Multidimensional Representations for the Gesture Phase Segmentation Problem
An Exploratory Study using Multilayer Perceptrons

Ricardo A. Feitosa, Jallysson M. Rocha, Clodoaldo A. M. Lima and Sarajane M. Peres
Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 03828-000, São Paulo-SP, Brazil

Keywords: Gesture Phase Segmentation, Multidimensional Representation, Multilayer Perceptron.

Abstract: Gesture analysis systems have been attracting a good deal of attention because of the improvements they have made to the interaction between humans, humans and machines, and humans and their environment. In this interaction, natural gesticulation can be regarded as a part of the linguistic system underlying the communication, and the whole information system that seeks to make use of this kind of interaction for making decisions, should be able to "interpret" it. This can be carried out through strategies for gesture phase segmentation. The establishment of an efficient data representation for gestures is a critical issue when undertaking this task. The chosen representation, as well as the way it is combined with analytical techniques, may or may not support the solution that is found. In this study, different forms of representation for gestures are applied to a Multilayer Perceptron to create a suitable environment for detecting the more discriminative representations. The results obtained in this study showed that spatial and temporal characteristics must be combined to build discriminatory gesture representation, for the context of gesture phase segmentation.

1 INTRODUCTION

The analysis of gestures is a task that is often carried out in contexts in which one intends to understand the meaning of a gesture and, on the basis of this understanding, make use of this meaning for some kind of decision making. In communication or interaction between people, the relationship between gestures and language is established in a natural way, either through a natural gesture or by composing a sign language. In contexts where the interaction between humans must be mediated by an information system or the interaction is established between humans and machines or environments, gestures should be considered as elements that convey information. If viewed as a system with a finite vocabulary, it is possible to define algorithms that can be exploited for the recognition of sign language, as in Ong and Ranganath, 2005. However, in the case of natural gesticulation, there is no finite or well defined vocabulary that can be analyzed. Gestures are uncertain and depend on the local and cultural diversity of the context in which they are made (Kim et al., 2007). In this scenario, the works in the literature study the movements of different parts of the body to analyze human behavior.

The context of gesture analysis studied in this work is constrained by the natural gesturing that is embedded in the communication. Within this context, the Theory of Gestures allows different types of analysis to be conducted. One of these analyses is the one conducted by Kendon, 1980, McNeill, 1992 and McNeill, 2015, who propose that a gesture is structured in phases. This means it allows an analysis to be undertaken of how the gestures are structured within the situation in which they are manifested.

The search for solutions for automated gesture phase segmentation requires the choice of a computational representation for the data under analysis – the gestures. This is a crucial choice and it must consider the data analysis technique that will be employed. Moreover, the domain in which the gestures are interpreted includes special features that must be well represented so that the computational algorithms are able to handle them properly. Within a linguistic system, executing and interpreting a gesture can entail different aspects, the most common of which are as follows: (1) spatial aspects that incorporate information about form, amplitude or direction; (2) temporal aspects that include information about both speed and acceleration, as well as frequency and periodicity; (3) structural aspects that hold the structural information and establish a link between the gestures

Feitosa, R., Rocha, J., Lima, C. and Peres, S.
Multidimensional Representations for the Gesture Phase Segmentation Problem - An Exploratory Study using Multilayer Perceptrons.
DOI: 10.5220/0006556003470354
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 347-354
ISBN: 978-989-758-275-2

and their constituent parts (Smith, 2011) (Dietterich, 2002), or between the gestures and the other elements of the system. The scope of this work is delimited by the study of computational representations that are capable of describing these aspects. Besides, since the objective is only to analyze the representations, the classical technique Multilayer Perceptrons were chosen to perform the pattern recognition.

This paper is structured as follows: Section 2 presents the basic concepts related to the solution proposed in this paper; Section 3 investigates the exploratory study under discussion when showing how gestures can be represented from spatial and temporal information; Section 4 describes the different features of the gestures that can be exploited to create multidimensional representations; Section 5 describes how the experiments were designed and carried out, and then analyzes the results obtained; finally, the Section 6 summarizes the conclusions and makes recommendations for further studies in the field.

