Quantitative analysis of non-verbal communication for competence analysis

Alvaro CEPERO [a] , Albert CLAPÉS [a,b] and Sergio ESCALERA [a,b]
[a] *Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona*
[b] *Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona*
E-mail: acepero13@gmail.com, aclapes@cvc.uab.cat, sergio@maia.ub.es

Abstract. Oral communication competence is defined on the top of relevant skills for professional and personal life. Because of the importance of communication in our daily activities it is crucial to study methods to improve our communication capability and therefore learn how to express ourselves better. In this paper, we propose a multi-modal RGB, depth, and audio data description and fusion approach in order to recognize behavioral cues and train classifiers able to predict the quality of oral presentations. The system is tested on real defenses from Bachelor's thesis presentations and presentations from an 8th semester Bachelor's class at Universitat de Barcelona. Using as ground truth the scores assigned by the teachers, our system achieved high classification rates categorizing and ranking the quality of presentations into different groups.

Keywords. Social Signal Processing, Multi-modal description and data fusion, Non-verbal communication, e-Learning

1. Introduction

Nowadays society is demanding new kind of competencies of its citizens and especially its professionals. With the implementation of the bachelor's degree in the European Higher Education Area, the concept of competencies became even more important in the educational field. One of the main goals of this plan is to provide specific skills and competencies to the student. Oral expression and communication is among the most relevant competencies in everyone's life. A nationwide survey conducted in 1988 by the American Society of Training and Development and the Department of Labor found that oral communication skills were ranked within the top five skills required of potential hires. However, an article in the September 2005 issue of the Hiragana Times states that "the generation raised by the one-way information provided on TV is poor in communication". Given the importance of communication in our daily life and the difficulty on the current society to train this competence, it is crucial to study methods to improve our communication skills and therefore learn how to express ourselves better.

In this context, Social Signal Processing is the field of study that analyzes communication signals by means of different sensors, such as microphones or cameras, and applies some kind of pattern recognition strategies. Some examples of application are social interaction analysis on small groups. According to the authors of [1] psychologists have grouped all possible non-verbal behavioral cues into five major classes, though most recent approaches are based on three of them: ***gestures and postures*** (considered as the

most reliable feature about people's attitude towards others), ***face and eye behavior*** and ***vocal behavior***. The two former kinds of communication can be evaluated as successful or not given certain criteria: knowledge transmission, richness of the language expression, discourse coherence and cohesion, and so forth. Whereas the verbal communication is often quite explicit to a human observer, non-verbal signals are relatively subtle regardless of the huge amount of information they provide about the communicating subject and are not always easily separable, so as to be clearly identified and evaluated individually, but they emerge as a whole and complex behavior. There is a vast literature in psychology studying the non-verbal component in communication act, but from the point of view of Artificial Intelligence (AI) we are at the beginning of a long way to go. While the verbal communication has been studied for many years by the Natural Language Processing field, the study of the non-verbal communication from a multimodal social signal point of view, including visual sources, is a relatively recent field of research.

Several works have been recently performed in the Social Signal Processing field in order to analyze the non-verbal communication of subjects in group interactions [13]. Most of these works focus on audio analysis, and main goals are based on dominance, influence, and leadership recognition. In the work of [2] the authors present a design and implementation of a platform for measuring and analyzing human behavior in organizational face-to-face settings using wearable electronic badges. The work of [5] presents the recognition of group actions in meetings using a multi-modal approach. Other recent approaches for dominance analysis in group interactions have been also proposed [3,4,8]. The work of [14] presents a bayesian framework that models dominance skills based on audio input sources. In [15], the authors model a multi-modal audio-video system to recognize leadership in group interactions. The system is defined based on simple multimodal features under controlled face-to-face interaction environments. In a similar scenario, the proposal of [16] defines multi-modal cues for the analysis of communication skills in a upper body setup. A more general purpose approach is presented in [17], where prosodic features are computed to define a set of relevant traits of subjects in oral communication settings. Very few works have been reported on the analysis of non-verbal communication as a competence skill in e-Learning scenarios. The authors of [18] presents a system based on audio analysis from mobile devices to analyze the communicative skills and provide relevant feedback to subjects that may suffer from communication problems and some degree of autism.

