Expert Systems With Applications 183 (2021) 115295

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Mexican sign language segmentation using color based neuronal networks to detect the individual skin color

Josué Espejel-Cabrera, Jair Cervantes *, Farid García-Lamont, José Sergio Ruiz Castilla, Laura D. Jalili

Department of Computer Science, Autonomous University of Mexico State, Prolongación de Av. Zumpango s/n, Fraccionamiento el Tejocote, Texcoco, State of México, 56255, Mexico

ARTICLE INFO

Keywords:
Mexican sign language
Classification
Segmentation

ABSTRACT

In recent years, the development of algorithms that assist in communicate with deaf people is an important challenge. The development of automatic systems to translate sign language is a current research topic. However, this involves several processes that range from video capture, pre-processing to identification or classification of the signal. The development of systems capable of extracting discriminative features that enhance the power of generalization of a classifier is even a very challenging problem. The meaning of a sign is the combination of the hand movement, hand shape, and the point of contact of the hand in the body. This paper presents a method to detect and translate hand gestures. First, we obtain 15 frames per word, obtaining 3 regions of interest (hands and face) from which we obtain geometric features. Finally we use several classifier techniques and present the experimental results.

1. Introduction

In recent years, artificial intelligence technologies have been developed to meet human needs. The areas where artificial intelligence systems have been developed range from medicine (Hamet & Tremblay, 2017), in angiograms (Kerkeni, Benabdallah, Manzanera, & Bedoui, 2016), retinal studies (Li et al., 2016), researches in biology (Cervantes, Lamont, Mazahua, Hidalgo, & Castilla, 2018), food industry (Mohanty, Jane, & Bonas, 2018) to applications such as agriculture. All these approaches increase the significance of artificial intelligence in nowadays duties and daily aspects. Artificial Intelligence approaches constitute one of the leading research areas. Computer vision particularly has the new developing technologies and auxiliaries like pedestrian detectors (Li et al., 2017), autonomous vehicles (Talebpour & Mahmassani, 2016). Automatic identification of carcinogenic melanomas through images, identify and classify cells, crop pests identification, and automatic translation of sign language through images and videos, among many others.

The development of vision systems that help to communicate with deaf people is crucial for the academic and intellectual development of deaf people. In general, the communication channel is built on a mutual language among people according to the country. But for deaf people, it is impossible to engage in dialogue in this way. Sign language is the only natural way to communicate with all people.

The development of technology to help deaf people communicate is of enormous importance in today's connected world. Current research can be divided according to the flow of communication as follows:

1. Communication from a deaf person to a person who speaks. In this case, the methods used consist of speech and text recognition techniques. Once the text or voice is recognized, a video dictionary is associated with the specific word and finally shows an avatar or person making the corresponding signal gesture.
2. Communication from a person who speaks to a deaf person. In this case, the research focused on interpreting or translate the signs from image sequences or video.

To face the second problem, current research focuses on three different branches:

This document is the results of the research project funded by the Autonomous University of Mexico State. This research was supported in part by grant: 4996/2020CIB

* Corresponding author.
E-mail addresses: jcervantesc@uaemex.mx (J. Cervantes), fgarcial@uaemex.mx (F. García-Lamont).

https://doi.org/10.1016/j.eswa.2021.115295
Received 18 January 2020; Received in revised form 8 September 2020; Accepted 25 May 2021
Available online 6 June 2021
0957-4174/

Table 1
Advantages and disadvantages of main techniques used in sign language translation.

Techniques	Sensors	PDI/Video	3D modeling
Advantages	High precision during the hand gesture. No losing of data during the hand movement or translation. Easy acquisition of info about every movement or flexion by any individual finger.	Image processing is used because of their low complex gathering. There is not needed a high cost cameras or accessories in a complex set. The images can be taken in almost any different place.	3D modeling is invariant to luminance issues, clothes, or any wearing accessories. The modeling of the hand can distinguish the palm or back of the hand and any separate finger. It can model the entire hand movement across the body and shown in different angles.
Disadvantages	It is necessary to have a device that interprets the signals. The cost of the sensors could be expensive and difficult to implement. The change or addition of a sensor could represent modify the code and/or re-wire of sensors.	The image processing is too sensitive to brilliant changes, blur, or not focused images. Accessories like necklaces, lens, or bracelets could introduce noise to the image. The background color similar to the foreground color could difficult the image segmentation.	To create a 3D environment there is needed a set of cameras or body sensors to create the environment. Either the images are taken in a closed area or outdoors where it is necessary to put the complete set of cameras, this difficult the change of place. It could be expensive to acquire the necessary stuff to perform experiments like cameras, sensors tripods, and necessary software.
References	(Tubaiz et al., 2015; Kumar et al., 2017; Kuroki et al., 2015)	(Ahmed et al., 2016; Tharwat et al., 2015; Koller et al., 2015)	(Elons et al., 2013; Tan et al., 2016; Romdhane et al., 2016; Kishore et al., 2018)

(1) The use of sensors in the deaf person that measure any movement and displacement of the hand.
(2) The use of 3D technology to translate the signals.
(3) the identification or translation of signs through images or videos

In the latter, the collected images are pre-processed with filters to remove noise, blur, or bad lighting. The processed image segmented into regions of interest that are commonly hands and face. From the segmented regions, the features that define the shape of the hands in each frame obtained are extracted. With the features, the classifiers are trained and tested on sets never seen before to validate the results (Ahmed, Chanda, & Mitra, 2016; Tharwat, Gaber, Hassanien, Shahin, & Refaat, 2015; Koller, Forster, & Ney, 2015).

Although these methods suppose in some cases a very high performance, the main disadvantages lie in the sensors and devices required to implement them. Furthermore, due to its complexity in both implementation and pricing, it is difficult to develop implementations that directly help deaf people. Video-based sign recognition systems are a clear advantage since only a cell phone with a camera is necessary. Some methods have been proposed in the literature. However, the use of excessive processes makes them unfeasible to be implemented in real-time.

In this paper, we develop a method based on video acquisition and image processing. From each video of the selected words, we get 15 frames, then we trained a neural network to detect skin color and segment the frames. This allows us to improve segmentation despite changes in global brightness conditions and reduce the processing time by segmenting the image based on skin color. The segmentation process obtains three regions of interest (hands and face). From the regions of interest, we extracted several geometric features. The geometric features allow us to capture different combinations of movements of the hand and points of contact of the hand with the body. Finally, we trained the models on the training data-set using several techniques of supervised learning. Finally, we get the performance on several precision metrics.

We structured this paper as follows; Section 2 describes the related work in sign language translation, Section 3, shows the proposed method, Section 4 explains the experimental results. Finally, Section 5 discusses the conclusions.

2. Related work

In recent years, many work have been developed to address the problem of identification and translation of sign language from images or videos.

The sign language identification methods in the literature can be divided into the following three groups: Methods based on Sensors, methods that use digital image processing, and video/images in 3D modeling.

2.1. Methods based on sensors

The sensors used in sign language are mainly used in gloves (Kumar, Gauba, Roy, & Dogra, 2017; Galka, Masior, Zaborski, & Barczewska, 2016; Abhishek, Qubeley, & Ho, 2016; Naglot & Kulkarni, 2016). The accelerometers, flex sensors, and switches are the most used sensors to translate the sign language. These sensors help to get information about the flexion or finger's stretching, translation or rotation of the entire hand, and the touch of the hand with some part of the body.

