Gesture Phase Segmentation Dataset: An Extension for Development of Gesture Analysis Models

Raúl A. Sánchez-Ancajima[1*], Sarajane Marques Peres[2], Javier A. López-Céspedes[3], José L. Saly-Rosas-Solano[4], Ronald M. Hernández[5] and Miguel A. Saavedra-López[6]

[1*]Universidad Nacional de Tumbes, Tumbes, Perú. rsanchez@untumbes.edu.pe, https://orcid.org/0000-0003-3341-7382

[2]University of São Paulo, São Paulo, Brazil. sarajane@usp.br, https://orcid.org/0000-0003-3551-6480

[3]Universidad Nacional de Tumbes, Tumbes, Perú. jlopezce@untumbes.edu.pe, https://orcid.org/0000-0003-2560-1876

[4]Universidad Nacional de Tumbes, Tumbes, Perú. jsalyrosass@untumbes.edu.pe, https://orcid.org/0000-0001-5457-8236

[5]Universidad Continental, Lima, Perú. ronald.hernandez@outlook.com.pe, https://orcid.org/0000-0003-1263-2454

[6]Universidad Nacional de Tumbes, Tumbes, Perú. saavedralopezmiguel@gmail.com, https://orcid.org/0000-0003-4913-933X

Received: August 08, 2022; Accepted: October 02, 2022; Published: November 30, 2022

Abstract

In recent years, experts in theory of gesture have been showing some interest in automating the discovery of gesture information. Such an automation can help them in reducing the inherent subjectivity of gesture studies. Usually, to produce information for linguistic and psycholinguistic studies, the researchers analyze a video of people speaking and gesturing. This annotation task is costly and it is the goal of automation. Such videos compose the datasets that allow the development of automated models capable to carry out part of the analysis of gestures. In this paper, we present a detailed documentation about the Gesture Phase Segmentation Dataset, publicized in UCI Machine Learning Repository, and an extension of such dataset. Such dataset is especially prepared to be used in the development of models capable to carry out the segmentation of gestures in their phases. The extended dataset is composed by nine videos of three people gesturing and telling stories. The data was captured with Microsoft Kinect Sensor and they are represented by spatial coordinates and temporal information (velocity and acceleration). The data are labeled following four phase of gesture (preparation, stroke, hold and retraction) and rest positions.

Keywords: Dataset, Gesture Studies, Phases of Gesture, Gesture Segmentation.

Journal of Interner Services and Information Security (JISIS), volume: 12, number: 4 (November), pp. 139-155
DOI: *10.58346/JISIS.2022.I4.010*

*Corresponding author: Universidad Nacional de Tumbes, Tumbes, Perú.

1 Introduction

He integration of heterogeneous data in different formats and from different communities requires a better understanding of the concept of a dataset, and of the key related concepts such as format, coding and version. The concept of a dataset is common in almost all scientific disciplines where data provide the empirical basis for research activities. Four basic characteristics can be identified as common to most definitions grouping, content, relationship, and purpose [1].

Research on gesture analysis is primarily focused on the exploration of gestures as a new way to provide systems with more natural methods for interaction. Within this scope, there are several goals to be achieved, such as the development of systems that allow interaction of simple gestures, and the development of systems that are able to recognize a predefined domain within some sign language, in order to provide a more natural interaction of deaf people, e.g., building a communication bridge between deaf people and people who are not able to understand sign language. Outside this domain, automated gesture analysis can also help to build multimodal tools for specific and specialized use, such as psycholinguistic analysis. Efforts in this direction point to the development of automated methods for feature extraction from a video stream. The different applications for automated gesture analysis have several types of analysis that can be performed: there are studies focusing on gesture recognition, others focusing on continuous gesture segmentation (i.e., segmentation of the gesture stream into gestural units), and even studies focusing on automated extraction of some gesture feature for future application of psycholinguistic analysis. On the other hand, there are other essential aspects to take into account, such as strategies for data acquisition and representation, as well as for evaluating the results of each type of analysis. Therefore, it can be observed that there is much research interest in gesture analysis.

There is a growing number of researches in the analysis of gestures whose objective is to support the interpretation of the role of gestures in the constitution of the discourse, thus allowing the establishment of relationships that improve the understanding of what is being spoken. Such research is based on the area of "gesture studies", an interdisciplinary area that aims to study the use of hands and other parts of the body for communicative purposes, combining knowledge from linguistics, psychology, social sciences and other areas [2].

The study of gestures is an interdisciplinary area that aims to analyze the use of hands and other body parts for communication. The study of gesticulation, i.e., the study of gestures that accompany speech [3] is an important topic that has been studied by researchers from various areas. The usual way to study gesticulation is to perform gesture analysis on recorded videos of people talking and gesturing [4].

Researchers in the area of gesture studies usually record videos of people speaking or gesturing in sign language, telling stories, in speech or in conversation and, from the recordings, analyze the gestures made by them. Such analysis requires a transcription of components of the videos, such as speech, tone of voice and gestures. One of the stages consists in the segmentation of gestures according to the basic phases: preparation, stroke, hold and retraction [5]. Such segmentation is usually performed manually by researchers, consisting of a high-cost process in relation to the time of effort. Thus, the interest in automated processes is gaining strength because it is of great help to such researchers.

Recent work has employed pattern recognition techniques and has focused on developing systems with various more natural user interfaces. These systems can use gestures for control interfaces or recognize sign language gestures, which can provide multimodal interaction systems; they consist of multimodal tools to help psycholinguists understand new aspects of speech analysis and to automate laborious tasks. Gestures are characterized by several aspects, mainly by movements and sequence of postures [6].

In this paper we describe the dataset for the gesture analysis research area, we worked with nine videos of three people telling three different stories, i.e. nine data templates were elaborated and obtained by a Microsoft Kinect SDK equipment, the main objective is to describe the dataset from its collection, preparation and analysis to the behavior and visualization of the data in each gestural phase.

