Multimedia Tools and Applications (2024) 83:19767–19785
https://doi.org/10.1007/s11042-023-16332-7

A multiple stream architecture for the recognition of signs in Brazilian sign language in the context of health

Diego R. B. da Silva[1] · Tiago Maritan U. de Araújo[2] · Thaís Gaudencio do Rêgo[2] · Manuella Aschoff Cavalcanti Brandão[2] · Luiz Marcos Garcia Gonçalves[1]

Received: 15 December 2021 / Revised: 11 July 2023 / Accepted: 17 July 2023 /
Published online: 28 July 2023

Abstract
Deaf people communicate naturally through sign languages and often face barriers to communicating with hearing people and accessing information in written languages. These communication difficulties are aggravated in the health domain, especially in a hospital emergency, when human sign language interpreters are unavailable. This paper proposes a solution for automatically recognizing signs in Brazilian Sign Language (Libras) in the health context to reduce this problem. The idea is that the system could assist in the communication between a Deaf patient and his doctor in the future. Our solution involves a multiple-stream architecture that combines convolutional and recurrent neural networks, dealing with sign languages' visual phonemes individual and specialized ways. The first stream uses the optical flow as input for capturing information about the "movement" of the sign; the second stream extracts kinematic and postural features, including "handshapes" and "facial expressions"; and the third stream process the raw RGB images to address additional attributes about the sign not captured in the previous streams. Thus, we can process more spatiotemporal features that discriminate the classes during the training stage. The computational results show that the solution can recognize signs in Libras in the health context, with an average accuracy, precision, recall, and f1-score of 99.80%, 99.81%, 99.80%, and 99.80%, respectively. Our system also performed better than other works in the literature, obtaining an average accuracy of 100% in an Argentine Sign Language (LSA) dataset, which is usually used for comparison purposes.

✉ Diego R. B. da Silva
diego.silva@lavid.ufpb.br

Tiago Maritan U. de Araújo
tiagomaritan@lavid.ufpb.br

Thaís Gaudencio do Rêgo
gaudenciothais@gmail.com

Manuella Aschoff Cavalcanti Brandão
manuella.lima@lavid.ufpb.br

Luiz Marcos Garcia Gonçalves
lmarcos@dca.ufrn.br

1 Universidade Federal do Rio Grande do Norte, Natal, Brazil

2 Universidade Federal da Paraíba, João Pessoa, Brazil

Keywords Sign language · Datasets · Deep learning · Neural networks · Libras

1 Introduction

According to a World Health Organization (WHO) report, approximately 360 million persons have disabling hearing loss, representing around 5.3% of the world's population [45]. In Brazil, according to the Demographic Census of the Brazilian Institute of Geography and Statistics (IBGE) of 2010, approximately 9.7 million people have a hearing impairment, representing around 5.1% of the population [14].

This representative portion of the population faces several barriers to accessing essential services, such as health, education, employment, and entertainment, among others [56]. One reason for this difficulty is that they communicate naturally through visual-spatial languages, called sign languages, and the spoken languages represent only a "second language". Consequently, they need help communicating with hearing people and reading and writing in spoken languages.

These communication difficulties can also be aggravated in health, especially in emergencies. An example of this situation would be a Deaf patient who is feeling sick and needs to go to the emergency but does not have a sign language interpreter who can accompany him to communicate with doctors, nurses, or other health professionals.

In Brazil, there are laws to force sign language interpreters' offered in public institutions and public health service concessionaires to address this issue. However, these laws generate a high demand for human interpreters and operating costs, especially considering Brazil has about 4,466 hospitals [18], 5,568 municipalities (or counties), and about 9.7 million people with some hearing impairment [14]. In global terms, there were around 16,500 hospitals in 2015, according to Cybermetrics Lab [35], and 360 million persons with hearing impairment, according to WHO [45], demonstrating the size and complexity involved in this theme.

One promising approach to address this problem pragmatically and reduce the communication and access to information issues of Deaf people is using machine translation (MT) tools to translate content between spoken languages and sign languages. However, this type of work is challenging because the translation is multi-modal, involving spoken and visual languages. Besides, these languages also differ in their structure over time. While spoken languages have a sequential structure (i.e., we produced phonemes sequentially over time), sign languages have a parallel structure (i.e., we can emit signs simultaneously involving various body parts). As a result, an MT system between these two types of languages must consider changes in the communication channel used (oral and visual) and their structure over time.

In the scientific literature, several works are addressing the machine translation from the text (or audio) in spoken languages into animations (or videos) in sign languages [3, 4, 28, 29, 37, 38, 43, 54]. Other works involve recognizing contents in sign languages (e.g., videos or images) into oral languages [1, 8, 15, 21, 22, 41, 48, 49, 53]. However, these solutions usually require some additional sensors or hardware (e.g., gloves, armbands, among others), making it challenging to use in a real scenario [5, 17, 26, 30–32, 63].

Besides, we did not find solutions or datasets to recognize Brazilian Sign Language (Libras), especially in the health domain. In addition, some signs in the health domain mainly depend on visual phonemes, such as facial expressions and body posture, which benefit from architectures that address these phonemes individually. For example, signs expressing pain are difficult to recognize without a component that acts directly on the signer's face.

To reduce this problem and help the inclusion and integration of Brazilian Deaf users, especially in the health domain, in this work, we propose a solution for the recognition of Libras signs in the health context. The idea is that the system can compose an automatic translation system of Libras content into Brazilian Portuguese in the future and thus assist in the communication between a Deaf patient and his doctor. Alternatively, the solution could be a first step to help a doctor who does not know Libras to understand some of the prominent patient's symptoms. Thus, it is possible to assist Deaf users in a consultation (e.g., telehealth or telemedicine), especially in emergencies or when human interpreters are unavailable.

The proposed solution combines convolutional neural networks and recurrent networks in multiple streams. The first stream uses the optical flow as input, capturing information about the "movement" of the sign, one of the prominent phonemes in sign languages. The second stream extracts kinematic and postural features (skeleton data), including handshapes and facial expressions. Finally, the third stream process the raw RGB images and can extract general and additional attributes about the sign not captured in the previous streams. This approach allows processing more spatiotemporal features that discriminate the classes during the training stage without increasing the dataset's size.

Since we did not find a dataset of signs in Brazilian Sign Language, we also created a new dataset for Libras signs recognition in the health context as an additional contribution to the work. This dataset consists of 5000 videos of 50 Libras signs extracted from day-to-day situations in the health domain and performed in an environment controlled ten times by 10 Libras interpreters. We selected the signs based on [2], which mapped Libras signs that Deaf people use more frequently in a hospital environment. Our dataset's size is similar to other datasets found in the literature for different sign languages, such as [20, 44, 46, 52]. However, these datasets were developed for general purposes rather than specific, as proposed in this work.

