with this subset of signs, we decided to build our own database with five samples of each sign and for only one signer. We therefore focus on the variation in signs not on variations of the signer.

Signs are recorded using the Kinect sensor and the *nuiCapture Analyze* software. Kinect sensor is a low cost RGB-D camera developed by Microsoft for the XBOX video game. The integrated hardware and software in Kinect allows the detection and recording of twenty specific points of the human body. These properties can be found in Cruz et al. (2012), Microsoft (2013) and Mankoff and Russo (2013).

For the generation of the database, two requirements were given: (i) The video must contain the sign space, which is from above the waist. Recording legs is not necessary, since they do not participate in the execution of signs. (ii) The distance should be enough to capture the movements of the arms. Based on these requirements, the distance was set as 1.9 m (6.2 feet) from the Kinect sensor. This value is within the recording range of the sensor, which is between 0.8 m (2.6 feet) and 3.5 m (11 feet).

Using these tools, we obtain three videos recorded at the same time for each sample sign. We can see one frame of one sign in Fig. 2: (a) color frame from the intensity video, (b) depth frame from the depth video, (c) skeleton frame from the skeleton video. The video frame rate is 30 frames per second and all the three videos are recorded in AVI format (audio and video format from Microsoft).

3.4. Detection of the ROI

Each sign has two Regions of Interest (ROI), one for each hand. In this section we describe how we extract the ROI from the video captured by the Kinect sensor and the *nuiCapture Analyze* software.

We first remove the background by using the depth information from the depth video, see Fig. 2(b). With the skeleton video, Fig. 2(c), it is possible to find the coordinates of the hands. In this skeleton video, right hand and wrist are always recorded with blue color, while left hand and wrist are recorded with orange color. With the information about the position of the hands detected in the skeleton video, we can make the segmentation of the hands in the RGB video, already without the background.

In this way, we can extract the ROI for both hands as shown in Fig. 2(e) and (f). Finally, we detect face and hands using an algorithm for skin color detection and convert these images to black and white format, see Fig. 2(g) and (h).

The detection and extraction of the ROI represent the pre-processing part of the methodology for feature extraction. Next, we describe the techniques employed to actually obtain the features and how they relate to the elements of the phonological structure.

4. Relating features to elements in BSL

We extract seven features related to the phonological structure of BSL as described in Section 3.1, which are summarized in Table 2. These features are described in the following next subsections. It is important to highlight that all the features are extracted after the summarization and for each representative frame identified in the summarization step.

4.1. Two-dimensional distance

For obtaining this feature, we calculate the two-dimensional Euclidean distance, in pixels, between both hands and shoulder center. We select the shoulder as reference because the change in its position is negligible when the signer is performing the sign. The positions of the hands and the shoulder are given by the average positions of the pixels in the block of pixels that compose these parts extracted from the skeleton image. We can see this skeleton image in Fig. 3(a).

Individual colors are assigned to hands and shoulder positions among twenty possible body parts detected by using the RGB-D

[3] *tr. v.* to commemorate.

Table 2
Extracted features from the region of interest are shown in the left column. Structural elements in BSL associated with these features are shown in the right column.

Sec.	Analyzed features	Elements in BSL
4.1	Distance between pixels	Articulation points
4.2	Distance in millimeters	Articulation points
4.3	Velocity	Type of movement, Orientation
4.4	Area of hands	Configuration of the hands
4.5	Corners average	Articulation points, Type of movement, Orientation
4.6	Detected lines	Configuration of the hands
4.7	SURF descriptors	Type of movement, Orientation

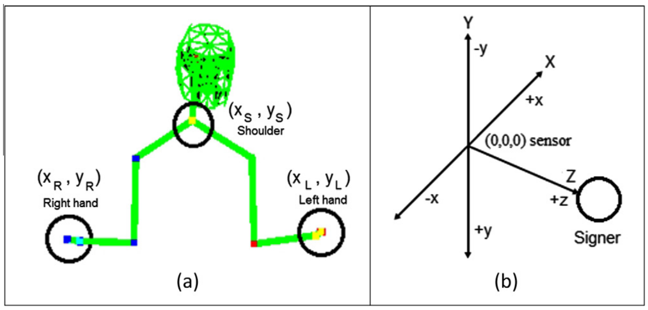


Fig. 3. (a) Feature 1 extracted: two-dimensional Euclidean distance between hands position and center of the shoulder. (b) Reference for feature 2, three-dimensional distance in millimeters.

sensor. In our case, the blue region represents the right hand, the orange region represents the left hand and the yellow color is for the shoulder center.

This feature provides information related to the articulation point of the sign in BSL.

4.2. Three-dimensional distance

Kinect records, using *nuiCapture Analyze*, the three-dimensional distance between sensor and signer in Matlab format (MathWorks, 2012). In Fig. 3(b) we can see that the origin of reference for these values is the sensor.

Three-dimensional distance provides information about the articulation point of the sign in BSL.

4.3. Velocity

We use the optical flow technique to calculate velocity for each sign. For more details about optical flow, see (Horn & Schunck, 1981). This feature provides vectors calculated using brightness difference between frames in a video. Thus, based on these vectors, we obtain information about hands tracking. In Fig. 4(a) we can see vectors calculated using this technique. This feature is related to the following structural elements in BSL: type of movement and orientation of the hands.

4.4. Area of hands

For this feature, we also use the optical flow technique, which allows to segment objects in an image based on image brightness discontinuity. In our region of interest, hands and, in some cases, the face, are the biggest objects. In Fig. 4(b) we can see these areas, all of them detecting hands. We observe, in frame 2, three detected objects: two hands and a face. This occurs because in the sign used as example, "to whip", the hands cross in front of the face. This characteristic can be preserved since this crossing is peculiar for this sign. Hands area is directly related to the Configuration of the hands in BSL.

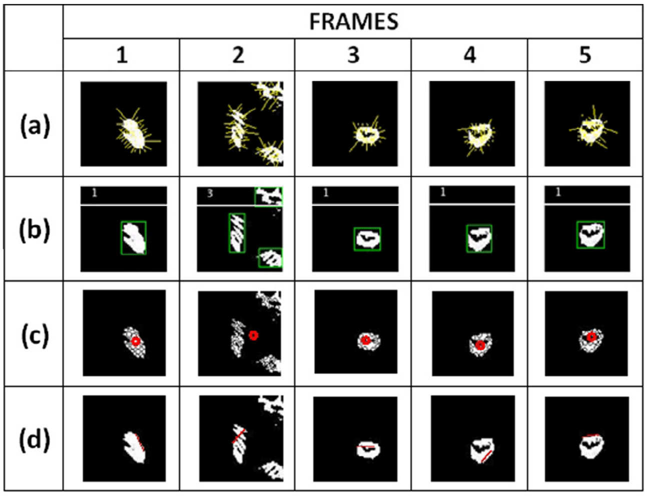


Fig. 4. Features extracted from right hand in the sign "to whip" in BSL: (a) Velocity, (b) Hands area, (c) Corners average and (d) Detected lines.