Researchers have not only focused on using basic sensors to solve this problem. At present, more sophisticated techniques and devices have been used such as leap motion (Quesada, López, & Guerrero, 2016), Mio armband (García et al., 2008; Abreu, Teixeira, Figueiredo, & Teichrieb, 2016), Gloves with flex sensors (Ali, Mushtaq, & Memon, 2016; Mohammad Iqbal & Supriyati, 2017; Tubaiz, Shanableh, & Assaleh, 2015; Galka et al., 2016).

With the examples mentioned above, the difficulty show implementing sensors it's found in the signals, the order of the signal codification, and the combination of different kind of flex sensors, accelerometers, etc., and also, the maintenance of these sensors. The sensors make it easier for us to obtain specific signals, but combining different types also leads to difficulties in interpreting the signals used for translation into sign language.

2.2. 3D modeling

The 3D hand model shows all the finger joints and can distinguish the palm or the back of the hand. Body representation can get the entire movement of the hands, head, feet, or can model a complete 3D avatar of a person and its movements. The use of infrared sensors, body markers, or multiple cameras can help to produce the expected model. The Kinect sensor can model a complete 3D avatar of a person and its movements, but the leap motion sensor only can create a 3D model of the hands in a specific position.

Current systems have also made use of other devices such as leap motion sensor (Mapari & Kharat, 2015; Mohandes, Aliyu, & Deriche, 2015; Chong & Lee, 2018), Kinect sensor (Aliyu, Mohandes, Deriche, & Badran, 2016; Sun, Zhang, Bao, & Xu, 2013), infrared cameras (Kumar, Kishore, Sastry, Kumar, & Kumar, 2018Li Li, Yu, Wu, Su, & Ji, 2015).

However, the sensors used have the disadvantage of the infrared

Table 2
List of words used in the experiments.

Greetings	good morning, good afternoon, good night,thanks, please, see you,
Time	day, hour, week, minutes, seconds
Days of the week	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Months	January, February, March, April, May, June, July, August, September, October, November, December
School stuff	Qualification, Class, Notebook, School, Pencil, Lecture, Sign Language, Exam, Writing, Sharpener, Ruler, Colors
Parents and Relatives	Father, Mother, Brother, Sister, Cousin, Aunt, Uncle, Grandfather, Grandmother, Wife, Husband, Family, Son, Daughter, Man, Woman, Girlfriend, Boyfriend, Cousin, niece, nephew,
House stuff	Room, Bathroom, Kitchen, Bed, House, Closet, Curtains, Cradle, ceiling, Floor, Stairs, Broom, Lamp, Table, Wall, Living room, Glass, Window, Floor, Door
Adjectives	Adult, Young, Kid, Baby, Ugly, Beautiful, Bad, Deaf, Dumb, Strong, Fat, Tall, Small, Good, Blind, Weak, Thin
Meals	Spoon, Knife, Plate, Glass, Lunch, Dinner, Breakfast, Fork, Napkin, Salt, I want more, I didn't like it
Clothes	Earrings, Blouse, Boots, Shirt, Necklace, Glove, Pants, Pijama, Shorts, Suit, Dress, Shoes, Skirt, Underwear
Body parts	Beard, Mustache, Arm, Mouth, Hair, Head, Face, Feet, Teeth, Eyes, Eyebrow, Nose, Ears, Cheek, Fingernail
Vehicles	Airplane, Ship, bicycle, Truck, Helicopter, Car, Motorcycle, Taxi, Tractor, Train, Subway, Combi, Van
Places of interest	Airport, Library, Downtown, Cinema, Circus, Building, Hospital, Hotel, Market, Museum, Restaurant, School, Supermarket, Cafeteria
Pronouns	I, You, He, She, They(masculine), They(feminine), Us (Masculine), Us(Feminine), You(plural), No one, Someone,
Verbs	Hug, Love, Fix, Frighten, Help, Find, Shut up, Close, Believe, Eat, Stop, Sleep, Hit, Save, Play, Lift up, Cry, Lie, Hear, Forget, Make, Laugh, Trow, Sort, Clean
Jobs	Actor, Firefighter, Doctor, Teacher, Waiter, Police officer, President, Secretary, Carpenter, Mecanic, Shoemaker, Stylist, Seamstress
States of México	Aguascalientes, Baja California Nte. Baja California Sur, Campeche, Chihuahua, Chiapas, Coahuila, Colima, Durango, Guanajuato, Guerrero, Hidalgo, Jalisco, México, Michoacán, Morelos, Nayarit, Nuevo León, Oaxaca, Puebla, Querétaro, Quintana Roo, San Luis Potosí, Sinaloa, Sonora, Tabasco, Tamaulipas, Tlaxcala, Veracruz, Yucatán, Zacatecas

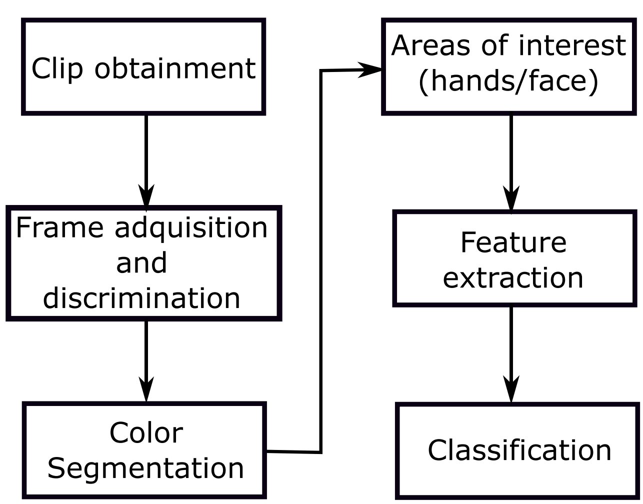

Fig. 1. Proposed methodology.

noise during the hand gesture compilation. Moreover, the use of two or more sensors such as leap motion increases the information obtained, causing a substantial increase in redundant data.

Fig. 2. Frame sequence of the videoclip.

2.3. *Methods that uses digital image processing and video techniques*

Another general area of research is based on translating sign language from images. Unlike the sensors whose work with electrical signals generated depending on the movement of the hand, this area doesn't need extra devices. Devices like smartphones are widely used for almost all people around the world. Although the first cellphones appeared whit only a few characteristics, the most recent models have the same capability as PCs. In general, we can use ordinary cameras, like cellphone cameras, and use image processing to get the necessary information.

Current research on translating sign language from images can be classified by the techniques used to solve the problem. Those that need to pre-process the image, segment it, and obtain different types of features (Raheja, Mishra, & Chaudhary, 2016; Kishore, Prasad, Kumar, & Sastry, 2016; Rao & Kishore, 2018) and techniques using convolutional neural networks (CNN) (Camgoz, Hadfield, Koller, & Bowden, 2017; Dudhal, Mathkar, Jain, Kadam, & Shirole, 2019; Koller, Zargaran, Ney, & Bowden, 2018).

However, the results obtained are not comparable between them due to multiple factors. One of the most important is that they have been developed to translate a specific sign language with a particular complexity. Furthermore, some research is only focused on identifying sign language of static signs like the alphabet and numbers (Tao, Leu, & Yin, 2018; Hamed, Belal, & Mahar, 2016; Oliveira et al., 2017), while others focus on words and phrases (Camgoz et al., 2017; Cui, Liu, & Zhang, 2017; Lim, Tan, & Tan, 2016) whose complexity is greater.

Table 1 shows the advantages and disadvantages of the methods used in the state of the art to face the problem of automatic identification of signs.Table 2.