Thus, we can highlight the following contributions of this work:

- Proposal of a multi-stream architecture based on deep learning for recognizing Libras signs in the health domain, which can be used to help Brazilian Deaf patients communicate with health professionals;
- The three different streams designed in the proposed model address prominent visual phonemes in Libras, such as "movement", "handshapes", and "facial expressions", presenting a high descriptive potential in health-related signs. Thus, we can process more spatiotemporal features that discriminate the classes during the training stage.;
- Creation of a new database consisting of 5,000 videos of 50 Libras signs extracted from day-to-day situations in the health domain;

The rest of the paper is structured as follows. Section 2 presents some related works. In Section 3, the proposed methodology is described in detail. Experimental evaluation and results are contained in Section 4. Finally, Section 5 presents the conclusions and future works.

2 Related work

The use of artificial intelligence techniques to address problems in the health domain has been growing recently in the scientific literature. Some works propose solutions to aid in the diagnosis [24, 36, 57, 59], others to aid in prevention and monitoring [6, 7, 13, 65]. However, it is challenging to find works that address the communication difficulties faced by Deaf people to communicate and receive care in the context of health. Some works address

the communication barriers of deaf people using machine translation or computer vision techniques, but they are generally designed for other domains not related to health care.

For example, Akmeliawati et al. [1] applied Artificial Neural Network (ANN) with 7392 samples to train a system to recognize 13 signs. Using a single ANN with 45 neurons at the input layer and 14 in the output layer with two hidden layers, they achieve a median accuracy of 96.02%. However, this method requires using a glove developed by the authors to extract the ANN input features.

Binh and Ejima [8] propose a Gesture Recognition Fuzzy Neural Network (GRFNN) to adapt fuzzy control for parameter learning. This approach can eliminate the need to preselect training attributes, improving gesture recognition accuracy. The GRFNN network achieves an accuracy of 92.19% for a dataset with 36 static hand gestures in American Sign Language (ASL).

In [21], Pugeault and Bowden proposed a system for real-time recognition of ASL signs using RGB images and deep images captured from Microsoft Kinect. They used the Random Forest (RF) technique in a dataset containing 131548 samples, achieving a median accuracy of 75%.

Pigou et al. [48] applied Convolutional Neural Network (CNN) in a dataset with 6600 samples of 20 distinct gestures, using 4600 samples for the training set and 2000 samples for the validation set, achieving a median accuracy of 91.7%. One of the disadvantages of this approach is the need to use in-depth image data extracted from Kinect, which is challenging to use in a real scenario.

In [41], Masood et al. proposed an architecture combining Convolutional Neural Networks 2D (CNN 2D) and Long Short-Term Memory (LSTM) to classify spatiotemporal features. The approach described initially in [23] aims to recognize sign language gestures (LSA). The dataset, introduced in [51], consists of 64 signs recorded from 10 LSA interpreters, resulting in 3200 videos. For each execution, they selected eight videos for training and two videos for the test. The approach uses a pre-trained network to perform transfer learning from Inception-V3 Network [58]. They proposed two solutions, one with more layers and a higher number of trainable parameters in the LSTM stage. The more complex model showed greater accuracy, with about 95.6% accuracy for the test set against 91.6% of the model with fewer parameters. Despite the high precision, a limitation of their experiments is that the test set was always composed of the same human interpreters as the training set. Thus, the performed tests do not include people who did not belong to the training and validation set, making it difficult to assess their models' generalization ability.

Simonyan and Zisserman [55] proposed a two-stream ConvNets architecture incorporating spatial and temporal networks through training deep CNNs using sampled frames and stacked optical flows. Their approach had competitive results and was better evaluated for the video classification problem in [66]. Wu et al. [64] work showed competitive results by expanding this architecture by adding LSTM networks to build a hybrid learning framework that can model essential aspects of the video data.

Wadhawan and Kumar [61] used a CNN architecture to recognize Indian Sign Language (ISL) and achieved good results. Parelli et al. [47] used 3D hand skeletal information for sign language (SL) recognition from RGB videos. They achieved good accuracy on a corpus of isolated signs of Greek SL and a dataset of continuous finger-spelling in ASL. Lu et al. [39] used Capsule Network (CapsNet) and Selective Kernel Network (SKNet) with attention mechanism to ISLR. This approach achieved a recognition accuracy of 98.88% in the experiments using their dataset.

Sharma and Kumar [53] proposed a model based on 3DCNN for isolated ASL recognition. The CNN is trained to classify 100 words on the Boston ASL (Lexicon Video Dataset) LVD

dataset with more than 3300 English words signed by six different signers. The proposed model is simple yet fast and accurate that outperforms state-of-the-art results on the LVD dataset in terms of precision (3.7%), recall (4.3%), and f-measure (3.9%). Dignan et al. [22] proposed a hybrid approach that takes advantage of low-cost sensory hardware and a deep learning-based approach to a sign-recognition system. This approach achieved an accuracy of 80% on the evaluation sets, but its applicability depends on specific hardware.

Rastgoo et al. [49] used hand keypoint coordinates as basic features to sign language recognition. Another work proposed by [50] used singular value decomposition (SVD) and long short-term memory for real-time isolated hand sign language recognition (IHSLR) from RGB video. The main contribution is that SVD is applied to the hand keypoint coordinates for more discriminative features.

A Graph Convolutional Networks (GCN) model is proposed by Vazquez-Enríquez et al. [60]. Their solution was designed to sign language recognition using a skeleton graph that includes body and finger joints. The results show that GCNs alone stand out as a viable technique for ISLR, especially when compared to 3D-CNN architecture since they can capture the internal relationship among semantically connected distant nodes in sign language dynamics. Boháček and Hrúz [9] presented a word-level sign language recognition system based on the Transformer model. They used body keypoints, and the main contribution is a pose normalization scheme that takes the hand and body poses in a separate and independent local coordinate system.

In [40], the authors use a hybrid architecture combining a 3D Convolutional Network and Bidirectional Convolutional Network LSTM (ConvLSTM). They used the model proposed by [67] pre-trained from the ISOGD dataset [62] and trained it to classify Libras signs. Their solution uses RGB-D images, i.e., RGB images with an additional depth channel. They performed computational tests with a non-public dataset containing 510 signs captured from 7 different interpreters with six repetitions for each sign. For this data set, the proposed model had an accuracy of 79.80%.

Analyzing these works, we can observe these works focus on isolated sign recognition for the general contents without addressing the specificities of the health domain, involve other sign languages, or are dependent on specific hardware, which makes its use in real scenarios more difficult [5, 17, 22, 26, 30–32, 63]. The only work [40] found for Libras was also developed for the general signs and was developed or tested in the health domain. In addition, their model was not designed to involve different streams to address the prominent Libras phonemes.