2 THEORETICAL BACKGROUND

This section defines the concepts underlying the two theories used in this study. The problem of gesture phase segmentation arises from the area of Gesture Studies, and is examined in Section 2.1. The potential value and drawbacks of the different forms of representation for the gesture phases are explored by solving the gesture phase segmentation problem through the use of classifiers. In this study, such classifiers are established by Multilayer Perceptrons, and their theoretical framework is described in Section 2.2.

2.1 Gesture Phase Segmentation

The study of gestures examines the movements of body parts in communication. The analysis conducted in this study is based on studies carried out by Kendon, 1980, McNeill, 1992 and McNeill, 2015. The data employed by these researchers, as well as that of others in this area, are based on videos of people speaking and gesticulating, which are converted into representations that facilitate an analysis of gestures. According to these authors, the following phases should be followed: *preparation*, this is the phase when the limbs of the body, for example the hands, move from a resting position (a period with no gesticulation) to a stroking movement; *pre-stroke hold*, the phase which represents a pause in the movement of the hands between the preparation and the stroking; *stroke*, is the gesture itself, or the period of gesticulation that conveys some information that has significant meaning during the execution; *post-stroke hold*: the phase that represents a pause in the movement of the hands between the stroke and the retraction; *retraction*, the phase during which the hand adopts a return to the rest position; *hold*, a phase added to this proposal to represent a period in which there is no movement, although there is the presence of information or meaning. It usually occurs between the preparation and retraction phases.

The period between the moment the hand leaves the resting position and returns to the resting position is called a *gesture unit* (Kita et al., 1998). The stroke is the only mandatory phase within the gesture unit. Figure 1 illustrates such phases by arranging them in a pattern along a gesture unit that is expressed in a persons natural gesture. Some frames from a video were extracted to compose this illustration in which someone makes a gesture related to an action (a “twist”). Each frame was taken from the period in the video where there was a reference to a gesture phase.

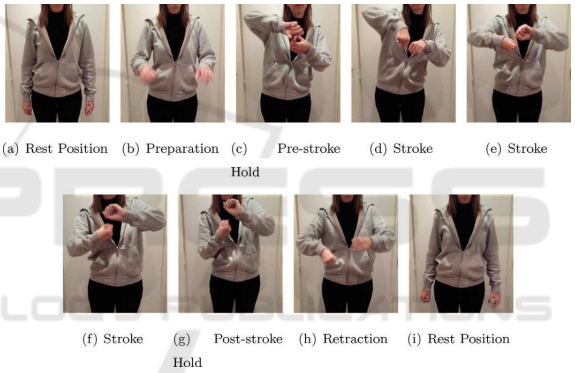

Figure 1: Illustration of a gesture unit (Madeo et al., 2016).

There is still difficulty in identifying phases with a similar configuration of limb features. There is a difficulty in differentiating between a hold and a rest, since both involve a reduced movement intensity. It is also hard to determine the transitions between the phases, or define the exact moment and frame in which the gesture can be found. For example, it might be from when the preparation phase enters the stroke phase, or when the movement has left the post-stroke hold phase and is about to enter the retraction phase.

Gesture analysis usually involves the following activities: (1) identifying the movements made by the users; (2) representing and recognizing movements based on a specific analytical model; and (3) making an evaluation of the results. After being extracted from the data sources, the gestures are captured on video and divided into frames. With regard to activities 2 and 3, a human specialist has to conduct an analysis of each frame extracted from the database to arrive at a verdict about which phase each frame belongs to, if the segmentation is carried out manually.

This manual activity must be undertaken by a specialists. The person carrying out this activity has to know the setting where the gestures were captured and their execution sequence to determine which phase of the gesture a given frame belongs to. It is difficult for specialists in this area to reach a consensus, since two experts can reach different conclusions about which phase of the gesture a frame belongs to, even though they have the same information (Kita et al., 1998). For these reasons, the manual segmentation of the gesture phases can take a long time to complete and have a low degree of efficiency.

The segmentation of gesture phases can also be carried out automatically by means of different techniques, and this is a means of overcoming or reducing the inherent problems of manual segmentation. Automatic segmentation usually relies on the same data sources as manual segmentation, but the decision about which phase of the gesture each frame should belong to is handled by heuristics, machine learning algorithms, grammars etc. The application of these techniques may require the collected gesture data to be preprocessed and transformed into a specific data representation so that they can be suitably analyzed.

