Presentation Skills Estimation Based on Video and Kinect Data Analysis

Vanessa Echeverría, Allan Avendaño, Katherine Chiluiza,
Aníbal Vásquez and Xavier Ochoa
Escuela Superior Politécnica del Litoral
Guayaquil, Guayas, Ecuador
vecheverria@cti.espol.edu.ec, aavendan@cti.espol.edu.ec, kchilui@cti.espol.edu.ec,
anibal.vasquez@cti.espol.edu.ec, xavier@cti.espol.edu.ec

ABSTRACT

This paper identifies, by means of video and Kinect data, a set of predictors that estimate the presentation skills of 448 individual students. Two evaluation criteria were predicted: eye contact and posture and body language. Machine-learning evaluations resulted in models that predicted the performance level (good or poor) of the presenters with 68% and 63% of correctly classified instances, for eye contact and postures and body language criteria, respectively. Furthermore, the results suggest that certain features, such as arms movement and smoothness, provide high significance on predicting the level of development for presentation skills. The paper finishes with conclusions and related ideas for future work.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Design Methodology—*Feature evaluation and selection, Pattern analysis*

General Terms

Human Factors, Measurement

Keywords

Multimodal; presentation skills; video features

1. INTRODUCTION

Currently, most universities around the world face the challenge of demonstrating the quality of their graduates and whether or not they fulfill the so called twenty-first Century competences. As indicated by [23], these competences are related, among others, to interpersonal skills that comprehend abilities to communicate and collaborate.

Communicating or doing presentations to a variety of audiences is one of the professional competences sought currently by business and industries; professional organizations and undergraduate program accreditation agencies (See [1], [11]). Instructors and students work hard to get evidence that demonstrate students reach a desired level of effective communication. Evidences are constructed mostly in the interactions that take place during class time, practice sessions, etc. Precisely, these interactions are used by instructors to measure, assess and give on-time feedback about the development of such competences. However, this process is a time-demanding and complex task that needs dedication and experience on the instructor side. For instance, when instructors assess presentations, they need to be alert about several verbal and non-verbal signals that happen in parallel, including: message clarity, expressiveness quality, gaze connection to the audience, hands and arms gestures, postures shifts, etc. [28]. The use of automated ways to keep up with this process, from the instructor perspective, is desirable.

In this sense, Multimodal learning analytics (MLA) is a promising area that builds upon the analysis of a combined variety of data sources, captured during learning interactions in similar settings as the one described above. Current multimedia processing technologies and machine learning techniques have progressed to a point where readily available algorithms can be used to process videos, audios, and other digital material; and produce rich features such as postures, gestures, skeletal models that are further used to support multimodal learning analytics. The efforts in researching in this area are still limited, mainly due to the fact that is a nascent intricate area; some examples of these efforts are the ones presented in [4] [29] [9].

This paper describes the use of existing multimedia processing technologies to produce a set of features from the multimodal dataset of students' recordings while doing presentations, provided by the Presentation Quality Challenge of MLA 2014. Given the complexity of the challenge, the paper uses only the video and Kinect media features to answer the following research question: which non-verbal characteristics of students are predictors of their level of skill development when doing presentations?

The paper is structured as follows: Section 2 presents work related to the extraction of non-verbal characteristics that are useful for this work; section 3 describes the multimodal dataset, the extracted features, algorithms and software used for the analysis. Section 4, presents the techniques used for classification and estimation of the level of development of presentation skills. Section 5 discusses the findings of previous section and section 6 presents the general conclusions of the work presented.

MLA'14, November 12, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2666633.2666641.