4.5. Corners average position

The area of the image where there is a large difference between pixels is named corner. Techniques to perform corner detection can be found in Harris and Stephens (1988), Shi and Tomasi (1994) and Rosten and Drummond (2006). In our approach, corners are detected using the Harris corner detector and an average of positions of these corners is the extracted feature. For more details about Harris corner detection technique, see (Harris & Stephens, 1988). This feature is related to the articulation points, the type of movement and orientation in BSL. An example of this feature is available in Fig. 4(c).

4.6. Detected lines

The sixth feature used in this work is given by the lines detected in each image. The technique used here is the Hough Transform, see (Illingworth & Kittler, 1988). This feature is an ordered pair (ρ, θ), because we use polar coordinates. In Fig. 4(d) we show the biggest lines extracted from the black and white image for each frame. These lines can contain information about the configuration of the hands in BSL.

4.7. Amount of common points between frames

The descriptor algorithm SURF (Speed-Up Robust Features) available in Matlab, (MathWorks, 2012), is used to extracted the amount of common points between frames. We can use SURF to find common points between representative frames of the sign using the template matching technique. At each two frames, SURF shows a different amount of points between them, showing the difference of movement between frames. This feature is related to the type of movement and orientation in BSL.

4.8. Assembling the feature vector

By relating each one of the seven features to each of the four structural elements in BSL, we can assemble the feature vector for each element. The number of features in each feature vector depends on the element under consideration. According to Table 2, we have the following feature vectors:

Table 3
Structure of feature vector for M features and n frames. In this study, we analyzed $M = 1, 2, 3$ features, depending on the element, and $n = 5$ frames.

Hand (right or left)			
1	...	n	Frames
1 a M	1 a M	1 a M	Features

Table 4
Structure of feature vector for M features and n frames with variable time added.

Hand (right or left)				
1	...	n		Frames
1 to M	1 to M	1 to M	t_1 to t_n	Features

Table 5
Structure of feature vector for M features and n frames with variable time multiplied.

Hand (right or left)			
1	...	n	Frames
$1.t_1$ to $M.t_M$	$1.t_1$ to $M.t_M$	$1.t_1$ to $M.t_M$	Features

- To the recognition of the element Articulation points, the following features are used (i) distance between pixels, (ii) distance in millimeters and (iii) corners average.
- To the recognition of the element Configuration of the hands, (i) area of hands and (ii) detected lines are used.
- To the element Type of movement we utilize (i) velocity, (ii) SURF descriptors and (iii) corners average to assemble the corresponding feature vector.
- Finally to the element Orientation, the feature vector is composed by the features (i) velocity and (ii) corners average.

The structure of the feature vector is shown in Table 3. Features (1 to M) are extracted from the first representative frame in the video ($n = 1$) and, according to the element under consideration, their values are appended sequentially in the vector. This is done for the n frames in sequence. We end up with a feature vector for each hand and for each one of the four structural elements, which are the input for the classification system presented in the next section.

The time variable is an important information to be considered in the recognition of dynamic signs. In our work, we add this information in two distinct ways in the feature vector. In the experiments, we investigate and compare the performance of the system when using

1. only the feature vector without the time information as in Table 3;
2. the feature vector with the time values of each frame added to it, see Table 4;
3. the feature vector with the time values of each frame multiplied by the values of the features, see Table 5.

5. Classification system

Fig. 5 shows a schematic of the implemented classification system. First, in (a) *Features*, we have as input the features extracted as described in Sections 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7. In (b) *Vector (n frames)*, a feature vector is assembled for each element in the phonological structure of the language, according to Table 2 and as described in Section 4.8. These vectors are the inputs of our classification system, in area (c). Our system classification is in fact implemented by Support Vector Machines (SVM) with two different kernels. This SVM identifies and classifies the linguistic elements based on the extracted features.

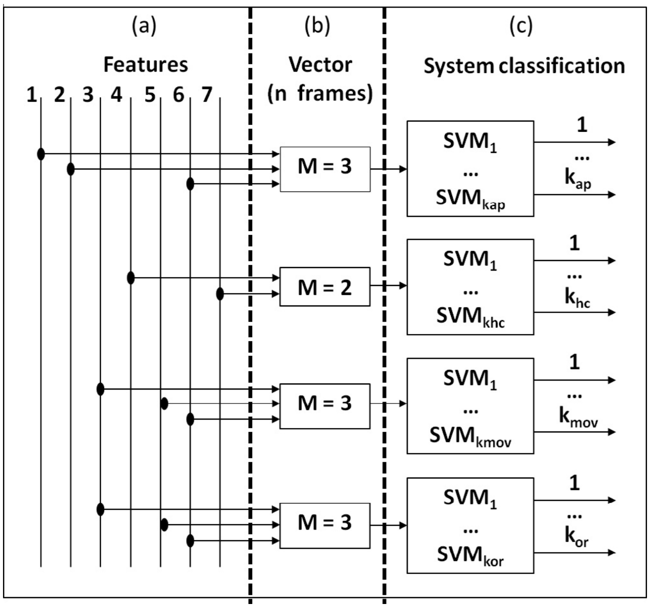


Fig. 5. System schematic. (a) Features: 7 extracted features. (b) Vector: see Table 3 for structure, n = 5 selected frames and M = numbers of features for a given element, see Table 2 for reference. (c) System classification: k_{ap} outputs for the articulation point ($k_{ap} = 5$ for right hand and $k_{ap} = 4$ for left hand); k_{hc} outputs for the configuration of the hands ($k_{hc} = 12$ for right hand and $k_{hc} = 10$ for left hand); k_{mov} outputs for the type of movement ($k_{mov} = 8$ for right and left hands); k_{or} outputs for the orientation ($k_{or} = 7$ for both right and left hands).

We compare results using SVM with two different kernels. First, we use a linear kernel, meaning dot product, as in Eq. (3). Second, we implement a radial basis function kernel as in Eq. (4).

$$f(x_i, x_j) = \sum_{i=1}^{S}\sum_{j=1}^{S} x_i \cdot x_j \tag{3}$$

$$f(x_i, x_j) = \sum_{i=1}^{S}\sum_{j=1}^{S} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \tag{4}$$

