Gesture Unit Segmentation Using Spatial-Temporal Information and Machine Learning

Priscilla K. Wagner, Sarajane M. Peres
Clodoaldo A. M. Lima, Fernando A. Freitas
Universidade de São Paulo
São Paulo, SP, Brazil

Renata Cristina Barros Madeo
Universidade Nove de Julho
São Paulo, SP, Brazil

Abstract

Currently, automated gesture analysis is being widely used in different research areas, such as human-computer interaction or human-behavior analysis. With regard to the latter area in particular, gesture analysis is closely related to studies on human communication. Linguists and psycholinguists analyze gestures from several standpoints, and one of them is the analysis of gesture segments. The aim of this paper is to outline an approach to automate gesture unit segmentation, as a way of assisting linguistic studies. This objective was attained by employing a Machine Learning technique with the aid of a spatial-temporal data representation.

1 Introduction

In the last few years, there has been an increase in research related to gesture studies, an interdisciplinary area that aims to analyze the use of hands and other parts of the body for communication. The study of gesticulation, i.e. the study of the gestures that accompany speech (McNeill 1992) is an important topic that has been studied by researchers from several areas. The usual way to study gesticulation is to carry out the gesture analysis on recorded videos of people talking and gesturing.

In these analysis, it is generally necessary to obtain a transcription of the gestures that are performed, which involves segmenting them into phases. This segmentation can be divided into two parts: the segmentation of gesture units, the period between two rest positions; and the segmentation of phases within the gesture unit (McNeill 1992), (Kendon 1980). These segmentations are usually called labeling and are carried out manually by researchers, which involves a slow and arduous procedure. In view of this, automated routines could greatly assist and accelerate research into gesture studies (Vinciarelli et al. 2008).

This paper discusses some of the results obtained from the employment of a Machine Learning (ML) technique – Multilayer Perceptron (MLP) – in addressing the gesture unit segmentation problem, as a first step in the phases of gesture segmentation. For this reason, the problem was modeled as a binary classification problem, in which each frame of a video is labeled as gesture unit or rest position.

In this paper, we address two research questions. The first is related to the exploration of different data representations. The second question is related to how to deal with the ambiguity inherent in the problem of analyzing gestures; even among experts, there is serious disagreement in determining when a gesture unit starts and ends. It is necessary to measure this ambiguity to evaluate the resulting MLP model correctly. We propose to use classical agreement coefficients (Artstein and Poesio 2008) in order to handle this topic.

This paper is structured as follows: Section 2 outlines theoretical concepts, which provide a background to the discussions that take place later in the paper, and examines some related studies; Section 3 defines the problem and describes the spatial-temporal data representation; a proof of concept is designed to test our hypothesis as outlined in Section 4; and, lastly, Section 5 concludes the study and makes suggestions for future work.

2 Literature Rewiew

In this section we provide a brief analysis of the theoretical concepts that are employed in our gesture unit segmentation approach and comment on some related studies in this field.

Background

The approach discussed in this paper is based on Machine Learning. The reason for choosing this type of strategy was largely based on the absence of well-defined models (or rules) that describe the behavior of gestures within the scope of gesticulation in natural conversation. Since Machine Learning techniques can discover patterns, they are potentially useful in this context. It is also necessary to present some aspects of the gesture analysis problem, in order to highlight the features that led to the decision making undertaken in this research: the study of spatial-temporal representations and the use of agreement coefficients.

Machine Learning Machine Learning is characterized by the development of methods and techniques to provide computer programs with the ability of enhancing its performance in a task, learning from experience (Mitchell 1997). This type of learning (inductive learning) can be achieved by supervised or unsupervised methods. In the supervised methods, the technique considers a labeled dataset and adjusts the parameters to minimize an objective function.

A well-known technique that implements inductive supervised learning is the Multilayer Perceptron (MLP). The MLP is a feedforward and multilayer neural network architecture, that is usually trained with the also well-known backpropagation method (or generalized delta rule). In the experiments discussed in this paper, the backpropagation method is implemented using gradient descent, and an adaptive learning rate. For further information about the MLP neural network see (Haykin 2008).