2. RELATED WORK

Good communicators demonstrate their presentations skills by means of their verbal characteristics but also through their non-verbal characteristics, like body language, eye contact with the audience, or even the space the presenter occupies on the stage [30]. These characteristics are perceived as important as what it is literally transmitted when speaking to an audience [10]. Most of the research related to non-verbal characteristics of people is associated to emotion and affective studies to mimic movements in robots or to create fluent interactions with alike systems. The understanding of such characteristics is a key element for automatic detection systems [26], where especial emphasis is put into spatio-temporal movement properties, such as the general energy shown, rate of movement or how much an expositor is pleasant or not during a presentation [8]. The spatial aspects of movement imply measuring the joint to joint distance by using a local coordinate system [2]. Other researchers [16] [5] analyze the power and fluency of an expositor by considering the temporal aspects during her/his presentation. These analyses are based on extracting cues like velocity and acceleration for each joint with respect to a common origin joint. Additionally, [21] propose a feature analysis based on spherical angles and angular velocities for human reconstruction poses and action recognition. In [7], authors explain the characteristics describing a poor and good performance of presenters, e.g. moving around too much, open body posture, etc.

A more qualitative analysis approach is the one followed by [18] [12], which base their research on Laban's theory explaining that there is a relationship between boundaries of the space occupied by a presenter and her/his personality while giving a speech. For instance, shyness could be described as a repetitive shrink of the space occupied by the presenter's body. The contraction and expansion of this space is called contraction index [12] which is measured by calculating the eccentricity of the ellipse formed by the head, spine and hands. A presenter's gestures could describe a degree of eccentricity drawn from the body bounds. The work presented in [25] describes a framework for expressive gesture quality analysis of humans using different dimensions such spatial aspects of movement, fluency, overall activity, smoothness, etc. These dimensions are used by a virtual agent that mimics how humans behave. They also mentioned that these characteristics were calculated using the Microsoft Kinect data. In [19], it is proposed a framework for capturing manually different features based on 3D points and Kinect skeleton data from one-minute presentations of 32 presenters. The features used in this study correspond to the angles comprised between upper body joints from Kinect data and Laban Space features. Additionally, they tagged manually all postures using the evaluation of an expert trainer.

Works related to capturing features that automatically describe the postures and gestures that differentiate skilled presenters are scarce. Most of the research have been conducted to classify pre-defined postures [13] or by tagging manually such postures [3] in different contexts.

This work combines the several features mentioned above for estimating presentation skills with a multimodal approach for the analysis. Current literature addressing this type of analysis has not been found in relation to the research question to be answered.

3. DATASET

This dataset includes 448 individual oral presentations of undergraduate students and their corresponding video and Kinect records. In addition, a human-coded information evaluating each presentation is provided. The following ten evaluation criteria were used in the human evaluation: voice volume, good pronunciation, structure and idea connections, appropriate language for the audience, visual design of slide presentations, slide readability and grammar, enthusiasm and self-confidence, postures and body language; and eye contact. The scores for each criterion go from 1 (low) to 4 (high). In this paper, the two last criteria were used as variables to be predicted using machine learning techniques.

In the following subsections a match between criterion to be predicted and techniques to extract features related to such criterion is presented.

4. EXTRACTED FEATURES

In order to obtain the features that were used to predict the quality of the oral presentations, each input data (video and Kinect) was analyzed. This section describes these features and the procedure used to extract them. Following, the set non-verbal characteristics measured from the set the videos and through the Kinect sensor.

4.1 Video features

Given the importance of eye contact as a criterion used in the human evaluation provided in the dataset, basic eye contact features were extracted from the videos. Luxand [15] was used as a solution to recognize faces and to detect features. This solution returns coordinates of 66 facial feature points including both eyes center and nose tip. The facial points extracted were used to estimate the presenter's gaze. This estimation considers the smallest distance from each eye center points to the nose tip point; thus, if the distance from the right eye center to nose tip is smaller than the distance from the left eye center to nose tip, then it is inferred that the presenter is watching to the right side of the audience. However, identifying a centered vision does not mean an equal distance from any eye center to the nose tip. Therefore, a range was set to delimit a maximum displacement from perfect centered gaze, and to distinguish among other possible sides.

Using these conditions, the estimation of presenter's gaze is calculated as the average for each of the three horizontal face displacements: center (C), right (R) and left (L). Additionally, some descriptive values were computed such as maximum (MAX), minimum (MIN) and average (AVG) of such displacements. A total of nine features were calculated using these video sources and they were named using the descriptive acronyms plus H, plus the acronym for the face displacement (e.g. head center: MAXHC, MINHC, AVGHC).