Examples of techniques applied to automated gesture segmentation include the HMM (Hidden Markov Model), which is employed by Yin and Davis, 2014 for identifying the gesture phases. In Popa et al., 2008 also used HMMs to identify strokes from a representation of gestures based on information about the trajectory of the hand. HMMs have often been used to solve the problem of gesture phase segmentation, as for example, the videos used in a lecture by Martell and Kroll, 2007 to identify the phases of preparation, stroke, hold, and retraction within a gesture unit. Ramakrishnan and Neff, 2013 made use of heuristics and SVM (Support Vector Machines). Initially the heuristics were used to identify the resting position and points of interest in the video database. The SVMs were then used to classify these points of interest in the preparation, stroke, and retraction phases. Heuristics were also used by Bryll et al., 2001 in their analysis of the organization of speech movements from the gestures of the hands, with a view to identifying the "holds". It is also worth noting the work of Gebre et al., 2012, who used logistic regression to detect strokes of the hand from frames which contained information about the gestures collected.

Gesture phase segmentation is an approach used for the analysis of gestures which can benefit professionals in a wide range of tasks. It can be stated that linguistic specialists have been the main beneficiaries of analyses based on the gesture phase segmentation. However, several other applications can be delineated.

With regard to applications involving a finite vocabulary, one can cite the work of Spano et al., 2012, which uses sub-gestures to recognize more complex gestures. Mention should also be made of Lee et al., 2016 and Madeo et al., 2010, in which the configurations and movements of the hands are explored. Another recent work involving the application of gesture analysis is that of Freitas et al., 2014, in which grammatical facial expressions are identified.

Many of the challenges with regard to the analysis of gestures concern natural gestures because of the innumerable signs and nuances that characterize them. This is so much so that works like that of Jacob and Wachs, 2014 investigate whether gesture phases are used to determine if a gesture was intentional or not. An interesting example that concerns natural gesticulation is the study carried out by Salem et al., 2012. In his work, the phases of the gestures are used to analyze human gestures and design a natural gesture for robots, as well as to investigate the use of this kind of gesture in the interaction between robots and humans.

2.2 Multilayer Perceptrons

Artificial Neural Networks (ANN) are based on a number of concepts (Fausett, 1994): *a)* neurons are the elements that process information, which are organized in an input layer (input neurons), representing the data that have to be processed, in one or more hidden layers (optional), representing linear or nonlinear mappings of the data space, and an output layer (output neurons), representing the ANN's response; *b)* the information is transmitted between the neurons through connections; *c)* each connection is associated with a weight that influences the transmitted information; *d)* the output of each neuron is determined by an activation function applied to its input.

The Multilayer Perceptron (MLP) is an ANN that is used for non-linearly separable classification problems and is characterized by its supervised learning. It consists of an input layer, one or more hidden layers and an output layer, the latter two formed by the neurons that will process the information in the ANN and determine the resulting value (Haykin et al., 2009). While using this network, the signal is propagated from the input layer to one or more hidden layers and then forwarded to the output layer, thus characterizing a feedforward network. In this type of network, it is embedded a unitary and positive input, called bias, for each neuron to increase the degree of freedom and adaptation of the neural network during the learning process (Haykin et al., 2009). The learning process uses the supervised backpropagation technique, based on error-correction rules. This technique consists of

two phases: propagation and backpropagation. In the former, an input data vector is shown to the input layer and its effect propagates through the network, layer by layer, and produces a set of outputs. Following this, the backpropagation phase adjusts the values of the synaptic weights on the basis of an error calculated as the difference between the obtained response and the desired response. This adjustment is made from the output layer to the input layer. This learning technique causes the network response to move toward the expected response (Haykin et al., 2009).

The MLP network usually uses a nonlinear activation function in its neurons. This nonlinearity is smooth and differentiable at any point; it is usually conferred by a *sigmoid* represented by the logistic function $y_j = \frac{1}{1+exp(-v_j)}$, in which v_j is the weighted sum of the inputs plus the bias of the neuron j and y_j is the output of the neuron. The network contains one or more hidden layers of neurons that are involved in learning complex tasks, and extracting the most significant characteristics of the input data. It also has a high degree of connectivity between its neurons. Any alteration of this connectivity should involve making a revision of the number of neurons or their weights. The error is calculated from the difference between the obtained response and the desired response, according to $e_j(n) = d_j(n) - y_j(n)$, in which $e_j(n)$ is the error in the neuron j at instant n. Thus, during the backpropagation of the error, the new values of the network weights can be defined by rules that take account of this error, in the procedure defined by the Delta Rule. In the Delta Rule, the value of weight w at instant $t+1$ is defined by $w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t)$, wherein $\Delta w_{ij}(t) = w_{ij} - a \times \frac{\partial EQM}{\partial w_{ij}}$, $EQM(n) = \frac{1}{2}\sum_{k=1}^{ns} {e_j}^2(n)$, a is the learning rate and EQM represents the mean square error.