Gesture Theory People commonly perform one or several movements or "excursion" with their hands, arms or even their bodies during a natural conversation or when giving a speech. According to Kendon (1980) and McNeill (1992), an excursion, particularly regarding the hands, refers to a movement from a rest position to some region in space, and then bringing them back to the same or another rest position. While the hands are away from the rest position, the movement is called a gesture unit. When these gestures co-occur with speech, they are called gesticulation.

In addition, according to Kendon and McNeill, a gesture unit can consist of one or more gestural phrases, which can be divided into phases: *preparation*, *stroke*, *hold* and *retraction*. *Stroke* defines the main movement in a gestures unit and carries a semantic meaning; *holds* are pauses during the phrase, in which the hand configuration used in the stroke is maintained; *preparation* and *retraction* are transitory phases between the gesture units and rest positions.

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

- the limits between the transitional phases and the rest positions: in fact, in the analysis of human gestures there is no precise boundary between the phases, since in real situations involving gesticulation, these limits are not apparent. Moreover, the differences shown by human coders (see Section 2) are closely related to this fact.
- the similarity between holds and rest positions: both phases are characterized by hands in a fixed configuration with an almost entire absence of movements. However, rest positions do not have any semantic content, whereas the interpretation of holds takes account of its meaning[1]. Hence, there is a problem that, in a automated analysis based on a kind of gesture representation that is devoid of semantic information, frames in a "hold segment" and frames in a rest position sequence may be too similar to allow them to be correctly recognized.

Figure 1 shows sequences of rest positions and gesture units taken from videos. In this diagram, it is possible to observe the problems outlined above.

Evaluation Strategy There are different methods of evaluating results in gesture analysis. Most studies conduct a frame-by-frame analysis, since they classify each frame as belonging to a specific phase (Martell and Kroll 2007), (Wilson, Bobick, and Cassell 1996), (Bryll, Quek, and Esposito 2001), (Gebre, Wittenburg, and Lenkiewicz 2012). In these cases, it is difficult to analyze when a segment is detected correctly, since it is hard to define the maximum number of incorrect frames that can be accepted within a segment. In addition, an error in the margin of a segment has the same weight as an error within the segment, which would be wrong from the perspective of gesture studies. However, Ramakrishnan (2011) identifies inflection frames that correspond to the beginning of each phase, and then classifies each segment as belonging to each phase. Thus, this author conducts an analysis of the segments, by taking into account the fact that there is no clear point where the segments start or end and by admitting some deviation at the borders.

Given this, it would be reasonable to argue that if a degree of ambiguity is revealed when the analytical results of the gesture are evaluated, this should be also considered in the conception and evaluation of the model that will be employed to analyze the gestures. In this paper, we propose to use agreement coefficients, in particular *Krippendorff's Alpha* (Kα) coefficient, to carry out this evaluation. The Kα was chosen because the α-type coefficient is the most commonly used measurement in agreement analysis, and because it minimizes coding biases (Artstein and Poesio 2008). The Kα coefficients range is from -1 to 1. Negative values indicate random classification (or labeling) or insufficient data. Values between 0 and 0.2 indicate slight agreement; between 0.2 and 0.4, fair; between 0.4 and 0.6, moderate; between 0.6 and 0.8, substantial; and, between 0.8 and 1, a perfect agreement (Artstein and Poesio 2008).

Related studies

Although there has been a considerable amount of work regarding gesture analysis within the gesture studies area (Kendon 1980), (McNeill 1992), (McNeill 2005), (Kita, van Gijn, and van der Hulst 1998), efforts for employing automated methods for gesture analysis are much more recent (Madeo, Wagner, and Peres 2013). The main studies in automated gesture phase segmentation are ones presented in (Martell and Kroll 2007) and (Ramakrishnan 2011).