Thus, this paper aims to investigate and propose a solution to deal with Libras isolated sign language recognition (ISLR) in the health context. Our solution does not require additional sensors or hardware and is based on images. We propose an architecture based on CNN and LSTM, containing three streams fed with RGB images, optical flows, and keypoints calculated from these images. Thus, we can address a sign's different phonemes (features), such as movements, handshapes, and facial expressions. Besides, we present a novel sign language dataset video, which may help develop and research other solutions by the community.

3 Methodology

In sign languages, the signs have three main parts: (i) Manual features involving gestures made with the hands and their location, (ii) Non-manual features such as facial expressions or body posture, which can form part of a sign or modify its meaning, and (iii) Fingerspelling,

where words are signed letter by letter in a manual alphabet [19]. However, this is an oversimplification since sign languages are as complex and expressive as spoken languages. Besides, each sign language has thousands of signs, differing from the next by minor changes in handshape, motion, position, non-manual features, or context [19].

Thus, to recognize signs in Libras, we proposed a multiple-stream architecture that addresses features from different sign parts. It uses the (i) RGB images, (ii) optical flows, and (iii) keypoints data as input, feeding the model in parallel with spatial components extracted from the raw images, the movement information encoded through the optical flow, and the body keypoints providing kinematic and postural information. Thus, the solution can address several distinct components (or phonemes) that discriminate a sign. Sections 3.1 to 3.4 presents the main details of the proposed system.

3.1 Dataset

Given the absence of datasets for Libras available in the literature, initially, we created a novel dataset of 50 Libras signs in the health domain. We selected these signs based on actual occurrences in the medical environment extracted from [2], which mapped Libras signs and their applicability for use in the clinical anamnesis of the nursing consultation for Deaf people.

The dataset contains 100 samples (videos) for each of the 50 signs, resulting in 5000 videos. We used 10 Libras specialists, which recorded all the signs ten times in different sessions. Each interpreter was chosen to represent the Brazilian ethical variability, presenting different skin shades, body types, and age ranges. We recorded the samples using a smartphone HD camera (720p) at 30 fps, with similar environment settings. Table 1 summarizes the recording environment settings.

According to Table 1, we used a white wall as a background, and the interpreter was located 0.4m from the wall. We placed the tripod 2.5m away from the interpreter at a height of 1.5m. We also used marks on the floor to facilitate the tripod's positioning, and interpreters maintained a similar and homogeneous distance and angle between the camera and the interpreter (see Table 2).

3.2 Pre-processing

In the preprocessing step, we convert each video in the dataset to 40 frames through subsampling. This step is necessary because convolutional networks typically require a known input size, at least the type used for the current study. However, some videos had less than 40 frames. In this case, we repeat the last one until we get 40 frames.

Table 1 Capture setup

Setup	value
Wall color	white
Distance from wall	0.4 m
Tripod distance from interpreter	2.5m
Tripod height	1.5m
Camera quality	HD (720p)
Interpreter shirt color	black
Illumination type	diffuse

Table 2 The characteristics of the data

	Avg	Max	Min
Duration	2.5s	1.2	3.2
Number of frames	125	40	260
FPS	29.5	29.1	29.9

Then, we use these frames to generate two more specific data sets. First, a dataset containing an optical flow calculated through OpenCV library [10]. The second is a dataset containing keypoints generated with OpenPose [11]. The new data sets aim to improve motion detection, postural body information, and the relationship between joints and facial expressions. As a result, it increases the number of features we can use to distinguish the signs without increasing the number of videos for each class in the raw data set.

We also cropped the region of interest for each frame, i.e., a rectangle around the signer, to decrease the data dimensionality. We also standardized its dimensions to 224x224 pixels. Therefore, the input shape is $40 \times 224 \times 224 \times 3$, to the RGB stream. We also perform a data augmentation step in the RGB images, introducing artificial noises to simulate light variation, loss of focus, spatial distortions, and other distortions in free cell phone recording.

To generate the optical flows, we have used the library OpenCV [10]. We down-sampled the original videos to 40 frames to perform this task. After that, we converted them to grayscale and applied the TV-L1 algorithm to calculate the optical flows. The resulting pixel values were then truncated to the range $[-20, 20]$ and scaled between -1 and 1. We used only two channels for this stream. Therefore, the input dimension with the optical flow is $40 \times 224 \times 224 \times 2$. Finally, we apply the one-hot coding to categorize variables (sign labels).

We generate the keypoints using the COCO model keypoints format, where we extract 25 keypoints from the body, 21 from each hand, and 70 from the face. Although the recording setup, only the upper body keypoints can be extracted, totaling 14 keypoints from the body (See Fig. 1). We stored the keypoints generated by the OpenPose in .json files, mapping each frame of the original video sequence to a single file. We run the Openpose without any additional configuration parameters. After generating the keypoint dataset, it was necessary to standardize the attributes (center of meaning and scaling to unit variance). This step is essential to help to reduce the network convergence time.

We also apply data augmentation to the keypoints. More specifically, we applied linear transformations that, given a probability p, perform the rotation, translation, and scale of each (x, y) coordinate of the keypoints. This type of transformation is helpful for the model to be able to recognize videos that are not in the same recording frame as the dataset (see Table 1).

3.3 The architecture

The proposed architecture is based on the structure of the signs in Libras, which consists of five phonemes: (i) hand shape, (ii) location, (iii) movement, (iv) orientation, and (v) non-manual expressions. This characteristic led to the hypothesis that a deep learning model exploring different phonemes in parallel would perform better than a more generalist architecture for classifying and recognizing video actions.

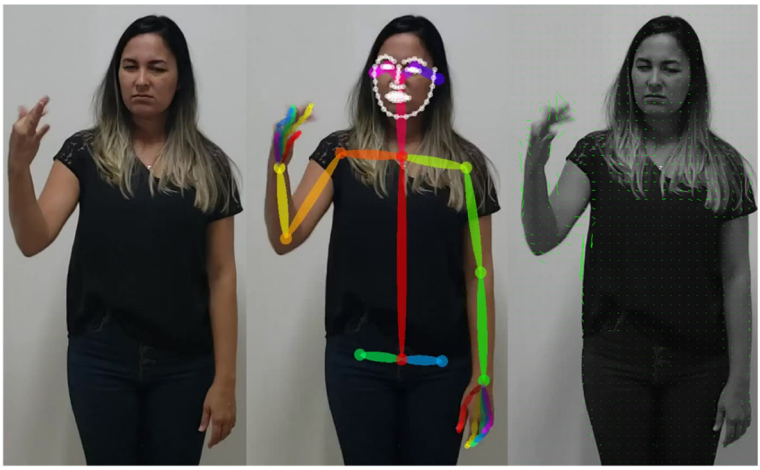

Fig. 1 Visualization of the three different types of data that feed the model. RGB image (left), keypoints poses (middle), and optical flow (right)

Then, we propose a three-stream architecture[1] based on (i) RGB images, (ii) optical flows, and (iii) keypoints. Thus, the model is powered in parallel with the spatial components of the image extracted from the RGB images, the encoded movement speed information through the optical flow, and body posture information computed through the OpenPose library. Thus, the solution can deal with multiple distinct components (or phonemes) that form a sign.