In this paper, we propose a multi-modal Audio-RGB-Depth system for oral expression and communication analysis. The system is based on capturing audio features, head pose, motion, and behavior patterns from RGB-Depth data, defining a set of high-level behavioral indicators and quantifying the level of presentation quality using different binary, multi-class, and ranking state-of-the-art statistical classifiers. Quantitative results on a novel multi-modal data set of university student defenses show accurate measurements of the proposed system and its reliability to be used in the training routine of non-verbal communication competence.

The rest of the paper is organized as follows. Section 2 presents the multi-modal audio-RGB-depth system for communication competence analysis. Section 3 evaluates the proposed methodology on a novel data set of student defenses. Finally, Section 4 concludes the paper and discusses future lines of research.

2. System

Our non-verbal communication framework for competence analysis is focused on the feature extraction and data fusion of different modalities, including RGB, depth, and audio, in order to recognize gestures, postures, and audio behavior-based cues. For this task, we separate the process in two different parts: first, the extraction of low-level features from the Audio-RGB-Depth data which defines the input audio source, face tracking system, and the skeletal body model; and second, the processing of these low-level features into high-level features to build the characteristics that codifies the subject's behavior. This indicators are then used to train strong classifiers able to predict the quality of the presentation given a ground truth defined by experts (teachers in this scenario). The different modules of our system are shown in Figure 1 and described next.

2.1. Low-level features

In order to analyze the behavior of the user towards the audience, we defined two sets of features: low and high-level. The low-level features are those characteristics that are extracted directly from the Kinect^{TM} SDK API, that is the RGB data, Depth and the raw audio. The way Kinect^{TM} provides multi-modal data relies on three hardware innovations working together:

• **A color VGA video camera**. It provides information to help on the facial recognition process by detecting three color components: red, green and blue. This component acts as a regular camera and it captures images around 30 frames per second, and projects at a 640×480 pixels resolution.

• **Infrared Sensor**. The Kinect^{TM} infrared sensor projects a pattern of infrared dots known as *structured light* to the scene. Then, each depth pixel is computed by sampling the derivative of the higher resolution infrared image taken in the infrared camera. This value is inversely proportional to the radius of each gaussian dot, which is linearly proportional to the actual depth.

• **Audio acquisition**. The Kinect^{TM} features a multi-array microphone that consists of four separate microphones spread out linearly at the bottom of the Kinect^{TM}, with each channel processing 16-bit audio at a sampling rate of 16 kHz. By comparing when each microphone captures the same audio signal, the microphone array can be used to determine the direction from which the signal is coming. In our system, the distance between the speaker and the Kinect^{TM} is less than 3.5 meters.

2.1.1. RGB-Depth features

We use the color or RGB data to perform face detection and facial description based on Viola & Jones face detection [12]. We combine this information with the depth information along with the RGB data of the detected face, providing a depth and spatial facial description of 121 representative landmarks of the human face.

We use the depth information to perform a skeleton tracking and to build a skeletal model. This model will yield the world coordinates of the user in real time. The Kinect^{TM} system defines 20 key points to determine a human skeleton. In our system we will focus only on the spatial position coordinates of hands, wrists, arms, elbow hip, shoulders, and head. The method to define the skeletal model is based on a previous limb-segmentation through Random Forest (RF) [11]. This process is performed computing random offsets of depth features as follows:

$$f_\theta(\mathbf{D},\mathbf{x}) = \mathbf{D}_{\left(\mathbf{x}+\frac{\mathbf{u}}{\mathbf{D_x}}\right)} - \mathbf{D}_{\left(\mathbf{x}+\frac{\mathbf{v}}{\mathbf{D_x}}\right)}, \tag{1}$$