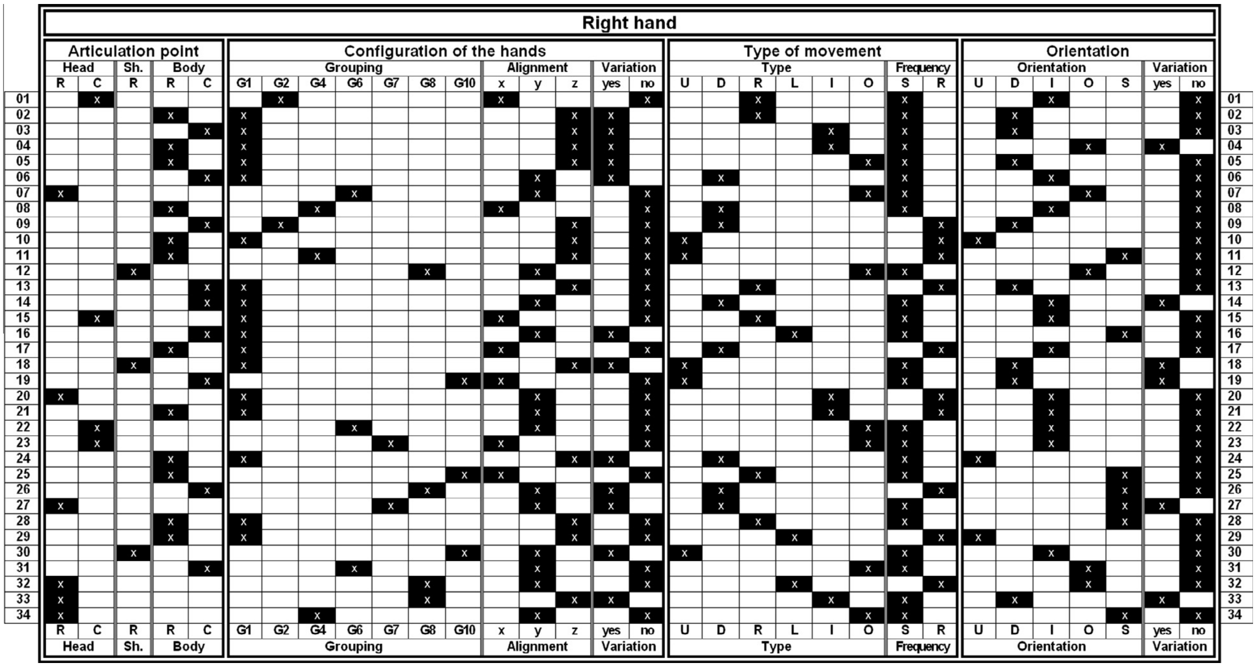
where x_i and x_j are feature vectors of signs in the database.

In Fig. 6 and 7 we show the attributes for system classification for each one of the 34 signs selected in this work.

6. Experiments and results

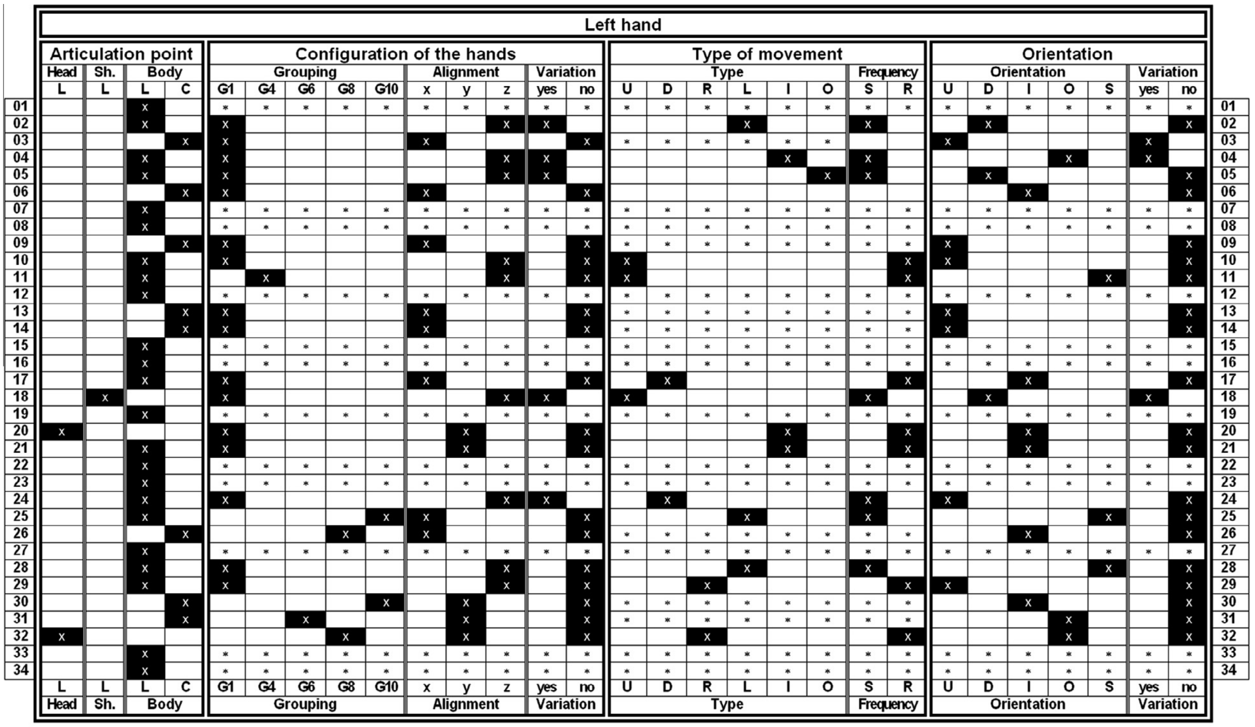
To evaluate our approach, we created a database with 34 specific signs. These signs were selected by a structural division done in Capovilla et al. (2012a), where authors describe 10.236 signs in BSL. Among these, 1.577 signs are mapped and 982 signs are composed by one or more than one of the 34 molecular morphemes chosen. Our database is composed by one example of each one of these 34 morphemes established. According to Capovilla et al. (2012a), it is possible to build other complex signs using these morphemes. In Fig. 1 we show the signs selected.

Remember that the database has five samples of each one of the 34 morphemes, making up a total of 170 examples in the database. We have performed 100 executions for each experiment, varying randomly the three samples of each sign used for training and the two samples used for testing the system, and we show averages and standard deviations for them, using statistical analysis



	Right hand																																
	Articulation point					Configuration of the hands												Type of movement								Orientation							
	Head		Sh.	Body		Grouping							Alignment			Variation		Type						Frequency		Orientation					Variation		
	R	C	R	R	C	G1	G2	G4	G6	G7	G8	G10	x	y	z	yes	no	U	D	R	L	I	O	S	R	U	D	I	O	S	yes	no	
01		x					x						x				x			x				x				x				x	01
02				x		x									x	x				x				x			x					x	02
03					x	x									x	x						x		x			x					x	03
04				x		x									x	x						x		x					x		x		04
05				x		x									x	x							x	x			x					x	05
06					x	x								x		x			x					x				x				x	06
07	x								x					x			x						x	x					x			x	07
08				x				x					x				x		x					x				x				x	08
09					x		x								x		x		x						x		x					x	09
10				x		x									x		x	x							x	x						x	10
11				x				x							x		x	x							x					x		x	11
12			x								x			x			x						x	x					x			x	12
13					x	x									x		x			x					x		x					x	13
14					x	x								x			x		x					x				x			x		14
15		x				x							x				x			x				x				x				x	15
16					x	x								x		x					x			x						x		x	16
17				x		x							x				x		x						x			x				x	17
18			x			x									x	x		x						x			x				x		18
19					x							x	x				x	x						x			x				x		19
20	x					x								x			x					x			x			x				x	20
21				x		x								x			x					x			x			x				x	21
22		x							x					x			x						x	x				x				x	22
23		x								x			x				x						x	x				x				x	23
24				x		x									x	x			x					x		x						x	24
25				x								x	x				x			x				x						x		x	25
26					x						x			x		x			x						x					x		x	26
27	x									x				x		x			x					x						x	x		27
28				x		x									x		x			x				x						x		x	28
29				x		x									x		x				x				x	x						x	29
30			x									x		x		x		x						x				x				x	30
31					x				x					x			x						x	x					x			x	31
32	x										x			x			x				x				x				x			x	32
33	x										x				x	x						x		x			x				x		33
34	x							x						x			x						x	x						x		x	34
	R	C	R	R	C	G1	G2	G4	G6	G7	G8	G10	x	y	z	yes	no	U	D	R	L	I	O	S	R	U	D	I	O	S	yes	no	
	Head		Sh.	Body		Grouping							Alignment			Variation		Type						Frequency		Orientation					Variation		