4.2 Kinect features

The extraction feature procedure from Kinect data followed two approaches to represent the posture and body language of presenters. The first approach focused on identifying the common postures appearing in the set of presentations provided. The second approach, echoed the work of [25], which extracted some features from the body joints of Kinect skeleton, based on Laban's theory measures (e.g.

spatial aspects of movement, temporal aspects of movement, fluency, etc.). In addition, shape-invariant time-scale features were also extracted following similar procedures as in [21].

4.2.1 Common Postures

In the context of this work, common postures of presenters were identified by using a classification method that employed fuzzy C-Means [14] for the extraction of prototypes and their further clusterization with the K-Means algorithm. The prototype, a generalization of a posture per person, used only the upper body of Kinect joint positions: right and left elbow; right and left wrist; right, center and left shoulder; spine; and, head. Additionally, only tracked positions from Kinect were used to avoid the generation of unusual postures or noise for the final classification. All skeleton joints were referenced to the spine joint to eliminate the spatial movement given by the local position of the Kinect camera and the presenter. Then, for each individual Kinect recording, all X and Y points were converted to a coordinate system based on the spine joint as the origin. It was not necessary to normalize the data because the prototypes were extracted for the presenter itself.

After processing the Kinect recordings, a fuzzy C-Means (FCM) algorithm with highest membership categorization was used to build prototypes and to extract five different postures per student. Five prototypes were enough to describe comprehensively the majority of postures each person in the dataset showed during a presentation. After applying the FCM, all the prototypes from all presenters were clustered using K-Means. A normalization over the data was accomplished by extracting the angles of the triangles per limb, which is composed by the arm joints (shoulder, elbow and wrist), and the limb orientation towards the shoulder. The relative data was used in each prototype and a new feature vector, representing each prototype, was calculated. The definition of the vector $X = (\alpha_L, \beta_L, \gamma_L, \alpha_R, \beta_R, \gamma_R, O_L, O_R)$ is defined as follows:

- α_L is the angle between the vectors defined from: the wrist to the shoulder, and the shoulder to the elbow in the left arm.
- β_L is the angle between the vectors defined from: the shoulder to the elbow, and the elbow to the wrist in the left arm.
- γ_L is the angle between the vectors defined from: the elbow to the wrist and the wrist and the shoulder in the left hand.
- α_R is the angle between the vectors defined from: the hand to the shoulder, and the shoulder to the elbow in the right arm.
- β_R is the angle between the vectors defined from: the shoulder to the elbow, and the elbow to the wrist in the right arm.
- γ_R is the angle between the vectors defined from: the elbow to the wrist and the wrist and the shoulder in the right wrist.
- O_L is the left arm orientation reference to the shoulder. It takes a value of one if it is located above of the reference and a minus one value, if it is located below of the reference.
- $O_R$is the right arm orientation reference to the shoulder. The assignation of values for O_R follows the same logic as in O_L.

The K-Means clusterization results in 24 postures and its corresponding centroids. The 24 postures were reduced to six postures after analyzing the actual postures appearing in the videos. The six postures that were finally used are listed below:

- arms down (AD)
- explaining with closed hands (EXPCH)
- pointing to presentation with one hand (PTPONEH)
- explaining with hands slightly separated (EXPHSS)
- explaining with one arm up (EXPONEAUP)
- pointing to presentation with two arms (PTPTWOA).

Each frame was classified according to this new set of postures and the percentage of each posture per student was calculated. Figures 1, 2 and 3 show video frames and its corresponding categorization.