3 GESTURE REPRESENTATION

In automatic gesture segmentation, preprocessing of the gestural data is usually necessary to form a representation that is appropriate for the automated segmentation technique that will be used.

In spatial information-based representations, features are used such as the coordinates (x,y,z) of the spatial position of the members that are monitored, the angulation between the members, the hand contours, among other aspects. The most used data representations combine these features, as in Kyan et al., 2015 where angulation in relation to the body is used for the representation of the positioning of the hands, wrists, elbows, shoulders, shoulder blades, hips, pelvis, knees, feet, ankles, spine and head.

The colored images comprises other information that has been used to construct representations. Liu et al., 2014 used this characteristic to accomplish the task of identifying the positioning of body parts. Dardas et al., 2014 and Xu and Lee, 2011 used the colored images to carry out gesture recognition tasks and find the contour of the hands. In Zhu et al., 2011 the information about contours is used to select points around the body under analysis. The outline of the limbs can also be represented by information extracted from the pixel matching, as investigated by Liang et al., 2014 and by Liu et al., 2015. Pixel matching is a term used by Liu et al., 2015 to define a technique where images are analyzed through of searching for pixels with similar characteristics (spatial coordinates and pixel tone).

In temporal information-based representations, the aspects are: *a)* the speed with which the monitored members move, *b)* acceleration, *c)* the trajectory and other information for which the order of the frames are significant. Information about the acceleration and speed of the gesture is frequently used (Bailador and Triviño, 2010), (Khan et al., 2012) and (Madeo et al., 2013).

4 MULTIDIMENSIONAL REPRESENTATIONS

A number of studies rely on combinations of information about gestures by making use of more than one type of spatial information and combining it with temporal information. Hachaj and Ogiela, 2014 use the spatial coordinates and angles to represent the hands, elbows, shoulders, thorax, hips, knees, feet, spine and head for gesture recognition and temporal information. The temporal information is derived from a timestamp attribute associated to each frame. The timestamp is used to check the delay among datapoints acquisition. The choice of characteristics that belong to different categories is strategic in the task of gesture phase segmentation.

When the gesture is analyzed as a video, the characteristics are extracted from the set of frames that make up the video. The initial data are obtained from the static images that record each moment of the gesture. As initial data, the coordinates and angulations might represent a point of interest of the gesture, as in Caramiaux et al., 2012 where the task of identifying the gestures involves the position of the hands and arms combined with the angulation and trajectory of the lower and upper limbs of the body. The task of obtaining information about the trajectory of

the hands and the position of parts of the body was also explored by Abid et al., 2015, for the recognition and classification of gestures in real time. Information about the trajectory, along with the positioning of objects that people are interacting with in the captured scenes, made up the representation of data used by Lücking et al., 2013 and Rosani et al., 2014, that also investigated the task of recognizing and classifying gestures in real time.

When frames are analyzed in sequence, temporal information can be extracted. For example, the sequence of different values of a given coordinate along the frames can be analyzed as a time series, where each scalar is a datapoint in the series. In a time series analysis, studies can be carried out by means of the phase space concept, which also provides an analysis that assumes that sequences of scalar measurements depend on previous states in the signal (Kantz and Schreiber, 2004). The excursion of a spatial coordinate along the frames can also be understood as a continuous time-varying signal. Thus, it is possible to consider filters being applied to smooth the signal.

5 EXPLORATORY STUDY

The aim of our study was to adopt an experimental strategy for studing multidimensional representations for gestures. These took account of the results of the gesture phase segmentation obtained from the use of these representations together with the MLP algorithm. This section outlines the dataset used in the experiments, describes the preprocessing procedures and the extraction of the characteristics, and conducts an analysis of the results generated by the MLP.