Fig. 6. Attributes for right hand *versus* 34 signs. *Articulation points*: 1. At right head. 2. At center head. 3. Sh.: shoulder. At right shoulder. 4. At right body. 5. At center body. *Configuration of the hands:* 1. G1. 2. G2. 3. G4. 4. G6. 5. G7. 6. G8. 7. G10. 8. Align in axis x. 9. Align in axis y. 10. Align in axis z. 11. Change in Configuration of the hands or alignment: yes. 12. Change in Configuration of the hands or alignment: No. *Type of movement:* 1. Up. 2. Down. 3. Right. 4. Left. 5. Inside. 6. Outside. 7. Frequency: simple. 8. Frequency: repeated. *Orientation:*1. Up. 2. Down. 3. Inside. 4. Outside. 5. To the side. 6. Variation: yes. 7. Variation: No.



	Left hand																													
	Articulation point				Configuration of the hands										Type of movement								Orientation							
	Head	Sh.	Body		Grouping					Alignment			Variation		Type						Frequency		Orientation					Variation		
	L	L	L	C	G1	G4	G6	G8	G10	x	y	z	yes	no	U	D	R	L	I	O	S	R	U	D	I	O	S	yes	no	
01			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	01
02			x		x							x	x					x			x			x					x	02
03				x	x					x				x	*	*	*	*	*	*			x					x		03
04			x		x							x	x						x		x					x		x		04
05			x		x							x	x							x	x			x					x	05
06				x	x					x				x											x				x	06
07			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	07
08			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	08
09				x	x					x				x	*	*	*	*	*	*	*	*	x						x	09
10			x		x							x		x	x							x	x						x	10
11			x			x						x		x	x							x					x		x	11
12			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	12
13				x	x					x				x	*	*	*	*	*	*	*	*	x						x	13
14				x	x					x				x	*	*	*	*	*	*	*	*	x						x	14
15			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	15
16			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	16
17			x		x					x				x		x						x			x				x	17
18		x			x							x	x		x						x			x				x		18
19			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	19
20	x				x						x			x					x			x			x				x	20
21			x		x						x			x					x			x			x				x	21
22			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	22
23			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	23
24			x		x							x	x			x					x		x						x	24
25			x						x	x				x				x			x						x		x	25
26				x				x		x				x	*	*	*	*	*	*	*	*			x				x	26
27			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	27
28			x		x							x		x				x			x						x		x	28
29			x		x							x		x			x					x	x						x	29
30				x					x		x			x	*	*	*	*	*	*	*	*			x				x	30
31				x			x				x			x	*	*	*	*	*	*	*	*				x			x	31
32	x							x			x			x			x					x				x			x	32
33			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	33
34			x		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	34
	L	L	L	C	G1	G4	G6	G8	G10	x	y	z	yes	no	U	D	R	L	I	O	S	R	U	D	I	O	S	yes	no	
	Head	Sh.	Body		Grouping					Alignment			Variation		Type						Frequency		Orientation					Variation		

Fig. 7. Attributes for left hand *versus* 34 signs. *Articulation points*: 1. At left head. 2. Sh.: shoulder. At left shoulder. 3. At left body. 4. At center body. *Configuration of the hands:* 1. G1. 2. G4. 3. G6. 4. G8. 5. G10. 6. Align in axis x. 7. Align in axis y. 8. Align in axis z. 9. Change in Configuration of the hands or alignment: yes. 10. Change in Configuration of the hands or alignment: No. *Type of movement:* 1. Up. 2. Down. 3. Right. 4. Left. 5. Inside. 6. Outside. 7. Frequency: simple. 8. Frequency: repeated. *Orientation:*1. Up. 2. Down. 3. Inside. 4. Outside. 5. To the side. 6. Variation: yes. 7. Variation: No.

ANOVA. We present results in box plot graphics. In the 100 executions, the sources of randomness are the training and test data and the kernel parameters of the SVM in the classification system.

First, in Section 6.1, we compare two different kernels in terms of the classification results for each element in BSL for both hands. Next, in Section 6.2, we present the results of the comparison between the feature vector with and without the time information, as explained in Tables 3–5.

6.1. Comparison between linear and RBF kernels

In Figs. 8 and 9, we present the classification results over the test data for the right and left hand respectively, using the feature vector only, as in Table 3.

We can see that, for the right hand, the average accuracy is higher for the RBF kernel in the elements Articulation points, Type of movement and Orientation. For the Configuration of the hands, the average accuracy is very similar. Also, for this element, the variation in accuracy is higher, showing the difficulty of correctly classifying the attributes of this element.

For the left hand, the RBF kernel again presents better results, specifically in the elements Configuration of the hands and Orientation. The linear kernel was better for the elements Articulation points and Type of movement. However, many signs are performed only with the right hand, with no participation of the left hand.

Based on these experiments, we selected the RBF kernel for the classification step.

6.2. Comparison of feature vectors

Using only the RBF kernel, we compared the feature vectors described in Tables 3–5,, to see the influence of the time information in the performance of the system.

For the right hand, the feature vector with the time information appended had a slightly superior accuracy, specially for the element Configuration of the hands. For the other elements, there were no statistical difference between adding time or not, and how we add

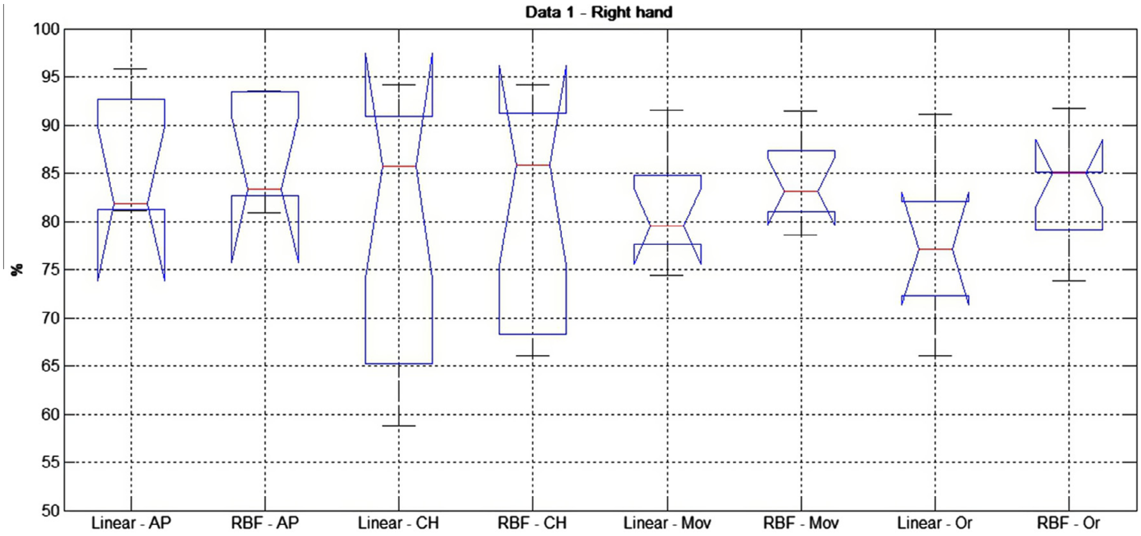


Fig. 8. Comparison between the linear and RBF kernels in the classification of the structural elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the right hand. Results obtained with the feature vector described in Table 3.

