Continuous Sign Recognition of Brazilian Sign Language in a Healthcare Setting

José Elías Yauri Vidalón and José Mario De Martino

***Abstract*—Communication is the basis of human society. The majority of people communicate using spoken language in oral or written form. However, sign language is the primary mode of communication for deaf people. In general, understanding spoken information is a major challenge for the deaf and hard of hearing people. Access to basic information and essential services is challenging for these individuals. For example, without translation support, carrying out simple tasks in a healthcare center such as asking for guidance or consulting with a doctor, can be hopelessly difficult. Computer-based sign language recognition technologies offer an alternative to mitigate the communication barrier faced by the deaf and hard of hearing people. Despite much effort, research in this field is still in its infancy and automatic recognition of continuous signing remains a major challenge. This paper presents an ongoing research project designed to recognize continuous signing of Brazilian Sign Language (Libras) in healthcare settings. Health emergency situations and dialogues inspire the vocabulary of the signs and sentences we are using to contribute to the field.**

***Index Terms*—Brazilian sign language, Libras, Continuous signing, Sign language recognition.**

I. Introduction

SINCE the origin of humanity, people have been using language to convey messages, concepts, ideas, moods, feelings, and emotions. While the majority of people use oral language as a system of communication, deaf people use sign language. Contrary to the common belief that sign language is a universal language, there are different sign languages scattered around the world, e.g., American Sign Language (ASL) in the USA [1], British Sign Language (BSL) in England, German Sign Language (GLS) in Germany, Portuguese Sign Language (PSL) in Portugal, Brazilian Sign Language (Libras) in Brazil [2], etc. Each sign language has particularities that makes it unintelligible to others.

Sign language is a visual-spatial language that uses agreed gestures to convey meaning. Gestures can be manual and non-manual. Manual gestures are performed by movements of the fingers, hands, and arms, while non-manual gestures are composed by movements of body and head, eye-gaze orientations and facial expressions. Gestures can also be static or dynamic. The former consists of the single positioning of body or limbs forming a posture, e.g., pointing. The later consists of a set of positions that change over time, e.g., clapping. A sign consists of a gesture, or gestures performed simultaneously, which has an agreed meaning by the deaf community. During translation, the meaning of a sign can be interpreted as one or more words of an oral language.

In sign language, a sign can be strictly manual or non-manual, or a combination thereof. Manual signs can be one-handed (those performed only by the dominant hand) or two-handed (those performed by both hands simultaneously, where the second hand is called the non-dominant hand). Non-manual signs usually involve features such as mouth movements and facial expressions. As mentioned before, according to the temporal variation, a sign can be static or dynamic. For example, in Libras, the sign representing the letter A of the alphabet is a static sign, and the sign GRIPE [*Flu*] is a dynamic sign which also has facial expressions (Fig. 2d). Additionally, similar to conversations in oral language where the speaker says a sequence of words to construct clauses and sentences, a signer uses a sequence of signs to convey meaning (e.g., the sequence of signs HOJE EU ADOECER GRIPE in Libras means [*Today I got sick with the flu*], Fig. 2a–d).

Wherever there are deaf people, sign language emerges spontaneously. Sign language is acquired by children born into deaf families and transmitted from generation to generation, primarily through special schools and deaf adults. However, life is not easy for the deaf. Access to public services and information in a world dominated by oral communication is difficult for them. For instance, accessing basic services like healthcare, education, legal and other services without a sign language interpreter can be stressful or impossible for many deaf people, who feel marginalized, ignored and isolated by society. Although interpreters can be of great help in better communication between deaf people and those without hearing impairment, the lack of interpreters in number and fluency limits their availability to a few situations. Computer-based sign language recognition (SLR) systems have become a promising technology that can help overcome these constraints. The main goal of SLR is to be able to recognize and translate sign language.

Since the Nineties, many SLR approaches have been proposed [3]. Most proposals have focused on isolated signs (both static and dynamic), while only a few approaches have focused on continuous signing (e.g., sentences, phrases, and discourses). When the recognition process has to deal with dynamic signs, sophisticated methods are required to manage the temporal information.

Although sophisticated data capture devices can help in the recognition process, for instance, data gloves that provide detailed information about the hand and fingers, or an elec-

The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vanessa Testoni.

José Elías Yauri Vidalón and José Mario De Martino are with the Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil (e-mails: {elias,martino}@dca.fee.unicamp.br.)

This paper is part of the Special Issue on Vision-based Human Activity Recognition.

Digital Object Identifier 10.14209/jcis.2015.10

tromagnetic 3D position tracker that tracks a specific limb of the body, recognition methods based on video cameras (called *vision-based* methods) are still preferred since cameras are less expensive and do not interfere with the signing process (the signer is not required to wear sensors). Despite all the technological advances, the SLR process is still challenging, mainly due to the simultaneous-sequential nature of sign production which conveys meaning through many modes at once [1]; the signer's inflections and variability [2]; the high degree of freedom in human movement [4]; and the large vocabulary of signs [3].