3. Proposed method

The proposed methodology is shown in Fig. 1. The proposed method comprises of 7 steps, whose are described detailed in the next linesFig. 2.

3.1. *Clip obtainment*

First, we consider 249 words of the Mexican Sign Language (MSL) divided as shown in Table ref words. Eleven people helped collect videos of the selected words. During the taking of videos, we use a specific background and clothing. The color of the background and clothes were black to enhance the areas of interest for this research (hands and face). We recorded the videos with a simple video camera on a tripod. In every sign, the rest position, or initial hand's position, the hands are near the legs and began the movement depending on the specific hand gesture. The gesture time was between 1 and 3 s. Each signal depends on every person's hand speed, detail during the gesture, and its complexity.

We collected 2480 videos of sign words, with standard consumer-grade cameras. We recollected videos from the "Centro Educativo para el Sordo" (CES) institute. The signs made in these videos come from eleven different individuals on average for each of the 249 words collected. Each video contains just one sign of the Mexican language. In videos, the individual is in front of the camera, with the left and the right hands-on rest. Then the individual begins the movement of the hand gesture, finishing the individual with its hands in rest.

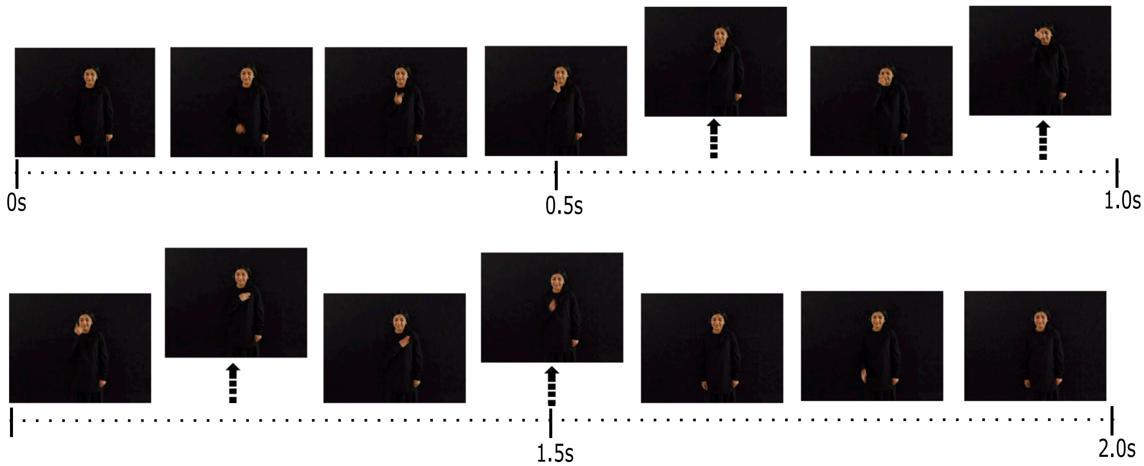

Fig. 3. Selecting and removing frames from the original sequence of frames.

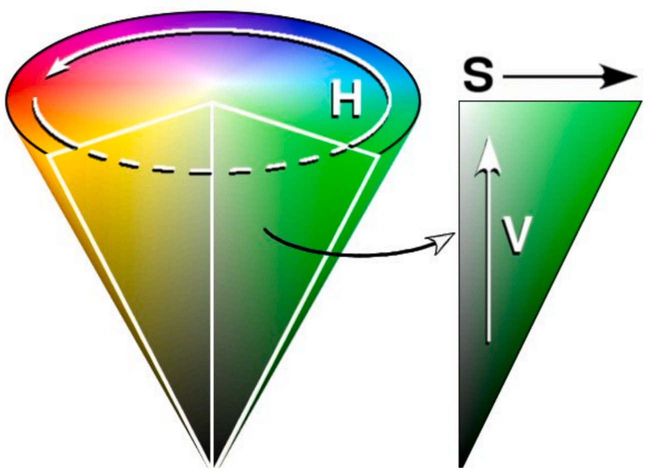

Fig. 4. HSV Space.

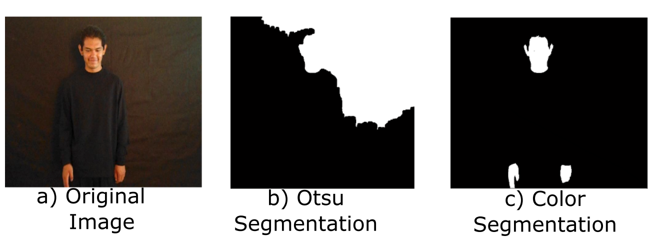

Fig. 5. Regions of interest.

3.2. *Frame acquisition and discrimination*

From the videos of the individuals, we get a frame sequence. These frames describe the sequence of the gesture. The average of frames in each sign is 15 frames per video.

However, not all frames contain relevant data. For example, the initial and final frames, both show the individual with the hands in the rest position. If we know that the initial and final hand position of the hands is almost the same, then we can ignore that section in the frame sequence. Counting with this particular point, we implement a method to acquire the main section of the frame sequence, which is the middle part of the sequence, because it contains the principal hand movements to denote the sign, as we show in Fig. 3.

As we note, the frames f obtained compound the middle section a can explain the principle hand movement and get with it the entire representation of the sign gesture. We chose the frames by the rule of Eq. (1)

$$D = \sum f \in F\left\{f_{\frac{n}{2}-3}, f_{\frac{n}{2}-1}, f_{\frac{n}{2}+1}, f_{\frac{n}{2}+3}\right\} \tag{1}$$

where n is the number of frames obtained.

3.3. *Segmentation techniques*

In this section, we introduced a new method used to segment the images. Initially, in the experiments, we use another kind of segmentation techniques like Otsu and cluster segmentation, but the results were very poor. Although the background is black for all images, the lighting or reflection did not allow segmenting the images.

Fig. 5 shows the original image (a) and the segmentation Otsu method used as (b). The Otsu method is based on maximizing entropy. In general, the Otsu method has excellent results, but in the proposed image data-set has no good results. Since the brightness of the images is not controlled and the image is converted to grayscale. Some parts of the region of interest have similar hues to the background of the image. This causes a segmentation of poor quality. In our experiments, the classical techniques tested obtained similar results.

The proposed segmentation method is based on the color space HSV. The advantages of HSV space color against the brightness fluctuations on images makes it a good choice. In HSV space color, the color is defined by a specific region of the base of a cone denoted for its angle α, the change of brightness or darkness is denoted for the height h of the cone in the same angle α of the base. Then if we find the specific color denoted by α, the differences of brightness do not affect, therefore if we find the specific α we can segment the regions belonging to the regions of interest.

Finally, we designed a technique of color segmentation to face these disadvantages. The segmentation algorithm that we used in the experiments is described entirely in this section.