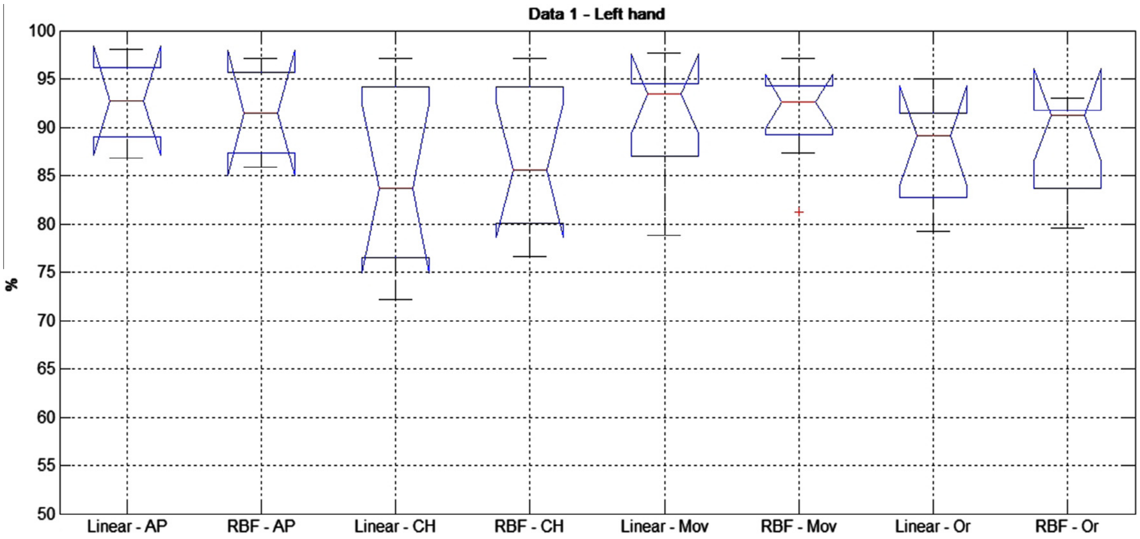


Fig. 9. Comparison between the linear and RBF kernels in the classification of the structural elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the left hand. Results obtained with the feature vector described in Table 3.

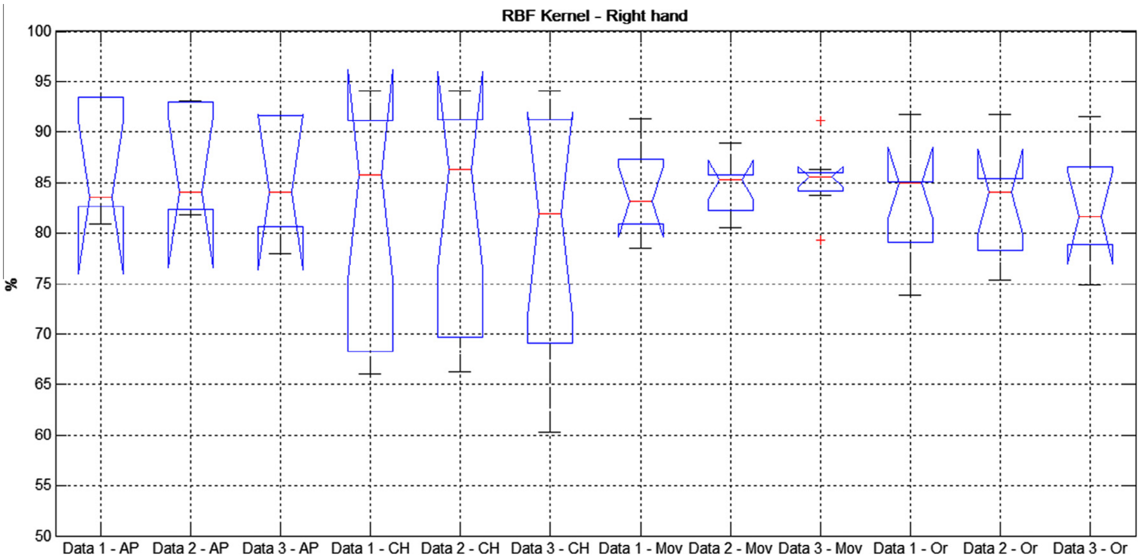


Fig. 10. Accuracy obtained by using the feature vectors described in Tables 3–5 for the RBF kernel in the classification of the elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the right hand.

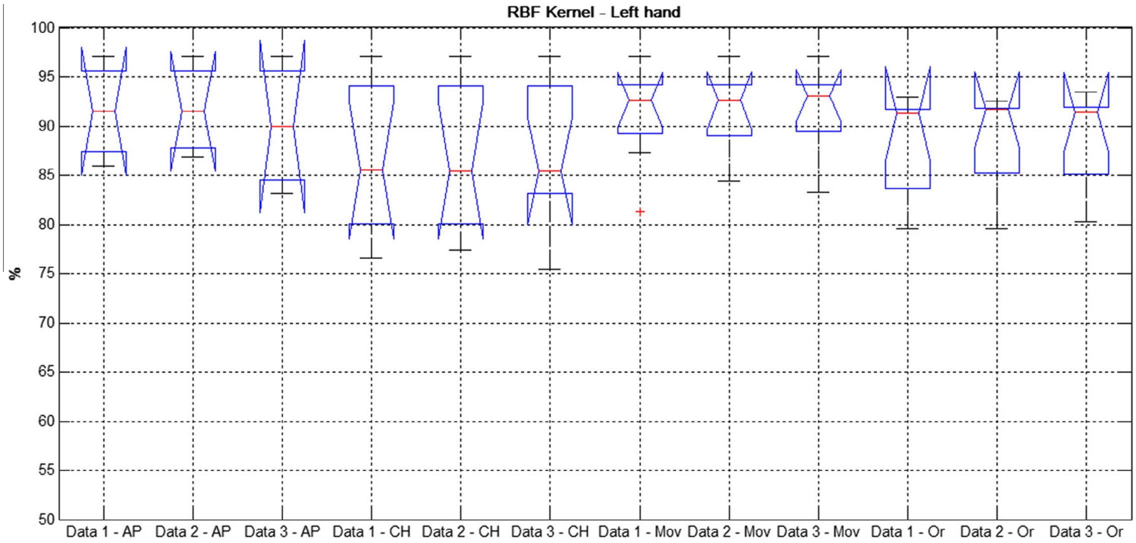


Fig. 11. Accuracy obtained by using the feature vectors described in Tables 3–5 for the RBF kernel in the classification of the elements (i) Articulation points (AP), (ii) Configuration of the hands (CH), (iii) Type of movement (Mov) and (iv) Orientation (Or) for the left hand.

this time information in the feature vector (see Figs. 10 and 11). Nevertheless, based on these experiments, we decided to adopt the feature vector with the time information appended as in Table 4.