2 Gesture Theory

The study of gestures is an interdisciplinary area that studies the use of the hands and other body parts for communication. When gestures are performed together with speech, they are called gestures [3]. Studies and analysis on gestural language are performed from videos of people talking and gesturing, therefore, this analysis requires a pre-processing that involves the segmentation of gesture in phases.

People often make one or more movements with their hands, arms, or even their bodies during a natural conversation or when giving a speech. According to Kendon [5] and McNeill [3], an excursion, particularly with respect to the hands, refers to a movement from a rest position to some region in space, and then bringing back to the same or another rest position. While the hands are away from the rest position, the movement is called a unit gesture. When these gestures co-occur with speech, they are called gesticulation. Furthermore, according to Kendon [5] and McNeill [3], a gesture unit can consist of one or more gestural phrases, which can be divided into phases: preparation, stroke, hold and retraction. The stroke phase defines the main movement in a gesture unit and has a semantic meaning; holds are pauses during the phrase, in which the hand configuration used in the stroke is maintained; preparation and retraction are transitional phases between gesture units and rest positions.

In light of this, when considering automated segmentation and unit gesture, there are two special issues that need to be addressed:

- Boundaries between transition phases and resting positions: in fact, in the analysis of human gestures there is no precise boundary between phases, since in real situations involving gesticulation, these boundaries are not evident. Moreover, the differences shown by human coders are closely related to this fact.
- The similarity between maintain (keep still) and rest position: the two phases are characterized by the hands in a fixed configuration with almost total absence of movements. However, rest positions have no semantic content, while the interpretation of hold takes into account their meaning. Thus, there is a problem, in an automated analysis based on a type of gesture representation that is devoid of semantic information, framing in a "stop segment" and frame a sequence of rest position that may be too similar to allow them to be correctly recognized.

Eisenstein [7], states that an analysis of gesticulation can help in the analysis of speech or voice. The dataset considers videos of people describing the operation of mechanical devices, considering 1137 nominal syntagms. Martell and J. Kroll [8] consider as a dataset a 22-minute video of a teacher giving a lecture. Wilson et. al [9] consider as a dataset two videos, with approximately four minutes each, in which the participants must recount an experience in which they were in danger. Madeo [6] studies the segmentation of gesture phases in natural gesture situations, i.e., situations in which gestures are produced in conjunction with speech, and to obtain data he opts for the "storytelling" situation; people read the story presented to them in small frames and in front of them there is a data capture device called Microsoft Kinect.

3 Gesture Phases

The segmentation of gesture phases considers two important concepts [3].

1) Gestural Unit (G-unit): The period of time between two rest positions, i.e. a position in which there is no execution of a gesture.
2) Gestural Phrase (G-phrase): The gestural phrase occurs within a gestural unit and is composed of one or more gesture phases.

Therefore, Kendon [3] and McNeill [5] define five phases of gesture and these can be organized in a hierarchy as in Fig. 1:

1) Preparation: Occurs when the hands move to the position in which the gesture is executed.
2) Hold pre-stroke: Represents a pause after the preparation, maintaining the position and configuration of the hands.
3) Stroke: Where the movement that expresses the semantics of the gesture occurs.
4) Hold post-stroke: Represents a pause after the stroke, maintaining the position and configuration of the hands.
5) Retraction: In that in which the return of the hands to the resting position occurs.

The definition of gesture phases to establish an expressive phase, can be composed by a stroke, also preceded and succeeded by dependent holds [10], as in the original definition of Kendon [3] and McNeill [5], or by an independent hold, which occurs when the semantic content of gesture is expressed by a pause. In this paper we are considering the use of independent hold and do not consider dependent holds. The gesture phases are exemplified in a simplified form in Fig. 2.

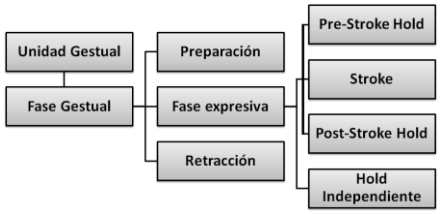


Figure 1: Hierarchy of Gesture Phases [10], Comprising Gesture Unit, Gesture Phase and Gesture Phases [6].

Figure 2: Gestures Accompanying a Speech: The First and the Last Frames Belong to the Rest Position, the Others Belong to the Gestural Unit [11].

4 Definition of the Gesture Phase Segmentation Problem

You have a video represented by a sequence of frames of size $S = \{f_1, f_2, \ldots, f_n\}$ of size nThe video is introduced in a segmentation strategy in order to identify the phases of the gesture. The *frame* sequence consists, physically, of a sequence of static images in RGB. Particularly in this work the *frame* sequence

contains images of a person telling a story. The segmentation problem consists of receiving the representation of a frame as input and between the f_i as input and between the classification of these *frames* as a *class* $c_i = \{D, P, S, H, R\}$corresponding to Rest, Preparation, Stroke, Hold and Retraction. This classification problem is divided into smaller subproblems:

1) Classification of the rest position: entry $f_i \in S$and exit $c_i = \{E, G\}$ where $G \supset \{P, S, H, R\}$, corresponding to the gestural unit.
2) Holds classification: input $f_i \in S_G$where $S_G \supset G$and output $c_i = \{H, D\}$ where $D \supset \{P, S, R\}$, corresponding to the dynamic phases.
3) Strokes Rating: input $f_i \in S_M$ where $S_M \supset M$, and output $c_i = \{S, T\}$ where $T \supset \{P, R\}$, corresponding to the transition phase.
4) Classification Preparation and Shrinkage: input $f_i \in S_T$where $S_T \supset T$and output $c_i = \{P, R\}$.

Fig. 3 illustrates this strategy, which is proposed by Madeo [12] to divide the classification problem into subproblems.

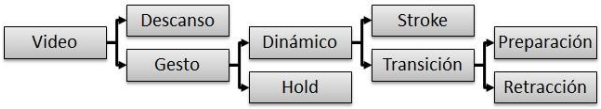


Figure 3: Strategy for Classification of Gesture Phases [12].