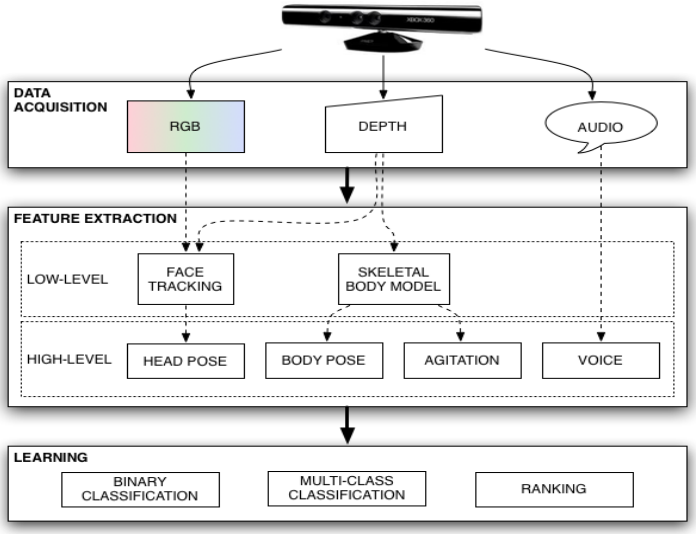

Figure 1. System modules for non-verbal communication analysis.

where $\theta = (\mathbf{u}, \mathbf{v})$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ is a pair of offsets, depth invariant. Thus, each θ determines two new pixels relative to $\mathbf{x}$, the depth difference of which accounts for the value of $f_\theta(\mathbf{D}, \mathbf{x})$. Using this set of random depth features, Random Forest is trained for a set of trees, where each tree consists of split and leaf nodes (the root is also a split node). Finally, we obtain a final pixel probability of body part membership l_i as follows:

$$P(l_i|\mathbf{D}, \mathbf{x}) = \frac{1}{\tau} \sum_{j=1}^{\tau} P_j(l_i|\mathbf{D}, \mathbf{x}), \tag{2}$$

where $P(l_i|\mathbf{D}, \mathbf{x})$ is the PDF stored at the leaf, reached by the pixel for classification $(\mathbf{D}, \mathbf{x})$ and traced through the tree j, $j \in \tau$.

Once this procedure is applied, we have positive probabilities for those pixels belonging to each limb of the human body. Then, mean shift is used to estimate human joints and representing the body in skeletal form.

2.1.2. Audio features

From the the raw audio obtained from the Kinect^{TM} we compute three types of low-level features per frame [6]. The first feature is the widely used short-term energy. Energy is the most common feature for speech/silence detection. The second feature is Spectral Flatness Measure. Spectral Flatness is a measure of the noisiness of spectrum. The third feature is the most dominant frequency component of the speech frame spectrum, which can be very useful in discriminating between speech and silence frames. These low-level features will be used later to compute the 'speaking' high-level feature indicator.

2.2. High-level features

The high-level features or meta-characteristics are built from the low-level features described in previous section in order to define speaker communication indicators. The set of behavioral indicators we considered in our framework are described next:

1. Facing towards: The average number of frames the user is looking at the tribunal/public. In order to analyze whether the user is looking at the tribunal or not we use the face detection, and if it is found in frontal view we use the implementation of the face tracking system provided by the Microsoft SDK to compute the nose's vector direction. We consider that the user is looking at the public if the angle formed between the nose and within an approximately 30 degrees range.

2. Crossed arms: The average number of frames in which the user is with his/her arms crossed. In order to determine if arms are crossed, the x coordinate of the right hand must be lower than the x coordinate of the spine, besides the x coordinate of the left hand must be greater than the spine, and finally the difference between the right hand x coordinate and the left hand x coordinate must be greater than the half of the forearm's length.

3. Pointing: The average time the user is pointing towards the blackboard. In order to know whether the user is pointing or not, firstly we discard those situations where the hand is closer to the body than the elbow. Then the distance between the hand and the hip is computed and divided by the forearm's length. Moreover, in order to avoid situations where the user seems to be pointing to the tribunal, we divide this distance by the difference in z-axis of both hand and hip, and finally we normalize by finding the inverse of this division. We found that values ranging from 0.0039 to 1 indicates that the user is pointing the blackboard with high precision.

4. Speaking: The average time the user is speaking. Once low-level short-term energy, Spectral Flatness, and most dominant frequency component features have been computed, we use the implementation of the VAD [1] algorithm [6] for speaking recognition. We use the three sets of low-level features previously described and take 30 frames for threshold initialization. For each incoming speech frame the three features are computed. The audio frame is marked as a speech frame if more than one of the features values fall over the pre-computed threshold. We consider that there is voice activity only if there are 5 or more successive frames marked as speech.