6.3. Sign recognition

The experiments in the previous sections show that the accuracy in identifying and classifying the attributes in the structural elements was high. However, this information should be integrated somehow in order to correctly recognize one sign among the others. This can be done in different ways. In this section we perform experiments by using a simple method for combining this information: we compare the output of the classifiers for each attribute of each element in both hands to the values in the lookup tables presented in Figs. 6 and 7.

We calculate the Hamming distance between the output vector and all the lines in the lookup tables, finding the closest k points (with $k = 1, 3, 5$).

Using the RBF kernel and the feature vector in Table 4, the average accuracy is shown in Table 6. In this Table, we compute the accuracy by considering that a correct recognition is achieved when the correct sign is within the k closest neighbors in the lookup table, and we vary k from 1 to 5.

Sign Language Recognition is a very complex and challenging problem, requiring the integration of feature extraction techniques and technology and classification methods. Nonetheless, this experiment shows that extracting features and relating them to the structural elements of the language does help the recognition of the signs. The previous sections showed that the accuracy for classifying attributes of the structural elements is high, which means that the feature extraction methodology proposed is indeed useful to recognize the elements composing the sign. However, the integration of this information for actual sign recognition is still a challenge. Although some signs are correctly recognized with a high rate, other signs still need additional information such that correct recognition be achieved.

Table 6
Average accuracy for the recognition of each one of the 34 signs in the database, based on the closest point in the lookup table.

	Signs	$k = 1$ (%)	$k = 3$ (%)	$k = 5$ (%)
1.	Person	95	100	100
2.	To spread	51	76	93
3.	To copy	67	79	93
4.	To catch	0	6	7
5.	To gather	63	80	87
6.	To disappear	94	97	100
7.	To look	98	100	100
8.	Fair	74	100	100
9.	Truth	43	84	96
10.	Weight	88	98	99
11.	Justice	24	65	72
12.	Who	66	100	100
13.	Nothing	81	96	98
14.	To believe	80	92	97
15.	To forget	75	100	100
16.	To love	20	100	100
17.	To afflict	43	79	91
18.	To commemorate	7	16	21
19.	Rancor	20	99	100
20.	Assembly meeting	7	51	70
21.	To compare	30	79	92
22.	To scream	98	100	100
23.	To speak	70	100	100
24.	To absorb	86	98	97
24.	To absorb	17	43	55
26.	To quarrel	18	45	69
27.	Perspicacious	22	67	100
28.	To shine	88	100	100
29.	Maid	35	70	64
30.	To replace	38	66	78
31.	Prison	37	49	66
32.	Television	0	4	6
33.	Yesterday	0	2	5
34.	Future	1	1	1

If a new sign should be added to the sign recognition system, we just need to add the recorded sign in the database and its description in terms of the structural elements as in Figs. 6 and 7. The number of attributes and elements is always fixed, thus the number of classes in the classification system does not grow with adding new signs. Therefore, in our approach scalability can in principle be manageable given that BSL has more than 10,000 signs.

7. Conclusions

In this paper we have presented a methodology for feature extraction in Brazilian Sign Language that explores the phonological structure of the language and relied on RGB-D sensor for obtaining data. From the RGB-D images we have obtained seven vision-based features. We have related these features to structural elements based on shape, movement and position of the hands. The experiments show that the attributes of these elements can be successfully recognized in terms of the features obtained from the RGB-D images, with accuracy results individually above 80% on average.

RGB-D sensors show a great potential as a tool in hand gesture recognition in general, and in Sign Language Recognition specifically, because it produces three important types of image at the same time: intensity image, depth image and skeleton image. Joining information from these three images provide great power and flexibility in which we can improve algorithms for image processing.

The sign recognition rate was high for some signs and low for others, but in general we can conclude that the proposed feature extraction methodology and the decomposition of the signs into their phonological structural can help expert systems designed for SLR. The proposed methodology can also be applicable not only to BSL but also to other sign languages in the world, because all sign languages can be described in terms of their specific phonological structure. This methodology shows great potential in SLR, but also some challenges. The structure of sign language is very complex as described in Quadros and Karnopp (2004), which was simplified and summarized in the paper. The proposed extracted features could be selected by using feature selection techniques based on quality measures, such as ranking with F-Score or the Pearson correlation coefficient. Feature selection is left for future work. Another important point to be explored is the study of the classification system performance based on the extracted features. However, given the complexity of sign languages in general, perhaps a more fruitful avenue would be exploring a more complex description of the phonological structure in order to improve sign recognition. This paper showed that even a simplified use of this structure is beneficial in the design of systems for SLR. Finally, other ideas about how to integrate the information acquired in the classification of elements into sign recognition should be further explored.

Acknowledgments

The authors would like to thank the support given by the following Brazilian agencies: State of Minas Gerais Research Foundation – FAPEMIG; Coordination for the Improvement of Higher Level Personnel – CAPES; National Council of Scientific and Technological Development – CNPq (Grants 312276/2013-3); and by the Federal Institute of Minas Gerais (IFMG).

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

References

Ahmed, T. (2012). A neural network based real time hand gesture recognition system. *International journal of computer applications* (Vol. 59, pp. 17–22). New York, USA: Foundation of Computer Science.

AL-Rousan, M., Assaleh, K., & Tala'a, A. (2009). Video-based signer-independent arabic sign language recognition using hidden markov models. *Applied Soft Computing, 9*, 990–999.

Aran, O., Burger, T., Caplier, A., & Akarun, L. (2009). A belief-based sequential fusion approach for fusing manual signs and non-manual signals. *Pattern Recognition, 42*, 812–822.

Auephanwiriyakul, S., Phitakwinai, S., Suttapak, W., Chanda, P., & Theera-Umpon, N. (2013). Thai sign language translation using scale invariant feature transform and hidden markov models. *Pattern Recognition Letters, 34*, 1291–1298.

Bossard, B., Braffort, A., & Jardino, M. (2003). Some issues in sign language processing. In A. Camurri & G. Volpe (Eds.), *Gesture Workshop* (pp. 90–100). Springer.

Boulares, M., & Jemni, M. (2012). 3D motion trajectory analysis approach to improve sign language 3d-based content recognition. *Procedia Computer Science, 13*, 133–143. Proceedings of the International Neural Network Society Winter Conference (INNS-WC2012).

Brashear, H., Starner, T., Lukowicz, P., & Junker, H. (2003). Using multiple sensors for mobile sign language recognition. In *Proceedings of the seventh IEEE international symposium on wearable computers* (pp. 45–52). Washington, DC, USA: IEEE Computer Society.