Martell and Kroll (2007) considered a corpus in which gesture units are already segmented and used a Hidden Markov Model to classify each frame in preparation, stroke, hold or retraction phase. Ramakrishnan (2011), summarized in (Ramakrishnan and Neff 2013), identified the most frequent rest positions for each person. Following this, gesture phase segmentation was performed by using heuristics to identify the hold phases and inflection frames, and Support Vector Machines (SVM) to identify preparation, stroke, and retraction. There are other studies that either discuss gesture unit segmentation or specific tasks within gesture phase segmentation. With regard to gesture unit segmentation, the authors in (Madeo, Lima, and Peres 2013) used SVM to classify each frame as either rest position or gesture unit. Wilson, Bobick and Cassel (1996) used a heuristic method for performing gesture unit segmentation. In identifying the specific phases, Gebre, Wittenburg and Lenkiewicz (2012) used Logistic Regression to detect stroking gestures. Bryll, Quek and Esposito (2001) applied a heuristic method to detect hand holds in natural conversation.

[1]This assertion about semantic content and holds is not a consensus in the theory of gesture area.

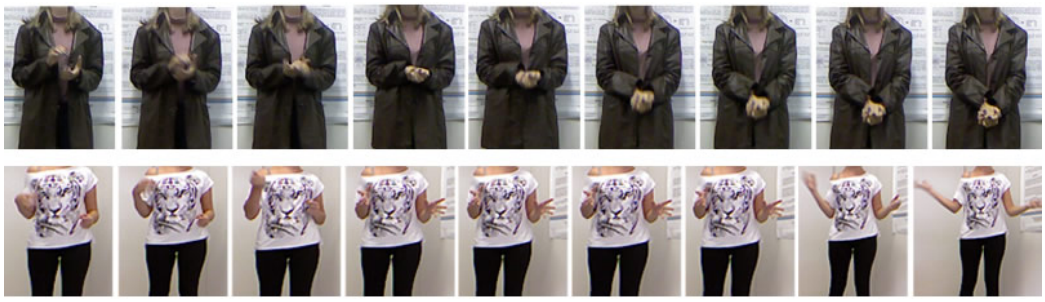

Figure 1: In the first line of frames, there is a rest position sequence from 4^{th} to 9^{th} frames. It is difficult to determine the first frame of the sequence, and there is a hands displacement into the sequence. In the second line, there is a hold sequence from 4^{th} to 7^{th} frames. It is a period with an almost entire absence of movements into a gesture unit.

Although such studies are not directly comparable, it is useful to observe the performance which were already reached in this kind of problem. The results obtained in each study are listed in Table 1.

Table 1: Results obtained in the related works (F-score).

Id.	Gesture Unit	Preparation	Stroke	Hold	Retraction
			Phases		
1		0.54	0.59	0.36	0.67
2*	0.88	0.74	0.79	0.95	0.86
2**		0.64	0.78	0.75	0.78
3	0.87				
4	0.81				
5			0.39		
6				0.84	

1 (Martell and Kroll 2007) 2 (Ramakrishnan 2011)

3 (Madeo, Lima, and Peres 2013) 4 (Wilson, Bobick, and Cassell 1996)

5 (Gebre, Wittenburg, and Lenkiewicz 2012) 6 (Bryll, Quek, and Esposito 2001)

* user dependent approach ** user independent approach

Bryll et al (2001) also highlights the fact that in gesture phase segmentation different coders produce different results. There are a few studies that discuss this problem, such as Kita et al (1998) and Martell (2005). Kita et al (1998) reported that there may be an average inter-coder disagreement of up to 28% regarding gesture phase segmentation, or up to 42% if the gesturing consists of sign language. Martell (2005) conducted a comparative analysis of inter-coder agreement and intra-coder agreement (for two different codings made by the same coder at different times) in gesture unit segmentation and gesture movement description, and obtained an average inter-coder agreement of 39% and an average intra-coder agreement of 13.4%[2].

3 Gesture Unit Segmentation

In this work, a video is the input for the gesture unit segmentation. This consists of a sequence $S = \{f_1, f_2, \cdots, f_N\}$ of RGB image frames. The segmentation problem can be defined as a classification problem, which concerns the problem of receiving a frame f_i from S as input, and classifying it according to one of the classes in $C = \{+1, -1\}$. This means that if the chosen class is +, $f_i \in$ a *gesture unit* video segment; otherwise, $f_i \in$ a *rest position* video segment. It is therefore, a binary classification problem.