In this paper, we present an ongoing research project that aims to develop an SLR system for continuous sentences of Brazilian sign language (Libras). Although there is no official number of Libras speakers, according to the 2010 Brazilian census [5], the number of Brazilian deaf and hard of hearing people was estimated at almost two million. In our proposal, instead of trying to recognize only isolated signs or single combinations of them, we aim to develop a system that recognizes the basic, real-life vocabulary and continuous sentences used in a specific domain. Among the major difficulties reported by the deaf individuals themselves [6], we chose the area of "health emergencies in a hospital" since it provides high social value. Therefore, 58 signs from Libras were selected to provide a basic vocabulary. We then recorded video data of isolated signs and 15 sentences by using a single Kinect sensor [7]. Next, we describe our approach to recognizing isolated signs and our attempt to recognize signs in sentences based on the Hidden Markov Model (HMM) [8]. The remainder of this paper is organized as follows: Section II summarizes the related works, Section III presents our proposal, and finally, the conclusions are exposed in Section IV.

II. Related work

From a linguistic point of view, in 1960 Stokoe [9] was the first to demonstrate that signs (in the ASL lexicon) are comprised of a relatively small number of meaningless subunits (like phonemes in speech) that may be recombined to produce a potentially large lexicon. He proposed three sign parameters: hand shape, hand location, and hand movement. Later, parameters of hand orientation and facial expression were introduced into the sign language phonology. Although both hands might be involved in the formation of many signs, there is only one primary active articulator in the lexical item, the dominant hand. The non-dominant hand either articulates nothing, or mimics what the dominant hand is doing, or serves as a place of articulation [10]. Parameters related to the hand establish the manual signal features (MS), while the other parameters define the non-manual signal features (NMS).

During the Eighties, besides the simultaneous sign production theory proposed by Stoke, theories of the sequential nature of sign production were proposed. Models like the movement-hold [11], posture-movement [12], posture-detention [13], and others were proposed to describe signs in a sequence of subunits or feature bundles, one of which can be independently affected by morphological processes and phonological rules. The value of the linguistic basis lies on the fact that it can help to understand, decompose, and model the signs in order to improve the recognition process.

From a computational point of view, the first SLR systems emerged in the early Nineties [3], most of them focused on isolated signs, both static and dynamic, and a few on recognizing continuous sentences. In the following paragraphs we discuss the works that inspired our proposal on continuous sign language recognition.

Starner and Pentland [14] presented the seminal work in recognizing ASL sentences using HMMs [8]. In this approach, the signer wears two distinctly colored gloves for each hand and sits in front of a camera. Based on a vocabulary of 40 signs, they tested 99 sentences of constrained structure (i.e., a personal pronoun, verb, noun, adjective, and personal pronoun again, in this order). Each sign is modeled using an HMM of four states and multidimensional Gaussian observations of the 2D features extracted from the hand. They achieved a recognition rate of 97% for sentences. They used a rigid grammar model for constructing the sentence, however the sentence structure in sign language may have a flexible word order.

Later, Bauer and Heinz [15] proposed a system to recognize German Sign Language (GSL) sentences based on HMMs. The signer wears colored cotton gloves in order to reduce the complexity of the hand feature extraction and tracking. Because they map the entire sentence using HMMs, the variation produced between the transitions of two consecutive signs is also incorporated into the model parameters. Afterward the model parameters of the single sign (also modeled as an HMM) is reconstructed from these data in order to recognize the sign. So, in order to detect the sign boundaries in the sentences, they take into account all possible initial and ending locations of the sign; and the path search is optimized by means of a beam search algorithm. Based on a vocabulary of 97 signs, they achieve a recognition rate of 91.7%. In a subsequent work, Bauer and Kraiss [16] proposed the extraction of subunits of a sign in order to find similar feature vectors by using the k-means algorithm (a feature vector is encoded according to its cluster). Each cluster/subunit is modeled as an HMM and a sign consists of a concatenation of these subunits. From 12 signs, they extracted 10 subunits, achieving a recognition rate of 80.8%. The subunits themselves are different from sign phonemes because they were determined via clustering instead of a linguistic approach. Moreover, the number of signs and subunits are not enough to draw generalizations.

Vogler and Metaxas [4] proposed to decompose signs based on the Movement-Hold model [11] in order to recognize ASL sentences. Movements are segments in which some aspects of the signer's configuration changes; while Holds are segments in which all aspects of the signer's configuration remain stationary. Thus, a sign is broken into movement and hold segments, each of which is considered a subunit or phoneme. The features are provided by a sophisticated electromagnetic system which gives the 3D location and motion of the hand and arm. Since this approach takes into account a two-handed sign, Vogler and Metaxas modeled the sign subunits by using parallel HMMs, i.e., one model for each hand. From

a vocabulary of 22 signs and 99 sentences, they showed that the parallel model of the two hands results in better recognition than the single modeling of the dominant hand —at a recognition rate of 84.85% and 80.81%, respectively. In a later work, in order to make the simultaneous aspects of ASL more tractable, Vogler and Metaxas [17] proposed to model the hand shape parameter of a sign by using additional independent channels. Using data provided by an electronic glove, they modeled the hand shape based on the degree of openness of the finger [10], achieving a recognition rate of 87.88%.