The architecture is based on the I3D network that processes the RGB and optical flow streams and an LSTM network that processes the keypoints stream. Figure 2 presents a schematic view of the proposed architecture. Each part of this architecture is detailed in Sections 3.4 to 3.6.

3.4 I3D

The I3D model proposed in [12] is generally applied to video classification problems. It uses 3D convolutions that inflate the 2D convolutional filters into 3D using the weights and biases of an Inception v1 model pre-trained on ImageNet, as shown in Fig. 3. It took the weights and biases of each 2D layer and stacked them up to form the third dimension to produce the deep spatiotemporal descriptors. Therefore, the spatial and temporal CNNs process a stack of consecutive RGB frames and optical flow images

Its architecture was trained on the Kinetics dataset, a massive compilation of YouTube URLs for over 400 human actions and over 400 video samples per action. It is a complex and costly architecture to be trained from scratch, requiring considerable computational infrastructure. However, it is an attractive architecture for transfer learning using the publicly available I3D model due to its high accuracy and the extensive data set that the authors trained it.

Our solution freezes the weights of the initial and intermediate layers and trains only the final dense layers responsible for producing the probability distribution over the classes. The input size is $40 \times 224 \times 224 \times 3$, where the first value represents the number of sampled

[1] The multi-stream architectures present multiple channels with different data and processing that are merged using concatenative, additive, subtractive multiplicative, statistical, among others

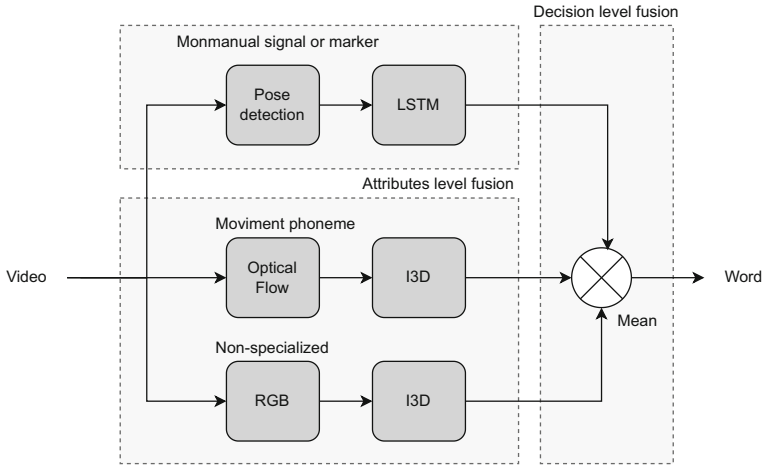

Fig. 2 Schematic view of the proposed solution with I3D and LSTM

frames from the input video, and the others represent the frame's width, height, and number of channels, respectively.

3.5 LSTM

LSTM [27] networks (Fig. 4) use specialized units called memory cells that can selectively retain and forget information over time, making them particularly effective at processing long sequences. This ability is achieved by adding gates that control the information flow between cells, allowing the network to learn when to retain or forget certain information.

Given $x_1, x_2, \ldots, x_m$, h_{t-1}, and c_{t-1}, where m is the length of the sequence, $x_i \in R^d$ is the feature vector obtained by concatenating features, and h_{t-1} and c_{t-1} are the previous hidden state and cell state of the LSTM cell (with h_o and c_o initialized as zero vectors), the new hidden state and new cell state are computed using the following equations:

$$\hat{c}_t = tanh(W_c[h_{t_i}, x_t] + b_c) \tag{1}$$

$$i_t = \sigma(W_i[h_{t_i}, x_t] + b_i) \tag{2}$$

$$f_t = \sigma(W_f[h_{t_i}, x_t] + b_f) \tag{3}$$

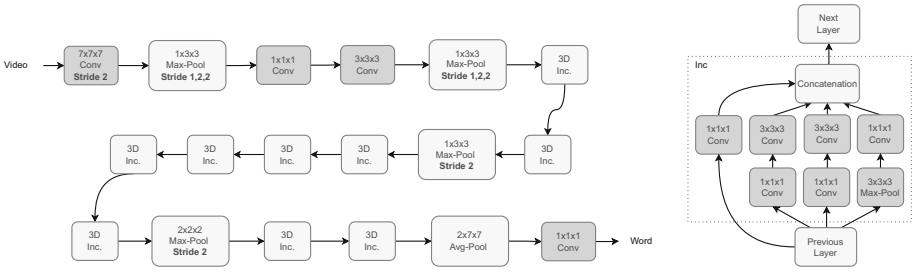

Fig. 3 The architecture of Inflated Inception-V1 model

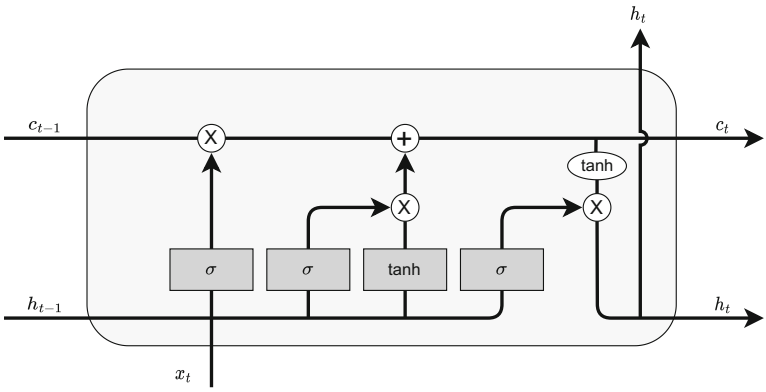

Fig. 4 The architecture of the LSTM model

$$o_t = \sigma(W_o[h_{t_i}, x_t] + b_o) \tag{4}$$

$$c_t = i_t * \hat{c}_t + f_t * c_{t-1} \tag{5}$$

$$h_t = tanh(c_t) * o_t \tag{6}$$

W_f, W_i, W_o, $W_c \in R^{(N+d)} \times N$ are weight matrices and b_f, b_i, b_o, $b_c \in R^N$ are bias vectors. Weight matrices and bias vectors are initialized randomly and learned through an optimization process during the training phase, where N is the size of the LSTM layer, and d is the dimension of the input feature vector.