Capovilla, F. C., Raphael, W. D., & Maurício, A. C. L. (2012). Novo Deit-Libras: Dicionário Enciclopédico Ilustrado Trilíngue da Língua Brasileira de Sinais (Libras) baseado em Linguística e Neurociências Cognitivas, Volume I: Sinais de A a H. volume I. Edusp.

Capovilla, F. C., Raphael, W. D., & Maurício, A. C. L. (2012). Novo Deit-Libras: Dicionário Enciclopédico Ilustrado Trilíngue da Língua Brasileira de Sinais (Libras) baseado em Linguística e Neurociências Cognitivas, Volume I: Sinais de I a Z. volume II. Edusp.

Caridakis, G., Karpouzis, K., Drosopoulos, A., & Kollias, S. (2012). Non parametric, self organizing, scalable modeling of spatiotemporal inputs: The sign language paradigm. *Neural Networks, 36*, 157–166.

Chaaraoui, A. A., Padilla-López, J. R., Climent-Pérez, P., & Flórez-Revuelta, F. (2014). Evolutionary joint selection to improve human action recognition with rgb-d devices. *Expert Systems with Applications, 41*, 786–794. Methods and Applications of Artificial and Computational Intelligence.

Chakraborty, P., Sarawgi, P., Mehrotra, A., Agarwal, G., & Pradhan, R. (2008). Hand gesture recognition: A comparative study. *International MultiConference of Engineers and Computer Scientists, 1*, 388–393.

Chen, L., Wei, H., & Ferryman, J. (2013). A survey of human motion analysis using depth imagery. *Pattern Recognition Letters, 34*, 1995–2006.

Chunli, W., Wen, G., & Jiyong, M. (2002). A real-time large vocabulary recognition system for chinese sign language. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and sign language in human-computer interaction. Lecture notes in computer science* (Vol. 2298, pp. 86–95). Berlin, Heidelberg: Springer.

Cooper, H., & Bowden, R. (2007). Large lexicon detection of sign language. In M. Lew, N. Sebe, T. Huang, & E. Bakker (Eds.), *Human-computer interaction. Lecture notes in computer science* (Vol. 4796, pp. 88–97). Berlin, Heidelberg: Springer.

Cooper, H., Pugeault, N., & Bowden, R. (2011). Reading the signs: A video based sign dictionary. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 914–919).

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

Cruz, L., Lucio, D., & Velho, L. (2012). Kinect and rgbd images: Challenges and applications. In *Proceedings of the 2012 25th SIBGRAPI conference on graphics, patterns and images tutorials* (pp. 36–49). Washington, DC, USA: IEEE Computer Society.

Cui, Y., & Weng, J. (2000). Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding, 78*, 157–176.

Dias, J., de Souza, K. P., & Pistori, H. (Eds.). (2006). Conjunto de Treinamento para Algoritmos de Reconhecimento de LIBRAS, II Workshop de Visão Computacional. UFSCar.

Dihl, L., & Musse, S. R. (2014). Recovering 3d human pose based on biomechanical constraints, postures comfort and image shading. *Expert Systems with Applications, 41*, 6305–6314.

Dipietro, L., Sabatini, A. M., & Dario, P. (2008). A survey of glove-based systems and their applications. *Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, 38*, 461–482.

Diraco, G., Leone, A., & Siciliano, P. (2013). Human posture recognition with a time-of-flight 3d sensor for in-home applications. *Expert Systems with Applications, 40*, 744–751.

Dreuw, P., Stein, D., & Ney, H. (2009). Enhancing a sign language translation system with vision-based features. In M. Sales Dias, S. Gibet, M. Wanderley, & R. Bastos (Eds.), *Gesture-based human-computer interaction and simulation. Lecture notes in computer science* (Vol. 5085, pp. 108–113). Berlin, Heidelberg: Springer.

Duduchi, M., & Capovilla, F. C. (2006). Buscasigno: A construção de uma interface computacional para o acesso ao léxico da língua de sinais brasileira. In *Proceedings of VII Brazilian symposium on human factors in computing systems* (pp. 21–30). New York, NY, USA: ACM.

Escalera, S., González, J., Guyon, I., & Moeslund, T. (2013). Multi-modal challenge – 2013.

Frati, V., & Prattichizzo, D. (2011). Using kinect for hand tracking and rendering in wearable haptics. In *2011 IEEE World Haptics Conference (WHC)* (pp. 317–321).

Freitas, A. R. R., Guimaraes, F. G., Pedrosa Silva, R. C., & Souza, M. J. F. (2014). Memetic self-adaptive evolution strategies applied to the maximum diversity problem. *Optimization Letters, 8*, 705–714.

Futane, P. R., Dharaskar, R. V., & Thakare, V. M. (2012). A comparative study for approaches for hand sign language. In *IJCA proceedings on national conference on innovative paradigms in engineering and technology (NCIPET 2012)* (pp. 36–39). New York, USA: Foundation of Computer Science.

Haberdar, H., & Albayrak, S. (2005). Real time isolated turkish sign language recognition from video using hidden markov models with global features. In *Proceedings of the 20th international conference on computer and information sciences* (pp. 677–687). Berlin, Heidelberg: Springer-Verlag.

Haberdar, H., & Albayrak, S. (2006). A two-stage visual turkish sign language recognition system based on global and local features. In F. Esposito, Z. W. Ras, D. Malerba, & G. Semeraro (Eds.), *Foundations of intelligent systems. Lecture notes in computer science* (Vol. 4203, pp. 29–37). Berlin, Heidelberg: Springer.

Harris, C., & Stephens, M. (1988). A combined corner and edge detection. In *Proceedings of the fourth alvey vision conference* (pp. 147–151).

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence, 17*, 185–203.

Huang, C. L., & Huang, W. Y. (1998). Sign language recognition using model-based tracking and a 3d hopfield neural network. *Machine Vision and Applications, 10*, 292–307.

Illingworth, J., & Kittler, J. (1988). A survey of the hough transform. *Computer Vision, Graphics, and Image Processing, 44*, 87–116.

Johnston, A. T. (2003). Language standardization and signed language dictionaries. *Sign Language Studies*, 431–468.

Karami, A., Zanj, B., & Sarkaleh, A. K. (2011). Persian sign language (psl) recognition using wavelet transform and neural networks. *Expert Systems with Applications, 38*, 2661–2667.

Karmokar, B. C., Alam, K. M. R., & Siddiquee, M. K. (2012). Bangladeshi sign language recognition employing neural network ensemble. *International journal of computer applications* (Vol. 58, pp. 43–46). New York, USA: Foundation of Computer Science.

Kawulok, M. (2008). Dynamic skin detection in color images for sign language recognition. In A. Elmoataz, O. Lezoray, F. Nouboud, & D. Mammass (Eds.), *Image and signal processing. Lecture notes in computer science* (Vol. 5099, pp. 112–119). Berlin, Heidelberg: Springer.

Kuo, C. C., Glover, F., & Dhir, K. S. (1993). Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sciences, 24*, 1171–1185.