Yang et al. [18] proposed a technique to recognize signs in sentences by using an adaptive threshold based on Conditional Random Field (CRF) [19]. They constructed a dictionary to distinguish between sign and non-sign patterns. No transition models or grammar rules are required to spot the signs, however the threshold fitting can be difficult to achieve. With 48 ASL signs, they formed 98 sentences (ranging from three to eight signs per sentence), achieving a recognition rate of 87%. In their later work, Yang et al. [20] worked to simultaneously recognize signs and finger spellings in sentences. Using a basis of 24 signs and 17 alphabetic ALS letters, they experimented with 98 sentences. By using Hierarchical-CRF and BoostMap embedding methods, they recognized signs and finger spellings at rates of 83% and 78%, respectively.

More recent research has focused on the integration of non-manual signal features (NMS) of signs such as body postures and facial expressions. The importance of NMS lies in the fact that it can completely change the meaning of the sign [1][2]. In this way, Yang et al. [21] presented a framework that recognizes both manual and non-manual signs in three steps. Firstly, a Hierarchical-CRF is used to detect segments of manual signals. Next, the BoostMap embedding method is used to detect hand shapes in segmented signs and to recognize finger spellings. Finally, Support Vector Machine (SVM) is applied to recognize facial expressions if there is any ambiguity in the two previous steps. Using this approach, data were collected by using multiple cameras: two orthogonal cameras (frontal and lateral view) focused on manual data, with a specific frontal camera focused on the face to capture facial expressions. Using the basis of 24 signs, 17 alphabetic letters and 5 facial expressions from ASL, they tested 98 sentences, achieving a recognition rate of 84%.

With new developments in sensor technology, which includes features that go beyond traditional RGB cameras, new possibilities for data gathering and interactions have become available. Zafrulla et al. [22] used Kinect to capture data and from it developed a sentence verification system for an electronic game designed for deaf children. Taking the RGB-D (color and depth range data) image and the skeleton information provided by the Kinect device, Zafrulla et al. extracted features from the depth image and the skeleton data. Working with 60 sentences of constrained structure based on a vocabulary of 19 signs, they achieved a recognition rate of 51.50% and 76.12% for signers who were both seated and standing, respectively.

Regarding the works concerned with Brazilian Sign Language (Libras) recognition, Pizzolato et al. [23] proposed recognizing 15 finger-spelled words of Portuguese which in turn are based on 17 static signs and one dynamic sign (the sign for the letter J) of the Libras alphabet by using a two-layer Artificial Neural Network (ANN). Only one finger-spelled word starts with J, and it is modeled as three static hand postures. The classification was performed in two stages: first, words with similar hand postures are grouped together for preliminary ANN classification; next, another ANN is applied to disambiguate some confusion between letters. Once the letters of the sequence have been identified, these letters are turned into HMMs (one HMM model for each word, in which the number of states depends on the number of letters of the word). They achieved a recognition rate of 91.1%. More recently, Souza and Pizzolato [24] presented a system able to recognize both finger-spelled words and isolated signs. They worked with 46 hand shapes and 13 Libras signs. They used SVM to classify hand shapes, whereas the signs were classified by using hidden-CRF (one model for the whole sign).

After reviewing the literature, we summarize the major issues in continuous sign language recognition (SLR):

- The simultaneous-sequential nature of sign production (the combination of manual and non-manual parameters while the sign is performed) challenges any SLR algorithm.
- The high degree of freedom (DoF) of human movement leads to partial or complete occlusion of body parts.
- The high degree of freedom (DoF) of the hand movement produces similar hand shapes and self-occlusion of fingers.
- The motion and appearance of the sign may vary significantly even for the same signer.
- In sentences, a sign is affected by the preceding signs which leads to co-articulated movement between signs. As a result, there are not always clear boundaries between two adjacent signs.
- The strong signer dependency on recognition systems. The recognition accuracy decreases dramatically when the system is tested with a signer whose data have not been used to train the system.
- The lack of attention to non-manual features. How to identify which elements are important to the sign and which elements are coincidental is a major concern. Additionally, the merging of manual and non-manual features is still unresolved.
- The linguistic properties of sign language are still under study and discussion. Moreover, each sign language has its own lexicon and particular grammar properties.
- Public availability of data sets are limited both in quantity and quality for the recognition task.

III. SLR System Overview

The main goal of our project is to develop a system which recognizes continuous signing of Libras [2]. In order to make more tractable the issues described in the previous section, we propose:

- To work within a real-life lexical domain. Due to its high social value, we have chosen the lexicon of healthcare

centers. There is evidence that access to health services for the deaf is difficult [6]. Deaf people can feel misunderstood, marginalized, and frustrated when they seek medical attention.