Our solution uses deep LSTM RNNs built by stacking 3 LSTM layers. Each layer uses 64 hidden units and a dropout rate of 0.5. We feed this network with 137 keypoints coordinates (x, y), and this input is propagated through LSTM layers. The encodings produced by the last LTSM layer are fed to a dense layer with 64 units. At the end of all layers, we have a fully connected layer using the softmax activation function, which produces a distribution over the sign classes according to the following equation:

$$s(z) = \frac{e^{z_i}}{\sum_{j=0}^{k} e^{z_j}} \tag{7}$$

The exponential function is executed on each element z_i of the input vector z, and the output values are normalized by dividing by the sum of the exponentials. The normalization is applied such that the output vector p_i sums up to 1.

3.6 Decision fusion

Our solution uses decision-level fusion methods to merge the predictions produced by the classifiers based on the I3D and LSTM networks. In other words, the decision is the average of the final layers containing the probability distributions of each model. We choose this type of fusion over attribute-level fusion methods due to its higher resistance to overfitting and shorter training time.

Algorithm 1 describes the proposed solution.

Algorithm 1 Multiple Stream Recognition System

procedure LIBRASMULTISTREAM(V)
  $V = sample(V, n)$ ▷ Frame sampling
  $K = \{k_1, k_2, ..., k_n\}$ ▷ Keypoints
  $O = \{o_1, o_2, ..., o_n\}$ ▷ Optical flows
  for each frame i in V **do**
    $f_i = resize(f_i)$ ▷ Resize to 224x224
    $k_i = PoseEstimation(f_i)$ ▷ Compute keypoints
    $o_i = OpticalFlow(f_i)$ ▷ Compute optical flows
    $f_i = crop(f_i, k_i)$ ▷ crop signer
    $f_i = normalize(f_i)$ ▷ normalize image
  end for
  $w_i = ThreeStream(V, O, K)$ ▷ run model
  return w_i ▷ return predicted word
end procedure

3.7 Training

For training, we used the Keras library [16] with the hyperparameters presented in Table 3. We trained the model with a batch size of 4 and used 200 iterations. We use 70% (3500 videos) of the dataset for training, 20% (1000 videos) for validation, and the remaining 10% (500 videos) for testing, as summarized in Table 4.

Another important detail is that the interpreter chosen for testing did not participate in the training. Thus, we used nine interpreters for training and validation and one for testing, which allows a more accurate measurement of the model's generalizability since it reduces the chance of overfitting.

We perform all computational experiments on a computer with an Intel i3-3240 (3.40 GHz) processor with 32 GB of RAM and a GPU Nvidia Quadro P6000. All experiments were performed on Ubuntu 18.04 LTS, Python 3.8, CUDA 11.7 and cuDNN 8.5.

4 Results and discussions

In this section, the results of the presented solution are presented and discussed in detail. We evaluated the model using videos of a human interpreter whom the model had not seen during training. Thus, given that there are n signers, we trained the model with $(n-1)$ signers and evaluated over the nth signer. The metrics used are typical of a multiclass classification problem, such as accuracy, precision, recall, and f1-score. In addition, the LSA64 dataset

Table 3 Hyperparameters values used in the training

Hyperparameters	Value
Learning rate	10e-4
Weight decay	10e-6
Optimizer	Adam
Frames per video	40
Image dimension	(224,224,3)
Batch size	4
Dropout	0.5

Table 4 Dataset division between training, validation, and test

	Amount of samples
Training set	3500
Validation set	1000
Testing set	500

was also used to analyze the behavior of the proposed model in other sign languages and to compare it with other related works.

The values of the hyperparameters used, the division of the dataset between training, validation, and test sets, and information related to the computational resources, technologies, and libraries used are presented in Section 3.7.

As performed in [57], Table 5 presents metrics like training loss, testing loss, training accuracy, and testing accuracy obtained during the training process (see Table 5). Table 6 shows the average accuracy values obtained during this experiment.

According to Table 6, we can note that the proposed solution obtained a good overall result, with an average accuracy of 99.80% in top-1 and 100.0% in top-5. We can also observe that the isolated streams already have high average accuracy, consistently above 95%, and that adding a more specialized stream focusing on a particular visual phoneme increases the accuracy.

In addition, the body posture and spatial relations between joints have a high power of discrimination, at least within the universe of signs in the data set. This stream addressed the main visual phoneme responsible for the recognition errors, the phoneme of non-manual expressions, which involves facial and body expressions.

We have also calculated the proposed model's precision, recall, and f1-score on the proposed dataset. The definition of these metrics is presented in (8) to (11).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{11}$$

Where:

- TN (True Negative): These are instances where the model correctly identifies that a given Libras' sign is not present in a video;
- TP (True Positive): These are instances where the model correctly identifies the presence of a specific Libras' sign in a video;

Table 5 The training, validation accuracy, and loss

	Loss	Acc	Val Loss	Val Acc
image	0.016	99.58	0.0001	97.42
flow	0.013	99.66	0.0006	98.99
pose	0.007	99.82	0.005	99.58

Table 6 Results of proposed methodology over the proposed dataset and a benchmark dataset

	Dataset results	
	LIBRAS50	LSA64
Image	95.32	97.79
Flow	95.31	97.83
Body skeleton	99.20	99.68
Hand skeleton	98.00	99.06
Image + Flow	96.12	98.44
Image + Flow + Body skeleton	98.80	100.0
Image + Flow + Hand skeleton	99.20	100.0
Image + Flow + Body, Hand skeleton	99.80	100.0

– FP (False Positive): These are instances where the model incorrectly identifies the presence of a specific Libras' sign in a video;
– FN (False Negative): These are instances where the model fails to identify the presence of a specific Libras' sign in a video.

Table 7 presents the results of the model's accuracy, precision, recall, and f1-score. According to Table 7, given that it is not possible to optimize the recall and accuracy metrics at the same time, and as the f1-score metric is the harmonic mean between these two metrics, we can conclude that the number of false positives and false negatives is reduced when using a multi-stream model.

In addition, we can note that the proposed solution recognizes the signs performed by a human user who did not participate in the training with a performance of almost 100%. The recognition failures were concentrated between the sign "pain" and other signs expressing pain or ache, such as "toothache", "headache" and "stomach ache". These misclassifications might be caused by many factors, such as the model requiring more data for training and minor differences in execution. Another possible contributing factor is that some human interpreters use different hands to perform the same sign or have different proficiency levels in sign language. One of our proposals for future work is a deeper investigation into the factors that may have caused these errors.