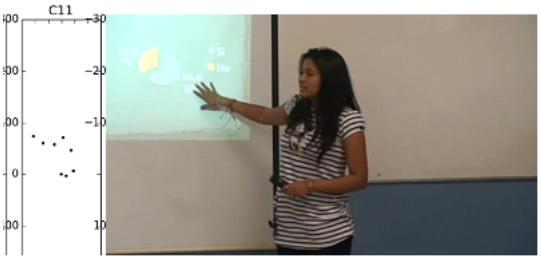

Figure 1: *Video frame categorized as C11 after K-Means was performed. The final classification is* ***pointing to presentation with one hand****.*

Figure 2: *Video frame categorized as C15 after K-Means was performed. The final classification is* ***explaining with hands slightly separated****.*

4.2.2 Extracted features based on Laban's Theory and shape-invariant time-scale

For the following features, the upper body set of 9 joints from Kinect records were used. These joints are both wrists, elbows, shoulders; and, hip center, spine, and shoulder center. For accuracy purposes, wrist features were used instead

Figure 3: *Video frame categorized as C23 after K-Means was performed. The final classification is* ***explaining with closed hands****.*

of hand features because the former were easily tracked. However, in the rest of the paper wrist features will be referred as hand features.

The following non-verbal characteristics were used in the feature extraction: spatial aspects of movement; temporal aspects of movement; fluency, smoothness and impulsivity; energy and power and overall activity. The detailed description of each characteristic and its associated extracting technique are described below. From this point on, the extracted features referring to limbs are named as right (R) or left (L); likewise, descriptive values for the measures of non-verbal characteristics such as standard deviation (STD), maximum (MAX), minimum (MIN), average (AVG) and skewness (SK) are used as prefixes of the calculated measures.

Spatial aspects of movement.

This set of features is calculated considering a local system of coordinates, centered at the spine joint per frame. From this joint the Euclidean distances (D) between each hand (H) and elbow (E) were computed, as in [6], resulting for instance in MAXLHD (Maximum Euclidean distance between left hand and spine). Similarly, descriptive values of distances between hands are calculated (i.e.MAXD2H,SKD2H) as in [2] and the average of frames where open (O) or closed (C) hands were computed using as reference a fixed threshold (AVGOH, AVGCH).

Finally, a contraction index (CI) [12], explained at section 2 and its descriptive values are calculated (i.e. AVGCI, MAXCI, etc.).

Temporal aspects of movement.

Another set of features is based on the movement execution along the time, suggesting the sense of power during a presentation [16]. Consequently, the first derivative of the distance from a common origin (spine joint) up to left and right wrist and elbow joints was calculated, resulting in the velocity (V) from consecutive frames (i.e. MAXRHV stands for the maximum right hand velocity).

Fluency, smoothness and impulsivity.

According to [6], fluency was measured by obtaining the sum of variance (SV) of the norms of the motion vectors. A motion vector is estimated by obtaining the norm of the average velocity per joint and per second. According to [18] the following features that are also related to fluency were computed: the area covered by presenter's hands, for one (1F) and fifteen (15F) frames as window sample size (e.g. AVG1F, MAXG1F, etc.). Additionally, the majority (MJ) of area values that lies out of a confidence interval for each sampling window were calculated (e.g. MJ15F). Finally, smoothness (SMTH) and impulsivity (IMP), which correlate to slow and fast wrists movements during a short period, respectively, were calculated as described in [17].

Energy and power.

Following the methodology described in [2], energy and power, from body movement, were estimated by calculating the second derivative of distance with respect to left and right hand and elbow joints. The average, maximum and minimum acceleration (A) were obtained from these joints (i.e. MAXRHA stands for maximum right hand acceleration).

Overall activity.

As described in [6], the overall activity is a characteristic that was measured by following two calculations. The first one calculates the sum of motion vectors (SMV). The second one calculates the quantity of motion by taking the average of the motion vectors (i.e. AVGMV relates the average of the norm of the motion vectors).

Shape-invariant time-scale features.

As stated by [21], original positions (x,y,z coordinates) of joints can be converted into spherical coordinates, resulting in shape-invariant time-scale data. This conversion is useful to avoid differences related to individual height of the presenter. Thus, hand and elbow joints were converted into spherical angles (ϕ and θ) to denote spherical positions. The angular velocity (ϕ_ω and θ_ω) was also calculated to consider the movement over the time of such angles. For instance, the minimum angle of the left elbow is referred as MINLEϕ and its angular velocity as MINLEϕ_ω.