- To deal with the basic vocabulary used in healthcare settings. The case study of "health emergencies in a hospital" provides us with a basic vocabulary of signs and sentences with different level of difficulty. For example, there are one-handed signs and two-handed signs that may differ only in some parameters of the sign or require touching a specific body part which may imply local movement of fingers, but may also include major or minor non-manual signals.
- To use RGB-D cameras to collect a database of signs and sentences. Depth cameras reduce the effects of lighting variation, occlusion and cluttered background that affect color cameras. According to our research, there are no databases on the domain of healthcare settings. Moreover, the data will be useful for further research and testing.
- To develop a signer-independent recognition system. We are working with two signers, however just one signer is used for learning and tuning the model parameters. We intend to use the other signer during the testing phase for performance evaluation and providing feedback on the model.
- To use linguistic foundations of sign language to model and represent features of both signs and sentences. The sign's linguistics provide decomposition tools and signal process production representations.
- To model sign features by using probabilistic time-series data models since they have to deal with varying information over time.

Next, we describe the architecture of the system and the stages of the proposed recognition system, as well as the preliminary results achieved so far.

A. System Structure

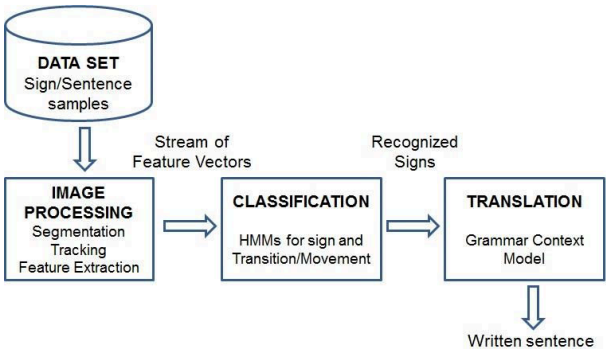


Fig. 1. Overview of the system architecture.

The structure of the proposed continuous sign/sentence recognition system is shown in Fig. 1. In summary, the collected sign and sentence samples are sent into the image processing module to extract salient features. These features are sent into the classification module to build models of both signs and transition/movements in order to segment continuous sentences. In our approach, based on phonetic transcription of the sign [13], we select frames that provide us a short representation of the sign motion path. Other motions are considered transition movements. The recognized signs are then sent to the translation module which uses language models to provide a written sentence in Portuguese.

B. Preliminary Results and Discussions

Here we describe the proposals and preliminary results accomplished in each module.

1) Data Set: The data set consists of signs and sentences collected by a Kinect sensor [7]. For each sample data, we recorded the color, depth, and skeleton information. To further simplify the image processing, signers wore a black sweater and stood in front of a single Kinect at a distance of 1.2–1.8 meters away.

The vocabulary basis of 58 signs is shown in Table I. Notice that the signs are written in uppercase and their translations in brackets. To make more understandable the chosen signs, they are classified into parts of speech categories (e.g., pronoun, noun, adjective, verb, and adverb). This helps to understand how signs can be joined together to make readable sentences using the grammar of a particular sign language.

TABLE I
SIGNS IN OUR DATA SET

Pronouns	EU [I], MEU [My], VOCÊ [You], SEU [Your], ELE [He], DELE [His]
Nouns	NOME [Name], HOSPITAL [Hospital], GRIPE [Flu], CORAÇÃO [Heart], DOR [Pain, Ache], DOENÇA [Disease], MÉDICO [Physician, Doctor], INJEÇÃO [Injection, Shot], ESTETOSCÓPIO [Stethoscope], CONSULTA MÉDICA [Medical consultation], SANGUE [Blood], RECEITA MÉDICA [Medical prescription], ENFERMAGEM [Nursing], EXAME [Medical Exam], COMPRIMIDO [Pill], FEBRE [Fever], PEITO [Chest], ESTÔMAGO [Stomach], DENTE [Tooth], REMÉDIO [Medicine], CABEÇA [Head], HOMEM [Male], MULHER [Female], ANO [Year], PASSADO [Past], FUTURO [Future]
Verbs	TER [Have], NÃO TER [Do not have], IR [Go], VIR [Come], DOER [Ache], SENTIR [Feel], AGENDAR [Schedule], INJETAR [Inject], VACINAR [Vaccinate], CONSULTAR [Consult], CURAR [Cure], ADOECER [Get sick], QUERER [Want], NÃO QUERER [Do not want]
Adjectives	INFLAMADO [Inflamed], SAUDÁVEL [Healthy], DOENTE [Sick], POUCO [Little, Few], BEM [Well], MAL [Not well]
Adverbs	ONTEM [Yesterday], HOJE [Today], AMANHÃ [Tomorrow], AGORA [Now], AQUI [Here], MUITO [Much, Many]

These collected signs present unique difficulties; for instance, the one-handed signs for I and my have the same location, but different hand configuration; the two-handed signs for physician and year have the same location and some touching between hands, but different movements and hand

configurations; and the signs for flu, tooth, and get sick rely on facial expressions.