A software application based on Microsoft Kinect™sensor was implemented to obtain (from each frame): 3-dimensional coordinates (x, y, z) from six points of interest in the human body (hands, wrists, head and chest); a RGB image frame; and an associated timestamp. The RGB frame and the timestamp are an useful means of supporting the manual labeling process.

The representations used in this paper include spatial and temporal information extracted from the gestures. This information has been represented by using the following: 3-dimensional coordinates (x, y, z) from hands and wrists; scalar velocity and acceleration calculated in relation to the movements of the hands and wrists; and a windowed strategy that allows information from past and future frames to be used to create an implicit temporal representation.

Feature extraction consists of: a pre-processing phase, which aims at making the representation invariant to the lateral displacement and distance of the user in relation to the camera; and a velocity and acceleration extraction phase. In the pre-processing phase, the hands and wrists coordinates are subtracted from the chest coordinates in each frame, and this new position is divided by the distance between the points of the head and chest. Details about the pre-processing and the feature extraction procedures can be analyzed in (Madeo, Lima, and Peres 2013).

In light of the standardized raw data and the velocity and acceleration extracted from it, the complete representation of an isolated frame[3] is a 20-dimensional vector with scalar velocity, scalar acceleration, and the (x, y, z)-coordinates of left hand, right hand, left wrist and right wrist.

Nevertheless, it is worth taking account of information about prior and post frames in the analysis of each frame, to obtain temporal information about the gestures. Thus, a windowed strategy was applied to make each data represen-

[2]This assumes there is an exact correspondence between the gesture unit segmentation and all the attributes for the description of the body gesture movements.

[3]In fact, the representation of an isolated frame considers three frames in the raw data due to a frame displacement required in the velocity calculus.

tation possible. The vector representation is composed of a sub-sequence of frames $S' = \{f_1, f_2, \cdots, f_n\}$ where n is the size of a window centered in a frame of interest f_{int}. Table 2 illustrates the framework of window, including the features of one frame. The dimension of the vector representation rises n times in a windowed representation.

Table 2: Example of a window with $n = 5$ centered in the frame f_{int}. Legend: l – left; r – rigth; h – hand; w – wrist; v – velocity; a – accelaration; x,y,z – coordinates.

f_{int-2}		f_{int-1}		f_{int}		f_{int+1}		f_{int+2}	
vlh	alh	xlh	ylh	zlh	vrh	arh	xrh	yrh	zrh
vlw	alw	xlw	ylw	zlw	vrw	arw	xrw	yrw	zrw

The information described above was arranged into four data representations to study how the spatial and temporal information can help overcome the problem of gesture unit segmentation:

- **Data Vector 1** – Velocity and Acceleration (Wagner, Madeo, and Peres 2013): a $8n$-dimensional vector including scalar velocity and acceleration related to left hand, right hand, left wrist and right wrist;
- **Data Vector 2** – Position: a $12n$-dimensional vector including the (x, y, z)-coordinates of left hand, right hand, left wrist and right wrist, in each frame;
- **Data Vector 3** – Velocity, Acceleration and Position: a $20n$-dimensional vector combining the vectors of **Data Vector 1** and **Data Vector 2**;
- **Data Vector 4** – Velocity, Acceleration and Position of the Central Frame: $8n + 3$-dimensional vector composed of the vector of **Data Vector 1** and the (x, y, z)-coordinates of the frame f_{int}.

4 Proof of concept

We have designed a proof of concept based on experiments carried out in a storytelling context to validate our approach. The subject of comics was chosen because the narrative is well-known and thus, the storytelling would be an easy task. In this section, there is a description of the datasets, the experiments and the obtained results.

Datasets

The study of the problem of gesture unit segmentation in natural gesticulation requires a dataset composed of videos of people talking and gesturing. The data consists of streams of gestures, captured with Microsoft Kinect™ sensor.

Different data capture sessions were carried out, and in each session a storyteller read a comic and told the story in front of the sensor device. Three different stories and three different users were included: **user A** told **story 1**, **story 2** and **story 3**; **user B** and **user C** told **story 1** and **story 3**. Each captured video was represented in accordance with the four different data representations, as described in Section 3. Table 3 provides information about the captured videos.