5. Upper agitation: The average of the displacement in real coordinates of arms, wrist, and hands while hands are above the head. Namely if the user left hand or right hand is above his/her head then the magnitude is computed as the difference between frames of the distance from the wrist, hand or arm point to the hip point (taken as a reference point).

6. Middle agitation: The average of the displacement in real coordinates of arms, wrist and hands while hands are below the head and above the hip. Namely if the user left hand or right hand is between his/her head and his/her hip then the magnitude is computed as the difference between frames of the distance from the wrist, hand or arm point to the hip point (taken as a reference point).

7. Bottom agitation: The average of the displacement in real coordinates of arms, wrist and hands while hands are below the hip. Namely if the user left hand or right hand is below his/her hip then the magnitude is computed as the difference between frames of the distance from the wrist, hand or arm point to the hip point (taken as a reference point).

8. Agitation while speaking: The average of the displacement in real coordinates of arms, wrist and hands while the user is speaking combining with the response of speaking indicador.

9. Agitation while not speaking: The average of the displacement in real coordinates of arms, wrist and hands while the user is not speaking combining with the response of speaking indicador.

Some examples of the detected low and high-level features are shown in Figure 2. Once the multi-modal high-level behavioral indicators have been automatically computed, we assign the feature vector of nine values to each student presentation. Then, the score assigned by the teacher is stored as the ground truth for that particular data sample. In the next section, before the presentation of the experimental results, we describe the novel data set we recorded, and describe the different statistical classifiers we

[1] https://github.com/shriphani/Listener/blob/master/VAD.py

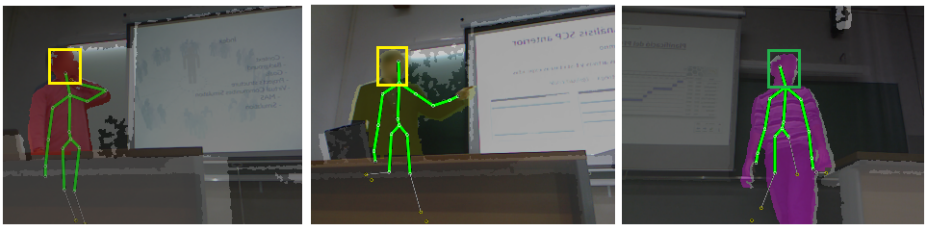

Figure 2. Examples of low-level features extraction. Depth and RGB aligned maps are superimposed with transparency. Color in the human body indicates user detection at pixel level using Random Forest. In this case, different colors indicate different user identifiers. Detected skeletons are drawn in green. Detected faces are also marked. Yellow color of faces indicates that the speaker is not looking at the tribunal/public and green marked face indicates that the speaker is facing towards the tribunal.

considered to validate our framework, mainly Adaboost and Support Vector Machines classifiers, which are used for binary classification of two groups of quality presentations, multi-class classification into several groups, analysis of feature selection of more relevant indicators, and finally, learning to rank presentations based on quality.

3. Results

In order to present the results, we first describe the data, settings, and evaluation measurements of the performed experiments.

3.1. Data

The analyzed data consists on 24 recorded videos of 13 Bachelor's Thesis presentations and 11 presentations from an 8th semester Bachelor's class at Universitat de Barcelona. All the videos were recorded with a Kinect^{TM} device at 14 FPS. The videos were recorded on three different classrooms. All the videos were recorded with the user facing the tribunal. The details of the data set are shown in Table 1. For each presentation the vector of nine high-level communication features is computed and the score assigned by the teacher regarding the presentation quality is stored as the ground truth. Some examples of the recorded scenarios are shown in Figure 3.