5 Set of Dice

In the area of computational learning, usually, every element or observation belonging to a data set is a point *in* R^D. That is, a dataset is a sample, whose $\{x_i\}_{i=1}^N \subset R^D$whose $n \in N$ elements follow some probability law, in general unknown. In general, there is a lot of literature concerning various computational learning tasks with such a data set. These data sets are of the form:

$$T = \{S_i\}_{i=1}^N \quad (1)$$

Where N is the number of observations and each observation $S_i$is a set of points [13].

Thus a set of data Xcan be defined as:

$$X = \begin{matrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{matrix} = \begin{matrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2.1} & x_{2.2} & \cdots & x_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{matrix} \quad (2)$$

Where x_j is a vector of p coordinates and n is a number of elements of the data set considering that each vector represents a data of that set and each coordinate of that vector represents a descriptive attribute of the data. The data set X resides in the space R^p, and this space is referred to by data analysis algorithms as "data space", "input space" or "vector space" [14].

For this work we have three users, which we will call user A, user B and user C; and we also have three different stories and we will denote them as H1, H2, H3; then each of these stories is gesticulated or told by each user, i.e. user A tells the story H1, story H2 and story H3; user B and C do the same and so we have nine videos, then each user labels the nine videos, that is to say that for each video there are three labelers, then a fourth encoder is included which also labels the nine videos so that each video now has four labelings, therefore this allows comparisons between the four labelings of each video, and obtain a fifth labeling which is obtained by majority, that is to say that the common labeling is chosen in each

of the *frames* that make up the videos. After that, the concordance coefficients and divergence coefficient are obtained for each pair of encoders (Table III).

6 Kinect and the Behavior of its Data

A. Kinect

It is a depth camera, based on an infrared emitter and a camera (Fig. 4). Normal camera produce images in which each pixel records the color of light (RGB) bouncing off objects. Kinect, on the other hand, records the distance of objects in the scene, creating a depth image. Infrared light is used for this purpose, which does not capture the appearance of the objects but their position in the scene. This depth image is displayed in black and white with some distortion. The lightest parts are the closest and the darkest parts are the furthest away. By using Processing (or other similar development environments) it is possible to create programs that give us the distance of users even in motion.

Figure 4: Kinect Without its Plastic Housing, Showing from Left to Right the Infrared (IR) Emitter, the RGB Camera and the IR Camera [15].

The coordinates are in the "real world", so the Kinect uses a Euclidean coordinate system that is in three dimensions to map the space that its sensors "can see". Its positive axes are such as shown in Fig. 5, the z coordinate represents the distance from the sensor [16].

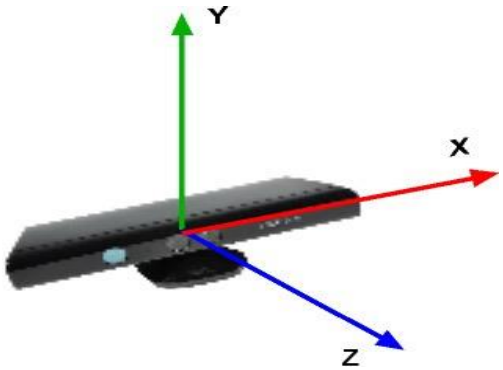


Figure 5: Kinect Euclidean Coordinate System [16].

B. Data Stability and Instability

For each video frame, the three-dimensional coordinates of each articulation point of interest and the image associated with that *frame* are stored, but due to a limitation of the device, it is not possible to simultaneously record the data and the image, therefore, there is a discrepancy of milliseconds to hundredths of a second between the image and the corresponding data [6].

In Fig. 6, we show the behavior of the data obtained from the movement of the left hand of user A when telling *story 1*, this only for a part of the video (900 *frames*), the black colored l´ınea defines the phases: Rest = 1, Preparation = 2, Stroke = 3, Hold = 4, and Retraction = 5; the red colored line is the coordinate z which represents the distance between the user and the kinect sensor, there is not much variation because in all cases users A, B and C had no displacement, the green colored line is the coordinate y which represents the height of the articulation point of interest (left hand in this case) with respect to the position of the kinect sensor and finally the blue line represents the horizontal coordinate with respect to the kinect sensor. x which is the horizontal with respect to the sensor.

Now it can be observed that there is enough variation for the stroke phases, a little less for the phases of preparation, retraction and hold, finally very little variation for the rest phase, this behavior is observed in all the stretch of the video which has 1742 *frames*, here is only shown a part corresponding to 900 *frames.*

This gives us an idea of the behavior of the data, an idea of the stability and instability of the data obtained from the kinect, however it cannot be stated reliably that the data set is stable or unstable, it is necessary to perform other studies, such as genetic algorithms, where rules can be established in each phase that would be defined by intervals.

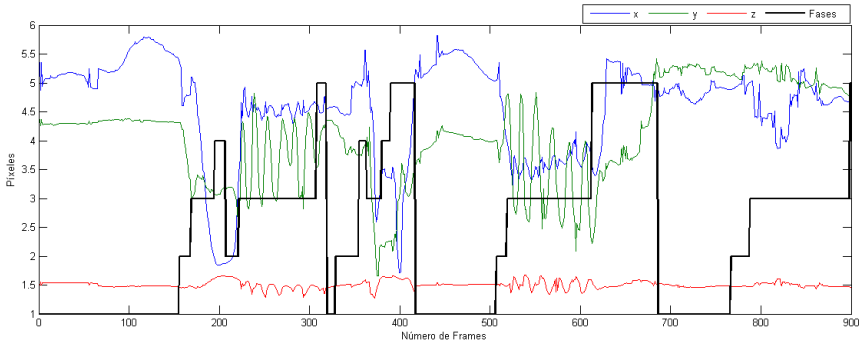


Figure 6: Left Hand Movement of User A in Part of Video 1.

7 Descriptive Analysis

A. Description of the Data Set

To study the segmentation of gestural units in natural gesturing situations, it is necessary to obtain a dataset composed of videos of people talking and gesturing. The videos used in this work refer to the context of storytelling, in which a user reads stories in small frames and then tells the story in front of a Microsoft Kinect device. Three different stories told by the same user in two data capture sessions were considered: A1 and A2 refers to videos from the first session, A3 refers to a video captured in the second session.