3.3.1. *Color segmentation*

The HSV (Hue, Saturation, Value) (see Fig. 4) space of color, is one of the most used color spaces in Digital Process Image. HSV space is less susceptible to luminance variance in images that RGB (Raheja et al., 2016). In experiments carried out, we found that the process followed during the video obtainment, the control of the clothing's color, and the background, enhance the segmentation of the region of interest. Moreover, in our case, the pre-processing of images doesn't improve the accuracy of the classifiers. The controlled environment decreased enough the possible issues/noise in the videos. Moreover, the luminosity

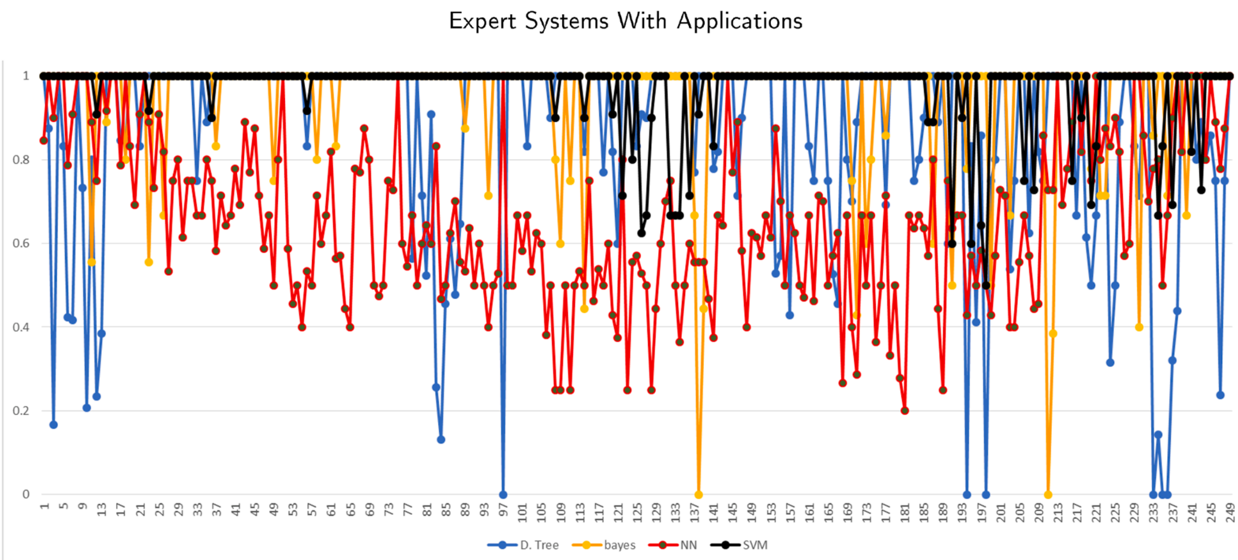

Fig. 6. Comparison of performances using precision measure.

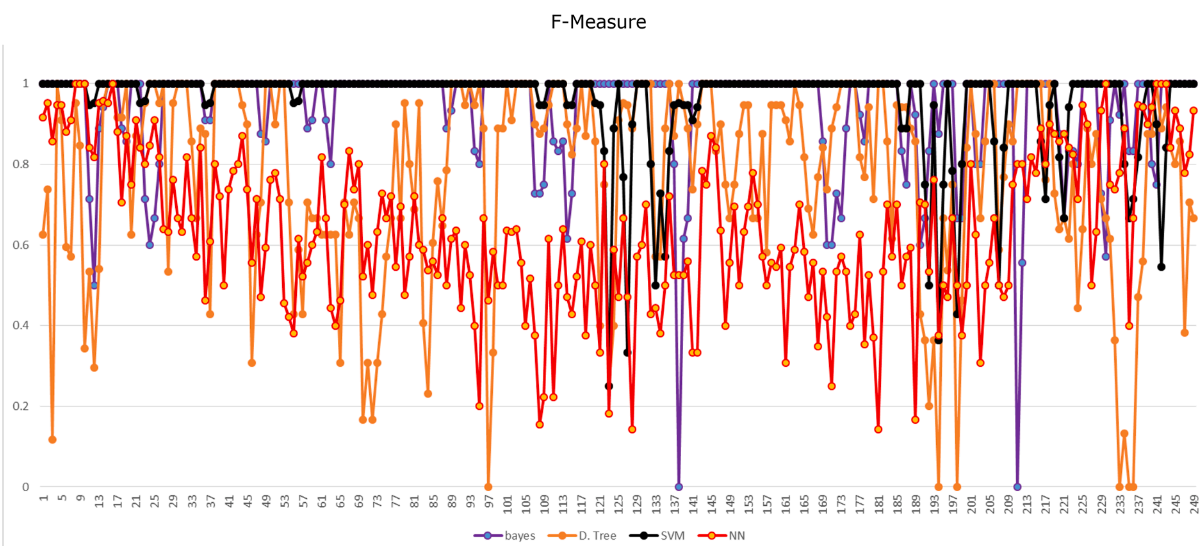

Fig. 7. Comparison of performances using f-measure.

variations were eliminate using the HSV space.

The Skin color segmentation process begins with the train of a neural network (NN) to find the colors in HSV. The proposed skin color segmentation process is shown in the Algorithm 1.

Algorithm 1: Skin Color Segmentation algorithm

Initialize: Create a Neuronal Network with a 5×5 *neuron matrix*
Take the space HSV for train the neurons in the next process:
Take the Hue(H) in the (HSV) space as a vector $\varphi[\sin\theta, \cos\theta]$
For *angle* $\theta \in \{0, 15, 30, \ldots, 360\}$
$\varphi_i = \mathrm{H}\theta$
Train $\varphi_i \rightarrow$ NN
End For
For f = 0: n ≜ For every frame
$\mathrm{f}[R, G, B] \rightarrow f[H, S, V]$ ≜Convert from RGB into HSV Space
For p(x,y), $\forall x = 1, \ldots, n,\ y = 1, \ldots, m$ ≜For every pixel in the frame
$p_{x,y} \rightarrow NN = p_{x,y}$(new) ≜The excited neuron represent the color will assign
End For
$\mathrm{f}[H, S, V] \rightarrow f[R, G, B]$ ≜Image Segmented
End For ≜ Details in García-Lamont, Cervantes, and López-Chau, 2016

3.4. Areas of interest

The signs of the Mexican sign language are described by a configuration of shape, place, and movement of the hand. To get information about these configurations it is necessary to define the regions of interest. The regions of interest are described by the configuration of the hands and head (3 regions). However, the hands are not always separated from the body or the head. In some signs, the hands are together or one hand is close to the head (2 regions). In other cases, due to the sign movements hands passes near to the head, combining both hands and face in the same region of interest (1 region). In the case of 1 or 2 regions of interest, the system infers that the hands or the face are too close or over the other regions. The shape of the hand, place of the fingers, their joint's flexion, or relaxation, all together give us specific information about the sign meaning.

However, in the words of the MSL, there are also movements and changes in the place of the hand. The regions of interest of the frames combined during the movement are not the same for all the examples. Each person has individual manners during the hand movement just like

Table 3
Table comparison of the bad performance of the compared classifiers.

Bayes				D. Tree				SVM				NN				
Precision	F-Measure	Recall	ROC	Precision	F-Measure	Recall	ROC	Precision	F-Measure	Recall	ROC	Precision	F-Measure	Recall	ROC	Class
0.556	0.714	1	0.998	0.8	0.533	0.4	0.959	1	0.947	0.9	1	0.889	0.842	0.8	0.923	11
0.55	0.71	1	0.99	1	0.952	0.909	1	1	0.957	1	1	0.889	0.8	0.727	1	23
0.429	0.6	1	0.998	0.889	0.889	0.889	0.995	1	1	1	1	0.286	0.25	0.222	0.994	171
0.5	0.667	1	0.999	1	0.364	0.222	0.923	0.6	0.75	1	0.999	0.636	0.7	0.778	0.953	191
0.385	0.556	1	0.995	1	1	1	1	1	1	1	1	0.727	0.8	0.889	0.999	212
0.4	0.571	1	1	0.714	0.667	0.625	0.914	1	1	1	1	1	1	1	1	230
1	1	1	1	0.909	0.952	1	1	0.625	0.769	1	0.999	0.529	0.667	0.9	0.999	126
1	1	1	1	0.9	0.947	1	1	0.667	0.333	0.222	0.998	0.5	0.471	0.444	0.998	127
1	1	1	1	0.833	0.667	0.556	0.986	0.6	0.76	1	0.999	0.571	0.5	0.444	0.999	195
1	1	1	1	0.857	0.75	0.667	0.98	0.643	0.783	1	0.999	0.583	0.667	0.778	0.998	197
1	0.667	0.5	0.996	0	0	0	0.979	0.5	0.429	0.375	0.999	0.5	0.5	0.5	0.997	198

Table 4
General results of the classifiers with different measures performance.