Table 3: Dataset description: videos A1 and A2 were captured on **Session 1**; video B1 was captured on **Session 2**; videos A3 and B3 were captured on **Session 3**; and, videos C1 and C3 were captured on **Session 4**.

Video (User/Story)	Length (seconds)	(-) frames	(+) frames	% (+) frames
A1	58	698	1049	60.05%
A2	42	468	796	62.98%
A3	61	598	1236	67.39%
B1	36	80	993	92.55%
B3	48	157	1267	88.98%
C1	37	232	879	79.12%
C3	48	350	1098	75.83%

The MLP was applied as the segmentation model. Since it is a supervised strategy, it requires a labeled training set. This resource was provided by three human coders (**coder 1**, **coder 2** and **coder 3**) which have manually labeled the videos using two classes: *gesture unit* and *rest position*. As described earlier, this is a subjective process and the experiment requires an assurance that these data are reliable. The reliability of the labeling process was measured by using the (Kα) coefficient and the percentage of labeling divergence, and the resulting measurements are listed in Table 4.

Table 4: Agreement coefficients (first numbers in cells) and percentages of divergence (second numbers in cells).

Coders	Videos						
	A1	A2	A3	B1	B3	C1	C3
1 e 2	0.92	0.88	0.80	0.73	0.66	0.73	0.86
	4.64	5.93	9.92	5.66	**7.23**	10.52	5.52
1 e 3	**0.92**	0.91	**0.91**	0.78	0.5	0.82	**0.92**
	4.29	4.19	**4.96**	6.55	10.88	6.38	**2.62**
2 e 3	0.89	**0.93**	0.85	**0.88**	**0.67**	**0.85**	0.88
	5.84	**3.48**	6.71	**5.21**	7.72	**5.94**	4.56

From Table 4, it can be noted that the labelings are reliable, since the percentage of divergence is low and the Kα coefficients are around 0.8. This threshold indicates that there is a substantial or perfect agreement between the labelings. This means that, any of the three labelings is suitable to train the MLP models. However, the agreements concerning coder 3 gave slightly higher values, and thus, the labeling produced by coder 3 was chosen to train the MLP models.

Experiments and Results

The experiments have been carried out to determine the performance of the models built with MLP and using the data representations described in Section 3. In the case of Data Vector 1, we are revisiting previously published results (Wagner, Madeo, and Peres 2013).

Two sets of experiments have been carried out specifically for this paper with three data representations (Data Vectors 1, 2 and 3). Experiment 1 consists of training the MLPs at the beginning of the video (70% of the frames) and testing at the end of the same video (30%). The objective was to evaluate the ability of MLP to segment the gesture units by adopting an user-story-session-dependent approach. Exper-

Table 5: **Results for experiments**. Parameters – window size in frames, number of neurons at hidden layer and learning rate; and result – F-score.

Data Vector	Video	Window size	Hidden Neurons (#)	Learning rate	F-score	Video	Window size	Hidden Neurons (#)	Learning rate	F-score
Previous work										
1	A1	40	17	0.125	0.9041	B1	75	24	0.0625	0.3000
	A2	65	22	0.5	0.6440	B3	65	22	1	0.5060
	A3	60	21	1	0.7362	C1	45	18	0.0625	0.7848
	A1/A2	45	18	0.5	0.7964	C3	50	20	0.0625	0.6868
Experiment 1										
2	A1	40	17	0.0312	0.9264	B1	75	24	0.0156	**0.8461**
	A2	40	17	1	0.8889	B3	50	20	0.25	**0.7786**
	A3	45	18	0.5	0.9129	C1	50	20	0.0156	**0.8163**
						C3	50	20	0.0156	**0.9800**
3	A1	50	20	0.125	0.9287	B1	50	20	0.0625	0.7333
	A2	60	21	0.0625	**0.9313**	B3	45	18	0.0156	0.7241
	A3	40	17	0.25	**0.9139**	C1	65	22	0.025	0.8081
						C3	45	18	0.125	0.9333
4	A1	50	17	0.0625	**0.9339**	B1	60	21	0.25	0.3809
	A2	50	20	0.0156	0.8279	B3	50	20	0.0625	0.7543
	A3	75	24	0.0156	0.8778	C1	40	17	0.0312	0.8148
						C3	40	17	1	0.8224
Experiment 2										
2	A1/A2	40	17	0.0156	0.8234					
3	A1/A2	55	20	0.5	0.8473					
4	A1/A2	40	17	0.125	**0.8750**					

iment 2 aims to evaluate the MLP performance by using video A1 for the training and video A2 for the testing; in an user-session-dependent and story-independent approach.