The Microsoft Kinect allows to obtain:

- The three-dimensional coordinates x, y, z of the articulation points chosen to compose the representation of the gesture (hand, pulses, head and trunk) associated with a time stamp.
- The RGB image corresponding to each frame, associated to each set of coordinates by time stamp.

Thus, for each *frame*, the three-dimensional coordinates of the articulation points and the image associated with that *frame* are stored. Table I presents the information on the fifth labeling of the videos obtained by majority, including the total number of *frames* in the video (n-Frames), the number of rest *frames* (n-Rest), the number of *frames* corresponding to the gestures (n-Gesture) specifying also the number of *frames* of each Phase - Preparation (P), Stroke (S), Hold (H) and Retraction (R).

Table I: Description of the Labeling of the Videos that make up the Dataset, Obtained by a Majority Vote by Comparing Four Previous Labels

User	**History**	**#frms**	**#desc**	**#gestures**			
				P	**S**	**H**	**R**
A	1	1747	706		686	43	
		1264	468	189	452		99
		1834	599	211	649		241
B	1	1114		368	402		
		999		162	345	99	
		1424		421	532		182
C	1	1112	241		413	145	146
		1082	429	201		123	201
		1448	358	184	607	101	198

B. Data Labeling

In the case of gesture unit segmentation, the phases are manually attributed to each gesture unit, i.e. to each *frame* of the video. For this, an encoder examines the *frames*, attributing a label to each one. As it is a subjective process, in this work, as mentioned above, the videos were labeled by four coders. With the four labelings it is possible to obtain a fifth labeling which is obtained by majority for each label. Research using manually labeled data needs to ensure that these data are reliable. It is possible to measure the reliability of the labeling through a concordance analysis obtained from the attribution of labels by different coders [17].

C. Concordance Coefficient Krippendorf's Alpha

The data we collect in a research project generally reflect our understanding of the subject under investigation and reflect our interpretation of the phenomena. However, it is common in some cases to wonder about the quality of the data set obtained. This and other questions can be detrimental to the integrity of scientific investigations if they are not resolved at the appropriate stage.

Krippendorff's alpha coefficient is a statistical measure that quantifies the agreement between two or more coders, and is regularly used by researchers in the area of content analysis. Since the 1970s, it has been used in content analysis on textual units [18]. Concordance expresses "the extent to which two or more coding u observers agree with each other", [19] and it should also be clear that concordance is different from correlation [20].

The calculation of Krippendorf's alpha coefficient can be obtained with statistical software and its formula is as follows:

$$\alpha = \frac{D_o}{D_e} = \frac{rm-1}{m-1}\frac{\Sigma_i\Sigma_b\Sigma_{i>b}n_{bi}n_{ci}}{\Sigma_b\Sigma_{i>b}n_bn_c} \quad (3)$$

Where:

- D_oDiscrepancies observed.
- D_e Expected discrepancies.
- r number of total configurations.
- m total number of observations made for each record.
- $\Sigma_i\Sigma_b\Sigma_{i>b}n_{bi}n_{ci}$ sum of products of the discrepancies observed between the coders or observers.
- $\Sigma_b\Sigma_{i>b}n_bn_c$ The sum of the products of the total frequencies of occurrence of each configuration [19].

To quantify the degree of agreement between different coders, it is possible to use a concordance coefficient. For the present work, Krippendorff's Alpha coefficient ($K\alpha$). The choice of this coefficient is because it efficiently minimizes coder bias [21]. The values of these coefficients range from -1 y 1knowing that negative values indicate insufficient data or random coding (Table II).

Thus, the concordance between the encoders that produced four labelings was analyzed for each of the videos, opting to perform the pairwise analysis to visualize the best labeling and then obtain a fifth labeling by majority for further studies. Table III shows the coefficients of concordance. $K\alpha$.

Table II: Interpretation of the Krippendorf's Alpha Concordance Coefficient Values [21].

Concordance Krippendorff's ($K\alpha$)	**Strength of concordance**
0-0.2	Weak
0.2-0.4	Reasonable
0.4-0.6	Moderate
0.6-0.8	Considerable
0.8-1	Perfect

8 Data Visualization

With the Kinect, the three-dimensional coordinates of the points of the joints chosen to track the movements made by a person are obtained. x, y, z of the joint points chosen to track the movements made by a person, the coordinates are organized in a.csv file extension where each line corresponds to the coordinates of the points obtained in a capture time unit and in a time stamp that shows when it was captured. Then, for each video *frame*, the image associated to each *frame* and the three-dimensional coordinates of each articulation point of interest are stored, in this case right and left hand, right and left pulse, head and forehead of the person (at the height of the spine).

From the frame images and timestamps, it is possible to manually attribute the class of each data, i.e. to each frame. To execute this classification process, an analyst (in this case users A, B and C) examines the images in sequence, identifying the class of each *frame*. For each examined image, the analyst attributes a class in the data line corresponding to the timestamps of the image in the file containing the coordinates of each articulation point.

Table III: Concordance Coefficients and Percentage of Divergence

Coders	Videos								
	History 1			History 2			History 3		
	A1	B1	C1	A2	B2	C2	A3	B3	C3
1 y 2	0.85	0.47	0.65	0.82	0.65	0.71	0.82	0.52	0.73
	10.07	38.24	26.89	12.97	26.93	21.63	13.47	35.81	20.72
1 y 3	0.88	0.49	0.64	0.77	0.67	0.72	0.76	0.51	0.69
	11.10	38.24	27.43	16.22	25.63	20.52	17.39	37.22	23.62
1 y 4	0.82	0.51	0.81	0.63	0.23	0.63	0.73	0.54	0.85
	12.36	34.74	14.30	26.98	57.26	27.63	20.56	33.36	10.57
2 y 3	0.83	0.52	0.73	0.81	0.67	0.73	0.77	0.52	0.80
	11.10	36.63	20.86	12.5	25.93	19.41	16.96	37.5	15.74
2 y 4	0.82	0.72	0.63	0.67	0.20	0.59	0.73	0.65	0.67
	12.42	20.20	28.51	23.73	58.75	30.96	20.07	26.26	24.56
3 y 4	0.78	0.55	0.63	0.62	0.24	0.64	0.71	0.49	0.64
	14.94	33.21	28.60	12.97	55.46	26.16	21.76	38.34	26.93