5.1 Dataset and Preprocessing

The experiments were carried out in a controlled environment to provide a more wide-ranging analysis of the results obtained and get a greater dominance over the preprocessing processes. Following these guidelines, one of the videos of the *Gesture Phase Segmentation Data Set*[1] were used. Such a dataset consists of seven videos showing the gesticulations of people during a storytelling activity. It is a dataset with frames labeled according to the phases of gesture: rest position, preparation, stroke, hold and retraction. The video chosen for analysis by this work lasts for 60 seconds and generated a total of 1747 frames distributed as follows: 698 in rest position; 163 in preparation; 656 in stroke; 39 in hold and 191 in retraction.

The video was recorded with the aid of MS Kinect sensor, and relied on its ability to track the human body in the captured images and provide coordinates (x,y,z) of points of interest in the body. Six points of interest were traced: the left hand, right hand, left wrist, right wrist, head and spine. The temporal characteristics represented by vector and scalar velocities, together with vector and scalar acceleration were extracted from these data. Besides, the angulation between theses points and the axis x were extracted. Thus, there is a representation of both spatial and temporal features in the database. The spatial aspects were extracted from each frame. The coordinate x refers to the position of the point of interest in the frame in a vertical direction; the coordinate y refers to the position in the horizontal direction; the coordinate z is expressed in millimeters in the form of the distance between the sensor and the monitored point of interest. The temporal aspects are the speed and acceleration. Such aspects are expressed in terms of scalar and vector measurements of the right and left hands and right and left wrists. The velocity is obtained from the displacement of the points of interest in the time of three frames. The acceleration is obtained from the velocity measurements.

The data were normalized nullifying body displacements and the variations in the distance from the gesture to the sensor. This also prevented the risk of the algorithm being influenced by the movement of a member when this movement did not in fact exist. For example, when a person raises his right hand and waves their whole body briefly forward or backward, the right hand does not move in the light of this *de facto* gesture. However, there is some variation in the hands coordinates that may influence whether or not the algorithm treats this variation as a gesture. The Wavelet filter was also employed for the data in an attempt to make an improvement in discriminating the phases of the gestures (Semmlow and Griffel, 2014). The effect of the filter on the data is illustrated in Figure 2. The information about angulation was also added; this is calculated from the spatial coordinates of the hands and the spine and the spatial coordinates of the hands and their respective wrists.

5.2 Design of the Experiments

The experiments were implemented by using MATLAB® environment. A number of MLP were trained and tested.

Tests on spatial data both without a filter and with a Wavelet filter. There are 4,950 test cases for this

[1] ⟨https://archive.ics.uci.edu/ml/datasets/Gesture+Phase+Segmentation⟩, (Lichman, 2013) and (Madeo et al., 2013).

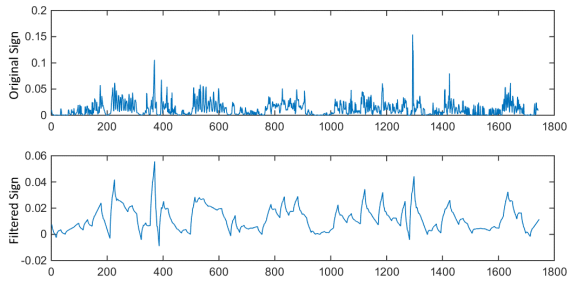

Figure 2: Scalar speed signal of the right hand when making a gesture: the first figure is the original signal; the second figure is the filtered signal.

scenario. These include tests with different combinations of gesture features: tests covering all the data, tests without the information of the *z* coordinate, tests with only *x* and *y* of the right hand, and tests with all the coordinates of the right hand. Each previous case also included tests with variations in the MLP optimization algorithm (Levenberg-Marquardt algorithm, gradient descent – GD, GD with adaptive learning rate, GD with momentum and GD with both adaptive learning rate and momentum) and in the MLP meta-parameters. Such meta-parameters were used as follows: initial learning rate varying in 0.1, 0.01 and 0.001; number of neurons of the hidden layer varying in 10, $2 \times \sqrt{x}$, in which x is the number of network inputs, and $n_2 = (x \times y)/2$, in which y is the number of network outputs. Finally, variations in the division of the dataset between training, validation and testing were added to each previous case: k-fold with k=10 and k=3, and holdout with divisions in the proportions of 70-15-15, 15-70-15, 15-15-70, 60-20-20, 20-60-20, 20-20-60, 40-30-30, 30-40-30 and 30-30-40 for the training, validation and testing phases.