We have also compared our solution with some related works in the literature. These works generally use the data set presented in [52], which consists of 64 signs from the Argentine Sign Language (LSA) for comparison purposes. Thus, we replicate the methodology proposed using this data set. Table 8 presents the result of this comparison. The results show that despite

Table 7 The overall classification results for the classification

Attributes	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Image	95.32	94.00	95.00	94.00
Flow	95.31	96.00	95.00	95.00
Body skeleton	99.20	99.27	92.20	99.20
Hand skeleton	98.00	98.53	98.00	97.77
Image + Flow	96.12	95.00	96.00	95.00
Image + Flow + Body skeleton	98.80	99.00	98.80	98.78
Image + Flow + Body, Hand skeleton	99.80	99.81	99.80	99.80

Table 8 Comparison with Related Works

Author, Year	Classifier used	Multiple stream	Sign Language	Performance
[41]	CNN-2D + LSTM	–	LSA	95.60%
[34]	CNN-2D	RGB+Pose	LSA	98.09%
[33]	CNN-2D + LSTM	RGB+Flow+Pose	LSA	99.84%
[40]	CNN-3D + ConvLSTM	RGB+Depth	Libras	79.80%
[25]	CNN-2D	RGB	ASL	95,83%
[42]	CNN-2D	RGB	ISL	93.89%
Our solution	CNN-3D + LSTM	RGB+Flow+Pose	Libras	**99.80%**
Our solution	CNN-3D + LSTM	RGB+Flow+Pose	LSA	**100.0%**

the dataset LSA64 having more classes, the model still presents better results compared to the dataset presented in this work. This highlights a difficulty in recognizing signs of the health domain.

According to Table 8, the results show that the proposed methodology achieved an accuracy of 100% using this data set. Also, our solution had an accuracy of 20 percentual points higher than the only work developed for the Libras language.

Besides, we believe that the performance of the solution (an average accuracy of 99.80%) also indicates that the solution can assist in the communication process between doctors and Deaf patients in the process of identifying symptoms. Thus, we could have a complementary and practical alternative, especially if we consider the difficulty of providing human Libras interpreters available in all Brazilian hospitals and doctors' offices on all telehealth platforms. However, we think that it is still necessary to conduct tests with doctors and Deaf people in a real scenario for further validation of the solution. This investigation is one of our proposals for future work.

We also perform a computational test to evaluate our model's average recognition time. Table 9 presents the average results. According to Table 9, despite this high accuracy, the proposed architecture presents a high computational cost due to its complex pipeline. Each stream (or model) needs to perform specific transformations to adapt the data (40 frames per video) to the methods or architectures employed in each stream.

Individually, the stream based on the I3D network is the most performative. It requires only the extraction, subsampling, and calculation of optical flows, which are relatively fast operations, resulting in a model system with less than 1 second response time. However, with the introduction of a stream based on the detection of poses (keypoints stream), the computational cost has increased, going on average from 1 to 4.6 seconds to perform the inference, since OpenPose can process on average 17 frames per second, being the component with the highest cost of processing the model.

Table 9 Average times of each stage of the model's pipeline (40 frames per video)

Attribute	Pre-processing (s)	Inference time (s)	Total time (s)	Accuracy (%)
RGB+Flow	0.230	0.226	0.500	96.34
Pose	3.1	0.044	3.2	98.00
RGB+Flow+Pose	3.3	0.320	4.6	99.80

However, there is a linear relationship between the number of GPUs and the processing time of OpenPose. Thus, this model's deployment could be feasible in an infrastructure based on GPUs (e.g., a multi-GPU environment or a cluster of GPUs), reducing response time while maintaining high accuracy. Nonetheless, testing in a multi-GPU environment was not possible because we did not have such a structure for evaluation. However, it is also one of our proposals for future work.

5 Conclusion and future work

In this paper, we proposed a multiple stream model to recognize Libras' signs in the health context. Our solution does not need additional hardware or capture sensors (e.g., gloves, armbands), entirely based on images. Recognizing signs is a complex classification task since the signs have different phonemes, such as handshape, movement, orientation, and non-manual expressions. Then, a single phoneme variation can generate a different sign.

To the best of our knowledge, our work was the first to address sign recognition in Brazilian Sign Language applied in the health context. Furthermore, one of the main differentials of our architecture is that the three different streams of the solution were designed to address the prominent visual phonemes in Libras, such as "movement", "handshapes", and "facial expressions", presenting a high descriptive potential in health-related signs. Thus, we can process more spatiotemporal features that discriminate the classes during the training stage.

In addition, another significant contribution of the work was the creation of a new database consisting of 5,000 videos of 50 Libras signs performed by highly qualified linguists and sign language specialists, and extracted from day-to-day situations in the health domain. Datasets in Libras are scarce and often need more quality and standardization, especially in the health domain. To the best of our knowledge, there is no dataset with this characteristic and data volume available in the literature, which may assist the development of solutions in helping in the communication of Deaf people in the health context and further research on the subject.

To validate the solution, initially, we have performed some computational tests using the Libras database created in the health domain. The results show that the test set's best accuracy was around 99.80%, considering a scenario where we separate the interpreter in the test set, not using him in the training set. Additionally, the results suggest the effectiveness in addressing different Libras phonemes using specialized flows.

Thus, we believe that the performance of the solution (an average accuracy of 99.80%) also indicates that the solution can assist the communication between Deaf patients and health professionals, especially in identifying symptoms.

We also performed a comparative analysis with other works in the scientific literature using the Argentine Sign Language (LSA) dataset presented in [52], which is usually used for comparison purposes. The results show that our system also performed better than other works in the literature, obtaining an average accuracy of 100% in the LSA dataset.

One of the limitations of this study is the size of the database. Although the signs were defined through a study that mapped the most frequent vocabulary in a clinical anamnesis context, a large universe of signs still needs to be covered to increase coverage and enable a better evaluation of the solution and its applicability in an end-user solution.

The recognition failures were concentrated between the sign "pain" and other signs expressing pain or ache. As mentioned in Section 5, these misclassifications might be caused by many factors, and performing a better investigation of these factors is one of our proposals for future work.

Another proposal for future work is to include new streams to address other sign elements (or phonemes). In addition, we also plan to extend the comparative analysis with other works in the literature.

We also plan to perform a more robust analysis of the dataset quality, quantifying this through a metric that captures the spatial-temporal variability. Another proposal is to increase the size of the database, including other signs in the health context.

Finally, another proposal for future work is to perform tests with Deaf users and health professionals to assess the solution's application in a real-world usage scenario.

Acknowledgements This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We gratefully acknowledge NVIDIA Corporation's support with the donation of a Quadro P6000 used for this research.