In the present work the points of articulation that are used to compose the representation of the gestures are:

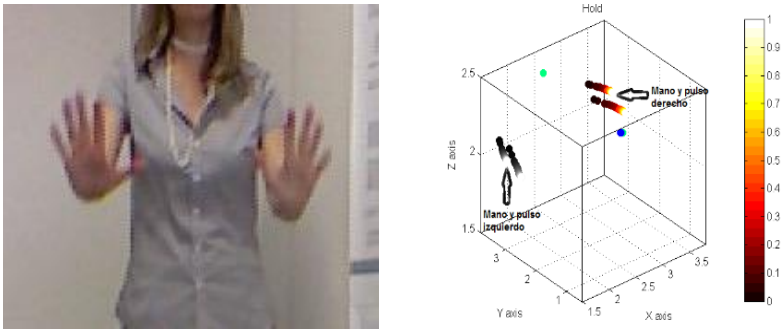
- Hands: The articulation points corresponding to the left and right hands provide the spatial location of the hands during gesture execution.
- Pulses: The position of the pulses can provide information about internal movements of each hand through the relationship of their position to the corresponding hand articulation point.
- Head: The position of the head can be used as a reference point to analyze the speed and distance of the hands in relation to the head.
- Trunk: The articulation point corresponding to the trunk can be used as an alternative reference point to the head point, since the head can execute gestures not linked to the position of the hands, while the trunk, on the other hand, normally remains stable in relation to the gesture.

According to what has been established above we can explain the behavior of the data geometrically, for some gesture phases.

Well, we must say that the colors that were assigned for the left hand and pulse is gray; the color for the trunk (spine) and head is a scale of green and blue (cold colors) the color for the right hand and pulse is a scale of yellow and orange with black as the lowest value and white as the highest value (warm colors).

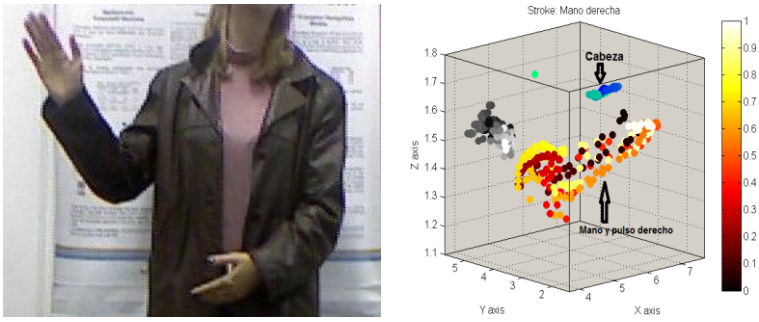
Then in Fig. 7 it is observed that the points of gray color and color scale of yellows and oranges are distributed with similar and small intensity of color, which indicates that the hands are holding their position briefly, that is, they pause, which is a Hold phase, except for the last points where there is certainly a transition stage towards another phase, which could possibly be a Stroke. With respect to the head and the front maintain their position since the points maintain their color which is blue and green respectively, that is to say there is no movement.

In Fig. 8 it can be observed that the yellow scale points (right hand) vary clearly, which indicates that there is well defined movement in the right hand, something that does not occur with the gray points that remain together and almost the same color, indicating that the movement for the left hand is less. The color for the points of the trunk (spine) is maintained and for the head some bluer points appear, which means that there is a slight movement in the head accompanying the movement of the right hand.



(a) Hold Phase Frame. (b) Hold Phase Points.

Figure 7: Hold Phase Description.

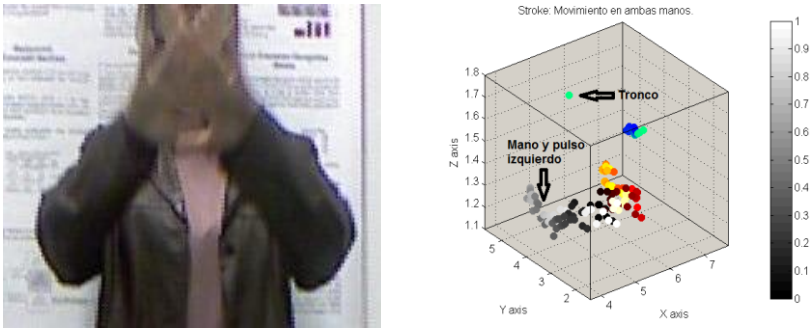


(a) Stroke Phase *Frame*. (b) Stroke Phase Points.

Figure 8: Stroke Phase Description Right Hand Movement.

In Fig. 9, it can be observed that the distribution of points both in their location and in their color scale is similar in intensity, those of gray color are of the left hand and pulse, those of yellow and orange color are of the right hand and pulse, which indicates that they are in movement and this defines a stroke, accompanied by a slight movement in the head, while the point of the trunk remains in its location.

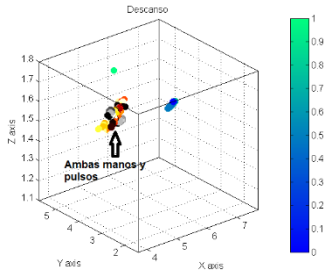
In Fig. 10 it can be seen that the points of the hands, pulses are kept together, the points of the trunk and head are kept in position, indicating that there is no movement and this is the resting phase.



(a) Stroke Frame. (b) Stroke Points.

Figure 9: Stroke Phase Description Stroke Movement in Both Hands.

(a) Rest Phase Frame.



(b) Rest Phase Points.

Figure 10: Rest Phase Description.