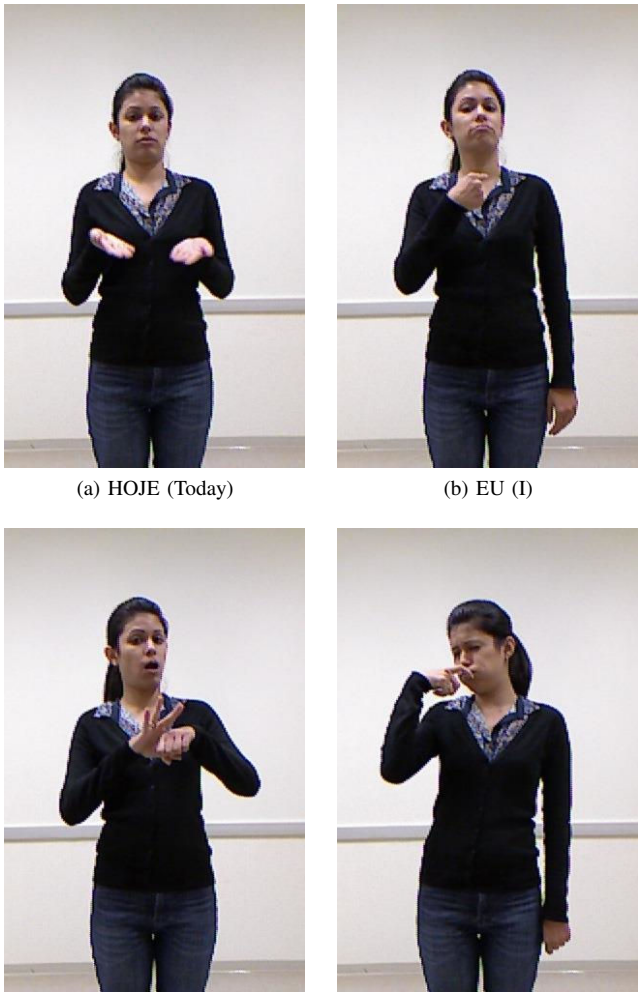
The 15 sentences collected and explored in this work are shown in Table II. Sentence length varies from three to eight signs each. Pronouns are usually omitted depending on the topic, and the word order follows the grammatical rules of Libras, which is different from Portuguese grammar. As we mentioned before, when a sentence is translated, the number of words could vary in order to maintain the meaning of the message. Snapshots of the sentence HOJE EU ADOECER GRIPE, which means *Today I got sick with the flu*, are presented in Fig. 2. The sign for EU [*I*] is one-handed (Fig. 2b), while signs HOJE [*Today*] and ADOECER [*To get sick*] are two-handed (Fig. 2.a and Fig. 2c); and the sign for GRIPE [*Flu*] is one-handed combined with facial expressions (Fig. 2d).

TABLE II
SENTENCES IN OUR DATA SET

DOR DENTE ONTEM EU TER [Yesterday I had a toothache]
MEU MÉDICO BOM [My doctor is good]
HOJE EU ADOECER GRIPE [Today I got sick with the flu]
EU AQUI HOSPITAL, CONSULTA MEDICA TER [Today I am in the hospital, because I have a medical consultation]
PASSADO MUITOS ANOS DOENÇA CORAÇÃO TER, AGORA CURAR [Many years ago, I had a heart illness, but now I am cured]
SENTIR MAL EU, AGORA IR HOSPITAL [I am feeling sick, now I am going to the hospital]
ESTETOSCÓPIO MÉDICO TER [The doctor has a stethoscope]
EXAME SANGUE MEU [This is my blood exam]
RECEITA-MÉDICA DELE TER MUITO COMPRIMIDO [His medical prescription consists of many pills]
AMANHÃ VACINAR EU HOSPITAL [Tomorrow, I am going to get vaccinated at the hospital]
EU DOENTE, MINHA CABEÇA DOER [I am sick, I have a headache]
EU TER MUITA FEBRE [I have a high fever]
ONTEM DOR ESTÔMAGO EU TER [Yesterday I had a stomachache]
ENFERMAGEM HOMEM INJEÇÃO EU [A male nurse gave me a shot]
ENFERMAGEM MULHER VACINA ELE [A female nurse gave him a vaccination]

The data were collected from two signers (one male and one female), and each sign was recorded at least five times. Sentences do not have any type of cues or delays between signs, i.e., they are performed naturally.

2) Image Processing: The main goal of this module is the extraction of distinguishing features in order to recognize signs. To begin with, we are interested in describing manual features, so hands should be properly detected, segmented, and represented.