5. EVALUATION AND RESULTS

5.1 Evaluation methodology

This section presents the evaluation of the applied machine learning techniques used for calculating the predictors scores features obtained with the evaluation criteria described in section 4. A basic setup was established previous to the evaluation procedure, which consist of the following steps: data normalization, feature selection, and classification using a machine learning technique that better fits the data. The classification uses a training set sample and ultimately a classification report is generated from a predicted set of values using a sample test set. The basic setup procedure was carried out using Python and the Scikit-learn library [22].

In the context of this work, the feature selection was applied by using a recursive feature elimination (RFE) via cross-validation, accuracy scores and a weighted logistic regression classifier. This specific setup was chosen after evaluating alternate setups and its results. Once the ranking of features was returned by RFE, a selected set of features were used for training the classifier using cross validation. The predicted values were matched with the ground truth building the confusion matrix and generating values such as average accuracy, precision, recall and the Receiver Operating Characteristic (ROC) value from confusion matrix.

Two evaluation approaches were conducted in this part of the study. The first approach analyzed the video features using the eye contact criterion (EC) and the second approach analyzed the Kinect features using the body language criterion (PBL). Both EC and PBL are evaluation criteria used by human coders. Table 1 shows the data source and the human coded criterion for the evaluation approaches.

The scores obtained using human coded criterion EC and PBL, were converted into a nominal scale with two possible classes: Poor and good application of both criterion. The class values and range of the original scores for each criterion are also presented in Table 1.

Data source	Human coded criterion	Class	Range of values
Video	Eye Contact (EC)	Poor	$0 < x \leq 2.67$
		Good	$2.67 < x \leq 4$
Kinect	Body and Posture Language (BPL)	Poor	$1 < x \leq 3$
		Good	$3 < x \leq 4$

Table 1: *Evaluation Approaches with their corresponding data source, initial number of features, related criterion, class and range of values per class.*

5.2 Results

5.2.1 Video features approach

Following the setup previously explained, the video features were normalized and then the RFE was performed using the accuracy as key point for recursive elimination, five-fold cross-validation and a weighted logistic regression classifier. The RFE resulted in the selection of nine features.

Next, a weighted logistic regression classifier was trained and tested using ten-fold cross-validation, resulting in a built classifier with a general accuracy of 0.68 and a ROC score of 0.65. Table 2 shows the nine most relevant features for predicting EC. Note that AVGHL (Average of presenter's head pointing to the left side) has the highest coefficient for the logistic decision function. In addition, as can be noted in Table 5, 61% of the cases that fall in the poor class are correctly classified, whereas, 68% of the good cases are classified accordingly.

Rank	Feature	Coefficient
1	AVGHL	150.73
2	MINHL	18.70
3	MINHC	16.04
4	MINHR	4.73
5	AVGHC	3.84
6	MAXHR	1.57
7	MAXHL	-5.44
8	MAXHC	-5.57
9	AVGHR	-50.12

Table 2: *Video ranking coefficients of the features in the logistic regression decision function.*

5.2.2 Kinect features approach

115 features from Kinect were used to perform the RFE, a similar process as described in the first evaluation approach was followed. The results of RFE did not show relevant features for the classification procedure. Therefore, another well-known procedure for selecting features is based on the Information Gain. Then, a tree classifier was built and the gain information for each feature was calculated. Features with an information gain higher than 0.0087 were selected. Table 3 shows the joint types, measures obtained from related joints and the name of selected features according to the description in section 4.2 . From this table, it is evident that most of the average related and spherical angles features were selected.

The feature space was reduced to 53 and a weighted logistic regression classifier was built upon these selected features. After training and testing the classifier using ten-fold cross-validation, the general accuracy resulting was 0.63 and a ROC score of 0.62. Note that in Table 4 the highest coefficient appearing in this table, for the logistic regression decision function, corresponds to the average right hand theta angle (AVGRHθ) and the minimum is related to the average right hand acceleration (AVGRHA). Table 5 shows the precision and recall for this classifier per category. Results indicate that, for this case, the percentage of correct classification for the poor category increased, and it decreased for the good category.

6. DISCUSSION

Results presented in previous section showed the existence of good predictors for grading a presenters' performance, according to eye contact and body posture language scores, assigned by experts.