Table 1. Details of the data set

Total videos	Final projects	Class projects	Total frames
24	13	11	150000

3.2. Settings

In order to train the the multi-modal features to be able to classify the quality of the presentations we use different classifiers. Specifically we selected Gentle Adaboost classifier [7] with decision stumps, Support Vector Machines with Radial Basis Function kernel (binary and one-versus-one multi-class) from *LibSVM* [9] and Ranking Support Vector Machines [10]. Adaboost is used in two ways, first to obtain a classifier which is able to separate between two differentiated groups: "good" presentations and "bad" presentations, and also as a feature selection method in order to analyze the relevance of each high-level communication indicator. We also analyzed the weight assigned to the features in the case of SVM to analyze the most relevant indicators by this classifier. Moreover, SVM classifier is tested in three scenarios: binary classification, multi-class classification, and ranking. The parameters used are summarized in Table 2.

Figure 3. Some examples of the presentations of our data set.

Table 2. Methods and parameters

Method	No. Examples	Parameters
Adaboost	24	No. of max. iterations (decision stumps): 50
LibSVM	24	C=128 ; γ=0.5; Radial Basis Function

3.3. Validation and measurements

In order to measure the generalization capability of our system we perform a leave-one-out validation model: a single observation (presentation) from the original sample is taken as the test data, and the remaining observations as the training data. This process is repeated as many times as observations we have, and the average number of hits is stored. This measurement is applied for binary and multi-class classification for different degrees of quality defined for the presentations. In the case of ranking SVM, the error in the prediction is calculated as the ratio between by how many positions did the classifier failed predicting the correct position and the maximum number of displacement errors (prediction error). This metric is detailed in the ranking experiment section.

3.4. Experiments

In order to validate the proposed system, we perform four analysis: a) binary classification into ”high quality” and ”low quality” presentations, b) multi-class classification into three and four categories of quality, c) analysis of feature selection relevance and classification with different feature subsets, and d) raking of presentations based on quality.

3.4.1. Binary classification

In order to train our model in the binary classification problem we consider a good presentation if its grade (in a scale from 6 to 10, being 10 the greatest grade possible) is greater or equal than 8.0, anything lower than 8.0 is considered as a ”low quality” presentation. First two rows of Table 3 show the results for binary classification using Adaboost and SVM, respectively. Best result is bolded. One can see the high recognition rate of both approaches, automatically splitting the presentations in two quality categories. In the case of SVM, the achieved accuracy is upon 90%.

Table 3. Binary and multi-class classification results

No. of classes	Method	Accuracy	'Bad'	'Average'	'Good'	'Excellent'
2	Adaboost	83.3%	6.0 - 7.9	-	8.0 - 10	-
2	SVM-RBF	**91.66%**	6.0 - 7.9	-	8.0 - 10	-
3	SVM-RBF	75%	6.0 - 7.4	7.5 - 8.4	8.5 - 10	-
3	SVM-RBF	**83.3%**	6.0 - 7.9	8.0 - 8.9	9.0 - 10	-
4	SVM-RBF	**67%**	6.0 - 6.9	7.0 - 7.9	8.0 - 8.9	9.0 - 10

3.4.2. Multi-class classification

In order to increase the set of possible quality categories, we designed the experiment in the multi-class case with three and four qualification groups. For the case of three groups we defined different ranges to split the presentations. The ranges of scores used to defined the three and for quality categories (namely bad, average, good, and excellent) are shown in Table 3. The results applying multi-class one-versus-one SVM in this case are shown in the same table. Best results are bolded. One can see that although the performance is decreased because of the increment in the number of categories, we are able to correlate in a percentage of 83.3% and 67% with the opinions of the teachers for the case of three and four qualification categories respectively.

3.4.3. Feature selection and relevance

We also perform feature selection and weight analysis on the high-level features selected by Adaboost and SVM classifiers in order to analyze their relevance for discriminating among groups of presentations. We focused on the binary classification problem, and for all the iterations of the leave-one-out evaluation we save the alpha weight value assigned for the selected features by Adaboost and the weights assigned by SVM. These values are normalized in order to compute the percentage of relevance of each feature in relation to the rest for each classifier. Results are summarized in Table 4. For each classifier the four features selected with the highest score are bolded. One can see that both classifiers correlate in the relevance of different features. In particular, 'Facing towards' and 'Speaking' is selected with high scores by both classifiers. Additionally, Adaboost gives high scores to the 'Upper agitation' and 'Agitation while speaking', whereas SVM also assigns as relevant the 'Middle agitation' and 'Bottom agitation' indicators. Thus, agitation indicators become relevant for both learning strategies.