(a) HOJE (Today) (b) EU (I)

(c) ADOECER (Get sick) (d) GRIPE (Flu)

Fig. 2. Still images of the sentence HOJE EU ADOECER GRIPE which means *Today I got sick with the flu.*

An input of sample data S_i consists of a sequence of color, depth and skeleton frames captured by Kinect; all were registered. The color data consists of a collection of RGB images $S_{i_{color}} = \{f_{c_1}, f_{c_2}, \ldots, f_{c_n}\}$; the depth data of depth range images $S_{i_{depth}} = \{f_{d_1}, f_{d_2}, \ldots, f_{d_n}\}$; and the skeleton of 2D and 3D locations of 20 joints $S_{i_{skel}} = \{f_{s_1}, f_{s_2}, \ldots, f_{s_n}\}$. Both color and depth images are 640×480 in size.

Since not every frame carries relevant information for our purposes, we choose key frames from the sample data by comparing the entropy between adjacent color frames. Entropy measures the average information of an image and may be computed by using the histogram of the intensity levels of the image. Images with high entropy convey more information than images with low entropy.

Then, the extracted frames are processed as follows:

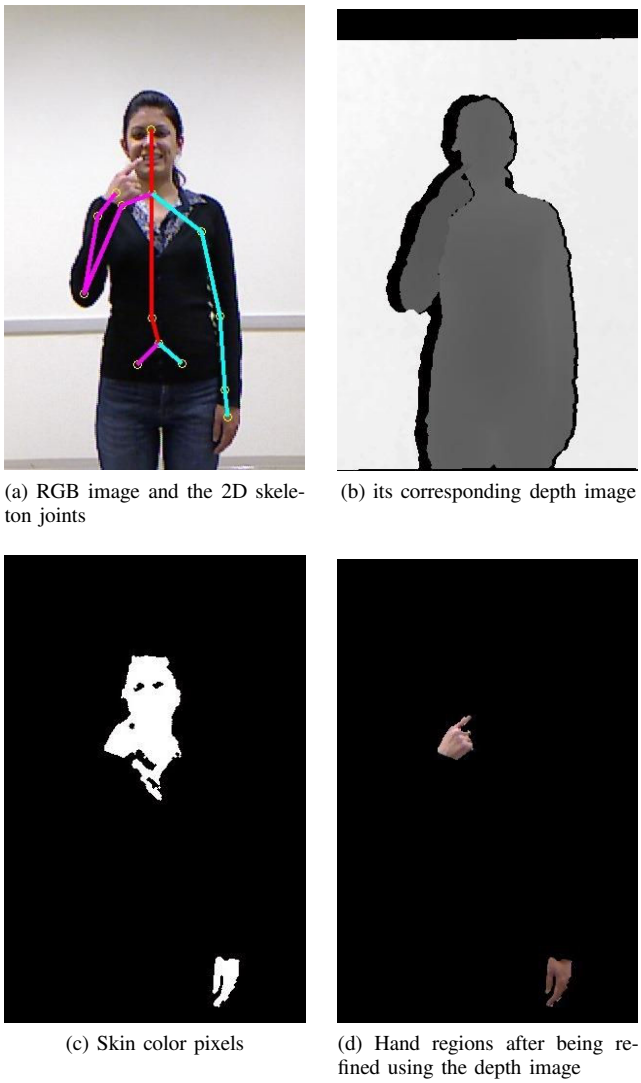
- Removing the background pixels of images based on a depth threshold (e.g., greater than 1.8 m) with the aim of reducing the next computational overhead.
- Locating the hands in the image space based on the skeleton data.
- Segmenting and extracting the hand pixels based on both

color and depth images.
- Computing the features.

In order to reduce the noise in the depth images provided by Kinect, in a pre-processing step we applied a median filter [25]. Since the noise pattern presents the characteristics of salt and pepper noise, the median filter provides an effective way to fill the missing depth values without blurring the image. Also, we remove the background, and in our approach the depth threshold is calculated in the first frame as the depth of the head joint plus 0.1 m, $th_{depth} = f_{d_1}(f_{s_1}(HeadJoint)) + 0.1$.

The hand location and tracking is performed by Kinect, however to reduce the jittering of the joints in the video, filtering such as a mean value between previous and subsequent location is applied. After detecting the hand location, the pixels of the hands are segmented by using a skin classifier in the color frame. Next, these pixels are improved through a conjunction operation with a depth mask extracted in the depth frame. To develop the skin color model, we gather skin pixels from the detected signer face [26] taken from the first two frames (we assume the skin color of the hands have a tone similar to the skin of the face). A parametric skin detector is computed as a mixture of Gaussian in the normalized RGB color space [27]. A pixel that belongs to the skin color model is classified as skin, otherwise, it is a non-skin pixel. Secondly, by using as seed the 2D skeleton hand location, a region growing algorithm operates on both color and depth frames: in the color frame, pixels that meet the skin detector's parameters are classified as skin; in the depth frame, pixels that meet a threshold (e.g., 0.05m) form a depth mask. Additionally, we also use spatial coherence to reduce the growth region. Thirdly, since the returned growth region is a binary image, the final hand region is the logical conjunction operation between the color and the depth masks. Applying morphological operations over the resulting pixels improve the shape of the hand region.