6.1 Eye contact

Even though the gaze related features were calculated using simple techniques, the accuracy reached in the model is in a good level of acceptance.

These results show that the feature AVGHL, average of presenter's head pointing to the left side, was the best predictor for a good performance class. This can be explained with a brief inspection into the video dataset, which reveals that most (56%) of the presenters were located at the left hand side of the slide presentation, meaning that she or he needed to gaze to the left side of the audience.

In contrast, the average of presenter's head pointing to the right side (AVGHR) was the worst predictor for the good performance class. It can be observed in table 2 that this feature has a negative coefficient. This behavior may be attributed to the presenter's position. That is, if the presenter turns her or his head to the right side, where it is supposed there is no audience, the AVGHR value would increase, then a poor performance would be predicted.

One would expect that a combined balance between the three head positions (Left, center, right) would be relevant, for maintaining good eye contact with the audience. However, results do not correspond to this expectation. The prediction model in this study is limited by the way the videos were captured. For example, the position of the teacher in the classroom was not provided in the dataset and the video setting, while capturing the data, was not always the same. Therefore, these aspects should be considered when gathering video data for building prediction models as the one portrayed in this article.

Joint types	Measures	Selected features
left hand	ϕ and θ ϕ_ω and θ_ω	MINLHϕ, AVGLHϕ, MAXLHϕ, AVGLHθ, STDLHθ, MAXLHθ, AVGLHϕ_ω, MAXLHθ_ω
	velocity and acceleration	AVGLHV, AVGLHA
right hand	ϕ and θ ϕ_ω and θ_ω	AVGRHϕ, STDRHϕ, MAXRHϕ, AVGRHθ, AVGRHϕ_ω, STDRHθ_ω, MAXRHθ_ω
	velocity and acceleration	MAXRHV, AVGRHA, MAXRHA
left elbow	ϕ and θ ϕ_ω and θ_ω	MINLEϕ, AVGLEϕ, MAXLEϕ, STDLEθ, MAXLEθ, MINLEθ, AVGREϕ_ω, STDLEθ_ω, MAXLEθ_ω
	velocity	AVGLEV
right elbow	ϕ and θ ϕ_ω and θ_ω	MINREϕ, AVGREϕ, MAXREϕ, MINREθ, AVGREθ, AVGREϕ_ω, MAXREϕ_ω, MINREθ_ω, STDREθ_ω
	acceleration	AVGREA
left hand, right hand	distance between two hands	SKD2H, AVGOH, AVGCH
	smoothness	SMTH
head, left hand, right hand	area covered by hands	AVG1F, STD1F, AVG15F, STD15F
upper body	Rigid stance, Open body posture, Hand/arm gestures to emphasize point	AD, PTPONEH, EXPONEAUP, PTPTWOA

Table 3: *Selected features after performing the Information Gain feature selection from Kinect dataset to predict body and posture language criterion.*

Rank	Feature	Coefficient
1	AVGRHθ	338.27
2	STDLEθ	207.74
3	MAXLEθ_ω	105.63
4	MAXLE θ	92.52
5	STDRHθ	84.46
6	MAXLEϕ	83.67
7	MINREϕ	73.71
8	MAXREϕ_ω	64.17
9	MINLHϕ	61.86
10	AVGRHA	51.44

Table 4: *Top ten Kinect ranking coefficients of the features in the logistic regression decision function for predicting bod*

	Class	Precision	Recall	Examples
Video features	Poor	0.48	0.61	147
	Good	0.78	0.68	301
Kinect features	Poor	0.66	0.70	260
	Good	0.55	0.50	188

Table 5: *Summary of the results per evaluation approach.*

6.2 Body and Posture Language

The classification procedure revealed, as it is shown in table 3, that the shape-invariant time-scale features are good predictors for the classifier. Moreover, in table 4 these features occupy nine of the top ten ranking coefficients that predict body and posture language. Thereby, using spherical angles to give a better precision of the data, before applying any mathematical formulation, was a good strategy to overcome issues related to the height of presenters.