Li, Y. (2012). Hand gesture recognition using kinect. In *2012 IEEE third international conference on software engineering and service science (ICSESS)* (pp. 196–199).

Liu, X., & Fujimura, K. (2004). Hand gesture recognition using depth data. In *Proceedings of the sixth IEEE international conference on Automatic face and gesture recognition* (pp. 529–534). Washington, DC, USA: IEEE Computer Society.

Loeding, B. L., Sarkar, S., Parashar, A., & Karshmer, A. I. (2004). Progress in automated computer recognition of sign language. In K. Miesenberger, J. Klaus, W. Zagler, & D. Burger (Eds.), *Computers helping people with special needs. Lecture notes in computer science* (Vol. 3118, pp. 1079–1087). Berlin, Heidelberg: Springer.

Mankoff, K. D., & Russo, T. A. (2013). The kinect: A low-cost, high-resolution, short-range 3d camera. *Earth Surface Processes and Landforms, 38*, 926–936.

Maraqa, M., Al-Zboun, F., Dhyabat, M., & Zitar, R. A. (2012). Recognition of arabic sign language (arsl) using recurrent neural networks. *Journal of Intelligent Learning Systems and Applications, 4*, 41–52.

MathWorks (2012). Matlab r2012a (7.14.0.739).

Microsoft (2013). Microsoft kinect for windows.

Ong, E. J., Cooper, H., Pugeault, N., & Bowden, R. (2012). Sign language recognition using sequential pattern trees. In *2012 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2200–2207).

Ong, S. C. W., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 873–891.

Oszust, M., & Wysocki, M. (2013). Polish sign language words recognition with kinect. In *2013 The sixth international conference on human system interaction (HSI)* (pp. 219–226).

Palacios, J. M., Sagüés, C., Montijano, E., & Llorente, S. (2013). Human-computer interaction based on hand gestures using rgb-d sensors. *Sensors, 13*, 11842–11860.

Parton, B. S. (2006). Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence. *Journal of Deaf Studies and Deaf Education, 11*, 94–101.

Parvini, F., Mcleod, D., Shahabi, C., Navai, B., Zali, B., & Ghandeharizadeh, S. (2009). An approach to glove-based gesture recognition. In *Proceedings of the 13th international conference on human-computer interaction. Part II: Novel interaction methods and techniques* (pp. 236–245). Berlin, Heidelberg: Springer-Verlag.

Phadtare, L., Kushalnagar, R., & Cahill, N. (2012). Detecting hand-palm orientation and hand shapes for sign language gesture recognition using 3d images. In *Image processing workshop (WNYIPW)* (pp. 29–32). Western New York.

Quadros, R. M., & Karnopp, L. B. (2004). Língua de sinais brasileira: estudos linguísticos. Artmed.

Quadros, R. M., Oliveira, J., & Miranda, R. D. (2007). Núcleo de aquisição de línguas de sinais - universidade federal de santa catarina – identificador de sinais.

Ramey, A., González-Pacheco, V., & Salichs, M. A. (2011). Integration of a low-cost rgb-d sensor in a social robot for gesture recognition. In *Proceedings of the sixth international conference on human-robot interaction* (pp. 229–230). New York, NY, USA: ACM.

Ramirez-Giraldo, D., Molina-Giraldo, S., Alvarez-Meza, A., Daza-Santacoloma, G., & Castellanos-Dominguez, G. (2012). Kernel based hand gesture recognition using kinect sensor. In *2012 XVII Symposium of image, signal processing, and artificial vision (STSIVA)* (pp. 158–161).

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

Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. In *Proceedings of the ninth European conference on computer vision – volume part I* (pp. 430–443). Berlin, Heidelberg: Springer-Verlag.

Schemer, G. M. (2003). From variant to standard: An overview of the standardization process of the lexicon of sign language of the Netherlands. *Sign Language Studies*, 469–486.

Shi, J., & Tomasi, C. (1994). Good features to track. In *IEEE computer society conference on computer vision and pattern recognition, 1994. Proceedings CVPR* (pp. 593–600). IEEE.

Sole, M., & Tsoeu, M. (2011). Sign language recognition using the extreme learning machine. In *AFRICON* (pp. 1–6).

Starner, T., & Pentland, A. (1995). Real-time american sign language recognition from video using hidden markov models. In *International Symposium on Computer Vision, 1995. Proceedings* (pp. 265–270).

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

Suarez, J., & Murphy, R. (2012). Hand gesture recognition with depth images: A review. In *RO-MAN: The 21st IEEE international symposium on robot and human interactive communication, 2012 IEEE* (pp. 411–417).

Trindade, P., Lobo, J., & Barreto, J. (2012). Hand gesture recognition using color and depth images enhanced with hand angular pose data. In *2012 IEEE conference on multisensor fusion and integration for intelligent systems (MFI)* (pp. 71–76).

Uebersax, D., Gall, J., Van den Bergh, M., & Van Gool, L. (2011). Real-time sign language letter and word recognition from depth data. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 383–390).

Van Cleve, J. V. (2003). Lexicography and the university: Making the gaulladet dictionary of american sign language. *Sign Language Studies, 4*, 487–500.

Wang, Q., Chen, X., Zhang, L. G., Wang, C., & Gao, W. (2007). Viewpoint invariant sign language recognition. *Computer Vision and Image Understanding, 108*, 87–97.

Watson, R., & College, T. (1993). A survey of gesture recognition techniques. Technical Report. Trinity College Dublin.

Yang, H. D., & Lee, S. W. (2013). Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters, 34*, 2051–2056.

Yang, R., Sarkar, S., Loeding, B., & Karshmer, A. I. (2006). Efficient generation of large amounts of training data for sign language recognition: A semi-automatic tool. In K. Miesenberger, J. Klaus, W. Zagler, & A. Karshmer (Eds.), *Computers helping people with special needs*. *Lecture notes in computer science* (Vol. 4061, pp. 635–642). Berlin, Heidelberg: Springer.

Yin, P., Starner, T., Hamilton, H., Essa, I., & Rehg, J. M. (2009). Learning the basic units in american sign language using discriminative segmental feature selection. In *Proceedings of the 2009 IEEE international conference on acoustics, speech and signal processing* (pp. 4757–4760). Washington, DC, USA: IEEE Computer Society.

Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., & Presti, P. (2011). American sign language recognition with the kinect. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 279–286). New York, NY, USA: ACM.

Zahedi, M., Keysers, D., Deselaers, T., & Ney, H. (2005a). Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Proceedings of the 27th DAGM conference on pattern recognition* (pp. 401–408). Berlin, Heidelberg: Springer-Verlag.

Zahedi, M., Keysers, D., & Ney, H. (2005b). Appearance-based recognition of words in american sign language. In *Proceedings of the second iberian conference on pattern recognition and image analysis – volume part I* (pp. 511–519). Berlin, Heidelberg: Springer-Verlag.

Zaki, M. M., & Shaheen, S. I. (2011). Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters, 32*, 572–577.

Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K., & Yang, J. (2011). A framework for hand gesture recognition based on accelerometer and emg sensors. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 41*, 1064–1076.

Zhang, C., Yang, X., & Tian, Y. (2013). Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1–8).