9 Velocity and Acceleration

A. Theoretical Basis

We know that a scalar quantity is one that does not need a direction to be understood (mass, height, temperature, etc.) whereas a vector quantity needs a direction to know where it is going (velocity, acceleration, weight, force, etc.). In addition, the displacement of an object can be described by a vector, and for this we need a coordinate system, which is the three-dimensional space that has the axes x, y, z.

Kinematics is a branch of physics that studies the laws of motion of bodies without attending to the causes that provoke it. Therefore, kinematics only studies the motion itself, unlike dynamics, which studies the interactions that produce it. The concepts with which kinematics works are several: point that refers to an element without volume located in space; reference system is that coordinate system with respect to which the position of the points and time is given; position or location with respect to a reference system; and time [15].

B. Data Representation

A suitable way to represent the videos to be processed by the classification algorithm is to use the position information $3D$ to create a normalized vector representation: for each frame, the position of the hands and pulses is subtracted from the position of the front of the subject, and this new 3-dimensional position is divided by the distance between the head and the front of the subject. From the normalized vector representation, new information is created by estimating the velocity and acceleration measurements. For velocity, the estimate is given by:

$$v_{i,i-d} = \frac{\Delta_{r_{i,i-d}}}{t_i - t_{i-d}} \quad (4)$$

Where t is the time stamp of the frame i, d *is the frame* offset and $\Delta_{r_{i,i-d}}$ is the Euclidean distance between the normalized position of the $3D$ normalized position of the point of interest in the frame i y $i - d$ the frame. The acceleration is calculated through:

$$a_{i,i-1} = \frac{v_i - v_{i-1}}{t_i - t_{i-1}} \quad (5)$$

Furthermore, the proposed approach considers a windowing strategy, using past and/or future *frame* information to represent each *frame* of interest due to the intrinsic temporal aspects of the gesture phase segmentation problem [6].

Then, from the above it is understood that in order to represent the gestures, it is possible to consider the movements made in the narration of the stories, represented by the speed and acceleration assumed by each point of articulation. Thus, the positions of the head and trunk are considered to perform a normalization of the positions of hands and pulses, by means of:

$$p_{norm} = \frac{p - p_t}{||p_c - p_t||} \quad (6)$$

Where p is the position of an articulation point, p_{norm} is the normalized position of the articulation point, and p, p_t is the position of the trunk and p_c is the position of the head.

From the normalized points, velocity and acceleration information are calculated to represent the motion of hands and pulses. The vector velocity is given by the following equation:

$$v_{vet} = \frac{p_i - p_{i-d}}{t_i - t_{i-d}} \quad (7)$$

Where p_i is the vector representing the position of a hinge point considering the axes x, y y z axes in the i-the last frame of the video, t_i corresponds to the time (in hundredths of seconds) in which the -th frame of the video i-nth *frame* of the video, and the d is the displacement in $frames$ used to calculate the velocity.

From the vector velocity, the vector acceleration is obtained, which consists of:

$$a_{vet} = \frac{v_i - v_{i-1}}{t_i - t_{i-1}} \quad (8)$$

In addition, scalar measurements can also be obtained from vector measurements, in this case the scalar measurement for both velocity and acceleration is given by:

$$v_{esc} = \sum_{i=1}^{N} v_i^2 \quad (9)$$

Where v is the vector measurement under consideration, i.e., velocity or acceleration; each of these measurements can be interpreted as a signal: a sequence of right-hand scalar velocity measurements for each *frame* forms the signal corresponding to the right-hand scalar velocity of that sequence of *frames* [6].

The following Figures represent some velocity and acceleration signals. In Fig. 11, that the blue line represents the scalar velocity and the red line represents the gesture phases, in this case we considered the values: Rest = 0.01, Preparation = 0.02, Stroke = 0.03, Hold = 0.04, Retraction = 0.05; this with the objective of graphing and analyzing the behavior of the speed during the narration of the *story1* by user B, then it can be observed that the speed for the rest and hold phases is equal, while for the stroke phases the speed is very variable, while for the preparation and retraction phases the speed increases or decreases moderately.

In Fig. 12, it is observed that the acceleration values are much smaller than those of the velocity, they vary from zero to 0.012; the behavior of the acceleration is similar to that of the velocity, since for the rest and hold phases it is almost zero, and for the stroke phases it increases considerably while in the preparation and retraction phases it varies moderately.

C. Vector Velocity and Acceleration

As mentioned in the previous section and defined the vector velocity and acceleration in equations (7) and (8) respectively, we now analyze the behavior of the data obtained from the kinect for the mentioned variables. The program $movimiento.m$ made in matlab R2012b, allows us to visualize the variation of

all the points in R^3 of the velocity and vector acceleration during the whole video, observing also the phase in which the movement occurs, for this section we worked with the data of hand movements: right and left; this for both variables.

We must say that the negative values that we observe in Figures 13, 14, 15 and 16, represent the directed distance yi in the direction of the directional vector $\vec{j} = (0, -1, 0)$ since the coordinate y observed in the mentioned figures is of the vector velocity and vector acceleration vectors.

It is decided to evaluate the coordinate y of the vector velocity and vector acceleration vectors with respect to the gesture phases, because it represents the height corresponding to the position of the kinect sensor, as mentioned in the kinect section the z-axis represents the distance of the user from its position to the sensor position, because here there is no variation since in all the cases of counting stories the users did not move at any time, this can also be observed in Fig. 6, the coordinate z is represented by the red line.

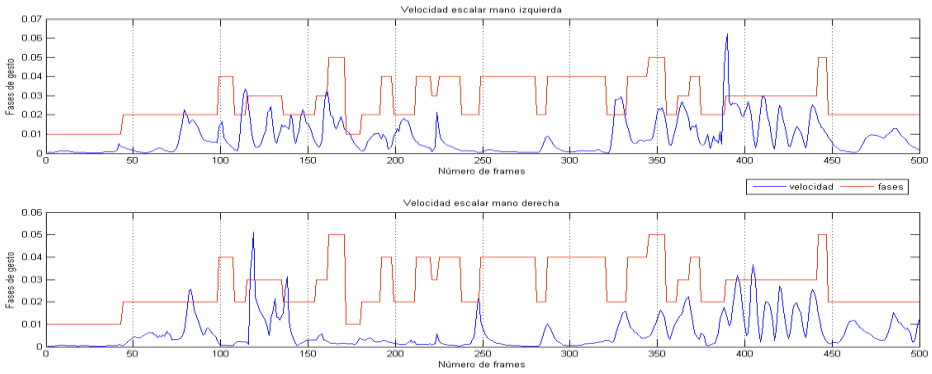


Figure 11: Scalar Velocity of Left Hand and Right Hand in a Section of Video B1.