Finally, in order to analyze the generalization capability of the most relevant features, we reproduced the binary classification of presentations with subsets of high-level features. Figure 4 shows the results. The first bars corresponds to the previous results using the complete set of nine high-level behavioral indicators. Second and last sets of bars shows the classification results of the leave-one-out experiments when classifiers only consider the subsets of four and two most relevant features based on the analysis shown in Table 4. Note that although the performance is reduced because of the use of a reduced feature set, we are able to correlate upon 80% of the times with the teaching scores only considering the two most discriminate features.

3.4.4. Ranking

The goal of Rank SVM [10] is to predict multivariate or structured outputs. In this case, we use the ground truth grade value of each presentation to generate pairwise preference constraints based on a training set ordered by quality of presentation in descendent order.

Table 4. Percentage of relevance assigned to the high-level features by Adaboost and SVM classifiers

Feature	Meaning	Adaboost	SVM-RBF
1	Facing towards	**21.23%**	**22.47%**
2	Crossed arms	2.12%	4.87%
3	Pointing	1.99%	0.75%
4	Speaking	**20.71%**	**24.21%**
5	Upper agitation	**23.71%**	1.37%
6	Middle agitation	0.77%	**14.89%**
7	Bottom agitation	0.71%	**19.21%**
8	Agitation while speaking	**28.77%**	6.12%
9	Agitation while not speaking	0.00%	6.12%

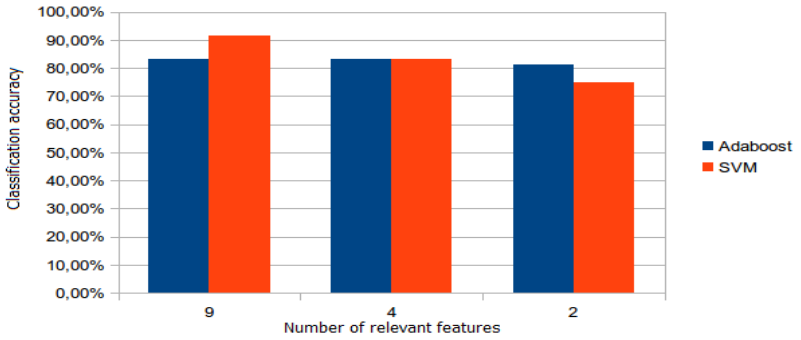

Figure 4. Binary classification with different feature subsets.

For this experiment, we defined different number of splits of our data, namely 3, 5, and 7 fold cross-validation, so that different number of test samples are ordered. In this case, we compute the recognition error E_ε as the ratio in percentage between by how many positions did the classifier failed predicting the correct position and the maximum number of displacement errors, defined as follows:

$$E_\varepsilon = \frac{m}{2(\sum_{i=0}^{n/2-1} N-(2i+1))-N+n} \cdot 100,$$

where m is the number of missed positions, N is total of test samples at each iteration of a K-fold experiment, and n is the number of different scores within the test samples. Then, the classification performance ζ is defined as $\zeta = 100 - E_\varepsilon$. The results of this experiments are shown in Table 5. One can see that for different number of $K \in \{2,3,5\}$, corresponding to rank at each iteration of the fold 12, 8, and 5 test samples respectively, we achieve high recognition rates, approximately in the range of 70%-80% of performance.

Table 5. Ranking of presentation results

2-fold		3-fold		5-fold	
E_ε	ζ	E_ε	ζ	E_ε	ζ
25%	75%	33%	67%	18%	82%

4. Conclusion and future work

We presented an automatic system for categorization of presentations as a e-Learning tool for evaluating the non-verbal communication competence. We performed multimodal human behavior analysis from RGB, depth, and audio data and defined a set of high-level behavior indicators. We recorded a novel data set of oral presentations, and based on the score defined by the experts (teachers in our case) we trained binary, multiclass, and ranking classifiers to evaluate the performance of our system. We analyzed the most discriminative features that correlate to the observers opinion, and achieved classification rates upon 90% categorizing two levels of presentation quality, and upon 80% and 70% classifying the quality of the presentations in three and four groups respectively. The results show the feasibility of our system to be applied as an automatic tool useful for user feedback in training scenarios, as well as for evaluation purposes. As a future work we plan to increase the amount of behavioral patterns with temporal constraints, precise facial expressions, as well as to extend the number of samples so that a precise score for each presentation could be assigned. Finally, we plan to apply the system in real scenarios defining a useful protocol for user feedback and including the framework as a e-Learning tool in the training of non-verbal communication competencies.