Classifier	Precision	AUC	Recall	F-Measure
Bayesian	93.74%	99.9%	93.70%	93.45%
Decision Tree	77.22%	95.17%	77.10%	77.72%
SVM	96.27%	99.9%	96.30%	95.90%
NN	64.89%	99.2%	64.90%	64.10%

manners in writing, speaking, and other aspects. Not always the hands of the people have the same shape, but like any specific letter that has a general stroke, MSL words have a specific hand movement. The goal of the proposed method is to find these patterns to correctly classify the signs of the MSL.

Finally, in Fig. 5 *c* can see the final segmentation of the areas of interest (hands and face), the Algorithm 1 proposed for color segmentation is described in the Skin Color Segmentation algorithm.

3.5. Feature extraction

The proposed method extracts geometrical features of every region of interest. Some frames do not have three regions due to one hand is close to the other or hands close to the face. To avoid issues during the feature's extraction, we replicate one region with zeros if the frame have not all the regions of interest, we detail the replicate method in the next Algorithm 2, obtaining the features in vector v

Algorithm 2: Feature's Extraction

```
Features r_1, r_2, ...r_p ∈ R          ≜The hands and face
  for f = 0 : n          ≜every frame in the clip
  for r = 1:p
    a← Hu moments(r)
    b← Fourier Descriptors(r)
    c← Ellipse(r)
    d← Gupta Descriptors(r)
    ⋮
    q← Flusser moments(r)
  return a + b+c + d+ … + q          ≜ The entire features array
```

3.5.1. Geometric features

In Digital Image Processing, the area of interest in the image has some features, which may change depending on the rotation, translation, or scaling of the image. However, we also can compute invariant measurements and can be used to determine the shape of the image. For this paper, we use the features detailed below.

In the experiments we use different techniques of features extraction as moments that are invariant to translation, rotation and inversion. The central moments are invariant to displacement and Hu moments are invariant to displacement, rotation and scale and can be calculated as is showed in detail in (Hu, 1962). The Fourier descriptors are used to determine the contour of the shape by two parametric equations. More details can be found in (Zahn & Roskies, 1972). The ellipsity determines the similarity of the border of the shape with the geometrical form (Sladoje & Žinić, 1997). The gupta moments are invariant to translation and scaling (Gupta, 1987). Sometimes it is necessary to obtain invariant characteristics to translation, rotation, scale and related transformations. Flusser moments obtain characteristics derived from the central second and third order moments that are invariant to affine transformations (Flusser, 1993).

Finally, the geometric feature vector X_g obtained can be represented as:

$$Xg = \left[x_{gb}, x_{Hu}, x_F, x_G, x_{DF}\right] \quad (2)$$

where x_{gb} represents the elemental geometric features ($x_{gb} = [x_1, \ldots, x_p]$), x_{Hu} represents the Hu invariant features ($x_{Hu} = [x_{p+1}, \ldots, x_q]$), x_F represents the Flusser invariant moments ($x_F = [x_{q+1}, \ldots, x_r]$), x_G represents the Gupta moments ($x_G = [x_{r+1}, \ldots, x_s]$), x_{DF} represents the first Fourier

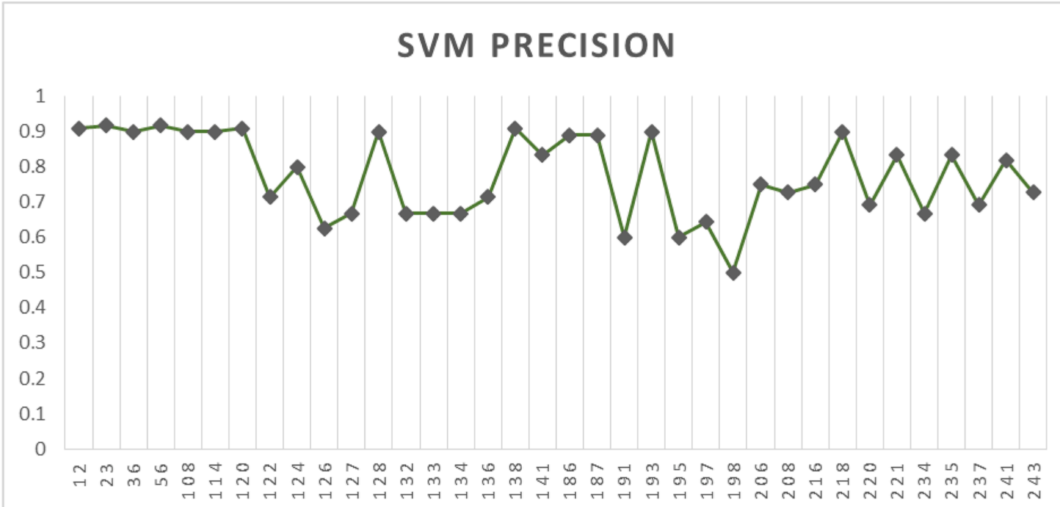

Fig. 8. Lower accuracy results with SVM.

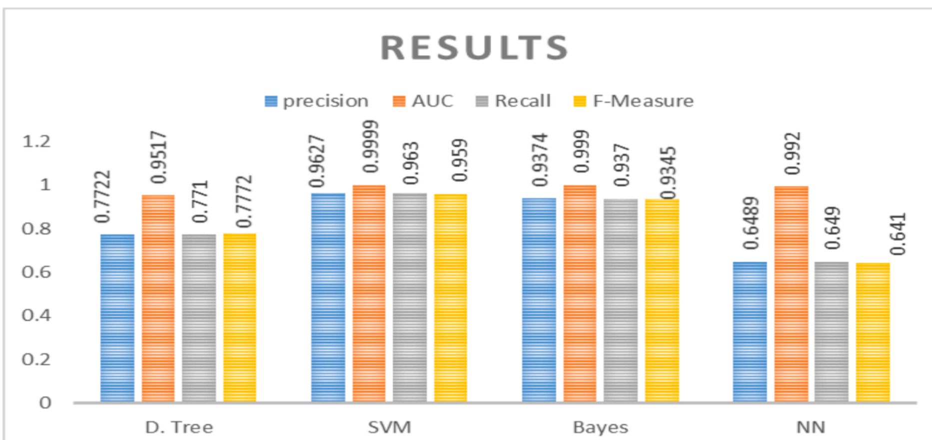

Fig. 9. General results of the classifiers with different measures performance.

Table 5
General Technique comparison.