All the experiments were carried out with the aid of Matlab®. We have applied the MLP neural network by using the following set of parameters: adaptive learning rate strategy; initial learning rate from 1 to 0.01 varying by a divide-by-two decrement rate; amount of epochs from 300 to 1300, varying by 100 steps; and number of neurons in the hidden layer determined by a heuristic rule (geometric mean between the amount of neurons in the input and output layers).The size of the window was determined in the range of 40 to 75[4], varying by 5 frames. The best MLP architectures were chosen through the performance metric *F-score* (Han, Kamber, and Pei 2006), since the datasets were unbalanced.

The models were tested every 10 training epochs to avoid overfitting. Table 5 shows the results obtained with the best model for each video and each data representation, in each set of experiments, considering the test partitions. Notice that, for Experiment 1, the models built with Data Vector 2 obtained the best results for videos from user B and C; this users are characterized by performances with few sequences of rest positions. The best results for user A were obtained with data representation that considers information about trajectory/position (using coordinates) and behavior (using velocity and acceleration). Regarding to Experiment 2, the data representations that include trajectory/position information reached results strongly better than those obtained only considering the velocity and acceleration information.

The classifications obtained with the best models were used to evaluate the model agreement with regard to the human coders. Table 6 shows the agreement coefficients between the best models and each coder. When these results are assessed, it is clear that almost all the models are compatible with the understanding of the problem provided by the coders, when position/trajectory are used. Overall, the coefficients Kα assume values which are similar to that values provided in the analysis of agreement among the human coders; in some cases, the agreement between the MLP and the coders 1 or 2 is higher than between the model and the coder 3 (whose labeling was used in the training). This means that the MLP learning is reasonable, and in this standpoint, the segmentation models present consistent results. There are some negative coefficients that were obtained in the analysis of video B1. In fact, the test set for this video is really difficult, even to the human coders.

Moreover, the best segmentation results were also evaluated by means of metrics that are generally used by specialists in gesture analysis. Some segmentation errors are expected due to the subjectivity of the problem, as stated in Section 2: transition errors; hold frames labeled as rest position. This analysis was applied for two models, as an example. The results are shown in Table 7.

5 Conclusions

In this paper, gesture unit segmentation problem was discussed. The results were analyzed with classical measurements used in binary classifiers and gesture analysis. The experiments covered story-dependent and independent approaches and shown that using position/trajectory information is useful to reach good results. The comparison with previous results show that position/trajectory information

[4]We have tested smaller windows in previous work, but the results were not so good. See (Madeo, Lima, and Peres 2013)

Table 6: Agreement evaluation between the best segmentation models and the human coders.

Previous work							
Coder (Data Vector)	**Videos** **A1**	**A2**	**A3**	**B1**	**B3**	**C1**	**C3**
1(1)	0.80	0.42	0.56	0.00	0.29	0.67	0.52
2(1)	0.77	0.42	0.16	0.61	0.28	0.57	0.44
3(1)	0.83	0.47	0.58	0.27	0.45	0.75	0.59
Video		**Coder 1**		**Coder 1**		**Coder 1**	
A1/A2		0.66		0.63		0.65	
Experiment 1							
1(2)	0.85	0.77	0.89	-0.01	0.66	0.89	0.89
2(2)	0.86	0.79	0.58	-0.03	0.56	0.83	0.79
3(2)	0.86	0.84	0.87	0.83	0.73	0.78	0.97
1(3)	0.87	0.83	0.88	-0.02	0.56	0.86	0.84
2(3)	0.89	0.85	0.49	-0.04	0.59	0.80	0.72
3(3)	0.87	0.90	0.87	0.72	0.67	0.77	0.91
1(4)	0.85	0.79	0.80	0.00	0.47	0.72	0.69
2(4)	0.86	0.75	0.42	0.45	0.52	0.69	0.59
3(4)	0.88	0.76	0.81	0.35	0.71	0.78	0.76