Tests with spatial coordinates and angulation, also with and without a filter. These scenarios include *a)* cases with the coordinates of all points of interest and angles, *b)* tests with only the angles, and *c)* tests with all angles except those that are formed between the wrist and the hand. Here, the same MLP variations were applied to each variation referred to, making a total of 2,970 test cases.

Tests on temporal data. These tests are conducted as follows: *a)* scalar velocity, scalar acceleration, vector velocity and vector acceleration of a gesture; *b)* only vector velocity and vector acceleration; *c)* and only scalar velocity and scalar acceleration. The same MLP test combinations were applied to each of these variations, making a total of 4,950 test cases.

The gesture phase segmentation was modeled as a multi-class classifying problem, with five possible responses. The highest value among the five possible outputs was regarded as the "final class". As a result, it was possible to determine which models yielded the results that were closer to those expected in the gesture phase segmentation. Confusion matrices were created and precision, recall, accuracy and the F1-Score measurements were obtained from such a matrix. As it was designed as a multi-class problem, the evaluating measurements were calculated for each class. All evaluations were made considering the test division of the dataset.

5.3 Results and Discussion

Table 1 lists the best results obtained from all the experiments. These four best results were obtained from models created with spatial aspects. The angles was always added to the model, it was obtained with Levenberg-Marquardt algorithm, under the 10-fold cross validation strategy.

Table 1: Best scenarios: # - result identification; Ft - filter; Gc - characteristics used in the gestures representation; Lr - learning rate; Hs - number of neurons in the hidden layer.

#	Ft	Gc	Lr	Hs
1	wavelet	all - ang of w-h	0.001	383
2	-	all	0.001	383
3	-	all	0.001	401
4	-	all	0.1	401

Table 2 shows the F1-Scores of these four best scenarios. These high scores show the significant potential value of MLP for application to gesture phase segmentation. Table 3 shows the degree of accuracy of these four best results. The same high values were observed, what shows a good MLP response rate both for the identification of frames per class and for the entire classification of the model.

Table 2: F1-Score of the best results identified for the phases: Rest, Preparation, Stroke, Hold and Retraction.

#	Rest	Prep	Stroke	Hold	Retrac
1	0.869	0.831	0.888	0.974	0.917
2	0.874	0.869	0.850	0.963	0.825
3	0.834	0.557	0.939	0.951	0.893
4	0.802	0.504	0.902	0.963	0.786

Table 3: Accuracy of the best results identified for the phases: Rest, Preparation, Stroke, Hold and Retraction.

#	Total	Rest	Prep	Stroke	Hold	Retrac
1	0.827	0.784	0.847	0.849	0.949	0.864
2	0.861	0.838	0.933	0.877	1.000	0.801
3	0.804	0.729	0.693	0.898	1.000	0.806
4	0.736	0.679	0.571	0.826	1.000	0.723

The use of the Wavelet filter proved useful for fine-tuning the results. However, although it was employed in the best results obtained, when it was not

used, the results obtained were not very different from when it was implemented. The extent of the difference can be observed in the ranking of these four best results. Besides, it is worth noting that the best results were obtained when the information about angulation was added to the data representation. This information added an important discriminative feature which improved the results of the classification.

Classifiers built on data representations without the use of angulation were also successful. A good result obtained in this case is shown in Table 4. Achieving a F1-score of 0.65 in the identification of the preparation phase is not a bad score, considering that this is a transitional phase between the absence of movement and movement, which usually makes it difficult to classify and differentiate it from a rest or stroke phases. The configuration of such scenario was: no filter, all the spatial coordinates, initial learning rate in 0.01, 355 neurons in the hidden layer, Levenberg-Marquardt algorithm and training, validation and test with the 70-15-15 dataset division.

Table 4: Results obtained without information about angles.