Technique	Methodology	accuracy	Reference
3D Modeling	C3D model using, normal cameras, infrared cameras, corporal sensors	99.35%, 97%, 99.10%	(Kumar et al., 2018; Aliyu et al., 2016; Li et al., 2015)
Sensors	Use globes with flex sensors, accelerometers and contact sensors	99.3%, 98.9%, 99.75%	(Ali et al., 2016; Tubaiz et al., 2015; Galka et al., 2016)
PDI	Use Images from the entire	97.5%, 92.8%, 93.9%	(Raheja et al., 2016; Dudhal et al., 2019; Camgoz et al., 2017)
Proposed	Use Skin color segmentation to obtain the face and the hands region to obtain the geometrical features	96.27%	

descriptors ($x_{DF} = [x_{s+1}, \ldots, x_t]$).

3.6. *Classification techniques*

The total amount of frames from the 249 videos is 2241, the size of the frames is 640 $\times$ 480 pixels in.jpg format, the data-set obtained with the features selected is a matrix of size 2241 $\times$ 684. The principal classification method is SVM, and we compare the results of the proposed classifier with Naive Bayes, Decision Tree, and Neuronal Networks.

3.6.1. *Support Vector Machines (SVM)*

In last years, SVM is one of the most used classification techniques.

The key characteristics of SVM's are the use of kernels on non-linear data sets, the absence of local minimums, and the solution depends on a small subset. The discriminative power of the model lies in the optimality of the margin. These characteristics allow SVMs to obtain very competitive results compared to other classifiers. Formally, SVM can be defined as follows:

Assuming that a training data set $X$is given as:

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n) \tag{3}$$

i.e. $X = \{x_i, y_i\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^d$ and $y_i \in (+1, -1)$. Training an SVM allows you to solve a quadratic programming problem as follows:

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i y_i \alpha_j y_j \mathbf{K}\langle x_i \cdot x_j\rangle + \sum_{i=1}^{l}\alpha_i$$
$$\text{subject to}: \sum_{i=1}^{l}\alpha_i y_i = 0, C \geqslant \alpha_i \geqslant 0, i = 1, 2, \ldots l \tag{4}$$

where $C > 0, \alpha_i = [\alpha_1, \alpha_2, \ldots, \alpha_l]^T, \alpha_i \geqslant 0, i = 1, 2, \ldots, l$, are coefficients that correspond to x_i, x_i with α_i nonzero which are called Support Vectors (SV).

SVMs are linear classifiers, that is, they are applicable when the two classes of data in the training set are linearly separable. When this is not possible, a function called kernel is used to transform the input space into a higher-dimensional space, where the two classes can be separated linearly after the transformation. However, the choice of a function is restricted to those that satisfy the conditions of Mercer (Vapnik, 1995).

Let S be the set of SVs obtained after the training, then the optimal hyperplane is given by:

$$\sum_{i \in S}(\alpha_i y_i)K(\mathbf{x}_i, \mathbf{x}_j) + b = 0 \tag{5}$$

the optimal decision function is defined as

$$f(x) = sign(\sum_{i \in S}(\alpha_i y_i)K(\mathbf{x}_i, \mathbf{x}_j) + b) \tag{6}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_l]$are the input data, α_i and y_i are the Lagrange multipliers. A new x object can be classified using (6).

The vector $\mathbf{x}_i$ is given in the form of dot product. There is a Lagrange α multiplier for each training data. When the maximum margin of the hyperplane is found, only the data most near the hyperplane satisfy $\alpha > 0$. These points are the support vectors (*SV*).

3.6.2. Neural networks

A neural network is a set of interconnected artificial neurons that use mathematical models to process information. The multiple connections between the neurons form an adaptive system whose weights are updated using a particular learning algorithm. Neural networks have been used in several application fields with different learning algorithms (Portillo, Cabanes, Marcos, & Zubizarreta, 2009; Macías, Roca, Fals, Fernández, & Muro, 2013; Rossomando, Soria, & Carelli, 2010; Gil & Páez, 2007).

One of the most used learning algorithms is the *backpropagation* (BP). To carry out the learning process, the BP learning algorithm iteratively changes the weights between the neurons minimizing the quadratic error between the desired output and that obtained with the current weights. Each of the examples of the training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$are used to adjust the weights in the network. When an example is presented, the signal is propagated forward of the network until the output is obtained. The output of the j^{th} hidden unit is calculated as:

$$o_{nj}^{h} = f_j^h\left(net_{nj}^{h}\right) = \frac{1}{1 + exp(-net_{nj}^{h})} \tag{7}$$

where $net_{nj}^{h} = \sum w_{ji}^{h} x_{ni} + \theta_j^h$. w_{ji}^{h} is the weight of the connection from the i^{th} input neuron to the j^{th} hidden neuron. θ_j^h y f_j^h represent the bias and activation function of the j^{th} hidden neuron. On the other hand, the output of the k^{th} neuron is given by $\theta_j h$ and $f_j h$

$$o_{nk}^{o} = f_k^o\left(net_{nj}^{o}\right) = \frac{1}{1 + exp(-net_{nk}^{o})} \tag{8}$$

where the superscripts h and o refer to the quantities in the hidden and output layers respectively. The error between the current output and the desired output is calculated to adjust the weights using $E_n = \frac{1}{2}\sum_{k=1}^{C}\left(t_{nk} - o_{nk}^{o}\right)$. The adjustment procedure is obtained from the descending gradient method to reduce the magnitude of the error. The procedure is first applied to the weights in the output layer and back-propagated through the net until the weights in the first layer have been adjusted $\Delta w_{kj}^{o} = -\eta \frac{\partial E_n}{\partial w_{kj}^{o}}$ and $\Delta w_{ji}^{h} = -\eta \frac{\partial E_n}{\partial w_{ji}^{h}}$. This procedure is performed for each example in the data set until a stop criterion is met. For an in-depth study of the algorithm you can refer to (Rumelhart & Hinton, 1986; Werbos, 1994).

3.6.3. Naive Bayes

Bayesian classifiers are based on Bayes decision theory. Bayes' principle provides a fundamental methodology for solving pattern classification problems when the probability distribution of the patterns is known. A Bayesian classifier uses a probabilistic approach to assign the class to an example. Let C be the class of an object that belongs to a set of m classes $(C_1, C_2, \ldots, C_m)$and $X_k$an object with k features $X_k = [x_1, x_2, \ldots, x_k]$, in our case it is the set of features that define a sign. The Bayesian classifier calculates the subsequent conditional probability $p(C_i|X_k)$ using the Bayes rule:

$$p(C_i|X_k) = \frac{p(X_k|C_i)p(C_i)}{p(X_k)}, i = 1, 2, \ldots, m. \tag{9}$$

The conditional probabilities in Eq. (9) $p(X_k|C_i), p(C_i)$ and $p(X_k)$ are calculated from training data. According to Bayes (Ng & Jordan, 2002; Russell & Norvig, 2003) theory, for a given observation (X_k, the class to which it belongs is given by the maximum posterior probability:

$$f(X_k) = \underset{i}{\operatorname{argmax}}\, p(C_i|X_k) \tag{10}$$

For a more thorough study you can refer to (Ng & Jordan, 2002; Russell & Norvig, 2003).

4. Experimental results

In the experiments, each image was segmented and the geometric features were obtained for every area of interest. From each image, we get one vector with 3 sets of geometric features (57×2241). The selected features were the Fourier descriptors, Hu moments, Ellipse, Gupta descriptors, Flusser moments.

The final Data-Set obtained of every word is a set of 4 images per word, and 3 regions each, and the final array of size 684×2241. The Data-set was used to train and test all the classifiers used in this paper. In the experiments, we use cross-validation to get the best parameters for each classification technique.