Fig. 3 illustrates the hand segmentation process. The input data consist of color, depth, and 2D skeleton information (Fig. 3a–b). The skin detector is applied over the color image which detects skin pixels (Fig. 3c). Since skin pixels that do not belong to the hand region might be considered part of it, taking into account the depth image, those pixels are removed, resulting in a refined hand region (Fig. 3d).



(a) RGB image and the 2D skeleton joints

(b) its corresponding depth image

(c) Skin color pixels

(d) Hand regions after being refined using the depth image

Fig. 3. Hand region segmentation.

After extracting salient regions, we perform feature extractions to adequately describe the signs. From the obtained 2D hand region, we compute its centroid $HC = (x_c, y_c)$ and area $HA = \#pixels$. Based on the centroid interframes, we also compute the orientation of motion of the hand $HM = arctan(y_k - y_{k-1}/x_k - x_{k-1})$, where (x_k, y_k) indicates the hand location in the kth frame. So far, the feature vector is 4-dimensional.

From the 3D skeleton data (Fig. 3a), we use as features: the location of the wrist $Wr = (x, y, z)$, elbow $El = (x, y, z)$, and shoulder $Sh = (x, y, z)$; the orientation vector shoulder-to-elbow $Sh \rightarrow El$ and elbow-to-wrist $El \rightarrow Wr$; the angle between the shoulder-center, shoulder and elbow joint $\angle ScShEl$, and the angle between the shoulder, elbow and wrist joint $\angle ShElWr$; the distance between the wrist to head and distance between both wrists (notice that we computed a 18-dimensional feature vector from the 3D joint data for each hand). So, the final feature vector has 45 dimensions.

3) Classification: Extracted manual features are modeled as probabilistic time series data in the classification module. In current experiments, we use HMMs in the manner suggested by Rabiner [8]: one model for each class and with state transition from left to right. In our approach, the number of states for each HMM is equal to the number of segments in the phonetic representation for each sign [13] plus two (we add the initial and the final state, since our data begin and end with a hand posture on the side of the body), while the number of emission symbols varies from eight to 16. Models are being trained and tested with Murphy's HMM Toolbox [28]. Our next concern is how to isolate and recognize a sign within a continuous sentence, so we plan to search for an algorithm based on movement patterns that takes into account the previously learned signs. The initial experiments produced promising results; although, in the current stage of our research, we have not yet analyzed enough cases to allow for a solid accuracy analysis and comparison with other

approaches.

4) Translation: Since each sign language has its own grammar rules, we expect that this translation module (under development) uses the recognized signs in the previous module and produces an understandable sentence. The main requirement is to maintain the meaning closest to that of the original signed sentence, so machine translation techniques will be used.

IV. Conclusion

This paper presents an ongoing research project that aims to recognize continuous signing of Brazilian sign language. Contrary to most works centered on recognizing isolated signs, our project focuses on the challenges of continuous sentences, since real-life communication is fluid, continuous and expressive.

The vocabulary basis was taken from signs and sentences used in daily conversations of deaf people in a medical care facility. In this regard, we are working with 58 signs and 15 sentences modeled with HMMs. Our expectation is that our system will be able to recognize sentences and provide a reliable translation.

Acknowledgment

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

References

[1] C. Valli, *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, 2000. ISBN 1563680971

[2] R. M. de Quadros and L. B. Karnopp, *Língua de Sinais Brasileira - Estudos Linguísticos*. Porto Alegre: Artmed, 2004. ISBN 978-85-363-0380-6

[3] H. Cooper, B. Holt, and R. Bowden, "Sign Language Recognition," in *Visual Analysis of Humans*, T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds. London: Springer, 2011, ch. 27, pp. 539–562. ISBN 978-0-85729-996-3

[4] C. Vogler and D. Metaxas, "A Framework for Recognizing the Simultaneous Aspects of American Sign Language," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 358–384, 2001.

[5] Instituto Brasileiro de Geografia e Estatística. (2010) Atlas do Censo Demográfico 2010. [Accessed on October 21, 2014]. [Online]. Available: http://censo2010.ibge.gov.br/apps/atlas/

[6] A. G. Steinberg, S. Barnett, H. E. Meador, E. A. Wiggins, and P. Zazove, "Health Care System Accessibility. Experiences and Perceptions of Deaf People." *Journal of general internal medicine*, vol. 21, no. 3, pp. 260–6, mar 2006.

[7] A. Fossati, J. Gall, H. Grabner, X. Ren, and K. Konolige, *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. Springer Publishing Company, Incorporated, 2013. ISBN 1447146395, 9781447146391

[8] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. doi: 10.1109/5.18626

[9] W. Stokoe, "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf," *Studies in linguistics: Occasional papers*, 1960.

[10] W. Sandler and D. Lillo-Martin, *Sign Language and Linguistic Universals*. Cambridge University Press, 2006. ISBN 9780521483957

[11] S. K. Liddell and R. E. Johnson, "American Sign Language: The Phonological Base," *Sign Lang. Stud.*, vol. 64, pp. 195–277, 1989.