Similarly, the average of open and closed hands (AVGOH, AVGCH), the skewness between two hands (SKD2H), and the area covered by hands (AVG1F, AVG15F) were significant predictors of the model. These predictors could be related to an adequate movement of hands; thus, they might be perceived by humans as good indicators of PBL. These findings go in line with [7], which stated that some affective postures, where upper limbs appear, are perceived as positive when communicating with other people. Conversely, static body postures are linked to negative basic emotions, such as sadness or anger, apparently regardless of the cultural context (See [20]

Interestingly, the smoothness of the presenter's movements (SMTH) was also a predictor in the classifier. This could be interpreted as if a presenter makes abrupt movements or moves around rigidly, it might be perceived as an undesirable characteristic during a presentation. In [27] [16] [5] [24] smoothness was identified as a good characteristic of performance scoring while dancing or for sentiment detection; however, no studies were found, where such feature is related to good presentation skills.

Moreover, the features extracted from common postures such as arms down (AD), pointing to presentation with one hand (PTPOH), explaining with one hand up (EXPONEAUP) and pointing to presentation with two arms (PTPTWOA) were also selected as predictor characteristics, which agree with the study of [7], where certain set of postures for pre-

sentations were identified. There were missing postures like explaining with hands slightly separated (EXPHSS) and explaining with closed hands (EXPCH) that were not selected as predictors of the model. Future work could analyze individually each of the six extracted postures to study their relation to good or bad presentation skills.

Finally, other set of features that were likely to relate to presentation skills, such as those linked to the overall activity characteristic, were less informative for the purpose of this research.

7. CONCLUSIONS

This paper aimed to answer the following question: which non-verbal characteristics of students are good predictors of their level of development of presentation skills?

It can be concluded that the measures related to eye contact; arm movement; smoothness and fluency in the stage, while communicating; and, a set of body postures that helps emphasizing points, of what it is uttered, are good estimators of level of development in presentations skills.

Nevertheless, if other dataset were used, as source, to create a similar model, additional predictors might appear. Especially if other video sources like eye tracking is included in the analysis. Additionally, a fixed setting for recording the presentations could improve the accuracy of the predicted models.

As a contribution from this study, analyzing each presenter created a category of common postures. Such categories could leave an open door for this research area due to the lack of defined postures for presentations. An improvement in the methodology for creating these categories would be the use of shape-invariant time-scale features before the clusterization (spherical angles of points). The use of these new space and dimension could generalize, in a better way, the categories.

Further research could improve the prediction of presenter's performance by matching the posture with the presenter's speech and the content of the slide presentation.

As a final conclusion, the results of this work could be used to improve the performance of automatic estimation tools. These tools could provide feedback to the students beforehand their actual presentations, as well as to alleviate the load on the side of instructors when assessing them.

8. ACKNOWLEDGMENTS

The authors want to acknowledge the support of the VLIR-UOS project ZEIN2010RIP09 and the SENESCYT Project "Andamios".

9. REFERENCES

[1] Abet. Accreditation policy and procedure manual 2013 - 2014, 2014.

[2] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In A. Paiva, R. Prada, and R. Picard, editors, *Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science*, pages 59–70. Springer Berlin Heidelberg, 2007.

[3] S. Bhattacharya, B. Czejdo, and N. Perez. Gesture classification with machine learning using kinect sensor data. In *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, pages 348–351, Nov 2012.

[4] P. Blikstein. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 102–106. ACM, 2013.

[5] M. Cakmak, S. S. Srinivasa, M. K. Lee, S. Kiesler, and J. Forlizzi. Using spatial and temporal contrast for fluent robot-human hand-overs. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 489–496. ACM, 2011.

[6] G. Caridakis, A. Raouzaiou, E. Bevacqua, M. Mancini, K. Karpouzis, L. Malatesta, and C. Pelachaud. Virtual agent multimodal mimicry of humans. *Language Resources and Evaluation*, 41(3-4):367–388, 2007.

[7] M. Cavanagh, M. Bower, R. Moloney, and N. Sweller. The effect over time of a video-based reflection system on preservice teachers' oral presentations. *Australian Journal of Teacher Education*, 39(6):1, 2014.