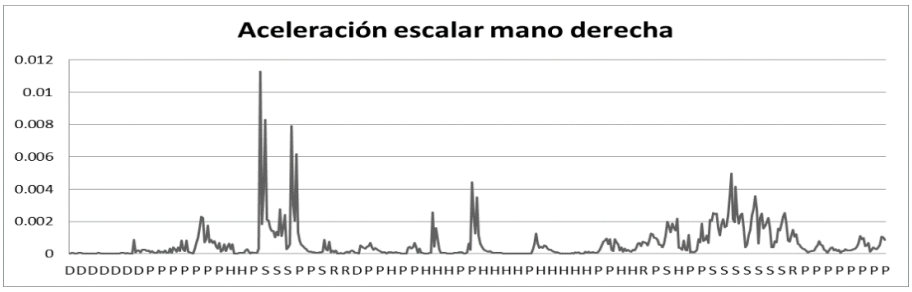
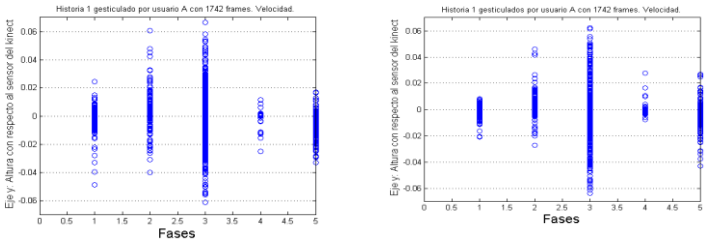


Figure 12: Right Hand Scalar Acceleration in a Stretch of Video B1.

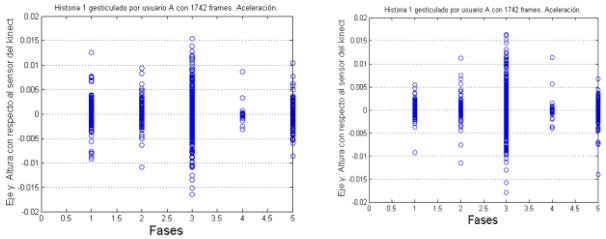
In Fig. 13 we have the representation of the coordinate $y$height of the right hand position (a) and the left hand position (b) at the i-th *frame* of the video with respect to the location of the kinect sensor; on the horizontal axis we have the phases, recall that for graphing purposes we labeled´ as follows: Rest = 1, Preparation = 2, Stroke = 3, Hold = 4, Retraction = 5. In this case we consider the values of the coordinate y of the velocity vector describing the movements of user A's hands when telling the *story1*. Figure (a) and figure (b) show that there is a high variation in the stroke phase, since movement is described in both hands, there is also coherence for the data of the rest and hold phases, since the variation is lower than those of the other phases, and for the retraction phase there is a moderate variation together with the hold phase, meaning that there is a transition stage between them.



(a) Right Hand. (b) Left Hand.

Figure 13: Coordinate y of the Velocity Vector Describing the Movement of User A's Hands

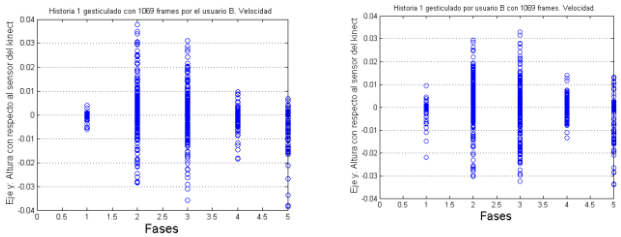
In Fig. 14, we have in analogous form the values of the coordinate and the acceleration vector, it can be observed that there is a difference between the maximum and minimum values that describe the movement of the acceleration of the hands, because for Fig. 13 it varies from 0.07 to 0.07 while for the acceleration it varies between 0.02 and 0.02, however the behavior for each one of the phases is similar to the behavior of the velocity of Fig. 13. $movimiento.m$ from the first *frame* to the last *frame*.



(a) Right Hand. (b) Left Hand.

Figure 14: Coordinate y of the Acceleration Vector Describing the Motion of User A's Hands

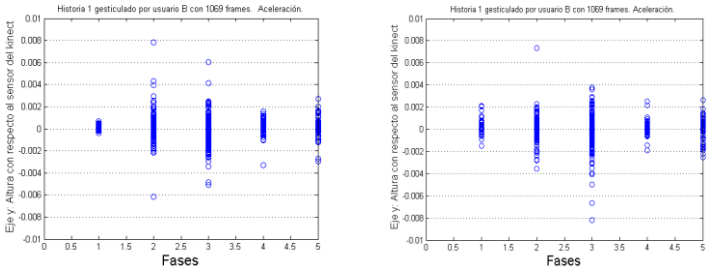
In Fig. 15 and 16, the variation of the coordinate of the velocity and acceleration vectors respectively, which describe the movement of the hands of user B when telling Story1, is presented. y of the velocity and acceleration vectors respectively, which describe the hand movement of user B when telling the *Story1*, there is a difference of *frames* between users A and B, even being the same story, this because they used different times when telling the *Story1*, however it is observed that for the preparation and stroke phases the behavior of the height of the hands with respect to the position of the kinect sensor is similar. The rest phase for the right hand is well defined; in the case of the left hand the rest and hold phase have most of their values between the interval 0.01 to 0.01, which confirms minimal movement. The behavior of the retraction phase for both hands is very similar as well as the stroke phase.



(a) Right Hand. (b) Left Hand.