Total	Rest	Prep	Stroke	Hold	Retrac
F1-Score					
-	0.937	0.650	0.937	0.951	0.804
Accuracy					
0.902	0.999	0.571	0.934	1.000	0.696

With regard to the second best results, these included those obtained by the models created from the use of temporal features (velocity and acceleration). These results were positive and were in the middle range between the models that used angulation and those that only had spatial coordinates. Table 5 lists four of these satisfactory results. They were all obtained from the use of the speed and acceleration data, by means of a MLP optimized with the Levenberg-Marquardt algorithm, initial learning rating in 0.01 or 0.001, and 472 neurons on hidden layer. In the first three results were obtained using 10-fold cross-validation, and in the fourth case the data was divided in proportions of 70-15-15 in the holdout strategy.

Table 5: F1-Score and accuracy of the best results. Models with temporal information-based representation.

#	Total	Rest	Prep	Stroke	Hold	Retrac
F1-Score						
5	-	0.974	0.883	0.962	0.769	0.851
6	-	0.970	0.798	0.939	0.825	0.844
7	-	0.967	0.828	0.958	0.895	0.856
8	-	0.934	0.799	0.876	0.800	0.802
Accuracy						
5	0.918	0.967	0.877	0.934	0.769	0.749
6	0.924	0.967	0.920	0.921	0.846	0.796
7	0.935	0.941	0.933	0.947	0.872	0.885
8	0.834	0.939	0.742	0.779	0.923	0.702

6 CONCLUSION

Gesture phase segmentation as a support for gesture analysis and a means of recognizing gesture patterns that are carried out automatically, requires the creation of a suitable representation for the data in question. This representation must be discriminant enough to allow machine learning algorithms to achieve the segmentation and, hence discover knowledge about gesture patterns. This study covered the construction of representations that take account of spatial and temporal features. The results described in this paper confirm that the most significant results were achieved through the complete representation of the gesture (including all the descriptive characteristics). However, the representation that includes spatial aspects may impose a user-dependent or discourse context-dependent bias to the classifiers. It means that the classifiers performance could deteriorate if diferent contexts were considered. Experiments considering more than one video are being carried out in order to verify such hypothesis. It is also planned to assemble models using the concept of rectangular data windows formed by data from a sequence of *n* frames. By employing such a concept, it should be possible to create the conditions for the MLP algorithm to include information both before and after the frame has been analyzed, and thus increase the temporal character of the representation of the gesture.

REFERENCES

Abid, M., Petriu, E., and Amjadian, E. (2015). Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar. *IEEE Trans. Instrum. Meas.*, 64(3):596–605.

Bailador, G. and Triviño, G. (2010). Pattern recognition using temporal fuzzy automata. *Fuzzy Sets and Syst.*, 161(1):37–55.

Bryll, R., Quek, F., and Esposito, A. (2001). Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Commun.*

Caramiaux, B., Wanderley, M. M., and Bevilacqua, F. (2012). Segmenting and parsing instrumentalists' gestures. *J. of New Music Research*, 41(1):13–29.

Dardas, N. H., Silva, J. M., and El-Saddik, A. (2014). Target-shooting exergame with a hand gesture control. *Multimedia Tools Application*, 70(3):2211–2233.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Proc. of Joint Structural, Syntactic, and Statistical Pattern Recognit. Int. Workshops*, pages 15–30. Springer.

Fausett, L., editor (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Freitas, F., Peres, S., Lima, C., and Barbosa, F. (2014). Grammatical facial expressions recognition with machine learning. In *Proc. of 27th Florida Artificial Intell. Research Society Conf.*, pages 180–185. AAAI.

Gebre, B. G., Wittenburg, P., and Lenkiewicz, P. (2012). Towards automatic gesture stroke detection. In *8th Int. Conf. on Language Resources and Evaluation*, pages 231–235. European Language Resources Association.

Hachaj, T. and Ogiela, M. R. (2014). Rule-based approach to recognizing human body poses and gestures in real time. *Multimedia Systems*, 20(1):81–99.

Haykin, S. S., Haykin, S. S., Haykin, S. S., and Haykin, S. S. (2009). *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA:.

Jacob, M. G. and Wachs, J. P. (2014). Context-based hand gesture recognition for the operating room. *Pattern Recognition Letters*, 36:196 – 203.

Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis*, volume 7. Cambridge university press.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The Relationship of verbal and nonverbal communication*, pages 207–227.

Khan, S., Bailey, D., and Gupta, G. S. (2012). Detecting pauses in continuous sign language. In *Proc. of Int. Conf. on Mechatronics and Mach. Vision in Practice*, pages 11–15. IEEE.