[8] S. A. Etemad. *Perceptually Guided Processing of Style and Affect in Human Motion for Multimedia Applications*. PhD thesis, Carleton University, 2014.

[9] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester. Modeling student programming with multimodal learning analytics. In *Proceeding of the 44th ACM technical symposium on Computer science education*, pages 736–736. ACM, 2013.

[10] L. Hsu. The impact of perceived teachers nonverbal immediacy on students motivation for learning english. *Asian EFL Journal*, 12(4):p188–204, 2010.

[11] A. f. C. M. A. Joint Task Force on Computing Curricula and I. C. Society. *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*. ACM, New York, NY, USA, 2013. 999133.

[12] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on*, 4(1):15–33, 2013.

[13] T.-L. Le, M.-Q. Nguyen, and T.-T.-M. Nguyen. Human posture recognition using human skeleton provided by kinect. In *Computing, Management and Telecommunications (ComManTel), 2013 International Conference on*, pages 340–345, Jan 2013.

[14] M.-J. Lesot, L. Mouillet, and B. Bouchon-Meunier. Fuzzy prototypes based on typicality degrees. In B. Reusch, editor, *Computational Intelligence, Theory and Applications*, volume 33 of *Advances in Soft Computing*, pages 125–138. Springer Berlin Heidelberg, 2005.

[15] Luxand. Luxand - face recognition, face detection and facial feature detection technologies, 2014.

[16] M. Mancini and G. Castellano. Real-time analysis and synthesis of emotional gesture expressivity, 2007.

[17] B. Mazzarino and M. Mancini. The need for impulsivity & smoothness - improving hci by qualitatively measuring new high-level human motion features. In *SIGMAP*, pages 62–67, 2009.

[18] J. Newlove and J. Dalby. *Laban for All*. A Nick Hern book. Nick Hern, 2004.

[19] A.-T. Nguyen, W. Chen, and M. Rauterberg. Online feedback system for public speakers. In *E-Learning, E-Management and E-Services (IS3e), 2012 IEEE Symposium on*, pages 1–5, Oct 2012.

[20] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: A survey. In T. Huang, A. Nijholt, M. Pantic, and A. Pentland, editors, *Artificial Intelligence for Human Computing*, volume 4451 of *Lecture Notes in Computer Science*, pages 47–71. Springer Berlin Heidelberg, 2007.

[21] G. Papadopoulos, A. Axenopoulos, and P. Daras. Real-time skeleton-tracking-based human action recognition using kinect data. In C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, editors, *MultiMedia Modeling*, volume 8325 of *Lecture Notes in Computer Science*, pages 473–483. Springer International Publishing, 2014.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] J. W. Pellegrino, M. L. Hilton, et al. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press, 2013.

[24] S. Piana, M. Mancini, A. Camurri, G. Varni, and G. Volpe. Automated analysis of non-verbal expressive gesture. In T. Bosse, D. J. Cook, M. Neerincx, and F. Sadri, editors, *Human Aspects in Ambient Intelligence*, volume 8 of *Atlantis Ambient and Pervasive Intelligence*, pages 41–54. Atlantis Press, 2013.

[25] S. P. R. Niewiadomski, M. Mancini. Human and virtual agent expressive gesture quality analysis and synthesis. *Coverbal Synchrony in Human-Machine Interaction, CRC Press*, 2013.

[26] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer. Building autonomous sensitive artificial listeners. *Affective Computing, IEEE Transactions on*, 3(2):165–183, April 2012.

[27] D. Tardieu, X. Siebert, B. Mazzarino, R. Chessini, J. Dubois, S. Dupont, G. Varni, and A. Visentin. Browsing a dance video collection: dance analysis and interface design. *Journal on Multimodal User Interfaces*, 4(1):37–46, 2010.

[28] L. Wilbanks. Are you communicating? *IT Professional*, 14(5):0060–61, 2012.

[29] M. Worsley. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 353–356. ACM, 2012.

[30] D. York. *Investigating a Relationship between Nonverbal Communication and Student Learning*. PhD thesis, Lindenwood University, 2013.