Experiment 2			
Coder	**Data Vector 2**	**Data Vector 3**	**Data Vector 4**
1	0.73	0.77	0.80
2	0.75	0.76	0.77
3	0.69	0.75	0.80

Table 7: Evaluation from experts' viewpoint. GU-RP: gesture unit labeled as rest position. RP-GU: rest position labeled as gesture unit.

Video	**Incorrect Frames**	**Transition Errors**	**GU-RP**	**RP-GU**
Experiment 1 - Data Vector 4				
A1	29 (of 511)	18 of 29	18 of 29	11 of 29
Experiment 2 - Data Vector 3				
A1/A2	142 (of 1206)	61 of 142	99* of 142	43 of 142

* Among the 99 frames, 10 were *hold*.

with velocity and acceleration in the data representation overcomes the representation built with only velocity and acceleration. Moreover, the agreement coefficient analysis shows that the MLP models are able to understand the problem as well as the human coders. The next steps in this research include: (a) verifying the performance of our solution in user and session independent approaches, since people present different gesticulation behavior in different days, and this fact must influence the segmentation models performance; (b) carrying out a new test to verify how a different human gesture analyzer evaluates the gesture unit segmentation performed by human coders and by MLP models.

6 Acknowledgments

The authors thank to Tutorial Education Program of Brazilian Education Ministry (PET/MEC).

References

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.* 34(4):555–596.

Bryll, R.; Quek, F.; and Esposito, A. 2001. Automatic hand hold detection in natural conversation. In *Proc. of the IEEE Workshop on Cues in Communication*, 1–6. IEEE.

Gebre, B.; Wittenburg, P.; and Lenkiewicz, P. 2012. Towards automatic gesture stroke detection. In *Proc. of the 8th Int. Conf. on Language Resources and Evaluation*, 231–235. Istanbul, Turkey: European Language Resources Association.

Han, J.; Kamber, M.; and Pei, J. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition.

Haykin, S. 2008. *Neural networks and learning machines*. Prentice Hall, 3 edition.

Kendon, A. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Key, M. R., ed., *The Relationship of verbal and nonverbal communication*. The Hague, The Netherlands: Mounton Publishers. 207–227.

Kita, S.; van Gijn, I.; and van der Hulst, H. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In *Gesture and Sign Language in Human-Computer Interaction*, volume 1371 of *LNCS*. Springer. 23–35.

Madeo, R.; Lima, C.; and Peres, S. 2013. Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions. In *Proc. of 28th Annual ACM Symp. on Applied Computing*, 114–121. ACM.

Madeo, R.; Wagner, P.; and Peres, S. 2013. A review of temporal aspects of hand gesture analysis applied to discourse analysis and natural conversation. *ArXiv e-prints*.

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

Martell, C. 2005. *FORM: An Experiment in the Annotation of the Kinematics of Gesture*. Ph.D. Dissertation, Univ. of Pennsylvania.

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

McNeill, D. 2005. *Gesture and Thought*. Univ. of Chicago Press.

Mitchell, T. M. 1997. *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc.

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

Ramakrishnan, A. S. 2011. Segmentation of hand gestures using motion capture data. Master's thesis, Univ. of California.

Vinciarelli, A.; Pantic, M.; Bourlard, H.; and Pentland, A. 2008. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proc. of the 16th ACM Int. Conf. on Multimedia*, 1061–1070. ACM.

Wagner, P. K.; Madeo, R. C. B.; and Peres, S. M.and Lima, C. A. M. 2013. Segmentação de unidades gestuais com multilayer perceptrons. In *Proc. of. X Enc. Nac. de Inteligência Artificial e Computacional*.

Wilson, A.; Bobick, A.; and Cassell, J. 1996. Recovering the temporal structure of natural gesture. In *Proc. of the 2nd Int. Conf. on Automatic Face and Gesture Recognition*, 66–71. IEEE.