Author contributions Diego R. B. da Silva and Tiago Maritan U. de Araújo conceived and designed the approach. Thaís Gaudencio do Rêgo contributed to the experimental design of the study. Manuella Aschoff Cavalcanti Brandão helped with data collection. The first draft of the manuscript was written by Diego R. B. da Silva. Tiago Maritan U. de Araújo and Luiz Marcos G. Gonçalves thoroughly corrected the manuscript. All authors read and approved the final manuscript.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Availability of data and material Data generated or used during the study is available from the corresponding author by request.

Declarations

Ethics approval Not applicable.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Akmeliawati R, Ooi MPL, Kuang YC (2007) Real-time Malaysian sign language translation using colour segmentation and neural network. In: 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007. IEEE. https://doi.org/10.1109/imtc.2007.379311
2. Aragão JDS, Francisco ISXD, Coura AS, Sousa FSD, Batista JDL, Magalhões IMDO (2007) A content validity study of signs, symptoms and diseases/health problems expressed in LIBRAS. Revista Latino-Americana de Enfermagem 23:1014–1023. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-11692015000601014&nrm=iso
3. Araujo T, Ferreira F, Silva D, Oliveira L, Falcão E, Martins V, Portela I, Nóbrega Y, Lima H, Souza Filho G, Tavares T, Duarte A (2014) An approach to generate and embed sign language video tracks into multimedia contents. Inf Sci 281:762. https://doi.org/10.1016/j.ins.2014.04.008
4. de Araújo TMU, Ferreira FLS, dos S. Silva DAN, Lemos FH, Neto GPB, Omaia D, de Souza Filho GL, Tavares TA (2012) Automatic generation of Brazilian sign language windows for digital TV systems. J Braz Comput Soc 19:107–125
5. Bessa Carneiro S, De M. Santos EDF, De A. Barbosa TM, Ferreira JO, Soares Alcalá SG, Da Rocha AF (2016) Static gestures recognition for Brazilian sign language with kinect sensor. In: 2016 IEEE Sensors. pp 1–3
6. Bhatti UA, Huang M, Wang H, Zhang Y, Mehmood A, Di W (2018) Recommendation system for immunization coverage and monitoring. Hum Vaccin Immunother 14(1):165–171. https://doi.org/10.1080/21645515.2017.1379639. (**PMID: 29068748**)
7. Bhatti UA, Huang M, Wu D, Zhang Y, Mehmood A, Han H (2019) Recommendation system using feature extraction and pattern recognition in clinical care systems. Enterprise Information Systems 13(3):329–351. https://doi.org/10.1080/17517575.2018.1557256

8. Binh ND, Ejima T (2005) Real-time Malaysian sign language translation using colour segmentation and neural network. In: Proceeding of ICGST International Conference Graphics, Vision and Image Processing. pp 1–6
9. Boháček M, Hrúz M (2022) Sign pose-based transformer for word-level sign language recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp 182–191
10. Bradski G (2000) The OpenCV Library. Dr. Dobb's Journal of Software Tools
11. Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y (2018) Openpose: realtime multi-person 2D pose estimation using part affinity fields. CoRR abs/1812.08008. http://arxiv.org/abs/1812.08008
12. Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset
13. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F (2021) AI applications to medical images: from machine learning to deep learning. Physica Med 83:9–24
14. Cavararo R (2010) Características gerais da população, religião e pessoas com defciência. Instituto Brasileiro de Geografa e Estatística (IBGE). https://biblioteca.ibge.gov.br/visualizacao/periodicos/94/cd_2010_religiao_deficiencia.pdf
15. Cheok MJ, Omar Z, Jaward MH (2017) A review of hand gesture and sign language recognition techniques. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-017-0705-5
16. Chollet F et al (2015) Keras. https://keras.io
17. Chuan C, Regina E, Guardino C (2014) American sign language recognition using leap motion sensor. In: 2014 13th International Conference on Machine Learning and Applications. pp 541–544
18. CNsaúde: Cenário dos Hospitais no Brasil. S.N. (2022). http://cnsaude.org.br/wp-content/uploads/2022/07/CNSAUDE-FBH-CENARIOS-2022.pdf
19. Cooper H, Holt B, Bowden R (2011) Sign language recognition. In: Visual Analysis of Humans. Springer London, pp 539–562. https://doi.org/10.1007/978-0-85729-997-0-27
20. Cooper H, Ong E, Pugeault N, Bowden R (2017) Sign language recognition using sub-units. pp 89–118. https://doi.org/10.1007/978-3-319-57021-1_3
21. Cooper H, Pugeault N, Bowden R (2011). Reading the signs: a video based sign dictionary. https://doi.org/10.1109/iccvw.2011.6130349
22. Dignan C, Perez E, Ahmad I, Huber M, Clark A (2022) An AI-based approach for improved sign language recognition using multiple videos. Multimed Tools Appl 81(24):34525–34546. https://doi.org/10.1007/s11042-021-11830-y
23. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2014) Long-term recurrent convolutional networks for visual recognition and description. Preprint at http://arxiv.org/abs/1411.4389
24. Elemento O, Leslie C, Lundin J, Tourassi G (2021) Artificial intelligence in cancer research, diagnosis and therapy. Nat Rev Cancer 21(12):747–752
25. Fakhfakh S, Jemaa YB (2022) Deep learning shape trajectories for isolated word sign language recognition. Int Arab J Inf Technol 19(4):660–666
26. Galicia R, Carranza O, Jiménez ED, Rivera GE (2015) Mexican sign language recognition using movement sensor. In: 2015 IEEE 24th International Symposium on Industrial Electronics (ISIE). pp 573–578
27. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
28. Huenerfauth M (2004). A multi-path architecture for machine translation of English text into American sign language animation. https://doi.org/10.3115/1614038.1614043
29. Huenerfauth M (2008) Generating American sign language animation: overcoming misconceptions and technical challenges. Univ Access Inf Soc 6:419–434. https://doi.org/10.1007/s10209-007-0095-7
30. Jani AB, Kotak NA, Roy AK (2018) Sensor based hand gesture recognition system for English alphabets used in sign language of deaf-mute people. In: 2018 IEEE Sensors. pp 1–4
31. Kau L, Su W, Yu P, Wei S (2015) A real-time portable sign language translation system. In: 2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS). pp 1–4
32. Kaya F, Tuncer AF, Yildiz ŞK (2018) Detection of the Turkish sign language alphabet with strain sensor based data glove. In: 2018 26th Signal Processing and Communications Applications Conference (SIU). pp 1–4
33. Konstantinidis D, Dimitropoulos K, Daras P (2018) A deep learning approach for analyzing video and skeletal features in sign language recognition. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST). pp 1–6. https://doi.org/10.1109/IST.2018.8577085
34. Konstantinidis D, Dimitropoulos K, Daras P (2018) Sign language recognition based on hand and body skeletal data. In: 2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON). pp 1–4. https://doi.org/10.1109/3DTV.2018.8478467