[12] D. Perlmutter, "Sonority and Syllable Structure in American Sign Language," *Linguistic Inquiry*, vol. 23, no. 3, pp. 407–442, 1992.

[13] R. E. Johnson and S. K. Liddell, "A Segmental Framework for Representing Signs Phonetically," *Sign Lang. Stud.*, vol. 11, no. 3, pp. 408–463, 2011.

[14] T. Starner and A. Pentland, "Real-Time American Sign Language Recognition from Video using Hidden Markov Models," in *Proceedings of International Symposium on Computer Vision - ISCV*. IEEE Comput. Soc. Press, 1995. doi: 10.1109/ISCV.1995.477012. ISBN 0-8186-7190-4 pp. 265–270.

[15] B. Bauer and H. Hienz, "Relevant Features for Video-based Continuous Sign Language Recognition," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000. doi: 10.1109/AFGR.2000.840672 pp. 440–445.

[16] B. Bauer and K.-F. Kraiss, "Towards an Automatic Sign Language Recognition System Using Subunits," in *Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, ser. GW '01. London, UK, UK: Springer-Verlag, 2002. ISBN 3-540-43678-2 pp. 64–75.

[17] C. Vogler and D. Metaxas, "Handshapes and Movements: Multiple-Channel American Sign Language Recognition," in *Gesture-Based Communication in Human-Computer Interaction SE-23*, ser. Lecture Notes in Computer Science, A. Camurri and G. Volpe, Eds. Springer Berlin Heidelberg, 2004, vol. 2915, pp. 247–258. ISBN 978-3-540-21072-6

[18] H.-D. Yang, S. Sclaroff, and S.-W. Lee, "Sign Language Spotting with a Threshold Model Based on Conditional Random Fields," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1264–1277, 2009. doi: 10.1109/TPAMI.2008.172

[19] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. ISBN 1-55860-778-1 pp. 282–289.

[20] H.-D. Yang and S.-W. Lee, "Robust Sign Language Recognition with Hierarchical Conditional Random Fields," in *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ser. ICPR '10. Washington, DC, USA: IEEE Computer Society, 2010. doi: 10.1109/ICPR.2010.539. ISBN 978-0-7695-4109-9 pp. 2202–2205.

[21] ——, "Robust Sign Language Recognition by Combining Manual and Non-manual Features Based on Conditional Random Field and Support Vector Machine," *Pattern Recogn. Lett.*, vol. 34, no. 16, pp. 2051–2056, 2013. doi: 10.1016/j.patrec.2013.06.022

[22] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American Sign Language Recognition with the Kinect," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ser. ICMI '11. New York, NY, USA: ACM, 2011. doi: 10.1145/2070481.2070532. ISBN 978-1-4503-0641-6 pp. 279–286.

[23] E. B. Pizzolato, M. dos Santos Anjo, and G. C. Pedroso, "Automatic Recognition of Finger Spelling for LIBRAS Based on a Two-layer Architecture," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, ser. SAC '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1774088.1774290. ISBN 978-1-60558-639-7 pp. 969–973.

[24] C. R. de Souza and E. B. Pizzolato, "Sign Language Recognition with Support Vector Machines and Hidden Conditional Random Fields: Going from Fingerspelling to Natural Articulated Words," in *Machine Learning and Data Mining in Pattern Recognition*, ser. Lecture Notes in Computer Science, P. Perner, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, vol. 7988, pp. 84–98. ISBN 978-3-642-39711-0

[25] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice-Hall, Inc., 2008. ISBN 9780131687288

[26] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004. doi: 10.1023/B:VISI.0000013087.49260.fb

[27] M. J. Jones and J. M. Rehg, "Statistical Color Models with Application to Skin Detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002. doi: 10.1023/A:1013200319198

[28] K. Murphy. (2005) Hidden Markov Model (HMM) Toolbox for Matlab. [Accessed on May 09, 2014]. [Online]. Available: http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html

José Mario De Martino is a professor in the Department of Computer Engineering and Industrial Automation at the School of Electrical and Computer Engineering at the University of Campinas, Brazil. He received his B.S., M.S., and Ph.D. in Electrical Engineering from the University of Campinas. From 1988 to 1992 he worked as a researcher at the Computer Graphics Center in Darmstadt, Germany. Currently, his main areas of interests include: computer image synthesis, computer animation, image recognition, signing avatars, sign language recognition, virtual and augmented reality, digital gaming, and scientific visualization.

José Elías Yauri Vidalón is currently a Ph.D. student in the Department of Computer Engineering and Industrial Automation at the School of Electrical and Computer Engineering at the University of Campinas, Brazil. He received his M.S. in Electrical Engineering from the University of Campinas, in 2013, and his B.S. in Informatics Engineering from the National University San Cristobal de Huamanga, Peru, in 2001. His research interests include computer vision, machine learning, active learning and domain adaptation, and their applications to sign language recognition.