Kim, D., Song, J., and Kim, D. (2007). Simultaneous gesture segmentation and recognition based on forward spotting accumulative {HMMs}. *Pattern Recognit.*, 40(11):3012 – 3026.

Kita, S., Gijn, I., and Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Proc. of Int. Gesture Workshop Bielefeld*, pages 23–35. Springer.

Kyan, M., Sun, G., Li, H., Zhong, L., Muneesawang, P., Dong, N., Elder, B., and Guan, L. (2015). An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment. *ACM Trans. on Intell. Syst. Technol.*, 6(2):23:1–23:37.

Lee, G. C., Yeh, F.-H., and Hsiao, Y.-H. (2016). Kinect-based taiwanese sign-language recognition system. *Multimedia Tools and Applications*, 75(1):261–279.

Liang, H., Yuan, J., and Thalmann, D. (2014). Parsing the hand in depth images. *IEEE Trans. Multimedia*, 16(5):1241–1253.

Lichman, M. (2013). UCI machine learning repository.

Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., and Yan, S. (2014). Fashion parsing with weak color-category labels. *IEEE Trans. Multimedia*, 16(1):253–265.

Liu, S., Liang, X., Liu, L., Lu, K., Lin, L., Cao, X., and Yan, S. (2015). Fashion parsing with video context. *IEEE Trans. Multimedia*, 17(8):1347–1358.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2013). Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *J. on Multimodal User Interfaces*, 7(1-2).

Madeo, R. C. B., Peres, S. M., Bíscaro, H. H., Dias, D. B., and Boscarioli, C. (2010). A committee machine implementing the pattern recognition module for fingerspelling applications. In *Proc. of the ACM Symposium on Applied Computing*, pages 954–958.

Madeo, R. C. B., Peres, S. M., and Lima, C. A. (2016). Gesture phase segmentation using support vector machines. *Expert Syst Appl.*, 56:100 – 115. In press.

Madeo, R. C. B., Wagner, P. K., and Peres, S. M. (2013). A review of temporal aspects of hand gesture analysis applied to discourse analysis and natural conversation. *Int. J. of C. Sci. & Inf. Tech.*, 5(4).

Martell, C. and Kroll, J. (2007). Corpus-based gesture analysis: an extension of the form dataset for the automatic detection of phases in a gesture. *Int. J. of Semantic Computing*, 1(04):521–536.

McNeill, D. (1992). Hand and mind: What the hands reveal about thought.

McNeill, D. (2015). *Why We Gesture: The Surprising Role of Hand Movements in Communication*. Cambridge University Press.

Ong, S. C. and Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):873–891.

Popa, D., Simion, G., Gui, V., and Otesteanu, M. (2008). Real time trajectory based hand gesture recognition. *WSEAS Trans. on Inf. Sci. and Appl.*, 5(4):532–546.

Ramakrishnan, A. S. and Neff, M. (2013). Segmentation of hand gestures using motion capture data. In *Proc. of the Int. Conf. on Autonomous Agents and Multi-agent Systems*, pages 1249–1250.

Rosani, A., Conci, N., and Natale, F. G. B. D. (2014). Human behavior recognition using a context-free grammar. *J. of Electronic Imaging*, 23(3):033016.

Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., and Joublin, F. (2012). Generation and evaluation of communicative robot gesture. *Int. J. of Social Robotics*, 4(2):201–217.

Semmlow, J. and Griffel, B. (2014). *Biosignal and Medical Image Processing, Third Edition*. Taylor & Francis.

Smith, N. A. (2011). *Linguistic Structure Prediction*. Morgan & Claypool.

Spano, L. D., Cisternino, A., and Paternò, F. (2012). A compositional model for gesture definition. In *Int. Conf. on Human-Centred Softw. Eng.*, pages 34–52. Springer.

Xu, W. and Lee, E.-J. (2011). Hand gesture recognition using improved hidden markov models. *J. of Korea Multimedia Soc.*, 14(7):866–871.

Yin, Y. and Davis, R. (2014). Real-time continuous gesture recognition for natural human-computer interaction. In *IEEE Symp. on Visual Languages and Human-Centric Computing*, pages 113–120.

Zhu, L., Chen, Y., Lin, C., and Yuille, A. L. (2011). Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. *Int. J. of Computing Vision*, 93(1):1–21.