Figure 15: Coordinate y of the Velocity Vector Describing the Movement of User B's Hands

In the case of Fig. 16, the behavior of the coordinate of the acceleration vector is represented. y of the acceleration vector, we see that its variation interval is a little less than 0.01 to 0.01, similarity is observed for the phases of preparation and stroke, similarity for the case of retraction and hold, while for the rest phase of the right hand (a) the variation is minimal which is consistent with the velocity of Fig. 15(a).



(a) Right Hand. (b) Left Hand

Figure 16: Coordinate y of the Acceleration Vector Describing the Motion of User B's Hands

10 Conclusions

The objective of the work is to describe the data set for gesture analysis from its collection, preparation and analysis to the behavior and visualization of the same in each gesture phase; to meet the objective of the work, we have nine videos of three users telling stories, which were recorded by a Microsoft Kinect equipment, which gives us numerical data of the points of articulation of the joints of interest of each encoder such as hands, pulses, head and trunk, this is in three-dimensional space. R^3These data are then analyzed geometrically as shown in Figures 7, 8, 9 and 10.

Figures 13, 14, 15 and 16 also show the behavior of the coordinate of the velocity and acceleration vectors. y_i of the velocity and acceleration vectors is also shown in Figures 13, 14, 15 and 16, since a program called $movimiento.m$ This is one of the very important contributions since each data sheet has an average of 1500 rows by 32 columns, including the time and phase label column.

It is important to highlight that the dataset composed of nine videos on the storytelling context, now with five different labelings for each video will alloẃ addressing the phase segmentation problem to validate results on the mentioned context.

Likewise, we have the result of the encoders with the concordance coefficient and divergence coefficient, which allows quantifying the coincidences and differences of the encoders at the moment of labeling the videos, being the phases of preparation, hold and retraction where there is more subjectivity, it can be proposed as a future work to argue other characteristics that could be important to determine the trajectory of the hands during the movement to differentiate the phases of preparation and retraction, for example the angular velocity of the hands in relation to the pulses can be useful to identify internal movements of the hands, helping to identify strokes and the end of the preparation.

Another important point is to try to study each articulation point individually, i.e. for example one hand can be at rest while the other is executing a gesture, therefore it would be very important to create strategies to study the phases of each hand individually.

The context of the work was storytelling, well it would be interesting to apply other contexts such as interviews and debates, consider a larger number of users in different sessions for each of them, include

mood changes among other situations and then confirm if the strategy in fact needs to be applied to each video individually or if it is possible to create models for a single user or for a single context being these independent, considering a larger data set.

Finally, the use of the Microsoft Kinect device to obtain data could be a limitation, because from the point of view of the area of gesture studies, it would be interesting to be able to perform an analysis of the videos through the vision of a computer, extracting data from existing videos.

References

[1] Renear, A.H., Sacchi, S., & Wickett, K.M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1-4.

[2] Kendon, A. (1996). An agenda for gesture studies. *Semiotic review of books*, *7*(3), 8-12.

[3] D. McNeill, *Hand and mind: What the hands reveal about thought*. IL, USA: Univ. of Chicago Press, 1992.

[4] Wagner, P.K., Peres, S.M., Madeo, R.C.B., de Moraes Lima, C.A., & de Almeida Freitas, F. (2014). Gesture unit segmentation using spatial-temporal information and machine learning. *In The Twenty-Seventh International Flairs Conference*.

[5] Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, *25*(1980), 207-227.

[6] Madeo, R.C.B. (2013). *Support Vector Machines and Gesture Analysis: incorporating temporal aspects* (Doctoral dissertation, PhD thesis, Universidade de Sao Paulo).

[7] Eisenstein, J. (2008). Gesture in automatic discourse processing.

[8] Martell, C., & Kroll, J. (2007). Corpus-based gesture analysis: an extension of the form dataset for the automatic detection of phases in a gesture. *International Journal of Semantic Computing*, *1*(04), 521-536.

[9] Wilson, A.D., Bobick, A.F., & Cassell, J. (1996). Recovering the temporal structure of natural gesture. *In IEEE Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 66-71.

[10] Kita, S., Gijn, I.V., & Hulst, H.V.D. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. *In International Gesture Workshop*, Springer, Berlin, Heidelberg, 23-35.

[11] Wagner, P.K., Madeo, R.C., Peres, S.M., & Lima, C.A. (2013). Segmentation of Gestural Units with Multilayer Perceptrons.

[12] Madeo, R.C., Lima, C.A., & Peres, S.M. (2013). Gesture unit segmentation using support vector machines: segmenting gestures from rest positions. *In Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 46-52.

[13] Guevara Díaz, J.L. (2015). *Supervised learning models using kernel methods, fuzzy sets and probability measures* (Doctoral dissertation, Universidade de São Paulo).

[14] Peres, S.M., Rocha, T., Bíscaro, H.H., Madeo, R.C.B., & Boscarioli, C. (2012). Tutorial on Fuzzy-c-Means and Fuzzy Learning Vector Quantization: hybrid approaches to clustering and classification tasks. *Revista de Informática Teórica e Aplicada*, *19*(1), 120-163.

[15] Ramos Gutiérrez, D. (2013). Kinematic study of the human body using Kinect.

[16] Wildman, W., Programming for kinect 4- kinect app with skeleton tracking [on-line], 2013.

[17] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

[18] Gwet, K.L. (2011). On the Krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October*, *2*(2011).

[19] Olalla, M.D.G. (2003). *Construction of joint activity and transfer of control in a parent-child interactive game situation* (Doctoral dissertation, Universitat Rovira i Virgili).

[20] Portillo, J.D. (2011). *Practical guide for the biostatistics course applied to health sciences*. Instituto Nacional de Gestión Sanitaria, Servicio de Recursos Documentales y Apoyo Institucional.

[21] Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, *34*(4), 555-596.

[22] Boualem, A., De Runz, C., & Ayaida, M. (2022). Partial Paving Strategy: Application to optimize the Area Coverage Problem in Mobile Wireless Sensor Networks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 13(2), 1-22.