We consider the general precision of every classifier, the F-Measure, and the AUC. To validate the results, we use k-fold cross-validation with $k = 10$.

The general precision is the number of true positives (TP) divided by the sum of true positives and false positives (FP), as follows:

$$precision = \frac{TP}{TP + FP} \tag{11}$$

The Recall is obtained by the number of true negatives (TN) divided by the sum of true negatives and false negatives (FN), as follows:

$$Recall = \frac{TN}{TN + FN} \tag{12}$$

The F-Measure is a measure composed by the precision obtained and the Recall:

$$F - Measure = 2 \cdot \frac{precision \cdot Recall}{precision + Recall} \tag{13}$$

The AUC(Area Under Curve) denotes the capability of the classifier to distinguish and correctly classify the inputs

It is very difficult to show all the results. The number of classes or signs words to be identified is too long. Table 4, and Fig. 9, show the average performance obtained of the 249 signs of the Mexican sign language, using different performance metrics. Figs. 6, and 7, show the precision, and F-measure results on the 249 signs respectively. The Figures show the measures obtained with different classifiers. The lines in blue show the results obtained with the decision tree, in red the results obtained with neural networks, in yellow the results obtained with the Bayesian classifier, and, in black the results obtained with SVM. It is clear to note that in most cases, the performance with SVM improves the results obtained in comparison with the other classifiers. On average, the best performance was obtained by SVM. However, in some cases, the SVM performance is low. Fig. 6 shows the worst results obtained with SVM. Table 3 shows the results obtained with the four classification methods, denoting some of the worst results obtained with SVM.

Finally, Table 4 shows the general results of the classifiers, its precision, the sensitivity, and the AUC(area under curve) of the ROC curve. The worst performance is shown in Decision Tree with 77.22% (Precision), and the best performance is shown in the proposed classifier SVM, with a precision of 96.27%.

The experimental results show that the best performances were obtained with the SVM and the worst performances were obtained with the NN. Fig. 8 shows the numbers of the signs where the SVM performance was lower. In the Figure, it is possible to see that only 9 classes perform less than 70% accuracy. Fig. 9 shows the general accuracy obtained with the proposed classifiers and also shows the experimental results obtained with AUC-ROC measure.

Table 5 shows the results reported by other methods. The table shows the results obtained with 3D-based techniques, sensor-based techniques, and techniques similar to our proposal. However, this Table is just representative. It is not possible to directly compare the performances of these techniques with our proposal. Since the data set for each of the authors is different in size and complexity. In addition to this, each sign of each language has its complexity. Since each sign language has its own rules, complexity, and movements, then a general comparison is difficult to make.

In Table 3 we note that class 181 has the lowest precision with 20% using Neuronal networks. That is the sign with the worst performance, also in the general performance per classifier, shows that a multilayer perceptron achieved the lowest accuracy. The architecture used in this paper is a multilayer perceptron with 2 hidden layers 80 neurons and 120 respectively in the hidden layers and sigmoidal activation functions. Although we use a grid search to tune the best parameters, we can not obtain similar performance as the other classifiers, and maybe the use of a different architecture could improve the performance in the MSL.

5. Conclusions

In this paper, we propose a robust method of chromatic segmentation based on Mexican Sign Language in the HSV space. The proposed system uses a NN to automatically detect the skin color in the images. This allows us to improve the segmentation despite changes in global brightness conditions. Also, the extracted features in this paper allow us to obtain a good performance without making use of techniques for feature selection.

Fig. 6 shows that results in all the classifiers is acceptable. However, the performance obtained with SVM outperforms the other classifier used in this paper. Figs. 6, and 7 show that the line corresponding to SVM (black) has minimal variations to precision and F-measure metrics.

On the another hand, although the results are promising, the experiments were carried out only on 249 words. A current challenge is the identification of phrases. In this case, identification is difficult and is required to implement algorithms for natural language processing. Furthermore, the technique of selecting frames in this paper could not work for the identification of phrases, because only a sequence of frames is considered for words and not for phrases.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abhishek, K. S., Qubeley, L. C. F., & Ho, D. (2016). Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In *2016 IEEE international conference on electron devices and solid-state circuits (EDSSC)*. IEEE.

Abreu, J. G., Teixeira, J. M., Figueiredo, L. S., & Teichrieb, V. (2016). Evaluating sign language recognition using the myo armband. In *2016 XVIII symposium on virtual and augmented reality (SVR)*. IEEE.

Ahmed, W., Chanda, K., & Mitra, S. (2016). Vision based hand gesture recognition using dynamic time warping for indian sign language. In *2016 international conference on information science (ICIS)*. IEEE.

Ali, A., Mushtaq, A., & Memon, A. Monna (2016). Hand gesture interpretation using sensing glove integrated with machine learning algorithms. *International Journal of Mechanical and Mechatronics Engineering, 10*(11), 1857–1861.

Aliyu, S., Mohandes, M., Deriche, M., & Badran, S. (2016). Arabie sign language recognition using the microsoft kinect. In *2016 13th international multi-conference on systems, signals & devices (SSD)*. IEEE.

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). SubUNets: End-to-end hand shape and continuous sign language recognition. In *2017 ieee international conference on computer vision (iccv)*. IEEE.

Cervantes, J., Lamont, F. G., Mazahua, L. R., Hidalgo, A. Z., & Castilla, J. S. R. (2018). Complex identification of plants from leaves. In *Intelligent computing methodologies* (pp. 376–387). Springer International Publishing.

Chong, T.-W., & Lee, B.-G. (2018). American sign language recognition using leap motion controller with machine learning approach. *Sensors, 18*(10), 3554.

Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. *IEEE Conference on Computer Vision and Pattern Recognition*, 7361–7369.

Dudhal, A., Mathkar, H., Jain, A., Kadam, O., & Shirole, M. (2019). Hybrid SIFT feature extraction approach for indian sign language recognition system based on CNN. In *Proceedings of the international conference on ISMAC in computational vision and bio-engineering 2018 (ISMAC-CVB)* (pp. 727–738). Springer International Publishing.

Elons, A. S., Abull-ela, M., & Tolba, M. (2013). A proposed PCNN features quality optimization technique for pose-invariant 3d arabic sign language recognition. *Applied Soft Computing, 13*(4), 1646–1660.

Flusser, S. T. J. (1993). Pattern recognition by affine moment invariants. *Pattern Recognition, 26*(1), 167–174.

Galka, J., Masior, M., Zaborski, M., & Barczewska, K. (2016). Inertial motion sensing glove for sign language gesture acquisition and recognition. *IEEE Sensors Journal, 16*(16), 6310–6316.

García, J., Molina, J. P., Martínez, D., García, A. S., González, P., & Vanderdonckt, J. (2008). Prototyping and evaluating glove-based multimodal interfaces. *Journal on Multimodal User Interfaces, 2*(1), 43–52.

García-Lamont, F., Cervantes, J., & López-Chau, A. (2016). Human mimic color perception for segmentation of color images using a three-layered self-organizing map previously trained to classify color chromaticity. *Neural Computing and Applications, 30*(3), 871–889. https://doi.org/10.1007/s00521-016-2714-9

Gil, R. V., & Páez, D. G. (2007). Identificación de sistemas dinámicos utilizando redes neuronales RBF. *Revista Iberoamericana de Automática e Informática Industrial RIAI, 4*(2), 32–42.