References

[1] *Vinciarelli, A.; Salamin, H.; Pantic, M., "Social Signal Processing: Understanding social interactions through nonverbal behavior analysis", CVPR Workshops, pp.42–49, 2009.*

[2] *Olguin, D.O.; Waber, B.N.; et.al., "Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior", SMCB, vol.39, no.1, pp.43–55, 2009.*

[3] *Escalera, S.; Bar, X.; Vitri, J.; Radeva, P.; Raducanu, B., "Social Network Extraction and Analysis Based on Multimodal Dyadic Interaction", Sensors, vol.12, no.2, pp.1702–1719, 2012.*

[4] *Ponce, V.; Gorga, M.; Bar, X.; Escalera, S., "Human Behavior Analysis from Video Data using Bag-of-Gestures", International Joint Conference on Artificial Intelligence, IJCAI, 2011.*

[5] *McCowan, I.; Gatica-Perez, D.; Bengio, S.; Lathoud, G.; Barnard, M.; Zhang, D., "Automatic analysis of multimodal group actions in meetings", PAMI, vol.27, no.3, pp.305–317, 2005.*

[6] *Moattar, M. H.; Homayounpour, M.M.; Kalantari, N.K., "A new approach for robust realtime Voice Activity Detection using spectral pattern", ICASSP, pp.4478–4481, 2010.*

[7] *Friedman, J.; Hastie, T.; Tibshirani, R., "Additive Logistic Regression: a Statistical View of Boosting", nnals of Statistics, vol.29, 1998.*

[8] *Escalera, S.; Pujol, O.; Radeva, P.; Vitrià, J.; Anguera, M.T., "Automatic Detection of Dominance and Expected Interest", EURASIP Journal on Advances in Signal Processing, 2010.*

[9] *Chang, H.; Lin, C., "LIBSVM: a Library for Support Vector Machines", 2001.*

[10] *Joachims, T., "Training Linear SVMs in Linear Time", ACM Data Mining (KDD), 2006.*

[11] *Shotton, J.;, Fitzgibbon, A. W.; Cook, M.; Sharp, T., "Real-time human pose recognition in parts from single depth images", CVPR, pp.1297-1304, 2011.*

[12] *Viola, P.; Jones, M., "Robust Real-Time Face Detection Authors: Paul Viola Microsoft Research", International Journal of Computer Vision, vol.57, no.2, pp.137–154, 2004.*

[13] *Vinciarelli, A.; Pantic, M.; Bourlard, H., "Social Signal Proc.: Survey of an Emerging Domain", 2008.*

[14] *Pan, W.; Dong, W.; Cebrian, M.; Kim, T.; Pentland, A., "Influnce model (dominance) audio: Modeling Dynamical Influence in Human Interaction", MIT Press, 2011.*

[15] *Sanchez-Cortes, D.; Aran, O.; Jayagopi, D.; Schmid Mast, M.; Gatica-Perez, D., "Emergent Leaders through Looking and Speaking: From Audio-Visual Data to Multimodal Recognition", JMUI, 2012.*

[16] *Marcos-Ramiro, A.; Pizarro-Perez, D.; Marron-Romera, M.; Nguyen, L.; Gatica-Perez, D., "Body Communicative Cue Extraction for Conversational Analysis", IEEE Face and Gesture Recognition, 2013.*

[17] *Mohammadi, G.; Vinciarelli, A. "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", IEEE Transactions on Affective Computing, vol.3, no.3, pp.273–284, 2012.*

[18] *Tanaka, H.; Neubig, G.; Tomoki, T.; Campbell, N.; Nakamura, S., "Non-verbal cognitive skills and autistic conditions: An analysis and training tool", Cognitive Infocommunications, pp.41–46, 2012.*