35. Lab C (2015) Ranking web of world hospitals. https://hospitals.webometrics.info/
36. Li T, Li J, Liu J, Huang M, Chen YW, Bhatti UA (2022) Robust watermarking algorithm for medical images based on log-polar transform. EURASIP J Wirel Commun Netw 2022(1):24. https://doi.org/10.1186/s13638-022-02106-6
37. López-Ludeña V, Morcillo C, López JC, Barra-Chicote R, Cordoba R, Hernandez R (2014) Translating bus information into sign language for deaf people. Eng Appl Artif Intell 32. https://doi.org/10.1016/j.engappai.2014.02.006
38. López-Ludeña V, Morcillo C, López JC, Ferreiro E, Ferreiros J, Hernandez R (2014) Methodology for developing an advanced communications system for the deaf in a new domain. Knowl-Based Syst 56:240–252. https://doi.org/10.1016/j.knosys.2013.11.017
39. Lu J, Nguyen M, Yan WQ (2021) Sign language recognition from digital videos using deep learning methods. In: Nguyen M, Yan WQ, Ho H (eds) Geometry and Vision. Springer International Publishing, Cham, pp 108–118
40. Machado MC (2018) Classificação automática de sinais visuais da língua brasileira de sinais representados por caracterização espaço-temporal. Master's thesis. https://tede.ufam.edu.br/handle/tede/6645. Instituto de Computação
41. Masood S, Srivastava A, Thuwal H, Ahmad M (2018) Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. pp 623–632. https://doi.org/10.1007/978-981-10-7566-7_63
42. Mistree K, Thakor D, Bhatt B (2021) Towards Indian sign language sentence recognition using INSIGN-VID: Indian sign language video dataset. Int J Adv Comput Sci Appl 12(8)
43. Morrissey S, Way A (2013) Manual labour: tackling machine translation for sign languages. Mach Transl 27. https://doi.org/10.1007/s10590-012-9133-1
44. Ong EJ, Koller O, Pugeault N, Bowden R (2014). Sign spotting using hierarchical sequential patterns with temporal intervals. https://doi.org/10.1109/CVPR.2014.248
45. World Health Organization (2013) Millions of people in the world have hearing loss that can be treated or prevented. WHO. encurtador.com.br/qOXZ8
46. Oszust M, Wysocki M (2013) Polish sign language words recognition with Kinect. In: 2013 6th International Conference on Human System Interactions (HSI). pp 219–226
47. Parelli M, Papadimitriou K, Potamianos G, Pavlakos G, Maragos P (2020) Exploiting 3D hand pose estimation in deep learning-based sign language recognition from RGB videos. In: Bartoli A, Fusiello A (eds) Computer Vision - ECCV 2020 Workshops. Springer International Publishing, Cham, pp 249–263
48. Pigou L, Dieleman S, Kindermans PJ, Schrauwen B (2015) Sign language recognition using convolutional neural networks. In: Computer Vision - ECCV 2014 Workshops. Springer International Publishing, pp 572–578. https://doi.org/10.1007/978-3-319-16178-5_40
49. Rastgoo R, Kiani K, Escalera S (2021) Hand pose aware multimodal isolated sign language recognition. Multimed Tools Appl 80(1):127–163. https://doi.org/10.1007/s11042-020-09700-0
50. Rastgoo R, Kiani K, Escalera S (2022) Real-time isolated hand sign language recognition using deep networks and SVD. J Ambient Intell Humaniz Comput 13(1):591–611. https://doi.org/10.1007/s12652-021-02920-8
51. Ronchetti F, Quiroga F, Estrebou C, Lanzarini L, Rosete A (2016) LSA64: a dataset of Argentinian sign language. XX II Congreso Argentino de Ciencias de la Computación (CACIC)
52. Ronchetti F, Quiroga F, Estrebou C, Lanzarini L, Rosete A (2016) LSA64: an Argentinian sign language dataset
53. Sharma S, Kumar K (2021) ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. Multimed Tools Appl 80(17):26319–26331. https://doi.org/10.1007/s11042-021-10768-5
54. Shoaib U, Ahmad N, Prinetto P, Tiotto G (2013) Integrating multiwordnet with Italian sign language lexical resources. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2013.09.027
55. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Preprint at http://arxiv.org/abs/1406.2199
56. de Souza MFNS, Araújo AMB, Sandes LFF, Freitas DA, Soares WD, de Mello Vianna RS, de Sousa ÁAD (2017) Principais dificuldades e obstáculos enfrentados pela comunidade surda no acesso à saúde: uma revisão integrativa de literatura. Revista CEFAC 19(3)395–405. https://doi.org/10.1590/1982-0216201719317116
57. Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ (2021) Classification of skin disease using deep learning neural networks with mobilenet V2 and LSTM. Sensors 21(8). https://doi.org/10.3390/s21082852. https://www.mdpi.com/1424-8220/21/8/2852
58. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. Preprint at http://arxiv.org/abs/1512.00567

59. Tran WT, Sadeghi-Naini A, Lu FI, Gandhi S, Meti N, Brackstone M, Rakovitch E, Curpen B (2021) Computational radiology in breast cancer screening and diagnosis using artificial intelligence. Can Assoc Radiol J 72(1):98–108
60. Vazquez-Enriquez M, Alba-Castro JL, Docio-Fernandez L, Rodriguez-Banga E (2021) Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp 3462–3471
61. Wadhawan A, Kumar P (2020) Deep learning-based sign language recognition system for static signs. Neural Comput Appl 32(12):7957–7968. https://doi.org/10.1007/s00521-019-04691-y
62. Wan J, Li SZ, Zhao Y, Zhou S, Guyon I, Escalera S (2016) Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp 761–769
63. Wu J, Sun L, Jafari R (2016) A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors. IEEE J Biomed Health Inform 20(5):1281–1290
64. Wu Z, Wang X, Jiang YG, Ye H, Xue X (2015) Modeling spatial-temporal clues in a hybrid deep learning framework for video classification
65. Yadav A, Verma D, Kumar A, Kumar P, Solanki P (2021) The perspectives of biomarker-based electrochemical immunosensors, artificial intelligence and the internet of medical things towardáCOVID-19 diagnosis and management. Mater Today Chem 20:100443
66. Ye H, Wu Z, Zhao RW, Wang X, Jiang YG, Xue X (2015) Evaluating two-stream CNN for video classification. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ICMR '15. Association for Computing Machinery, New York, NY, USA, pp 435–442. https://doi.org/10.1145/2671188.2749406
67. Zhang L, Zhu G, Shen P, Song J, Shah SA, Bennamoun M (2017) Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp 3120–3128