Gupta, S. M. L. (1987). Contour sequence moments for the classification of closed planar shapes. *Pattern Recognition, 20*, Article 267272.

Hamed, A., Belal, N. A., & Mahar, K. M. (2016). Arabic sign language alphabet recognition based on HOG-PCA using microsoft kinect in complex backgrounds. In *2016 IEEE 6th international conference on advanced computing (IACC)*. IEEE.

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism, 69*, S36–S40.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory, 8*(2), 179–187.

Kerkeni, A., Benabdallah, A., Manzanera, A., & Bedoui, M. H. (2016). A coronary artery segmentation method based on multiscale analysis and region growing. *Computerized Medical Imaging and Graphics, 48*, 49–61.

Kishore, P., Prasad, M., Kumar, D. A., & Sastry, A. (2016). Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. In *2016 IEEE 6th international conference on advanced computing (IACC)*. IEEE.

Kishore, P. V. V., Kumar, D. A., Sastry, A. S. C. S., & Kumar, E. K. (2018). Motionlets matching with adaptive kernels for 3-d indian sign language recognition. *IEEE Sensors Journal, 18*(8), 3327–3337.

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding, 141*, 108–125.

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2018). Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *International Journal of Computer Vision, 126*(12), 1311–1325.

Kumar, E. K., Kishore, P. V. V., Sastry, A. S. C. S., Kumar, M. T. K., & Kumar, D. A. (2018). Training CNNs for 3-d sign language recognition with color texture coded joint angular displacement maps. *IEEE Signal Processing Letters, 25*(5), 645–649.
Kumar, P., Gauba, H., Roy, P. P., & Dogra, D. P. (2017). A multimodal framework for sensor based sign language recognition. *Neurocomputing, 259*, 21–38.
Kuroki, K., Zhou, Y., Cheng, Z., Lu, Z., Zhou, Y., & Jing, L. (2015). A remote conversation support system for deaf-mute persons based on bimanual gestures recognition using finger-worn devices. In *2015 IEEE international conference on pervasive computing and communication workshops (PerCom workshops)*. IEEE.
Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2017). Scale-aware fast r-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, 1.
Li, Q., Feng, B., Xie, L., Liang, P., Zhang, H., & Wang, T. (2016). A cross-modality learning approach for vessel segmentation in retinal images. *IEEE Transactions on Medical Imaging, 35*(1), 109–118.
Li, S.-Z., Yu, B., Wu, W., Su, S.-Z., & Ji, R.-R. (2015). Feature learning based on SAE–PCA network for human gesture recognition in RGBD images. *Neurocomputing, 151*, 565–573.
Lim, K. M., Tan, A. W., & Tan, S. C. (2016). A feature covariance matrix with serial particle filter for isolated sign language recognition. *Expert Systems with Applications, 54*, 208–218.
Macías, E. J., Roca, A. S., Fals, H. C., Fernández, J. B., & Muro, J. C. S.-D. (2013). Emisión acústica y redes neuronales para modelado y caracterización del proceso de soldadura por fricción agitación. *Revista Iberoamericana de Automática e Informática Industrial RIAI, 10*(4), 434–440.
Mapari, R. B., & Kharat, G. (2015). Real time human pose recognition using leap motion sensor. In *2015 IEEE international conference on research in computational intelligence and communication networks (ICRCICN)*. IEEE.
Mohammad Iqbal, T.L. & Endang Supriyati (2017). Sibi blue: Developing indonesiansign language recognition system based on the mobile communication platform. International Journal of Information Technology, Computer Science and Open Source.
Mohandes, M., Aliyu, S., & Deriche, M. (2015). Prototype arabic sign language recognition using multi-sensor data fusion of two leap motion controllers. In *2015 IEEE 12th international multi-conference on systems, signals & devices (SSD15)*. IEEE.
Mohanty, P., Jane, N. A., & Bonas, M. C. (2018). In *Identifying variables to define innovator group in the healthy food industry: A fuzzy approach In Lecture notes in computer science* (pp. 440–445). Springer International Publishing.
Naglot, D., & Kulkarni, M. (2016). Real time sign language recognition using the leap motion controller. In *2016 international conference on inventive computation technologies (ICICT)*. IEEE.
Ng, M., & Jordan, A. Y. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems, 14*, 841–848. MIT-Press.
Oliveira, M., Chatbri, H., Little, S., Ferstl, Y., O'Connor, N. E., & Sutherland, A. (2017). Irish sign language recognition using principal component analysis and convolutional neural networks. In *2017 international conference on digital image computing: Techniques and applications (DICTA)*. IEEE.
Portillo, E., Cabanes, I., Marcos, M., & Zubizarreta, A. (2009). Aplicación de redes neuronales en la detección de regímenes degradados en el proceso wedm. *Revista Iberoamericana de Automática e Informática Industrial RIAI, 6*(1), 39–50.
Quesada, L., López, G., & Guerrero, L. (2016). Improving deaf people accessibility and communication through automatic sign language recognition using novel technologies. In *Advances in intelligent systems and computing* (pp. 497–507). Springer International Publishing.
Raheja, J. L., Mishra, A., & Chaudhary, A. (2016). Indian sign language recognition using SVM. *Pattern Recognition and Image Analysis, 26*(2), 434–441.
Rao, G. A., & Kishore, P. (2018). Selfie video based continuous indian sign language recognition system. *Ain Shams Engineering Journal, 9*(4), 1929–1939.
Romdhane, N. B., Mliki, H., Beji, R. E., & Hammami, M. (2016). Combined 2d/3d traffic signs recognition and distance estimation. In *2016 IEEE intelligent vehicles symposium (IV)*. IEEE.
Rossomando, F. G., Soria, C., & Carelli, R. (2010). Control de robots móviles con incertidumbres dinámicas usando redes de base radial. *Revista Iberoamericana de Automática e Informática Industrial RIAI, 7*(4), 28–35.
Rumelhart, G. R., & Hinton, D. E. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.
Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Prentice Hall.
Sladoje, N., & Žinić, J. (1997). Ellipses estimation from their digitization. In *Discrete geometry for computer imagery* (pp. 187–198). Berlin Heidelberg: Springer.
Sun, C., Zhang, T., Bao, B.-K., & Xu, C. (2013). Latent support vector machine for sign language recognition with kinect. In *2013 IEEE international conference on image processing*. IEEE.
Talebpour, A., & Mahmassani, H. S. (2016). Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C: Emerging Technologies, 71*, 143–163.
Tan, M., Wang, B., Wu, Z., Wang, J., & Pan, G. (2016). Weakly supervised metric learning for traffic sign recognition in a LIDAR-equipped vehicle. *IEEE Transactions on Intelligent Transportation Systems, 17*(5), 1415–1427.
Tao, W., Leu, M. C., & Yin, Z. (2018). American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence, 76*, 202–213.
Tharwat, A., Gaber, T., Hassanien, A. E., Shahin, M. K., & Refaat, B. (2015). SIFT-based arabic sign language recognition system. In *Advances in intelligent systems and computing (359–370)*. Springer International Publishing.
Tubaiz, N., Shanableh, T., & Assaleh, K. (2015). Glove-based continuous arabic sign language recognition in user-dependent mode. *IEEE Transactions on Human-Machine Systems, 45*(4), 526–533.
Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.
Werbos, P. J. (1994). *The roots of backpropagation. From Ordered Derivatives to Neural Networks and Political Forecasting.* New York, NY: John Wiley & Sons Inc.
Zahn, C. T., & Roskies, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers, C-21(3)*, 269–281.